Bayesian and Variational Inference for Reinforcement Learning

Mattie Fellows

St Cross College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2021

Abstract

This thesis explores Bayesian and variational inference in the context of solving the reinforcement learning (RL) problem. Recent advances in developing state-of-the-art algorithms suitable for continuous control introduce regularisation into the reinforcement learning objective. Analysis has revealed that introducing regularisation defines a metaphorical probabilistic inference problem whose solution is used to learn policies with improved exploration properties. Solving this inference problem enables the application of powerful optimisation tools such as variational inference to reinforcement learning (RL). However, existing inference frameworks and their algorithms pose significant challenges for learning optimal policies, e.g., the lack of mode capturing behaviour and difficulties learning deterministic policies. Our first contribution is a theoretically grounded variational inferenence for reinforcement learning (VIREL) framework that utilises a parametrised Q-function to summarise future dynamics of the underlying Markov decision process (MDP), generalising existing approaches. Our framework resolves theoretical issues of existing approaches and an empirical evaluation demonstrate that actor-critic algorithms derived from VIREL outperform state-of-the-art methods based on soft value functions in several domains, even with approximations.

Our analysis reveals that the issues associated with regularised RL stem from the construction of the metaphorical inference problem, which does not capture the uncertainty in the RL problem like a true Bayesian approach would. Moreover, to define and solve this inference problem we must make incorrect assumptions about the Markov decision process (MDP). We show that VIREL resolves these issues by mimicking the adaptive behaviour of a policy that arises from taking a true Bayesian approach to RL. However the type of exploration used by VIREL policies is still suboptimal in comparison to true Bayesian policies is does not capturing uncertainty in the MDP. This motivates taking a Bayesian approach to RL.

The de-facto objective used in RL is frequentist. In this thesis, we argue for starting from the Bayesian objective instead. A key theme of this thesis is that, unlike frequentist approaches, a Bayesian approach to RL adheres to the likelihood principle: in defining the RL objective, frequentist approaches assume exact prior knowledge of the environment which is unrealistic for most applications. In contrast, Bayesian methods characterise uncertainty in the MDP and only condition on knowledge that the agent has, even under approximation. Our analysis reveals that many of the pathologies of frequentist RL stem from breaking the likelihood principle and are immediately resolved by starting with a Bayesian RL objective instead.

Finally, we introduce a novel perspective on Bayesian RL; whereas existing approaches infer a posterior over the transition distribution or Q-function, we characterise the uncertainty in the Bellman operator. Our Bayesian Bellman operator (BBO) framework is motivated by the insight that when bootstrapping is introduced, model-free approaches actually infer a posterior over Bellman operators, not value functions. In this thesis, we use BBO to provide a rigorous theoretical analysis of model-free Bayesian RL to better understand its relationship to established frequentist RL methodologies. We prove that Bayesian solutions are consistent with frequentist RL solutions, even when approximate inference is used, and derive conditions for which convergence properties hold. Empirically, we demonstrate that algorithms derived from the BBO framework have sophisticated deep exploration properties that enable them to solve continuous control tasks at which state-of-the-art regularised actor-critic algorithms fail catastrophically.

Bayesian and Variational Inference for Reinforcement Learning



Mattie Fellows

St Cross College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2021

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) award number OUCL/2016/MGF and has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement number 637713). The experiments were made possible by a generous equipment grant from NVIDIA. My collaborator Anuj Mahajan has received funding by Google DeepMind and the Drapers Scholarship.

Abstract

This thesis explores Bayesian and variational inference in the context of solving the reinforcement learning (RL) problem. Recent advances in developing state-of-the-art algorithms suitable for continuous control introduce regularisation into the reinforcement learning objective. Analysis has revealed that introducing regularisation defines a metaphorical probabilistic inference problem whose solution is used to learn policies with improved exploration properties. Solving this inference problem enables the application of powerful optimisation tools such as variational inference to RL. However, existing inference frameworks and their algorithms pose significant challenges for learning optimal policies, e.g., the lack of mode capturing behaviour and difficulties learning deterministic policies. Our first contribution is a theoretically grounded VIREL framework that utilises a parametrised Q-function to summarise future dynamics of the underlying Markov decision process (MDP), generalising existing approaches. Our framework resolves theoretical issues of existing approaches and an empirical evaluation demonstrate that actor-critic algorithms derived from VIREL outperform state-of-the-art methods based on soft value functions in several domains, even with approximations.

Our analysis reveals that the issues associated with regularised RL stem from the construction of the metaphorical inference problem, which does not capture the uncertainty in the RL problem like a true Bayesian approach would. Moreover, to define and solve this inference problem we must make incorrect assumptions about the MDP. We show that VIREL resolves these issues by mimicking the adaptive behaviour of a policy that arises from taking a true Bayesian approach to RL. However the type of exploration used by VIREL policies is still suboptimal in comparison to true Bayesian policies is does not capturing uncertainty in the MDP. This motivates taking a Bayesian approach to RL.

The de-facto objective used in RL is frequentist. In this thesis, we argue for starting from the Bayesian objective instead. A key theme of this thesis is that, unlike frequentist approaches, a Bayesian approach to RL adheres to the likelihood principle: in defining the RL objective, frequentist approaches assume exact prior knowledge of the environment which is unrealistic for most applications. In contrast, Bayesian methods characterise uncertainty in the MDP and only condition on knowledge that the agent has, even under approximation. Our analysis reveals that many of the pathologies of frequentist RL stem

from breaking the likelihood principle and are immediately resolved by starting with a Bayesian RL objective instead.

Finally, we introduce a novel perspective on Bayesian RL; whereas existing approaches infer a posterior over the transition distribution or Q-function, we characterise the uncertainty in the Bellman operator. Our Bayesian Bellman operator (BBO) framework is motivated by the insight that when bootstrapping is introduced, model-free approaches actually infer a posterior over Bellman operators, not value functions. In this thesis, we use BBO to provide a rigorous theoretical analysis of model-free Bayesian RL to better understand its relationship to established frequentist RL methodologies. We prove that Bayesian solutions are consistent with frequentist RL solutions, even when approximate inference is used, and derive conditions for which convergence properties hold. Empirically, we demonstrate that algorithms derived from the BBO framework have sophisticated deep exploration properties that enable them to solve continuous control tasks at which state-of-the-art regularised actor-critic algorithms fail catastrophically.

Contents

List of Figures **xi**

1 Introduction **1**

1.1 A Motivating Example 2

1.2 Roadmap: A Bayesian Agenda 3

1.2.1 Pathologies of Frequentist Methods 3

1.2.2 A Case for Bayesian RL 5

1.2.3 New Perspectives on Bayesian RL 6

1.3 AI Safety and Uncertainty 7

2 Mathematical Preliminaries **9**

2.1 Measures and Probability Spaces 9

2.2 Function Spaces 11

2.3 Convergence of Random Variables 12

2.3.1 Pointwise and Uniform Convergence 12

2.3.2 Almost Sure Convergence 13

2.3.3 Convergence of Probability Measures 14

2.4 An Important Theorem 17

3 Reinforcement Learning **21**

3.1 The Reinforcement Learning Problem 22

3.2 Value Functions and Bellman Equations 23

3.3 A Taxonomy of RL 26

3.3.1 Planning vs Learning 26

3.3.2 Model-based vs Model-Free 29

3.3.3 Value Based vs Policy Based 31
3.3.4 Bayesian vs Frequentist RL 33
3.4 The Exploration/Exploitation Dilemma 34
3.4.1 Residual Stochasticity 37
3.4.2 Dithering Exploration 37
3.4.3 Uncertainty Driven Exploration 38
3.5 Practical Policy Evaluation 39
3.5.1 TD Learning 39
3.5.2 Direct Methods 41
3.5.3 Convergence of Policy Evaluation 43
3.6 A Guide to this Thesis 51

4 Characterising Uncertainty 53
4.1 Bayesian Inference 53
4.2 Likelihood Principle 55
4.3 Approximate Inference 58
4.3.1 Variational Inference 59
4.3.2 Marginal Likelihood and Expectation Maximisation 60
4.3.3 Randomised Priors 61

5 Control as Inference 65
5.1 Existing Approaches 66
5.1.1 Maximum Marginal Likelihood as Control 67
5.1.2 Posterior Inference as Control 71
5.1.3 A Counterexample for MERL Optimal Policies 74
5.2 VIREL 76
5.2.1 The VIREL Posterior 77
5.2.2 Comparing MERL with VIREL 80
5.2.3 The VIREL Objective 81
5.3 VIREL for Control 83
5.3.1 Variational Actor-critic 83

5.3.2 Empirical Evaluation 86
5.4 Fundamental Issues with RL as Inference 90
5.4.1 Pathologies of RL as Inference 90
5.4.2 Should we just be Bayesian? 91

6 Bayesian Reinforcement Learning 93
6.1 The Bayesian RL Objective 93
6.1.1 Bayes-optimal Policies are Admissible 96
6.1.2 Bayes-optimal Policies Explore Optimally 99
6.2 Practical Bayesian Reinforcement Learning 103
6.2.1 Model-based Bayesian RL 103
6.2.2 Posterior Sampling and Optimism 105
6.2.3 Model-free Bayesian RL 107
6.3 Bayesian RL and the Likelihood Principle 110
6.3.1 Bayesian vs Frequentist RL 112

7 Bayesian Bellman Operators 115
7.1 Uncertainty in the Bellman Operator 116
7.2 BBO Model 118
7.2.1 Consistency of the Posterior 123
7.2.2 Gaussian BBO 125
7.2.3 Control as Inference from a Bayesian Perspective 128
7.3 Approximate BBO 129
7.3.1 Connection to Frequentist TD Learning 129
7.3.2 Randomised Priors BBO 131
7.3.3 Convergence Analysis 135
7.3.4 Empirical Evaluation of Analysis 138
7.4 BBO for Control 141
7.4.1 Bayesian Bellman Q-learning 142
7.4.2 Bayesian Bellman Actor-critic 145
7.4.3 Continuous Control Experiments 149
7.5 Insights for Future Research 153

8 Conclusion: A Bayesian Manifesto **155**

Appendices

A Probabilistic Measure Theory **161**

A.1 Probability Spaces 161

A.2 Random Variables 163

B Proofs **167**

B.1 Proof of Theorem 5 167

B.2 Proof of Theorem 6 173

B.2.1 A Frequentist Analysis of Theorem 6 177

C Linear BBO **179**

C.1 The Linear Model 179

C.2 Deriving a Least-Squared Algorithm 180

C.3 Recovering TDC/GTD2 182

C.4 Linear Policy Evaluation Experiments 183

C.4.1 Hyperparameters 186

D Experimental Details **187**

D.1 VIREL Continuous Control Experiments 188

D.2 BBO Policy Evaluation Experiments 188

D.2.1 Tsitsiklis' Triangle Counterexample 188

D.2.2 Neural Network Function Approximators 190

D.3 BBO Continuous Control Experiments 193

D.3.1 BBAC MountainCar-Continuous-v0 194

D.3.2 BBAC cartpole 195

Bibliography **197**

List of Figures

1.1 Mushroom bandit problem 2

2.1 Sketch of Theorem 1. 17

3.1 Taxonomy of RL algorithms 27
3.2 Value iteration (left) vs policy iteration (right) 28
3.3 Alpine exploration MDP 35

4.1 A simple inference problem 53
4.2 Variational inference 59
4.3 Marginal likelihood 60

5.1 Sketch of a variational distribution learnt using a marginal likelihood vs RL objective 71
5.2 A discrete counterexample MDP for MERL optimal policies 75
5.3 Graphical models for MERL and VIREL 80
5.4 MuJoCo continuous control experiments for VIREL. 88

6.1 Graphical model of Bayesian RL. 110

7.1 Graphical Model for BBO. 118
7.2 Sketch of Tangent Cones in Ω 136
7.3 Tsitsiklis Counterexample. 138
7.4 Nonlinear policy evaluation results. 140
7.5 Schematic of RP-BBAC. 147
7.6 Continuous control with sparse reward. 151

7.7 Diagnostics throughout learning of BBAC. (Top) Expected Q-values for each ensemble member ($L = 8$) over environment states $(position, velocity)$. (Bottom) State visitation. 152

7.8 Diagnostics throughout learning of SAC. (Top) Ensemble values over environment state $(position, velocity)$. (Bottom) State visitation. 152

7.9 Diagnostics throughout learning of BAC. (Top) Expected Q-values for each ensemble member ($L = 8$) over environment states $(position, velocity)$. (Bottom) State visitation. 152

C.1 Linear policy evaluation results. 184

D.1 Tsitsiklis' triangle MDP. 189

D.2 Evaluation of BBAC's sensitivity to randomized prior hyperparameters. We vary (a) ensemble size L, (b) prior scale σ, and (c) regularization weight λ, while keeping other hyperparameters fixed. 195

Abbreviations and Acroynms

ELBO evidence lower bound.

MERL maximum entropy reinforcement learning.

VIREL variational inferenence for reinforcement learning.

BAMDP Bayes adaptive Markov decision process.

BBO Bayesian Bellman operators.

EM expectation maximisation.

KL Kullback–Leibler.

LML log-marginal likelihood.

MAP maximum a posteriori.

MDP Markov decision process.

MML maximum marginal likelihood.

MSBBE mean squared Bayesian Bellman error.

MSBE mean squared Bellman error.

MSE mean squared error.

MSPBE mean squared projected Bellman error.

ODE ordinary differential equation.

RL reinforcement learning.

RP randomised priors.

SAC soft actor-critic.

TD temporal difference.

VI variational inference.

1 Introduction

Reinforcement learning (RL) is appealing because it provides a framework for understanding how intelligent, autonomous behaviour can emerge in an environment with minimal or no prior knowledge. When we reflect on ourselves, we recognise that our capability for intelligent behaviour emerges from our abilities to specify desires, plan over long timescales to achieve those desires, form and update beliefs about the world we live in based our experience, incorporate existing knowledge into our beliefs and be curious about new experiences that we believe could be rewarding or provide information about the world so as to better understand our desires and how to achieve them. It is remarkable that these capabilities emerge from beings whose nervous systems are composed of less than a few kilograms of matter and have relatively modest material requirements in the form of food, water and shelter.

When understanding the challenges faced by intelligent agents, we must recognise that they never have complete knowledge of their environment. All of the assets listed above rely on a capability to reason over uncertainty. For intelligent behaviour to emerge, it is therefore essential that an agent can characterise and reason over their uncertainty and adapt their behaviour accordingly. If this were not the case, the best an agent can do is to respond in a pre-determined way without adapting its behaviour in response to any new observations. We cannot deem this behaviour as intelligent.

1.1 A Motivating Example

To illustrate the necessity for intelligent agents to characterise uncertainty, we introduce a simple bandit decision problem in Fig. 1.1. Consider an amateur mycologist who is interested in foraging for a rare mushroom. To forage for the mushroom, they usually visit a forest to the east of their home where they know that the probability of finding the mushroom is 0.5.

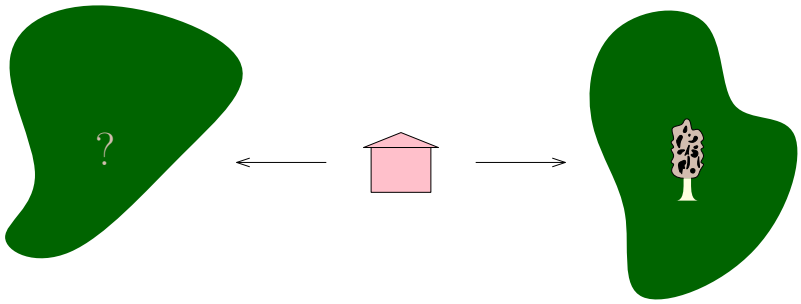


Figure 1.1: Mushroom bandit problem

One day, the mycologist is made aware of a forest to the west of their home. They must chose to visit one of the two forests. We represent this decision by the action set $\{\leftarrow, \rightarrow\}$: if the agent chooses $\leftarrow$, they travel to the forest west of their home and if they chose $\rightarrow$, they travel to the forest east of their home. If they find the mushroom, they receive a reward of $+1$ or else they receive as reward of 0.

Clearly the mycologist should always take the action that maximises their expected reward, however they cannot compute the expected reward for taking action $\leftarrow$ as they are uncertain a priori about the probability that they will find a mushroom there. Without characterising and reasoning over this uncertainty, the mycologist can only take a sequence of actions in a pre-determined way. For example: visit the east forest on Tuesdays and the west forest on Thursdays. As this behaviour is arbitrary we cannot call it intelligent.

In reality, the mycologist reasons over their uncertainty in the environment. They may have very strong beliefs that it is not worth exploring the action $\leftarrow$ (for example, if they know that in comparison to a universal average, the probability 0.5 of finding this mushroom extraordinarily high or they know that the west forest does not contain trees that the mushroom is likely to grow with), which would inform them to ignore action $\leftarrow$.

Otherwise, by choosing the action $\leftarrow$, they will make observations that give new them information about the probability of finding the mushroom in the west forest.

In all cases, the mycologist reasons over their existing prior knowledge to infer their belief over the probability of finding a mushroom in the west forest. Once these beliefs have been inferred, they weigh up the expected benefits of each action and choose the one they believe is best. If this action provides new information, then that information updates their prior beliefs and the process continues. The mycologist's behaviour is thus adaptive, as their future actions adapt to their changing beliefs which reflects any new information they receive from observations.

1.2 Roadmap: A Bayesian Agenda

Our mushroom bandit example from Fig. 1.1 motivates the need for a unified framework that enables us to make optimal decisions under uncertainty. Despite this, at present, the overwhelming majority of research in RL takes a frequentist objective as a foundation which assumes complete knowledge of the environment. In this thesis, we intend to highlight the inherent pathologies of that arise from attempting to solve the frequentist RL objective when an agent is uncertain about its environment. We then make both theoretical and practical arguments for considering a Bayesian approach instead before proposing a new methodology for carrying out Bayesian RL in the model-free setting.

1.2.1 Pathologies of Frequentist Methods

In Chapter 3 we review existing frequentist approaches for solving the RL problem. A key observation is that the frequentist RL objective can only be optimised if the agent has complete knowledge of the environment. For this reason, the frequentist RL objective is a planning objective that is not suitable for learning. Despite this, the majority of research in RL develops methods that attempt to solve this objective when the agent is uncertain about the environment's reward function and transition distribution.

When developing algorithms under the frequentist paradigm, researchers are haunted by the inherent assumptions needed to optimise an objective that assumes knowledge of

unknown quantities. This leads to a number of pathologies including the double sampling problem [Baird, 1995], algorithms that are provably divergent [Tsitsiklis and Van Roy, 1997] and overestimation of Q-functions [Hasselt et al., 2016, Fujimoto et al., 2018a]. The most emblematic pathology of frequentist RL is the exploration/exploitation dilemma [Sutton and Barto, 2018]. As the agent is faced with uncertainty, it must balance choosing actions to learn a model of the unknown dynamics (or random variables thereof) with optimising the objective as best it can using its incomplete knowledge of the environment. A frequentist epistemology gives no guide on how to address balance. Battling these pathologies through heuristics is a major focus of frequentist RL approaches.

In Chapter 5 we study the control as inference framework [Levine, 2018a], which is a framework for generating state-of-the-art regularised algorithms by finding a solution to an equivalent metaphorical inference problem. The goal of these approaches is to make powerful optimisation tools such as variational inference applicable for solving RL problem. These approaches also introduce a regularisation term into the frequentist RL objective, which encourages exploration by penalising premature collapse of a policies to a deterministic Dirac-delta.

Despite the framework's name, control as inference is frequentist by nature as the inference procedure does not characterise uncertainty in the environment. We reveal two pathologies in these approaches: the capacity for risk-seeking behaviour and the potential for learning policies that are far from optimal due to bias being introduced into the RL objective. We propose VIREL, a theoretically grounded probabilistic inference framework that resolves these issues and derive a family of actor-critic methods including a scheme for adaptive exploration. An empirical evaluation demonstrates that actor-critic algorithms from this family outperform state-of-the-art methods based on soft value functions [Haarnoja et al., 2018a] in several domains.

An analysis reveals that the heuristics introduced by the VIREL framework mean that policies eventually tend to a Dirac-delta distribution in the limit increasing samples from the environment and rate of this collapse is regulated by a measure of optimality. VIREL optimal policies thus approximate some of the properties of a Bayesian approach, namely

a Bernstein von-Mise theorem [Vaart, 1998] and the introduction of an adaptive scheme for exploration. A conclusion we draw is that if the success of methods under the control as inference paradigm is determined by how well they approximate certain properties that are natural to a Bayesian approach, then why not start with a Bayesian approach in the first place?

1.2.2 A Case for Bayesian RL

In this thesis we make a case for taking a Bayesian approach as foundation for solving the RL problem. A Bayesian epistemology treats probability as a degree of belief in an unknown variable [Savage, 1954] and captures uncertainty in an intuitive and principled way. Bayesian approaches infer a posterior over random variable(s), which characterises the agent's uncertainty in their value given observations from the environment. We introduce Bayesian statistics in Chapter 4 where we formalise the notion of making decisions under uncertainty in our mushroom bandit example.

In Chapter 6 we show that Bayesian RL clarifies the goal of learning optimal policies under uncertainty. Whilst we are not the first to consider a Bayesian approach to RL [Bellman, 1956, Duff, 2002], there exist deep connections between Bayesian inference and making decisions under uncertainty that have been largely ignored by the RL community. Our goal in Chapter 6 is to provide an exposition and clarification of these results in the context of solving the RL problem. In particular, we show that many of Bayesian RL's benefits of stem from the likelihood principle. This implies the conditionality principle holds, which asserts that Bayesian methods should only condition on observed data and not reason over hypothetical or future observations. As a posterior is used to marginalise over all possible values according to the belief in their value, a Bayesian RL objective can be derived that doesn't rely on quantities that are unknown to the agent a priori. Many of the frequentist pathologies arise because this principle is broken because the frequentist RL objective assumes complete knowledge of the environment. We provide a novel analysis in Section 6.3 showing that even under tractable approximation, Bayesian RL approaches still retain the conditionality principle.

It is well known that optimal policies under a Bayesian objective optimally balance exploration with exploitation: in many cases they are referred to as a gold-standard for exploration [Osband et al., 2018, Zintgraf et al., 2020]. Agents explore to reduce uncertainty in the environment, but only in so far as that exploration leads to greater expected returns under their current belief. As well as formalising and clarifying this notion of optimality, we also apply established results from decision theory to show that these policies are admissible [Wald, 1950]. Admissibility justifies the introduction of a prior and states that any Bayesian optimal policy cannot be pareto improved upon by any other non-Bayesian agent given the prior over beliefs. In particular, we use these insights to show that frequentist approaches that use Monte Carlo sampling of returns implicitly assume a Dirichlet process prior. This prior is a particularly poor choice for solving the RL problem under uncertainty as it actively directs the agent towards trajectories that it has already encountered, encouraging pure exploitation.

The likelihood principle also provides justification for the derivation of model-free Bayesian RL. We provide a formal argument to show that characterising uncertainty over any random variable that is sufficient to solve the RL problem provides the same information regardless of the random variable. Consequently, inferring a posterior over a Q-function provides the same information for solving the Bayesian RL objective as inferring a posterior over the transition dynamics and reward, provided the underlying model of the environment is consistent. This argument blurs the lines between established frequentist notions of model-free and model-based RL.

1.2.3 New Perspectives on Bayesian RL

The centrepiece of this thesis is the development of a novel framework for Bayesian RL in Chapter 7. Whereas existing approaches infer a posterior over the transition distribution or Q-function, we characterise the uncertainty in the Bellman operator. Our Bayesian Bellman operator framework is motivated by the insight that when bootstrapping is introduced, model-free approaches actually infer a posterior over Bellman operators, not value functions. The likelihood principle justifies characterising the uncertainty in the Bellman operator as it is sufficient for solving the RL problem. We provide a rigorous

theoretical analysis of model-free Bayesian RL to better understand its relationship to established frequentist RL methodologies, proving that Bayesian solutions are consistent with frequentist RL solutions, even when approximate inference is used, and derive conditions for which convergence properties hold.

Moreover, we show that by taking a Bayesian approach we forfeit nothing as many of the algorithms that fix the pathologies inherent in frequentist approaches using heuristics arise naturally from our model-free Bayesian RL framework. In particular, we can recover a family of temporal-difference learning algorithms from a maximum-a-posteriori point estimate of our posterior. These algorithms avoid the double sampling problem whilst being provably convergent. In addition, our analysis reveals that many existing pathologies associated with the maximum entropy formulation of the control as inference framework arise because they make incorrect a priori assumptions about the environment that cannot be corrected by observations.

As calculating and marginalising using the exact posterior is typically intractable, we introduce approximate inference and posterior sampling to find an approximate solution to the Bayesian RL objective. Empirically, we demonstrate that algorithms derived from the approximate BBO framework still have sophisticated deep exploration properties that enable them to solve continuous control tasks at which state-of-the-art regularised actor-critic algorithms from the RL a inference framework fail catastrophically.

1.3 AI Safety and Uncertainty

Many of the benefits afforded by Bayesian methods play a significant role in a wide range of real-world applications of RL where identifying decisions that are being made under high uncertainty is critical. Incorporating prior knowledge can have substantial impact on speed and accuracy, for example in medical applications of RL [Yu et al., 2019]. Furthermore, safe exploration and convergence guarantees are crucial in many physical-world domains such as robotics or autonomous vehicles [Kiran et al., 2020], where undesirable actions and algorithmic divergence could cause direct human harm

or even fatalities. In these situations, it is crucial that systems identify decisions that are being made under high uncertainty so that they can hand control over to an expert.

It is inevitable that investment bankers will always look to fields in applied statistics such as artificial intelligence when developing algorithms for trading. Financial applications of deep RL [Deng et al., 2017] have the potential to destabilise economies. For this reason, precise uncertainty quantification can enable safer applications of RL for trading [Jiang et al., 2017, Spooner et al., 2018]. By promoting a Bayeisan approach to RL and providing the foundations of practical methods to quantify uncertainty, we hope that those working in finance might adopt a Bayesian approach to minimise the risks associated with their activities.

Finally, we remark that although our contributions in this thesis are primarily theoretical, stronger theoretical guarantees can increase confidence in algorithms and can accelerate the implementation of RL for real-world applications. The societal effects of rapidly increased automation are uncertain and disputed [Smith and Anderson, 2017]. Positive effects include claims of increased job creation, productivity [Acemoglu and Restrepo, 2018] and even techno-utopianism [Srnicek and Williams, 2015]. A critical approach to these claims is important, especially given the gulf between the ideals and the reality of what automation has achieved so far [Greenfield, 2018] and the potential automation has for increased inequality [Acemoglu and Restrepo, 2020] and discrimination [Mehrabi et al., 2019].

2

Mathematical Preliminaries

2.1 Measures and Probability Spaces

By taking a more rigorous measure theoretical approach to probability in this thesis we gain access to powerful and indispensable tools for analysis. We also avoid having to distinguish between discrete, continuous and mixed state and/or action spaces, meaning that we don't have to give a separate treatment for each case: in particular, by using measure theory we avoid the cumbersome linear algebra associated with the state-transition matrix that is used to prove theoretical results in discrete Markov decision processes in favour of more concise and general results. For readers not familiar with measure theory, we provide a brief introduction in Appendix A.

We denote a probability spaces as the triple (O, Σ, P) where O is the sample space of outcomes, Σ is a σ-algebra of events and P is a probability measure on (O, Σ). We use the notation $P(\mathcal{E}) \coloneqq \int_{\mathcal{E}} dP(o)$ to define the probability of an event $\mathcal{E} \in \Sigma$ according to the distribution P. To ease notation, we write integrals over the whole sample space as $\int$ where the specific sample space will be clear from context, for example, for an O-valued variable, we write $\int = \int_O$ and we write integrals with respect to the real line as $\int = \int_{\mathbb{R}}$.

We write $Q \ll P$ if a measure Q is absolutely continuous with respect to a measure P, that is $Q(A) = 0$ for all $A \in \{A' : P(A') = 0\}$. Equality of two measures $P = Q$ implies that $P \ll Q$ and $Q \ll P$. We write the (n-dimensional) Lebesgue measure as λ, the counting measure as $\#$ and the density of the distribution P using the lower case $p(o) \coloneqq \frac{dP(o)}{d\lambda(o)}$ for continuous distributions and $p(o) \coloneqq \frac{dP(o)}{d\#(o)}$ for discrete distributions[1]. The dimensionality of λ will be clear from context.

Any function $f : \mathcal{O} \to \mathbb{R}$ is P-integrable if

$$\int |f(o)| dP(o) < \infty.$$

Random variables are measurable functions of outcomes $X : O \to \mathcal{X}$ which induce a probability space $(\mathcal{X}, \Sigma_{\mathcal{X}}, P_X)$ where $\mathcal{X} = X(\mathcal{O}) \coloneqq \{X(o) : o \in \mathcal{O}\}$ is the image of $\mathcal{O}$, $X^{-1}(\mathcal{E}) \coloneqq \{o \in \mathcal{O} : X(o) \in \mathcal{E}\} \in \Sigma$ is the pre-image of any $\mathcal{E} \in \Sigma_{\mathcal{X}}$ and the distribution P_X of X is the *pushforward* of P through X. For any $\mathcal{E} \in \Sigma_{\mathcal{X}}$ and function $f : \mathcal{X} \to$ such that $f \circ X$ is P-integrable, the pushforward P_X satisfies:

$$\int_{X^{-1}(\mathcal{E})} f \circ X(o) dP(o) = \int_{\mathcal{E}} f(x) dP_X(x), \tag{2.1}$$

If P_X is known, we refer to $P(o)$ as the *pullback* measure of P_X through X[1]. When carrying out analysis of the random variable X, Eq. (2.1) assures us that it suffices to consider the probability space $(\mathcal{X}, \Sigma_{\mathcal{X}}, P_X)$ defined by the variable using the distribution P_X rather than relating the analysis back to the original probability space (O, Σ, P). We write the expectation of the random variable X as:

$$\mathbb{E}_P[X] \coloneqq \int x dP_X(x) = \int X(o) dP(o),$$

where the second equality follows from Eq. (2.1).

We denote the Dirac delta measure at a point $a \in O$ as δ_a which has property:

$$\int_A f(o) d\delta_a(o) = \begin{cases} f(a), & a \in A, \\ 0, & a \notin A. \end{cases}$$

[1]The density is the same as the probability mass function from elementary probability theory in the discrete setting

2.2 Function Spaces

We denote a parametric function using the subscript notation $f_\omega : \mathcal{X} \to \mathbb{R}$ parametrised by a vector $\omega \in \Omega$. An intuitive interpretation of the parameter vector ω is that it indexes a particular function in the function space, for example in neural networks the parameter vector is the set of weights associated with each layer and each $\omega \in \Omega$ represents a specific instantiation of the network. We make occasional use of the space of linear functions $f_\omega = \omega^\top v$ where $v : \mathcal{X} \to \mathbb{R}^n$ is a basis vector and $\Omega \subseteq \mathbb{R}^n$.

We write the p-norm of a vector space $\Omega \subseteq \mathbb{R}^n$ as:

$$\|\omega\|_p \coloneqq \left(\sum_{i=1}^{n} |\omega_i|^p\right)^{\frac{1}{p}},$$

which formalises the notion of the distance of the vector from the origin. The Euclidean norm can be recovered by choosing $p = 2$ and we define the ∞-norm as:

$$\|\omega\|_\infty \coloneqq \max_{i \in [1:n]} \omega_i.$$

We also introduce the matrix norm, which for any $M \times A$ matrix A is defined as:

$$\|A\|_2^2 \coloneqq \sup_{x \in \mathbb{R}^M} \frac{\|Ax\|_2^2}{\|x\|_2^2}.$$

We can generalise the principles of the p-norm to function spaces using the Lebesgue norm. We make use of the L^2 Lebesgue norm, which we define with respect to measure m for a function $f : \mathcal{X} \to \mathbb{R}$ as:

$$\|f\|_m \coloneqq \left(\int f(x)^2 dm(x)\right)^{\frac{1}{2}}.$$

Consider two functions: $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{X} \to \mathbb{R}$. We can use the L^2 Lebesgue norm to introduce a notion of distance between functions:

$$\|f - g\|_m = \left(\int \left(f(x) - g(x)\right)^2 dm(x)\right)^{\frac{1}{2}}.$$

Whenever $\|f - g\|_m = 0$, we have that $f = g$ m-almost everywhere. This means that $f = g$ except possibly over sets that have measure zero according to m, i.e.:

$$\|f - g\|_m = 0 \iff \int_E \left(f(x) - g(x)\right) dm(x) = 0 \; \forall \; E \in \{E' : \; m(E') \neq 0\}.$$

We note that for parametrised functions, there may be different parametrisations which lead to m-almost everywhere equality, that is there may exists two (or more) parameters $\omega_1, \omega_2 \in \Omega$ such that $\|f_{\omega_1} - f_{\omega_2}\|_m = 0$. This is typically the case for neural networks where there often exists several or even infinite modes of symmetry over the network weights.

Finally, we can extend the notion of the ∞-norm to function spaces. In this thesis, we only require the ∞-norm for functions $f : \mathcal{X} \to \mathbb{R}$ where $\mathcal{X}$ is compact and f is bounded and upper semi-continuous. Under these conditions, the ∞-norm is defined as the maximum value of f on $\mathcal{X}$:

$$\|f\|_\infty \coloneqq \max_{x \in \mathcal{X}} |f(x)|.$$

Whenever $\|f - g\|_\infty = 0$, we have that $f(x) = g(x)$ for all $x \in \mathcal{X}$, which means that $f = g$ *pointwise*.

2.3 Convergence of Random Variables

We now characterise and introduce notation for the different modes of convergence of a sequence of random variables $\{X_N : \mathcal{O} \to \mathcal{X}\}_{N \in \mathbb{Z}^+}$ as $N \to \infty$.

2.3.1 Pointwise and Uniform Convergence

A sequence of random variables converges pointwise to a random variable $X : \mathcal{O} \to \mathcal{X}$ on $\mathcal{O}' \subseteq \mathcal{O}$ if $\{o \in \mathcal{O}' : \lim_{N \to \infty} X_N(o) = X(o)\} = \mathcal{O}'$. We denote pointwise convergence as:

$$X_N \to X.$$

For some specific cases where real-valued random variables are used, we need to strengthen pointwise convergence to uniform convergence by ensuring that for every scalar $\varepsilon > 0$, a sequence of random variables remain uniformly bounded within a margin ε of their limiting value after a fixed number of steps in the sequence. Uniform convergence for real valued functions is formally defined as:

Definition 1 (Uniform Convergence)**.** *A sequence* $\{X_N : \mathcal{O} \to \mathbb{R}\}_{N \in \mathbb{Z}^+}$ *of real-valued random variables uniformly to* $X : \mathcal{O} \to \mathbb{R}$ *on* $\mathcal{O}' \subseteq \mathcal{O}$ *if for every* $\varepsilon > 0$ *there exists some natural number* K *such that for all* $o \in \mathcal{O}'$ *and* $N \geq K$*,*

$$|X_N(o) - X(o)| < \varepsilon.$$

An equivalent definition of uniform converge is:

$$\lim_{N \to \infty} \|X_N - X\|_\infty = 0.$$

We denote uniform convergence as:

$$X_N \xrightarrow{\text{unif}} X.$$

2.3.2 Almost Sure Convergence

Instead of requiring that a random variable converges on all points in the outcome space, we can ignore convergence on events that have measure zero according to P as they will never occur. We then say that a sequence of random variables converges P-almost surely if convergence holds over a set with probability 1, that is:

$$P(\{o \in \mathcal{O} : \lim_{N \to \infty} X_N(o) = X(o)\}) = 1 \tag{2.2}$$

We denote P-almost sure convergence as:

$$X_N \xrightarrow[P-\text{a.s.}]{} X.$$

The definition of P-almost sure convergence in Eq. (2.2) restricts pointwise convergence to sets that have measure 1. Just like in Section 2.3.1, we can strengthen this notion of convergence by requiring uniform convergence over sets that have measure 1, which we define as uniform P-almost sure convergence. We only require uniform P-almost sure convergence for real-valued random variables, which is formally defined as:

Definition 2 (Uniform P-almost Sure Convergence)**.** *A sequence* $\{X_N : \mathcal{O} \to \mathbb{R}\}$ *of real-valued random variables converges uniformly to* $X : \mathcal{O} \to \mathbb{R}$ $P_{\mathcal{D}}$*-almost surely if for every* $\varepsilon > 0$ *there exists some natural number* K *such that for all* $o \in \mathcal{O}$ *and* $N \geq K$*,*

$$|X_N(o) - X(o)| < \varepsilon,$$

except possibly a some subset $\mathcal{O}' \subset \mathcal{O}$ such that $P(\mathcal{O}') = 0$. An equivalent definition of uniform P-almost sure converge is:

$$P\left(\left\{\mathcal{O}' : \lim_{N\to\infty} \|X_N - X\|_\infty = 0\right\}\right) = 1.$$

We denote uniform P-almost sure convergence as:

$$X_N \xrightarrow[P-\text{a.s.}]{\text{unif}} X.$$

2.3.3 Convergence of Probability Measures

Instead of characterising the convergence of random variables themselves, we can provide conditions for which their distributions converge instead. The weak convergence of a distribution is a collection equivalent conditions known as the *Portmaneau Theorem*:

Proposition 1 (Portmaneau Theorem)**.** *A sequence of distributions $\{P_{X,N}\}_{N\in\mathbb{Z}^+}$ over a measurable space $(\mathcal{X}, \Sigma_\mathcal{X})$ converges weakly to a distribution P_X if either:*

i $\lim_{N\to\infty} \mathbb{E}_{P_{X,N}}[f] = \mathbb{E}_{P_X}[f]$ *for all uniformly continuous f.*

ii $\limsup_{N\to\infty} P_{X,N}(A) \leq P_X(A)$ *for all closed $A \in \mathcal{X}$.*

iii $\liminf_{N\to\infty} P_{X,N}(A) \geq P_X(A)$ *for all open $A \in \mathcal{X}$.*

iv $\lim_{N\to\infty} P_{X,N}(A) = P_X(A)$ *for all P_X-continuity sets*[2] *$A \in \mathcal{X}$.*

Proving one of these conditions holds implies that all of these conditions hold.

Proof. This result is central to probabilistic measure theory and is established is almost all texts on the subject. See, for example, Billingsley [1999] Theorem 2.1. □

We denote weak convergence as:

$$P_{X,N} \rightharpoonup P_X.$$

Remark. *A key property of almost sure convergence is that it implies weak convergence (see, for example, Chung and Zhong [2001] Theorem 4.1.2) however the converse is not true. For this reason, almost sure convergence is sometimes called strong convergence.*

[2] A continuity set of distribution P_X is any Borel set $A \in \mathcal{X}$ whose boundary has measure zero according to $P_X : P_X(\partial A) = 0$

In this thesis we are also concerned with the convergence of conditional distributions. In particular, we study the convergence properties of a posterior distribution $P_{X,N}(\cdot|\mathcal{D})$ which defines a probability measure over $(\mathcal{X}, \Sigma_{\mathcal{X}})$ that conditions on data $\mathcal{D} \sim P_{\mathcal{D}}$. As the data is a random variable, we combine two modes of convergence to characterise the overall convergence of the measure:

Definition 3 (Weak $P_{\mathcal{D}}$-almost sure convergence:)**.** *A sequence of conditional distributions $P_{X,N}(\cdot|\mathcal{D})$ converges weakly to P_X $P_{\mathcal{D}}$-almost surely if:*

$$P_{\mathcal{D}}\left(\{\mathcal{D}' : P_{X,N}(\cdot|\mathcal{D}) \rightharpoonup P_X\}\right) = 1.$$

We denote weak $P_{\mathcal{D}}$-almost sure convergence as:

$$P_{X,N}(\cdot|\mathcal{D}) \xrightarrow[P_{\mathcal{D}}-\text{a.s}]{} P_X.$$

Finally, a stronger notion of convergence can be obtained by introducing a statistical distance measure between distributions. For illustrative purposes, consider two distributions P and Q defined over a measurable space $(\mathcal{X}, \Sigma_{\mathcal{X}})$ with densities p and q respectively. In this thesis, we use the Kullback–Leibler (KL) divergence:

$$\text{KL}(P \parallel Q) \coloneqq \int \log\left(\frac{p(x)}{q(x)}\right) dP(x),$$

and the total variational distance:

$$\text{TV}(P \parallel Q) \coloneqq \sup_{\mathcal{E}\in\Sigma} |P(\mathcal{E}) - Q(\mathcal{E})|.$$

Both the KL divergence and total variational distance enjoy the following properties, which justify their use as a measure of distance between two probability measures:

Proposition 2 (Properties of Statistical Distances)**.** *Let D be either the KL or total variational distance then:*

i $D(P \parallel Q) = 0 \implies P = Q$,

ii $D(P \parallel Q) \geq 0$.

Moreover, the total variational distance is a true metric, which has the properties:

iii $TV(P \parallel Q) = TV(Q \parallel P)$,

iv $TV(P \parallel Q) \leq TV(Q \parallel P) + TV(P \parallel R)$.

Proof. See Rachev [1991]. □

Definition 4 (Convergence in Statistical Distance)**.** *Let D be a statistical distance. A sequence $P_{X,N}$ of distributions over a measurable space $(\mathcal{X}, \Sigma_{\mathcal{X}})$ converges in a statistical distance D to P_X if:*

$$\lim_{N \to \infty} D(P_{X,N} \parallel P_X) = 0.$$

We denote convergence in KL divergence as:

$$P_{X,N} \xrightarrow{\text{KL}} P_X,$$

and convergence in total variational distance as:

$$P_{X,N} \xrightarrow{\text{TV}} P_X.$$

We can relate convergence in KL divergence to the total variational distance and weak convergence via the following proposition:

Proposition 3 (Relationship Between Convergence of Measures)**.** *Convergence in KL divergence implies convergence in total variational distance which implies weak convergence, that is for sequence $P_{X,N}$ of distributions and P_X defined over a measurable space $(\mathcal{X}, \Sigma_{\mathcal{X}})$:*

$$P_{X,N} \xrightarrow{KL} P_X \implies P_{X,N} \xrightarrow{TV} P_X \implies P_{X,N} \rightharpoonup P_X.$$

Proof. The proof of $P_{X,N} \xrightarrow{\text{KL}} P_X \implies P_{X,N} \xrightarrow{\text{TV}} P_X$ can be shown via Pinkster's inequality:

$$0 \leq \text{TV}(P_{X,N} \parallel P_X) \leq \sqrt{\frac{1}{2}\text{KL}(P_{X,N} \parallel P_X)},$$

hence $\lim_{N \to \infty} \text{KL}(P_{X,N} \parallel P_X) = 0 \implies \lim_{N \to \infty} \text{TV}(P_{X,N} \parallel P_X) = 0$ as required.

We now prove that $P_{X,N} \xrightarrow{\text{TV}} P_X \implies P_{X,N} \rightharpoonup P_X$. As the total variational distance is a metric over probability measures, we see from Proposition 2 property i that $\lim_{N \to \infty} \text{TV}(P_{X,N} \parallel P_X) = 0$ implies $P_X = P_{X,N}$, and hence properties i to iv of Proposition 1 trivially hold, satisfying the conditions necessary for weak convergence. □

2.4 An Important Theorem

The technique of introducing bounds on $P_N(\cdot|y)$ in Theorem 1 came from a discussion with Piotr Miłoś, who suggested this approach to solve a similar problem.

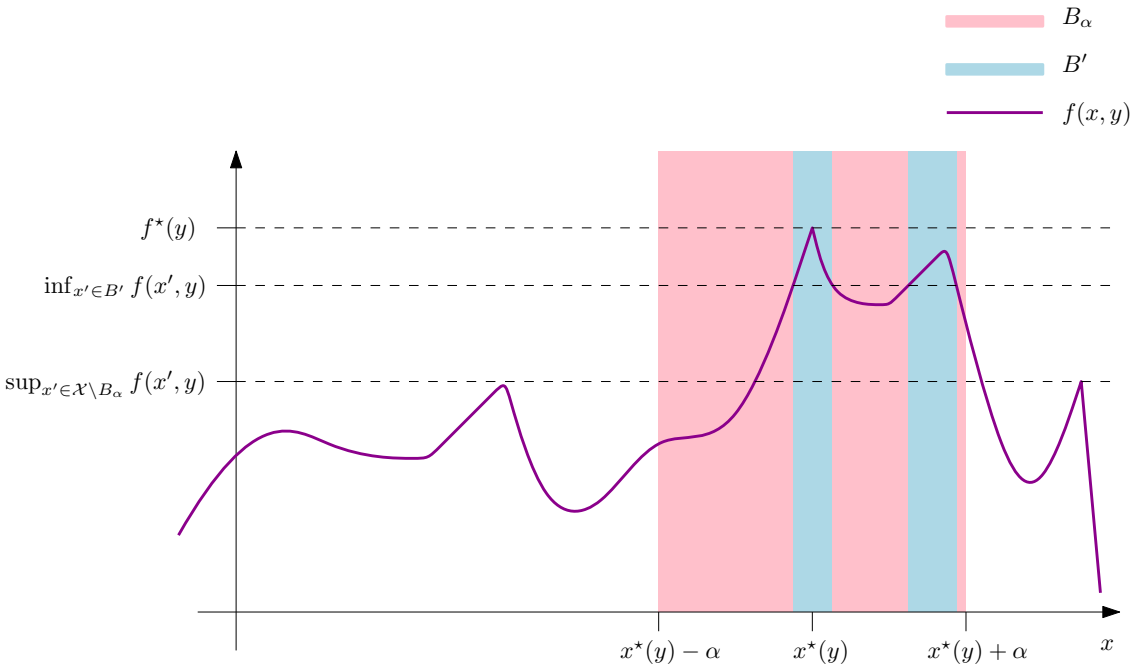


Figure 2.1: Sketch of Theorem 1.

We now prove a theorem that is key for establishing many theoretical results in this thesis. The theorem's proof may be tricky to follow for those not familiar with advanced real analysis and measure theory, so we provide an intuitive sketch in Fig. 2.1.

Theorem 1 (An Important Convergence Theorem)**.** *Let $(\mathcal{Y}, \Sigma_\mathcal{Y}, P_\mathcal{Y})$ and $(\mathcal{X}, \Sigma_\mathcal{X}, P_{\mathcal{X}|\mathcal{Y}})$ be probability spaces where $P_{\mathcal{X}|\mathcal{Y}}(\cdot|y)$ is a conditional distribution. Let $f_N : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a sequence of real-valued functions that converge to f and $\epsilon_N \geq 0$ be a sequence of scalars such that $\lim_{N\to\infty} \epsilon_N = 0$. Assume that $P_\mathcal{Y}$-almost surely:*

i $(\mathcal{X}, d_\mathcal{X})$ *is a compact metric space and $P_{\mathcal{X}|\mathcal{Y}}$ has support $\mathcal{X}$;*

ii $f_N \xrightarrow[P_{\mathcal{X}|\mathcal{Y}}-a.s.]{} f$. *and f_N and f are bounded and continuous in $\mathcal{X}$ for all N and;*

iii f *has a unique maximum $x^\star(y)$ over $\mathcal{X}$, that is $x^\star(y) = \arg\max_{x'\in\mathcal{X}} f(x', y)$ where $|x^\star(y)| = 1$*

Then the probability measure $P_N(\cdot|y)$ defined as:

$$P_N(A|y) = \frac{\int_A \exp\left(\frac{f_N(x,y)}{\epsilon_N}\right) dP_{\mathcal{X}|\mathcal{Y}}(x|y)}{\int \exp\left(\frac{f_N(x,y)}{\epsilon_N}\right) dP_{\mathcal{X}|\mathcal{Y}}(x|y)},$$

converges weak $P_{\mathcal{Y}}$-almost surely to a Dirac-delta measure centred on $x^\star(y)$, that is:

$$P_N(\cdot|y) \xrightarrow[P_{\mathcal{Y}}-a.s]{} \delta_{x^\star(y)}.$$

Proof. Using the portmanteau lemma from Proposition 1, it suffices to show that:

$$\limsup_{N\to\infty} P_N(A|y) \leq \delta_{x^\star(y)}(A), \tag{2.3}$$

for any closed set $A \in \mathcal{X}$ $P_{\mathcal{Y}}$-almost surely. Let E' be any closed set such that $x^\star(y) \in E'$. Eq. (2.3) trivially holds for E' as $\delta_{x^\star(y)}(E') = 1$ and $P_N(A'|y) \leq 1$ for all N from Kolmogorov's axioms. We are therefore left to prove:

$$\limsup_{N\to\infty} P_N(E|s) \leq \delta_{x^\star(y)}(E) = 0, \quad P_{\mathcal{Y}} - a.s.$$

where E is any closed set such that $x^\star(y) \neq E$.

As f_N and f are bounded, there exists a scalar constant $C^\star$ such that $f < C^\star$ and $f_N < C^\star$ for all N . Using $C^\star$, we can re-write $P_N(E|y)$ as:

$$P_N(E|y) = \frac{\int_E \exp\left(\frac{f_N(x,y)-C^\star}{\epsilon_N}\right) dP_{\mathcal{X}|\mathcal{Y}}(x|y)}{\int \exp\left(\frac{f_N(x,y)-C^\star}{\epsilon_N}\right) dP_{\mathcal{X}|\mathcal{Y}}(x|y)}. \tag{2.4}$$

We make two observations. Firstly, $f_N < C^\star$ implies that $\exp\left(\frac{f(x,y)-C^\star}{\epsilon_N}\right)$ and $\exp\left(\frac{f_N(x,y)-C^\star}{\epsilon_N}\right)$ are bounded for all N. Secondly, $\lim_{N\to\infty} \epsilon_N = 0$ and $f_N \xrightarrow[P_{\mathcal{X},\mathcal{Y}}-\text{a.s.}]{} f$ imply:

$$\exp\left(\frac{f_N(x,y) - C^\star}{\epsilon_N}\right) \xrightarrow[P_{\mathcal{X},\mathcal{Y}}-\text{a.s.}]{} \lim_{N\to\infty} \exp\left(\frac{f(x,y) - C^\star}{\epsilon_N}\right) = 0.$$

These observations are sufficient conditions for applying the dominated convergent theorem to Eq. (2.4), hence

$$\begin{aligned}
\lim_{N\to\infty} P_N(E|y) &= \frac{\int_E \lim_{N\to\infty} \exp\left(\frac{f(x,y)-C^\star}{\epsilon_N}\right) dP_{\mathcal{X}|\mathcal{Y}}(x|y)}{\int \lim_{N\to\infty} \exp\left(\frac{f(x,y)-C^\star}{\epsilon_N}\right) dP_{\mathcal{X}|\mathcal{Y}}(x|y)}, \\
&= \lim_{N\to\infty} \frac{\int_E \exp\left(\frac{f(x,y)-C^\star}{\epsilon_N}\right) dP_{\mathcal{X}|\mathcal{Y}}(x|y)}{\int \exp\left(\frac{f(x,y)-C^\star}{\epsilon_N}\right) dP_{\mathcal{X}|\mathcal{Y}}(x|y)}, \\
&= \lim_{N\to\infty} \frac{\int_E \exp\left(\frac{f(x,y)}{\epsilon_N}\right) dP_{\mathcal{X}|\mathcal{Y}}(x|y)}{\int \exp\left(\frac{f(x,y)}{\epsilon_N}\right) dP_{\mathcal{X}|\mathcal{Y}}(x|y)}.
\end{aligned} \tag{2.5}$$

We now bound the numerator and denominator in Eq. (2.5) to establish $\lim_{N\to\infty} P_N(E|s) = 0$. Define $f^\star(y) \coloneqq f(x^\star(y), y)$. Consider a small open ball of radius $\alpha > 0$ centred on $x^\star(y)$:

$$B_\alpha \coloneqq \{x : d_{\mathcal{X}}(x, x^\star(y)) < \alpha\}.$$

We claim that $\sup_{x'\in\mathcal{X}\setminus B_\alpha} f(x', y) < f^\star(y)$ for any $\alpha > 0$. To verify this, we use a proof by contradiction, assuming that $\sup_{x'\in\mathcal{X}\setminus B_\alpha} f(x', y) = f^\star(y)$. As $\mathcal{X} \setminus B_\alpha$ is compact and f is continuous, $f^\star(y) \in f(\mathcal{X} \setminus B_\alpha)$ hence there exists some $x' \in \mathcal{X} \setminus B_\alpha$ such that $f(x') = f^\star(y)$. This contradicts the definition of $f^\star(y)$, which is a unique maximum at $x^\star(y) \in B_\alpha$.

As $E \subset \mathcal{X}$ is closed and does not contain $x^\star(y)$, there exists some α such that $E \subseteq \mathcal{X}\setminus B_\alpha$, and hence $\sup_{x'\in E} f(x', y) \leq \sup_{x'\in\mathcal{X}\setminus B_\alpha} f(x', y)$. We can use this to upper bound the numerator of Eq. (2.5):

$$\begin{aligned}
\int_E \exp\left(\frac{f(x,y)}{\epsilon_N}\right) dP_{\mathcal{X}|\mathcal{Y}}(x|y) &\leq \int_E \exp\left(\frac{\sup_{x'\in E} f(x',y)}{\epsilon_N}\right) dP_{\mathcal{X}|\mathcal{Y}}(x|y), \\
&= \exp\left(\frac{\sup_{x'\in E} f(x',y)}{\epsilon_N}\right) P_{\mathcal{X}|\mathcal{Y}}(E|y), \\
&\leq \exp\left(\frac{\sup_{x'\in\mathcal{X}\setminus B_\alpha} f(x',y)}{\epsilon_N}\right) P_{\mathcal{X}|\mathcal{Y}}(E|y),
\end{aligned}$$

As f is continuous, there exists some open set $B' \subset B_\alpha$ such that $\inf_{x'\in B'} f(x', y) > \sup_{x'\in\mathcal{X}\setminus B_\alpha} f(x', y)$, hence we can lower bound the denominator of Eq. (2.5):

$$\begin{aligned}
\int \exp\left(\frac{f(x,y)}{\epsilon_N}\right) dP_{\mathcal{X}|\mathcal{Y}}(x|y) &\geq \int_{B'} \exp\left(\frac{f(x,y)}{\epsilon_N}\right) dP_{\mathcal{X}|\mathcal{Y}}(x|y), \\
&\geq \int_{B'} \exp\left(\frac{\inf_{x'\in B'} f(x',y)}{\epsilon_N}\right) dP_{\mathcal{X}|\mathcal{Y}}(x|y), \\
&= \exp\left(\frac{\inf_{x'\in B'} f(x',y)}{\epsilon_N}\right) P_{\mathcal{X}|\mathcal{Y}}(B'|y).
\end{aligned}$$

Using these lower and upper bounds, we can upper bound $P_N(E|y)$ as:

$$P_N(E|y) \leq \frac{P_{\mathcal{X}|\mathcal{Y}}(E|y)}{P_{\mathcal{X}|\mathcal{Y}}(B'|y)} \exp\left(\frac{\sup_{x'\in\mathcal{X}\setminus B_\alpha} f(x',y) - \inf_{x'\in B'} f(x',y)}{\epsilon_N}\right).$$

As $\inf_{x' \in B'} f(x', y) > \sup_{x' \in \mathcal{X} \setminus B_\alpha} f(x', y)$ and $\lim_{N \to \infty} \epsilon_N = 0$, taking the limit $N \to \infty$ establishes our result:

$$\lim_{N \to \infty} P_N(E|y) = 0, \quad P_{\mathcal{Y}} - a.s.$$

which follows immediately from $\lim_{N \to \infty} P_N(E|y) = 0 \implies \limsup_{N \to \infty} P_N(E|y) = 0$. □

Although Theorem 1 may seem abstract, it has powerful consequences depending on the semantics of $P_N(\cdot|y)$: when $P_N(\cdot|y)$ is a Boltzmann distribution, Theorem 1 formally establishes its limiting $\max$-operator like behaviour as its temperature tends to zero. When $P_N(\cdot|y)$ is a posterior, Theorem 1 establishes Bernstein-von-Mises type theorems [Vaart, 1998, Kleijn and van der Vaart, 2012], proving that Bayesian and frequentist approaches are consistent. If $P_N(\cdot|y)$ is an adaptive policy that depends on samples from the environment, then Theorem 1 can be used to show the emergence of deterministic behaviour from stochasticity in the limit of increasing environment samples.

3
Reinforcement Learning

In reinforcement learning (RL), we are concerned with the problem of an agent learning to behave optimally in an environment. We assume that the agent is in sequential decision making process known as a Markov decision process (MDP) [Szepesvári, 2010, Puterman, 2014, Sutton and Barto, 2018]. Formally, an MDP is a tuple: $\langle \mathcal{S}, \mathcal{A}, r, P(\cdot|s,a), P_0(S), \gamma \rangle$ where $\mathcal{S}$ is the set of states in the agent's environment, $\mathcal{A}$ is the set of actions available to the agent, $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the reward function, $P(\cdot|s,a)$ is the state transition distribution, P_0 is the initial state distribution and $\gamma \in [0,1)$ is the discount factor. In this chapter we will assume all distributions have densities for the sake of exposition so that our presentation can be easily related to standard probabilistic notation in the RL literature, however this need not be the case - for example it is possible that distributions such as Dirac delta measures can represent policies, transition distributions and initial state distributions

At each timestep t, the agent is in state $s_t \in \mathcal{S}$ and chooses an available action according to a policy $a_t \sim \pi(\cdot|s_t)$. We overload notation slightly by denoting both the density and distribution (should it exist) of the policy as π. The space of policies is Π. When deriving policy gradient algorithms, we will assume the policy space is a parametric space of distributions $\Pi \coloneqq \{\pi_\theta : \theta \in \Theta\}$ parametrised by a metric space (Θ, d_Θ). Given an action $a_t \in \mathcal{A}$, the agent transitions to the next state in the environment according to the state

transition distribution $s_{t+1} \sim P(\cdot|s_t, a_t)$ and receives a scalar reward for its behaviour $r_t \coloneqq r(s_{t+1}, a_t, s_t)$. The agent starts in an initial state according to $s_0 \sim P_0(\cdot)$ and its policy induces a trajectory of state-actions: $\tau_N = (s_0, a_0, s_1, a_1, ...a_{N-1}, s_N) \in \mathcal{T}_N \coloneqq \mathcal{S} \times \bigtimes_{i=1}^{N}(\mathcal{S} \times \mathcal{A})$ with the distribution: $P_N^\pi(T_N)$ and density:

$$p_N^\pi(\tau_N) \coloneqq p_0(s_0) \prod_{t=0}^{N-1} \pi(a_t|s_t)p(s_{t+1}|s_t, a_t).$$

We will make use of the state-action transition distribution: $P^\pi(\cdot|s, a)$ which satisfies:

$$dP^\pi(s', a'|s, a) = dP(s'|s, a)d\pi(a'|s').$$

Finally, we introduce the distribution ρ_N^π which characterises the probability of being in state $s \in \mathcal{S}$ after N timesteps for an agent following a policy π:

$$d\rho_N^\pi(s) \coloneqq \int\int dP^\pi(s|s_N, a_N)d\pi(a_N|s_N)dP_N^\pi(\tau_N). \tag{3.1}$$

3.1 The Reinforcement Learning Problem

In this thesis, we will be concerned with the infinite horizon discounted setting, and our focus will be on control. To ensure that our problem is well defined, we assume the standard regularity conditions:

Assumption 1 (RL Regularity Conditions)**.** *i) rewards are bounded with* $|r(s', a, s)| \leq r_{max} < \infty$ *for all* $(s', a, s) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$*; ii)* $(\mathcal{S}, d_\mathcal{S})$ *and* $(\mathcal{A}, d_\mathcal{A})$ *are compact metric spaces; and either: iii a)* $\mathcal{A}$ *is discrete; or iii b)* $\mathbb{E}_{P(s'|s,a)}[r(s', a, s)]$ *is upper semi-continuous on* $\mathcal{A}$ *for any* $s \in \mathcal{S}$ *and* $P(B|s, a)$ *is continuous on* $\mathcal{A}$ *for any* $s \in \mathcal{S}$ *and Borel subset* $B \in \mathcal{S}$*.*

Assumption 1 is fundamental to all results in this thesis, hence we assume it holds hereafter without further reference. For most applications in RL, it suffices to use compact Euclidean spaces or finite discrete spaces for our state and action spaces, which satisfy Assumption 1 $ii)$. In verifying Assumption 1 $iiib)$, it is often easier to verify the more restrictive condition of continuity, which implies upper semi-continuity by

definition. We discuss Borel subsets and other aspects of probabilistic measure theory in Appendix A.

The goal of the agent is to find an optimal policy $\pi^\star$ that maximises the expected discounted *return*, that is the total expected discounted reward in the limit of $N \to \infty$. We can formalise this notion by defining the RL objective that any optimal policy will maximise:

$$J^\pi \coloneqq \lim_{N\to\infty} \mathbb{E}_{P_N^\pi}\left[\sum_{t=0}^{N-1} \gamma^t r_t\right]. \tag{3.2}$$

Using the formula for the summation of a geometric series, we can show that the discounted return is bounded under Assumption 1:

$$\left|\sum_{t=0}^{\infty} \gamma^t r_t\right| \leq \frac{r_{\max}}{1-\gamma}.$$

In addition, it is well known that for a Markov chain defined by an initial state distribution and sequence of transition distributions that there exists a unique probability measure for the infinite sequence: $P_\mathcal{T}^\pi \coloneqq \lim_{N\to\infty} P_N^\pi$ that agrees with the finite-dimensional distributions (see Theorem 3 in Chapter 22 and Definition 10 in Chapter 21 of Fristedt and Gray [1997]). It therefore follows that the limit in Eq. (3.2) exists and we require no further special considerations to discuss infinite length trajectories: $\tau \coloneqq \lim_{N\to\infty} \tau_N \in \mathcal{T}$ whose underlying probability space: $(\mathcal{T}, \Sigma, P_\mathcal{T}^\pi)$ is well defined where $\mathcal{T} \coloneqq \lim_{N\to\infty} \mathcal{T}_N$ is the infinite Cartesian product of the state-action space. Assumption 1 also guarantees the existence of an optimal policy for the RL problem:

Proposition 4 (Existence of Optimal Policy, Puterman)**.** *Under Assumption 1, an optimal stationary policy $\pi^\star \in \arg\max_{\pi\in\Pi} J^\pi$ exists for the MDP*

Proof. The result is established in Theorems 6.2.10 and 6.2.12 of Puterman [2014]. □

3.2 Value Functions and Bellman Equations

Many algorithms and analyses in reinforcement learning rely on the ability to characterise the expected return of a policy for any state $s \in \mathcal{S}$ of the MDP through a *value*

function:

$$V^{\pi}(s) \coloneqq \mathbb{E}_{P_{\mathcal{T}}^{\pi}}\left[\sum_{t=0}^{\infty}\gamma^{t}r_{t}\middle|s_{0}=s\right].$$

Finding a value function for a particular policy is known as *policy evaluation* and is an important goal in RL in itself. In the control setting, we require that an agent in state $s \in \mathcal{S}$ can characterise the expected return of a policy as a function of its actions $a \in \mathcal{A}$ through an *action-value* function:

$$Q^{\pi}(s,a) \coloneqq \mathbb{E}_{P_{\mathcal{T}}^{\pi}}\left[\sum_{t=0}^{\infty}\gamma^{t}r_{t}\middle|s_{0}=s,a_{0}=a\right].$$

We refer to action-value functions as Q-functions in this thesis as they can be used to compare the *quality* of each action that an agent can take in terms of the expected return whilst following a policy π. We can relate the value function to the Q-function by taking expectations over actions using the policy:

$$V^{\pi}(s) = \int Q^{\pi}(s,a)d\pi(a|s).$$

To learn Q-functions, we will make use of the Bellman operator for a policy π, which takes a function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as its input:

$$\mathcal{B}[f](s,a) \coloneqq \int_{\mathcal{S}\times\mathcal{A}} \left(r(s',a,s) + \gamma f(s',a')\right) dP^{\pi}(a',s'|s,a).$$

Any Q-function satisfies a Bellman equation:

$$\mathcal{B}[Q^{\pi}] = Q^{\pi}.$$

To solve the reinforcement learning problem, our goal is to find an optimal policy $\pi^{\star}$. We therefore pay special attention to optimal Q-function: $Q^{\star} \coloneqq Q^{\pi^{\star}}$ and the optimal Bellman operator:

$$\mathcal{B}^{\star}[f](s,a) \coloneqq \int_{\mathcal{S}} \left(r(s',a,s) + \gamma \max_{a'\in\mathcal{A}} f(s',a')\right) dP(s'|s,a).$$

The optimal Q-function satisfies an optimal Bellman equation $\mathcal{B}^{\star}[Q^{\star}] = Q^{\star}$. There always exists a deterministic optimal policy $\delta_{a^{\star}(s)}$ where $a^{\star}(s) \in \arg\max_{a\in\mathcal{A}} Q^{\star}(a,s)$ is an arbitrary element of the set of optimal actions for each state: if this were not the case, it would be possible to find a policy π' with an action a' such that $Q^{\pi'}(s,a') > Q^{\star}(s,a^{\star}(s))$ for some $s \in \mathcal{S}$, which contradicts the definition of $Q^{\star}$. A key property of Bellman operators is that they are contraction mappings in the ∞-norm:

Proposition 5 (Contraction of Bellman Operators)**.** *The Bellman operator $\mathcal{B}[\cdot]$ and optimal Bellman operator $\mathcal{B}^\star[\cdot]$ are contraction mappings in the ∞-norm.*

Proof. We start by proving the contraction mapping of $\mathcal{B}[\cdot]$. To prove this result, consider any two bounded, continuous functions: $f_1 : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and $f_2 : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. From the definition of the Bellman operator, it follows:

$$\begin{aligned}
\|\mathcal{B}[f_1] - \mathcal{B}[f_2]\|_\infty &= \|\mathbb{E}_{P^\pi}\left[r + \gamma f_1\right] - \mathbb{E}_{P^\pi}\left[r + \gamma f_2\right]\|_\infty, \\
&= \|\gamma \mathbb{E}_{P^\pi}\left[f_1 - f_2\right]\|_\infty, \\
&\leq \|\gamma \max_{(s',a') \in \mathcal{S}\times\mathcal{A}} \left(f_1(s',a') - f_2(s',a')\right)\|_\infty, \\
&\leq \|\gamma \max_{(s',a') \in \mathcal{S}\times\mathcal{A}} |f_1(s',a') - f_2(s',a')|\,\|_\infty, \\
&= \gamma \|f_1 - f_2\|_\infty,
\end{aligned}$$

as required. For the optimal Bellman operator, a similar analysis holds:

$$\begin{aligned}
\|\mathcal{B}^\star[f_1] - \mathcal{B}^\star[f_2]\|_\infty &= \left\|\mathbb{E}_P\left[r + \gamma \max_{a' \in \mathcal{A}} f_1(a', \cdot)\right] - \mathbb{E}_P\left[r + \gamma \max_{a' \in \mathcal{A}} f_2(a', \cdot)\right]\right\|_\infty, \\
&= \left\|\mathbb{E}_P\left[\gamma\left(\max_{a' \in \mathcal{A}} f_1(a', \cdot) - \max_{a' \in \mathcal{A}} f_2(a', \cdot)\right)\right]\right\|_\infty.
\end{aligned}$$

As f_1 and f_2 are bounded, there exists some $|C| < \infty$ such that $f_1 \geq C$ and $f_2 \geq C$, hence:

$$\begin{aligned}
\|\mathcal{B}^\star[f_1] - \mathcal{B}^\star[f_2]\|_\infty &= \left\|\mathbb{E}_P\left[\gamma\left(\max_{a' \in \mathcal{A}} f_1(a', \cdot) - C - \left(\max_{a' \in \mathcal{A}} f_2(a', \cdot) - C\right)\right)\right]\right\|_\infty, \\
&= \left\|\mathbb{E}_P\left[\gamma\left(\max_{a' \in \mathcal{A}} \left(f_1(a', \cdot) - C\right) - \max_{a' \in \mathcal{A}} \left(f_2(a', \cdot) - C\right)\right)\right]\right\|_\infty, \\
&= \left\|\mathbb{E}_P\left[\gamma\left(\max_{a' \in \mathcal{A}} |f_1(a', \cdot) - C| - \max_{a' \in \mathcal{A}} |f_2(a', \cdot) - C|\right)\right]\right\|_\infty, \\
&\leq \left\|\mathbb{E}_P\left[\gamma\left|\max_{a' \in \mathcal{A}} |f_1(a', \cdot) - C| - \max_{a' \in \mathcal{A}} |f_2(a', \cdot) - C|\right|\right]\right\|_\infty,
\end{aligned}$$

where the penultimate line follows from $f_1 - C \geq 0$ and $f_2 - C \geq 0$. Applying the reverse triangle inequality yields:

$$\begin{aligned}
\|\mathcal{B}^\star[f_1] - \mathcal{B}^\star[f_2]\|_\infty &\leq \left\|\gamma \mathbb{E}_P\left[\max_{a' \in \mathcal{A}} |f_1(a', \cdot) - C - f_2(a', \cdot) + C|\right]\right\|_\infty, \\
&= \left\|\gamma \mathbb{E}_P\left[\max_{a' \in \mathcal{A}} |f_1(a', \cdot) - f_2(a', \cdot)|\right]\right\|_\infty, \\
&\leq \left\|\gamma \max_{(s',a') \in \mathcal{S}\times\mathcal{A}} |f_1(s', a') - f_2(s', a')|\right\|_\infty, \\
&= \gamma \|f_1 - f_2\|_\infty,
\end{aligned}$$

as required. □

3.3 A Taxonomy of RL

There exist several approaches to solving the RL problem according to which attributes of the MDP are available to the agent, the type of MDP the agent is in and whether or not we privilege data or computational efficiency. A schematic is shown in Fig. 3.1 which classifies RL algorithms according to three key attributes, where example algorithms are shown in each category, which we now review:

3.3.1 Planning vs Learning

If the agent has access to the MDP, we can solve the control problem or carry out policy evaluation without ever interacting with the environment. In this situation, solving an RL problem is known as *planning*. Perhaps the most well-known approach to planning is dynamic programming. We demonstrated in Proposition 5 that Bellman operators are contraction mappings, hence from Banach's fixed point theorem, iteratively applying a Bellman operator for a policy π to an initial estimate Q_0 recovers the true Q-function, that is $\lim_{k\to\infty} Q_k = Q_\pi$ where $Q_{k+1} = \mathcal{B}[Q_k]$. Dynamic programming refers to a collection of methods that exploit this insight to solve the RL problem using the MDP's state-action transition distribution and reward function.

If the state-action space is finite, then each Q_k can be represented as a vector in $\mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$, where each element represents $Q^k(s,a)$. In this case, the expectation under P^π required by the Bellman operator $\mathcal{B}[\cdot]$ can be carried out exactly by multiplying by the $|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}\times\mathcal{A}|$ state-action transition matrix where each element corresponds to a particular $p^\pi(s',a'|s,a)$. In the more general case, the expectation may not have an analytic solution, and so some degree of estimation (for example, by a Monte Carlo method) is required to evaluate the expectations in the Bellman operator.

We have already demonstrated that it is always possible to construct an optimal deterministic policy from the optimal Q-function, hence if we can learn $Q^\star$ then we have solved the RL problem. Approaches that learn $Q^\star$ directly by iteratively applying the

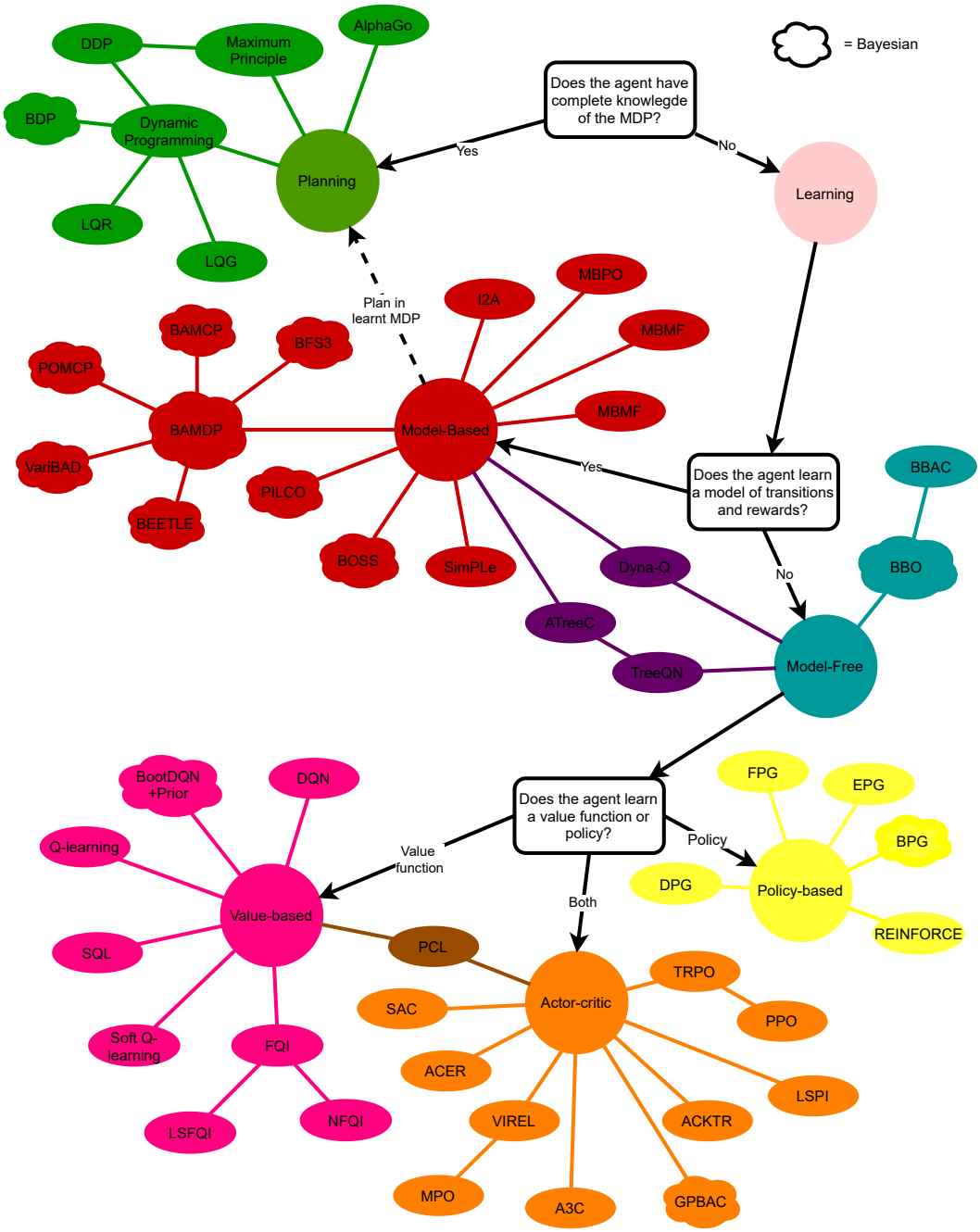


Figure 3.1: Taxonomy of RL algorithms

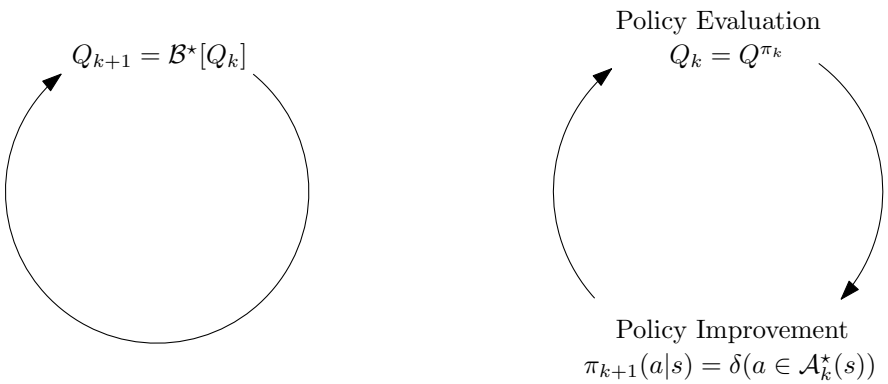


Figure 3.2: Value iteration (left) vs policy iteration (right)

optimal Bellman operator $\mathcal{B}^\star[\cdot]$ fall under the category of *value iteration*. Alternatively, approaches known as *policy iteration* carry out policy evaluation for a particular policy π_k to learn its Q-function Q^{π_k}. A *policy improvement* step then generates a better policy by constructing a policy that always takes an action that maximises the new Q-function: $\pi_{k+1}(\cdot|s) = \delta_{\mathcal{A}_k^\star(s)}$ where $\mathcal{A}_k^\star(s) \in \arg_{a' \in \mathcal{A}} \max Q^{\pi_k}(s, a')$ is an arbitrary selection of any element of $\arg_{a' \in \mathcal{A}} \max Q^{\pi_k}(s, a')$. A comparison of value iteration and policy iteration is shown schematically in Fig. 3.2.

In the policy iteration scheme introduced above, policy evaluation and policy improvement are carried out exactly and alternately. We can conceive of a more general scheme, known as *generalised policy iteration*, where policy evaluation and improvement can interact independently of each other and don't have to be carried out until completion. Generalised policy iteration is a powerful concept in RL and unifies the two approaches of policy iteration and value iteration into a single framework: each value iteration step $Q_{k+1} = \mathcal{B}^\star[Q_k]$ can be interpreted as carrying out a policy improvement step (via the optimal Bellman operator's use of the max operator) and then a applying single iteration of policy evaluation via the improved policy's Bellman operator. In contrast, finding $Q_K = Q^{\pi_k}$ in policy iteration's policy evaluation step involves iteratively applying the Bellman operator $\mathcal{B}^{\pi_k}[\cdot]$ until convergence. Only then is policy improvement carried out.

Although dynamic programming has limited practical value due to its high computational cost and assumption of an exact model of the MDP, generalised policy iteration serves as a guiding theoretical ideal for many algorithms, even when the state-transition or reward functions are unknown a priori: almost all algorithms in RL fit into the generalised policy iteration framework, with the distinction between then being how the policy evaluation and policy improvement steps are carried out and what approximations are used [Sutton and Barto, 2018]. We refer to algorithms that solve the RL problem without complete knowledge of the MDP as *learning*, which reflects the necessity of the agent to learn some aspect of the MDP to find an optimal policy.

A second family of planning algorithms is based around tree search: here, the agent builds a tree where the root node is typically the current state that the agent is in and each node represents a subsequent state of the MDP. The goal is to estimate the Q-function for the tree's nodes using samples of returns. At each iteration, the agent traverses the current tree structure by taking an action according to a tree-policy whilst maintaining a log of rewards. When an agent reaches a node that has been expanded - that is, there are still actions available in that state that the agent has yet to take - it selects a random action from the set of unexplored actions and transitions to the new state, adding a new leaf node to the tree that represents this state. A Monte Carlo simulation is then started from this new leaf node and the return is used to update the Q-function estimates of all parent nodes traversed. We will not be concerned with tree-search methods in this thesis and have included their description for completeness. We refer interested readers to Browne et al. [2012] for an introduction to these approaches and Świechowski et al. [2021] for a survey of state-of-the-art tree search algorithms.

3.3.2 Model-based vs Model-Free

In the more general and realistic situation where an agent does not have a complete model of the MDP, solving the RL problem involves learning one or more unknown quantities in the MDP. Learning algorithms can be further categorised into two branches depending upon what is learnt: in *model-based* approaches the agent learns a model of the transition distribution P and/or the reward function r directly from environment samples. Once a

model of the environment has been learnt, the agent must still learn an optimal policy. The RL problem then reduces to a planning problem in the learnt MDP. As samples from the MDP are used to learn the reward function and transition distribution directly, model-based approaches are highly sample efficient as an arbitrary number of samples can be drawn from the learnt model without interacting with the environment at any point during the planning process to learn an optimal policy and/or value function.

Model-based methods are particularly suitable for MDPs with deterministic transition distributions where learning the exact transition dynamics reduces to learning a mapping $\mathcal{S} \times \mathcal{A} \to \mathcal{S}$, which can be represented by a neural network. Unfortunately, learning high dimensional, multi-modal conditional distributions that may be typical of more realistic MDPs quickly becomes intractable [Song et al., 2013]. Whilst simpler, tractable models such as Gaussians can still be used, the optimal policy obtained from planning in the learnt model can be far from optimal, especially when the transition distribution is multi-modal. Hence designing model-based RL algorithms that can scale to environments with complex transition distributions and continuous/large state-action spaces remains a key challenge.

Instead of learning a model of the MDP, in *model-free* approaches the agent uses samples from the environment to estimate expectations required to solve a Bellman equation or optimise the reinforcement learning objective directly. This approach avoids the issues associated with learning the transition dynamics, allowing model-free algorithms to scale to larger domains at the expense of increased sample inefficiency: rather than learning a fixed distribution that can be sampled from to solve the planning problem, samples have less persistence as they are thrown away after updating the value functions or policy of interest in an iteration step. The boundary between model-based and model-free reinforcement learning becomes blurred however when replay buffers are used [Vanseijen and Sutton, 2015]: here, samples from the MDP are stored in a data set and are reused for model-free algorithms, improving the sample efficiency. Indeed, a replay buffer can be understood as a non-parametric analogue of a tabular model that grows with increasing data [Lin, 1992, Moerland et al., 2020].

3.3.3 Value Based vs Policy Based

Within the class of model-free RL algorithms, we can make two further distinctions that mirror the distinction between policy improvement and value iteration in planning.

Methods that attempt to learn value or Q-functions are known as *value-based methods*. Most value-based approaches use principles from Q-learning [Watkins, 1989, Watkins and Dayan, 1992, Mnih et al., 2015]. In Q-learning, the expectations required to evaluate the optimal Bellman operator in value iteration algorithms are estimated with samples from the environment, which requires the use of a $\max$ operator over the Q-function estimate. We describe Q-learning in more detail in Section 3.5.1. Because of this operator, Q-learning is most appropriate for domains with small and discrete action spaces. In continuous domains, solving the optimisation problem required to evaluate a $\max$ operator is typically NP-hard when non-linear families of function approximators are used to represent the Q-functions [Hochbaum, 2007].

An alternative class of algorithms known as *policy-based* methods assume a parametric differentiable policy π_θ parametrised by $\theta \in \Theta$. Under this assumption, derivatives of the reinforcement learning objective can be obtained exactly using the policy gradient theorem [Williams et al., 1993, Sutton et al., 2000]:

$$\nabla_\theta J(\theta) = \sum_{t=0}^{\infty} \gamma^t \int \int Q^{\pi_\theta}(s,a) \nabla_\theta \log \pi_\theta(a|s) d\pi_\theta(s|s) d\rho_t^{\pi_\theta}(s), \tag{3.3}$$

where recall that $\rho_t^{\pi_\theta}$ is the distribution over states after t timesteps defined in Eq. (3.1). Observe that Eq. (3.1) requires the existence of $\lim_{t\to\infty} \rho_t^{\pi_\theta}$ in order for the gradient to be well defined. We will give conditions for and characterise the convergence of the state distribution in Section 3.5.3. Perhaps due to a lack of awareness of the true form of the policy gradient theorem for the discounted case in Eq. (3.3), many state-of-the-art algorithms (including PPO [Schulman et al., 2017], TRPO [Schulman et al., 2015], SAC [Haarnoja et al., 2018a], A3C [Mnih et al., 2016], ACKTR [Wu et al., 2017], ACER[Wang et al., 2016]) use the gradient:

$$\nabla_\theta J(\theta) = \sum_{t=0}^{\infty} \int \int Q^{\pi_\theta}(s,a) \nabla_\theta \log \pi_\theta(a|s) d\pi_\theta(a|s) d\rho_t^{\pi_\theta}(s), \tag{3.4}$$

thereby ignoring the discount factor γ in the outer summation of Eq. (3.3). Whilst using the form in Eq. (3.4) is theoretically problematic as it does not correspond to a derivative the gradient of any function [Nota and Thomas, 2020], it appears to make little difference practice. Moreover, as the introduction of the discounting acts like sampling from an (unnormalised) geometric distribution, sampling the true gradient from Eq. (3.3) would involve terminating a trajectory with probability $1 - \gamma$ which would reduce the sample efficiency.

A more fundamental problem with using the policy gradient in Eq. (3.3) is that a sample from the Q-function for the current policy π_θ is required. Algorithms like REINFORCE [Williams et al., 1993] estimate the Q-function using samples of returns, which introduces two major drawbacks. Firstly, they suffer from high variance. For most RL problems, it is prohibitively costly to sample more than a few trajectories for an update, especially if they are long enough to mitigate the effect of bias due to truncation. Moreover, the sampled rewards from each trajectory are often reused to estimate Q-functions for each timestep along the trajectory, which further compounds the issue with variance. The second major drawback of policy gradient methods is their sample inefficiency. After an update to the parameters is carried out, all of the data used to compute the Q-function estimates are neglected and new samples are required for the new policy. Using importance sampling can help mitigate this at the expense of increased variance. For these two reasons, pure policy gradient approaches are seldom used in modern algorithms.

A way to tackle both the variance and sample inefficiency issues of policy gradient approaches is to introduce a Q-function approximator known as a *critic* $Q^{\pi_\theta} \approx \hat{Q}_\omega$. The policy π_θ is known as an *actor* and is learnt by using the appropriate policy gradient with Q^{π_θ} substituted for the critic $\hat{Q}_\omega$. Algorithms that interleave successive policy iteration and policy evaluation steps are known as *actor-critic* algorithms [Konda and Tsitsiklis, 2000]. Actor-critic methods have improved variance because learning a critic avoids the single-sample estimator problem of using the return; instead, $\hat{Q}_\omega$ is estimated from from several samples from multiple trajectories in the environment. Similarly, the critic never expires in the same way a trajectory does; when carrying out policy evaluation with new samples the most up to date form of the function approximator is used as the

initial estimate. This means that the influence of samples extends beyond a single update. Finally, actor-critic algorithms are suitable for all action spaces as they don't require prohibitively expensive operators like $\max$ and $\arg\max$, but instead carry out partial policy improvement and policy evaluation.

When taking a Bayesian approach to solving the RL problem using an actor-critic, we show in Section 7.4.2 that there exists a formal justification for ignoring the 'true' policy gradient in Eq. (3.3) and instead it suffices to find $\arg\max_{\pi\in\Pi} \mathbb{E}_{\pi\rho}\left[\hat{Q}_\omega\right]$ as a policy improvement step, where ρ is a distribution with support over $\mathcal{S}$.

3.3.4 Bayesian vs Frequentist RL

A final distinction that we make is between Bayesian approaches and frequentist approaches to RL. To elucidate the difference between the two, we consider the problem of estimating an unknown Q-function from a finite number of unbiased samples $\mathcal{D}$. Under a frequentist epistemology, the dataset of samples are assumed to be random and the uncertainty in the Q-function is simply the distribution that is used to generate samples: $\mathcal{D} \sim P_{\mathcal{D}}$. According to a frequentist approach, the underlying Q-function is not a random object and hence the goal would be to infer a point estimate of its value. One of the defining features of frequentist approaches is that they rely on hypothetical data to make inferences about the Q-function. If we wanted to characterise the uncertainty in some estimate of the Q-function at (s, a), say $Q^\pi(s, a)$, a frequentist approach would generate a confidence interval $I^\alpha(s, a) \subseteq \mathbb{R}$

$$I^\alpha(s, a) \in \{I : P_{\mathcal{D}}(Q^\pi(s, a) \in I(\mathcal{D})) = 1 - \alpha\}.$$

$I^\alpha(s, a)$ can be interpreted as an interval within which the true value $Q^\pi(s, a)$ would lie $1 - \alpha$ of the time if we were to sample an arbitrary dataset from the sampling distribution. A frequentist epistemology therefore relies on hypothetical data that has not been observed to evaluate the measure $P_{\mathcal{D}}(Q^\pi(s, a) \in I(\mathcal{D}))$ and to make predictions, interpretting probability as a long term frequency. A key observation is that if we are in the learning problem, we can't even evaluate these confidence intervals as we don't know the true sampling distribution in the first place. To make any progress, we must break

away from a frequentist epistemology and reason over the distribution $P_{\mathcal{D}}$ using a prior, which violates the assumption that the Q-function is a fixed, non random object.

In contrast, a Bayesian epistemology interprets probability as the degree of belief in an unknown quantity. When considering the problem of estimating an unknown Q-function, a Bayesian approach considers the observed dataset $\mathcal{D}$ to be fixed and the function Q^π to be random. As Q^π is random function, the goal of Bayesian statistics is to infer a distribution known as a *posterior* $P(Q^\pi|\mathcal{D})$, which characterises the uncertainty in Q^π given the observed dataset $\mathcal{D}$. Unlike in frequentist approaches, the uncertainty is not predicated on hypothetical datasets and the long run frequency of observed random variables. Instead, the posterior is only conditioned on observed data and is updated whenever new data is available to reflect the change in belief due to new evidence. This makes Bayesian approaches particularly natural for the RL setting as we seek to make decisions that can be inferred from relatively few incomplete instances of data drawn from underlying distributions that are constantly being updated. We formalise Bayesian inference in Chapter 4 before providing a detailed exposition of its application to RL in Chapter 6.

3.4 The Exploration/Exploitation Dilemma

As we discussed in Section 3.3.1, if the agent has complete knowledge of the MDP then samples can be drawn arbitrarily from the state-transition distribution and reward function to solve the planning problem. If the agent is in a learning problem, then the data that it receives from the environment is controlled by the sequences of actions selected by their policy and the consequent states that it observes. To this effect, the agent faces an exploration problem as it must explore the state-action space sufficiently to be able to gather diverse enough data to be able to find an optimal policy. At the same time, the agent must eventually exploit the knowledge that it has gained from exploring the MDP as exploring indefinitely will be sub-optimal. Balancing exploration with exploitation is known as the *exploration/exploitation* dilemma.

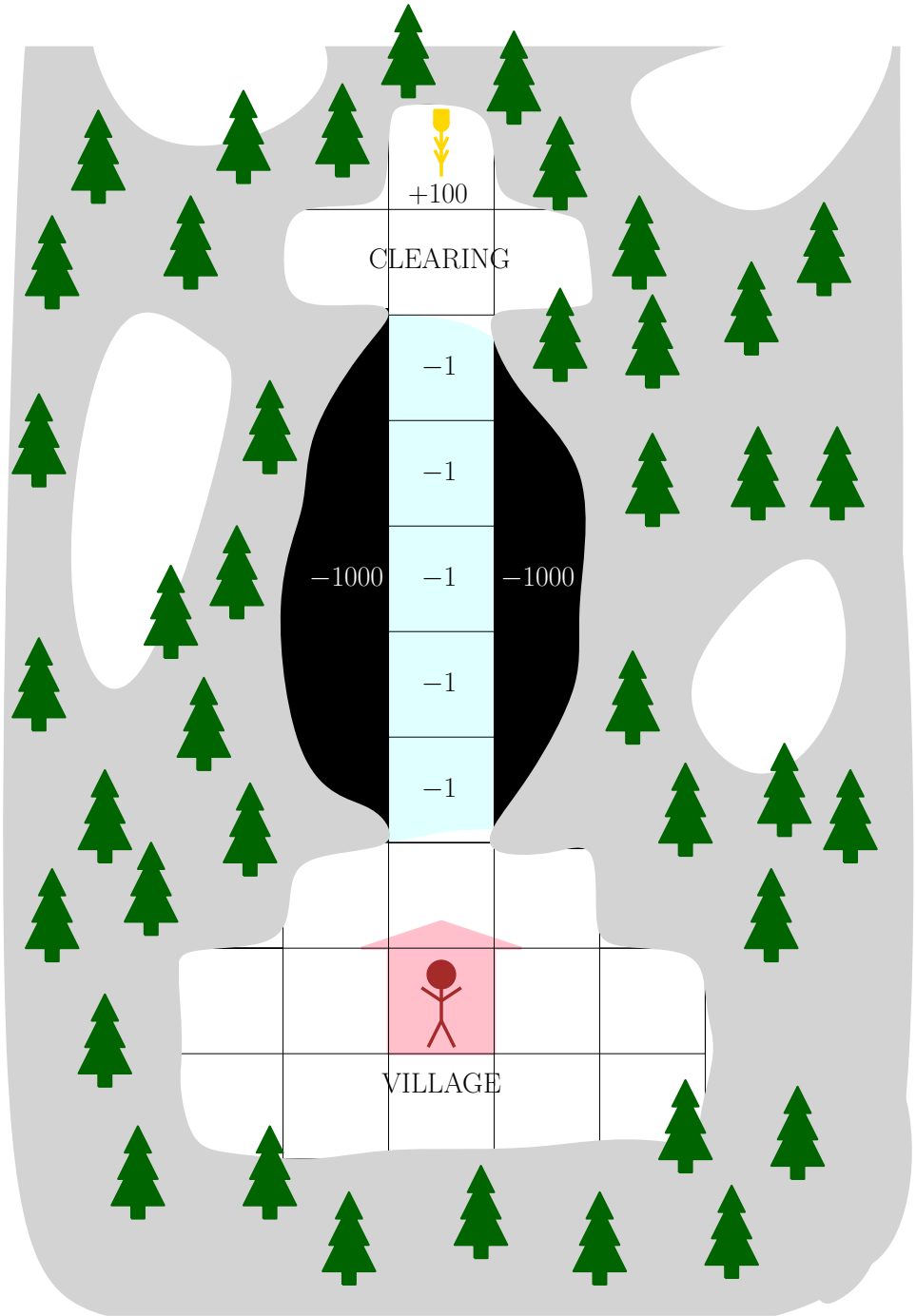


Figure 3.3: Alpine exploration MDP

Developing methods to tackle the exploration/exploitation dilemma will be a key theme of this thesis. To introduce and compare exploration approaches, we present a simple discrete alpine exploration MDP in Fig. 3.3 which is a gridworld that models an inhabitant of a remote mountain village. The inhabitant is particularly fond of making a delicious alpine liquor that requires a rare golden flower. Unfortunately, the flower only grows in a high altitude clearing that can be accessed by a treacherous mountain pass. The MDP's action space is the set $\mathcal{A} = \{\uparrow, \rightarrow, \downarrow, \leftarrow\}$ which represents the movements that the agent can take. The agent's state is represented a vector $s = (s_1, s_2)$ where $s_1 \in \mathcal{S}_1 \coloneqq \{1 : 22\}$ encodes which square of 22 the agent is located and $s_2 \in \mathcal{S}_2 \coloneqq \{0, 1\}$ is a binary variable which encodes whether or not the flower is present in the clearing. The overall state space is $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2$. The agent's initial position is always in their house, which is represented as the pink square the middle of the village as shown in Fig. 3.3. Initially $s_2 = 1$, representing the presence of the flower in the clearing.

The agent's goal is to retrieve the golden flower and bring it back to their house to prepare the liquor. The agent moves deterministically one square in the direction of the action arrow provided there is a square to move into. If not, the agent stays put. Once the agent is in the same square as the golden flower, $s_2 = 0$ representing the flower being picked. If the agent returns to their house on the pink square with $s_2 = 0$, the state becomes $s_2 = 1$ representing the growth of a new flower for the agent to pick and the agent receives a reward of $+100$. In order to reach the flower, the agent must traverse a treacherous mountain pass connecting the village to the clearing, represented here by blue squares. The agent receives a reward of -1 every time they move into a blue square as shown in Fig. 3.3. If the agent falls off the mountain pass by choosing an action $a \in \{\leftarrow, \rightarrow\}$ on a blue square, they stay in the same square but receive a reward of -1000. The reward function for all other state-action-state interactions is 0. To learn an optimal policy, the agent must learn a path to the golden flower and back home via the mountain pass.

We assume a priori that the agent has no knowledge of the MDP. The MDP therefore represents a challenging exploration problem: by taking any action, the agent can initially learn the transition dynamics within the village and find policies that always lead to an expected return of 0. Eventually, the agent may discover the mountain pass. Until the

flower is discovered and returned to the house, any policy that involves traversing the mountain pass, even without falling, will yield a lower expected return than a policy that keeps the agent walking around the village. To discover the flower and return home, several directed exploration steps are required that will initially yield policies with lower returns than the best optimal policy that can be constructed from the observed data. We now review several approaches to exploration in light of this example:

3.4.1 Residual Stochasticity

The simplest exploration strategy relies on the residual stochasticity present in the agent's policy. This method is most appropriate for approaches that learn a policy directly such as policy gradients and actor-critic. As an example, consider a class of Gaussian policies with a parametric mean and covariance matrix. During training, exploration depends on the fact that policy gradients aren't optimised to completion before new samples are obtained, hence the policy will still be a Gaussian with a non-zero variance. In reality, relying on residual stochasticity does not provide sufficient exploration as the policy may still collapse too quickly to a Dirac-delta distribution. For this reason, residual stochasticity alone is seldom used in practice.

3.4.2 Dithering Exploration

Another simple exploration method injects noise around the agent's estimate of the optimal policy. For approaches such as Q-learning where taking an $\arg\max$ is computationally feasible, ϵ-greedy exploration constructs an optimal policy according to the current optimal Q-function estimate $\pi_{k+1}(\cdot|s) = \delta_{\mathcal{A}_k^\star(s)}$ where recall $\mathcal{A}_k^\star(s) \in \arg_{a' \in \mathcal{A}} \max Q^{\pi_k}(s, a')$ is an arbitrary selection of any element of $\arg_{a' \in \mathcal{A}} \max Q^{\pi_k}(s, a')$. The policy π_{k+1} is then followed with probability ϵ and a random action is selected with probability $1 - \epsilon$. A related approach that is more suitable for continuous domains is known as Boltzmann exploration. A Boltzmann policy is constructed using the Q-function approximator with density:

$$\pi_{\text{Boltz}}(a|s) \propto \exp\left(\frac{\hat{Q}_\omega(s, a)}{\alpha}\right).$$

The scalar $\alpha \in (0, \infty)$ is the temperature parameter which controls how peaked the distribution is. We can interpret the Boltzmann distribution as being a soft maximum which is controlled by the temperature; in the limit $\alpha \to 0$, the policy concentrates on the maximum of $\hat{Q}_\omega(s, a)$ where as when $\alpha \to \infty$, the policy tends towards a uniform distribution.

A major issue with dithering exploration is that the agent has no way of preventing actions being repeatedly selected once it has learned that they are sub-optimal. For example in our alpine MDP, if the agent is using ϵ-greedy and is on a blue square, it will continue take an action from the set $\{\leftarrow, \rightarrow\}$ with probability $0.5 \cdot (1 - \epsilon)$ even if it has learnt that these actions always lead to a low return due to falling off the mountain pass. One approach to overcome this issue is to introduce an adaptive schedule for the exploration parameter, for example monotonically increasing the value of ϵ towards 1 or the temperature parameter α towards 0, thereby curtailing the degree of exploration. This introduces a further issue as it is difficult to set such a schedule reliably without prior knowledge of the MDP.

Even if a good schedule is available, dithering is inherently undirected in its exploration, which means that exploration is directed in all directions indiscriminately. To understand the effect that this has on sample inefficiency, consider the alpine MDP example. Using a pure exploration policy (that is setting $\epsilon = 0$ for an ϵ-greedy approach), the agent's probability of the agent taking the trajectory of 16 actions to return home with the flower and discover a reward of $+100$ is still only 0.25^{16}. Despite this shortcoming, dithering exploration is still a popular choice in RL due to its simplicity and computational efficiency.

3.4.3 Uncertainty Driven Exploration

One of the fundamental reasons for the failure of dithering exploration is that the agent cannot explore in a directed manner based on its uncertainty in the environment. One of the key principles of effective uncertainty driven exploration is that an agent should explore to reduce its uncertainty in the MDP, but only in so far as that exploration leads to greater expected returns. This allows an agent to cross off actions that it is certain will lead to low returns (for example falling from the mountain pass in the alpine MDP)

and instead focus exploration on parts of the MDP that offer potential for high returns. Uncertainty driven exploration is therefore always *adaptive* as it must vary according to the samples the agent has observed. For this reason, uncertainty driven exploration is naturally aligned with a Bayesian epistemology which, as we discussed in Section 3.3.4, characterises uncertainty through a posterior that conditions only on observed data. We formalise this notion through the conditionality principle in Section 4.2 and demonstrate that frequentist methods don't adhere to it in Section 6.3.1.

We also distinguish between deep and myopic uncertainty driven exploration. With reference to the alpine MDP, a myopic exploration scheme would only characterise the uncertainty in the expected one step reward and adapt exploration based on that. We see that this is still insufficient to learn an optimal policy as the agent would never explore beyond the first square of the mountain pass because the immediate reward for staying in the village is higher. In contrast, a deep exploration strategy would characterise the uncertainty in the full expected return, optimising exploration not just for immediate reward, but for exploring unknown transition dynamics and rewards many steps into its future [Osband et al., 2019].

3.5 Practical Policy Evaluation

All methods for model-free RL introduced thus far rely on obtaining an estimate of a Q-function. As discussed in Section 3.3.3, estimating a Q-function with sampled returns is typically impractical due to high variance and the necessity for several steps of interaction with the environment to obtain enough reward samples to reduce the bias in the estimate. To circumvent this issue, practical approaches use a combination of sampled rewards with a Q-function approximator to estimate the true Q-function. We now describe two of the most frequently encountered approaches for practical policy evaluation.

3.5.1 TD Learning

When learning Q-functions, a convenient norm to measure the distance between the true Q-function and the Q-function approximator is the mean squared error (MSE):

$\omega^\star \in \arg\min_{\omega\in\Omega} \text{MSE}(\omega)$, where

$$\text{MSE}(\omega) \coloneqq \|Q^\pi - \hat{Q}_\omega\|_\varsigma^2, \tag{3.5}$$

and ς is a sampling distribution with support over state-actions. Under the assumption that we have access to samples from the true Q-function, stochastic gradient descent can be used to minimise the mean-squared error by obtaining estimate of the gradient:

$$\nabla_\omega \text{MSE}(\omega) \propto -\mathbb{E}_\varsigma \left[\left(Q^\pi - \hat{Q}_\omega \right) \nabla_\omega \hat{Q}_\omega \right].$$

The MSE still can't be evaluated because it requires Q^π. temporal difference (TD) approaches [Sutton, 1988] address this by replacing it with an approximation obtained by applying the Bellman equation:

$$Q^\pi = \mathbb{E}_{P^\pi} \left[r + \gamma Q^\pi \right] \approx \mathbb{E}_{P^\pi} \left[r + \gamma \hat{Q}_\omega \right].$$

The concept of approximating the Q-function with immediate reward and estimating future returns with the Q-function approximator is known *bootstrapping*. When an agent in state s having taken action a, it transitions to the next state according to the state transition distribution $s' \sim P(\cdot|s,a)$, choses an action $a \sim \pi(\cdot|s)$, observes a reward $r = r(s', a, s)$, and then selects a new action $a' \sim \pi(\cdot|s')$, forming a tuple of observations (s', a', r, s, a). The TD algorithm uses these observations to obtain gradient estimates to minimise the mean-squared error via stochastic gradient descent, yielding a sequence of parameter updates:

$$\omega_{k+1} = \omega_k + \alpha_k \left(r + \gamma \hat{Q}_{\omega_k}(s', a') - \hat{Q}_{\omega_k}(s, a) \right) \nabla_\omega \hat{Q}_{\omega_k}(s, a), \quad \text{(TD)} \tag{3.6}$$

where the tuple index is determined by initial the initial state-action pair (s, a) according the sampling distribution $(s, a) \sim \varsigma$ and α_k controls the stepsize of the update. If the agent is learning *online*, the updates occur simultaneously with observed samples. Otherwise the tuples are stored in a data structure for updating after the agent has finished interacting with the environment.

For value-based approaches, we are concerned with learning an optimal Q-function. To accommodate this, a special case of TD-learning can be derived that approximates the

application of the optimal Bellman operator, forming a family of approaches known as *Q-Learning* [Watkins, 1989, Watkins and Dayan, 1992, Mnih et al., 2015]. Here updates require the application of the $\max$ operator:

$$\omega_{k+1} = \omega_k + \alpha \left(r + \gamma \max_{a' \in \mathcal{A}} \hat{Q}_{\omega_k}(s', a') - \hat{Q}_{\omega_k}(s, a) \right) \nabla_\omega \hat{Q}_{\omega_k}(s, a). \quad (Q\text{-Learning})$$

Instead of sampling from an optimal policy, which is not possible a priori, agents using methods such as Q-learning must sample from a separate sampling policies when interacting with the environment. TD methods that interact with the environment using a policy that differs from the policy of the Q-function that they are attempting to estimate are known as *off-policy* methods. We later discuss in Section 3.5.3 how off-policy sampling can negatively affect the long term convergence properties of TD methods.

TD-learning can be further generalised by recognising that the Bellman operator can be iteratively applied T times to the Q-function to obtain a T-step approximation:

$$Q^\pi = \mathbb{E}_{P_T^\pi} \left[\sum_{t=0}^{T-1} \gamma^t r_t + \gamma Q^\pi(s_T, a_T) \right] \approx \mathbb{E}_{P_T^\pi} \left[\sum_{t=0}^{T-1} \gamma^t r_t + \gamma^T \hat{Q}_\omega(s_T, a_T) \right],$$

which yields the T-step TD update:

$$\omega_{k+1} = \omega_k + \alpha \left(\sum_{t=0}^{T-1} \gamma^t r_t + \gamma^T \hat{Q}_{\omega_k}(s_T, a_T) \right) \nabla_\omega \hat{Q}_{\omega_k}(s_0, a_0). \quad (T\text{-step TD})$$

The choice of T thus controls the influence of sampled returns vs approximation for boostrapping: with larger T, the TD-approximation becomes less biased as a longer return of sampled rewards is used, but estimating its value will induce greater variance and can estimation can only commence after T interactions with the environment. Observe that when $T \to \infty$, we recover a Monte Carlo method where only the agent's returns are used, which further highlights the impracticality of this approach. Like in most practical RL applications, we will set $T = 1$ for the rest of this thesis.

3.5.2 Direct Methods

Instead of attempting to approximate the Q-function in the MSE of Eq. (3.5) we can exploit the insight that any Q-function must satisfy a Bellman equation: if we consider a family of parametric function approximators $\hat{Q}_\omega$ parametrised by $\omega \in \Omega$, solving

a Bellman equation is then equivalent to finding a parametrisation $\omega^\star \in \Omega$ such that $\mathcal{B}[\hat{Q}_{\omega^\star}] = \hat{Q}_{\omega^\star}$. This problem is equivalent to finding the minimiser of the mean squared Bellman error (MSBE):

$$\text{MSBE}(\omega) \coloneqq \|\mathcal{B}[\hat{Q}_\omega] - \hat{Q}_\omega\|_\varsigma^2,$$

where ς is a sampling distribution with support over the state-actions. Direct methods minimise the MSBE by estimating the gradient:

$$\begin{aligned}\nabla_\omega \text{MSBE}(\omega) &\propto \mathbb{E}_\varsigma \left[\left(\mathcal{B}[\hat{Q}_\omega] - \hat{Q}_\omega\right)\left(\nabla_\omega \mathcal{B}[\hat{Q}_\omega] - \nabla_\omega \hat{Q}_\omega\right)\right], \\ &= \mathbb{E}_\varsigma \left[\left(\mathbb{E}_{P^\pi}\left[r + \gamma \hat{Q}_\omega\right] - \hat{Q}_\omega\right)\left(\gamma \mathbb{E}_{P^\pi}[\nabla_\omega \hat{Q}_\omega] - \nabla_\omega \hat{Q}_\omega\right)\right]. \end{aligned} \tag{3.7}$$

Unlike in TD-learning, obtaining unbiased estimates of the gradient in Eq. (3.7) requires two samples from the state-transition distribution: one for the expectation in the Bellman operator: $\mathbb{E}_{P^\pi}\left[r + \gamma \hat{Q}_\omega\right]$ and another for the expectation in the gradient of the Bellman operator: $\mathbb{E}_{P^\pi}[\nabla_\omega \hat{Q}_\omega]$. To this end, a direct method using stochastic gradient descent to minimise the MSBE would sample tuple of observations $\langle s'', a'', s', a', r, s, a\rangle$ and update the Q-function approximator parameters according to:

$$\begin{aligned}&\omega_{k+1} = \omega_k \\ &\quad + \alpha \left(r + \gamma \hat{Q}_{\omega_k}(s', a') - \hat{Q}_{\omega_k}(s, a)\right)\left(\nabla_\omega \hat{Q}_{\omega_k}(s, a) - \gamma \nabla_\omega \hat{Q}_{\omega_k}(s'', a'')\right). \quad \text{(Direct)}\end{aligned} \tag{3.8}$$

Unfortunately, obtaining two samples from the state transition distribution requires the agent to choose its state arbitrarily, which is often impossible for most practical RL problems of interest. This is known as the *double sampling* problem [Baird, 1995], and prevents direct methods being applied in practice.

The double sampling problem reveals a more fundamental issue with the policy evaluation methods described thus far: by not accounting for uncertainty in the MDP, such approaches rely on being able to sample arbitrarily from unknown quantities. This is because frequentist methods don't have to adhere to the likelihood principle. We will show in Chapter 6 that this issue can be resolved by taking a Bayesian approach, as by accounting for uncertainty in unknown quantities such as state transition distributions, we only need to characterise our degree of belief in their value via a posterior to be able to carry out policy evaluation.

3.5.3 Convergence of Policy Evaluation

As direct methods employ stochastic gradient descent on the MSBE objective, their convergence properties can be established if the standard Robbins-Munro conditions are satisfied [Robbins and Monro, 1951]. To understand the convergence properties of TD-learning approaches, we can compare the update in Eq. (3.8) to Eq. (3.6), which reveals that the TD update is the same as the direct udpate with the problematic $\gamma \nabla_\omega \hat{Q}_{\omega_k}(s'', a'')$ gradient term dropped. On the one hand this avoids double sampling problem that renders direct methods impractical, on the other the TD updates are no longer true gradients of any fixed objective, which raises a question about their convergence properties.

To analyse the convergence of TD-learning, we can derive the underlying ordinary differential equation (ODE) of the TD updates:

$$\frac{\partial \omega}{\partial t} = \mathbb{E}_\varsigma \left[\left(\mathcal{B}[\hat{Q}_\omega] - \hat{Q}_\omega \right) \nabla_\omega \hat{Q}_\omega \right]. \tag{3.9}$$

Stability of any TD algorithm is predicated on the stability of this underlying ODE. Exact analysis of the Eq. (3.9) is possible for two specific classes of function approximator: the tabular setting where $\hat{Q}_\omega(s,a) = \omega^\top \mathbb{1}(s,a)$ and the linear setting where $\hat{Q}_\omega(s,a) = \omega^\top v(s,a)$. Our analysis will focus on the linear setting as it includes the tabular setting as a specific case when the basis vector is the indicator vector: $v = \mathbb{1}$. The fixed point of the ODE can be found exactly by setting Eq. (3.9) to zero and rearranging for $\omega^\star$, yielding:

$$\begin{aligned}
\mathbb{E}_\varsigma \left[\left(\mathbb{E}_{P^\pi} \left[r + \gamma \omega^{\star\top} v' \right] - \omega^{\star\top} v \right) v \right] &= 0, \\
\mathbb{E}_\varsigma \left[\left(\gamma \mathbb{E}_{P^\pi} \left[v' \right] - v \right) v^\top \right] \omega^\star &= \mathbb{E}_\varsigma \left[v \mathbb{E}_{P^\pi} \left[r \right] \right], \\
\omega^\star &= \left(\mathbb{E}_\varsigma \left[\left(\gamma \mathbb{E}_{P^\pi} \left[v' \right] - v \right) v^\top \right] \right)^{-1} \mathbb{E}_\varsigma \left[v \mathbb{E}_{P^\pi} \left[r \right] \right],
\end{aligned} \tag{3.10}$$

where we have introduced the shorthand $v' \coloneqq v(s', a')$. For the tabular setting, we can substitute for $v = \mathbb{1}$ in Eq. (3.10) to show that the fixed point is the exact solution to the Bellman equation $\hat{Q}_{\omega^\star} = \mathcal{B}[\hat{Q}_{\omega^\star}]$, which confirms the intuition that a tabular function approximator can exactly represent the true Q-function. For more general linear basis vectors, the true Q-function may not lie within the function approximator space. In this

case, we see that the TD fixed point in Eq. (3.10) is the solution to a *projected* Bellman equation $\hat{Q}_{\omega^\star} = \mathcal{P} \circ \mathcal{B}[\hat{Q}_{\omega^\star}]$ where $\mathcal{P} \circ \mathcal{B}[\hat{Q}_\omega]$ the projection of the Bellman operator $\mathcal{B}[\hat{Q}_\omega]$ onto the space of representable functions in the family of function approximators according to the $\|\cdot\|_\varsigma^2$ norm:

$$\mathcal{P} \circ \mathcal{B}[\hat{Q}_\omega] \coloneqq \left\{ \hat{Q}_{\omega'} : \omega' \in \arg\min_{\omega' \in \Omega} \|\mathcal{B}[\hat{Q}_\omega] - \hat{Q}_{\omega'}\|_\varsigma^2. \right\}. \tag{3.11}$$

If the space of function approximators can represent the Bellman operator exactly (such as in the tabular setting), then the projection is the identity $\mathcal{P} \circ \mathcal{B}[\hat{Q}_{\omega^\star}] = \mathcal{B}[\hat{Q}_{\omega^\star}]$. Observe from Eq. (3.9) that the TD fixed point also satisfies a projected Bellman equation $\hat{Q}_{\omega^\star} = \mathcal{P} \circ \mathcal{B}[\hat{Q}_{\omega^\star}]$ for nonlinear function approximators should the projection $\mathcal{P} \circ \mathcal{B}[\hat{Q}_{\omega^\star}]$ exist.

For the tabular setting, convergence of the TD algorithm has been established regardless of the sampling distribution ς [Sutton, 1988, Dayan, 1992]. We therefore focus on the more general linear case, where our analysis will show that to guarantee convergence, it is sufficient to assume that $d\varsigma = d\pi d\rho^\pi$ where is ρ^π the *stationary distribution* of the Markov chain induced by P_0 and the state-state transition distribution $P^\pi(\cdot|s) = \int P(\cdot|s,a)d\pi(a|s)$. We introduce the notation ξ^π for the distribution $d\xi^\pi \coloneqq dP^\pi d\pi d\rho^\pi$.

If we are to use the stationary distribution as a norm, we also need to be able to sample from it. In general, the stationary distribution does not have a known form a priori because the state transition distribution is unknown. However by introducing an assumption of ergodicity, we can handle both the question of the stationary distribution's existence and the ability to asymptotically sample from it:

Assumption 2 (Ergodicity)**.** *We assume that the MDP is ergodic for any policy π, that is there exists a unique stationary distribution ρ^π over the measurable space $(S, \Sigma_\mathcal{S})$ that, for any measurable function $f : \mathcal{S} \to \mathbb{R}$, satisfies:*

$$\int f(s) d\rho^\pi(s) = \int f(s') dP^\pi(s'|s) d\rho^\pi(s), \tag{3.12}$$

and to which the N-step state distribution convergences:

$$\rho_N^\pi \xrightarrow{TV} \rho^\pi.$$

Taking a frequentist view, ergodicity means that if we allow the agent to interact with the environment for a long enough period of time, the samples of states that the agent is in will closer approximate samples from the steady state distribution as the distribution over states ρ_N^π converges to the stationary distribution in total variation (recall from Proposition 2 that $\lim_{N\to\infty} \text{TV}(\rho^\pi \parallel \rho_N^\pi) = 0 \implies \lim_{N\to\infty} \rho_N^\pi = \rho^\pi$). Proving Assumption 8 holds typically requires verifying that the Markov chain is aperiodic and positive Harris recurrent [Harris, 1956, Gilks et al., 1995, Meyn et al., 2009]

For linear function approximators, we now analyse the stability of the TD fixed point and convergence of TD learning. These results have been established before in discrete state MDPs [Tsitsiklis and Van Roy, 1997, Borkar and Meyn, 2000] but, as far as the author is aware, never for the more general state-spaces considered in this thesis. Our analysis follows the stochastic approximation approach of [Borkar, 2008]. As we would like to prove that the long term behaviour of our stochastic updates to is characterised by the underlying ODE, it is crucial that we first establish that this ODE has an asymptotically stable equilibrium or else there can exist no stable fixed point for the algorithm to converge to. We prove stability in Lemma 1 by analysing the Jacobian of the ODE in Eq. (3.9):

Lemma 1. *Under Assumption 8 with $d\varsigma = d\pi d\rho^\pi$, for linear function approximators, the Jacobian $J(\omega) \coloneqq \nabla_\omega \left(\frac{\partial \omega}{\partial t}\right)$ of the ODE in Eq.* (3.9) *has the property:*

$$\omega^\top J(\omega)\omega \leq (\gamma - 1)\mathbb{E}_{\rho^\pi \pi}\left[(\omega^\top v)^2\right] \leq 0, \tag{3.13}$$

for all non-zero $\omega \in \mathbb{R}^n$.

Proof. Firstly, we must derive the Jacobian for a linear function approximator $\hat{Q}_\omega = \omega^\top v$:

$$\begin{aligned} J(\omega) \coloneqq \nabla_\omega \left(\frac{\partial \omega}{\partial t}\right) &= \mathbb{E}_{\rho^\pi \pi}\left[\left(\gamma \mathbb{E}_{P^\pi}\left[v'\right] - v\right) v^\top\right], \\ &= \gamma \mathbb{E}_{\rho^\pi \pi}\left[\mathbb{E}_{P^\pi}\left[v'\right] v^\top\right] - \mathbb{E}_{\rho^\pi \pi}\left[v v^\top\right]. \end{aligned}$$

Now, we analyse $\omega^\top J(\omega)\omega$:

$$\begin{aligned} \omega^\top J(\omega)\omega &= \omega^\top \gamma \mathbb{E}_{\rho^\pi \pi}\left[\mathbb{E}_{P^\pi}\left[v'\right] v^\top\right] \omega - \omega^\top \mathbb{E}_{\rho^\pi \pi}\left[v v^\top\right] \omega, \\ &= \gamma \mathbb{E}_{\rho^\pi \pi}\left[\mathbb{E}_{P^\pi}\left[\omega^\top v'\right] \omega^\top v\right] - \mathbb{E}_{\rho^\pi \pi}\left[(\omega^\top v)^2\right]. \end{aligned} \tag{3.14}$$

In order to establish the bound in Eq. (3.13), we apply the Cauchy-Schwartz inequality to the first term of Eq. (3.14):

$$\begin{aligned}\mathbb{E}_{\rho^\pi\pi}\left[\mathbb{E}_{P^\pi}\left[\omega^\top v'\right]\omega^\top v\right] &= \mathbb{E}_{\xi^\pi}\left[\omega^\top v'\omega^\top v\right], \\ &\leq \mathbb{E}_{\xi^\pi}\left[\left|\omega^\top v'\omega^\top v\right|\right], \\ &\leq \sqrt{\mathbb{E}_{\xi^\pi}\left[(\omega^\top v')^2\right]\mathbb{E}_{\xi^\pi}\left[(\omega^\top v)^2\right]}, \\ &= \sqrt{\mathbb{E}_{\xi^\pi}\left[(\omega^\top v')^2\right]\mathbb{E}_{\rho^\pi\pi}\left[(\omega^\top v)^2\right]}\end{aligned}$$

Under Assumption 8 the MDP is ergodic, allowing us to apply Eq. (3.12):

$$\mathbb{E}_{\xi^\pi}\left[(\omega^\top v')^2\right] = \mathbb{E}_{\rho^\pi\pi}\left[\mathbb{E}_{P^\pi}\left[(\omega^\top v')^2\right]\right] = \mathbb{E}_{\rho^\pi\pi}\left[(\omega^\top v)^2\right],$$

yielding the inequality:

$$\mathbb{E}_{\rho^\pi\pi}\left[\mathbb{E}_{P^\pi}\left[\omega^\top v'\right]\omega^\top v\right] \leq \sqrt{\mathbb{E}_{\rho^\pi\pi}\left[(\omega^\top v)^2\right]\mathbb{E}_{\rho^\pi\pi}\left[(\omega^\top v)^2\right]} = \mathbb{E}_{\rho^\pi\pi}\left[(\omega^\top v)^2\right],$$

which we apply to Eq. (3.14). Using the fact that $\gamma \in [0, 1)$ and $(\omega^\top v)^2 \geq 0$ yields our desired result:

$$\omega^\top J(\omega)\omega \leq (\gamma - 1)\mathbb{E}_{\rho^\pi\pi}\left[(\omega^\top v)^2\right] \leq 0.$$

□

We see from Lemma 1 that the Jacobian is negative definite if $\mathbb{E}_{\rho^\pi\pi}\left[vv^\top\right]$ is positive definite, providing a necessary and sufficient condition for the stability of the fixed point. Another assumption we must make to prove convergence of TD is that the variance of the samples for the updates is finite. Together these form a set of regularity assumptions that ensure the stochastic updates have a stable attractor and remain stable during the algorithm:

Assumption 3 (Regularity of Stochastic Approximation)**.** *We make the following regularity assumptions:*

i $\mathbb{E}_{\rho^\pi\pi}\left[vv^\top\right]$ *is positive definite.*

ii *The random variables* $a \coloneqq rv$ *and* $B \coloneqq \gamma vv'^\top - vv^\top$ *have finite variance, that is* $\mathbb{E}_{\xi^\pi}[\|a - \mathbb{E}_{\xi^\pi}[a]\|_2^2] < \infty$ *and* $\mathbb{E}_{\xi^\pi}[\|B - \mathbb{E}_{\xi^\pi}[B]\|_2^2] < \infty$.

Finally, we must make the standard Robbins-Monro assumption [Robbins and Monro, 1951] for the sequence of stepsizes in our updates:

Assumption 4 (Robbins-Monro Stepsize Condition)**.** *The sequence of stepsizes* $\{\alpha_k\}_{k\in\mathbb{Z}_0^+}$ *are positive scalars satisfying:*

$$\sum_{k=0}^{\infty}\alpha_k=\infty,\quad \sum_{k=0}^{\infty}\alpha_k^2<\infty.$$

Intuitively, the first condition of Assumption 4 $\sum_{k=1}^{\infty}\alpha_k$ ensures that the updates can reach any part of Ω. If this condition does not hold, the sequence of updates $\{\omega_k\}_{k\in\mathbb{Z}_0^+}$ is bounded which implies that there may be exist initial parameters values ω_0 that can never reach an equilibrium point ω^*. The second condition of Assumption 4 $\sum_{k=1}^{\infty}\alpha_k^2$ ensures that the influence of the stochastic noise in the updates eventually diminishes, allowing ω_k to converge to the fixed point exactly. Given our assumptions, we are now ready to state our convergence result:

Theorem 2 (Convergence of Linear TD)**.** *Under Assumptions 8, 3 and 4, the sequence of parameter updates* $\{\omega_k\}_{k\in\mathbb{Z}_0^+}$*:*

$$\omega_{k+1}=\omega_k+\alpha_k\left(r(s',a,s)+\gamma\omega_k^\top v(s',a')-\omega_k^\top v(s,a)\right)v(s,a), \tag{3.15}$$

where $(s',a',s,a)\sim\xi^\pi$*, converges to the TD fixed point:*

$$\omega^\star:=\left(\mathbb{E}_{\rho^\pi\pi}\left[\left(\gamma\mathbb{E}_{P^\pi}\left[v'\right]-v\right)v^\top\right]\right)^{-1}\mathbb{E}_{\rho^\pi\pi}\left[v\mathbb{E}_{P^\pi}\left[r\right]\right],$$

almost surely, that is:

$$\omega_k\xrightarrow[\xi^\pi-a.s.]{}\omega^\star. \tag{3.16}$$

Proof. To apply stochastic approximation analysis, we re-write Eq. (3.15) as:

$$\omega_{k+1}=\omega_k+\alpha_k\left(h(\omega_k)+\mathcal{M}_k\right),$$

where

$$h(\omega_k):=\mathbb{E}_{\xi^\pi}\left[\left(r+\gamma\omega_k^\top v'-\omega_k^\top v\right)v\right],$$

and

$$\mathcal{M}_k \coloneqq \left(r + \gamma\omega_k^\top v' - \omega_k^\top v\right) v - h(\omega_k),$$

is a Martingale. Theorem 2 from Borkar [2008] states that Eq. (3.16) holds if:

1. $h(\cdot)$ is Lipschitz.
2. The sequence of stepsizes $\{\alpha_k\}_{k\in\mathbb{Z}_0^+}$ satisfy Assumption 4.
3. $\{\mathcal{M}_{k\in\mathbb{Z}_0^+}\}$ is a Martingale difference sequence with respect to the filtration of increasing σ-algebras:

   $$\mathcal{F}_n \coloneqq \sigma(\{\omega_0, \mathcal{M}_0, ...\omega_k, \mathcal{M}_k\}),$$

   that is:

   $$\mathbb{E}[\mathcal{M}_k|\mathcal{F}_{n-1}] = 0 \quad \text{a.s.}, \quad n > 0.$$

4. Each $\mathcal{M}_k$ is square-integrable with:

   $$\mathbb{E}[\|\mathcal{M}_k\|_2^2|\mathcal{F}_{n-1}] \leq K(1 + \|\omega_k\|_2^2) \quad \text{a.s.}, \quad n > 0.$$

   where K>0 is some finite constant.

5. The sequence of updates remain bounded, that is

   $$\sup_{n\geq 0}\|\omega_k\| < \infty, \quad \text{a.s.}$$

Condition 1 trivially follows from the use of a linear function approximator and condition 3 is satisfied by construction of the Martingale. We now use the triangle inequality to verify condition 4:

$$\begin{aligned}
\|\mathcal{M}_k\|_2^2 &= \|\left(r + \gamma\omega_k^\top v' - \omega_k^\top v\right) v - h(\omega_k)\|_2^2, \\
&= \left\|\left(r + \gamma\omega_k^\top v' - \omega_k^\top v\right) v - \mathbb{E}_{\xi^\pi}\left[\left(r + \gamma\omega_k^\top v' - \omega_k^\top v\right) v\right]\right\|_2^2, \\
&= \left\|rv - \mathbb{E}_{\rho^\pi\pi}\left[rv\right] + \gamma\omega_k^\top\left(\left(v' - v\right)v - \mathbb{E}_{\xi^\pi}\left[\left(\gamma v' - v\right)v\right]\right)\right\|_2^2, \\
&\leq \|rv - \mathbb{E}_{\rho^\pi\pi}\left[rv\right]\|_2^2 + \gamma^2\left\|\omega_k^\top\left(\left(v' - v\right)v - \mathbb{E}_{\xi^\pi}\left[\left(\gamma v' - v\right)v\right]\right)\right\|_2^2, \\
&\leq \|rv - \mathbb{E}_{\rho^\pi\pi}\left[rv\right]\|_2^2 + \gamma^2\|\omega_k\|_2^2\,\|\left(\left(v' - v\right)v - \mathbb{E}_{\xi^\pi}\left[\left(\gamma v' - v\right)v\right]\right)\|_2^2.
\end{aligned}$$

Taking expectations and using Assumption 3 yields:

$$\begin{aligned}
&\mathbb{E}[\|\mathcal{M}_k\|_2^2|\mathcal{F}_{n-1}] \\
&\quad\leq \mathbb{E}_{\rho^\pi\pi}\left[\|rv - \mathbb{E}_{\rho^\pi\pi}[rv]\|_2^2\right] + \gamma^2\|\omega_k\|_2^2\mathbb{E}_{\xi^\pi}\left[\|((v'-v)\,v - \mathbb{E}\left[(\gamma v'-v)\,v\right])\|_2^2\right], \\
&\quad\leq K(1+\|\omega_k\|_2^2),
\end{aligned}$$

where

$$K = \max\left(\mathbb{E}_{\rho^\pi\pi}\left[\|rv - \mathbb{E}_{\rho^\pi\pi}[rv]\|_2^2\right], \gamma^2\mathbb{E}_{\xi^\pi}\left[\|((v'-v)\,v - \mathbb{E}\left[(\gamma v'-v)\,v\right])\|_2^2\right]\right),$$

as required.

Verifying condition 5 is non-trivial, however Theorem 7 of Borkar [2008] states that it is satisfied if, in addition to conditions 1-4 being satisfied:

5*. The functions $h_c(\omega) \coloneqq \frac{h(c\omega)}{c},\ c \geq 1$ satisfy $h_c \xrightarrow{\text{unif}} h_\infty$ for some continuous h_∞. The ODE :

$$\frac{\partial\omega(t)}{\partial t} = h_\infty(\omega),$$

has the origin as its unique globally asymptotic stable equilibrium.

We see that the first requirement of condition $5^\star$ is satisfied by the linearity of $h(\omega)$:

$$\begin{aligned}
h_c(\omega) &= \mathbb{E}_{\xi^\pi}\left[\left(\frac{r}{c} + \gamma\omega_k^\top v' - \omega_k^\top v\right)v\right], \\
\implies h_c(\omega) &\xrightarrow{\text{unif}} h_\infty(\omega) = \mathbb{E}_{\xi^\pi}\left[\left(\gamma\omega_k^\top v' - \omega_k^\top v\right)v\right].
\end{aligned}$$

Finally, the ODE $\frac{\partial\omega(t)}{\partial t} = h_\infty(\omega)$ has a solution at $\omega = 0$ and the same Jacobian as the TD ODE in Eq. (3.9), hence we can use Lemma 1 with Assumption 3 to show that the Jacobian $\nabla_\omega h_\infty(\omega)$ is positive definite everywhere, implying that $\omega = 0$ is a unique globally asymptotic stable equilibrium. Condition $5^\star$ is therefore satisfied and our main result follows immediately. □

We pay particular attention to the assumption that samples are obtained from the MDP's stationary distribution, which is essential for proving that the Jacobian in Lemma 1 is positive definite, and therefore establishing the stability of TD fixed point. We refer to methods that sample from ρ^π (at least in the limit of infinite data) by following policy

π when learning Q^π as *on-policy* methods. In contrast, in *off-policy* methods such as Q-learning, agents sample from a policy π' that is different from the Q-function that they are learning Q^π. This means that their state occupation distribution ρ^π_N will not converge to ρ^π in the limit of $N \to \infty$ and our convergence analysis of Theorem 2 will no longer hold. Moreover, there exist several counter example MDPs that are provably divergent for linear TD methods when agents sample off-policy [Tsitsiklis and Roy, 1994, Gordon, 1995, Boyan and Moore, 1995, Baird, 1995].

We might be tempted to hypothesise that the result of Theorem 2 can be extended to any smooth nonlinear function approximator. Unfortunately, there too exist MDPs that are provably divergence for classes of nonlinear function approximators [Tsitsiklis and Van Roy, 1997]. Despite this, many popular TD algorithms still use nonlinear deep neural networks to approximate value functions, often in combination with off-policy sampling [Mnih et al., 2015, Schulman et al., 2015, 2017, Haarnoja et al., 2017, 2018a]. At first glance, this suggests that issues of convergence might not matter in practice for TD-learning, however on closer inspection, many of these approaches introduce a separate set of target parameters $\omega' \in \Omega$ which prove essential to these algorithms' stability. The target parameters remain fixed during the TD update, acting like a supervised target for which gradient descent is performed on:

$$\omega_{k+1} = \omega_k + \alpha_k \left(r + \gamma \hat{Q}_{\omega'}(s', a') - \hat{Q}_{\omega_k}(s, a)\right) \nabla_\omega \hat{Q}_{\omega_k}(s, a). \quad \text{(TD-Target)} \quad (3.17)$$

Unlike in classic TD-learning, the update in Eq. (3.17) is an unbiased estimate of a true gradient on the TD-target objective:

$$\text{TD-Target}(\omega) = \|\mathcal{B}[\hat{Q}_{\omega'}] - \hat{Q}_\omega\|^2_\varsigma \quad (3.18)$$

The target parameters are updated to the current approximator parameters every T_{target} timesteps:

$$\omega' = \begin{cases} \omega_k, & k \bmod T_{\text{target}} = 0, \\ \omega', & \text{otherwise.} \end{cases}$$

In Chapter 7, we will prove that this separation of timescales allows formal convergence properties to emerge, with the parameters ω_k being pseudo-equilibrated from the perspective of the target parameters and the parameters ω' being pseudo-stationary from the perspective of the approximator parameters ω_k. Moreover, we will show that TD-target objective and the introduction of target parameters arise naturally from a Bayesian approach to policy evaluation in situations where approximate inference is required.

3.6 A Guide to this Thesis

Within this taxonomy of RL algorithms, all algorithms presented in this thesis can be classified as model-free RL: in Chapter 5 we will focus on a particular family of model-free approaches that introduce a degree of regularisation into the policy gradient update. The additional regularisation introduced into the objective allows us to make an analogy between calculating an optimal policy for the RL problem and inferring a posterior in Bayesian inference. For this reason, this family of approaches is commonly known as *control as inference* [Levine, 2018b]. To avoid a common misconception regarding control as inference, we clarify here that there is nothing inherently Bayesian about this approach as it does not characterise uncertainty in any aspect of the MDP and any connection to Bayesian inference is metaphorical. However, by making this connection, powerful tools used to approximate the inference procedure are automatically applicable to the RL problem. Our VIREL framework adopts a variational inference (VI) approach to finding an exploration policy whilst avoiding some of the pathologies of other approaches such as a biased solution to the RL problem or an objective that encourages risk seeking behaviour.

In Chapter 6, we study at methods that use Bayesian inference to characterise the uncertainty in the MDP. Our aim of this chapter is to provide a formal exposition of the benefits of Bayesian RL collecting theories from optimal decision theory and applying them to the RL context as well as clarifying what is means for a policy to be optimal under a Bayesian approach. We use these theories to analyse why a frequentist approach to the reinforcement learning problem is not appropriate in the learning setting and explore

pathologies that arise from doing so. We also review existing approaches to Bayesian RL.

As we discussed in Section 3.5 an essential aspect of modern, practical RL algorithms in the model-free paradigm is the use of bootstrapping in policy evaluation. Despite its ubiquity, there has been no investigation exploring the effect that bootstrapping has on Bayesian model-free RL. To address this issue, we develop the Bayesian Bellman operator (BBO) framework in Chapter 7 which accounts for bootstrapping by modeling the uncertainty of a function approximator under a Bellman operator. Moreover, we show that unlike the TD approaches discussed in Section 3.5.3, policy evaluation under the BBO is convergent and is consistent with frequentist approaches, converging to a TD fixed point under a Gaussian likelihood model. In addition to unifying several existing model-free approaches in a theoretically sound framework, BBO also allows us to develop a novel actor-critic algorithm that combines the benefits of a regularised RL with a true Bayesian exploration regime.

4

Characterising Uncertainty

4.1 Bayesian Inference

A Bayesian epistemology treats probabilities as a degree of belief in unknown quantities. A Bayesian characterises the uncertainty in these quantities by inferring a distribution over their value, known as a posterior, given observations from an environment and a model of the underlying process. We now look to a simple example to help illustrate the concepts of Bayesian inference: consider the graphical model in Fig. 4.1, which defines a probabilistic model of the observations $P_X(\cdot|\phi)$. In our graphical models we symbolise observed variables as double pink circles and latent variables as white circles. Given N independent identically distributed (i.i.d.) observations of the random variable X, we can form a dataset $\mathcal{D}_N := \{x_i\}_{i\in 1:N}$. The dataset has a likelihood defined by the density:

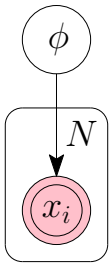


Figure 4.1: A simple inference problem

$$p(\mathcal{D}_N|\phi) = \prod_{i=1}^{N} p_X(x_i|\phi).$$

Given these observations, we would like to infer the conditional probability of the latent variable $\phi \in \Phi$ conditioned on our dataset: $P_\Phi(\cdot|\mathcal{D}_N)$. $P_\Phi(\cdot|\mathcal{D}_N)$is known as the *posterior*, and characterises the belief in the unknown variable ϕ.

In order to derive the posterior, we must first specify a prior over the hidden variable: P_Φ. A prior should summarise all pre-existing knowledge about ϕ before receiving new samples from the environment. The prior belief, in conjunction with the likelihood of sampled data, can be used to infer the posterior using Bayes' rule, which has the density:

$$p_\Phi(\phi|\mathcal{D}_N) = \frac{p(\mathcal{D}_N|\phi)p_\Phi(\phi)}{p(\mathcal{D}_N)}.$$

$p(\mathcal{D}_N)$ is known as the *evidence* and can be calculated by marginalising over the joint:

$$p(\mathcal{D}_N) = \int p(\mathcal{D}_N|\phi)dP_\Phi(\phi). \tag{4.1}$$

The posterior thus characterises the uncertainty in the unknown latent variable ϕ through its distribution given all observations ϕ, a model of the process and our prior beliefs in its value. When reasoning over an unknown variable, we can use the posterior to marginalise over all possible values it might take according to our current belief, thereby formally accounting for our uncertainty in its value. Posterior marginalisation also allows us to characterise the uncertainty of random variables that depend on the unknown latent variables. In our example, we can characterise the uncertainty in the variable X by marginalising over all possible models:

$$P_X(\cdot|\mathcal{D}_N) = \int P_X(\cdot|\mathcal{D}_N)dP_\Phi(\phi|\mathcal{D}_N).$$

The resulting distribution $P_X(\cdot|\mathcal{D}_N)$ is thus a posterior over X. $P_X(\cdot|\mathcal{D}_N)$ is sometimes referred to as the posterior predictive, as it also allows us to predict the probability of any new set of datapoints given our observations.

Due to the use of a prior, Bayesian statistics is often referred to as subjective probability: agents with different priors will infer different posterior beliefs over quantities given exactly the same data and so may reach different conclusions over finite samples. As more data is available to the agent, the prior beliefs will be either be confirmed or overwhelmed by the data and we might wonder whether agreement between agents with different priors can emerge. A formal, positive answer to this questions arises from the

Bernstein von-Mises theorem Vaart [1998], a consequence of which is that under certain regularity assumptions, in the limit of infinite observations, a Bayesian agent's posterior will concentrate on the maximum likelihood estimator regardless of their initial choice of prior.

4.2 Likelihood Principle

A more philosophical arguement for considering Bayesian approaches in this thesis, but one with profound practical consequences, is that Bayesian probability satisfies the *likelihood principle* [Birnbaum, 1962, Gandenberger, 2015]. Formally, the likelihood principle states that when inferring a variable ϕ having observed x, all relevant information is contained in the likelihood function for the observed x. Furthermore, two likelihood functions contain the same information about ϕ if they are proportional to each other.

The likelihood principle follows from two related and desirable properties: the sufficiency and conditionality principles [Birnbaum, 1962]. The sufficiency principle states that if we are inferring ϕ from a variables X and $T \circ X$ is a sufficient statistic, then inferring ϕ from T contains all the same information. The sufficiency principle is a consequence of the existence of pushforward distributions on transformations of variables, which we investigate further in Section 6.3 for Bayesian RL. The conditionality principle states that inferences about ϕ should only condition on the data observed and not on hypothetical or future data. A Bayesian approach naturally satisfied the conditionality principle as all inferences are made using a posterior distribution, which only conditions on observed data. In the context of RL the likelihood principle asserts that an agent should make decisions based only the data that it has observed, given the same model of the environment, it should not matter which random variable is observed provided the variable is sufficient for solving the RL problem.

To elucidate the likelihood principle and introduce the key Bayesian themes of this thesis, we return to the simple bandit problem presented in Chapter 1. Recall that we are faced with choosing an action $a \in \mathcal{A} = \{\leftarrow, \rightarrow\}$ and then receive a random reward $r \in \{0, 1\}$

with probability $P(r = 1|a, \phi_{\leftarrow}, \phi_{\rightarrow}) = \phi_a$. For this example, we assume we know the value of the parameter $\phi_{\rightarrow} = 0.32$, but are uncertain about $\phi_{\leftarrow}$.

We extend and adapt the example of Lindley and Phillips [1976]. Consider the binomial probability space $(O, \Sigma_O, P_O(\cdot|\phi_{\leftarrow}))$ where each outcome $o = \{r_1, r_2, r_3, ...\}$ is an infinite number of i.i.d. samples from $P(r = 1| \leftarrow, \phi_{\leftarrow}])$. We assume that the agent has access to a single observation from either the random variable $x_1 \in X_1$ that counts the number of rewards sampled until five rewards of 1 are encountered: $X_1(o) = \sum_{k=10}^{\infty} k\mathbb{1}(o \in \{o' : \sum_{i=1}^{k} r_i = 5\})$ or the random variable $x_2 \in X_2$ that counts the sum of rewards of the first fifteen samples: $X_2(o) = \sum_{i=1}^{15} r_i$. The probability densities of the random variables X_1 and X_2 are:

$$p_{X_1}(x_1|\phi_{\leftarrow}) = \binom{x_1 - 1}{4} \phi_{\leftarrow}^5 (1 - \phi_{\leftarrow})^{x_1 - 5},$$
$$p_{X_2}(x_2|\phi_{\leftarrow}) = \binom{15}{x_2} \phi_{\leftarrow}^{x_2} (1 - \phi_{\leftarrow})^{15 - x_2}.$$

Under a Bayesian approach, we introduce a prior $p(\phi_{\leftarrow})$ which represents our belief in $\phi_{\leftarrow}$ before any data is observed and infer the posterior over the unknown parameter $\phi_{\leftarrow}$ using Bayes' rule:

$$p(\phi_{\leftarrow}|x_1) \propto p(\phi_{\leftarrow}) \phi_{\leftarrow}^5 (1 - \phi_{\leftarrow})^{x_1 - 5},$$
$$p(\phi_{\leftarrow}|x_2) \propto p(\phi_{\leftarrow}) \phi_{\leftarrow}^{x_2} (1 - \phi_{\leftarrow})^{15 - x_2}.$$

To decide which action to take, the agent can marginalise over $\phi_{\leftarrow}$ using these posteriors to determine its best expected reward for choosing $\leftarrow$: $r(\leftarrow, x_i) = \int \phi_{\leftarrow} dP(\phi_{\leftarrow}|x_i)$. Given the observed random variable x_i, the agent thus picks:

$$a_{\text{Bayes}}(x_i) = \begin{cases} \leftarrow, & r(\leftarrow, x_i) > \phi_{\rightarrow}, \\ \rightarrow, & \text{otherwise.} \end{cases}$$

Under a frequentist approach, we must introduce a null hypothesis, $\mathcal{H}_0 : \quad \phi_{\leftarrow} < \phi_{\rightarrow}$ and an alternative hypothesis $\mathcal{H}_1 : \quad \phi_{\leftarrow} \geq \phi_{\rightarrow}$. Frequentist statistics does not specify a single formal decision bound based on the observed data and we must choose a reasonable

method. In our case, we suggest using the decision rule that rejects the null hypothesis if the observed variable is greater than the mode value under $\phi_{\leftarrow} = \phi_{\rightarrow}$, that is:

$$a_{\text{freq}}(x_i) = \begin{cases} \leftarrow, & x_i > \arg\max_{x'} p_{X_i}(x'|\phi_{\rightarrow}), \\ \rightarrow, & \text{otherwise.} \end{cases}$$

As X_1 is distributed according to a negative binomial distribution and X_2 is a binomial variable, the modes have established forms (see, for example, [DeGroot, 2004]):

$$a_{\text{freq}}(x_1) = \begin{cases} \leftarrow, & x_1 > 10 + \left\lfloor \frac{10\phi_{\rightarrow}}{1-\phi_{\rightarrow}} \right\rfloor, \\ \rightarrow, & \text{otherwise.} \end{cases}$$

$$a_{\text{freq}}(x_2) = \begin{cases} \leftarrow, & x_2 > \lceil 16\phi_{\rightarrow} \rceil - 1, \\ \rightarrow, & \text{otherwise.} \end{cases}$$

We now show that given the same observed data, a frequentist agent may chose different actions according to which random variable they observe, whereas a Bayesian agent choses the same action regardless. As an example, we consider an outcome where the agent has observed fifteen samples where there are five observations of $r_i = 1$ and ten observations of $r_i = 0$, hence $x_1 = 15$ and $x_2 = 5$. A frequentist agent observing the random variable $x_1 = 15$ will select the action $\leftarrow$, whereas a frequentist agent observing the random variable $x_2 = 5$ will select the action $\rightarrow$. For the Bayesian agent, the posterior for either random variable is:

$$p(\phi_{\leftarrow}|x_i) \propto p(\phi_{\leftarrow})\phi_{\leftarrow}^{5}(1-\phi_{\leftarrow})^{10},$$

and hence the Bayesian agent always makes the same decision regardless of the random variable chosen.

Our example highlights why frequentist approaches may not adhere to the likelihood principle. X_1 and X_2 are sufficient statistics as both contain the same information about the relative frequencies of rewards in the outcome (there were five rewards of 1 and ten rewards of 0) and this information is sufficient for inferring $\phi_{\leftarrow}$. We should therefore expect the agent to make the same decision regardless of which variable it observes. Despite this, the frequentist agent makes different inferences depending on whether it observes X_1 or X_2. We would have faced similar difficulties using the common p-value

approaches for a decision bound instead [Lindley and Phillips, 1976, Murphy, 2012]. If we look at the support for the variables X_1 and X_2, we see that X_1 is defined over the space of all observed rewards $\{r_1, r_2, ...\}$ whereas the variable X_2 is defined over the first fifteen reward observations $\{r_1, r_2, ...r_{15}\}$. In our example, the same inferences can be made from X_1 and X_2 after fifteen observations, however frequentist approaches allow decision rules based on the entire support of X_1, which will take into account unknown future observations that the agent has yet to observe. In this sense, frequentist decision rules can break the conditionality principle as they require unobserved data to make decisions in the present.

According to Edwards et al. [1963]: *'It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience'*. In the context of RL, this insight is particularly pertinent as frequentist agents that continue to collect data about the environment beyond a point at which inference can be made are wasteful in their exploration.

4.3 Approximate Inference

If the likelihood and the prior are conjugate, it is possible to obtain exact analytic expressions for the posterior: in the context of regression, this is possible if the mean function is in a linear function space, i.e. $\mu_\phi = \phi^\top b$, and the prior is Gaussian. Unfortunately linear Gaussian models are restrictive in the their ability to represent many functions of interest, and so we often require more complicated families of function spaces, such a neural networks. For nonlinear function spaces, the posterior cannot be calculated exactly as the integration in Eq. (4.1) required to calculate the evidence is intractable.

One solution is to obtain a point estimate that best represents the latent variable. A point estimate known as the maximum a posteriori (MAP) estimate can be obtained by finding the set of hiddens variables that maximise the posterior:

$$\Phi_{\text{MAP}} \coloneqq \arg\max_{\phi \in \Phi} p_\Phi(\phi|\mathcal{D}_N).$$

Unfortunately, by using point estimates we forfeit the ability to characterise the uncertainty that we have in an unknown variable. For this reason, MAP estimates are not considered truly Bayesian. We now introduce VI and randomised priors (RP), two approaches for approximating intractable inference problems that scale well to large datasets and estimate the uncertainty in the posterior.

4.3.1 Variational Inference

In VI [Jordan, 1999, Beal, 2003] we introduce a tractable (parametric[1]) variational posterior to approximate the true posterior $P_\Phi(\cdot|\mathcal{D}_N) \approx \hat{P}_\psi$, parametrised by $\psi \in \Psi$. The graphical model of a VI problem is shown in Fig. 4.2. We represent variational parameters as blue boxes.

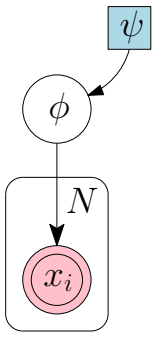


Figure 4.2: Variational inference

Recall from Section 2.3.3 that if D is a statistical distance, then $D(P_\Phi(\cdot|x)\|\hat{P}_\psi) = 0$ implies that $P_\Phi(\cdot|x) = \hat{P}_\psi$. Hence if we can find a set of variational parameters that minimises a convenient statistical distance between the variational and true posterior, our a tractable variational distribution will be equivalent to the intractable posterior. The most common statistical distance used for VI is the KL dive, which is well defined if $\hat{P}_\psi$ is absolutely continuous with respect to the true posterior. Under this condition, the KL divergence offers an intuitive measure of similarity between the posterior and variational distribution that we seek to minimise:

$$\psi^* \in \arg\min_{\psi\in\Psi} \mathrm{KL}(\hat{P}_\psi \parallel P_\Phi(\cdot|x)). \tag{4.2}$$

Unfortunately the KL divergence as written in Eq. (4.2) is still intractable, however we

[1] Although nonparametric VI models exist [Gershman et al.], the majority of approaches, including this thesis, assume a parametric distribution.

can expand it to derive a tractable, equivalent alternative objective:

$$\begin{aligned}
\mathrm{KL}(\hat{P}_\psi \parallel P_\Phi(\cdot|x)) &= \int \log\left(\frac{\hat{p}_\psi(\phi)}{p_\Phi(\phi|\mathcal{D}_N)}\right) d\hat{P}_\psi(\phi), \\
&= \int \log\left(\frac{\hat{p}_\psi(\phi)p(\mathcal{D}_N)}{p(\mathcal{D}_N,\phi)}\right) d\hat{P}_\psi(\phi), \\
&= \log p(\mathcal{D}_N) - \int \log\left(\frac{p(\mathcal{D}_N,\phi)}{\hat{p}_\psi(\phi)}\right) d\hat{P}_\psi(\phi), \\
&= \log p(\mathcal{D}_N) - \mathcal{L}(\psi|\mathcal{D}_N),
\end{aligned}$$

where the second term $\mathcal{L}(\psi|\mathcal{D}_N)$ is the evidence lower bound (ELBO) . As the log evidence $\log p(\mathcal{D}_N)$ is independent of ψ and finite, minimising $\mathrm{KL}(\hat{P}_\psi \parallel P_\Phi(\cdot|x))$ is equivalent to maximising the ELBO , which is tractable. We write the ELBO as:

$$\begin{aligned}
\mathcal{L}(\psi|\mathcal{D}_N) &\coloneqq \int \log\left(\frac{p(\mathcal{D}_N,\phi)}{\hat{p}_\psi(\phi)}\right) d\hat{P}_\psi(\phi), \\
&= \int \log\left(\frac{p(\mathcal{D}_N|\phi)p_\Phi(\phi)}{\hat{p}_\psi(\phi)}\right) d\hat{P}_\psi(\phi), \\
&= \left(\int \log\left(p(\mathcal{D}_N|\phi)\right) + \log\left(\frac{p_\Phi(\phi)}{\hat{p}_\psi(\phi)}\right) d\hat{P}_\psi(\phi)\right), \\
&= \int \log p(\mathcal{D}_N|\phi) d\hat{P}_\psi(\phi) + \mathrm{KL}(q_\psi \parallel p_\Phi),
\end{aligned}$$

where the KL divergence is often interpreted as a regularisation term.

4.3.2 Marginal Likelihood and Expectation Maximisation

For certain problems we may want to optimise the likelihood with respect to some non-Bayesian parameters ψ but still marginalise over a set of Bayesian parameters ϕ. The likelihood for this problem is:

$$p_X(x|\phi,\psi),$$

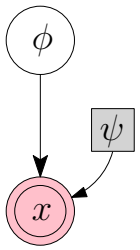


Figure 4.3: Marginal likelihood

which is represented by the graphical model in Fig. 4.3. We represent non-Bayesian parameters as grey boxes. As ϕ is a Bayesian parameter, we can infer the posterior over its value $P_\Phi(\cdot|x)$ as outlined in Section 4.1. We then use this posterior to marginalise over ϕ, obtaining the marginal likelihood:

$$p_X(x|\psi) = \int p_X(x|\phi,\psi) dP_\Phi(\phi|x).$$

Maximising the marginal likelihood allows us to obtain a point estimate that is partially Bayesian, thus accounting for the uncertainty in part of our model. This can be useful in situations where it is impracticable to infer a posterior over all parameters of a model. For mathematical convenience, the log-marginal likelihood (LML) $\log p_X(x|\psi)$ is optimised instead.

If the posterior is intractable, it is not possible to maximise the LML directly. We can borrow the techniques introduced in Section 4.3.1 to derive a tractable lower bound instead. For a probability distribution $\hat{P}_\Phi$ with support over Φ we can rewrite the LML as a difference of two divergences [Jordan, 1999],

$$\begin{aligned}\log p_X(x|\psi) =& \int \log\left(\frac{p_X(x|\phi,\psi)p_\Phi(\phi)}{\hat{p}_\Phi(\phi)}\right) d\hat{P}_\Phi(\phi) - \int \log\left(\frac{p_\Phi(\phi|x)}{\hat{p}_\Phi(\phi)}\right) d\hat{P}_\Phi(\phi), \\ =& \mathcal{L}(\psi|x) + \mathrm{KL}(\hat{P}_\Phi \parallel P_\Phi(\cdot|x)).\end{aligned}$$

When the ELBO and marginal likelihood are identical, the resulting KL divergence between the distribution $\hat{P}_\Phi$ and the posterior $P_\Phi(\cdot|x)$ is zero, implying that $\hat{P}_\Phi = P_\Phi(\cdot|x)$. For this reason, $\hat{P}_\Phi$ behaves in exactly the same way as a variational posterior distribution.

Maximising the LML now reduces to maximising the ELBO, which can be achieved iteratively using the expectation maximisation (EM) algorithm [Dempster et al., 1977, Wu, 1983]: an expectation step (E-step) finds the posterior for the current set of model parameters and then a maximisation step (M-step) maximises the ELBO with respect to ω while keeping $\hat{P}_\Phi$ fixed as the posterior from the E-step.

4.3.3 Randomised Priors

We now present an alternative approach to approximate inference known as randomised priors (RP). The essence of RP is that we characterise uncertainty by injecting noise into the MAP estimate through the prior. Although this approach was first proposed by Osband et al. [2018], we present novel derivation here.

To motivate RP, we consider a linear Gaussian regression model: $P_Y(\cdot|x,\phi) = \mathcal{N}(\phi^\top v(x), \sigma^2)$ for some fixed σ^2 where $\phi \in \mathbb{R}^n$ and $v : \mathcal{X} \to \mathbb{R}^n$ is a basis vector.

We introduce a conjugate prior, which is also Gaussian: $P_\Phi = \mathcal{N}(\phi_0, \Sigma_0)$. Given a dataset $\mathcal{D}_N \coloneqq \{x_i, y_i\}_{i\in\{1:N\}}$ of observations, a linear Gaussian regression model with a conjugate prior is known to have an analytic posterior which is also Gaussian (see, for example, Murphy [2012] for a derivation):

$$P_\Phi(\cdot|\mathcal{D}_N) = \mathcal{N}(\phi_N, \Sigma_N).$$

where

$$\Sigma_N = \left(\Sigma_0^{-1} + \frac{1}{\sigma^2}\sum_{i=1}^{N} v(x_i)v^\top(x_i)\right)^{-1},$$
$$\phi_N = \Sigma_N\left(\Sigma_0^{-1}\phi_0 + \frac{1}{\sigma^2}\sum_{i=1}^{N} v(x_i)y_i\right).$$

To derive the sampling process for RP we introduce a change of variables $\phi = g_N : \mathcal{E} \to \Phi$. P_E is the corresponding pullback measure of the posterior, which satisfies:

$$\int f(\phi)dP_\Phi(\phi|\mathcal{D}_N) = \int f(g_N(\epsilon))dP_E(\epsilon). \tag{4.3}$$

Like the related concept of normalising flows [Rezende and Mohamed, 2015], we hope to find a tractable transformation of variables g_N such that the pullback P_E has a convenient form. From Eq. (4.3), sampling from $\epsilon \sim P_E$ and then applying the transformation of variables $\phi = g_N(\epsilon)$ is equivalent to sampling from the posterior.

Consider the function $g_N(\epsilon) \coloneqq \phi_N + \Sigma_N\Sigma_0^{-1}\epsilon$. For the linear Gaussian model, we can find an analytic solution for the pullback measure through this function by using the change of variables formula:

$$\begin{aligned}
&\int f(\phi)dP_\Phi(\phi|\mathcal{D}_N) \\
&= \frac{1}{\sqrt{(2\pi)^m \det\left(\Sigma_0\Sigma_N^{-1}\Sigma_0\right)}} \int f\circ g(\epsilon)\exp\left(-\epsilon^\top\Sigma_0^{-1}\Sigma_N\Sigma_0^{-1}\epsilon\right)d\lambda(\epsilon), \\
&= \int f\circ g(\epsilon)dP_E(\epsilon),
\end{aligned}$$

where $P_E = \mathcal{N}(0, \Sigma_0^{-1}\Sigma_N\Sigma_0^{-1})$. In the original presentation of the algorithm Osband et al. [2018], this distribution is approximated as $P_E \approx \mathcal{N}(0, I\sigma_{\text{rp}}^2)$, where σ_{rp}^2 is a hyper-parameter

To derive a randomisation of the MAP estimate, we rewrite the transformation function as:

$$\begin{aligned} g_N(\epsilon) &= \phi_N + \Sigma_N \Sigma_0^{-1} \epsilon, \\ &= \Sigma_N \left(\Sigma_0^{-1} \phi_0 + \frac{1}{\sigma^2} \sum_{i=1}^{N} v(x_i) y_i + \Sigma_0^{-1} \epsilon \right), \\ &= \Sigma_N \left(\Sigma_0^{-1} (\phi_0 + \epsilon) + \frac{1}{\sigma^2} \sum_{i=1}^{N} v(x_i) y_i \right). \end{aligned} \tag{4.4}$$

As shown by Osband et al. [2018], Eq. (4.4) is the solution to the optimisation problem:

$$\begin{aligned} g_N(\epsilon) &\in \arg\min_{\phi} \left(\frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y_i - \phi^\top v(x_i) \right)^2 + ((\phi - \epsilon) - \phi_0)^\top \Sigma_0^{-1} ((\phi - \epsilon) - \phi_0) \right), \\ &= \arg\max_{\phi} \log \left(p(\mathcal{D}_N | \phi) p_\Phi(\phi - \epsilon) \right), \end{aligned}$$

which we can interpret as the MAP estimate obtained by maximising an objective with a randomised prior $p_\Phi(\phi - \epsilon)$. To derive a prior-randomised form for the posterior, we re-write Eq. (4.3) using the Dirac-delta measure:

$$\begin{aligned} \int f(g_N(\epsilon)) dP_E(\epsilon) &= \int f(\phi) \int d\delta_{g_N(\epsilon)} dP_E(\epsilon), \\ \implies p_\Phi(\phi | \mathcal{D}_{N,\omega}) &= \int \delta_{g_N(\epsilon)} dP_E(\epsilon). \end{aligned} \tag{4.5}$$

Taking a step back to analyse Eq. (4.5), we see that the posterior for a linear Gaussian model can be written as an MAP estimate $\delta_{g_N(\epsilon)}$ that has been randomised by a pullback measure that is superficially similar to the prior. For nonlinear Gaussian models, the true pullback measure P_E under the transformation of variables $g_N(\epsilon)$ may not exist or may be intractable. In this case Eq. (4.5) does not hold exactly, but may still provide a good approximation:

$$P_\Phi(\cdot | \mathcal{D}_{N,\omega}) \approx \hat{P}_\Phi(\cdot | \mathcal{D}_{N,\omega}) = \int \delta_{g_N(\epsilon)} dP_E(\epsilon),$$

When sampling the RP approximate posterior, we muust first sample a variable $\epsilon \sim P_E$ and then solve the corresponding prior-randomised MAP maximisation problem $g_N(\epsilon) = \arg\max_\phi \log \left(p(\mathcal{D}_N | \phi) p_\Phi(\phi - \epsilon) \right)$. This method is intractable for nonlinear models as solving the nonlinear optimisation problem is NP-hard. To circumvent this issue, we can

use an ensembling method to sample from the approximate posterior instead. Consider the expectation under the approximate posterior for a test function $f(\phi)$. Given a set of L samples $\epsilon_l \sim P_E$, we can employ a Monte-Carlo sampling approach to estimate this expectation:

$$\int f(\phi) d\hat{P}_{\Phi}(\phi|\mathcal{D}_{N,\omega}) = \int f \circ g_N(\epsilon) dP_E(\epsilon) \approx \frac{1}{L}\sum_{l=1}^{L} f \circ g_N(\epsilon_l) = \frac{1}{L}\sum_{l=1}^{L} f(\phi_i),$$

where each $\phi_l \in g_N(\epsilon_l)$ is a solution to one of L individual optimisation problem $g_N(\epsilon_l)$. Therefore by maintaining a set of L optimal parameters, $\{\phi_l\}_{l=1:L}$, we can always obtain an estimation to posterior marginalisation. This approach is known as an *ensembling*.

5
Control as Inference

This chapter expands the presentation of the VIREL *framework in [Fellows et al., 2019]. Empirical work was carried out in collaboration with Anuj Mahajan.*

Control as inference attempts to recast the RL problem as one of inference. In doing so, we hope to exploit powerful algorithms that exist to obtain tractable inference approximations in the context of RL, opening up a space of novel algorithms and furthering our understanding of the control problem. As we shall demonstrate, many of these approaches introduce a form of regularisation to the control problem, which has led to a new generation of deep RL algorithms with proven empirical benefits over their unregularised analogues Haarnoja et al. [2017, 2018a], Abdolmaleki et al. [2018a], Haarnoja et al. [2019].

Unfortunately, several of these algorithms introduce a number of theoretical issues due to the additional regularisation in the objective or the choice of underlying divergence used to obtain tractable algorithms. We now review the two major existing families of approaches to control as inference which characterise learning an optimal policy as either posterior inference or maximum marginal likelihood (MML) estimation with the aim to elucidate the various shortfalls of each approach

5.1 Existing Approaches

Existing approaches to control as inference make an analogy between Bayesian inference and solving the control problem by introducing a random binary variable $\mathcal{O} \in \{0, 1\}$ with density:

$$p(\mathcal{O}|s_{t+1}, a_t, s_t) \propto \exp\left(\frac{\gamma^t r(s_{t+1}, a_t, s_t)}{\alpha}\right)^{\mathcal{O}} \left(1 - \exp\left(\frac{\gamma^t r(s_{t+1}, a_t, s_t)}{\alpha}\right)\right)^{1-\mathcal{O}},$$

where $\alpha \geq 0$ is a temperature coefficient. The role of the temperature coefficient has a subtly different meaning depending upon whether a MML or posterior inference problem is defined. We will discuss the semantics of α in both contexts.

If $p(\mathcal{O}|s', a, s)$ was a density of an actual random variable, we would require that all rewards are negative to avoid violating the Kolmogorov axioms (see Definition 8 in Appendix A.2 for an overview), a condition that can be satisfied if rewards are upper-bounded without changing the optimal policy. However, we emphasise that because control as inference is only a metaphorical construction designed to borrow established optimisation techniques from VI, it is of little concern whether $p(\mathcal{O}|s', a, s)$ characterises a proper distribution provided the objectives that we derive are well-defined.

To derive an inference problem whose solution generates an optimal policy for the RL problem, we condition on $\mathcal{O} = 1$. Given an entire trajectory of length, we can find the probability $\mathcal{O} = 1$ as:

$$p(\mathcal{O} = 1|\tau) \propto \prod_{t=0}^{\infty} \exp\left(\frac{\gamma^t r_t}{\alpha}\right) = \exp\left(\sum_{t=0}^{\infty} \frac{\gamma^t r_t}{\alpha}\right). \tag{5.1}$$

In relation to classic Bayesian inference, the distribution $p(\mathcal{O} = 1|\tau)$ is treated as the *likelihood*, although any semantic similarity is only superficial. For notational convenience, we drop the explicit notation $\mathcal{O} = 1$ and assume that our distributions are conditioned on $\mathcal{O} = 1$ from hereon out. A choice of prior over the trajectory completes the model. We now detail how solving either a problem of posterior inference or MML using this model generates an RL objective from which an optimal policy can be learnt.

5.1.1 Maximum Marginal Likelihood as Control

Studying Eq. (5.1), we see that the use of an exponential transformation on the rewards ensures that $p(\mathcal{O} = 1|\tau)$ is a monotonic transformation of the total return for the trajectory. To derive an MML problem, we treat the trajectory distribution $P_{\mathcal{T}}^{\pi_\phi}$ for a parametrised policy π_ϕ as a prior over trajectories :

$$p_\phi(\tau) = p_{\mathcal{T}}^{\pi_\phi}(\tau) = p_0(s_0) \prod_{t=0}^{\infty} \pi_\phi(a_t|s_t) p(s_{t+1}|s_t, a_t).$$

Using the prior to marginalise over all trajectories, we obtain the marginal likelihood density:

$$\begin{aligned} p_\phi(\mathcal{O}) &= \int p(\mathcal{O}|\tau) dP_\phi(\tau), \\ &= \int \exp\left(\sum_{t=0}^{\infty} \frac{\gamma^t r_t}{\alpha}\right) dP_\phi(\tau). \end{aligned} \tag{5.2}$$

Because of our choice of prior, the objective in Eq. (5.2) is identical to the frequentist RL objective defined in Eq. (3.2), with the application of an exponential transformation on the return. As the exponential function is monotonic, the transformation does not affect the set of optimal policies:

$$\phi^\star \in \arg\max_{\phi \in \Phi} p_\phi(\mathcal{O}) \iff \phi^\star \in \arg\max_{\phi \in \Phi} J(\phi).$$

The marginal likelihood in Eq. (5.2) is no easier to optimise than the frequentist RL objective, however due to our inference formulation, we have the advantage of being able to draw on the powerful optimisation tools that have been used to solve MML problems. In particular, we can introduce a variational distribution $\hat{P}_\theta$ parametrised by $\theta \in \Theta$ to derive an ELBO objective on the marginal likelihood (see Section 4.3.2 for a derivation of this bound):

$$\begin{aligned} \mathcal{L}(\phi, \theta) &= \alpha \int \log\left(\frac{p(\mathcal{O}|\tau) p_\phi(\tau)}{\hat{p}_\theta(\tau)}\right) d\hat{P}_\theta(\tau), \\ &= \int \left(\sum_{t=0}^{\infty} \gamma^t r_t\right) d\hat{P}_\theta(\tau) - \alpha \mathrm{KL}(\hat{P}_\theta \parallel P_\phi). \end{aligned} \tag{5.3}$$

To optimise our ELBO objective, we seek a solution to the bilevel optimisation problem:

$$\phi^\star \in \arg\max_{\phi\in\Phi} \left(\max_{\theta\in\Theta} \mathcal{L}(\phi, \theta) \right).$$

If there always exists some prior parametrisation $\phi^\star$ such that $\lim_{\phi\to\phi^\star} \mathrm{KL}(\hat{P}_\theta \parallel P_\phi) = 0$ for any variational parameter $\theta \in \Theta$ the resulting set of parameters

$$\Theta^\star(\theta) \coloneqq \{\phi^\star : \lim_{\phi\to\phi^\star} \mathrm{KL}(\hat{P}_\theta \parallel P_\phi) = 0\},$$

will maximise the outer level optimisation problem. To ensure that a such set of solutions exists, we can introduce a variational policy π_θ from the same parametric family as π_ϕ and choose the variational distribution to be the distribution of a trajectory under π_θ:

$$\hat{p}_\theta(\tau) \coloneqq p_0(s_0) \prod_{t=0}^{\infty} \pi_\theta(a_t|s_t) p(s_{t+1}|s_t, a_t). \tag{5.4}$$

Using the variational form given in Eq. (5.4), the ELBO objective from Eq. (5.3) becomes exactly the classic RL objective augmented with a KL divergence regularisation term whose influence is controlled by α. To highlight this distinction, we can re-write the ELBO objective in Eq. (5.3) as:

$$\int \left(\sum_{t=0}^{\infty} \left(\gamma^t r_t - \alpha \mathrm{KL}(\pi_\theta(\cdot|s_t) \parallel \pi_\phi(\cdot|s_t)) \right) \right) d\hat{P}_\theta(\tau). \tag{5.5}$$

Compared to the classic RL objective, the additional KL divergence introduced penalises variational policy updates that stray too far from π_ϕ in the inner optimisation, $\theta^\star \in \arg\max \mathcal{L}(\theta, \phi)$. The degree of this penalisation is controlled by the parameters α where the classic RL objective can be recovered in the limit $\alpha \to 0$. Observe that any optimal policy $\pi_{\phi^\star}$ such that $\phi^\star \in \arg\max_{\phi\in\Phi} (\max_{\theta\in\Theta} \mathcal{L}(\phi, \theta))$ is a solution to the frequentist RL objective whenever $\Theta = \Theta$: the optimal outer level solution is obtained by simply setting $\phi = \theta$ for any $\theta \in \Theta$. Substituting for $\phi = \theta$ into $\mathcal{L}(\phi, \theta)$ yields:

$$\phi^\star \in \arg\max_{\theta\in\Theta} \int \sum_{t=0}^{\infty} \gamma^t r_t d\hat{P}_\theta(\tau) = \arg\max_{\theta\in\Theta} J(\theta),$$

hence additional the KL regularisation does not introduce bias into RL objective if the variational distribution is chosen to be Eq. (5.4).

Existing algorithms find an optimal policy using (variational) EM [Dempster et al., 1977, Wu, 1983], an algorithm designed to tackle the underlying MML problem by iteratively optimising $\mathcal{L}(\phi, \theta)$ for θ and ϕ. Existing algorithms thus vary in their choice of prior, variational distribution and method of optimising $\mathcal{L}(\phi, \theta)$ in each iterative step Dayan and Hinton [1997], Toussaint and Storkey [2006], Peters and Schaal [2007], Hachiya et al. [2009], Neumann [2011], Abdolmaleki et al. [2018b].

An unfortunate drawback of algorithms under the MML umbrella is that they can generate policies that are risk-seeking [Levine, 2018a]. This stems from the definition of the underlying marginal likelihood that these methods maximise, which we repeat from Eq. (5.2) for convenience:

$$p_\phi(\mathcal{O}) = \int \exp\left(\sum_{t=0}^{\infty} \frac{\gamma^t r_t}{\alpha}\right) dP_\phi(\tau).$$

In comparison to the classic RL objective, the exponential transformation is a re-weighting on rewards that minimises the influence of very low returns. Whereas a trajectory with a return of $\sum_{t=0}^{\infty} \frac{\gamma^t r_t}{\alpha} = -10^6$ would be penalised heavily under the classic RL objective, the same trajectory would only a incur cost of $\exp(-10^6) \approx 0$ under the marginal likelihood objective. The exponential transformation also increases the influence of trajectories with very high returns on the optimisation problem, thus encouraging policies to generate trajectories high returns. This strategy is risky, as even small variations from the exact optimal policy could lead to returns whose cost has been downplayed by exponential re-weighting.

Following the analysis of Neumann [2011], we can probe this issue further by analysing the variational distribution $\hat{P}_\theta$ that was introduced to solve the MML problem. Recall from Section 4.3.2 that the role of the variational distribution $\hat{P}_\theta$ is to approximate an intractable posterior. From Bayes' rule, this posterior has the density:

$$\begin{aligned} p_\phi(\tau|\mathcal{O}) &\propto p(\mathcal{O}|\tau) p_\phi(\tau) \\ &= \exp\left(\sum_{t=0}^{\infty} \frac{\gamma^t r_t}{\alpha}\right) p_\phi(\tau), \\ &= \exp\left(\frac{R(\tau)}{\alpha}\right) p_\phi(\tau), \end{aligned}$$

where $R(\tau)$ is the discounted return. In Section 4.3.1 we proved that introducing a variational distribution $\hat{P}_\theta$ and optimising the ELBO is equivalent to minimising the divergence $\mathrm{KL}(P_\phi(\cdot|\mathcal{O}) \parallel \hat{P}_\theta)$. This form of KL divergence encourages the variational posterior to be moment-matching, spreading mass across the entirety of target distribution's density [Murphy, 2012]. In contrast, minimising the RL objective can be shown to be equivalent to minimising a mode-seeking form of KL divergence:

$$
\begin{aligned}
\theta^\star &\in \arg\max_{\theta\in\Theta} J(\theta), \\
&= \arg\max_{\theta\in\Theta} \int R(\tau) dP_\theta(\tau), \\
&= \arg\min_{\theta\in\Theta} \int -\log \exp\left(\frac{R(\tau)}{\alpha}\right) dP_\theta(\tau), \\
&= \arg\min_{\theta\in\Theta} \left(\int \log\left(\frac{p_\theta(\tau)}{p_\theta(\tau)}\right) - \log\exp\left(\frac{R(\tau)}{\alpha}\right) dP_\theta(\tau) \right), \\
&= \arg\min_{\theta\in\Theta} \left(\int \log\left(\frac{p_\theta(\tau)}{p_\theta(\tau|\mathcal{O})}\right) dP_\theta(\tau) \right), \\
&= \arg\min_{\theta\in\Theta} \mathrm{KL}(P_\theta \parallel P_\theta(\cdot|\mathcal{O})),
\end{aligned}
$$

where:

$$
p_\theta(\tau|\mathcal{O}) \propto p(\mathcal{O}|\tau) p_\theta(\tau) = \exp\left(\frac{R(\tau)}{\alpha}\right) p_\theta(\tau).
$$

If the optimal policy can be represented accurately in the class of policy distributions, the problem is fully observable, the learning algorithm converges to a global minimum and the returns can be represented exactly, then the optimal policy is the same for minimising both divergences. In the more realistic setting where these conditions are not satisfied, we provide a sketch in Fig. 5.1 to explain the difference between optimising these two variations of KL divergence.

The sketch shows a simplified contour plot of $p_\theta(\tau|\mathcal{O})$, overlayed with two example Gaussian variational distributions learnt from the RL and marginal likelihood objectives. When using the mode-seeking RL objective, the variational policy will be attracted towards a peak of the target density to avoid placing mass in regions where $p_\theta(\tau|\mathcal{O})$ is low. Conversely, under the mass-covering marginal likelihood objective, the variational policy will spread its mass evenly across both peaks of $p_\theta(\tau|\mathcal{O})$. As $p_\theta(\tau|\mathcal{O})$ is proportional to

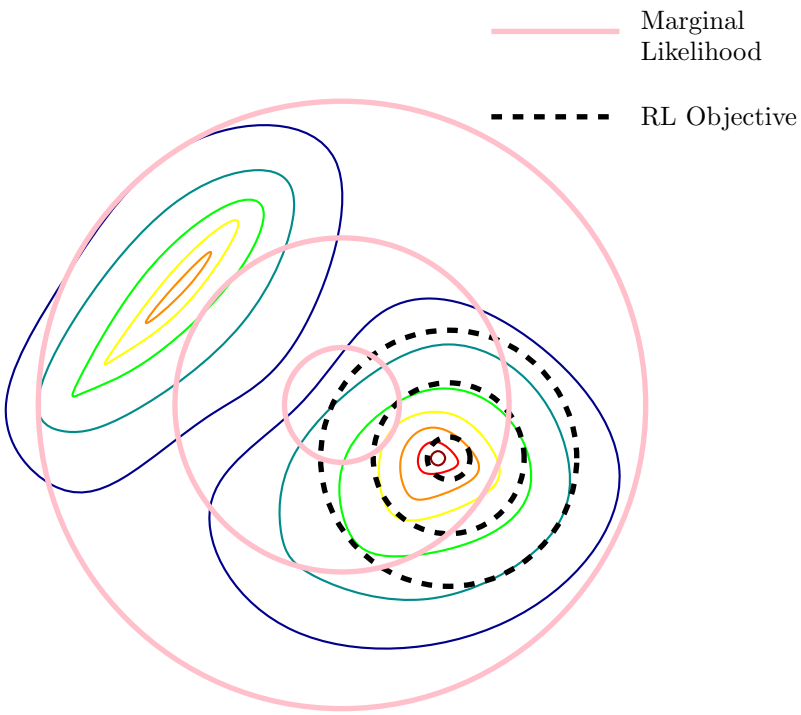


Figure 5.1: Sketch of a variational distribution learnt using a marginal likelihood vs RL objective

the trajectory-weighted returns, its mean will be centred on a region of very low expected return. Typically, a deterministic policy is constructed from the variational mean policy for deployment. In the example given, this will direct the agent towards trajectories with very low returns as only regions of high return are taken into account by the variational objective. For these reasons minimising a mass-covering divergence like the marginal likelihood objective is often inappropriate for RL.

5.1.2 Posterior Inference as Control

To avoid the risk-seeking behaviour associated with the MML approach, posterior inference as control aims to infer the posterior $P_\phi(\cdot|\mathcal{O})$ by minimising the mode-seeking form of KL divergence. The derivation of the posterior inference as control framework follows the MML approach introduced in Section 4.3.2: the prior P_ϕ is

$$p_\phi(\tau) = p_{\mathcal{T}}^{\pi_\phi}(\tau) = p_0(s_0) \prod_{t=0}^{\infty} \pi_\phi(a_t|s_t) p(s_{t+1}|s_t, a_t),$$

a variational policy π_θ is introduced and the variational distribution $\hat{P}_\theta$ is the distribution over trajectories under π_θ:

$$\hat{p}_\theta(\tau) = p_0(s_0) \prod_{t=0}^{\infty} \pi_\theta(a_t|s_t) p(s_{t+1}|s_t, a_t).$$

Unlike the MML approach, our goal is to infer the posterior $p_\phi(\tau|\mathcal{O}) \propto p(\mathcal{O}|\tau)p_\phi(\tau)$ by minimising the reverse KL divergence $\arg\min_{\theta\in\Theta} \mathrm{KL}(\hat{P}_\theta \parallel P_\phi(\cdot|\mathcal{O}))$ thereby avoiding the issues that stem from minimising the moment-matching forward KL divergence. For posterior inference as control, the KL divergence has a tractable form:

$$\begin{aligned}
\mathrm{KL}(\hat{P}_\theta \parallel P_\phi(\cdot|\mathcal{O})) &= \int \log \hat{p}_\theta(\tau) d\hat{P}_\theta(\tau) - \int \log p_\phi(\tau|\mathcal{O}) d\hat{P}_\theta(\tau), \\
&= \int \log \left(\frac{\hat{p}_\theta(\tau)}{p_\phi(\tau)} \right) d\hat{P}_\theta(\tau) - \int \log p(\mathcal{O}|\tau) d\hat{P}_\theta(\tau) + \log p_\phi(\mathcal{O}), \\
&= \mathrm{KL}(\hat{P}_\theta \parallel P_\phi) - \int \log p(\mathcal{O}|\tau) d\hat{P}_\theta(\tau) + \log p_\phi(\mathcal{O}).
\end{aligned}$$

As the prior is considered fixed and our goal is to infer the variational posterior, we can ignore the *log*-marginal likelihood term $\log p_\phi(\mathcal{O})$ as it has no dependence on θ. Scaling by α, we can thus maximise the same ELBO from the MML approach:

$$\begin{aligned}
\mathcal{L}(\theta|\phi) &= \int \log p(\mathcal{O}|\tau) d\hat{P}_\theta(\tau) - \alpha \mathrm{KL}(\hat{P}_\theta \parallel P_\phi), \\
&= \int \sum_{t=0}^{\infty} \gamma^t r_t d\hat{P}_\theta(\tau) - \alpha \mathrm{KL}(\hat{P}_\theta \parallel P_\phi), \\
&= \int \left(\sum_{t=0}^{\infty} \left(\gamma^t r_t - \alpha \mathrm{KL}(\pi_\theta(\cdot|s_t) \parallel \pi_\phi(\cdot|s_t)) \right) \right) d\hat{P}_\theta(\tau). \qquad (5.6)
\end{aligned}$$

Although both the MML and posterior inference approaches use the same objective, in posterior inference the prior is considered fixed and hence the objective is only maximised with respect to the variational parameters. The prior is a free choice, however the overwhelming majority of algorithms under the posterior inference as control approach maximise the maximum entropy reinforcement learning (MERL) objective [Toussaint, 2009, Rawlik et al., 2012, Haarnoja et al., 2018a, 2017, Chen et al., 2021], which is derived by assuming a uniform prior policy, $\pi_\phi(\cdot|s) = \mathrm{Unif}(\cdot)$ [Levine, 2018a]. Making this substitution into Eq. (5.6) yields the MERL objective:

$$J_{\text{merl}}(\theta) \coloneqq \int \left(\sum_{t=0}^{\infty} \left(\gamma^t r_t + \alpha \mathcal{H}(\pi_\phi(\cdot|s_t)) \right) \right) d\hat{P}_\theta(\tau), \qquad (5.7)$$

where $\mathcal{H}(\cdot)$ denotes the (differential) entropy. Sometimes the MERL objective is written in a different but equivalent form, where the expectation over π_θ in the entropy is absorbed into the outer expectation over trajectories:

$$J_{\text{merl}}(\theta) = \int \left(\sum_{t=0}^{\infty} \left(\gamma^t r_t - \alpha \log \pi_\phi(a_t|s_t) \right) \right) d\hat{P}_\theta(\tau).$$

We see from Eq. (5.7) that the MERL approach introduces an additional entropy term into the RL objective that encourages exploration by preventing policies collapsing prematurely: when a policy is very confident about a particular action, it will tend towards a Dirac-delta distribution. This will cause the entropy term to become very large and negative, penalising this confidence. The entropy term is the largest when all actions have equal probability, encouraging policies to have a large variance (and hence encouraging exploration) until they are confident enough that the increase in expected reward outweighs the penalisation factor of the entropy term.

The MERL objective was first studied by Ziebart et al. [2008, 2010] who showed that it can be solved using the same techniques introduced in Chapter 3 to solve the unregularised RL problem. The MERL optimal policy $\pi^\star_{\text{merl}} \in \arg\max_{\theta \in \Theta} J_{\text{merl}}(\theta)$ has the density:

$$\pi^\star_{\text{merl}}(a|s) \propto \exp\left(\frac{Q^\star_{\text{soft}}(s,a)}{\alpha} \right),$$

where

$$Q^\star_{\text{soft}}(s,a) \coloneqq \mathbb{E}_{P^{\pi^\star_{\text{merl}}}(\tau|s,a)} \left[r(s_1,a,s) + \sum_{t=1}^{\infty} \left(\gamma^t r_t + \alpha \mathcal{H}(\pi^\star_{\text{merl}}(\cdot|s_t)) \right) \right]$$

In posterior inference as control, we see that by controlling the degree of regularisation, the role of α is to *soften* the max operation used to define an optimal policy and optimal Q-function under the unregularised RL objective: indeed, our result in Theorem 1 can be used to formally verify that in the limit $\alpha \to 0$, $\pi^\star_{\text{merl}} \rightharpoonup \pi^\star$ and $Q^\star_{\text{soft}} \to Q^\star$. For this reason, $Q^\star_{\text{soft}}$ is known as a soft optimal Q-function. Any soft optimal Q-function also satisfies a soft optimal Bellman equation $Q^\star_{\text{soft}} = \mathcal{B}^\star_{\text{soft}}[Q^\star_{\text{soft}}]$ [Ziebart, 2010] where the soft optimal Bellman operator is defined as:

$$\mathcal{B}^\star_{\text{soft}}[Q^\star_{\text{soft}}](s,a) = \mathbb{E}_{P(s'|s,a)} \left[r(s',a,s) + \alpha\gamma \log \int \exp\left(\frac{Q^\star_{\text{soft}}(s',a')}{\alpha} \right) d\text{Unif}(a') \right].$$

Similarly, for any other policy π, we can define the soft Q-function as

$$Q^{\pi}_{\text{soft}}(s,a) \coloneqq \mathbb{E}_{P^{\pi}(\tau|s,a)}\left[r(s_1,a,s) + \sum_{t=1}^{\infty}\left(\gamma^t r_t + \alpha\mathcal{H}(\pi(\cdot|s_t))\right)\right],$$

and the soft Bellman operator as:

$$\mathcal{B}_{\text{soft}}[Q^{\star}_{\text{soft}}](s,a) = \mathbb{E}_{P(s'|s,a)}\left[r(s',a,s) + \gamma\int\left(Q^{\pi}_{\text{soft}}(s',a')\alpha - \alpha\log\pi(a'|s')\right)d\pi(a'|s')\right].$$

Fox et al. [2016], Haarnoja et al. [2018a] prove an analogous result to Proposition 5 for the soft Bellman operators, showing that they are contraction mappings. This justifies applying techniques developed to solve the RL in the maximum entropy setting. Soft analogues of established RL algorithms such as soft Q-learning [Haarnoja et al., 2017], soft actor-critic (SAC) [Haarnoja et al., 2018a,b, 2019] have shown strong empirical performance across MuJoCo continuous control tasks [Brockman et al., 2016].

To avoid minimising the mass-spreading form of KL divergence, we have shown that MERL policies optimise the ELBO with a fixed prior which introduces bias into the RL objective in the form of an additional entropy term. This bias encourages stochasticity in policies which can lead to improved exploration during training, however the optimal policy under the MERL is not naturally deterministic. A common solution is to construct a deterministic behavioural policy at test time from the mode (or mean) of the exploration policy, which is then used for deployment [Haarnoja et al., 2017, 2018a]. Because of the bias introduced into the objective, this deterministic behavioural policy is not guaranteed to be an optimal policy for the MDP that the agent is in and, as we show by way of counter example in Section 5.1.3, can lead to the worst possible actions being taken for some MDPs. Constraining the variational policies to the set of Dirac-delta distributions does not solve this problem either, since it renders the inference procedure intractable [Rawlik et al., 2010, 2012].

5.1.3 A Counterexample for MERL Optimal Policies

This counterexample was devised by Anuj Mahajan.

To demonstrate how poorly a MERL optimal policy can perform, consider the discrete state MDP as shown in Fig. 5.2, with action set $\mathcal{A} = \{\uparrow, \downarrow, a_1, \ldots a_M\}$ and state set

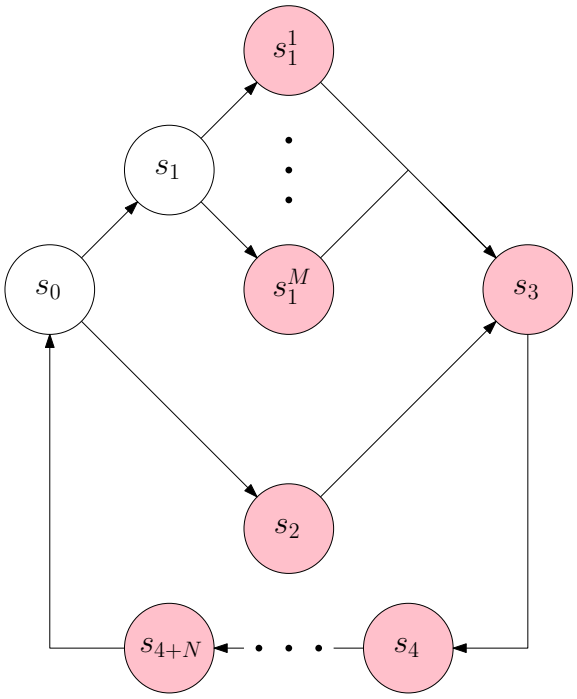


Figure 5.2: A discrete counterexample MDP for MERL optimal policies

$\mathcal{S} = \{s_0, s_1, s_1^1, \cdots s_1^M, s_2, s_3, s_4, \cdots s_{4+N}\}$. All state transitions from a pink state are deterministic and independent of the action taken. In a white state, the agent transitions deterministically to its action state, that is $P(s_1|s_0, \uparrow) = 1$, $P(s_2|s_0, \downarrow) = 1$ and $P(s_1^m|s_1, a_m) = 1 \; \forall \; m \in \{1 : M\}$. The reward function is $r(s_0, \downarrow) = 1$ and zero otherwise. The RL optimal policy has $\pi^*(a_2|s_0) = 1$.

Let $\theta \coloneqq (p_0, p_1^1, \cdots p_1^M) \in \Theta \coloneqq [0, 1]^{M+1}$ define a space of policies as $\Pi \coloneqq \{\pi_\theta : \theta \in \Theta\}$ with $\pi_\theta(\uparrow |s_0) = p_0$, $\pi_\theta(\downarrow |s_0) = (1 - p_0)$ and $\pi_\theta(a_1^m|s_1) = p_1^m$. Clearly Π contains the optimal policy, which is obtained for any θ such that $p_0 = 0$. Assuming $N >> 4$, we can evaluate the MERL objective $J_{\text{merl}}(\theta)$ for any temperature α and $\theta \in \Theta$:

$$J_{\text{merl}}(\theta) = (1 - p_0)(1 - \alpha \log(1 - p_0)) - p_0 \left(\alpha \log p_1 + \alpha\gamma \sum_{m=1}^{M} p_1^m \log p_1^m \right). \quad (5.8)$$

We now find the optimal MERL policy. As $p_1^m = \frac{1}{M}$ maximises the final term in Eq. (5.8), we can substituting for $p_1^m = \frac{1}{M}$ and take derivatives of Eq. (5.8) with respect to p_0, which

we set to zero find the MERL optimal policy parameter $p_0^* = \pi^*_{\text{merl}}(\uparrow | s_0)$:

$$1 - \alpha \log(1 - p_0^*) = \alpha \left(\gamma \log M - \log p_0^*\right),$$
$$\implies p_0^* = \frac{1}{M^{-\gamma} \exp\left(\frac{1}{\alpha}\right) + 1},$$

hence, for any $M^{-\gamma} \exp\left(\frac{1}{\alpha}\right) < 1$, we have $p_0^* > \frac{1}{2}$ and so π^* cannot be recovered from π^*_{merl}, even using a deterministic policy constructed from the mode action. The degree to which the modal policy varies from the optimal unregularised policy depends on both the value of α and M, the later controlling the number of states with sub-optimal reward.

Our counterexample illustrates that when there are large regions of the state-space with sub-optimal reward, the temperature must be comparatively small to compensate, hence algorithms derived from the posterior inference become very sensitive to temperature. This issue stems from the static prior over trajectories required to define the posterior inference problem, which introduces a long-term expected regularisation term into the RL objective. In MERL, policies thus optimise for entropy not only in their current state, but also for the expected entropy over all future states. For MDPs where large regions of the state-space are sub-optimal, the regularisation term will dominate the objective and optimal agents will take actions towards regions of the MDP with high expected entropy, which can mask low reward. While initially beneficial for exploration, this can lead to sub-optimal polices being learnt in complex domains as there is often too little a priori knowledge about the MDP to make it possible to choose an appropriate value or schedule for α.

5.2 VIREL

Given the shortcomings of the MML and posterior inference approaches, we are motivated to develop an alternative control as inference framework that combines the benefits of both. Namely, our framework should satisfy three key desiderata:

i the optimal policy under our framework is a deterministic optimal policy for the RL problem;

ii non-optimal policies are stochastic and encourage exploration of the state-action space through regularisation, and;

iii inferring our policy permits the application of powerful and tractable optimisation algorithms from VI that optimise the risk-neutral form of KL divergence.

To satisfy these three criteria, we propose the variational inferenence for reinforcement learning (VIREL) framework.

5.2.1 The VIREL Posterior

Like in previous approaches, it is possible to construct a metaphorical posterior inference problem that corresponds to finding an optimal policy under our framework. Unlike in previous work, in VIREL we infer a posterior over actions, not entire trajectories. To derive the VIREL posterior, we follow the exposition in Section 5.1 by introducing a prior policy over actions: $\pi(\cdot|s)$ and a binary variable $\mathcal{O} \in \{0, 1\}$. The likelihood conditioned on $\mathcal{O} = 1$ has a density:

$$p_\omega(\mathcal{O}|s, a) \propto \exp\left(\frac{\hat{Q}_\omega(s, a) - \max_{a' \in \mathcal{A}} \hat{Q}_\omega(s, a')}{\alpha_\omega}\right), \tag{5.9}$$

where $\hat{Q}_\omega$ is a Q-function approximator and α_ω is a temperature parameter that satisfies the following optimality condition:

Assumption 5 (Temperature Optimality Condition)**.** *Let ρ be a sampling distribution with support over $\mathcal{S}$ and $\Omega^\star$ be the set of parameters that parametrise an optimal Q-function $\rho\pi$-almost surely, that is $\Omega^\star \coloneqq \left\{\omega^\star : \hat{Q}_{\omega^\star} = Q^\star, \quad \rho\pi - a.s.\right\}$. We assume that $\alpha_\omega \geq 0$ is continuous in ω with $\alpha_{\omega^\star} = 0$ iff $\omega^\star \in \Omega^\star$.*

A natural objective that satisfies Assumption 5 for any set of bounded Q-function approximators is the mean-squared optimal Bellman error:

$$\alpha_\omega = \left\|\mathcal{B}^\star\left[\hat{Q}_\omega\right] - \hat{Q}_\omega\right\|^2_{\rho\pi}.$$

Because $\hat{Q}_\omega$ and (by definition) $\mathcal{B}^\star\left[\hat{Q}_\omega\right]$ are bounded, we can prove that $\lim_{\omega \to \omega^\star} \alpha_\omega = 0$ holds for this choice of temperature by applying the dominated convergence theo-

rem:

$$\lim_{\omega \to \omega^\star} \alpha_\omega = \left\| \lim_{\omega \to \omega^\star} \left(\mathcal{B}^\star \left[\hat{Q}_\omega \right] - \hat{Q}_\omega \right) \right\|_{\rho\pi}^2 = 0 \implies \hat{Q}_{\omega^\star} = Q^\star, \quad \rho\pi - a.s.$$

Under Assumption 5, the temperature parameter α_ω becomes a measure of the optimality of the Q-function approximator, which we can exploit to introduces a natural schedule for regularisation. Assumption 5 also clarifies the semantics of our model: the likelihood in Eq. (5.9) represents a measure of how optimal an action a is for an agent in state s according to its current estimate of the optimal Q-function, with $p_\omega(\mathcal{O}|s,a) = 1$ for any optimal action $a \in \arg\max \hat{Q}_\omega(s,a)$ under the Q-function approximator. As our approximator becomes more accurate, $\hat{Q}_\omega \approx Q^\star$ implies that $\alpha_\omega \approx 0$, and hence the likelihood $p_\omega(\mathcal{O}|s,a)$ penalises sub-optimal actions more severely. Of course, throughout this discussion we have assumed that we can represent the Q-function approximator exactly. We now formalise this assumption:

Assumption 6 (Representability of Approximator)**.** *There exists at least one $\omega^\star \in \Omega$ such that $\hat{Q}_{\omega^\star} = Q^\star$ $\rho\pi - a.s.$.*

In practice, we can't expect Assumption 6 to hold for arbitrary families of function approximators. This is of no concern as we can re-define Assumptions 5 and 6 using a projected condition instead, that is replacing the condition $\hat{Q}_{\omega^\star} = Q^\star$ with $\mathcal{P}[Q^\star] = \hat{Q}_{\omega^\star}$ where recall that $\mathcal{P}[Q^\star]$ is the projection operator defined in Eq. (3.11). Any optimal TD fixed point will satisfy Assumption 6 under these conditions and a natural choice of temperature is the mean-squared optimal projected Bellman error:

$$\alpha_\omega = \|\mathcal{P} \circ \hat{B}^\star[\hat{Q}_\omega] - \hat{Q}_\omega\|_{\rho\pi}^2.$$

Under this modification, all theoretical results developed in this section hold, however the VIREL optimal policies will be optimal with respect to the optimal projected Q-function.

We define the VIREL policy to be the posterior over actions $\pi_\omega(\cdot|s) \coloneqq P_\omega(\cdot|s, \mathcal{O})$ which, from Bayes' rule, has the density:

$$\begin{aligned}\pi_\omega(a|s) \coloneqq p_\omega(a|s, \mathcal{O}) &= \frac{p_\omega(\mathcal{O}|s,a)\pi(a|s)}{p(\mathcal{O}|s)}, \\ &\propto \exp\left(\frac{\hat{Q}_\omega(s,a)}{\alpha_\omega} - \log \pi(a|s)\right). \end{aligned} \tag{5.10}$$

In the special case of a uniform prior $\pi(\cdot|s) = \text{Unif}(\cdot)$, the VIREL policy becomes a Boltzmann policy:

$$\pi_\omega(a|s) \propto \exp\left(\frac{\hat{Q}_\omega(s,a)}{\alpha_\omega}\right).$$

with a temperature α_ω.

We now reflect on how the VIREL policy fulfils desiderata i and ii. We can analyse the optimal VIREL policy by taking the limit $\omega \to \omega^\star$, where $\omega^\star$ is any optimal parametrisation satisfying $\hat{B}^\star[\hat{Q}_{\omega^\star}] = \hat{Q}_{\omega^\star}$ $\rho\pi$-almost surely. Under Assumption 5, the temperature of the VIREL policy tends to zero, that is $\lim_{\omega\to\omega^\star} \alpha_\omega = 0$. We therefore expect the VIREL policy to tend towards a Dirac-delta distribution in the same limit:

$$\pi_\omega(\cdot|s) \xrightarrow[\rho-a.s.]{} \delta_{a_{\omega^\star}(s)},$$

where $a_{\omega^\star}(s) \in \arg\max_{a'\in\mathcal{A}'} Q^\star(s,a')$ and $\mathcal{A}' \subseteq \mathcal{A}$ is a set such that $\rho\pi(\mathcal{A}' \times \mathcal{S}) = 1$. As $\delta_{a_{\omega^\star}(s)}$ defines an optimal RL policy, the optimal policy under VIREL is thus optimal under the RL objective $\rho\pi$-almost surely, hence i is satisfied. We formally verify this claim in Theorem 3 using our important convergence theorem from Section 2.4.

Our main proof applies Theorem 1 to the VIREL policy. We must therefore establish the appropriate regularity assumptions in the VIREL framework:

Assumption 7 (VIREL Regularity Assumptions)**.** *Our model satisfies ρ-almost surely:*

i $(\mathcal{A}, d_\mathcal{A})$ *is a compact metric space and the prior $\pi(\cdot|s)$ has support $\mathcal{A}$;*

ii $\hat{Q}_\omega$ *and $Q^\star$ are bounded and continuous in $\mathcal{A}$ for all $\omega \in \Omega$, and;*

iii For every $s \in \mathcal{S}$, $Q^\star$ has a unique maximum $a^\star(s)$, that is $a^\star(s) = \arg\max_{a'\in\mathcal{A}} Q^\star(s,a')$ where $|a^\star(s)| = 1$.

Theorem 3 (Convergence of VIREL Optimal Policy)**.** *Let ω_N be a sequence of parameters that converge to $\omega^\star$, that is $\lim_{N\to\infty}\omega_N = \omega^\star$. Under Assumptions 5-7, the* VIREL *optimal policy converges weak ρ-almost surely to the RL optimal policy, that is:*

$$\pi_{\omega_N}(\cdot|s) \xrightarrow[\rho-a.s.]{} \delta_{a_{\omega^\star}(s)}.$$

Proof. Assumption 7 i and iii establish conditions i and iii of Theorem 1, and the boundedness part of condition ii. Under Assumption 6, $\omega_N \to \omega^\star$ implies $\hat{Q}_{\omega_N} \xrightarrow[\rho\pi-a.s.]{} Q^\star$, hence condition ii is satisfied. In addition $\hat{Q}_{\omega_N} \xrightarrow[\rho\pi-a.s.]{} Q^\star$ implies $\lim_{N\to\infty}\alpha_{\omega_N} = 0$ from Assumption 5, and so all conditions for applying Theorem 1 are fulfilled. Our result hold immediately. □

For any non-optimal VIREL policy, Assumption 6 implies that $\alpha_\omega > 0$. From the definition of the posterior in Eq. (5.10) we see that $\pi_{\omega_N}(\cdot|s)$ is non-deterministic for any $\alpha_\omega > 0$ and hence induces exploration *for all non-optimal* ω. The VIREL policy therefore satisfies desideratum ii as $\pi_{\omega_N}(\cdot|s)$ only collapses to a deterministic policy once the RL problem is solved and is stochastic otherwise.

5.2.2 Comparing MERL with VIREL

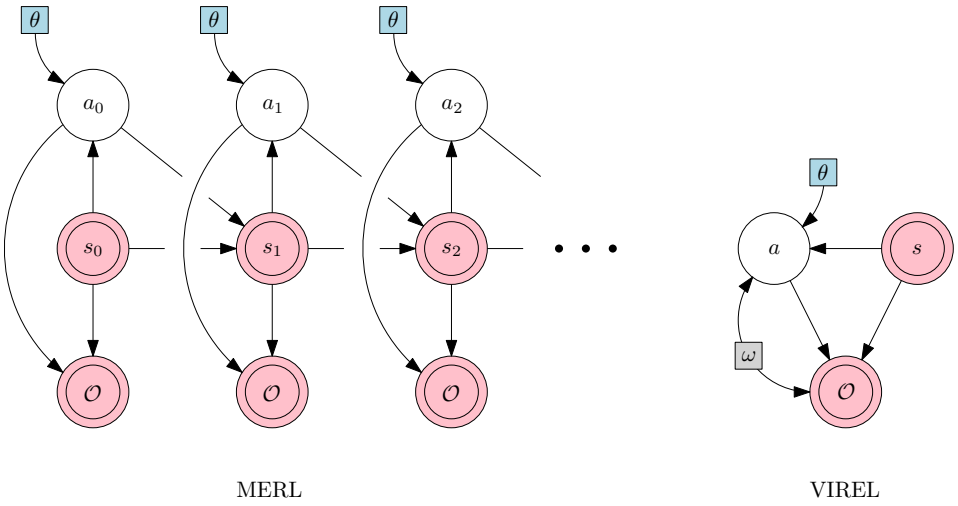


Figure 5.3: Graphical models for MERL and VIREL

To compare MERL and VIREL we consider the underlying probabilistic inference proposed by the two approaches and present their graphical models in Fig. 5.3. Comparing the

graphs in Fig. 5.3, we observe two key differences. Firstly as both a and $\mathcal{O}$ condition on ω, function approximation in VIREL is accounted for by the inference problem whereas it is a separate issue in MERL. Secondly, MERL models exponential *cumulative* rewards over entire trajectories. By contrast, VIREL's variational policy models a single step and a function approximator is used to model future *expected* rewards. The resulting KL divergence minimisation for MERL is therefore much more sensitive to the value of temperature, as this affects how much future entropy influences the variational policy. For VIREL, temperature is defined by the model, and updates to the variational policy will not be as sensitive to errors in its value or linear scaling as its influence only extends to a single interaction. We hypothesise that VIREL may afford advantages in higher dimensional domains where there is greater chance of encountering large regions of state-action space with sub-optimal reward; like our counterexample from Section 5.1.3, the temperature α must be comparatively small to balance the influence of entropy in these regions to prevent MERL algorithms from learning sub-optimal policies.

5.2.3 The VIREL Objective

In the context of an RL algorithm, the VIREL policy serves a role as both an exploration policy, enabling the agent to gather data to solve the optimal Bellman equation, and a solution, asymptotically curtailing exploration according to the agent's ability to represent an optimal Q-function and concentrating on an RL optimal policy. How we solve the optimal Bellman equation given our exploration policy is an open choice and VIREL is compatible with any practical method from the inventory of RL algorithms such as Q-learning or actor-critic.

One problem remains: calculating the normalisation constant to sample directly from the VIREL policy in Eq. (5.10) is intractable for many MDPs and function approximators. Because our framework generates the VIREL policy from a posterior inference problem, powerful and proven approxiate inference methods are applicable to approximate π_ω. In VIREL, we introduce a *variational policy* $\pi_\theta \approx \pi_\omega$, parametrised by $\theta \in \Theta$, and aim to find a parameterisation that minimises the expected KL divergence:

$\mathbb{E}_{\rho(s)}\left[\mathrm{KL}(\pi_\omega(\cdot|s) \parallel \pi_\theta(\cdot|s))\right]$. Under exact representability conditions, any policy $\pi_{\theta^\star}$ such that $\theta^\star \in \arg\min_{\theta\in\Theta} \mathbb{E}_{\rho(s)}\left[\mathrm{KL}(\pi_\omega(\cdot|s) \parallel \pi_\theta(\cdot|s))\right]$ satisfies

$$\mathbb{E}_{\rho(s)}\left[\mathrm{KL}(\pi_\omega(\cdot|s) \parallel \pi_{\theta^\star}(\cdot|s))\right] = 0,$$

which implies that $\pi_{\theta^\star} = \pi_\omega$ ρ-almost surely.

As this objective is still intractable, we can derive a tractable, expected ELBO objective as an equivalent alternative. From Section 4.3.1,

$$\begin{aligned}
&\arg\min_{\theta\in\Theta} \mathbb{E}_{\rho(s)}\left[\mathrm{KL}(\pi_\omega(\cdot|s) \parallel \pi_\theta(\cdot|s))\right] \\
&\qquad = \arg\min_{\theta\in\Theta} \mathbb{E}_{\rho(s)}\left[\int \log p_\omega(\mathcal{O}|s,a) d\pi_\theta(a|s) - \mathrm{KL}(\pi_\theta(\cdot|s) \parallel \pi(\cdot|s))\right], \\
&\qquad = \arg\min_{\theta\in\Theta} \mathbb{E}_{\rho(s)}\left[\frac{1}{\alpha_\omega}\int \hat{Q}_\omega(s,a) d\pi_\theta(a|s) - \mathrm{KL}(\pi_\theta(\cdot|s) \parallel \pi(\cdot|s))\right],
\end{aligned}$$

which is equivalent to minimising the VIREL objective:

$$\mathcal{L}(\theta|\omega) = \mathbb{E}_{\rho\pi_\theta}\left[\hat{Q}_\omega(s,a) - \alpha_\omega \log\left(\frac{\pi_\theta(a|s)}{\pi(a|s)}\right)\right]. \tag{5.11}$$

This yields a tractable way to estimate π_ω at any point during our an RL algorithm by minimising Eq. (5.11) for θ. Our objective satisfies desideratum iii, as we minimise the mode-seeking direction of KL divergence, $\mathrm{KL}(\pi_\omega(\cdot|s) \parallel \pi_\theta(\cdot|s))$, thereby avoiding the risk-seeking tendencies associated with MML methods.

We can interpret the VIREL objective as a variational RL objective: the true Q-function is replaced by an approximation $\hat{Q}_\omega$ and an additional regularisation term is introduced that is scaled by a measure of how far from optimal the approximator is. When the RL problem is solved and $\alpha_\omega = 0$, our approximator is the optimal value function, $\hat{Q}_{\omega^*} = Q^*$ and we recover the true RL objective:

$$\mathcal{L}(\theta|\omega^\star) = \mathbb{E}_{\rho\pi_\theta}\left[Q^\star(s,a)\right].$$

π^* can be inferred exactly by finding $\max_{a'} \hat{Q}_{\omega^*}(a', s)$ or by using a pure policy gradient $\nabla_\theta \mathbb{E}_{\rho\pi_\theta}\left[Q^\star(s,a)\right]$ to minimise $\mathcal{L}(\theta|\omega^\star)$. Whenever $\alpha_\omega > 0$, finding $\theta \in \arg\max_{\theta'\in\Theta} \mathcal{L}(\theta'|\omega)$ corresponds to carrying out a regularised *policy improvement* step.

5.3 VIREL for Control

We now use the VIREL framework to develop a family of algorithms suitable for solving the control problem. Inferring the VIREL policy π_ω by maximising the VIREL objective is at the heart of any VIREL approach as it generates exploration policies. Regardless of the algorithm chosen to learn $\hat{Q}_{\omega^\star}$ we can always infer the $\pi_{\theta^\star}$ for any $\omega \in \Omega$ using a simple stochastic gradient ascent on $\mathcal{L}(\theta|\omega)$:

$$\theta_{k+1} = \theta_k + \beta_k \nabla_\theta \log \pi_{\theta_k}(a|s) \left(\hat{Q}_\omega(s,a) - \alpha_\omega \left(\log \left(\frac{\pi_{\theta_k}(a|s)}{\pi(a|s)} \right) - 1 \right) \right), \tag{5.12}$$

where $s \sim \rho$, $a \sim \pi_{\theta_k}(\cdot|s)$ and β_k is a stepsize. We can interpret the update in Eq. (5.12) as a regularised variational actor update and, keeping with established RL nomenclature, we refer to $\hat{Q}_\omega$ as the *critic* and π_θ as the *actor*.

To learn the critic, we can adapt any RL algorithm using the VIREL policy to explore, for example, we can apply VIREL to the DQN algorithm by updating $\hat{Q}_\omega$ using

$$\omega_{k+1} = \omega_k + \zeta_k \left(r(s',a,s) + \gamma \max_{a' \in \mathcal{A}} \hat{Q}_{\omega'}(s',a') - \hat{Q}_{\omega_k}(s,a) \right) \nabla_\omega \hat{Q}_{\omega_k}(s,a),$$

where $s \sim \rho$, $a \sim \pi(\cdot|s)$ and $s' \sim P(\cdot|s,a)$. When using replay buffers, ρ is a uniform distribution the collected data. Here ω' are lagged parameters updated on a slower timescale to ω_k as outlined in Section 3.5.3. In this way, the VIREL policy replaces the traditional ϵ-greedy exploration used in Q-learning algorithms.

5.3.1 Variational Actor-critic

Most control as inference algorithms are designed to perform on high dimensional continuous control tasks where algorithms such as Q-learning are infeasible due to the $\max$ operator. For this reason, we develop an actor-critic approach that avoids using this operator.

To motivate our actor-critic, observe that a choice of temperature satisfying Assumption 5 is:

$$\alpha_\omega = \left\| \mathcal{B}_\omega \left[\hat{Q}_\omega \right] - \hat{Q}_\omega \right\|_{\rho\pi}^2, \tag{5.13}$$

where $\mathcal{B}_\omega\left[\hat{Q}_\omega\right]$ is the Bellman operator for π_ω. To confirm this choice of temperature, observe that $\omega^\star$ such that $\hat{Q}_{\omega^\star} = Q^{\pi_{\omega^\star}}$ implies that $\alpha_{\omega^\star} = 0$ and from Theorem 3 $\pi_{\omega^\star} = \delta_{a_{\omega^\star}(s)}$, hence an optimal an optimal policy for the RL problem can be learn using the temperature in Eq. (5.13).

One problem remains: evaluating $\mathcal{B}_\omega\left[\hat{Q}_\omega\right]$ requires sampling from π_ω, which may be intractable. Instead, we can use our variational policy to approximate the expectations under π_ω in $\mathcal{B}_\omega\left[\hat{Q}_\omega\right]$ with expectations under π_θ. The subsequent variational temperature is:

$$\alpha_{\omega,\theta} = \left\|\mathcal{B}_\theta\left[\hat{Q}_\omega\right] - \hat{Q}_\omega\right\|^2_{\rho\pi} \approx \alpha_\omega,$$

where:

$$\mathcal{B}_\theta\left[\hat{Q}_\omega\right](s,a) = \int\left(r(s',a,s) + \gamma\int\hat{Q}(s',a')d\pi_\theta(a'|s')\right)dP(s'|s,a).$$

Introducing the variational approximation $\alpha_\omega \approx \alpha_{\omega,\theta}$ into the VIREL objective from Eq. (5.11) yields an approximate VIREL objective:

$$\hat{\mathcal{L}}(\theta|\omega) = \mathbb{E}_{\rho\pi_\theta}\left[\hat{Q}_\omega(s,a) - \alpha_{\omega,\theta}\log\left(\frac{\pi_\theta(a|s)}{\pi(a|s)}\right)\right]. \tag{5.14}$$

We now provide a formal justification for making this variational approximation, proving that it does not affect the optimality of the solution:

Proposition 6. *For any $\theta \in \Theta$ assume there exists some $\omega \in \Omega$ such that $\hat{Q}_\omega = Q^{\pi_\theta}$ $\rho\pi - a.s.$. Then any solution to the nested optimisation problem:*

$$\theta^\star \in \arg\max_{\theta\in\Theta}\hat{\mathcal{L}}(\theta|\omega^\star) \quad \textit{s.t.} \quad \omega^\star(\theta) \in \arg\min_{\omega\in\Omega}\alpha_{\omega,\theta},$$

parametrises an optimal policy for the RL problem.

Proof. Under our assumption of representability, there exists an $\omega^\star \in \Omega$ such that $\hat{Q}_{\omega^\star} = Q^{\pi_\theta} = \mathcal{B}[\hat{Q}_{\omega^\star}]$ $\rho\pi$-almost surely. This implies:

$$\|\mathcal{B}[\hat{Q}_{\omega^\star}] - \hat{Q}_{\omega^\star}\|^2_{\rho,\pi} = 0 = \min_{\omega\in\Omega}\alpha_{\omega,\theta}.$$

Substituting for $\omega = \omega^\star$ in Eq. (5.15) yields:

$$\hat{\mathcal{L}}(\theta|\omega^\star) = \mathbb{E}_{\rho\pi_\theta}\left[\hat{Q}_{\omega^\star} - \alpha_{\omega^\star,\theta}\log\left(\frac{\pi_\theta(a|s)}{\pi(a|s)}\right)\right] = \mathbb{E}_{\rho\pi_\theta}\left[Q^{\pi_\theta}\right],$$

and as ρ has support over $\mathcal{S}$:

$$\arg\min_{\theta\in\Theta} \hat{\mathcal{L}}(\theta|\omega^\star) = \arg\max_{\theta\in\Theta} J(\theta),$$

as required. □

Solving the nested optimisation problem in Proposition 6 can be tackled in exactly the same way as solving the nested policy evaluation and policy improvement optimisation problem to obtain an optimal RL solution. Here the policy evaluation objective is the variation temperature: $\alpha_{\omega,\theta}$ and the policy improvement objective is the variational VIREL objective: $\hat{\mathcal{L}}(\theta|\omega)$. This permits the derivation of a VIREL actor-critic algorithm, interleaving policy evaluation $\omega \in \arg\min_{\omega'\in\Omega} \alpha_{\omega',\theta}$ and regularised VIREL policy improvement steps $\theta \in \arg\max_{\theta'\in\Theta} \hat{\mathcal{L}}(\theta|\omega)$ until a solution is reached. When optimising the VIREL policy improvement and policy evaluation objectives using stochastic gradient descent, the familiar actor-critic parameter updates emerge, with additional regularisation in the actor update:

$$\theta_{k+1} = \theta_k + \beta_k\left(\hat{Q}_\omega(s,a) - \alpha_\omega\left(\log\left(\frac{\pi_{\theta_k}(a|s)}{\pi(a|s)}\right) - 1\right)\right)\nabla_\theta \log \pi_{\theta_k}(a|s), \quad \textbf{(actor)}$$

$$\omega_{k+1} = \omega_k + \zeta_k\left(r(s',a'',s) + \hat{Q}_{\omega'}(s',a') - \hat{Q}_{\omega_k}(s,a'')\right)\nabla_\omega \hat{Q}_{\omega_k}(s,a''), \quad \textbf{(critic)}$$

where $s \sim \rho$, $a \sim \pi_{\theta_k}(\cdot|s)$, $a'' \sim \pi_{\theta_{k+1}}(\cdot|s)$, $s' \sim P(\cdot|s,a)$ and $a' \sim \pi(\cdot|s')$. Again ,we have introduced a set of lagged parameters ω' that are updated on a slower timescale to ω_k. To maintain the separation of variables to solve the nested optimisation problem in Proposition 6, we formally require that $\lim_{k\to\infty} {}^{\zeta_k}/_{\beta_k} = 0$, meaning that our critic is updated asymptotically faster than our actor. We discuss the theoretical reasons and intuition behind introducing separation of timescales to solve nested optimisation problems in Section 7.3.

5.3.2 Empirical Evaluation

The aim of our empirical evaluation is threefold: firstly, as explained in Section 5.2.2, algorithms using soft value functions cannot be recovered from VIREL as they optimise for long term entropy which is incompatible with the VIREL graphical model. We therefore demonstrate that using hard value functions does not affect performance. Secondly, we provide evidence for our hypothesis stated in Section 5.2.2 that using algorithms from the MERL framework may harm performance in higher dimensional tasks due to sensitivity of the temperature parameter. Thirdly, we show that even under all practical approximations, the variational actor-critic derived in Section 5.3.1 still outperforms other established advanced actor-critic methods.

We compare our methods to SAC and DDPG [Lillicrap et al., 2015] algorithms on MuJoCo tasks in OpenAI gym [Brockman et al., 2016]. We use SAC as a baseline because Haarnoja et al. [2018a] show that it outperforms PPO [Schulman et al., 2017], soft Q-learning [Haarnoja et al., 2017] and TD3 [Fujimoto et al., 2018b]. We compare to DDPG [Lillicrap et al., 2015] because, like our methods, it can learn deterministic optimal policies.

To compare to MERL, we assume a uniform prior policy $\pi(\cdot|s) = \text{Unif}(\cdot)$, under which the VIREL objective uses entropy regularisation:

$$\hat{\mathcal{L}}(\theta|\omega) = \mathbb{E}_{\rho\pi_\theta}\left[\hat{Q}_\omega(s,a) - \alpha_{\omega,\theta}\log\pi_\theta(a|s)\right]. \tag{5.15}$$

We assume that ρ is the distribution induced by using a replay buffer of samples $\mathcal{D}$ and we use a Gaussian variational policy $\pi_\theta = \mathcal{N}(\mu_\theta, \Sigma_\theta)$, which enables the use of the reparametrisation trick [Kingma and Welling, 2014] in our objectives.

We consider two variants of the VIREL actor-critic: In the first, called *virel*, we keep the scale of the entropy term in the gradient update for the actor constant: α; in the second, called *beta*, we use an estimate of the temperature $\alpha_{\omega,\theta}$ using a fixed number of samples N_ε drawn from the replay buffer $\mathcal{D}$. We also multiply the estimate by a scalar $\lambda \approx \frac{1-\gamma}{r_{avg}}$, where r_{avg} is the average reward observed. Multiplying the temperature by this fixed scalar does not change the theory of VIREL and allows us to better balance the influence

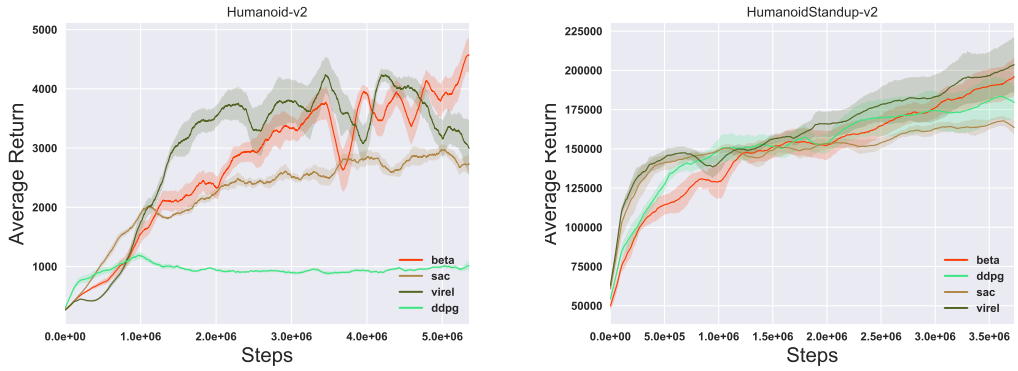
of entropy in proportion to the expected returns as λ^{-1} roughly captures the order of magnitude of the expected returns. We found performance is poor and unstable without λ.

Algorithm 1 Variational Actor-Critic: *virel*

Initialize parameter vectors ϕ, $\bar{\phi}$, θ, ω, $\mathcal{D} \leftarrow \{\}$
for each iteration **do**
  for each environment step **do**
    $a_t \sim \pi^q(a|s;\theta)$
    $s_{t+1} \sim p(s_{t+1}|s_t, a_t)$
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(s_t, a_t, r_t, s_{t+1})\}$
  end for
  for each gradient step **do**
    $\phi \leftarrow \phi - \beta_V \hat{\nabla}_\phi J^V(\phi)$ (target critic step)
    $\omega \leftarrow \omega - \beta_Q \hat{\nabla}_\omega J^Q(\omega)$ (critic step)
    $\theta \leftarrow \theta - \beta_{\pi^q} \hat{\nabla}_\theta J_{\textit{virel}}(\theta)$ (actor step)
    $\bar{\phi} \leftarrow \tau\phi + (1-\tau)\bar{\phi}$
  end for
end for

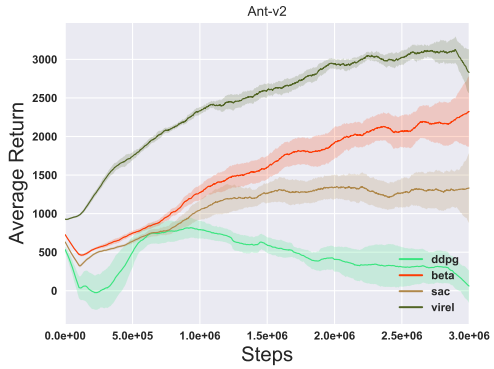
Algorithm 2 Variational Actor-Critic: *beta*

Initialize parameter vectors ϕ, $\bar{\phi}$, θ, ω, $\mathcal{D} \leftarrow \{\}$
for each iteration **do**
  for each environment step **do**
    $a_t \sim \pi^q(a|s;\theta)$
    $s_{t+1} \sim p(s_{t+1}|s_t, a_t)$
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(s_t, a_t, r_t, s_{t+1})\}$
  end for
  for each gradient step **do**
    $\alpha_\omega \leftarrow \mathbb{E}_\mathcal{D}\left[\left(r_t + \gamma \hat{V}_\phi(s_{t+1}) - \hat{Q}_\omega(s_t, a_t)\right)^2\right]$
    $\phi \leftarrow \phi - \beta_V \hat{\nabla}_\phi J^V(\phi)$ (target critic step)
    $\omega \leftarrow \omega - \beta_Q \hat{\nabla}_\omega J^Q(\omega)$ (critic step)
    $\theta \leftarrow \theta - \beta_{\pi^q} \hat{\nabla}_\theta J_{\textit{beta}}(\theta)$ (actor step)
    $\bar{\phi} \leftarrow \tau\phi + (1-\tau)\bar{\phi}$
  end for
end for

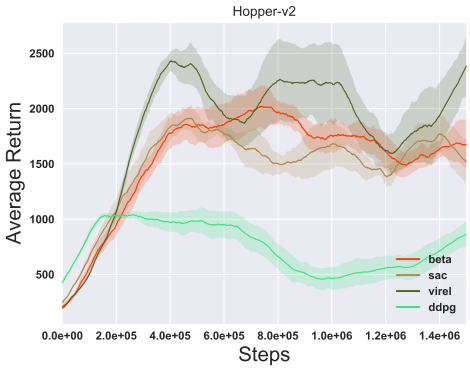


(a) Humanoid

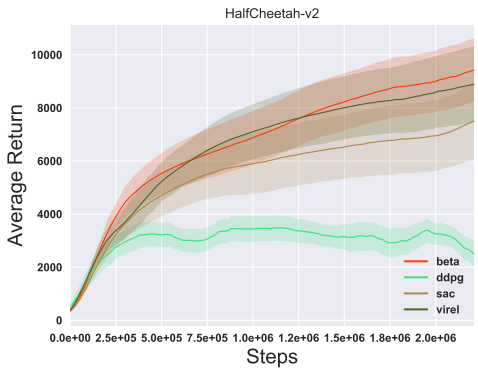
(b) HumanoidStandup



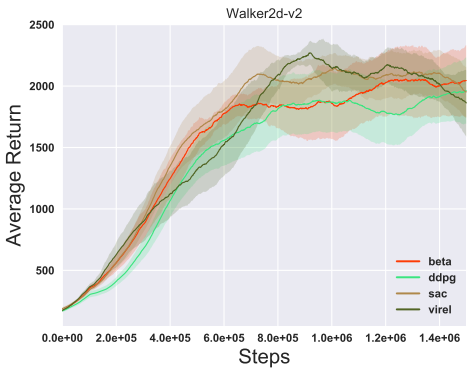
(c) Ant



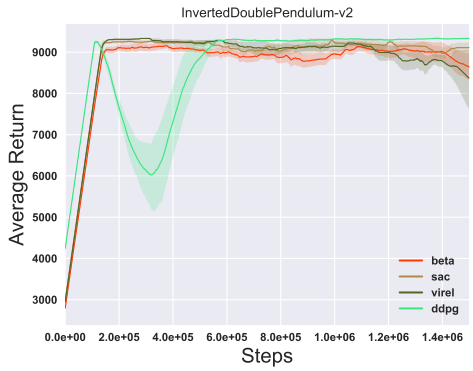
(d) Hopper



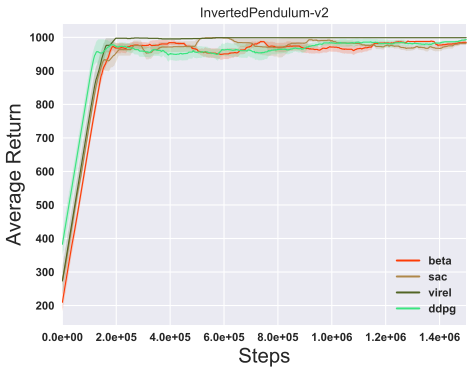
(e) HalfCheetah



(f) Walker



(g) InvertedDoublePendulum



(h) InvertedPendulum

Figure 5.4: MuJoCo continuous control experiments for VIREL.

To reduce variance, all algorithms use a value function network $\hat{V}_\phi$ as a baseline with target parameters $\bar{\phi}$. The respective objectives are:

$$
\begin{aligned}
J^V(\phi) =& \mathbb{E}_{s_t \sim \mathcal{D}} \left[\frac{1}{2} \left(V_\phi(s_t) - \mathbb{E}_{a_t \sim \pi_\theta} \left[\hat{Q}_\omega(s_t, a_t) \right] \right)^2 \right], \\
J^Q(\omega) =& \mathbb{E}_{(s_t, a_t, r_t, s_{t+1}) \sim \mathcal{D}} \left[\frac{1}{2} \left(r_t + \gamma \hat{V}_{\bar{\phi}}(s_{t+1}) - \hat{Q}_\omega(s_t, a_t) \right)^2 \right], \\
J_{virel}(\theta) =& \mathbb{E}_{s_t, a_t \sim \mathcal{D}} \Big[\log \pi_\theta(a_t|s_t) (\alpha - (\hat{Q}_\omega(s_t, a_t) - \hat{V}_{\bar{\phi}}(s_t))) \Big], \\
J_{beta}(\theta) =& \mathbb{E}_{s_t, a_t \sim \mathcal{D}} \left[\log \pi_\theta(a_t|s_t) \left(\frac{1-\gamma}{r_{avg}} \alpha_{\omega,\theta} - (\hat{Q}_\omega(s_t, a_t) - \hat{V}_{\bar{\phi}}(s_t)) \right) \right],
\end{aligned}
$$

and Algorithms 1 and 2 show the pseudocode for our variational actor-critic algorithms. $\hat{\nabla}$ denotes a gradient estimate. All experiments use five random initialisations. Hyperarameter values are given in Appendix D.1.

We plot the training curves for both VIREL algorithms and our baseslines in Fig. 5.4. Our results show that algorithms from the VIREL framework perform on par or better than the equivalent MERL approach, and so there is no disadvantage in using the VIREL framework, only Pareto improvement.

For high dimensional tasks such as Humanoid (action space dimensionality: 17, state space dimensionality: 376) and Ant (action space dimensionality: 8, state space dimensionality: 111), DDPG fails to learn any reasonable policy. We believe that this is because the Ornstein-Uhlenbeck noise that DDPG uses for exploration is insufficiently adaptive in high dimensions. While SAC performs better, *virel* and *beta* still significantly outperform it. As hypothesised in Section 5.2.2, we believe that this performance advantage arises because the gap between optimal unregularised (i.e. VIREL) policies and optimal variational policies learnt under MERL is sensitive to temperature α. This effect is exacerbated in high dimensions where there may be large regions of the state-action space with sub-optimal reward. Thus, as the state-action dimensionality increases, algorithms derived from VIREL outperform SAC and DDPG.

5.4 Fundamental Issues with RL as Inference

On a practical level, control as inference approaches only differ from existing methods by introducing regularisation into the RL objective. These approaches have two key benefits. Firstly, regularisation can be shown to aid the optimisation landscape and can offer improved convergence bounds in comparison to the unregularised objective [Vieillard et al., 2020]. Secondly, regularisation introduces stochasticity into the agent's exploration, preventing policies collapsing prematurely to deterministic Dirac-delta policies.

The first of these benefits mostly concerns the optimisation machinery used to solve the objectives and also applies more generally outside of the context of RL. The second of these benefits is specific to the RL problem and arises because of the exploration/exploitation dilemma. Despite clear improvements in MuJoCo control tasks, this approach is by no means perfect and, as we have shown in Section 5.1, can introduce new issues in its attempt to solve the exploration/exploitation dilemma. We now explore the deeper reasons behind these pathologies.

5.4.1 Pathologies of RL as Inference

To avoid the risk-seeking form of KL divergence minimisation associated with MML, MERL approaches construct a posterior inference problem to solve instead. In the MML problem, the posterior is inferred and the prior parameters are optimised for. In contrast, in the posterior inference problem the prior is fixed. As we discussed in Section 5.1.3, the prior controls the entropy regularisation introduced into the objective. Because the posterior is does not capture uncertainty in the MDP, it does not depend on the data that the agent receives. A consequence of this is that a Bernstein von-Mises type theorem [Vaart, 1998] can never hold, and so the influence of the prior in the objective is persistent and never diminishes with increasing samples from the environment. In practical terms, the lack of a Bernstein von-Mises type theorem means that exploration is not adaptive as the posterior cannot concentrate on a single solution as more data is obtained. In VIREL we corrected this pathology by introducing a heuristic policy whose temperature decreased according to how far the agent was from obtaining an optimal Q-function. An

additional problem with all control as inference approaches is that their policies encourage dithering exploration. This is an innate issue with Boltzmann-type policies for exploration as they cannot characterise uncertainty and so will repeatedly take actions that the agent knows are suboptimal, even with an adaptive temperature scheme.

In Section 7.2.3 we confirm that under a true Bayesian approach, the entropy term in the MERL framework is incorporated into the likelihood, not the prior, which explains why it cannot be corrected for by data. It is concerning that control as inference is solving an inference problem whose likelihood is always inconsistent with our prior beliefs no matter what the MDP. The benefits from frameworks like MERL thus arise from finding a solution that we know is incorrect in the hope its inaccuracy introduces some stochasticity that might be beneficial for exploration.

At this point, we conjecture that the exploration/exploitation dilemma only arises in the first place because of the use of a frequentist objective as a starting point for the RL problem, which assumes a priori knowledge in the environment that the agent does not have. This conjecture is confirmed in Chapter 6. From this perspective, in frameworks like MERL, agents find a solution that is incorrect under the knowledge they do have about the MDP to solve a problem that assumes knowledge that they don't have about the MDP, thus relying on one pathology to solve another! We believe this is a critical issue with control as inference and encourage researches to use regularisation only in the more general sense that it aids optimisation and not to rely on it for exploration.

5.4.2 Should we just be Bayesian?

The pathologies above stem from the fact that control as inference is not a true Bayesian method yet its solution still attempts to carry out Bayesian inference over a different problem that may imply assumptions that are inconsistent with solving an MDP. The VIREL framework resolves some of these issue by constructing a posterior that mimics the behaviour of a true posterior for the MDP, concentrating on an optimal policy as the Q-function approximator approaches optimality, thereby providing an adaptive exploration scheme.

By defining and satisfying a set of desiderata, the VIREL framework is using frequentist methods to solve a problem that arises from taking a frequentist view of the RL problem. If we had taken a Bayesian approach in the first place, we would have never have had to correct for these pathologies. Moreover, the VIREL solution is not perfect. It has no means of quantifying the uncertainty in the MDP and we have no guarantee that the rate at which the temperature adapts is optimal in any sense. By taking a frequentist approach, these imperfections are not revealed to us and only become apparent when we break away from the frequentist paradigm or learn the hard way by watching our algorithms fail in unpredictable ways in new domains.

Instead of pretending to be Bayesian and unintentionally introducing pathologies into our problem by using the control as inference framework, we suggest taking a true Bayesian approach for control. We characterise the benefits of Bayesian RL in Chapter 6 before developing a novel framework in Chapter 7 that characterises the uncertainty in the Bellman operator, thereby opening Bayesian up to practical methods such as bootstrapping.

6

Bayesian Reinforcement Learning

6.1 The Bayesian RL Objective

To motivate the Bayesian RL objective, we revisit the definition of the standard RL objective presented in Eq. (3.2):

$$J(\pi|P,r) \coloneqq \int \sum_{t=0}^{\infty} \gamma^t r_t dP_{\mathcal{T}}^{\pi}(\tau), \tag{6.1}$$

where we have emphasised that this objective conditions upon the transition distribution P and reward function r. From the perspective of decision theory, the RL objective in Eq. (6.1) is an example of a *frequentist utility function* of the decision policy π (which is equivalent to the negative a *frequentist risk* up to a constant) [Bernardo and Smith, 2000]. Unless we are solving a planning problem, the values of P and/or r are unknown a priori, which renders a purely frequentist approach to solving the RL problem somewhat unnatural as the objective in Eq. (6.1) cannot be evaluated. Because the frequentist RL objective does not account for any uncertainty in the MDP, a frequentist agent is always faced with an exploration/exploitation dilemma and must trade off accurately learning P and r (or random variables thereof) via exploration with exploiting its current knowledge of the MDP to act greedily according to its current estimate of $J(\pi|P,r)$. Moreover, the objective yields no guidance or insight in how to approach this balance, As discussed in discussed in Section 3.4, the frequentist RL objective's ubiquity in the

RL literature [Sutton and Barto, 2018, Puterman, 2014, Szepesvári, 2010] has had the unfortunate consequence that exploration presently used in practical RL algorithms is typically heuristic, unprincipled and does not take uncertainty into account.

To resolve these issues, we look to Bayesian decision theory. We first introduce a model of the unknown reward function and transition distribution. In this thesis, we will assume joint parametric models: $r(s', a, s|\phi)$ and $p(\cdot|s, a, \phi)$, whose parameters are elements of the same metric space (Φ, d_Φ). Under a parametric model, the frequentist RL objective can be re-written as:

$$J(\pi|\phi^\star) = \int \sum_{t=0}^{\infty} \gamma^t r_t(\phi^\star) dP^\pi_{\mathcal{T}}(\tau|\phi^\star), \tag{6.2}$$

where $\phi^\star$ parametrises the true value of the unknown reward and transition functions. Evaluating and optimising Eq. (6.2) is still not possible as $\phi^\star$ is unknown. To make progress, some form of distribution P_Φ over the parameter space must be introduced to average over all parameters to eliminate the objective's dependence on the unknown $\phi^\star$:

$$J_{\text{prior}}(\pi) = \int J(\pi|\phi) dP_\Phi(\phi). \tag{6.3}$$

From a Bayesian perspective, the distribution P_Φ is a prior, representing the agent's a priori belief over the space of MDPs: if perfect information about the MDP is available to the agent, the prior becomes a delta distribution $P_\Phi = \delta_{\phi^\star}$ centred on the true parametrisation of the transition distribution and reward function. In the learning setting, the prior should be chosen to reflect the agent's uncertainty in its environment. Eq. (6.3) is often known as either the *pre-posterior* or *Bayes* utility [Wald, 1950]. In this thesis, we will refer to $J_{\text{prior}}(\pi)$ as the prior RL objective and any resulting optimal policy:

$$\pi^\star_{\text{prior}} \in \Pi^\star_{\text{prior}} \coloneqq \arg\max_{\pi \in \Pi} J_{\text{prior}}(\pi),$$

as a prior-optimal policy.

Relying solely on a prior distribution however is clearly insufficient for reinforcement learning as a prior-optimal policy can only be optimal according to the agent's initial beliefs in the MDP. Once samples of data $\mathcal{D}$ from the environment are available, a

posterior can be maintained over the parameters $P_\Phi(\cdot|\mathcal{D})$ which updates the agent's beliefs about the MDP given this new information. It is therefore reasonable to expect that the objective should also be updated to reflect the information gained by this new data. Using the posterior, we can achieve this by marginalising over all reward functions and transition distributions in our model according to the agent's uncertainty in their value, allowing us to derive the Bayesian RL objective:

$$\begin{aligned} J_{\text{Bayes}}(\pi|\mathcal{D}) &\coloneqq \int J(\pi|\phi) dP_\Phi(\phi|\mathcal{D}), \\ &= \int \sum_{t=0}^{\infty} \gamma^t r_t(\mathcal{D}) dP_{\mathcal{T}}^{\pi}(\tau|\mathcal{D}), \end{aligned} \tag{6.4}$$

where $r_t(\mathcal{D}) \coloneqq \int r(s_{t+1}, a_t, s_t) dP_\Phi(\phi|\mathcal{D})$ is the reward predictive mean, $P_{\mathcal{T}}^{\pi}(\cdot|\mathcal{D})$ is the predictive distribution over trajectories, which has the density:

$$p_{\mathcal{T}}^{\pi}(\tau|\mathcal{D}) = p_0(s_0) \prod_{t=0}^{\infty} p(s_{t+1}|s_t, a_t, \mathcal{D}) \pi_(a_t|s_t),$$

and $p(s_{t+1}|s_t, a_t, \mathcal{D}) \coloneqq \int p(s_{t+1}|s_t, a_t, \phi) dP_\Phi(\phi|\mathcal{D})$ is the predictive transition density. We refer to any policy $\pi^\star_{\text{Bayes}}$ that optimises the Bayesian RL objective in Eq. (6.4) as a *Bayes-optimal* policy:

$$\pi^\star_{\text{Bayes}} \in \Pi^\star_{\text{Bayes}} \coloneqq \arg\max_{\pi \in \Pi} J_{\text{Bayes}}(\pi|\mathcal{D}).$$

Finding a Bayes-optimal policy is equivalent to solving a planning problem in the MDP defined by the predictive reward and transition distributions:

$$\text{BAMDP}(\mathcal{D}) \coloneqq \langle \mathcal{S}, \mathcal{A}, r(\cdot, \mathcal{D}), P(\cdot|s, a, \mathcal{D}), P_0, \gamma \rangle. \tag{6.5}$$

The MDP defined in Eq. (6.5) is known as a Bayes adaptive Markov decision process (BAMDP) [Bellman, 1956, Duff, 2002]. Unlike in the frequentist RL objective, solving the Bayesian RL problem is a planning problem and is always possible because the BAMDP only conditions on observed data. This also means that the Bayes-optimal policy is adaptive, changing to reflect the data observed by agent. As the BAMDP is obtained by marginalising over all possible MDPs using the posterior, a Bayes optimal policy will naturally balance exploration with exploitation, taking actions to reduce its uncertainty in the reward and transition distribution, but only insofar as those actions will lead to greater

expected according to the agent's beliefs in their value. Moreover, the exploration is not myopic: by definition any Bayes-optimal policy takes into account the uncertainty not only in immediate reward, but all future rewards and transitions. The key insight is that, rather than being a separate issue as seen from a frequentist perspective, exploration is inherent to the Bayesian RL objective by definition as Bayes-optimal policies are optimal given our current beliefs in the MDP.

Perhaps the most controversial issue in using Bayesian approaches is the introduction of a prior. In Section 6.1.1, we provide a formal justification for introducing a prior distribution in the context of learning problems. Moreover, we show in Section 6.1.1 and Section 7.3.1 that when a frequentist relies on samples from the environment to optimise $J(\pi|P, r)$ via a Monte Carlo approach and/or using samples learning to learn estimates of P and r (or random variables of P and r), both of these approaches implicitly rely on a prior to make up for the fact that the frequentist objective cannot be optimised a priori. Once a prior is accepted, there is little reason to use frequentist decision theory: in Section 6.1.1 we show that any admissible policy is Bayesian; in Section 6.1.2 we confirm that Bayesian policies optimally balance exploration with exploitation; and in Section 6.3 we prove that even under tractable approximation, Bayesian methods still adhere to the conditionality principle, whereas there are no such guarantees for frequentist approaches.

6.1.1 Bayes-optimal Policies are Admissible

We now provide a formal justification for Bayesian RL, proving that even under a frequentist epistemology, it is sufficient to only consider policies that optimise the Bayesian RL objective. To be able to compare policies, we introduce the concept of *admissibility*:

Definition 5 (Admissibility of a Policy)**.** *A policy π is dominated by another policy π' over a set Φ if for all $\phi \in \Phi$,*

$$J(\pi|\phi) \leq J(\pi'|\phi).$$

with strict inequality over $\Phi' \subseteq \Phi$. *An admissible policy is one that is not dominated by any other policy in* Π.

Intuitively, the concept of an admissible policy is related to pareto efficiency in that an admissible policy cannot be bettered in terms of expected return by any other policy across the entire parameter space Φ. Although any method that cannot learn admissible policies clearly should be avoided, an admissible policy may still perform poorly across some regions of the parameters space in comparison to other admissible policies. We now prove a celebrated result from decision theory [Wald, 1947, 1950, Bernardo and Smith, 2000] in the context of the reinforcement learning problem:

Proposition 7 (Admissibility of Bayesian Policies)**.** *Let* P_Φ *be a proper prior, then:*

i every prior-optimal policy $\pi^\star_{prior} \in \Pi^\star_{prior}$ *is admissible* P_Φ*-almost surely;*

ii every Bayes-optimal policy $\pi^\star_{Bayes} \in \Pi^\star_{Bayes}$ *is admissible* $P_\Phi(\cdot|\mathcal{D})$*-almost surely.*

Proof. Assume that proposition i doesn't hold. There thus exists a model with prior P_Φ, prior-optimal policy $\pi^\star_{\text{prior}}$ and policy

$$\pi' \in \Pi \setminus \arg\max_{\pi \in \Pi} J_{\text{prior}}(\pi),$$

such that $J(\pi'|\phi) > J(\pi^\star_{\text{prior}}|\phi)$ over some set Φ' with $P_\Phi(\Phi') > 0$ and $J(\pi'|\phi) = J(\pi^\star_{\text{prior}}|\phi)$ over $\Phi \setminus \Phi'$. As P_ϕ is proper, it then follows from this inequality:

$$\begin{aligned} &\int_{\Phi'} J(\pi'|\phi) dP_\Phi(\phi) > \int_{\Phi'} J(\pi^\star_{\text{prior}}|\phi) dP_\Phi(\phi), \\ \implies &\int_{\Phi} J(\pi'|\phi) dP_\Phi(\phi) > \int_{\Phi} J(\pi^\star_{\text{prior}}|\phi) dP_\Phi(\phi), \end{aligned}$$

which contradicts the definition of a prior-optimal policy.

Proposition ii is proved in exactly the same way for $\pi^\star_{\text{Bayes}}$, using $P_\Phi(\cdot|\mathcal{D})$ in place of the prior.

□

Proposition 7 assures us that any Bayesian approach will generate an admissible policy provided the prior is proper; for improper priors, we may forfeit admissibility. Moreover, Wald [1950] proved under relatively mild regularity assumptions that the converse result

also holds, that is any admissible decision rule is Bayesian with respect to some, possibly improper, prior. This final result ensures that it is sufficient only to consider the Bayesian RL objective in Eq. (6.4) and justifies our introduction of a prior, as any other approach will only yield policies that are dominated by Bayesian policies.

As a brief confirmation of this result, we now demonstrate that minimising the frequentist RL objective via a Monte Carlo gradient descent is actually a Bayesian approach using an improper prior. Firstly we re-write the frequentist objective as:

$$J(\pi) = \int R(\tau) dP_T^{\pi}(\tau)$$

where $R(\tau)$ is the return. If we consider a Dirichlet process prior over the returns R and the tth transition $P(s'|s,a,t)$, in the limit of the scaling parameter tending to 0, we obtain the following posteriors [Ferguson, 1973]:

$$dP(R|\tau,\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N} d\delta_{R_i}(R|\tau), \quad dP(s'|s,a,t,\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N} d\delta_{s_{t+1,i}}(s'|s,a,t),$$

where the dataset $\mathcal{D}$ consists of N trajectories the corresponding returns. Using the posteriors to marginalise over the returns and trajectories leads to the Monte Carlo objective:

$$J(\pi|\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=0}^{\infty}\gamma^t\left(\int r_i(s_{t+1,i},a,s_{t,i})d\pi(a|s_{t,i})\right)\right). \tag{6.6}$$

Finally, if we assume a parametric policy θ, taking gradients of Eq. (6.6) with respect to the policy parameters and using Monte Carlo samples from the policy yields the REINFORCE update [Williams et al., 1993]:

$$\nabla_\theta J(\theta|\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=0}^{\infty}\gamma^t r_i(s_{t+1,i},a_{t,i},s_{t,i})\nabla_\theta \log \pi_\theta(a_{t,i}|s_{t,i})\right).$$

By taking a Bayesian perspective, the shortcomings of using Monte Carlo sampling over returns to learning an optimal policy under uncertainty are made clear: firstly, the prior is improper as it is derived in the limit of the Dirichlet process prior's scaling parameter tending to zero [Terenin and Draper, 2017]. Under Proposition 7 this means that we cannot guarantee that a policy learnt from Monte Carlo methods will be admissible.

Secondly, the posteriors only place mass on observed datapoints, meaning that we cannot reason about uncertainty outside of the returns and transitions we have observed. This makes it impossible to quantify uncertainty over regions of the MDP that agent has yet to visit, preventing a policy from leading the agent in a directed manner towards unexplored regions of the MDP with potentially high returns. Moreover, it is well known that a Dirichlet process prior leads to a rich-gets-richer phenomenon in its predictive distribution [Teh, 2010], the strength of which is controlled by its scaling parameter, with the most extreme case occurring in the limit tending to zero. In the context of RL, a Dirichlet process prior is far from uninformative, as it will *actively direct* the agent towards trajectories that it has already encountered, encouraging pure exploitation.

Finally, in deriving a Monte Carlo gradient descent on the frequentist objective we have not represented the a priori information that the transition distribution and reward function are stationary across time. Instead, we derived a posterior over entire returns and a posterior for each transition distribution after t timesteps. By encoding this prior information through the choice of a model - that is specifying a model with a single reward function and a single transition distribution - we can use data available more efficiently: samples from the environment may be re-used across multiple timesteps if the same (or at least similar) state-action-state tuples are encountered.

6.1.2 Bayes-optimal Policies Explore Optimally

Following our justification for the introduction of a prior in Section 6.1.1, we can also confirm our intuition that a Bayes-optimal policy balances exploration with exploitation: the essential criteria of any such policy is that it explores to reduce uncertainty in the MDP, but only so far as that reduction in uncertainty can yield greater expected returns in future states. We now formalise this notion to demonstrate that is satisfied for Bayes-optimal policies.

Given a dataset of observations $\mathcal{D}$, let $\pi_0(\cdot|\mathcal{D})$ be the policy which is generated by optimising an objective $J(\pi|\mathcal{D})$. Let $\mathcal{D}_n$ be the predictive data that we observe over n future interactions with the environment and $\pi_n \coloneqq \pi(\cdot|\mathcal{D}, \mathcal{D}_n)$ be the corresponding sequence of policies generated from the objectives $J(\pi|\mathcal{D}, \mathcal{D}_n)$. As a prior P_Φ has been

justified, we can obtain a predictive distribution over the future data $P(\cdot|\mathcal{D}, \{\pi_i\}_{i\in 0:n-1})$ that conditions on the observed data $\mathcal{D}$ and sequence of policies $\{\pi_i\}_{i\in 0:n-1}$, with distribution:

$$p(\mathcal{D}_n|\mathcal{D}, \{\pi_i\}_{i\in 0:n-1}) = \prod_{i=0}^{n-1} \int p(s_{i+1}|s_i, a_i, \mathcal{D}, \mathcal{D}_i)p(r|s_{i+1}, a_i, s_i, \mathcal{D}, \mathcal{D}_i)d\pi_i(a_i|s_i).$$

where each $p(s_{i+1}|s_i, a_i, \mathcal{D}, \mathcal{D}_i)$ and $p(r|s_{i+1}, a_i, s_i, \mathcal{D}, \mathcal{D}_i)$ are obtained by marginalising over using the posterior for future observations $P(\cdot|\mathcal{D}, \mathcal{D}_i)$ rather than current posterior $P(\cdot|\mathcal{D})$.

Using the predictive distribution over future data, we can find the predictive objective that will generate the nth policy:

$$J(\pi_n|\mathcal{D}, \{\pi_i\}_{i\in 0:n-1})) \coloneqq \int J(\pi_n|\mathcal{D}, \mathcal{D}_n)dP(\mathcal{D}_n|\mathcal{D}, \{\pi_i\}_{i\in 0:n-1}).$$

Using this predictive objective, we can characterise the notion of optimal exploration: for any $m \geq 0$, any sequence of policies $\{\pi_i\}_{i\in 0:m}$ that maximises $J(\pi_m|\mathcal{D}, \{\pi_i\}_{i\in 0:m-1}))$ will have explored optimally to find a policy with the greatest predictive expected return according to the predicted belief after m timesteps. But we also require that the same sequence of policies should also maximise $J(\pi_k|\mathcal{D}, \{\pi_i\}_{i\in 0:k-1}))$ for any $k > m$. In other words, by following the sequence of policies $\{\pi_i\}_{i\in 0:m}$, will the agent have also explored optimally to find a policy with the greatest returns according to its future belief in the MDP after $k > m$ timesteps? To specify policies that balance exploration with exploitation perfectly, it is necessary that an objective can generate a single sequence of policies that are optimal according to $J(\pi_m|\mathcal{D}, \{\pi_i\}_{i\in 0:m-1}))$ for *every* $m > 0$. Policies that achieve this will gather data not only to obtain an optimal policy according to their current belief in the MDP, but also to obtain an optimal policy according to all predicted future beliefs in the MDP. We describe objectives whose policies explore in this way as *predictive-optimal*:

Definition 6 (Predictive-optimal exploration)**.** *We say that an objective $J(\pi|\mathcal{D})$ generates policies that are predictive-optimal with respect to a prior P_Φ and observed data $\mathcal{D}$ if for*

all $k > m$ and $m \geq 0$, any sequence of m optimal policies:

$$\{\pi_i^\star\}_{i\in 0:m} \in \arg\max_{\{\pi_i\}_{i\in 0:m}} J(\pi_m|\mathcal{D}, \{\pi_i\}_{i\in 0:m-1})),$$

and any sequence of k optimal policies:

$$\{\pi'^\star_i\}_{i\in 0:k} \in \arg\max_{\{\pi_i\}_{i\in 0:k}} J(\pi_k|\mathcal{D}, \{\pi_i\}_{i\in 0:k-1})),$$

should satisfy:

$$J(\pi'^\star_k|\mathcal{D}, \{\pi'^\star_i\}_{i\in 0:k-1})) = J(\pi'^\star_k|\mathcal{D}, \{\pi_i^\star\}_{i\in 0:m}, \{\pi'^\star_i\}_{i\in m+1:k-1})).$$

We remark that our analysis does not rely on reasoning over unobserved future data, but rather *predictive* future data based on our current observations, and so adheres to the likelihood principle introduced in Section 4.2. In this sense, we are asserting that all future predictions will be consistent with our current belief in the MDP.

If predictive-optimality does not hold, under our current beliefs, we predict that the agent will reach a timestep $k > m$ in the future where the data gathered from its choice of historical policies $\{\pi_i\}_{i\in 0:m}$ will be suboptimal for finding a policy that maximises the expected returns at time k. There would exist an alternative sequence of policies $\{\pi_i^\star\}_{i\in 0:m}$ and some $k > m$ such that:

$$\max_{\{\pi_i\}_{i\in m+1:k}} J(\pi_k|\mathcal{D}, \{\pi_i^\star\}_{i\in 0:m}, \{\pi_i\}_{i\in m+1:k-1})) > \max_{\{\pi_i\}_{i\in m+1:k}} J(\pi_k|\mathcal{D}, \{\pi_i\}_{i\in 0:m}, \{\pi_i\}_{i\in m+1:k-1})).$$

An agent following the sequence of alternative policies $\{\pi_i^\star\}_{i\in 0:m}$ would have better explored the MDP and, on average, would have been lead to regions where it could have gathered data to obtain a policy with higher expected returns at time k.

It is easy to show that the Bayesian RL objective is predictive-optimal and therefore, given a model, prior and observations, Bayes-optimal policies optimally balance exploration with exploitation. We start by deriving the predictive Bayesian objective:

$$
\begin{aligned}
J_{\text{Bayes}}(\pi_n|\mathcal{D}, \{\pi_i\}_{i\in 0:n-1}) &= \int J_{\text{Bayes}}(\pi_n|\mathcal{D}, \mathcal{D}_n) dP(\mathcal{D}_n|\mathcal{D}, \{\pi_i\}_{i\in 0:n-1}), \\
&= \int \int J(\pi_n|\phi) dP_\Phi(\phi|\mathcal{D}, \mathcal{D}_n,) dP(\mathcal{D}_n|\mathcal{D}, \{\pi_i\}_{i\in 0:n-1}), \\
&= \int J(\pi_n|\phi) dP_\Phi(\phi|\mathcal{D}),
\end{aligned}
$$

where we have used the result that:

$$
\begin{aligned}
&\int dP_\Phi(\phi|\mathcal{D}, \mathcal{D}_n, \{\pi_i\}_{i\in 0:n-1}) dP(\mathcal{D}_n|\mathcal{D}, \{\pi_i\}_{i\in 0:n-1}) \\
&\quad = dP_\Phi(\phi|\mathcal{D}, \{\pi_i\}_{i\in 0:n-1}), \\
&\quad = dP_\Phi(\phi|\mathcal{D}),
\end{aligned}
$$

in deriving the final line as the posterior given current observations does not depend on future policies. Thus, we obtain the satisfying result that under our current beliefs, the best policy in terms of minimising the predicted expected return after any number of future interactions n with the environment will still be a Bayes-optimal policy:

$$
\arg\min_{\pi_n, \pi_{n-1}, \ldots \pi_0} J_{\text{Bayes}}(\pi_n|\mathcal{D}, \{\pi_i\}_{i\in 0:n-1})) = \arg\min_{\pi_n} J_{\text{Bayes}}(\pi_n|\mathcal{D}) = \Pi^\star_{\text{Bayes}}, \tag{6.7}
$$

and hence Bayes-optimal policies are predictive-optimal, justifying their status as a gold standard for exploration. The result in Eq. (6.7) also provides a sanity check: until the agent actually observes the future data $\mathcal{D}_n$ and updates its posterior accordingly, it cannot gain any more information about what the actual Bayes-optimal policy will be after n timesteps, hence the best prediction is the current Bayes-optimal policy. If this was not the case, a Bayes-optimal policy would be neglecting information that it could use to immediately reduce uncertainty in the MDP to derive a better policy without obtaining any further data.

Our investigation into the properties of Bayes-optimal policies reveals the nature of exploration inherent in Bayesian approaches. Unlike heuristics required by frequentist methods, Bayesian exploration is automatic and optimal with respect to a prior and model. The performance of the agent ultimately rests on the quality of the model and prior: clearly an agent with an accurate model of the transition dynamics and a prior that is

concentrated around the true parametrisation will have an advantage over one with limited, or even incorrect, prior information. However as these factors are typically beyond our control, we are reassured that a Bayesian agent can do no better if it can represent the MDP to the best of its ability given our computational limitations and knowledge we have access to.

6.2 Practical Bayesian Reinforcement Learning

6.2.1 Model-based Bayesian RL

We have already shown in Section 6.1 that finding a Bayes-optimal policy is equivalent to solving a planning problem in a BAMDP. Unfortunately, solving BAMDPs in practice poses several critical challenges that make exact solutions infeasible for all but the very simplest domains: firstly, modelling the conditional transition distribution with sufficient accuracy becomes harder as the dimension d of the underlying probability space increases [Grünewälder et al., 2012]. As an example, Wasserman [2006] show that when using a nonparametric frequentist kernel approach to density estimation, even with an optimal bandwidth, the mean squared error scales as $\mathcal{O}\left(N^{-\frac{-1}{4+d}}\right)$ where N is the number of samples from the true density - to ensure a mean squared error of less than 0.1 when the target density is a multivariate Gaussian of dimension 10, the number of samples required is 842000 in comparison to 19 for a 2 dimensional problem. In a parametric approach, this implies that the number of parameters required to sufficiently represent more realistic multi-modal transition distributions will scale poorly with increasing dimension of the state-space.

Secondly, maintaining a posterior over these parameters also becomes infeasible: even in the discrete setting, maintaining a posterior over the true space of transition functions requires a $|\mathcal{S}| \times |\mathcal{A}| \times |\mathcal{S}|$ element lookup table, rendering exact model-based Bayesian approaches as impractical as tabular RL. In the continuous setting, finding the normalisation constant for the posterior requires high dimensional integrals over the parameter space if accurately capturing the true transition distribution means we cannot use simpler conjugate models.

Finally, even if it is possible to calculate and maintain the exact form of the posterior, carrying out the marginalisation required to derive the BAMDP and solve the underlying planning problem still requires the calculation of high dimensional integrals over the parameter space. For these reasons, existing model-based Bayesian reinforcement learning approaches are restricted to small and discrete state-action spaces [Dearden et al., 1999, Dei, Vlassis et al., 2012, Ghavamzadeh et al., 2015], even when approximations are used [Asmuth and Littman, 2011, Guez et al., 2012, 2013], or for maintaining posteriors over a small discrete set of tasks [Poupart et al., 2006, Brunskill, 2012].

One notable exception is the VariBAD framework [Zintgraf et al., 2020], which exploits a meta-learning setting to carry out approximate Bayesian inference. Here, parametric models of the reward function and transition distribution are introduced: $P_\phi(r|s,a,m)$ and $P_\phi(r|s,a,m)$ which are parametrised by $\phi \in \Phi$ and condition on the random variable m. The latent variable m can be seen as indexing the MDP the agent is in, and in a meta-learning context it is assumed a range of tasks is available for the agent to sample according to $m \sim P_M$. Outside of the meta-learning setting, P_M is simply a Dirac measure centred on the true MDP. To learn the parameters ϕ, the posterior $P_\Phi(\cdot|\mathcal{D})$ is used to obtain a marginal likelihood over trajectories:

$$p(\tau|\phi,\mathcal{D}) = \int p_\phi(\tau|m) dP(m|\mathcal{D}),$$

which is optimised for ϕ. In this way, VariBAD is a partially Bayesian approach, inferring a posterior over m, but not parameters ϕ, allowing us to control the dimensionality of m without forfeiting the expressiveness of our transition and reward models. As sampling from the exact posterior makes evaluating the marginal likelihood intractable, VariBAD also introduces an approximate parametric posterior over m parametrised by $\psi \in \Psi$, $\hat{p}_\psi(m|\mathcal{D})$, allowing us to maximise a tractable ELBO on the marginal likelihood to learn both ψ and ϕ (see Section 4.3.2 for details of this standard procedure). Finally a parametric policy is introduced that conditions on the approximate posterior $\hat{p}_\psi$, which is learnt by optimising any RL objective using the observed data $\mathcal{D}$.

As the VariBAD optimal policy is not obtained by marginalising over a space of MDPs nor is the uncertainty accounted for in all of the model's parameters, it is not Bayes-optimal. Despite forfeiting the theoretical guarantees of Bayes-optimality, the VariBAD still performs well in practice and is able to succeed in continuous domains that are intractable for other model-based Bayesian RL approaches.

6.2.2 Posterior Sampling and Optimism

To avoid the problem of marginalising over the space of MDPs using a posterior, we discuss two alternative exploration schemes that use the posterior in a more computationally efficient way at the expense of Bayes-optimality.

Posterior sampling is the natural extension of the Thompson sampling [Thomson, 1933] algorithm for solving bandit problems, applied to Markov decision problems [Strens, 2000, Osband et al., 2013]: at the start of an interaction with the environment, an agent samples a single MDP from its posterior and then follows the optimal policy according to that sampled MDP for T timesteps. Exploration in posterior sampling is thus driven by sampling policies $\pi_{\text{ps}} \sim P_\pi(\cdot|\mathcal{D})$ from the pushforward of the posterior through a change of variables $\pi^\star = \arg\max_{\pi\in\Pi} J(\pi|\phi)$, which satisfies:

$$\int_{\mathcal{A}} \arg\max_{\pi\in\Pi} J(\pi|\phi) dP_\Phi(\phi|\mathcal{D}) = P_\pi(\arg\max_{\pi\in\Pi} J(\pi|\mathcal{A})|\mathcal{D}),$$

and characterises the uncertainty over the space of optimal policies $\Pi^\star := \{\arg\max_{\pi\in\Pi} J(\pi|\phi) : \phi \in \Phi\}$. As the posterior over Φ concentrates, so too will the posterior over $\Pi^\star$, adaptively curtailing exploration as samples from $P_\pi(\cdot|\mathcal{D})$ will cluster around the true optimal policy of the underlying MDP.

A related approach drives exploration through the principle of *optimism in the face of uncertainty* [Kaelbling, 1994, Wang et al., 2005, Kolter and Ng, 2009, Ciosek et al., 2019a]. By introducing the change of variables $J^\star = \max_{\pi\in\Pi} J(\pi|\phi)$, we can derive the posterior over the optimal expected returns $P_{J^\star}(\cdot|\mathcal{D})$, which satisfies:

$$\int_{\mathcal{A}} \max_{\pi\in\Pi} J(\pi|\phi) dP_\Phi(\phi|\mathcal{D}) = P_{J^\star}(\max_{\pi\in\Pi} J(\pi|\mathcal{A})|\mathcal{D}),$$

An upper credible interval $\text{UCI}_{P_{J^\star}}(\mathcal{D})$ on the optimal expected returns can be constructed. Here a set of MDPs are selected whose expected return under an optimal policy according to our belief the posterior $P_{J^\star}(\cdot|\mathcal{D})$ lie within a top proportion of all MDPs. The MDPs in the upper credible interval are optimistic, as we believe they give the best expected returns under an optimal policy. We define the set of optimistic parametrisations as:

$$\Phi^{\text{optim}}(\mathcal{D}) \coloneqq \{\phi : \max_{\pi} J(\pi|\phi) \in \text{UCI}_{P_{J^\star}}(\mathcal{D})\},$$

and then construct an optimistic policy as:

$$\pi_{\text{optim}} \in \{\arg\max_{\pi \in \Pi} J(\pi|\phi) : \phi \in \Phi^{\text{optim}}(\mathcal{D})\}.$$

By following an optimistic policy π_{optim}, an agent explores by choosing actions that will lead to the greatest possible expected returns under its current beliefs. If that optimism turns out to be misguided, the posterior will be updated and the MDP will not be selected again. If the true expected return matches the optimistic expected return, the agent can exploit its knowledge of the optimistic MDP with the assurance that there is no better policy and so can stop exploring. In practice, a posterior and upper credible interval is typically maintained over an optimal Q-function and actions selected according to an exploration bonus which augments the predictive Q-function with an additional term that adds an optimistic value proportional to the state-action visitation frequencies of the agent.

Both posterior sampling and optimism are tractable approaches that afford the agent many of the benefits associated with Bayesian exploration [Osband et al., 2018, 2019] at the expense of forfeiting the optimality of a true Bayes-optimal policy [Zintgraf et al., 2020]. Both approaches can also achieve near optimal Bayesian regret bounds that scale as the square-root of samples [Osband et al., 2013, 2019]. Our preference in this thesis is for posterior sampling as, for all but the simplest MDPs, the construction of a sufficiently accurate upper credible interval requires sampling several MDPs from the posterior. This phenomenon has been well documented by Osband and Van Roy [2017] who identify a trade-off between statistical and computational efficiency in optimistic exploration due to the construction of sub-optimal credible intervals. In contrast, posterior sampling requires only a single sample at the start of each episode.

6.2.3 Model-free Bayesian RL

To alleviate the issues associated with modelling the transition distribution in a BAMDP, we can derive an analogue of model-free frequentist RL approaches. Given a policy π and an initial state-action pair (s, a), we denote the posterior over the reward-augmented trajectory $\tau_r \coloneqq (r_0, s_1, a_1, r_1, s_2...)$ given a starting state-action pair (s, a) as $P^{\pi}_{\mathcal{T}_r}(\cdot|s, a, \mathcal{D})$ which has density:

$$p^{\pi}_{\mathcal{T}_r}(\tau_r|s, a, \mathcal{D}) = p(s_1|s, a, \mathcal{D}) \prod_{t=1}^{\infty} p(r|s_t, a_{t-1}s_{t-1}, \mathcal{D}) p(s_{t+1}|s_t, a_t, \mathcal{D}) \pi(a_t|s_t).$$

We now make a transformation of variables, defining the variable:

$$Q = R(s, a, \tau_r) \coloneqq r_0(s_1, a, s) + \sum_{t=1}^{\infty} \gamma^t r_t$$

Using Eq. (2.1), we can re-write the Bayesian RL objective by finding the pushforward of $P^{\pi}_{\mathcal{T}_r}(\cdot|s, a, \mathcal{D})$ through the transformed random variable:

$$\begin{aligned} J_{\text{Bayes}}(\pi|\mathcal{D}) &= \int\int\int R(s, a, \tau_r) dP^{\pi}_{\mathcal{T}_r}(\tau_r|s, a, \mathcal{D}) d\pi(a|s) dP_0(s), \\ &= \int\int\int Q dP^{\pi}_{Q}(Q|s, a, \mathcal{D}) d\pi(a|s) dP_0(s), \\ &= \int\int Q^{\pi}(s, a, \mathcal{D}) d\pi(a|s) dP_0(s), \end{aligned} \tag{6.8}$$

where $P^{\pi}_{Q}(\cdot|s, a, \mathcal{D})$ is the pushforward distribution with mean:

$$Q^{\pi}(s, a, \mathcal{D}) \coloneqq \int Q dP^{\pi}_{Q}(Q|s, a, \mathcal{D}) d\pi(a|s).$$

$P^{\pi}_{Q}(\cdot|s, a, \mathcal{D})$ is thus a posterior over Q-functions as it characterises the uncertainty in the random variable Q. Eq. (6.8) reveals that instead of characterising the uncertainty in an MDP via a transition distribution and reward function, we can also solve a BAMDP by characterising the uncertainty over Q-functions. This is a direct consequence of the likelihood principle introduced in Section 4.2 and we discuss it further in Section 6.3.

Existing work in model-free Bayesian RL can be categorised by the choice of the posterior $P^{\pi}_{Q}(Q|s, a, \mathcal{D})$, choice of approximate inference (if any), the exploration method employed by any algorithms and the approach taken to optimise the Bayesian RL objective.

Existing nonparametric approaches use a Gaussian process prior over the Q-function. In GPTD , Engel et al. [2005] show that for any trajectory of length T, a Q-function satisfies

$$r_{T-1} = \Lambda_T Q_T + \eta_T, \tag{6.9}$$

where $r_{T-1} \coloneqq (r_0 \cdots r_{T-1})^\top$ is a vector of an observed sequence of rewards, $Q_T \coloneqq (Q^\pi(s_0, a_0) \cdots Q^\pi(s_T, a_T))^\top$ is a vector of action-values, η_T is a vector of residual noise and the matrix

$$\Lambda_T \coloneqq \begin{bmatrix} 1 & -\gamma & 0 & \cdots & 0 \\ 0 & 1 & -\gamma & \cdots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & \cdots & 1 & -\gamma \end{bmatrix}$$

applies the appropriate discount factors. A Gaussian prior is placed over each Q^π with residuals modelled as a Gaussian process (GP). The posterior over the Q-function can then be inferred for any new point $Q^\pi(s, a)$. Ghavamzadeh et al. [2016] provide a full exposition of Bayesian actor-critic and policy gradient algorithms derived under this framework.

While theoretically appealing, this approach suffers from a number of drawbacks. Firstly, the posterior mean which is used to infer the Q-function from the dataset requires storing and inverting a $T \times T$ matrix. To provide a practical solution to this problem, a sparse implementation is employed whereby samples are only used if the square distance metric over the GP kernel exceeds some value. As well as wasting samples, this introduces an error between the left and right hand sides of Eq. (6.9). Secondly, this formulation relies on Monte Carlo samples from the environment with no bootstrapping which, as we discussed in Section 3.3.3, is notoriously sample inefficient. Thirdly, batch updating is not possible, so the algorithm's performance is at the mercy of a single trajectory collected directly from its exploration policy, which does not exploit the GP's uncertainty estimates.

Some of these issues were later addressed by Chowdhary et al. [2013], whose batch GPQ algorithm can be viewed as a variant of fitted Q-iteration (FQI) [Riedmiller, 2005]. In this

approach, the Q-function is updated after each episode using a batch of samples from the environment to infer the posterior mean of the GP. Instead of using a single trajectory like in GPTD, targets are bootstrapped samples, which improves sample efficiency. However, as GPQ still must invert a large matrix every iteration, sparsity is needed to limit the number of data samples used to form the matrix. For MDPs with large state-action spaces, the number of samples may not be granular enough to accurately estimate the true optimal Q-function, especially if the kernel is poorly chosen. Heuristic exploration is achieved by using Bayesian credible intervals from the posterior to generate optimistic samples to encourage exploration in regions with high uncertainty. The algorithm can also suffer from over-exploration, which is a significant problem in high-dimensional domains.

Previous work using parametric models assume a Gaussian likelihood over the optimal Q-function:

$$P_{Q^\star}(\cdot|s,a,\phi) = \mathcal{N}(\hat{Q}_\omega(s,a), \sigma^2(s,a)).$$

A prior over parameters is introduced P_Φ and a Bayesian Q-learning algorithm is then derived where inferring the predictive optimal Q-function $Q^\star(s,a,\mathcal{D}) := \int Q dP_{Q^\star}(Q|s,a,\mathcal{D})$ using the posterior (or a tractable posterior approximation) replaces the frequentist policy evaluation step and finding the greedy policy according to a posterior sampling replaces the policy improvement step. When linear Gaussian models are used with a conjugate Gaussian prior, the posterior has an analytic form and Bayesian policy evaluation can be carried out exactly [Azizzadenesheli et al., 2018, Janz et al., 2019].

For nonlinear models that may better represent the true optimal Q-function, exact inference is intractable. Instead, some degree of approximate inference is required: Gal and Ghahramani [2016] use dropout as VI; Tang and Kucukelbir [2017] learn a variational posterior by minimising an ELBO; [Lipton et al., 2018] use independent factorized Gaussians as an approximate posterior, learnt via VI; and [Touati et al., 2019] carry out VI on the posterior via normalising flows [Rezende and Mohamed, 2015]. Unfortunately, all of these methods naïvely apply approximate inference to the Bellman error, treating $\mathcal{B}^\star[Q^\star](s,a)$ and $Q^\star(s,a)$ as independent variables. This leads to poor

uncertainty estimates as the Bellman error cannot correctly propagate the uncertainty [O'Donoghue et al., 2018, Osband et al., 2018]. Consequently, Osband et al. [2019] demonstrate that this can cause uncertainty estimates of $Q^\star(s,a)$ to be zero at some (s,a) and propose BootDQN+Prior as an alternative. We provide a full exposition of BootDQN+Prior in Section 7.4.2 where we derive it in a principled way from our Bayesian Bellman operators (BBO) framework.

For the purposes of this thesis, we highlight two key observations in relation to existing model-free Bayesian RL: firstly, like their frequentist counterparts, all model free Bayesian RL algorithms must utilise some degree of bootstrapping to be practical. Secondly, all existing approaches assume either a Gaussian process or parametric Gaussian regression model for the likelihood.

6.3 Bayesian RL and the Likelihood Principle

We now take a high-level view of the Bayesian RL problem to understand how these methods adhere to the likelihood principle introduced in Section 4.2.

Under parametric models considered in this thesis, we can interpret each ϕ as indexing a particular MDP in a space of MDPs Φ. Inferring a posterior $P_\Phi(\cdot|\ \mathcal{D})$ thus characterise the uncertainty in the MDP. When we take a model-based approach to Bayesian reinforcement learning, we introduce random variables of ϕ in the form of the reward function $r = r(\cdot,\phi)$ and the transition distribution $\mathcal{P} = P(\cdot|s,a,\phi)$, which defines a model of the environment, and infer the posteriors over these variables: $P_r(\cdot|s',a,s,\phi)$ and $P_\mathcal{P}(\cdot|s,a,\mathcal{D})$, which for any integrable f satisfy:

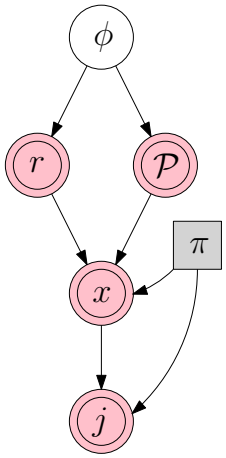


Figure 6.1: Graphical model of Bayesian RL.

$$\int f \circ r(s',a,s,\phi) dP_\phi(\phi|\mathcal{D}) = \int f(r) dP_r(r|s',a,s,\phi),$$
$$\int f \circ P(\cdot|s,a,\phi) dP_\Phi(\phi|\mathcal{D}) = \int f(\mathcal{P}) dP_\mathcal{P}(\mathcal{P}|s,a,\mathcal{D}).$$

In model-free approaches, a random variable $x = X(r, \mathcal{P}, \pi)$ is introduced which may take the form of a Q-function or, in the case of the BBO model introduced in Chapter 7, a Bellman operator. The graphical model of this chain of dependence is shown in Fig. 6.1. As r and $\mathcal{P}$ are random variables of ϕ, it suffices to model the mapping $X(\cdot, \pi) : \Phi \to \mathcal{X}$. Given the likelihood and a prior, the posterior $P_X(\cdot|\mathcal{D}, \pi)$ can be inferred.

If the RL objective can be written in terms of the random variable $X(\cdot|\pi)$, that is $J(\pi|X(\phi, \pi)) = J(\pi|\phi)$, then the Bayesian RL objective can be found by marginalising over x using the posterior $P_X(\cdot|\mathcal{D}, \pi)$:

$$J_{\text{Bayes}}(\pi|D) = \int J(\pi|x) dP_X(x|\mathcal{D}, \pi) = \int J(\pi|r, \mathcal{P}) dP_{r,\mathcal{P}}(r, \mathcal{P}|\mathcal{D}) = \int J(\pi|\phi) dP_\Phi(\phi|\mathcal{D}) \tag{6.10}$$

The chain of equalities in Eq. (6.10) is a manifestation of the likelihood principle applied to Bayesian RL: if X is a sufficient statistic for J given π then inferring a posterior over X given π provides the same information as inferring a posterior over ϕ (the sufficiency principle) and all inferences only condition on the observed data $\mathcal{D}$ and the known policy π (the conditionality principle).

In the context of RL, X is a sufficient statistic for J if given X and π we can evaluate $J(\pi|\phi)$ without any further information about the MDP. An example of a random variable that is a sufficient statistic for J is the Q-function, as given Q^π, we have :

$$J(\pi|Q^\pi) = \int \int Q^\pi(s, a) d\pi(a|s) dp_0(s),$$

and we need no further information from the environment to evaluate $J(\pi|Q^\pi)$ exactly.

The key insight from the likelihood principle is that provided a random variable X is sufficient for obtaining an optimal policy, we can find a Bayes-optimal policy by inferring a posterior over X and the corresponding Bayesian RL objective can be optimised exactly as it only conditions on observed data.

Moreover if our model of $X(\cdot|\pi)$ is consistent with a model of the environment dynamics $r(\cdot|\phi)$ and $\mathcal{P}(\cdot\phi)$, that is:

$$X(\phi|\pi) = X(r(\cdot|\phi), \mathcal{P}(\cdot|\phi)|\pi),$$

then inferring a posterior over the reward and environment dynamics provides the same information as inferring a posterior over the variable X. In this case, the Bayes-optimal policy in model-free and model-based Bayesian RL will be the same.

We conclude our discussion by reassessing the frequentist distinction that is made between model-free and model-based RL in light of the likelihood principle. Observe from our chain of equalities in Eq. (6.10) that in specifying a model to infer the posterior over Q, we make an implicit assumption about the underlying environment dynamics and thus the space of MDPs that we can represent. For these reasons, the name model-free is a misnomer as at the heart of any model of a Q-function is a model of the underlying MDP. Because a Q-function is sufficient for solving the RL problem, a posterior over its value provides exactly the same information as the corresponding model of the environment dynamics. We discuss this idea further in Section 7.5.

6.3.1 Bayesian vs Frequentist RL

A frequentist approach can always be recovered from a Bayesian approach by using the posterior $P_\Phi(\cdot|\mathcal{D}) = \delta_{\phi^\star}$ where $\phi^\star$ indexes the true MDP. As analysis in Eq. (6.10) holds for any posterior, we recover the frequentist RL objective:

$$J(\pi|\phi^\star) = J(\pi|x^\star) = J(\pi|r^\star, \mathcal{P}^\star). \tag{6.11}$$

where $r^\star, \mathcal{P}^\star$ are the true environment dynamics and $x^\star$ is a sufficient variable of the true MDP. If knowledge of $\phi^\star$ exists, the agent is in a planning problem and Bayesian and frequentist approaches agree on the equations in Eq. (6.11). Under these conditions, both the frequentist and Bayesian approaches satisfy the likelihood principle.

In the more general case, the agent does not have complete knowledge of $\phi^\star$ and is in a learning problem. In this situation, the frequentist approach does not satisfy the likelihood principle: by definition, the conditionality principle is not satisfied as all objectives in

Eq. (6.11) cannot be evaluated without $\phi^\star$. Pathologies can thus arise whereby agents reach different conclusions from the same data as they try to reason over information they don't have. Because of this, frequentist approaches are inappropriate for learning problems whereas Bayesian approaches are appropriate for both learning and planning problems.

Moreover, if we introduce a tractable posterior point estimate for the Bayesian approach, the conditionality principle is still satisfied. This follows immediately by substituting for estimate $P(\cdot|\mathcal{D}) \approx \delta_{\phi^\star(\mathcal{D}}$ in Eq. (6.10):

$$J(\pi|\mathcal{D}) \approx J(\pi|\phi^\star(\mathcal{D})) = J(\pi|x^\star(\mathcal{D})) = J(\pi|r^\star(\mathcal{D}), \mathcal{P}^\star(\mathcal{D})), \tag{6.12}$$

where $r^\star(\mathcal{D}) = r(\cdot|\phi^\star(\mathcal{D}))$, $\mathcal{P}^\star(\mathcal{D}) = \mathcal{P}(\cdot|\phi^\star(\mathcal{D}))$, $X^\star(\mathcal{D}) = X(\cdot|\phi^\star(\mathcal{D}))$ and $\phi^\star(\mathcal{D})$ is any point estimate derived from $P_\Phi(\cdot|\mathcal{D})$, for example the MAP estimate.

Like any approximation, by introducing a point estimate we may loose the sufficiency condition as our approximations of the sufficient random variables may not have all of the information needed to solve the underlying Bayesian RL problem, however Eq. (6.12) proves that we still retain the conditionality principle as all approximations only condition on $\mathcal{D}$.

In contrast, if we try deriving a point estimate under a frequentist approach given some data $\mathcal{D}$, say $\omega(\phi^\star, \mathcal{D}) \in \arg\min_{\omega\in\Omega} f(\omega; \phi^\star, \mathcal{D})$, our objective and hence point estimate can still depend on $\phi^\star$ and so we have no guarantee that the conditionality principle holds. A concrete example of where this happens in practice is the MSE and MSBE introduced in Section 3.5.2, which are used to learn a point estimate of a Q-function. Minimising either of these objectives is impossible for most MDPs as they still rely on knowledge of $\phi^\star$ that is unavailable. TD methods attempt to circumvent this issue by dropping gradient terms that cannot be evaluated, however convergence guarantees are lost.

In conclusion, Bayesian approaches have an advantage over frequentist approaches as even under an approximations that renders them as tractable as each other, Bayesian approaches still retain the conditionality principle.

7

Bayesian Bellman Operators

This chapter expands the presentation of Bayesian Bellman Operators in [Fellows et al., 2021]. Empirical work was carried out in collaboration with Kristian Hartikainen.

To motivate the BBO framework, we revisit the model-free Bayesian RL objective from Eq. (6.8):

$$J_{\text{Bayes}}(\pi|\mathcal{D}) = \int\int\int Q dP_Q^\pi(Q|s,a,\mathcal{D}) d\pi(a|s) dP_0(s).$$

The posterior $P_Q^\pi(\cdot|s,a,\mathcal{D})$ was derived by pushing forward the posterior distribution over trajectories through the returns given an initial state-action pair. As each sample $Q_i(s,a) = r(s_i',a,s) + \sum_{t=1}^{\infty}\gamma^t r_{i,t}$ is a noisy estimate of $Q^\pi(s,a)$, the posterior $P_Q^\pi(\cdot|s,a,\mathcal{D})$ is thus the posterior over the Q-function at (s,a). However, we observed in Section 6.2.3 that practical model-free Bayesian RL algorithms cannot rely on sampled returns and must introduce some degree of bootstrapping. For example, if a function approximator $\hat{Q}_\omega \approx Q^\pi$ is introduced, the samples can be approximated using the one-step TD estimate: $Q_i(s,a) \approx r(s_i',a,s) + \hat{Q}_\omega(s_i',a_i')$, introducing an error that is dependent on ω. Existing approaches do not account for this error's dependence on the function approximator. With bootstrapping, samples are no longer noisy estimates of a point $Q^\pi(s,a)$ and the resulting posterior is not $P_Q^\pi(\cdot|s,a,\mathcal{D})$ as it has dependence on $\hat{Q}_\omega$ due to the dataset. This issue is made worse when model-free Bayesian RL is applied to

learning an optimal Q-function (for example in a Bayesian Q-learning setting) as it is never possible to obtain samples from the optimal Q-function without an optimal policy, even when bootstrapping is avoided. These major theoretical issues raise the following questions:

1. *Do model-free Bayesian RL approaches that use bootstrapping still infer a posterior?*
2. *If it exists, how does this posterior relate to solving the RL problem?*
3. *What effect does approximate inference have on the solution?*
4. *Do methods that sample from an approximate posterior converge?*

We address these questions by introducing the BBO framework. In answer to Question 1, we show that by introducing bootstrapping, we actually infer a posterior over Bellman operators. We can use this posterior to marginalise over all Bellman operators to obtain a Bayesian Bellman operator. Our theoretical results provide answers to Questions 2-4, proving that the Bayesian Bellman operator can parametrise a TD fixed point as the number of samples $N \to \infty$ and is analogous to the projection operator used in convergent reinforcement learning. Our results hold even under posterior approximation.

7.1 Uncertainty in the Bellman Operator

To introduce the BBO framework, we consider the problem of finding a Q-function in a BAMDP. Mirroring our analysis of a frequentist MDP in Section 3.2, we can define the Bayesian Bellman operator as:

$$\mathcal{B}[Q^\pi](s,a,\mathcal{D}) = \int\int\int (r + \gamma Q^\pi(s',a',\mathcal{D}))dP^\pi(r,s',a'|s,a,\mathcal{D}),$$

where $dP^\pi(r,s',a'|s,a,\mathcal{D}) = d\pi(a'|s')dP(r|s's,a,\mathcal{D})dP(s'|s,a,\mathcal{D})$. Any Bayesian Q-function satisfies a Bayesian Bellman equation $\mathcal{B}[Q^\pi](\cdot,\mathcal{D}) = Q^\pi(\cdot,\mathcal{D})$. This insight affords a tractable approach for learning a Bayesian Q-function $Q^\pi(\cdot,\mathcal{D})$ by introducing a function approximator $\hat{Q}_\omega \approx Q^\pi(\cdot,\mathcal{D})$ and minimising the Bayesian analogue of the MSBE, which we define as the mean squared Bayesian Bellman error (MSBBE):

$$\text{MSBBE}(\omega|\mathcal{D}) \coloneqq \|\mathcal{B}[\hat{Q}_\omega](\cdot,\mathcal{D}) - \hat{Q}_\omega\|^2_\varsigma.$$

For any $\omega^\star$ such that $\mathcal{B}[\hat{Q}_{\omega^\star}](\cdot, \mathcal{D}) = \hat{Q}_{\omega^\star}$ we have MSBBE$(\omega|\mathcal{D}) = 0$ and $\hat{Q}_{\omega^\star} = Q^\pi(\cdot, \mathcal{D})$ ς-almost surely. Whilst evaluating the Bayesian Bellman operator for $\mathcal{B}[\hat{Q}_\omega](\cdot, \mathcal{D})$ is possible due to the conditionality principle, it requires marginalisation using the predictive reward and transition distribution. To avoid this we can introduce a change of variables using the empirical Bellman function:

$$b = b^\pi_\omega(r, s') \coloneqq r + \gamma \int \hat{Q}_\omega(s', a') d\pi(a'|s'), \tag{7.1}$$

and instead sample from the pushforward posterior $P_B(\cdot|s, a, \mathcal{D}_\omega)$, which satisfies:

$$\begin{aligned}
\mathcal{B}[\hat{Q}_\omega](s, a, \mathcal{D}) &= \int\int (r + \gamma \int \hat{Q}_\omega(s', a') d\pi(a'|s')) dP(r, s'|s, a, \mathcal{D}), \\
&= \int\int b^\pi_\omega(r, s') dP(r, s'|s, a, \mathcal{D}), \\
&= \int b dP_B(b|s, a, \mathcal{D}_\omega).
\end{aligned}$$

The pushforward posterior $P_B(\cdot|s, a, \mathcal{D}_\omega)$ is thus a posterior over Bellman operators, characterising the uncertainty in the true Bellman operator $\mathcal{B}[\hat{Q}_\omega]$ using the function approximator $\hat{Q}_\omega$. We can also derive a posterior over optimal Bellman equations by using the optimial empirical Bellman function for the change of variables:

$$b^\star = b^\star_\omega(r, s') \coloneqq r + \gamma \max_{a' \in \mathcal{A}} \hat{Q}_\omega(s', a'),$$

The resulting pushforward distribution $P_{B^\star}(\cdot|\mathcal{D}_\omega)$ characterises the uncertainty in the optimal Bellman operator and satisfies:

$$\begin{aligned}
\mathcal{B}^\star[\hat{Q}_\omega](s, a, \mathcal{D}) &\coloneqq \int\int (r + \gamma \max_{a' \in \mathcal{A}} \hat{Q}_\omega(s', a')) dP(r, s'|s, a, \mathcal{D}), \\
&= \int b^\star dP_{B^\star}(b^\star|s, a, \mathcal{D}_\omega),
\end{aligned} \tag{7.2}$$

where $dP(r, s'|s, a, \mathcal{D}) = dP(r|s's, a, \mathcal{D}) dP(s'|s, a, \mathcal{D})$.

The guiding principle behind modelling uncertainty in the Bellman operator is that Bayesian methods satisfy the likelihood principle and the Bellman operator is a sufficient statistic for evaluating the RL objective. As discussed in Section 4.2, this assures us that, given the same underlying model of the environment, observations from the random variables $b^\star$ or b in BBO give us the same information as observations from r and s'

when evaluating the uncertainty in the reinforcement learning objective. Modelling the uncertainty in these random variables is much simpler than the transition distribution as they are scalar-valued. Unlike modelling a Q-function, we can obtain samples of $b^\star$ or b by applying the relevant empirical bellman function just as we do in bootstrapping.

7.2 BBO Model

In BBO our aim is to infer the posterior over Bellman operators: $P_B(\cdot|s, a, \mathcal{D}_\omega)$. The graphical model for BBO is shown in Fig. 7.1. Observe that the empirical Bellman function and hence our dataset and posterior also has implicit dependence on π. We do not explicit write this dependence to help reduce notational clutter. For ease of exposition we assume a parametric model, with a likelihood: $p(b|s, a, \phi)$, and we write the likelihood's conditional mean as:

$$\hat{B}_\phi(s, a) \coloneqq \mathbb{E}_{P(b|s,a,\phi)}[b]$$

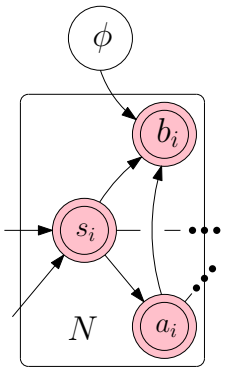


Figure 7.1: Graphical Model for BBO.

The conditional mean of the distribution $\hat{B}_\phi$ can be interpreted as a function space of approximators that represents a space of Bellman operators, each indexed by $\phi \in \Phi$. The choice of $p(b|s, a, \phi)$ should therefore ensure that the space of approximate Bellman operators characterised by $\hat{B}_\phi$ is expressive enough to sufficiently represent the true Bellman operator. We introduce a prior P_Φ to characterise our initial beliefs in the true Bellman operator.

To obtain samples of the random variable b, we follow the same bootstrapping procedure used by model-free BayesianRL approached: as an agent in state s_i interacts with the environment, it takes an action $a_i \sim \pi(\cdot|s_i)$, transitions to the next state according to the transition distribution $s'_i \sim P(\cdot|s_i, a_i)$ and observes a reward $r_i = r(s'_i, a_i, s_i)$. From Eq. (7.1) a sample of the random variable b is obtained by applying the empirical Bellman function $b_i = r_i + \gamma \int \hat{Q}_\omega(s'_i, a'_i) d\pi(a'|s'_i)$. When obtaining the sample b_i it suffices to approximate the inner expectation with a sample from the policy as we only require unbiased estimates of the true Bellman operator in BBO. After N interactions with

the environment, we collect the interactions to form a dataset of N samples: $\mathcal{D}_\omega^N \coloneqq \{(b_i, a_i, s_i)\}_{i \in 1:N}$.

We now make the following assumption regarding how states that the agent encounters are generated:

Assumption 8 (State Generating Distribution)**.** *Each state s_i is drawn either i) i.i.d. from a distribution $\rho(s)$ with support over $\mathcal{S}$ or ii) from an ergodic Markov chain with stationary distribution $\rho(s)$ defined over a σ-algebra that is countably generated from $\mathcal{S}$.*

Observe that the ergodic Markov chain in Assumption 8 does not have to be that followed by $\pi(a|s)$, hence BBO can accommodate off-policy sampling provided the underlying Markov chain is ergodic with a stationary distribution: $\rho(s)$. Using Bayes' rule, we can infer the posterior over parameters $P_\Phi(\cdot|\mathcal{D}_\omega^N)$, which has the same density regardless of our choice of state generating distribution:

Theorem 4 (BBO Posterior)**.** *Under a state generating distribution defined in Assumption 8, the BBO posterior has a density:*

$$p_\Phi(\phi|\mathcal{D}_\omega^N) = \frac{\prod_{i=1}^N p(b_i|s_i, a_i, \phi) p_\Phi(\phi)}{\int \prod_{i=1}^N p(b_i|s_i, a_i, \phi) dP_\Phi(\phi)}.$$

Proof. We start by deriving the likelihood of the data $\mathcal{D}_\omega^N$ for the i.i.d. case:

$$p(\mathcal{D}_\omega^N|\phi) = \prod_{i=1}^N \rho(s_i)\pi(a_i|s_i)p(b_i|s_i, a_i, \phi).$$

Using our prior $p_\Phi(\phi)$ and Bayes' rule, we can infer the posterior as:

$$\begin{aligned}
p_\Phi(\phi|\mathcal{D}_\omega^N) &= \frac{p(\mathcal{D}_\omega^N|\phi)p_\Phi(\phi)}{\int p(\mathcal{D}_\omega^N|\phi) dP_\Phi(\phi)}, \\
&= \frac{\prod_{i=1}^N \left(\rho(s_i)\pi(a_i|s_i)p(b_i|s_i, a_i, \phi)\right) p_\Phi(\phi)}{\int \prod_{i=1}^N \left(\rho(s_i)\pi(a_i|s_i)p(b_i|s_i, a_i, \phi)\right) dP_\Phi(\phi)}, \\
&= \frac{\prod_{j=1}^N (\rho(s_j)\pi(a_j|s_j)) \prod_{i=1}^N p(b_i|s_i, a_i, \phi) p_\Phi(\phi)}{\int \prod_{j=1}^N (\rho(s_j)\pi(a_j|s_j)) \prod_{i=1}^N p(b_i|s_i, a_i, \phi) dP_\Phi(\phi)}, \\
&= \frac{\prod_{j=1}^N (\rho(s_j)\pi(a_j|s_j)) \prod_{i=1}^N p(b_i|s_i, a_i, \phi) p_\Phi(\phi)}{\prod_{j=1}^N (\rho(s_j)\pi(a_j|s_j)) \int \prod_{i=1}^N p(b_i|s_i, a_i, \phi) dP_\Phi(\phi)}, \\
&= \frac{\prod_{i=1}^N p(b_i|s_i, a_i, \phi) p_\Phi(\phi)}{\int \prod_{i=1}^N p(b_i|s_i, a_i, \phi) dP_\Phi(\phi)}.
\end{aligned} \tag{7.3}$$

For sampling from an ergodic Markov chain, we denote the initial state-action density as $p_0(s,a)$ and the transition density as $p(s',a'|s,a)$. For sampling on-policy these distributions are defined as:

$$p_0(s,a) \coloneqq p_0(s)\pi(a|s)$$
$$p(s',a'|s,a) \coloneqq p(s'|s,a)\pi(a'|s').$$

We write our likelihood as:

$$\begin{aligned} p(\mathcal{D}_\omega^N|\phi) &= p_0(s_1,a_1)p(b_1|s_1,a_1,\phi)\prod_{i=2}^{N}(p(s_i\,a_i|s_{i-1},a_{i-1})p(b_i|s_i,a_i,\phi)), \\ &= p_0(s_1,a_1)\prod_{j=2}^{N}p(s_j\,a_j|s_{j-1},a_{j-1})\prod_{i=1}^{N}p(b_i|s_i,a_i,\phi), \\ &= p(S_N,A_N)\prod_{i=1}^{N}p(b_i|s_i,a_i,\phi), \end{aligned}$$

where $S_N \coloneqq \{s_1,...s_N\}$, $A_N \coloneqq \{a_1,...a_N\}$ and

$$p(S_N,A_N) = p_0(s_1,a_1)\prod_{j=2}^{N}p(s_j\,a_j|s_{j-1},a_{j-1}).$$

We now infer the posterior using Bayes' rule:

$$\begin{aligned} p_\Phi(\phi|\mathcal{D}_\omega^N) &= \frac{p(\mathcal{D}_\omega^N|\phi)p_\Phi(\phi)}{\int p(\mathcal{D}_\omega^N|\phi)dP_\Phi(\phi)}, \\ &= \frac{p(S_N,A_N)p(b_i|s_i,a_i,\phi)p_\Phi(\phi)}{p(S_N,A_N)\int\prod_{i=1}^{N}p(b_i|s_i,a_i,\phi)dP_\Phi(\phi)}, \\ &= \frac{\prod_{i=1}^{N}p(b_i|s_i,a_i,\phi)p_\Phi(\phi)}{\int\prod_{i=1}^{N}p(b_i|s_i,a_i,\phi)dP_\Phi(\phi)}, \end{aligned}$$

which has the same form as the i.i.d. case in Eq. (7.3). □

Finally, we can infer the posterior over Bellman operators $P_B(\cdot|s,a,\mathcal{D}_\omega^N)$ by marginalising over the parameters in the likelihood:

$$p_B(b|s,a,\mathcal{D}_\omega^N) = \int p(b|s,a,\phi)dP_\Phi(\phi|\mathcal{D}_\omega^N).$$

Having derived the BBO posterior, we are in a position to address our motivating Question 1:

1. *Do model-free Bayesian RL approaches that use bootstrapping still infer a posterior?*

Our analysis has revealed that by introducing bootstrapping, model-free Bayesian RL infers a posterior over *Bellman operators*, not a posterior over Q-functions. Unlike existing approaches, our posterior density is a function of ω, which correctly accounts for the dependence on $\hat{Q}_\omega$ in our data and can handle learning optimal Q-functions.

We highlight that both the likelihood and prior can also depend on ω if we want to encode prior knowledge of how the underlying Bellman operator varies according to the function approximator, however this is not necessary as the posterior automatically accounts for this dependence due to its conditioning on $\mathcal{D}_\omega^N$. As we anticipate that most applications of BBO will seek to learn an optimal policy, and hence an optimal Q-function, a prior P_Φ that incorporates any knowledge about the optimal Bellman operator will speed learning and give the agent an advantage.

Recall that deriving the posterior over Bellman operators was motivated by solving the problem of finding a Q-function approximator to satisfy the Bayesian Bellman equation introduced in Section 7.1. We therefore need to derive the Bayesian Bellman operator under the BBO model. Introducing the shorthand $\mathcal{B}_{\omega,N} \coloneqq \mathcal{B}[\hat{Q}_\omega](\cdot, \mathcal{D}_\omega^N)$ for the Bayesian Bellman operator, we can obtain an expression for its value under the BBO framework by deriving the the predictive mean:

$$\mathcal{B}_{\omega,N}(s,a) = \int b dP(b|s,a,\mathcal{D}_\omega^N) = \int \hat{B}_\phi(s,a) dP_\Phi(\phi|\mathcal{D}_\omega^N). \tag{7.4}$$

Eq. (7.4) confirms our intuition that the Bayesian Bellman operator is the Bayesian estimate of the true Bellman operator because we evaluate $\mathcal{B}_{\omega,N}(s,a)$ by marginalising over all Bellman operators that our model can represent according to the uncertainty in their value. We can re-write the MSBBE from Section 7.1 using our new notation:

$$\text{MSBBE}_N(\omega) = \|\mathcal{B}_{\omega,N} - \hat{Q}_\omega\|_\varsigma^2.$$

Although the MSBBE has a similar form and functionality to the MSBE, our theoretical results in Section 7.2.1 show its frequentist interpretation is closer to the mean squared projected Bellman error (MSPBE) used by convergent TD algorithms [Sutton et al.,

2009a]. We now derive the gradient of the MSBBE with respect to the function approximator parameters:

$$\begin{aligned}\nabla_\omega \text{MSBBE}_N(\omega) &\propto \mathbb{E}_\varsigma\left[\left(\hat{Q}_\omega - \mathbb{E}_{P_\Phi(\phi|\mathcal{D}_\omega^N)}\left[\hat{B}_\phi\right]\right)\left(\nabla_\omega \hat{Q}_\omega - \nabla_\omega \mathcal{B}_{\omega,N}\right)\right] \\ &= \mathbb{E}_\varsigma\left[\left(\hat{Q}_\omega - \mathbb{E}_{P_\Phi(\phi|\mathcal{D}_\omega^N)}\left[\hat{B}_\phi\right]\right)\left(\nabla_\omega \hat{Q}_\omega - \mathbb{E}_{P_\Phi(\phi|\mathcal{D}_\omega^N)}\left[\hat{B}_\phi \nabla_\omega \log p_\Phi(\phi|\mathcal{D}_\omega^N)\right]\right)\right]. \quad (7.5)\end{aligned}$$

If we can sample from the posterior then unbiased estimates of $\nabla_\omega \text{MSBBE}_N(\omega)$ can be obtained using Eq. (7.5), hence minimising the MSBBE via a stochastic gradient descent algorithm is convergent if the standard Robbins-Munro conditions are satisfied [Robbins and Monro, 1951]. As we should expect from the conditionality principle, minimising the MSBBE also avoids the double sampling problem encountered in frequentist RL that we discussed in Section 3.5.2 where to minimise the MSBE, two independent transition distribution samples $s'_1, s'_2 \sim P(\cdot|s,a)$ are required to obtain unbiased gradient estimates [Baird, 1995], which is not possible because $P(\cdot|s,a)$ is unknown a priori and agents can't arbitrarily select their state. In BBO, this issue is avoided because all Bayesian quantities condition only data that is observed, meaning that an unbiased gradient estimate requires drawing two independent approximate Bellman operators B_{ϕ_1} and B_{ϕ_2} from the posterior $\phi_1, \phi_2 \sim P_\Phi(\cdot|\mathcal{D}_\omega^N)$ instead.

Many existing parametric model-free Bayesian RL approaches attempt to carry out policy evaluation by minimising objectives similar to the MSBBE, using the posterior over Q-functions in place of the BBO posterior but with bootstrapped samples. As $P_Q(\cdot|s,a,\mathcal{D})$ has no dependence on ω, the gradient $\nabla_\omega \log p(\phi|\mathcal{D}_\omega^N)$ is not accounted for, leading to gradient terms being dropped in the update. Like in the TD methods discussed in Section 3.5.1, stochastic gradient descent using these updates does not optimise any objective and so may not converge to any solution. In a Bayesian setting, this problem also affects the quality of the exploration regime as the lack of convergence will prevent $P_B(\cdot|s,a,\mathcal{D}_\omega^N)$ from concentrating and reduces the posterior's ability to correctly characterise the uncertainty in the Bellman operator, potentially undermining the purported benefits of Bayesian exploration.

7.2.1 Consistency of the Posterior

To address Question 2, we develop a set of theoretical results to understand the relationship between BBO's posterior and the frequentist RL problem. By following a similar analysis to the Bayesian perspective presented in Section 7.1, we start by deriving the true distribution that the random variable b is sampled from. We write the frequentist Bellman operator for $\hat{Q}_\omega$ as an expectation of the empirical Bellman function b_ω:

$$\mathcal{B}[\hat{Q}_\omega](s,a) = \mathbb{E}_{P^\pi(s',a'|a,s)}\left[b_\omega(s',a',s,a)\right],\ b_\omega(s',a',s,a) \coloneqq r(s',a,s) + \gamma \hat{Q}_\omega(s',a'). \tag{7.6}$$

Like in our Bayesian analysis, we make a transformation of variables using the empirical Bellman function $b = b_\omega(s',a',s,a)$. The transformed variable has a conditional distribution $P_B(\cdot|s,a,\omega)$ which is the *pushforward* of the true state-action transition distribution $P^\pi(\cdot|s,a)$ under the empirical Bellman function $b_\omega(\cdot,s,a)$. For any P_B-integrable function $f : \mathbb{R} \to \mathbb{R}$, the pushforward satisfies:

$$\mathbb{E}_{P_B(b|s,a,\omega)}\left[f(b)\right] = \mathbb{E}_{P^\pi(s',a'|s,a)}\left[f \circ b_\omega(s',a',s,a)\right]. \tag{7.7}$$

As the pushforward $P_B(\cdot|s,a,\omega)$ is a distribution over empirical Bellman functions, each sample $b \sim P_B(\cdot|s,a,\omega)$ is a noisy sample of the Bellman operator at a point: $b_i = \mathcal{B}[\hat{Q}_\omega](s_i,a_i) + \eta_i$. To prove this, observe that taking expectations of b recovers $\mathcal{B}[\hat{Q}_\omega](s,a)$:

$$\mathbb{E}_{P_B(b|s,a,\omega)}[b] \underbrace{=}_{\text{Eq. (7.7)}} \mathbb{E}_{P^\pi(s',a'|s,a)}\left[b_\omega(s',a',s,a)\right] \underbrace{=}_{\text{Eq. (7.6)}} \mathcal{B}[\hat{Q}_\omega](s,a). \tag{7.8}$$

The aim of our analysis is to demonstrate consistency between the frequentist pushforward $P_B(\cdot|s,a,\omega)$ and the Bayesian posterior $P_B(\cdot|s,a,\mathcal{D}_\omega^N)$ in the limit of $N \to \infty$. We first introduce some regularity assumptions on our choice of Bayesian model:

Assumption 9 (Regularity of Model)**.** *i) $\hat{Q}_\omega$ is bounded and: (Φ, d_Φ) and (Ω, d_Ω) are compact metric spaces; ii) $\hat{B}_\phi$ is Lipschitz in ϕ, $P(\cdot|s,a,\phi)$ has finite variance and a density $p(b|s,a,\phi)$ which is Lipschitz in ϕ and bounded; and iii) $p(\phi) \propto \exp\left(-R(\phi)\right)$ where $R(\phi)$ is bounded and Lipschitz.*

Any Q-function is upper and lower bounded as:

$$\frac{r_{\min}}{1-\gamma} \leq Q^{\pi} \leq \frac{r_{\max}}{1-\gamma},$$

which can be used as a natural bound when designing the Q-function approximator to ensure Assumption 9 $i)$ is satisfied. As the Q-function approximator and reward function are bounded, it follows from Eq. (7.6) that each $b \sim P_B(\cdot|s,a,\omega)$ is bounded too, hence $P_B(\cdot|s,a,\omega)$ has finite variance. The assumption that the model likelihood $P(\cdot|s,a,\phi)$ has finite variance under Assumption 9 $ii)$ therefore does not affect its capacity to represent $P_B(\cdot|s,a,\omega)$. Finally, the prior being bounded under Assumption 9 $iii)$ avoids pathological cases where the prior places all its mass on a finite number of parametrisations.

Our main result is a Bernstein-von-Mises under misspecification theorem [Kleijn and van der Vaart, 2012] applied to reinforcement learning. We prove that the posterior asymptotically converges to a Dirac delta distribution centred on the set of parameters that minimise the KL divergence between the true and model distributions, which are the frequentist maximum likelihood parameters::

$$\begin{aligned}\phi^{\star}_{\omega} &\coloneqq \arg\min_{\phi\in\Phi} \mathrm{KL}(P_B(b,s,a|\omega) \parallel P(b,s,a|\phi)) \\ &= \arg\min_{\phi\in\Phi} \mathbb{E}_{P_B(b,s,a|\omega)} \left[-\log p(b,s,a|\phi)\right], \end{aligned} \tag{7.9}$$

where the expectation is taken with respect to true (frequentist) distribution that generates the observed data: $dP_B(b,s,a|\omega) = dP_B(b|s,a,\omega)d\pi(a|s)d\rho(s)$. If the BBO model can sufficiently represent the true conditional distribution then $\mathrm{KL}(P_B(\cdot|\omega) \parallel P(\cdot|\phi^{\star}_{\omega})) = 0 \implies P_B(\cdot|s,a,\omega) = P_B(\cdot|s,a,\phi^{\star}_{\omega})$.

We make a simplifying assumption that there is a single maximum likelihood parameter, which eases analysis and exposition of our results, allowing us to prove convergence of $P_{\Phi}(\cdot|\mathcal{D}^N_{\omega})$ to a single Dirac delta measure. In the more realistic situation where the KL divergence may have multiple, disjoint minimisers, our analysis holds, however the posterior converges to a weighted sum of Dirac delta measures centred on each element of the set of minimisers specific to the exact MDP being studied.

Assumption 10 (Single Minimiser)**.** *The set of maximum likelihood parameters $\phi^\star_\omega$ defined in Eq.* (7.9) *exists and is a singleton.*

We denote the true distribution of the complete data $\mathcal{D}_\omega := \{b_0, s_0, a_0, b_1, s_1, a_1 ...\}$ as $P_\mathcal{D}$, which characterises the nature of convergence. We are now ready to prove consistency between the Bayesian BBO framework and approaches that use the frequentist Bellman operator. The proof of this result is technical and details can be found in Appendix B.1.

Theorem 5. *Under Assumptions 8-10, in the limit $N \to \infty$: i) $P_\Phi(\cdot|\mathcal{D}^N_\omega) \xrightarrow[P_\mathcal{D}-a.s]{} \delta_{\phi^\star_\omega}$; ii) $\mathcal{B}_{\omega,N} \xrightarrow[P-a.s.]{} \hat{B}_{\phi^\star_\omega}$; and iii) MSBBE$_N(\omega) \xrightarrow[P-a.s.]{} \|\hat{Q}_\omega - \hat{B}_{\phi^\star_\omega}\|^2_\varsigma$.*

Proof. See Appendix B.1. □

Theorem 5 proves that under mild regularity assumptions, the posterior concentrates on the frequentist solution $\phi^\star_\omega$ and hence the Bayesian Bellman operator converges to the true Bellman operator: $\hat{B}_{\phi^\star_\omega}(s,a) = \mathbb{E}_{P(b|s,a,\phi^\star_\omega)}[b] = \mathbb{E}_{P_B(b|s,a,\omega)}[b] = \mathcal{B}[\hat{Q}_\omega](s,a)$ almost surely, which we can use to answer Question 2:

2. *If it exists, how does this posterior relate to solving the RL problem?*

As every Bellman operator characterises an MDP, any Bayesian RL solution obtained using the BBO posterior such as an optimal policy or value function is consistent with the true RL solution. When the true distribution is not in the model class, $B_{\phi^\star_\omega}$ converges to the closest representation of the true Bellman operator according to the parametrisation that maximises the likelihood $\mathbb{E}_{P_B(b,s,a|\omega)}[\log p(b,s,a|\phi)]$. This is analogous to frequentist convergent TD learning where the function approximator converges to a parametrisation that minimises the projection of the Bellman operator into the model class [Sutton et al., 2009a,b, Bhatnagar et al., 2009]. We now make this relationship precise by considering a Gaussian likelihood model.

7.2.2 Gaussian BBO

To showcase the power of Theorem 5 and to provide a direct comparison to existing frequentist approaches discussed in Section 6.2.3, we consider the nonlinear Gaussian

model $P(\cdot|s,a,\phi) = \mathcal{N}(\hat{B}_\phi(s,a), \sigma^2)$ that is commonly used for Bayesian regression [Murphy, 2012, Gal, 2016]. The mean is a nonlinear function approximator that best represents the Bellman operator $\hat{B}_\phi \approx \mathcal{B}[\hat{Q}_\omega]$ and the model variance $\sigma^2 > 0$ represents the aleatoric uncertainty in our samples.

Defining the log-normalisation constant as $c_{\text{norm}} \coloneqq \log \int_\phi \prod_{i=1}^N p(b_i|s_i,a_i,\phi) dP(\phi)$, we can derive the exact form of the $\log$-posterior density for our Gaussian model:

$$\begin{aligned}
p(\phi|\mathcal{D}_\omega^N) &= \exp(-c_{\text{norm}}) \prod_{i=1}^N p(b_i|s_i,a_i,\phi) p(\phi), \\
&= \exp(-c_{\text{norm}}) \prod_{i=1}^N \exp\left(-\frac{1}{2\sigma^2}(b_i - \hat{B}_\phi(s_i,a_i))^2\right) \exp\left(-R(\phi)\right), \\
&= \exp\left(-c_{\text{norm}} - \frac{1}{2\sigma^2}\sum_{i=1}^N (b_i - \hat{B}_\phi(s_i,a_i))^2 - R(\phi)\right), \\
\implies -\log p(\phi|\mathcal{D}_\omega^N) &= c_{\text{norm}} + \sum_{i=1}^N \frac{(b_i - \hat{B}_\phi(s_i,a_i))^2}{2\sigma^2} + R(\phi), \qquad (7.10)
\end{aligned}$$

Ignoring the $\log$-normalisation constant c_{norm}, the $\log$-posterior is simply a regularised empirical mean squared error between the empirical Bellman samples and the model mean $\hat{B}_\phi(s_i,a_i)$. Finding a MAP estimate of the posterior is therefore equivalent to minimising the empirical TD-target objective from Eq. (3.18) used by frequentist approaches with target parameters, with an additional regularisation term that is controlled by σ. We can provide a deeper connection to frequentist approaches by exploring the long term behaviour of the Gaussian BBO posterior. Firstly, we derive an expression the set of KL minimising parameters that we expect our posterior to concentrate on:

$$\begin{aligned}
\phi_\omega^\star &\coloneqq \arg\min_{\phi\in\Phi} \mathbf{KL}(P_B(b,s,a|\omega) \parallel P(b,s,a|\phi)) \\
&= \arg\min_{\phi\in\Phi} \mathbb{E}_{P_B(b,s,a|\omega)} \left[-\log p(b,s,a|\phi)\right], \\
&= \arg\min_{\phi\in\Phi} \mathbb{E}_{P_B(b,s,a|\omega)} \left[-\log \left(p(b|s,a,\phi)\rho(s)\pi(a|s)\right)\right], \\
&= \arg\min_{\phi\in\Phi} \mathbb{E}_{P_B(b,s,a|\omega)} \left[-\log p(b|s,a,\phi)\right], \\
&= \arg\min_{\phi\in\Phi} \mathbb{E}_{P_B(b,s,a|\omega)} \left[(b - \hat{B}_\phi(s,a))^2\right], \\
&= \arg\min_{\phi\in\Phi} \mathbb{E}_{\rho(s)} \left[\mathbb{E}_{\pi(a|s)} \left[\mathbb{E}_{P_B(b|s,a\omega)} \left[(b - \hat{B}_\phi(s,a))^2\right]\right]\right], \qquad (7.11)
\end{aligned}$$

where we have ignored terms independent of ϕ and substituted for the definition of the Gaussian density in deriving the penultimate line. Analysing the inner expectation in Eq. (7.11) with respect to $P_B(\cdot|s, a\omega)$, which we denote as $\mathbb{E}_B[\cdot]$ for convenience, yields:

$$\begin{aligned}\mathbb{E}_B\left[(b - \hat{B}_\phi)^2\right] &= \mathbb{E}_B\left[b^2\right] - 2\mathbb{E}_B\left[b\right]\hat{B}_\phi + \hat{B}_\phi^2, \\ &= \mathbb{E}_B\left[b^2\right] - 2\mathcal{B}[\hat{Q}_\omega]\hat{B}_\phi + \hat{B}_\phi^2,\end{aligned}$$

where we have used $\mathbb{E}_B\left[b\right] = \mathcal{B}[\hat{Q}_\omega]$ from Eq. (7.8). Denoting the conditional variance of b as $\mathbb{V}\left[b\right]$, we substitute for $\mathbb{E}_B\left[b^2\right] = \mathbb{V}\left[b\right] + \mathcal{B}[\hat{Q}_\omega]^2$:

$$\begin{aligned}\mathbb{E}_B\left[(b - \hat{B}_\phi)^2\right] &= \mathbb{V}\left[b\right] + \mathcal{B}[\hat{Q}_\omega]^2 - 2\mathcal{B}[\hat{Q}_\omega]\hat{B}_\phi + \hat{B}_\phi^2, \\ &= \mathbb{V}\left[b\right] + (\mathcal{B}[\hat{Q}_\omega] - \hat{B}_\phi)^2.\end{aligned}$$

As $\mathbb{V}\left[b\right]$ has no dependence on ϕ and we are finding $\arg\min_{\phi\in\Phi}$, we can ignore it from our derivation. Scaling by $\frac{1}{2}$ and substituting into Eq. (7.11) yields:

$$\phi^\star_\omega = \arg\min_{\phi\in\Phi} \|\mathcal{B}[\hat{Q}_\omega] - \hat{B}_\phi\|^2_{\rho,\pi}.$$

Theorem 5 proves that in the limit $N \to \infty$, the effect of the regularising prior diminishes and the Bayesian Bellman operator converges to the parametrisation: $\mathcal{B}_{\omega,N} \xrightarrow{a.s.} \hat{B}_{\phi^\star_\omega}$. As $\phi^\star_\omega$ is the set of parameters that minimise the mean squared error between the true Bellman operator and the approximator, $\hat{B}_{\phi^\star_\omega}$ is the *projection* of the true Bellman operator onto the space of functions represented by $\hat{B}_\phi$ as defined in Eq. (3.11):

$$\hat{B}_{\phi^\star_\omega} = \mathcal{P} \circ \mathcal{B}[\hat{Q}_\omega] \coloneqq \{\hat{B}_\phi : \phi \in \arg\min_{\phi\in\Phi} \|\mathcal{B}[\hat{Q}_\omega] - \hat{B}_\phi\|^2_{\rho,\pi}\}.$$

Finally, Theorem 5 proves that the MSBBE converges to the mean squared projected Bellman error (MSPBE):

$$\text{MSBBE}_N(\omega) \xrightarrow{P_\mathcal{D}-a.s.} \text{MSPBE}(\omega) \coloneqq \|\hat{Q}_\omega - \mathcal{P} \circ \mathcal{B}[\hat{Q}_\omega]\|^2_{\rho,\pi}$$

Recall from our analysis in Section 3.5.3, a solution $\hat{Q}_\omega = \mathcal{P} \circ \mathcal{B}[\hat{Q}_\omega]$ is a TD fixed point; hence any asymptotic MSBBE minimiser parametrises a TD fixed point should it exist. Our analysis has provided a rigorous proof that Bayesian approaches under the

BBO framework and frequentist TD approaches converge to the same solution. Bayesian approaches also afford us with uncertainty quantification, optimal exploration regimes and are provably convergent, which we demonstrate holds even under posterior approximation in Section 7.3. Finally, we provide a full exposition of the linear Gaussian BBO model in Appendix C as a case study.

7.2.3 Control as Inference from a Bayesian Perspective

We conclude this section by analysing the correct Bayesian interpretation of the MERL framework introduced in Section 5.1. We can derive a Bayesian version of algorithms from the MERL framework by introducing the empirical soft Bellman equation:

$$b_\omega^{\text{soft}}(r, s') = r + \gamma \int \left(\hat{Q}_\omega(s', a') - \alpha \log \pi(a'|s')\right) d\pi(a'|s').$$

Given a dataset of observations $\mathcal{D}_\omega^N \coloneqq \{b_i^{\text{soft}}\}_{i\in\{1:N\}}$ where $b_i^{\text{soft}} \coloneqq b_\omega^{\text{soft}}(r_i, s_i')$, for any model $P_B(\cdot|s, a, \phi)$ and prior P_Φ, the density of the posterior is:

$$p_\Phi(\phi|\mathcal{D}_\omega^N) \propto \prod_{i=1}^{N} p_B(b_i^{\text{soft}}|s_i, a_i, \phi) p_\Phi(\phi). \tag{7.12}$$

Observe from Eq. (7.12) that it is only the likelihood term that encodes the entropy regularisation through each empirical soft Bellman sample, which represents the precise knowledge that the agent is solving an entropy regularised RL problem. As this knowledge is encoded in the likelihood, it can never be overwhelmed by the data. This confirms our concerning conclusion from Section 5.4 that control as inference is solving an inference problem whose likelihood is always inconsistent with our prior beliefs no matter what the MDP.

A more principled approach to encoding the introduction of regularisation into the RL objective would be through the prior. In this way, data would eventually overwhelm this incorrect belief and the posterior can concentrate on the true MDP according to Theorem 5. The VIREL posterior thus provides a heuristic approach to approximating this behaviour.

7.3 Approximate BBO

We have demonstrated in Eq. (7.5) that if it is tractable to sample from the posterior, a simple convergent stochastic gradient descent algorithm can be used to minimise the MSBBE. Unfortunately, models like linear Gaussians that have analytic posteriors are often too simple to accurately represent the Bellman operator for domains of practical interest in RL. We now extend our analysis to include approximate inference approaches that can handle more expressive nonlinear function approximators for which the posterior normalisation is intractable. Throughout this section, we introduce two methods to learn a tractable posterior approximation: $\hat{P}_\Phi(\cdot|\mathcal{D}_\omega^N) \approx P(\cdot|\mathcal{D}_\omega^N)$. In particular, we will use the approximate posterior to derive a tractable approximation to the MSBBE:

$$\text{MSBBE}(\omega) \approx \|\hat{Q}_\omega - \int \hat{B}_\phi d\hat{P}_\Phi(\phi|\mathcal{D}_\omega^N)\|_\varsigma^2. \tag{7.13}$$

7.3.1 Connection to Frequentist TD Learning

If the reader has still yet to be convinced that the only sensible approach to solving an RL problem when the agent is uncertain of the environment is to be Bayesian, we hope to put their concerns to rest by demonstrating that well-principled, convergent TD methods can be derived from the BBO framework by introducing a MAP point estimate of the posterior:

$$\hat{P}_\Phi(\cdot|\mathcal{D}_\omega^N) = \delta_{\psi^\star(\mathcal{D}_\omega^N)}, \quad \psi^\star(\mathcal{D}_\omega^N) \in \arg\max_{\phi\in\Phi} \log p(\phi|\mathcal{D}_\omega^N).$$

As MAP approaches to approximate inference replace the posterior point estimates, they cannot characterise the uncertainty in the MDP and so will have no inherent exploration properties. Regardless, an MAP approximate posterior is still useful in settings outside of the control problem where we would like to carry out policy evaluation and cannot influence the policy collecting data. In these situations, obtaining uncertainty estimates may waste computational resources, yet an agent can still benefit from the regularisation,the ability to incorporate pre-existing knowledge that introducing a prior affords and the fact that the conditionality principle holds for approximate Bayesian

methods. Substituting for the MAP approximate posterior in Eq. (7.13) yields the approximate MSBBE objective:

$$\text{MSBBE}(\omega) \approx \|\hat{Q}_\omega - \hat{B}_{\psi^\star(\mathcal{D}_\omega^N)}\|_\varsigma^2. \tag{7.14}$$

If we make the additional assumption that $\hat{Q}_\omega$ and $\hat{B}_\phi$ share a function space, that is $\Phi = \Omega$ and $\hat{Q}_\omega = \hat{B}_\omega$ for any $\omega \in \Omega$ then, when a fixed point $\hat{Q}_{\omega^\star} = \hat{B}_{\psi^\star(\mathcal{D}_{\omega^\star}^N)}$ exists, minimising Eq. (7.14) is equivalent to finding an $\omega^\star$ such that $\psi^\star(\mathcal{D}_{\omega^\star}^N) = \omega^\star$. In this case, we can minimise the simpler auxiliary objective:

$$\mathcal{U}(\omega; \psi^\star) = \|\omega - \psi^\star(\mathcal{D}_\omega^N)\|_2^2.$$

which has the advantage that deterministic gradient updates can be obtained. The auxiliary objective $\mathcal{U}(\omega)$ can still provide an alternative auxilliary objective when there does not exist an $\omega^\star$ such that $\psi^\star(\mathcal{D}_{\omega^\star}^N) = \omega^\star$ as the minimum of Eq. (7.14) has the same solution as minimising $\mathcal{U}(\omega)$ for sufficiently smooth $\hat{B}_\phi$.

To connect BBO to TD learning approaches that minimise a mean-squared error, we use a Gaussian likelihood. Using the derivation of the log-posterior from Eq. (7.10) in Section 7.2.2, the MAP minimises the regularised empirical TD-target loss:

$$\psi^\star(\mathcal{D}_\omega^N) \in \arg\min_{\phi \in \Phi} \left(\sum_{i=1}^N \frac{(b_i - \hat{B}_\phi(s_i, a_i))^2}{2} + \sigma^2 R(\phi) \right).$$

Other TD losses of the form $\mathbb{E}_{P_B(b,s,a|\omega)}\left[f(b, s, a, \phi)\right]$ can be accommodated through the choice of likelihood $p(b|s, a, \phi) \propto \exp(-f(b, s, a, \phi))$.

To minimise the approximate MSBBE we must solve a system of equations that form a *bilevel* optimisation problem of the form:

$$\omega \in \arg\min_{\omega \in \Omega} \mathcal{U}(\omega; \psi^\star) \quad \text{s.t.} \quad \psi^\star \in \arg\min_{\phi \in \Phi} \mathcal{L}(\phi, \omega).$$

where $\mathcal{L}(\phi, \omega) = -\log p(\phi|\mathcal{D}_\omega^N)$ for the case of the MAP estimate. Solving bi-level optimisation problems is NP-hard [Bard, 1991]. To tackle this problem, we propose a

two timescale gradient descent approach, introducing a variable $\psi \in \Phi$ to track $\psi^\star(\mathcal{D}_\omega^N)$. Taking gradients of $\mathcal{U}(\omega; \psi^\star)$ and $\mathcal{L}(\psi, \omega)$ leads to the updates:

$$\psi_{k+1} = \mathcal{P}_\Phi\Big(\psi_k + \alpha_k \Big((r + \gamma \hat{Q}_{\omega_k}(s', a') - \hat{B}_{\psi_k}(s, a))\nabla_\phi \hat{B}_{\psi_k}(s, a)\Big) + \sigma^2 \nabla_\phi R(\psi_k)\Big), \quad \text{(fast)} \tag{7.15}$$

$$\omega_{k+1} = \mathcal{P}_\Omega\left(\omega_k + \beta_k\left(\psi_k - \omega_k\right)\right), \quad \text{(slow)} \tag{7.16}$$

which are updated on asymptotically faster and slower timescales with $\lim_{k\to\infty} \frac{\beta_k}{\alpha_k} = 0$. We have introduced projection operators $\mathcal{P}_\Omega(\cdot) \coloneqq \arg\min_{\omega\in\Omega} \|\cdot - \omega\|_2^2$ into our updates which keep the parameter updates within the compact parameter spaces. These are formally required to prove convergence and, provided Φ and Ω are large enough, we don't expect to require projection in practice.

Comparing MAP Gaussian BBO to the TD-target algorithm analysed in Section 3.5.3, we see that in the limit $\sigma^2 \to 0$, Eq. (7.15) is exactly update from Eq. (3.18). As the parameters ω_k are updated on an asymptomatically slower timescale in Eq. (7.16), they are lagged in comparison to ψ_k, playing exactly the same role as the lagged critic parameters that are necessary stabilise the TD algorithm updates when formal convergence cannot be proved. Our convergence analysis in Section 7.3.3 confirms that by taking a Bayesian approach to solving the RL objective, we can derive TD updates that are provably convergent with a greater generality of sampling regimes and function approximators than their frequentist counterparts. Finally, we show in Appendix C that the convergent off policy linear TD algorithms that minimise the MSPBE such as GTD, GTD2/TDC and Greedy-GQ [Sutton et al., 2009a,b, Maei et al., 2010] algorithms can be derived from a linear BBO model in the frequentist limit of $N \to \infty$.

7.3.2 Randomised Priors BBO

Although an MAP estimate allows a direct comparison between BBO and existing frequentist TD approaches, it is an approximation that does not characterise the uncertainty in the Bellman operator and so cannot be used with Bayesian exploration regimes motivated in Chapter 6. To address this issue, we approximate the posterior using RP by applying the method outlined in Section 4.3.3 to the BBO posterior. Although BBO is

compatible with any tractable approximate inference approach, existing work shows that RP uncertainty estimates are conservative [Pearce et al., 2019, Ciosek et al., 2020] with strong empirical performance in RL [Osband et al., 2018, 2019] for the Gaussian model that we provide as a continuing case study in this thesis.

Recall from Section 4.3.3 that RP randomises the MAP point estimate obtained in Section 7.3.1 by injecting noise into the prior. Here, a noise variable $\epsilon \in \mathcal{E}$ with distribution $P_E(\epsilon)$ where the density $p_E(\epsilon)$ has the same form as the prior is introduced, which defines a continuum of ϵ-randomised MAP estimates:

$$\psi^\star(\epsilon; \mathcal{D}_\omega^N) = \arg\max_{\phi \in \Phi} \mathcal{L}(\phi; \mathcal{D}_\omega^N, \epsilon),$$
$$\mathcal{L}(\phi; \mathcal{D}_\omega^N, \epsilon) \coloneqq \frac{1}{N}\left(R(\phi - \epsilon) - \sum_{i=1}^{N} \log p(b_i | s_i, a_i, \phi)\right). \tag{7.17}$$

It is implicitly assumed when using RP that the prior-randomised MAP estimate $\psi^\star(\epsilon; \mathcal{D}_\omega^N)$ is P_E-integrable and, for ease of analysis, that each $\arg\max_{\phi \in \Phi} \mathcal{L}(\phi; \mathcal{D}_\omega^N, \epsilon)$ is a singleton. The RP approximate posterior can then be constructed by averaging over all MAP estimates using P_E:

$$\hat{P}_\Phi(\cdot | \mathcal{D}_\omega^N) = \int \delta_{\psi^*(\epsilon; \mathcal{D}_\omega^N)} dP_E(\epsilon).$$

A naïve way to sample from the approximate posterior is to first sample $\epsilon \sim P_E$ and solve the optimisation problem $\psi^\star(\epsilon; \mathcal{D}_\omega^N) = \arg\max_{\phi \in \Phi} \mathcal{L}(\phi; \mathcal{D}_\omega^N, \epsilon)$ each time the posterior needs sampling. This approach is not tractable as solving this optimisation problem is NP-hard. RP in practice uses ensembling: L prior randomisations $\mathcal{E}_L \coloneqq \{\epsilon_l\}_{l=1:L}$ are first drawn from P_E. For each $l \in \{1 : L\}$, a set of solutions to the prior-randomised MAP objective are then found:

$$\psi_l^\star(\omega_l) \in \arg\min_{\phi \in \Phi} \mathcal{L}(\phi; \mathcal{D}_{\omega_l}^N, \epsilon_l)$$

To use RP for BBO, we write the Q-function approximator as an ensemble of L parameters $\Omega_L \coloneqq \{\omega_l\}_{l=1:L}$ where $\hat{Q}_\omega = \frac{1}{L}\sum_{l=1}^{L} \hat{Q}_{\omega_l}$ and require an assumption about the prior and the function spaces used for approximators:

Assumption 11 (RP Function Spaces)**.** *i)* $\hat{Q}_{\omega_l}$ *and* $\hat{B}_{\omega_l}$ *share a function space, that is* $\hat{Q}_{\omega_l'} = \hat{B}_{\omega_l'}$ *for any* $\omega_l' \in \Omega$*, where* $\Phi = \Omega \subset \mathbb{R}^n$ *is compact, convex with a smooth boundary. ii)* $\mathcal{E} \subseteq \mathbb{R}^n$ *and* $R(\phi - \epsilon)$ *is defined for any* $\phi \in \Phi, \epsilon \in \mathcal{E}$*.*

Assumption 11 $i)$ is easily satisfied by defining $\Omega = \Phi$ to be a closed ball of arbitrary radius, which should be applicable to the majority of cases where parametric function approximators are used. It is common in RP to use Gaussian a prior P_Φ, which technically violates the assumption that Φ is compact. We can still satisfy Assumption 11 $i)$ by defining P_Φ to be a truncated Gaussian. This won't affect the algorithm or objectives in practice as P_Φ simply adds a regularisation term and there are no restrictions on $\mathcal{E}$, hence P_E can still be Gaussian.

We now further justify our choice of RP by proving that the consistency results from Theorem 5 can be extended to a RP approximate posterior:

Corollary 5.1. *Under Assumptions 8-11, results i)-iii) of Theorem 5 hold with* $P(\cdot|\mathcal{D}_\omega^N)$ *replaced by the RP approximate posterior* $\hat{P}_\Phi(\cdot|\mathcal{D}_\omega^N)$ *both with or without ensembling.*

Proof. See Appendix B.1. □

Corollary 5.1 can be used to answer Question 3:

3. *What effect does approximate inference have on the solution?*

Our results shows that the difference between using the RP approximate posterior and the true posterior lies in their characterisation of uncertainty and not their asymptotic behaviour, hence any approach using an RP approximate posterior will be consistent with the equivalent frequentist approach. Observe that Corollary 5.1 also includes the MAP point estimate as a special case.

The RP approximate posterior $\hat{P}_\Phi(\cdot|\mathcal{D}_\omega^N)$ depends on the ensemble of Q-function approximators $\hat{Q}_{\omega_l}$ and like in Section 7.2, we must learn an ensemble of optimal parametrisations $\omega_l^\star$. RP treats learning each ω_l as a separate problem for each ensemble [Osband et al., 2018], leading to the randomised priors ensembled MSBBE:

$$\text{MSBBE}_{\text{RP}}(\omega_l) \coloneqq \|\hat{Q}_{\omega_l} - \hat{B}_{\psi_l^\star(\omega_l)}\|_\varsigma^2,$$

from which we can derive an auxiliary parameter objective using the same principles as for the MAP objective in Section 7.3.1:

$$\mathcal{U}(\omega_l; \psi_l^\star) = \|\omega_l - \psi_l^\star(\mathcal{D}_\omega^N)\|_2^2.$$

Optimising the system of equations $\arg\min_{\omega\in\Omega} \mathcal{U}(\omega_l; \psi_l^\star)$ such that $\psi_l^\star(\omega_l) \in \arg\min_{\phi\in\Phi} \mathcal{L}(\phi; \mathcal{D}_{\omega_l}^N, \epsilon_l)$ is exactly equivalent to finding a prior-randomised MAP solution for each ensemble, hence we propose adapting the two-timescale gradient approached to find the MAP solution that we developed in Section 7.3.1. We introduce an ensemble of parameters $\Psi_L \coloneqq \{\psi_l\}_{1:L}$ to track each $\psi_l^\star(\omega_l)$ and carry out a two-timescale gradient update for each $l \in \{1 : L\}$ on the RP objectives:

$$\psi_{l,k+1} = \mathcal{P}_\Omega \left(\psi_{l,k} - \alpha_k \nabla_\phi \left(R(\psi_{l,k} - \epsilon_l) - \log p(b|s, a, \psi_{l,k})\right)\right), \quad \text{(fast)} \tag{7.18}$$

$$\omega_{l,k+1} = \mathcal{P}_\Omega(\omega_{l,k} - \beta_k(\omega_{l,k} - \psi_{l,k})). \quad \text{(slow)} \tag{7.19}$$

From a Bayesian perspective, we are concerned with characterising the uncertainty after a *finite* number of samples $N < \infty$ and hence (b, s, a) should be drawn uniformly from the dataset $\mathcal{D}_{\omega_l}^N$ to form estimates of the summation in Eq. (7.17), which becomes intractable with large N. When compared to existing RL algorithms, sampling from $\mathcal{D}_{\omega_l}^N$ is analogous to sampling from a replay buffer [Mnih et al., 2015]. A frequentist analysis of our updates is also possible by considering samples that are drawn online from the underlying data generating distribution $(b_i, s_i, a_i) \sim P_B$ in the limit $N \to \infty$. We discuss this frequentist interpretation further in Appendix B.2.1.

As ω_l are updated on a slower timescale, they lag the parameters ψ_l. When deriving a Bayesian actor-critic algorithm in Section 7.4.2, we demonstrate that these parameters share a similar role to a *lagged critic*. We emphasise that there is no Bayesian explanation for these parameters under existing approaches: when applying approximate inference to the posterior over Q-functions P_Q^π, the RP solution $\psi_l^\star$ has no dependence on ω_l. Hence, minimising $\mathcal{U}(\omega_l; \psi_l^\star)$ and the approximate MSBBE has an exact solution by setting $\omega_l^\star = \psi_l^\star$. In this case, $\hat{Q}_{\omega_l^\star} = \hat{B}_{\psi_l^\star}$ meaning that existing approaches do not distinguish between the Q-function and Bellman operator approximators.

7.3.3 Convergence Analysis

To answer Question 4), we prove convergence of updates (7.18) and (7.19) using a straightforward application of two-timescale stochastic approximation [Borkar, 1997, 2008, Heusel et al., 2017] to BBO. Two timescale analysis extends the single timescale ODE method that we used to prove convergence of linear TD in Section 3.5.3. Our analysis proves that the faster timescale update (7.18) converges to an element in Ω using standard martingale arguments, viewing the parameter ω_l as quasi-static as it behaves like a constant. Since the perturbations are relatively small, the separation of timescales then ensures that ψ_l tracks $\psi_l^\star(\omega_l)$ whenever ω_l is updated in the slower timescale update (7.19), viewing the parameter ψ_l as quasi-equilibrated [Borkar, 2008].

Unlike in the single timescale case, our parameter updates makes use of a projection operator, and so to derive the limiting ODEs of our updates, we characterise the directional derivative of the projection operator in Proposition 8:

Proposition 8 (Existance of Directional Derivative, Shapiro [1988])**.** *The directional derivative of the projection operator at $\omega \in \Omega$ in the direction $y \in \mathbb{R}^n$ given by:*

$$\Gamma_\omega(y) \coloneqq \lim_{\epsilon \downarrow 0} \left(\frac{\mathcal{P}_\Omega(\omega + \epsilon y) - \omega}{\epsilon} \right)$$

which always exists under Assumption 11 and is equivalent to the projection of y onto the tangent cone $T_\Omega(\omega)$:

$$\Gamma_\omega(y) = \mathcal{P}_{T_\Omega(\omega)}(y) \coloneqq \underset{t \in T_\Omega(\omega)}{\arg\min} \|y - t\|_2^2$$

where $T_\Omega(\omega) \coloneqq \text{Closure}\left(\bigcup_{\varepsilon > 0} \bigcup_{\omega' \in \Omega} \frac{1}{\varepsilon}(\omega' - \omega) \right)$

Proof. See Shapiro [1988]. □

The directional derivative of the projection operator has been well studied [Zarantonello, 1971, Shapiro, 1987, 1988, Borkar, 2008] and we provide an informal sketch of its interpretation in Fig. 7.2 for the reader's intuition.

We see that for any point $\omega_1 \in \Omega$ in the interior of Ω, the projection operator is simply the identity function: $\Gamma_{\omega_1}(y) = y$ as $y \in T_\Omega(\omega_1)$. For any point ω_2 on the boundary of Ω, the tangent cone is the closure of the cone formed by all half-lines emanating from ω_2 intersecting Ω in at least one point distinct from ω_2. There are two cases to consider for boundary points. Firstly, when the directional vector y_1 defines a half-line from ω_2 that intersects Ω, the projection operator is the identity $\Gamma_{\omega_2}(y) = y$. In the second case when the directional vector y_2 defines a half-line from ω_2 that leaves Ω, $\Gamma_{\omega_2}(y_2)$ returns the nearest element along the boundary of $T_\Omega(\omega_2)$ (solid purple) according to the projection $\mathcal{P}_{T_\Omega(\omega_2)}(y_2)$.

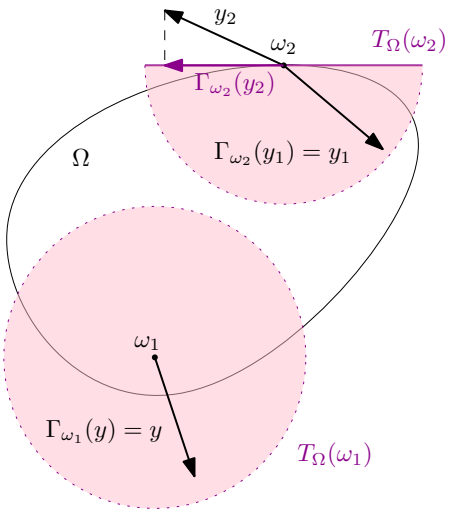


Figure 7.2: Sketch of Tangent Cones in Ω

Using the directional derivative of the projection operator, we can derive the limiting ODE of update (B.8):

$$\partial_t \psi_l(t) = -\Gamma_{\psi_l(t)}(\nabla_{\psi_l} \mathcal{L}(\psi_l(t); \mathcal{D}^N_{\omega_l}, \epsilon_l)) \tag{7.20}$$

Under standard ODE analysis, any equilibria of Eq. (7.20) satisfy $-\Gamma_{\psi_l(t)}(\nabla_{\psi_l} \mathcal{L}(\psi_l(t); \mathcal{D}^N_{\omega_l}, \epsilon_l)) = 0$ and we denote an asymptotically stable local equilibrium as $\psi_l^{\circledast}(\omega_l)$. Of course, there may be several or even infinite stable local equilibrium for the ODE, but as our assumptions only require the existence of at least one stable attractor within the domain of attraction defined by the initial parametrisations, we lose no generality by considering $\psi_l^{\circledast}(\omega_l)$ in isolation.

As the fast update converges asymptotically quicker than the slower update, we consider ψ_l to be equilibrated at $\psi_l^{\circledast}(\omega_l)$ when analysing the limiting ODE of the slow update. We make this argument rigorously in Theorem 6, from which we derive the limiting ODE for the update (7.19) as:

$$\partial_t \omega_l(t) = -\Gamma_{\omega_l(t)}(\omega_l(t) - \psi_l^{\circledast}(\omega_l(t))). \tag{7.21}$$

Crucially, this provides reassurance that our updates and any equilibria satisfying $\Gamma_{\omega_l(t)}(\omega_l(t) - \psi_l^{\circledast}(\omega_l)) = 0$ preserve the dependence of ψ_l on ω_l. We denote an asymptotically stable local equilibrium of Eq. (7.21) as $\omega_l^{\circledast}$. To prove convergence, we introduce the standard two-timescale regularity assumptions and derive the limiting ODEs of updates (7.18) and (7.19):

Assumption 12 (Two-timescale Regularity)**.** *i)* $\nabla_{\psi_l} R(\psi_l - \epsilon_l)$, $\nabla_{\psi_l} \log p(b_i|s_i, a_i, \psi_l)$ *and* $\Gamma_{\psi_l}\left[\nabla_{\psi_l}\mathcal{L}(\psi_l; \mathcal{D}_{\omega_l}^N, \epsilon)\right]$ *are Lipschitz in* ψ_l, $\Gamma_{\omega_l}\left[-(\omega_l - \psi_l)\right]$ *is Lipschitz in* ω_l *and* $(b_i, s_i, a_i) \sim \text{Unif}(\mathcal{D}_{\omega_l}^N)$*; ii)* $\psi^{\circledast}(\omega_l)$ *and* $\omega_l^{\circledast}$ *are local aysmptotically stable attractors of the limiting ODEs of updates (7.18) and (7.19) respectively and* $\psi_l^{\circledast}(\omega_l)$ *is Lipschitz in* ω_l*; and iii) The stepsizes satisfy:* $\lim_{k\to\infty} \frac{\beta_k}{\alpha_k} = 0$, $\sum_{k=1}^{\infty} \alpha_k = \sum_{k=1}^{\infty} \beta_k = \infty$, $\sum_{k=1}^{\infty} (\alpha_k^2 + \beta_k^2) < \infty$.

As $(\omega_l - \psi_l)$ is Lipschitz, the Lipschitzness of $\Gamma_{\psi_k}\left[\nabla_{\psi_l}\mathcal{L}(\psi_l; \mathcal{D}_{\omega_l}^N, \epsilon)\right]$ and $\Gamma_{\omega_l}\left[-(\omega_l - \psi_l)\right]$ follows from the assumption that $\nabla_{\psi_l}\mathcal{L}(\psi_l; \mathcal{D}_{\omega_l}^N, \epsilon)$ is Lipschitz provided Ω is sufficiently large to contain the entire limiting flow of the ODE: under these conditions, we expect gradients on the boundary of Φ and Ω to be directed into their interior. For this reason we are essentially encouraged to consider the non-projected form [Chung et al., 2019] in establishing Assumption 12 $i)$. As noted by Heusel et al. [2017], the condition of locally asymptotically stable ODEs in Assumption 12 $ii)$ can be ensured by an additional weight decay term in the loss function which increases the eigenvalues of the Hessian. This fits naturally in a Bayesian setting where prior regularisation introduces weight decay into our objectives. Finally, Assumption 12 $iii)$ extends the classic Robbins-Munro stepsize conditions [Robbins and Monro, 1951] to the two-timescale case. The additional assumption $\lim_{k\to\infty} \frac{\beta_k}{\alpha_k} = 0$ ensures that the faster timescale update converges asymptotically faster than the slower update..

We are now ready to provide an affirmative answer to Question 4. The details of this proof can be found in Appendix B.2 and extend the ODE method that we used for proving convergence of on-policy linear TD in Theorem 2 of Section 3.5.3 to two timescales.

4. *Do methods that sample from an approximate posterior converge?*

Theorem 6. *If Assumptions 8 to 12 hold, ψ_l and ω_l converge to $\psi_l^{\circledast}(\omega_l^{\circledast})$ and $\omega_l^{\circledast}$ almost surely.*

Proof. See Appendix B.2. □

Our analysis has proved that we can obtain convergent algorithms by introducing a separation of timescales. We observe that like Corollary 5.1, Theorem 6 also applies for the BBO MAP update that we derived in Section 7.3.1, which verifies our claim that by taking a Bayesian approach to RL affords convergent analogues of classic frequentist approaches.

7.3.4 Empirical Evaluation of Analysis

To confirm our convergence and consistency results under approximation, we evaluate BBO in several nonlinear policy evaluation experiments that are constructed to present a convergence challenge for TD algorithms. As the policy is fixed for these experiments, there is no benefit to characterising uncertainty in the MDP, hence all implementations use the MAP updates in Section 7.3.1 with a (truncated) Gaussian prior $p_\Phi(\phi) \propto \frac{1}{2\sigma_0^2}\|\phi - \phi_0\|_2^2$:

$$\psi_{k+1} = \mathcal{P}_\Phi\Big(\psi_k + \alpha_k \Big((r + \gamma \hat{Q}_{\omega_k}(s', a') - \hat{B}_{\psi_k}(s, a))\nabla_\phi \hat{B}_{\psi_k}(s, a)\Big) + \frac{\sigma^2}{\sigma_0^2}(\psi_k - \phi_0)\Big), \quad \text{(fast)} \tag{7.22}$$

$$\omega_{k+1} = \mathcal{P}_\Omega\left(\omega_k + \beta_k(\psi_k - \omega_k)\right). \quad \text{(slow)} \tag{7.23}$$

We set ϕ_0 to our initial estimate of the value function parameters which, when not fixed by the specification of the experiment, is initialised using a Glorot uniform initialisation. All practical details of our experiments can be found in Appendix D.2.

Firstly, we verify the convergence of nonlinear Gaussian BBO in the famous counterexample task of Tsitsiklis and Van Roy [1997], in which the TD(0) algorithm is provably divergent. We compare against the frequentist nonlinear GTD2 and TDC [Bhatnagar et al., 2009] which

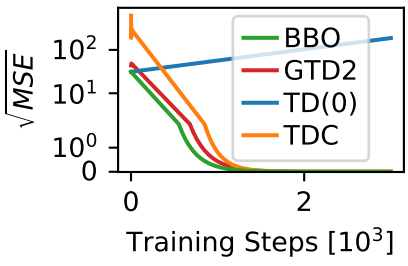


Figure 7.3: Tsitsiklis Counterexample.

extend the convergent linear GTD2 and TDC algorithms [Sutton et al., 2009b] to nonlinear function approximators by using a linear model with a basis vector that is a local linearisation of the value function. In Appendix C we derive these algorithms from the linear BBO model in the frequentist limit of $N \to \infty$. The results are presented in Fig. 7.3, where we plot the mean squared error between the true value function and the approximate value function vs training timesteps. As expected, TD(0) diverges, while BBO converges to the optimal solution slightly faster than convergent frequentist nonlinear TDC and GTD2 Bhatnagar et al. [2009] due to the additional regularisation that the prior provides.

Most interesting real-world tasks demand use of expressive function approximators such as neural networks. Despite their lack of theoretical convergence guarantees, neural networks have been successfully used in practice for estimating value functions in a wide range of recent reinforcement learning applications. Our proposed BBO algorithms provide provably convergent method for policy evaluation with runtime complexity that is linear in the dimension of the parameter space. In this experiment, we evaluate these properties empirically by applying BBO to a nonlinear regime with neural network function approximators.

We consider three policy evaluation tasks commonly used to test convergence of nonlinear TD using neural network function approximators: 20-Link Pendulum [Dann et al., 2014], Puddle World [Boyan and Moore, 1995], and Mountain Car [Boyan and Moore, 1995]. We refer to algorithms that update the parameters according to the updates in Eq. (7.22) and Eq. (7.23) as *gradient* BBO as they take the posterior's dependence of $\hat{Q}_\omega$ into account and hence minimise the MSBBE exactly. We also test a version where we ignore this dependence that we call *direct* BBO, ignoring the slow update and setting $\hat{B}_{\phi_k} = \hat{Q}_{\omega_k}$ and ignoring the slow update in Eq. (7.23). This is the update that existing model-free Bayesian RL approaches would take when they apply bootstrapping to their samples.

Our experiments are designed to: verify BBO's convergence and consistency properties in nonlinear regime, especially in cases where not all theoretical assumptions are fulfilled

exactly; investigate the effect sampling of from the BBO posterior (gradient methods) vs sampling from the posterior over Q-functions with bootstrapping (direct methods); and investigate the effect of additional regularization in BBO due to the prior.

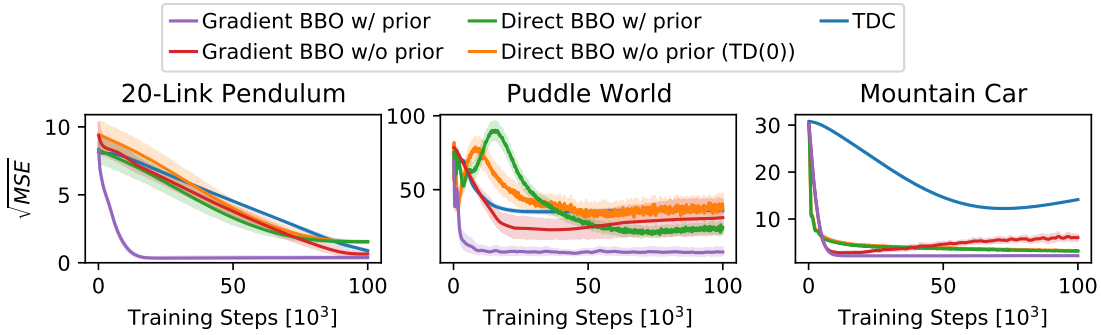


Figure 7.4: Nonlinear policy evaluation results.

The results are presented in Fig. 7.4. Gradient BBO with a prior quickly converges to a good solution in all tasks, outperforming the non-Bayesian and direct BBO variants as well as other nonlinear methods (note that direct BBO without prior corresponds to TD(0)) across all tasks. Given that both the bi-level optimization and the prior alone perform worse than Bayesian gradient BBO, the speed and quality of the solution can be attributed to their combination. Furthermore, while not a direct measure, the convergence to near-zero MSE provides empirical evidence of the algorithm's consistency. We conclude that:

i by ignoring the posterior's dependence on ω, existing model-free Bayesian approaches are less stable and perform poorly in comparison to the gradient based MSBBE minimisation approach of the BBO framework;

ii regularisation from a prior can improve performance of policy evaluation by aiding the optimisation landscape [Du et al., 2017]; and

iii better solutions in terms of mean squared error can be found using BBO instead of the local linearisation approach of nonlinear TDC/GTD2[Bhatnagar et al., 2009].

7.4 BBO for Control

A key motivation for establishing a Bayesian approach to reinforcement learning is the ability to develop algorithms where agents explore (near)-optimally when they are uncertain of the MDP. In this section, we develop Bayesian analogues of Q-learning and actor-critic frameworks for solving the control problem. In Section 7.4.1, we find that when using a nonlinear Gaussian likelihood, BBO is equivalent to the state-of-the-art BootDQN+Prior [Osband et al., 2018, 2019] algorithm. In particular, BootDQNprior+'s lagged target parameters, which are essential to its performance, arise from applying approximate inference to the BBO posterior and cannot be explained by existing model-free Bayesian RL theory which naïvely applies bootstrapping to the posterior over Q-functions. In Section 7.4.2 we extend the principles of Bayesian Q-learning to continuous domains to develop a Bayesian actor-critic algorithm that is suitable for continuous control.

For both our Bayesian Bellman Q-learning and actor-critic algorithms, we rely on posterior sampling as a tractable approach for exploration. We recall from Section 6.2.2 that the posterior sampling distribution P_π satisfies:

$$P_\pi(\arg\max_{\pi\in\Pi} J(\pi|\Phi)|\mathcal{D}_\omega^N) = \int \arg\max_{\pi\in\Pi} J(\pi|\phi) dP_\Phi(\phi|\mathcal{D}_\omega^N),$$

hence to carry out posterior sampling, we sample an MDP $\phi \in \Phi$ according to our beliefs and then follow the optimal policy for that ϕ. Whether we use a value-based or actor-critic approach to finding the optimal policy determines the class of algorithm. As we will be using RP to approximate the posterior, we can derive the RP approximate posterior sampling distribution by making the substitution $dP_\Phi(\phi|\mathcal{D}_\omega^N) = d\hat{P}_\Phi(\phi|\mathcal{D})$:

$$\begin{aligned}\int \arg\max_{\pi\in\Pi} J(\pi|\phi) d\hat{P}_\Phi(\phi|\mathcal{D}) &= \int \arg\max_{\pi\in\Pi} J(\pi|\phi_l^\star(\epsilon;\mathcal{D}_\omega^N)) dP_E(\epsilon), \\ &\approx \int \arg\max_{\pi\in\Pi} J(\pi|\phi_l^\star(\epsilon_l;\mathcal{D}_\omega^N)) d\text{Unif}(l), \end{aligned} \tag{7.24}$$

where we have used the RP ensembling approximation in deriving the final line.

7.4.1 Bayesian Bellman Q-learning

Bayesian value-based approaches to solving the control problem require characterising uncertainty in the optimal Q-function. Under the BBO framework, this is attained by inferring the posterior over the optimal Bellman operator: $P_{B^\star}(\cdot|\mathcal{D}_\omega^N)$. Recall from Section 7.1 that $P_{B^\star}(\cdot|\mathcal{D}_\omega^N)$ is derived by pushing the posterior over the state-transition distribution and reward function through the optimal empirical Bellman equation:

$$b_\omega^\star(r, s') \coloneqq r + \gamma \max_{a' \in \mathcal{A}} \hat{Q}_\omega(s', a'),$$

which can be modelled exactly as outlined in Section 7.2 using a dataset of empirical Bellman samples: $\mathcal{D}_\omega^N = \{b_i^\star, a_i, s_i\}_{i \in 1:N}$ where $b_i^\star \coloneqq r_i + \gamma \max_{a' \in \mathcal{A}} \hat{Q}_\omega(s_i', a')$. We introduce the simplified notation for the Bayesian optimal Bellman operator:

$$\mathcal{B}_{\omega,N}^\star(s, a) \coloneqq \mathcal{B}^\star[\hat{Q}_\omega](s, a, \mathcal{D}_\omega^N) = \int b^\star dP_{B^\star}(b^\star|\mathcal{D}_\omega^N),$$

and learn the optimal Q-function approximator by minimising the MSBBE:

$$\text{MSBBE}(\omega) = \|\mathcal{B}_{\omega,N}^\star - \hat{Q}_\omega\|_\varsigma^2. \tag{7.25}$$

It follows from the definition of the pushforward posterior $P_{B^\star}(\cdot|\mathcal{D}_\omega^N)$ in Eq. (7.2) that any function approximator satisfying $\hat{Q}_{\omega^\star} = \mathcal{B}_{\omega^\star,N}^\star$ will be a Bayesian optimal Q-function for the BAMDP that the BBO model can represent. Any greedy policy $\pi^\star(\cdot|s) = \delta_{a_N^\star(s)}$ where $a_N^\star(s) \in \arg\max_{a \in \mathcal{A}} \mathcal{B}_{\omega^\star,N}^\star(s, a)$ will therefore be a Bayes-optimal policy.

For many models, it may be intractable to carry out the marginalisation required to solve the Bayesian optimal Bellman equation $\hat{Q}_{\omega^\star} = \mathcal{B}_{\omega^\star,N}^\star$ exactly. Instead, we can combine approximate inference with posterior sampling to develop a Bayesian Q-learning style algorithm. To derive BootDQNprior+ [Osband et al., 2018, 2019] from BBO, we chose a Gaussian likelihood and prior model with densities:

$$p(b^\star|s, a, \phi) \propto \exp\left(-\frac{\left(b^\star - \hat{B}_\phi^\star(s, a)\right)^2}{2\sigma^2}\right), \quad p_\Phi(\phi) \propto \exp\left(-\frac{\|\phi - \phi_0\|_2^2}{2\sigma_0^2}\right), \tag{7.26}$$

and use randomised priors to approximate the posterior with a Gaussian noise distribution $P_E = \mathcal{N}(0, I)$ to match the prior. Following the derivation of ensembled randomised

priors BBO in Section 7.3.2, L prior randomisations $\mathcal{E}_L \coloneqq \{\epsilon_l\}_{l=1:L}$ are first drawn from P_E. We also maintain an ensemble of Bellman operator parameters $\Psi_L \coloneqq \{\psi_l\}_{1:L}$ and an ensemble of target parameters $\Omega_L \coloneqq \{\omega_l\}_{1:L}$. The approximate posterior is updated by carrying out a two-timescale gradient update on the RP objectives with the ensemble parameters for each $l \in \{1 : L\}$:

$$\psi_{l,k+1} = \mathcal{P}_\Omega\Big(\psi_{l,k} - \alpha_k\Big(\Big(\hat{B}^\star_{\psi_{l,k}}(s,a) - (r + \gamma \max_{a' \in \mathcal{A}} \hat{Q}_{\omega_{l,k}}(s',a'))\Big) \nabla_\phi \hat{B}^\star_{\psi_{l,k}}(s,a) + \frac{\sigma^2}{\sigma_0^2}(\psi_{l,k} - \epsilon_l - \phi_0)\Big)\Big), \quad \text{(fast)} \tag{7.27}$$

$$\omega_{l,k+1} = \mathcal{P}_\Omega(\omega_{l,k} - \beta_k(\omega_{l,k} - \psi_{l,k})), \quad \text{(slow)} \tag{7.28}$$

where (r, s', a, s) are drawn uniformly from the dataset. The updates in Eq. (7.28) and Eq. (7.27) were derived by substituting for our Gaussian model from Eq. (7.34) into the RP updates in Eq. (7.18) and Eq. (7.19). The fast Gaussian Bayesian Q-learning update in Eq. (7.27) is equivalent to the frequentist DQN update [Mnih et al., 2015] with an additional prior-randomised regularisation term. The slow update plays the same role as the target Q-function updates. We can therefore understand the role of updating the posterior in the Bayesian Q-learning algorithm as carrying out Bayesian policy evaluation.

If we try to derive a Bayesian Q-learning algorithm using the posterior over optimal Q-functions: $P_{Q^\star}(\cdot|\mathcal{D}^N)$, the solution to minimising the MSBBE in Eq. (7.25) is to set $\hat{Q}_\omega = \mathcal{B}^\star_N$. When applying approximate inference, this is equivalent to making the substitution $\hat{Q}_{\omega_{l,k}} = \hat{B}^\star_{\psi_{l,k}}$ in Eq. (7.27) and ignoring the slow update, which is trivially $\omega_{l,k+1} = \omega_{l,k} = \psi_{l,k}$ for all k. Hence, by aplying bootstrapping to the theory of existing model-based Bayesian RL, we would thus derive the direct BBO update for $\psi_{l,k+1}$

$$\psi_{l,k+1} = \mathcal{P}_\Omega\Big(\psi_{l,k} - \alpha_k\Big(\Big(\hat{B}^\star_{\psi_{l,k}}(s,a) - (r + \gamma \max_{a' \in \mathcal{A}} \hat{Q}_{\psi_{l,k}}(s',a'))\Big) \nabla_\phi \hat{B}^\star_{\psi_{l,k}}(s,a) + \frac{\sigma^2}{\sigma_0^2}(\psi_{l,k} - \epsilon_l - \phi_0)\Big)\Big), \quad \text{(Direct)} \tag{7.29}$$

which does not predict the existence of target parameters. Theoretically, this stems from the fact that existing approaches replace noisy samples from an optimal Q-function with

a bootstrapped approximation using $\hat{Q}_\omega$, without accounting for this dependence in the posterior. On a practical level, the direct update in Eq. (7.29) does not optimise any objective and is a regularised TD update, hence we cannot guarantee the convergence of $\lim_{k\to\infty} \psi_{l,k+1}$ for the reasons that we discussed in Section 3.5.3.

Algorithm 3 RP-BBQ

Initialise Ω_L, Ψ_L and $\mathcal{D} \leftarrow \varnothing$
Sample prior randomisations $\mathcal{E}_L \sim P_E$
Sample initial state $s \sim P_0$
while not converged **do**
  Sample ensemble $l \sim$ Unif
  for $n \in \{1, ...N_{\text{env}}\}$ **do**
    Take greedy action $a \in \arg\max_{a' \in \mathcal{A}} \hat{B}^\star_{\psi_l}(a', s)$
    Observe next state $s' \sim P(\cdot|s, a)$
    Observe reward $r = r(s', a, s)$
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{s, a, r, s'\}$
  end for
  $\Omega_L, \Psi_L \leftarrow$ UPDATERPPOSTERIOR$(\Omega_L, \Psi_L, \mathcal{E}_L, \mathcal{D})$
end while

To equip our agent with a posterior sampling exploration regime, we must derive the optimal policy for a given parametrisation drawn from the posterior $\phi \sim P_\Phi(\cdot|\mathcal{D}^N_{\omega^\star})$ where $\hat{Q}_{\omega^\star} = \mathcal{B}^\star_{\omega^\star,N}$, we can write the optimal Bellman operator as:

$$\int r + \gamma \max_{a' \in \mathcal{A}} \hat{Q}_{\omega^\star}(s', a') dP(r, s'|s, a, \phi) = \int b^\star dP(b^\star|s, a, \phi) = \hat{B}^\star_\phi(s, a),$$

hence an optimal policy can be derived as $\pi^\star_\phi(\cdot|s) = \delta_{a^\star_N(s,\phi)}$ where

$$a^\star_N(s, \phi) \in \arg\max_{a \in \mathcal{A}} \hat{B}^\star_\phi(s, a).$$

When applying randomised priors, we sample from the RP posterior at the start of each episode by choosing an ensemble $l \sim$ Unif. In accordance with Eq. (7.24), a posterior sampling policy can then be constructed for the sampled MDP by selecting a greedy action $a \in a^\star_N(s, \psi_l) = \arg\max_{a \in \mathcal{A}} \hat{B}^\star_{\psi_l}(s, a)$. We outline this procedure in Algorithm 3 which summarises our randomised priors Bayesian Bellman Q-learning (RP-BBQ) algorithm. When using the updates from Eq. (7.27) and Eq. (7.28) under the Gaussian model, Algorithm 3 is exactly the BootDQNprior+ algorithm.

7.4.2 Bayesian Bellman Actor-critic

Due its use of the $\arg\max$ operator, RP-BBQ (and hence BootDQNprior+) is not appropriate for continuous action or large discrete action domains as a nonlinear optimisation problem must be solved every time an action is sampled. To extend RP-BBQ to continuous domain we develop a randomised priors Bayesian Bellman actor-critic (RP-BBAC).

Posterior sampling exploration is at the heart of our control algorithms, for which we need to sample from the (approximate) posterior sampling distribution. This requires calculating the optimal policy for any sample ϕ from the posterior $\pi^\star_\phi \in \arg\max_{\pi\in\Pi} J(\pi|\phi)$. In an actor-critic approach, we introduce a parametrised policy π_θ and write the RL objective for ϕ as:

$$J(\theta|\phi) = \int\int Q^{\pi_\theta}(a, s, \phi) d\pi_\theta(a|s) dP_0(s). \tag{7.30}$$

Finding the gradient of Eq. (7.30) with respect to θ is the same a deriving the policy gradient from Eq. (3.3):

$$\nabla_\theta J(\theta|\phi) = \sum_{t=0}^{\infty} \gamma^t \int\int Q^{\pi_\theta}(s, a, \phi) \nabla_\theta \log \pi_\theta(a|s) d\pi_\theta(a|s) d\rho_t^{\pi_\theta}(s|\phi), \tag{7.31}$$

For a given parametrisation drawn from the posterior $\phi \sim P_\Phi(\cdot|\mathcal{D}^N_{\omega^\star})$ where $\hat{Q}_{\omega^\star} = \mathcal{B}_{\omega^\star,N}$, we can write Q-function in terms of the Bellman operator as:

$$Q^{\pi_\theta}(s, a, \phi) = \int r + \gamma \hat{Q}_{\omega^\star}(s', a') dP^{\pi_\theta}(r, s', a'|s, a, \phi) = \int b dP(b|s, a, \phi) = \hat{B}_\phi(s, a).$$

Substituting $Q^{\pi_\theta}(s, a, \phi) = \hat{B}_\phi(s, a)$ into Eq. (7.31) yields:

$$\nabla_\theta J(\theta|\phi) = \sum_{t=0}^{\infty} \gamma^t \int\int \hat{B}_\phi(s, a) \nabla_\theta \log \pi_\theta(a|s) d\pi_\theta(s|s) d\rho_t^{\pi_\theta}(s|\phi).$$

Just like in classic actor-critic algorithms, following the true policy gradient may be impractical because of the discount factor and the requirement to sample from the underlying ergodic state distribution $\rho_t^{\pi_\theta}(\cdot|\phi)$. In our Bayesian setting, we can derive an alternative. A policy $\pi^\star_\phi \in \arg\max_{\pi\in\Pi} J(\pi|\phi)$ can be found by optimising:

$$J(\pi|\phi) = \int\int Q^{\pi_\theta}(s, a, \phi) d\pi(a|s) d\rho(s)$$

for any sampling distribution ρ with support over $\mathcal{S}$. Substituting for $Q^{\pi_\theta}(s, a, \phi) = \hat{B}_\phi(s, a)$ yields:

$$J(\pi|\phi) = \int\int \hat{B}_\phi(s, a) d\pi(a|s) d\rho(s) = \mathbb{E}_{\rho\pi}\left[\hat{B}_\phi\right].$$

When taking a RP approach to approximate inference, we see that sampling from the approximate posterior sampling distribution in Eq. (7.24) requires finding

$$\arg\max_{\pi\in\Pi} J(\pi|\phi^\star(\epsilon; \mathcal{D}^N_{\omega^\star})) = \arg\max_{\pi\in\Pi} \mathbb{E}_{\rho\pi}\left[\hat{B}_{\phi^\star(\epsilon;\mathcal{D}^N_{\omega^\star})}\right]. \tag{7.32}$$

To solve the bi-level approximate inference problem, recall that we introduced a set of parameters $\psi \in \Phi$ to track $\phi^\star(\epsilon; \mathcal{D}^N_\omega)$. Solving Eq. (7.32) under RP approximate inference leads to the nested optimisation problem:

$$\begin{aligned}
\pi^\star &\in \arg\max_{\pi\in\Pi} J(\pi; \psi^\star), && \textbf{(Actor)}\\
\omega^\star &\in \arg\min_{\omega\in\Omega} \mathcal{U}(\omega; \psi^\star), && \textbf{(Target Critic)}\\
\psi^\star &\in \arg\min_{\psi\in\Psi} \mathcal{L}(\psi; \mathcal{D}^N_\omega). && \textbf{(Critic)}
\end{aligned}$$

We have labelled each objective to highlight the similarity between this system of objectives and the well-known family of actor-critic algorithms [Konda and Tsitsiklis, 2000]: $J(\pi; \psi^\star)$ is an *actor* objective to learn an optimal exploration policy. Observe that when introducing a parametrised policy π_θ, the actor gradient is:

$$\nabla_\theta J(\theta; \psi^\star) = \mathbb{E}_{\rho\pi}\left[\nabla_\theta \log \pi_\theta(a|s) \hat{B}_\phi\right], \tag{7.33}$$

which only requires samples from ρ, thus avoiding the issues associated with sampling from the classic policy gradient. Formally, this arises because of the separation of timescale introduced: as $\psi^\star$ tracks $\phi^\star(\epsilon; \mathcal{D}^N_\omega)$, it accounts for the dependence of $\phi^\star(\epsilon; \mathcal{D}^N_\omega)$ on the policy π. Moreover, as $\psi^\star$ is updated on a faster timescale, it can be seen as quasi-equilbriated from the perspective of the actor objective, and so does not require a gradient. The approximate inference parameters $\psi^\star$ share the same role as the *critic* parameters and the auxiliary objective parameters $\omega^\star$ can be interpreted as *target critic* parameters which lag the critic parameters.

To account for ensembling, we introduce an ensemble of parametric policies $\pi_{\theta_l}(a|s)$ parametrised by a set of parameters $\Theta_L \coloneqq \{\theta_l\}_{l=1:L}$. A schemactic of the ensembled actor-critic is shown in Fig. 7.5. For each ensemble $l \in \{1 : L\}$, the optimal exploration policy $\pi_{\theta_l^\star}$ is parametrised by the solution to its own optimisation problem: $\theta_l^\star \in \arg\max_{\theta\in\Theta} J(\theta|\psi_l)$, and can be interpreted as a prior-randomised actor. Every ensemble $l \in \{1...L\}$ also has its own critic B_{ψ_l} and target .

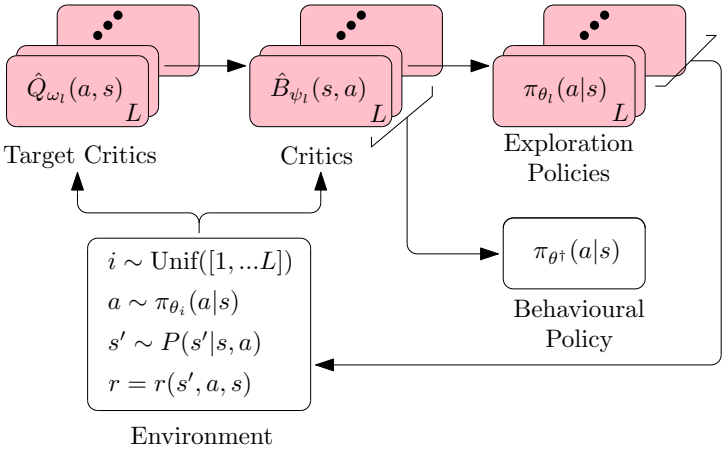


Figure 7.5: Schematic of RP-BBAC.

We can compare our actor-critic to the RP-BBQ and existing frequentist actor-critic algorithms by chosing a Gaussian likelihood and prior model with densities:

$$p(b|s,a,\phi) \propto \exp\left(-\frac{\left(b-\hat{B}_\phi(s,a)\right)^2}{2\sigma^2}\right), \quad p_\Phi(\phi) \propto \exp\left(-\frac{\|\phi-\phi_0\|_2^2}{2\sigma_0^2}\right), \tag{7.34}$$

and use randomised priors to approximate the posterior with a Gaussian noise distribution $P_E = \mathcal{N}(0, I)$ to match the prior. Plugging these densities into our derivation of the RP updates from Section 7.3.2 yields the Gaussian RP-BBAC critic and target critic updates:

$$\psi_{l,k+1} = \mathcal{P}_\Omega\Big(\psi_{l,k} - \alpha_k\Big(\Big(\hat{B}_{\psi_{l,k}}(s,a) - (r+\gamma\hat{Q}_{\omega_{l,k}}(s',a'))\Big)\nabla_\phi\hat{B}_{\psi_{l,k}}(s,a) + \frac{\sigma^2}{\sigma_0^2}(\psi_{l,k}-\epsilon_l-\phi_0)\Big)\Big), \quad \textbf{(Critic)} \tag{7.35}$$

$$\omega_{l,k+1} = \mathcal{P}_\Omega(\omega_{l,k} - \beta_k(\omega_{l,k}-\psi_{l,k})), \quad \textbf{(Target Critic)} \tag{7.36}$$

$$\theta_{l,k+1} = \mathcal{P}_\Theta(\theta_{l,k} + \zeta_k\nabla_\theta \log\pi_{\theta_{l,k}}(a|s)\hat{B}_{\phi_{l,k}}(s,a)), \quad \textbf{(Actor)} \tag{7.37}$$

where (s', a, s, r) are drawn from the dataset and $a' \sim \pi_{\theta_l}(\cdot|s')$.

In comparison to existing critic objectives, we can interpret the first term of Eq. (7.35) as the standard TD update using a target critic. The second term introduces regularisation from the prior and injects noise into the objective via $\epsilon_l \sim P_E(\cdot)$, allowing us to approximate the variance in the posterior. The target critic is updated on slower timescale Eq. (7.36), mimicking the updating of target critic parameters after a regular interval in frequentist approaches [Haarnoja et al., 2018a]. In order to maintain the separation of timescales to justify our actor objective gradient in Eq. (7.33), we require that the actor is updated on an asymptotically slower timescale than the critic, that is $\lim_{k\to\infty} \zeta_k/\alpha_k = 0$. Convergence of the update in Eq. (7.37) can be proved in a similar way to the convergence of RP approximate policy evaluation in Section 7.3.3.

Unlike frequentist approaches, an exploratory actor is selected at the start of each episode in accordance with our current uncertainty in the MDP characterised by the approximate RP posterior. Finally, exploration policies may not perform well at test time, so we learn a behaviour policy $\pi_{\theta^\dagger}(a|s)$ parametrised by $\theta^\dagger \in \Theta$ from the data collected by our exploration policies using the ensemble of critics: $\{\hat{B}_{\psi_l}\}_{l=1:L}$. From a Bayesian perspective, we can marginalise over all MDPs in our ensemble to find an approximation to the Bayesian Bellman operator, and then find a policy that is maximal . In this thesis we augment this Bayesian behaviour policy objective with regularisation:

$$J_{\text{Behavioural}}(\pi|\Psi_L) = \mathbb{E}_{\rho\pi}\left[\frac{1}{L}\sum_{l=1}^{L}\hat{B}_{\psi_l} + \alpha\mathcal{H}(\pi)\right],$$

allowing us to combine the exploratory benefits of posterior sampling with the faster convergence rates and algorithmic stability of regularised RL [Vieillard et al., 2020], however the choice of behavioural policy objective can be determined by the practitioner to suit the problem. We summarise the RP-BBAC algorithm in Algorithm 4.

Algorithm 4 RP-BBAC

Initialise $\Theta_L, \Omega_L, \Psi_L, \mathcal{E}_L, \theta^\dagger$ and $\mathcal{D} \leftarrow \varnothing$
Sample prior randomisations $\mathcal{E}_L \sim P_E$
Sample initial state $s \sim P_0$
while not converged **do**
  Sample policy $\theta_l \sim \text{Unif}(\Theta_L)$
  for $n \in \{1, ...N_{\text{env}}\}$ **do**
    Sample action $a \sim \pi_{\theta_l}(\cdot|s)$
    Observe next state $s' \sim P(\cdot|s, a)$
    Observe reward $r = r(s', a, s)$
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{s, a, r, s'\}$
  end for
  $\Theta_L, \Omega_L, \Psi_L \leftarrow \text{UPDATEPOSTERIOR}(\Theta_L, \Omega_L, \Psi_L, \mathcal{E}_L, \mathcal{D})$
  $\theta^\dagger \leftarrow \text{UPDATEBEHAVIOURALPOLICY}(\Psi_L, \mathcal{D})$
end while

For the same reasons as discussed in Section 7.4.1, RP-BBAC's explicit specification of lagged target critics is unique to BBO and corrects the theoretical issues raised by applying bootstrapping to existing model-free Bayesian RL theory, which does not account for the posterior's dependence on $\hat{Q}_\omega$. In reference to Section 3.4, RP-BBAC's exploration is both deep and adaptive as actions from an exploration policy are directed towards minimising epistemic uncertainty in the MDP and the posterior variance reduces in accordance with Corollary 5.1 as more data is collected.

7.4.3 Continuous Control Experiments

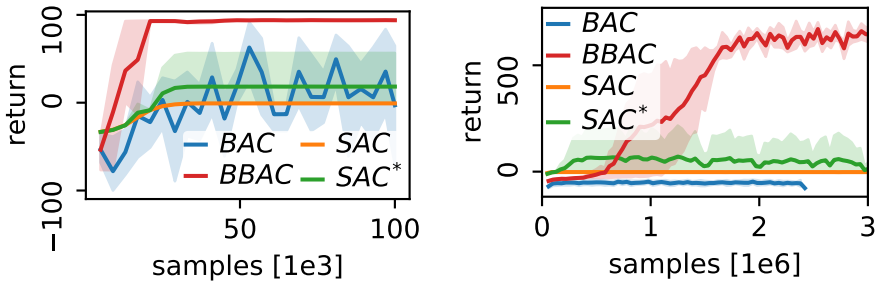
We now evaluate RP-BBAC in a continuous control task to showcase the benefits of BBO. The goals of our empirical evaluation are to test the performance of algorithms derived from the BBO framework in continuous control tasks where success requires solving a challenging deep exploration problem and to investigate the empirical implications of ignoring the theoretical issues raised by applying bootstrapping to existing model-free Bayesian RL. All practical details of our experiments can be found in Appendix D.3.

In many benchmark tasks for continuous RL, such as the locomotion tasks from MuJoCo Gym suite [Brockman et al., 2016], the environment reward is shaped to provide a smooth gradient towards a successful task completion and naïve Boltzmann-type dithering

exploration strategies from control as inference in Chapter 5 can provide a strong inductive bias. In practical real-world scenarios, dense rewards are difficult to specify by hand, especially when the task is learned from raw observations like images. To demonstrate the exploration benefits of RP-BBAC, we focus on a set of continuous analogues of the discrete experiments used to test BootDQN+Prior [Osband et al., 2018]: *MountainCar-Continuous-v0* from Gym benchmark suite and a slightly modified version of the *cartpole-swingup_sparse* from DeepMind Control Suite [Tassa et al., 2018]. Both environments have a sparse reward signal and penalize the agent proportional to the magnitude of executed actions, details of which can be found in Appendix D.3. As the agent is always initialised in the same state, it has to deeply explore costly states in a directed manner for hundreds of steps until it reaches the rewarding region of the state space. Our experiments are carried out in a tabula rasa setting where the prior parameters are drawn as randomly initialized neural networks, which controls for any advantage that may arise from a fortuitous prior parameter initialisation.

In our experiments, we compare RP-BBAC against two variants of the state-of-the-art soft actor-critic: SAC, which is the exact algorithm presented in [Haarnoja et al., 2018b]; and SAC*, a tailored version which uses a single Q-function to avoid *pessimistic underexploration* [Ciosek et al., 2019b] due to the use of the double-minimum-Q trick (see Appendix D.3 for details). To understand the practical implications of our theoretical results, we also compare against a Bayesian actor-critic (BAC) which is a variant of RP-BBAC where $\hat{Q}_{\omega_l^\star} = \hat{B}_{\psi_l^\star}$, thereby ignoring the target critic updates in Eq. (7.36). As we discussed in Section 7.4.2, BAC is the Bayesian actor-critic that results from applying RP approximate inference to the posterior over Q-functions used by existing model-free Bayesian approaches with bootstrapping, so its choice of a benchmark allows us to evaluate the empirical benefits of using BBO vs the existing model-free Bayesian RL framework.

The results are shown in Fig. 7.6. Due to the lack of smooth signal towards the task completion, SAC consistently fails to solve the tasks and converges to always executing the 0-action due to the action cost term, while SAC* achieves the goal in one out of five seeds. The performance of BAC is unstable and the algorithm fails to converge to



(a) MountainCar **(b)** Cartpole

Figure 7.6: Continuous control with sparse reward.

an optimal policy. RP-BBAC succeeds for all five seeds in both tasks. To probe the reasons behind the poor performance of BAC and SAC/SAC*, we carry out a state support analysis for *MountainCar-Continuous-v0*.

The analysis in Fig. 7.7 reveals the deep, adaptive nature of BBO's exploration, which leads agents to systematically explore regions of the state-action space with high uncertainty. In the beginning, even when the actions are costly and no positive rewards are encountered, there still exists ensemble members whose value are optimistic under uncertainty: at $T = 9000$, the agent has not yet discovered any reward from the environment, but the disagreement of ensemble members drives the algorithm to deeply explore the state space to reflect the uncertainty that agent has over undiscovered states that could yield high returns. At around $T = 14000$, the agent achieves the goal for the first time and the value functions start to shape towards optimal solution, demonstrating that the posterior concentrates in agreement with Theorem 5 thereby curtailing further exploration and preventing agent revisiting sub-optimal states.

The same analysis in Fig. 7.8 for SAC confirms the inefficiency of the exploration typical of control as inference: the agent eschews costly actions that would ultimately lead to rewarding states, repeatedly explores actions that lead to poor performance and rarely exploring beyond its initial state. This can also be seen in the value functions, which are prematurely driven to sub-optimal solutions. In comparison to RP-BBAC, SAC's exploration is myopic and is not adaptive, with the agent failing to curtail exploration towards the (sub-optimal) solution that its value functions have converged to,

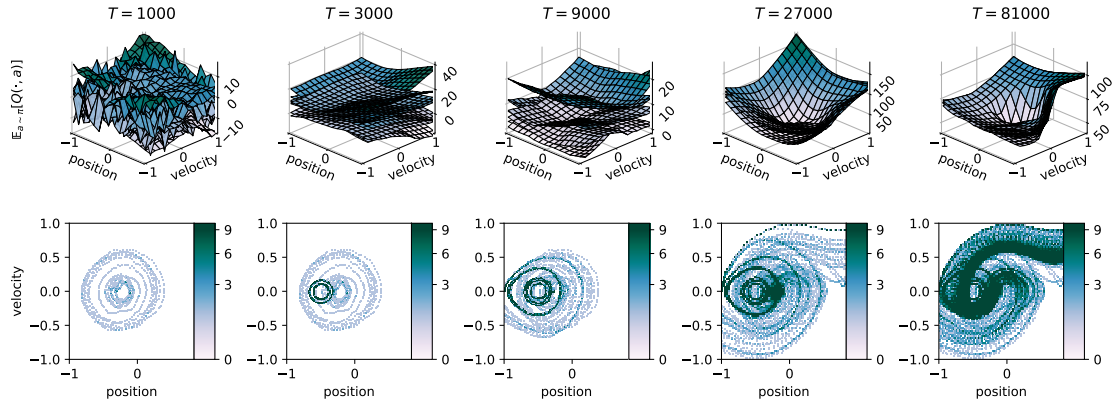


Figure 7.7: Diagnostics throughout learning of BBAC. (Top) Expected Q-values for each ensemble member ($L = 8$) over environment states $(position, velocity)$. (Bottom) State visitation.

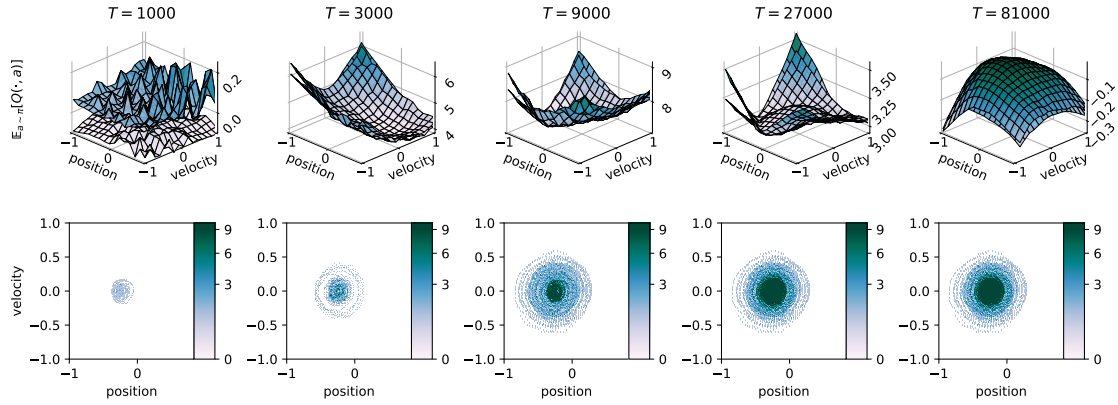


Figure 7.8: Diagnostics throughout learning of SAC. (Top) Ensemble values over environment state $(position, velocity)$. (Bottom) State visitation.

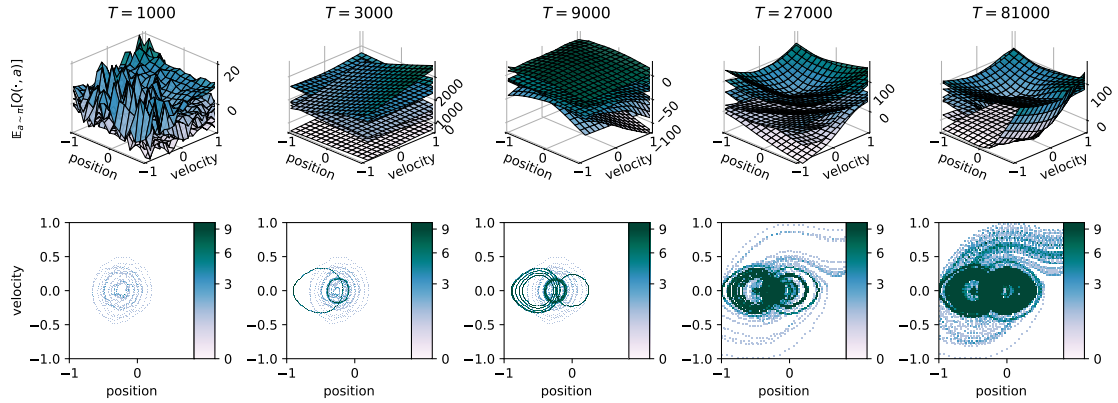


Figure 7.9: Diagnostics throughout learning of BAC. (Top) Expected Q-values for each ensemble member ($L = 8$) over environment states $(position, velocity)$. (Bottom) State visitation.

repeatedly taking actions that, even by its own incorrect belief in the MDP, knows are sub-optimal.

The analysis for BAC in Fig. 7.9 confirms that BAC initially explores similarly to BBAC: at $T = 9000$, the agent has not yet discovered any reward from the environment and the disagreement of ensemble members drives the algorithm to explore the state space similarly as with BBAC. As time progress, due to the convergence issues that stem from ignoring the posterior's dependence on ω, the ensembles never concentrate with increasing number of samples. This means that the approximate posterior cannot characterise the epistemic uncertainty well and the residual stochasticity continues to drive exploration when the agent should have learnt to ignore actions that can't lead to increased returns. In comparison to SAC, the initial uncertainty estimates enable the BAC agent to reach the goal but, because the Q-function learning is thus relatively unstable without target networks, spends more time re-exploring around the origin than BBAC.

In Appendix D.3.2, we also investigate RP-BBAC's sensitivity to randomized prior hyperparameters. The range of working hyperparameters is wide and easy to tune. We conclude from our experiments that:

i due to its sophisticated exploration, algorithms derived from the BBO framework can solve continuous control tasks at which state-of-the-art regularised actor-critic algorithms fail catastrophically,

ii the theoretical issues associated with applying bootstrapping to existing model-free Bayesian approaches prevent an agent accurately quantifying uncertainty which has negative empirical impact on performance. This verifies that it is essential for Bayesian model-free RL algorithms with bootstrapping to sample from the BBO posterior.

7.5 Insights for Future Research

We conclude by highlighting an issue with existing model-free Bayesian RL algorithms and sketching a direction for addressing it in future research.

From our exposition of the likelihood principle in Section 6.3, if a random variable $X(\cdot|\pi)$ is sufficient for evaluating the RL objective $J(\pi|\phi)$ and is consistent with the variables $\mathcal{P}$ and r representing the dynamics of the MDP, then the chain of equivalences holds:

$$J_{\text{Bayes}}(\pi|D)$$
$$= \int J(\pi|x)dP_X(x|\mathcal{D},\pi) = \int J(\pi|r,\mathcal{P})dP_{r,\mathcal{P}}(r,\mathcal{P}|\mathcal{D}) = \int J(\pi|\phi)dP_\Phi(\phi|\mathcal{D}).$$

The likelihood principle thus justifies model-free Bayesian RL approaches as there is no difference between inferring the posterior $P_X(\cdot|\mathcal{D},\pi)$ and $P_{r,\mathcal{P}}(\cdot|\mathcal{D})$ in terms of being able to evaluate the Bayesian RL objective. In BBO the random variable is the empirical Bellman equation.

When we derive the posterior for X we start by defining a model $P_X(\cdot|\phi)$ of the process. By introducing this model, we have implicitly made assumptions about the dynamics of environment. To see this, observe that in RL, we obtain samples of x using the transformation of variables $x = X(r,\mathcal{P},\pi)$ starting from samples from the reward function and transition distribution as the agent interacts with the environment. Under the appropriate assumptions of measurability, there thus exists a pullback measure $P_{r,\mathcal{P}}(\cdot|\phi)$ of $P_X(\cdot|\phi)$ through $X(r,\mathcal{P},\pi)$ which, for any integrable function f, satisfies:

$$\int f(x)dP_X(x|\phi,\pi) = \int f \circ X(r,\mathcal{P},\pi)dP_{r,\mathcal{P}}(r,\mathcal{P}|\phi).$$

Our choice of $P_X(\cdot|\phi)$ implicitly defines a space of reward functions and transition dynamics that our model represents, and hence the more expressive our density $p_X(x|\phi)$ is the more expressive the space of MDPs we are reasoning over when carrying out inference. Despite this, as we discussed in Section 6.2.3, all established parametric model-based Bayesian RL assumes a Gaussian distribution $P_X(\cdot|\phi) = \mathbb{N}(\mu_\phi,\sigma_\phi)$. An analysis by Janz et al. [2019] reveals that the pullback of a Gaussian model when $X(r,\mathcal{P},\pi) = Q^\pi$ can't even represent a space of MDPs with known deterministic transition distributions and a Gaussian reward model.

If we are to develop tractable methods for approximating Bayes-optimal policies by exploiting the likelihood principle using the BBO framework, it is essential that we move beyond Gaussian models and research alternative approaches for specifying model $P_B(\cdot|s,a,\phi)$ that are expressive enough to represent a diverse space of MDPs.

8

Conclusion: A Bayesian Manifesto

In this thesis, we have discussed some of the pathologies that plague researchers who develop algorithms to optimise the frequentist RL objective when an agent in uncertain about its environment. The double sampling problem [Baird, 1995], TD algorithms that are provably divergent [Tsitsiklis and Van Roy, 1997], overestimation of Q-functions [Hasselt et al., 2016, Fujimoto et al., 2018a] and having no formal solution to the exploration/exploitation dilemma [Sutton and Barto, 2018] are all symptoms of this approach. Many of the issues with frequentist RL stem from the fact that optimising the RL objective does not adhere to the likelihood principle as it assumes prior knowledge that is not available to the agent.

Under the frequentist paradigm, research is often directed towards developing heuristics that make up for the gap between the assumption of complete knowledge and the reality of an agent's situation. For the vast majority of cases, research has progressed by discovering heuristics that work well in a limited set of similar MDPs: as we have shown in Chapter 3, state-of-the-art algorithms derived from MERL are successful in MuJoCo continuous control tasks, but generate an incorrect optimal policy in our counterexample discrete MDP of Section 5.1.3 and can fail catastrophically in the simple MountainCar domain of Section 7.4.3 where exploration is not trivial.

We can connect the Bayesian and frequentist RL objectives by recognising that the Bayesian RL objective contains the frequentist RL objective as a special case corresponding to a Dirac-delta prior corresponding to the belief that agent knows the true MDP. The Bayesian RL objective is thus appropriate for both learning and planning whereas the frequentist objective is only appropriate for planning. Moreover, in our thesis, we demonstrated that solutions to the Bayesian RL objective are admissible Section 6.1.1, optimal under a prior over the environment Section 6.1.2 and that it doesn't matter which random variable we characterise uncertainty in provided it is sufficient to solving the RL problem Section 6.3. This means that, given a prior over MDPs, a Bayesian approach cannot be bettered by another approach over that space of MDPs. From this perspective, if a Bayes-optimal solution is always tractable, there is little rationality behind ignoring it. Of course this is rarely the case in practice, however tractable posterior approximations are well established [Jordan, 1999, Beal, 2003, Kingma and Welling, 2014, Rezende and Mohamed, 2015, Gal, 2016, Lakshminarayanan et al., 2017] and continue to be a major focus for research in the Bayesian machine learning community.

In light of this, instead of using the frequentist RL objective as a foundation and then developing heuristics to make up for the incorrect assumption of complete knowledge, we propose starting research from the Bayesian objective instead and then developing approximations to make solutions tractable. To support this proposal, we recognise that Bayesian approaches are intractable due to their use of a posterior, however even under approximation they still have provable benefits. We now outline these three benefits with representative examples.

Bayesian methods always satisfy the conditionality principle. As we demonstrated in Section 6.3.1, even when using a tractable estimate for the posterior, Bayesian approaches still adhere to the conditionality principle. This is because any Bayesian approximate method can only be derived from the posterior, which only conditions on observations. In contrast, a frequentist RL approach uses a point estimate as a prior, corresponding to the belief that the agent knows the true MDP. This affects the inference

procedure as the prior and posterior are the same, hence any quantity derived from the framework, even under approximation, may not condition on known observations.

As an example of this benefit, one of the greatest milestones in deep RL research was the development of the replay buffer and target parameters to stabilise TD learning [Mnih et al., 2015], which have become ubiquitous in model-free algorithms [Abdolmaleki et al., 2018b, Haarnoja et al., 2017, 2018a, 2019, Osband et al., 2016, 2019]. Remarkably, using a replay buffer and target parameters was not known to induce formal converge properties until quite recently [Yang et al., 2019]. The motivation for their introduction was heuristic [Mnih et al., 2015], using a biological metaphor in an attempt to resolve the pathologies present in the frequentist RL objective: namely the double sampling problem and lack of convergence guarantees for nonlinear TD methods. In contrast, we have shown in Chapter 7 that both replay buffers and target parameters arise naturally from BBO under point estimation. This is because under a Bayesian framework we never have to encounter any of the frequentist pathologies that arise from breaking the conditionality principle.

Frequentist heuristics can be harmful. In our attempt to resolve some of the pathologies associated with starting from the frequentist objective, the methods that we introduce can inadvertently introduce unforeseen pathologies. A frequentist epistemology can therefore lead to an unsatisfactory research paradigm whereby one heuristic works well for a subset of MDPs, but has unexpected consequences when introduced to another set of MDPs. In the RL literature, this is a common manifestation: even within material of this thesis we've seen in Chapter 5 that the MERL framework can resolve some of the pathologies of existing approaches in tackling the exploration/exploitation dilemma but in doing so its solution introduces surreptitious pathologies that don't manifest in the domains that they are evaluated in. The frequentist epistemology has the potential to introduce a false progress narrative whereby we embark on a chain of heuristics where each new method resolves the pathologies of existing methods whilst introducing new pathologies for future research to tackle.

Bayesian analysis is always possible. In contrast to frequentist heuristics, under a Bayesian epistemology, if our approximations introduce pathologies there is always a perfect solution that we can actually compare to that is free of issues. This stems from the conditionality principle, meaning that all Bayesian quantities can be evaluated under our current knowledge because we introduce a prior. As an example, Bayesian regret bounds in RL are easy to derive and analyse as, given a prior, they can be evaluated [Osband et al., 2013]. In contrast, the assumption of knowledge of the exact solution a priori means that true frequentist regret bounds can't be analysed outside of contrived and trivial domains. A common solution to this problem is to use a minimax analysis whereby the lowest regret that reported over a space of MDPs, but again this makes prior assumptions about the space of MDPs that the minimax analysis is carried over! [Murphy, 2012]. Even when the Bayesian solution is intractable, the Bayesian approach can still be used to analyse and reason over how our algorithms and approximations will perform.

As another example of this benefit, recall from we Section 6.1.1 that under a Bayesian analysis, policy gradient algorithms that maximise the frequentist RL objective using Monte Carlo samples specify a Dirichlet process prior. Under a Bayesian epistemology, a Dirichlet process prior represents the belief that exploration should be as minimal as possible. This is clearly inappropriate for any situation that is not a planning problem. This analysis was hidden from us under a frequentist perspective as we don't allow priors, hence minimising the frequentist RL objective using a Monte Carlo samples is uncontroversial. More concretely: by always assuming that the agent is in a planning problem, a frequentist approach to reinforcement learning cannot develop its own tools to understand why its methods don't work when the agent is in a learning problem without the use of heuristics.

There is thus little to be lost by abandoning the frequentist RL objective, and much to be gained.

Appendices

A

Probabilistic Measure Theory

This chapter introduces some basic but essential concepts from probabilistic measure theory. It loosely follows Williams [1991], which we refer the reader to for a more complete treatment of the subject.

A.1 Probability Spaces

To motivate a formal measure theoretic treatment of probability, consider the simple experiment of tossing a coin twice. The coin can either land heads H or tails T. This gives four possible outcomes, which we collect in a set of outcomes $O = \{HH, HT, TH, TT\}$ known as the sample space. We also need to define a *σ-Algebra* of possible events from these outcomes, which for our example we take to be the power set:

$$\begin{aligned}\Sigma = 2^O \coloneqq \{&\varnothing, \{HH\}, \{HT\}, \{TH\}, \{TT\},\\ &\{HH, HT\}, \{HH, TH\}, \{HH, TT\},\\ &\{HT, TH\}, \{HT, TT\}, \{TH, TT\},\\ &\{HH, HT, TH\}, \{HH, HT, TT\},\\ &\{HH, TH, TT\}, \{HT, TH, TT\},\\ &\{HH, HT, TH, TT\}\}\end{aligned}$$

We can then introduce a probability measure $P : \Sigma \to [0, 1]$ which measures the probability (however that is interpreted) that the outcome $o \in O$ belongs to an event $E \in \Sigma$. For example, we may be interested in the event

$$E = \textit{At least one tails} = \{HT, HT, TT\}.$$

which is clearly an element of Σ defined above. For a fair coin where every outcome has equal probability, the probability measure is simply:

$$P(\cdot) = \frac{\#(\cdot)}{\#(O)},$$

where $\#(\cdot)$ is the counting measure (i.e. a function that returns the number of elements in the input set). Hence, for the event $E = \textit{At least one tails}$, we have $P(E) = 0.75$.

Together, the triple (O, Σ, P) is known as a *probability space*. In order to define general probability spaces, we only need to make sure that the event space Σ is a σ-alegbra of the sample space and P is a probability measure. We now provide a formal definition of these two objects:

Definition 7 (σ-algebra)**.** *Let 2^O be the power set of a sample space O. $\Sigma \subseteq 2^O$ is a σ-algebra if:*

i Σ *contains the sample space:* $O \in \Sigma$,

ii Σ *is closed under complements: If* $E \in \Sigma$ *then* $O/E \in \Sigma$ *and,*

iii Σ *is closed under countable unions: whenever* $E_n \in \Sigma$ *where* $n \in \mathbb{N}$ *then* $\bigcup_n E_n \in \Sigma$.

Remark. *Conditions i) and ii) imply that Σ contains the empty set: $\varnothing \in \Sigma$ and conditions ii) and iii) imply that Σ is closed under countable intersections: $\bigcap_{n\in\mathbb{N}} E_n \in \Sigma$.*

Definition 8 (Probability measure)**.** *A function $P : \Sigma \to [0, 1]$ is a probability measure if:*

i $P(O) = 1$

ii P *is countably additive: for any sequence of disjoint*[1] *events* $E_n \in \Sigma$ *where* $n \in \mathbb{N}$, $P\left(\bigcup_N E_n\right) = \sum_n P\left(E_n\right)$

[1] Two sets A and B are disjoint if $A \cap B = \varnothing$

Remark. *The conditions of Definition 8 are commonly known as the Kolmogorov axioms [Kolmogorov et al., 1956] and although this formulation seems abstract, they afford the probability space with the following intuitive properties: Firstly, we see from the definition of P that the probability of any event must always be non-negative, which is important as negative probabilities have no meaning. Conditions i) and ii) can be used to derive the result that $P(\varnothing) = 0$ and together with condition i), this says that at least one event from the sample space will occur with probability 1. Finally, the monotonicity property can be derived from conditions i) and ii), which states that for any events $A \subseteq B$, then $p(A) \leq p(B)$.*

In the context of our example, we see that there are more σ-algebras than just 2^O, for example $\{\varnothing, \{HH\}, \{HT, TH, TT\}, \{HH, HT, TH, TT\}\}$ is a valid σ-algebra satisfying Definition 7. In general, for a countable set O, there are as many σ-algebras as there are partitions of O with the largest being the power set 2^O and the smallest being the trivial σ-algebra $\{\varnothing, O\}$. We also note that our probability measure was normalised by dividing by $\#(O)$ to ensure that condition i) of Definition 8 is satisfied. In order to demonstrate the importance of the monotonicity property, consider the two events:

$$E_1 = \text{At least one heads} = \{HH, HT, TH\},$$
$$E_2 = \text{Only one heads} = \{HT, TH\}.$$

Clearly any outcome that satisfies E_2 must also satisfy E_1 as $E_2 \subset E_1$, which is reflected in the fact that $p(E_1) = 0.75 > p(E_2) = 0.5$

A.2 Random Variables

For most applications, it is important to be able to work with functions whose inputs are random - i.e. functions whose domain is the sample space O. These functions are known as *random variables*. As an example, a random variable may be a function that returns number of heads of an outcome o:

$$X(o) = \begin{cases} 2, & o = HH, \\ 1, & o \in \{HT, TH\}, \\ 0, & \text{otherwise.} \end{cases} \tag{A.1}$$

Random variables may be defined as a measurable mapping from outcome space to any measurable space (O', Σ') where Σ' is a σ-alegbra of O'. Given a random variable $X : O \to O'$ we are often interested in the probability of an event in the new space: $E' \in \Sigma'$. We can relate the probability of E' back to the probability of the event in the original event space: $E = X^{-1}(E') \in \Sigma$ using a probability distribution: $P_X(E') \coloneqq P \circ X^{-1}(E')$. The notation $X^{-1}(E') \coloneqq \{o : X(o) = E'\}$ denotes the *pre-image* of E' and the probability P_X is formally known as the *pushforward* of the measure P under X. We now formalise the assumption that X is a measurable function:

Definition 9 (Measurable Function)**.** *A random variable $X : O \to O'$ is measurable if the pre-image of an event $E' \in \Sigma'$ is an element of Σ, that is:*

$$X^{-1}(E') \in \Sigma \; \forall \; E' \in \Sigma'$$

If X is not measurable, then there exists some events in the new event space Σ' whose pre-image we cannot find the probability of, and hence the distribution of X is not well defined.

Typically, random variables map to the real numbers: $X : O \to \mathbb{R}$ which is true of all random variables in this thesis. A useful tool to analyse random variables that map to reals is the Borel σ-algebra:

Definition 10 (Borel σ-algebra)**.** *A Borel σ-algebra $\mathfrak{B}(S)$ of a set S is defined as the smallest σ-algebra generated by all open subsets of S. We use the shorthand $\mathfrak{B} \coloneqq \mathfrak{B}(\mathbb{R})$ to denote the Borel σ-algebra of reals.*

Remark. *As $\mathfrak{B}$ is a σ-algebra, in addition to all open intervals of $\mathbb{R}$, it must also contain all closed intervals, all unions of closed and open intervals and all intersections of closed and open intervals of $\mathbb{R}$.*

Using the Borel σ-algebra, we can introduce a convenient definition of a real-valued random variable:

Definition 11 (Real-valued Random Variable)**.** *A real-valued random variable X of a probability space (O, Σ, P) is a function that is Σ-measurable, i.e.:*

$$X : O \to \mathbb{R}, \quad X^{-1} : \mathfrak{B} \to \Sigma$$

As we have discussed, the probability density of the random variable can be obtained by using the composition:

$$P \circ X^{-1} : \mathfrak{B} \xrightarrow{X^{-1}} \Sigma \xrightarrow{P} [0, 1],$$

however due to the nature of accounting for every possible set under Definition 7, working directly with σ-algebras is cumbersome, making it difficult to define X^{-1} on $\mathfrak{B}$. Fortunately, as $\mathfrak{B}$ can be generated by the set of intervals $\mathfrak{I} \coloneqq \{(-\infty, c] : c \in \mathbb{R}\}$, it suffices[2] to use a *cumulative distribution function* defined as the composition $P \circ X^{-1}$ restricted to $\mathfrak{I}$:

$$F_X(c) \coloneqq P \circ X^{-1}((-\infty, c]) : \mathfrak{I} \xrightarrow{X^{-1}} \Sigma \xrightarrow{P} [0, 1].$$

We now derive a distribution function for the example of the random variable that returns the number of heads from Eq. (A.1). It is simple to derive X^{-1} restricted to $\mathfrak{I}$:

$$X^{-1}((-\infty, c]) = \begin{cases} O, & c \geq 2, \\ \{HH, HT, TH\}, & 2 > c \geq 1, \\ \{TT\}, & 1 > c, \end{cases}$$

which leads to the distribution function:

$$F_X(c) = \begin{cases} 1, & c \geq 2 \\ 0.75, & 2 > c \geq 1 \\ 0.25, & 1 > c. \end{cases}$$

[2] See, for example, Williams [1991] for a rigorous proof of this result

B
Proofs

B.1 Proof of Theorem 5

Central to our proofs is the empirical log-likelihood (ELL):

$$\text{ELL}_N(\phi, \mathcal{D}_\omega^N) \coloneqq \frac{1}{N}\sum_{i=1}^{N} \log p(b_i|s_i, a_i, \phi).$$

Using this notation, we can write the posterior density as:

$$p(\phi|\mathcal{D}_\omega^N) = \frac{\exp\left(\frac{N}{\sigma^2}\text{ELL}_N(\phi, \mathcal{D}_\omega^N)\right) p_\Phi(\phi)}{\int_\Phi \exp\left(\frac{N}{\sigma^2}\text{ELL}_N(\phi, \mathcal{D}_\omega^N)\right) dP_\Phi(\phi)}$$

where λ is the Lebesgue measure. We now derive some useful properties

Proposition 9 (Useful Facts About Functions of ϕ)**.** *Under Assumption 9, $i)$ $\hat{B}_\phi$ is bounded; and $ii)$ $\mathbb{E}_{P_B(b,s,a,|\omega)}\left[|\log p(b|s,a,\phi)|\right] < \infty \; \forall \; \phi \in \Phi$.*

Proof. To prove $i)$, we recall that $\hat{B}_\phi$ is defined as the model's mean: $\hat{B}_\phi(s,a) \coloneqq \mathbb{E}_{P(b|s,a,\phi)}[b]$. If $\hat{B}_\phi(s,a)$ is not bounded for all (s,a), the variance of $P(b|s,a,\phi)$ would also not be bounded, which would lead to a contradiction as $P(b|s,a,\phi)$ has finite variance under Assumption 9. To prove $ii)$, we note that as $p(b|s,a,\phi)$ is bounded under Assumption 9, it must be P_B-integrable and hence $\mathbb{E}_{P_B(b,s,a,|\omega)}\left[|\log p(b|s,a,\phi)|\right] < \infty \; \forall \; \phi \in \Phi$. □

We can now use our important Theorem 1 to show that the posterior concentrates on the KL-minimising parameters:

Lemma 2 (Posterior Concentration)**.** *Under Assumptions 8-10, the posterior converges weak $P_{\mathcal{D}}$-almost surely to a Dirac delta distribution centered on the parameters that minimises the KL divergence:*

$$P_{\Phi}(\cdot|\mathcal{D}_{\omega}^{N}) \xrightarrow[P_{\mathcal{D}}-a.s]{} \delta_{\phi_{\omega}^{\star}}$$

where

$$\phi_{\omega}^{\star} = \arg\max_{\phi\in\Phi} \mathbb{E}_{P_B(b,s,a|\omega)} \left[\log p(b|s,a,\phi)\right]$$

Proof. Theorem 1 implies that our result holds if:

i (Φ, d_{Φ}) is a compact metric space and P_{Φ} has support Φ;

ii $\text{ELL}_N(\cdot, \mathcal{D}_{\omega}^{N}) \xrightarrow[P_{\mathcal{D}}-a.s.-\text{a.s.}]{} \mathbb{E}_{P_B(b,s,a|\omega)} \left[\log p(b|s,a,\phi)\right]$ and $\text{ELL}_N(\cdot, \mathcal{D}_{\omega}^{N})$ and $\mathbb{E}_{P_B(b,s,a|\omega)} \left[\log p(b|s,a,\phi)\right]$ are bounded and continuous in Φ for all N;

iii $\mathbb{E}_{P_B(b,s,a|\omega)} \left[\log p(b|s,a,\phi)\right]$ has a unique maximum $\phi^{\star}$ over Φ.

Conditions i and iii follow from Assumptions 9 and 10. $\text{ELL}_N(\cdot, \mathcal{D}_{\omega}^{N})$ and $\mathbb{E}_{P_B(b,s,a|\omega)} \left[\log p(b|s,a,\phi)\right]$ are bounded and continuous in Φ for all N from Assumption 9 and Proposition 9. We are therefore left to prove:

$$\text{ELL}_N(\phi, \mathcal{D}_{\omega}^{N}) = \frac{1}{N}\sum_{i=1}^{N} \log p(b_i|s_i,a_i,\phi) \xrightarrow[P_{\mathcal{D}}-a.s.]{} \mathbb{E}_{P_B(b,s,a|\omega)} \left[\log p(b|s,a,\phi)\right]. \quad \text{(B.1)}$$

To prove Eq. (B.4), we use the strong law of large numbers (SLLN) under the two sampling options presented in Assumption 8. As $\mathbb{E}_{P_B(b,s,a|\omega)}[|\log p(b|s,a,\phi)|] < \infty$ from Proposition 9, the pointwise SLLN holds for i.i.d. samples by (for example) Williams [1991] Theorem 14.5 or for sampling from an ergodic Markov chain under the conditions of Assumption 8 by Theorem 4.3 of Gilks et al. [1995]. Eq. (B.4) holds, completing our proof. □

Theorem 5. *Under Assumptions 8-10, in the limit $N \to \infty$: i) $P_{\Phi}(\cdot|\mathcal{D}_{\omega}^{N}) \xrightarrow[P_{\mathcal{D}}-a.s]{} \delta_{\phi_{\omega}^{\star}}$; ii) $\mathcal{B}_{\omega,N} \xrightarrow[P-a.s.]{} \hat{B}_{\phi_{\omega}^{\star}}$; and iii) MSBBE$_N(\omega) \xrightarrow[P-a.s.]{} \|\hat{Q}_{\omega} - \hat{B}_{\phi_{\omega}^{\star}}\|_{\varsigma}^{2}$.*

Proof.

$$\textbf{Claim } \boldsymbol{i}) : \quad P_\Phi(\cdot|\mathcal{D}_\omega^N) \xrightarrow[P_\mathcal{D}-\text{a.s}]{} \delta_{\phi_\omega^\star}.$$

Claim $i)$ follows immediately from Lemma 2.

$$\textbf{Claim } \boldsymbol{ii}) : \quad \mathcal{B}_{\omega,N} \xrightarrow[P-\text{a.s.}]{} \hat{B}_{\phi_\omega^\star}.$$

To prove $ii)$ we analyse the convergence of the Bayesian Bellman operator, writing it as an expectation:

$$\lim_{N\to\infty} \mathcal{B}_{\omega,N}^\star = \lim_{N\to\infty} \int_\Phi \hat{B}_\phi dP(\phi|\mathcal{D}_\omega^N).$$

From $i)$, the posterior converges weakly to a Dirac delta distribution centered on $\phi_\omega^\star$ $P_\mathcal{D}$-almost surely. From Proposition 9 $\hat{B}_\phi$ is bounded and from Assumption 9 is Lipschitz in ϕ, so we can apply the portmanteau theorem for weak convergence of measures [Billingsley, 1999] to find the limit:

$$\lim_{N\to\infty} \mathcal{B}_{\omega,N}^\star = \lim_{N\to\infty} \int_\Phi \hat{B}_\phi dP(\phi|\mathcal{D}_\omega^N) = \int_\Phi \hat{B}_\phi d\delta_{\phi_\omega^\star} = \hat{B}_{\phi_\omega^\star} \quad P_\mathcal{D} - a.s., \tag{B.2}$$

as required.

$$\textbf{Claim } \boldsymbol{iii}) : \quad \text{MSBBE}_N(\omega) \xrightarrow[P-\text{a.s.}]{} \|\hat{Q}_\omega - \hat{B}_{\phi_\omega^\star}\|_\varsigma^2.$$

To prove $iii)$, we start with the definition of the MSBBE in the limit $N \to \infty$:

$$\begin{aligned}\lim_{N\to\infty} \text{MSBBE}_N(\omega) &= \lim_{N\to\infty} \left\|\hat{Q}_\omega - \mathcal{B}_{\omega,N}^\star\right\|_{\rho,\pi}^2 \\ &= \lim_{N\to\infty} \frac{1}{2} \int_{\mathcal{S}\times\mathcal{A}} \left(\hat{Q}_\omega(s,a) - \mathcal{B}_{\omega,N}^\star(s,a)\right)^2 d(\rho(s)\pi(a|s)).\end{aligned} \tag{B.3}$$

To apply the dominated convergence theorem to Eq. (B.3), we must show the integrand is dominated and convergences pointwise [Bass, 2013]. Firstly, from Eq. (B.2), we have $\mathcal{B}_{\omega,N}^\star \xrightarrow{P_\mathcal{D}-a.s.} \hat{B}_{\phi_\omega^\star}$, which is bounded by Proposition 9. Now consider the Bayesian Bellman operator for finite N: $\mathcal{B}_{\omega,N}^\star = \int_\Phi \hat{B}_\phi dP(\phi|\mathcal{D}_\omega^N)$. As $\hat{B}_\phi$ is bounded, Φ is compact and $P(\phi|\mathcal{D}_\omega^N)$ is absolutely continuous with respect to the Lebesgue measure, it follows

that $\hat{B}_\phi$ is $P(\phi|\mathcal{D}_\omega^N)$-integrable hence $\mathcal{B}^\star_{\omega,N}$ is bounded for all N by some positive constant $\tilde{C}$.

Consider now the integrand in Eq. (B.3). Applying the triangle inequality, we have:

$$|\hat{Q}_\omega(s,a) - \mathcal{B}^\star_{\omega,N}(s,a)|^2 \leq (|\hat{Q}_\omega(s,a)| + |\mathcal{B}^\star_{\omega,N}(s,a)|)^2 \leq (|\hat{Q}_\omega(s,a)| + \tilde{C})^2.$$

From Assumption 9 $\hat{Q}_\omega$ is bounded and so the integrand is dominated by some finite constant. The dominated convergence theorem therefore applies for Eq. (B.3), hence

$$\lim_{N\to\infty} \text{MSBBE}_N(\omega) = \left\|\hat{Q}_\omega - \lim_{N\to\infty} \mathcal{B}^\star_{\omega,N}\right\|^2_{\rho,\pi} = \left\|\hat{Q}_\omega - \hat{B}_{\phi^\star_\omega}\right\|^2_{\rho,\pi} \quad P_{\mathcal{D}} - a.s,$$

as required. □

Finally, we prove our corollary, which establishes the results of Theorem 5 when RP approximate posterior is used in place of the true posterior. We must first prove the uniform almost sure convergence of the RP approximate inference objective. We show that it suffices to strengthen the pointwise convergence of the empirical log-likelihood that we established in Lemma 2 to uniform convergence.

Lemma 3 (Uniform Almost Sure Convergence of RP Objective)**.** *Under Assumptions 8 and 9,*

$$\mathcal{L}(\phi; \mathcal{D}_\omega^N, \epsilon) \xrightarrow[P_{\mathcal{D}}-a.s.]{\text{unif}} \mathbb{E}_{P_B(b|s,a,\omega)}\left[\log p(b|s,a,\phi)\right] \quad \forall\, \epsilon \in \mathcal{E}.$$

Proof. Applying the triangle inequality to Definition 2 yields:

$$\begin{aligned}
&\sup_{\phi\in\Phi} \left|\mathcal{L}(\phi; \mathcal{D}_\omega^N, \epsilon) - \mathbb{E}_{P_B(b|s,a,\omega)}\left[\log p(b|s,a,\phi)\right]\right| \\
&= \sup_{\phi\in\Phi} \left|\text{ELL}_N(\phi, \mathcal{D}_\omega^N) - \mathbb{E}_{P_B(b,s,a|\omega)}\left[\log p(b|s,a,\phi)\right] - \frac{1}{N} R(\phi - \epsilon)\right|, \\
&\leq \sup_{\phi\in\Phi} \left|\text{ELL}_N(\phi, \mathcal{D}_\omega^N) - \mathbb{E}_{P_B(b,s,a|\omega)}\left[\log p(b|s,a,\phi)\right]\right| + \frac{1}{N} \sup_{\phi\in\Phi} \left|R(\phi - \epsilon)\right|.
\end{aligned}$$

As $R(\phi-\epsilon)$ is bounded for all $\epsilon \in \mathcal{E}$, by Assumption 9, $\lim_{N\to\infty} \frac{1}{N} \sup_{\phi\in\Phi} \left|R(\phi-\epsilon)\right| = 0$, hence:

$$\begin{aligned}
&\lim_{N\to\infty} \sup_{\phi\in\Phi} \left|\mathcal{L}(\phi; \mathcal{D}_\omega^N, \epsilon) - \mathbb{E}_{P_B(b,s,a|\omega)}\left[\log p(b|s,a,\phi)\right]\right| \\
&\quad \leq \lim_{N\to\infty} \sup_{\phi\in\Phi} \left|\text{ELL}_N(\phi, \mathcal{D}_\omega^N) - \mathbb{E}_{P_B(b,s,a|\omega)}\left[\log p(b|s,a,\phi)\right]\right|.
\end{aligned}$$

We are therefore left to prove:

$$\text{ELL}_N(\phi, \mathcal{D}_\omega^N) \xrightarrow[P_{\mathcal{D}}-a.s.]{\text{unif}} \mathbb{E}_{P_B(b,s,a|\omega)} \left[\log p(b|s, a, \phi)\right] \tag{B.4}$$

Theorem 3 of Andrews [1992] states that (B.4) holds if i) Φ is totally bounded, ii) $\log p(b|s, a, \phi)$ is Lipschitz in ϕ and iii) the empirical mean converges pointwise almost surely, that is:

$$\text{ELL}_N(\phi, \mathcal{D}_\omega^N) \xrightarrow[P_{\mathcal{D}}-a.s.]{} \mathbb{E}_{P_B(b,s,a|\omega)} \left[\log p(b|s, a, \phi)\right]$$

Conditions i) and ii) are satisfied by Assumption 9 and we have already proved iii) in Lemma 2. The result holds immediately. □

Corollary 5.1. *Under Assumptions 8-11, results i)-iii) of Theorem 5 hold with $P(\cdot|\mathcal{D}_\omega^N)$ replaced by the RP approximate posterior $\hat{P}(\cdot|\mathcal{D}_\omega^N)$ both with or without ensembling.*

Proof. As results $ii)$ and $iii)$ of Theorem 5 follow directly from $i)$ under Assumptions 8-10, we only need to prove $i)$ still holds with the approximate posterior, that is:

$$\hat{P}(\cdot|\mathcal{D}_\omega^N) \xrightarrow[P_{\mathcal{D}}-\text{a.s}]{} \delta_{\phi_\omega^\star}.$$

We consider the case of using the exact RP posterior (without ensembling) $\hat{P}(\cdot|\mathcal{D}_\omega^N) \coloneqq \int \delta_{\psi^*(\epsilon;\mathcal{D}_\omega^N)} dP_E(\epsilon)$. From the Portmanteau Theorem for the weak convergence of measures (Proposition 1), it suffices to prove:

$$\int_\Phi f(\phi) d\hat{P}(\phi|\mathcal{D}_\omega^N) \xrightarrow[P_{\mathcal{D}}-a.s.]{} \int_\Phi f(\phi) d\delta_{\phi_\omega^\star} = f(\phi_\omega^\star). \tag{B.5}$$

for any Lipschitz, bounded function $f : \Phi \to \mathbb{R}$. Substituting for the definition of the RP approximate posterior $\hat{P}(\cdot|\mathcal{D}_\omega^N) \coloneqq \int \delta_{\psi^*(\epsilon;\mathcal{D}_\omega^N)} dP_E(\epsilon)$, proving (B.5) holds is equivalent to proving:

$$\int_\mathcal{E} f \circ \psi_N(\epsilon) dP_E(\epsilon) \xrightarrow[P_{\mathcal{D}}-a.s.]{} f(\phi_\omega^\star), \tag{B.6}$$

for any sequence $(\psi_N(\epsilon))$ where $\psi_N(\epsilon) \in \psi^*(\epsilon; \mathcal{D}_\omega^N)$. Under the definition of RP (see Section 4.3.3), it is implicitly assumed that $f \circ \psi_N : \mathcal{E} \to \mathbb{R}$ is P_E-integrable for any

bounded, Lipschitz $f : \Phi \to \mathbb{R}$. Hence, we can apply the dominated convergence theorem to (B.6) to derive an equivalent condition to prove that (B.5) holds:

$$f \circ \psi_N(\epsilon) \xrightarrow[P_{\mathcal{D}}-a.s.]{} f(\phi^\star_\omega) \quad \forall\, \epsilon \in \mathcal{E},$$

which, from the continuity of f, is equivalent to proving that any sequence of minimisers of $\mathcal{L}(\phi; \mathcal{D}^N_\omega, \epsilon)$ converge almost surely to the maximisers of $\mathbb{E}_{P_B(b|s,a,\omega)}[\log p(b|s,a,\phi)]$, that is:

$$\psi_N(\epsilon) \xrightarrow{P_{\mathcal{D}}-a.s.} \phi^\star_\omega \quad \forall\, \epsilon \in \mathcal{E}. \tag{B.7}$$

Theorem 7.33 of Rockafellar and Wets [1998] states that (B.7) holds if for all $\epsilon \in \mathcal{E}$: $a)$ $\mathcal{L}(\phi; \mathcal{D}^N_\omega, \epsilon)$ and $-\mathbb{E}_{P_B(b|s,a,\omega)}[\log p(b|s,a,\phi)]$ are proper, lower semi-continuous functions of ϕ; $b)$ the sequence $(\mathcal{L}(\phi; \mathcal{D}^N_\omega, \epsilon))$ is eventually level-bounded; and $c)$ $\mathcal{L}(\phi; \mathcal{D}^N_\omega, \epsilon)$ epi-converges to $-\mathbb{E}_{P_B(b|s,a,\omega)}[\log p(b|s,a,\phi)]$.

Condition $a)$ is trivially satisfied by the assumption of $p(b|s,a,\phi)$ being bounded and Lipschitz in ϕ in Assumption 9, and as a continuous function, is lower semi-continuous as any continuous mapping between any two metric spaces is proper. Recall the definition of $\mathcal{L}(\phi; \mathcal{D}^N_\omega, \epsilon)$:

$$\mathcal{L}(\phi; \mathcal{D}^N_\omega, \epsilon) \coloneqq -\frac{1}{N}\sum_{i=1}^{N} \log p(b_i|s_i,a_i,\phi) + \frac{1}{N} R(\phi - \epsilon).$$

As $p(b|s,a,\phi)$ is bounded and $R(\phi-\epsilon)$ is bounded by some $R_{\max} > 0$ under Assumption 9, there exists a $K > 0$ such that $|\log p(b_i|s_i,a_i,\phi)| \leq K$ which we use to bound $\mathcal{L}(\phi; \mathcal{D}^N_\omega, \epsilon)$:

$$\begin{aligned}
|\mathcal{L}(\phi; \mathcal{D}^N_\omega, \epsilon)| &\leq \frac{1}{N}\sum_{i=1}^{N} |\log p(b_i|s_i,a_i,\phi)| + \frac{1}{N}|R(\phi-\epsilon)|, \\
&\leq \frac{1}{N}\sum_{i=1}^{N} K + \frac{R_{\max}}{N}, \\
&= K + \frac{R_{\max}}{N}.
\end{aligned}$$

As $\mathcal{L}(\phi; \mathcal{D}^N_\omega, \epsilon)$ is bounded for all $N \in \mathbb{Z}^+$ and $\epsilon \in \mathcal{E}$, the sequence is trivially eventually level-bounded, hence $b)$ is satisfied.

To establish epi-convergence of condition $c)$, we use Proposition 7.15 a) of Rockafellar and Wets [1998] which strengthens the uniform convergence established in Lemma 3 to epi-convergence $P_{\mathcal{D}}$-almost surely provided each $\mathcal{L}(\phi; \mathcal{D}_{\omega}^{N}, \epsilon)$ is lower semi-continuous, which we have already proved. Condition $c)$ is satisfied and hence our desired result hold.

All arguments above also hold using the ensembled approximation $\hat{P}(\phi|\mathcal{D}_{\omega}^{N}) \approx \frac{1}{L}\sum_{l=1}^{N}\delta(\phi \in \psi(\mathcal{D}_{\omega}^{N}, \epsilon_l))$ if we replace expectations of any function $f(\epsilon)$ under $P_E(\epsilon)$: $\int_{\mathcal{E}} f(\epsilon) dP_E(\epsilon)$ with expectations over ensembles: $\frac{1}{L}\sum_{l=1}^{N} f(\epsilon_l)$. □

B.2 Proof of Theorem 6

As we are concerned with finite N in our Bayesian analysis, we analyse the RP objective ignoring the factor of $\frac{1}{N}$ as it leaves the solution unchanged:

$$\mathcal{L}(\phi; \mathcal{D}_{\omega_l}^{N}, \epsilon_l) \coloneqq R(\phi - \epsilon_l) - \sum_{i=1}^{N} \log p(b_i|s_i, a_i, \phi).$$

We repeat the sequence of updates from Section 7.3 here for convenience:

$$\psi_l = \mathcal{P}_{\Omega}\left(\psi_l - \alpha_k \nabla_{\psi_l}\left(R(\psi_l - \epsilon_l) - \log p(b_i|s_i, a_i, \psi_l)\right)\right), \quad \text{(fast)} \tag{B.8}$$

$$\omega_l = \mathcal{P}_{\Omega}(\omega_l - \beta_k(\omega_l - \psi_l)). \quad \text{(slow)} \tag{B.9}$$

We now use Assumption 12 to establish the boundedness of several quantities that will be essential for our main proof:

Proposition 10. *Under Assumptions 11 and 12, the quantities i)* $\nabla_{\psi_l} R(\psi_l - \epsilon_l)$ *and* $\nabla_{\psi_l} \log p(b_i|s_i, a_i, \psi_l)$*; ii)* $\sum_{j\neq i}^{N} \nabla_{\psi_l} \log p(b_j|s_j, a_j, \psi_l)$*; and iii)* $\nabla_{\psi_l} \mathcal{L}(\psi_l; \mathcal{D}_{\omega_l}^{N}, \epsilon_l)$ *are all bounded.*

Proof. We establish $i)$ by noting that from Assumption 12, $\nabla_{\psi_l} R(\psi_l - \epsilon_l)$ and $\nabla_{\psi_l} \log p(b_i|s_i, a_i, \psi_l)$ are Lipschitz in ψ_l, and from Assumption 11, Φ is compact, hence as any Lipschitz function defined on a compact set must be bounded, it follows that $\nabla_{\psi_l} R(\psi_l - \epsilon_l)$ and $\nabla_{\psi_l} \log p(b_i|s_i, a_i, \psi_l)$ are bounded. To prove $ii)$, we bound

$\sum_{j\neq i}^{N} \nabla_{\psi_l} \log p(b_j|s_j, a_j, \psi_l)$ using the triangle inequality:

$$\begin{aligned}\left|\sum_{j\neq i}^{N} \nabla_{\psi_l} \log p(b_j|s_j, a_j, \psi_l)\right| \leq & \sum_{j\neq i}^{N} \left|\nabla_{\psi_l} \log p(b_j|s_j, a_j, \psi_l)\right|, \\ < & N \sup_{j\in\{1:N\}} \left|\nabla_{\psi_l} \log p(b_j|s_j, a_j, \psi_l)\right|,\end{aligned}$$

which is bounded from $i)$ and the fact that N is finite in our Bayesian regime. To prove $iii)$, we establish a similar bound:

$$\begin{aligned}\left|\nabla_{\psi_l} \mathcal{L}(\psi_l; \mathcal{D}^N_{\omega_l}, \epsilon_l)\right| &= \left|\nabla_{\psi_l}\left(R(\psi_l - \epsilon_l) - \sum_{i=1}^{N} \log p(b_i|s_i, a_i, \psi_l)\right)\right|, \\ &\leq \left|\nabla_{\psi_l} R(\psi_l - \epsilon_l)\right| + \sum_{i=1}^{N} \left|\nabla_{\psi_l} \log p(b_i|s_i, a_i, \psi_l)\right|, \\ &\leq \left|\nabla_{\psi_l} R(\psi_l - \epsilon_l)\right| + N \sup_{i\in\{1:N\}} \left|\nabla_{\psi_l} \log p(b_i|s_i, a_i, \psi_l)\right|,\end{aligned}$$

which is bounded from $i)$ and N being finite. □

To ease the notational burden of our proof, we drop dependence on the ensemble index $l \in \{1 : L\}$ as the convergence proof is the same for all ensembles. We formalise updates (B.8) and (B.9) by analysing the recursive sequence:

$$\psi_{k+1} = \mathcal{P}_\Omega\left(\psi_k - \alpha_k \nabla_\psi \left(R(\psi_k - \epsilon) - \log p(b_i|s_i, a_i, \psi_k)\right)\right), \quad \text{(fast)} \tag{B.10}$$

$$\omega_{k+1} = \mathcal{P}_\Omega(\omega_k - \beta_k(\omega_k - \psi_k)). \quad \text{(slow)}$$

where $i \sim \text{Unif}(\{1 : N\})$.

Theorem 6. *If Assumptions 8 to 12 hold, ψ_l and ω_l converge to $\psi_l^{\circledast}(\omega_l^{\circledast})$ and $\omega_l^{\circledast}$ almost surely.*

Proof. Firstly, we define the martingale:

$$\mathcal{M}^i_k \coloneqq -\nabla_\psi \sum_{j\neq i}^{N} \log p(b_j|s_j, a_j, \psi_k).$$

Using the identity:

$$\nabla_\psi \left(R(\psi_k - \epsilon) - \log p(b_i|s_i, a_i, \psi_k)\right) = \nabla_\psi \mathcal{L}(\psi_k; \mathcal{D}^N_{\omega_k}, \epsilon) - \mathcal{M}^i_k \tag{B.11}$$

we re-write the recursive update in Eq. (B.10) as a martingale difference sequence:

$$\psi_{k+1} = \mathcal{P}_\Omega \left(\psi_k - \alpha_k (\nabla_\psi \mathcal{L}(\psi_k; \mathcal{D}^N_{\omega_k}, \epsilon) - \mathcal{M}^i_k) \right).$$

As $\mathcal{P}_\Omega$ has a well-defined directional derivative under Proposition 8, we can re-write the fast update using a series expansion about ψ_k:

$$\begin{aligned}
&\mathcal{P}_\Omega \left(\psi_k - \alpha_k (\nabla_\psi \mathcal{L}(\psi_k; \mathcal{D}^N_{\omega_k}, \epsilon) - \mathcal{M}^i_k) \right) \\
&\quad = \mathcal{P}_\Omega(\psi_k) - \Gamma_{\psi_k} \left[\alpha_k (\nabla_\psi \mathcal{L}(\psi_k; \mathcal{D}^N_{\omega_k}, \epsilon) - \mathcal{M}^i_k) \right] + o(\alpha_n^2), \\
&\quad = \psi_k + \alpha_k \Gamma_{\psi_k} \left[\mathcal{M}^i_k - \nabla_\psi \mathcal{L}(\psi_k; \mathcal{D}^N_{\omega_k}, \epsilon) \right] + o(\alpha_n^2), \\
&\quad = \psi_k + \alpha_k \left(\Gamma_{\psi_k} \left[\mathcal{M}^i_k \right] - \Gamma_{\psi_k} \left[\nabla_\psi \mathcal{L}(\psi_k; \mathcal{D}^N_{\omega_k}, \epsilon) \right] + o(\alpha_n) \right),
\end{aligned}$$

where we have used the identity $\mathcal{P}_\Omega(\psi_k) = \psi_k$ in deriving the second equality and the linearity of the derivative for the final equality. Consider now the slow update, which we can write using a similar expansion:

$$\begin{aligned}
\mathcal{P}_\Omega(\omega_k - \beta_k(\omega_k - \psi_k)) &= \mathcal{P}_\Omega(\omega_k) - \Gamma_{\psi_k} \left[\beta_k (\omega_k - \psi_k) \right] + o(\beta_k^2), \\
&= \omega_k + \beta_k \left(\Gamma_{\psi_k} \left[-(\omega_k - \psi_k) \right] + o(\beta_k) \right).
\end{aligned}$$

We write the two updates together here for clarity:

$$\psi_{n+1} = \psi_k + \alpha_k \left(\Gamma_{\psi_k} \left[\mathcal{M}^i_k \right] - \Gamma_{\psi_k} \left[\nabla_\psi \mathcal{L}(\psi_k; \mathcal{D}^N_{\omega_k}, \epsilon) \right] + o(\alpha_n) \right), \tag{B.12}$$

$$\omega_{n+1} = \omega_k + \beta_k \left(\Gamma_{\psi_k} \left[-(\omega_k - \psi_k) \right] + o(\beta_k) \right). \tag{B.13}$$

Our proof adapts Theorem 1 of Heusel et al. [2017], which is a generalisation of Borkar [2008] using the proofs of Karmakar and Bhatnagar [2018]. There is one subtle difference between the updates of Heusel et al. [2017] and the updates for BBO in Eqs. (B.12) and (B.13): in Eqs. (B.12) and (B.13), there is an additional term $o(\alpha_k)$ and $o(\beta_k)$ such that by Assumption 12, $\lim_{k\to\infty} o(\alpha_k) = o(\beta_k) = 0$. $o(\beta_k)$ will be absorbed in the error term ϵ_n in the slow update in Eq. 14 and $o(\alpha_k)$ will be absorbed in the error term ϵ'_n in the fast update in Eq. 17 of Karmakar and Bhatnagar [2018]. As $\lim_{n\to\infty} \epsilon_n = \epsilon'_n = 0$, these additional terms $o(\alpha_n)$ and $o(\beta_n)$ leave the proof of Theorem 5 of Karmakar and Bhatnagar [2018] unaffected. Theorem 1 of Heusel et al. [2017] therefore remains

unchanged, which uses Theorem 5 of Karmakar and Bhatnagar [2018] to establish that the sequences of Eqs. (B.12) and (B.13) converge almost surely to a set of local attractors of the underlying ODE:

$$\{\omega_k, \psi_k\} \xrightarrow{k\to\infty} \{\omega^{\circledast}, \psi^{\circledast}(\omega^{\circledast})\}\ a.s.$$

if

1. $\Gamma_{\psi_k}\left[-\nabla_{\psi}\mathcal{L}(\psi_k; \mathcal{D}^N_{\omega_k}, \epsilon)\right]$ is Lipschitz in ψ_k and $\Gamma_{\omega_k}\left[-(\omega_k - \psi_k)\right]$ is Lipschitz in ω_k under Assumption 12,
2. Stepsizes α_k and β_k satisfy Assumption 12
3. For all k, $(\Gamma_{\psi_k}[\mathcal{M}^i_k])$ is a martingale difference sequence with respect to the filtration of increasing σ-algebras:

$$\mathscr{F}_k \coloneqq \sigma\left(\psi_j, \Gamma_{\psi_j}\left[\mathcal{M}^i_j\right] : j = 1, ...k\right),$$

   with

$$\mathbb{E}\left[\left\|\Gamma_{\psi_k}\left[\mathcal{M}^i_k\right]\right\|^2 | \mathscr{F}_k\right] < \infty,$$

4. The limiting ODEs $\partial_t\psi(t) = -\Gamma_{\psi(t)}(\nabla_{\psi}\mathcal{L}(\psi(t); \mathcal{D}^N_{\omega}, \epsilon))$ and $\partial_t\omega(t) = -\Gamma_{\omega(t)}(\omega_l(t) - \psi^{\circledast}_l(\omega_l(t)))$ satisfy Assumption 12,
5. $\sup_n\|\psi_n\| < \infty$ and $\sup_n\|\omega_n\| < \infty$ a.s.

As 5) is satisfied by virtue of the projection operators, we are left to prove 3).

As $\Gamma_{\psi_k}[\cdot]$ is a linear operator with $\Gamma_{\psi_k}[0] = 0$, $(\Gamma_{\psi_k}[\mathcal{M}^i_k])$ is a Martingale difference sequence if $(\mathcal{M}^i_k)$ is a Martingale difference sequence. To prove 3) holds, it therefore suffices to show that $(\mathcal{M}^i_k)$ is a Martingale difference sequences with respect to the filtration of increasing σ-algebras:

$$\mathscr{F}'_k \coloneqq \sigma\left(\psi_j, \mathcal{M}^i_j : j = 1, ...k\right),$$

To prove this, we observe that: i) $\mathcal{M}^i_k$ is $\mathscr{F}'_k$-measurable for all k by construction, from Proposition 10 $\mathcal{M}^i_k$ is bounded, hence ii) $\mathbb{E}[\mathcal{M}^i_k] < \infty$ and from Eq. (B.11) iii) the

conditional expectation of $\mathcal{M}_k^i$ satisfies for all k:

$$\begin{aligned}
\mathbb{E}\left[\mathcal{M}_k^i|\mathscr{F}_k'\right] &= \nabla_\psi \mathcal{L}(\psi_k; \mathcal{D}_{\omega_k}^N, \epsilon) - \mathbb{E}\left[\nabla_\psi \left(R(\psi_k - \epsilon) - \log p(b_i|s_i, a_i, \psi_k)\right) \middle| \mathscr{F}_k'\right], \\
&= \nabla_\psi \mathcal{L}(\psi_k; \mathcal{D}_{\omega_k}^N, \epsilon) - \nabla_\psi R(\psi_k - \epsilon) + \mathbb{E}\left[\nabla_\psi \log p(b_i|s_i, a_i, \psi_k)|\mathscr{F}_k'\right], \\
&= \nabla_\psi \mathcal{L}(\psi_k; \mathcal{D}_{\omega_k}^N, \epsilon) - \nabla_\psi R(\psi_k - \epsilon) + \sum_{i=1}^N \nabla_\psi \log p(b_i|s_i, a_i, \psi_k), \\
&= 0.
\end{aligned}$$

Together i) to iii) form the definition of a martingale difference sequence with respect to the filtration $(\mathscr{F}_k')$ [Williams, 1991]. As we sample from a discrete uniform distribution and from Proposition 10 $\nabla_{\psi_l} R(\psi_l - \epsilon_l)$, $\nabla_{\psi_l} \log p(b_i|s_i, a_i, \psi_k)$ and $\nabla_\psi \mathcal{L}(\psi_k; \mathcal{D}_{\omega_k}^N, \epsilon)$ are bounded, the variance of the updates is finite:

$$\begin{aligned}
&\mathbb{E}\left[\|\mathcal{M}_k^i\|^2|\mathscr{F}_k'\right] \\
&\quad = \mathbb{E}\left[\left(\nabla_\psi \mathcal{L}(\psi_k; \mathcal{D}_{\omega_k}^N, \epsilon) - \nabla_\psi \left(R(\psi_k - \epsilon) - \log p(b_i|s_i, a_i, \psi_k)\right)\right)^2 \middle| \mathscr{F}_k'\right] < \infty,
\end{aligned}$$

and hence 3) is satisfied as required. □

B.2.1 A Frequentist Analysis of Theorem 6

In the frequentist regime, we are concerned with the objective in the limit $N \to \infty$ and we consider data to be arriving online. Our convergence analysis holds if we replace the assumption of uniform sampling from the dataset $\mathcal{D}_\omega^N$ with (assuming samples have finite variance) sampling i.i.d. from the underlying data distribution $(b_i, s_i, a_i) \sim P_B$. The condition of i.i.d. unbiased estimates leaves our analysis unchanged, however our proof can be extended to cases where the updates are sampled from a Markov chain. In particular, if assumptions (A1), (A5) and (A6)' of Karmakar and Bhatnagar [2018] are satisfied, Theorem 6 can be extended to the non-i.i.d. case. We do not discuss these assumptions further as they have been extensively discussed (with a case study) in Heusel et al. [2017] which sufficiently covers extending BBO to sampling from a Markov chain in the frequentist regime. Note that the ergodicity of the Markov chain in Assumption 8 greatly simplifies their verification as the set of ergodic occupancy measures is a singleton.

C
Linear BBO

To further highlight the relationship between BBO and convergent TD algorithms that minimise the mean squared projected Bellman operator, we explore the linear Gaussian regression model as a case study, through which we derive a regularised Bayesian TDC/GTD2 algorithm [Sutton et al., 2009b,a].

C.1 The Linear Model

We assume a linear Q-function approximator $\hat{Q}_\omega = \omega^\top v$ and a conjugate Bayesian Gaussian linear regression model, which has a likelihood:

$$P(\cdot|s,a,\phi) = \mathcal{N}(\phi^\top v(s,a), \sigma^2), \quad P_\Phi = \mathcal{N}(\phi_0, \Sigma_0),$$

where $\phi, \omega \in \mathbb{R}^n$ and $v : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a basis vector. For mathematical convenience, we form an $N \times n$ matrix of features $V_N \coloneqq (v_1 \cdots v_N)^\top$ from our data where $v_i \coloneqq v(s_i, a_i)$ and an N-dimensional vector of datapoints $\beta_\omega^N \coloneqq \{b_1, b_2, ...b_N\}$. Our Bayesian Gaussian linear regression model has been well studied (see Murphy [2012]) and the posteriors can be shown to be Gaussian:

$$P_\Phi(\cdot|\mathcal{D}_\omega^N) = \mathcal{N}(\phi_\omega^N, \Sigma_N), \quad P_B(\cdot|s,a,\mathcal{D}_\omega^N) = \mathcal{N}(v(s,a)^\top \phi_\omega^N, \sigma^2 + v(s,a)^\top \Sigma_N v(s,a)),$$

where

$$\Sigma_N = \left(\Sigma_0^{-1} + \frac{1}{\sigma^2} V_N^\top V_N\right)^{-1}, \tag{C.1}$$

$$\phi_\omega^N = \Sigma_N \left(\Sigma_0^{-1}\phi_0 + \frac{1}{\sigma^2} V_N^\top \beta_\omega^N\right). \tag{C.2}$$

The variance of $P_B(\cdot|s,a,\mathcal{D}_\omega^N)$ is composed of two terms: the first term, σ^2, is the aleatoric uncertainty of the data due to observation noise; the second term, $v(s,a)^\top \Sigma_N v(s,a)$, characterises the epistemic uncertainty, and grows whenever the test point $v(s,a)$ is far from observed data. An agent can gauge how worthwhile it is to explore a region of high epistemic uncertainty depending upon its potential for high returns. Areas of low epistemic uncertainty will be 'crossed-off' by the agent and are likely not to be explored again, only exploited if their predictive returns are high enough.

By definition, the predictive mean is the Bayesian Bellman operator $\mathcal{B}_{\omega,N}(s,a) = v(s,a)^\top \phi_\omega^N$ from which we derive the MSBBE for linear BBO:

$$\begin{aligned}
\text{MSBBE}_N(\omega) &\coloneqq \left\|\hat{Q}_\omega - \mathcal{B}_{\omega,N}\right\|_{\rho,\pi}^2, \\
&= \left\|v(s,a)^\top \omega - v(s,a)^\top \phi_\omega^N\right\|_{\rho,\pi}^2, \\
&= \left\|v(s,a)^\top (\omega - \phi_\omega^N)\right\|_{\rho,\pi}^2, \\
&= \frac{1}{2}(\omega - \phi_\omega^N)^\top \mathbb{E}_{\rho,\pi}\left[v(s,a)v(s,a)^\top\right](\omega - \phi_\omega^N).
\end{aligned} \tag{C.3}$$

C.2 Deriving a Least-Squared Algorithm

Although we can obtain an analytic expression for the gradient of the MSBBE, a simple approach for inferring the Bayesian Bellman operator is to solve the equation $\omega = \phi_\omega^N$: indeed, as a sanity check, we see from Eq. (C.3) that any $\omega^\star$ such that $\omega^\star = \phi_{\omega^\star}^N$ trivially results in an MSBBE of 0 and parametrises a global minimiser. Expanding our equation

using the definition of ϕ_ω^N from Eq. (C.2), we obtain:

$$\begin{aligned}\omega^\star &= \phi_{\omega^\star}^N, \\ &= \Sigma_N \left(\Sigma_0^{-1}\phi_0 + \frac{1}{\sigma^2} V_N^\top \beta_{\omega^\star}^N \right), \\ &= \Sigma_N \Sigma_0^{-1}\phi_0 + \frac{1}{\sigma^2}\Sigma_N V_N^\top \beta_{\omega^\star}^N, \\ &= \Sigma_N \Sigma_0^{-1}\phi_0 + \frac{1}{\sigma^2}\Sigma_N \sum_{i=1}^N v_i b_i. \end{aligned} \tag{C.4}$$

Introducing the shorthand $v_i' \coloneqq v(s_i', a_i')$, we substitute for the definition of the empirical Bellman equation from Eq. (7.6) using the linear function approximator, $b_i = r_i + \gamma {v_i'}^\top \omega$. This allows us to factorise Eq. (C.4):

$$\begin{aligned}\omega^\star =& \Sigma_N \Sigma_0^{-1}\phi_0 + \frac{1}{\sigma^2}\Sigma_N \sum_{i=1}^N v_i (r_i + \gamma {v_i'}^\top \omega^\star), \\ \Sigma_N^{-1}\omega^\star =& \Sigma_0^{-1}\phi_0 + \frac{1}{\sigma^2}\sum_{i=1}^N v_i r_i + \frac{\gamma}{\sigma^2}\left(\sum_{i=1}^N v_i {v_i'}^\top\right)\omega^\star, \\ \left(\Sigma_N^{-1} - \frac{\gamma}{\sigma^2}\sum_{i=1}^N v_i {v_i'}^\top\right)\omega^\star &= \Sigma_0^{-1}\phi_0 + \frac{1}{\sigma^2}\sum_{i=1}^N v_i r_i, \end{aligned}$$

Substituting for the definition of the predictive covariance from Eq. (C.1):

$$\Sigma_N^{-1} = \Sigma_0^{-1} + \frac{1}{\sigma^2} V_N^\top V_N = \Sigma_0^{-1} + \frac{1}{\sigma^2}\sum_{i=1}^N v_i v_i^\top,$$

yields:

$$\left(\Sigma_0^{-1} + \frac{1}{\sigma^2}\sum_{i=1}^N v_i \left(v_i - \gamma v_i'\right)^\top\right)\omega^\star = \Sigma_0^{-1}\phi_0 + \frac{1}{\sigma^2}\sum_{i=1}^N v_i r_i,$$

Define the matrix:

$$D_N \coloneqq \Sigma_0^{-1} + \frac{1}{\sigma^2}\sum_{i=1}^N v_i \left(v_i - \gamma v_i'\right)^\top,$$

and the vector:

$$\chi_N \coloneqq \Sigma_0^{-1}\phi_0 + \frac{1}{\sigma^2}\sum_{i=1}^N v_i r_i$$

with which we obtain a solution for $\omega^\star$:

$$\omega^\star = D_N^{-1}\chi_N$$

If the problem is tractable enough to store D_N^{-1}, then the exact solution $\omega^\star$ can be obtained using Sherman-Morrison updates as outlined in Algorithm 5. We use an importance weight $w_i \coloneqq \frac{\pi(a_i|s_i)}{\pi_e(a_i|s_i)}$ for the off-policy case when an exploratory policy π_e that is difference from the evaluation policy π is used to gather data. When the on-policy sampling is used, $w_i = 1$. If the prior takes a form that can be inverted with complexity $\mathcal{O}(n^2)$ or less, the overall complexity of Algorithm 5 is $\mathcal{O}(Nn^2)$.

Algorithm 5 BBO LSTD

for $i \in \{1, ...N\}$ **do**
  if $i == 1$ **then**
    $D^{-1} = \left(\sigma^2\Sigma_0^{-1}\right)^{-1}$
    $\chi = \sigma^2\Sigma_0^{-1}\phi_0$
  end if
  $\Delta = w_i(v_i - \gamma v_i')$
  $D^{-1} = D^{-1} - \frac{D^{-1}v_i\Delta^\top D^{-1}}{1+\Delta^\top D^{-1}v_i}$
  $\chi = \chi + w_i v_i r_i$
end for
$\omega^\star = D^{-1}\chi$
return $\omega^\star$

C.3 Recovering TDC/GTD2

To derive TDC/GTD2 from our Gaussian linear regression model, we observe that the Bayesian Bellman operator can be interpreted as an N-sample Monte-Carlo estimate of the projection operator with additional bias due to the prior:

$$\begin{aligned}\mathcal{B}_{\omega,N} &= v^\top \left(\sigma^2\Sigma_0^{-1} + V_N^\top V_N\right)^{-1}\left(\sigma^2\Sigma_0^{-1}\phi_0 + V_N^\top \beta_\omega^N\right), \\ &= v^\top \left(\frac{1}{N}\left(\sigma^2\Sigma_0^{-1} + \sum_{i=1}^N v_i v_i^\top\right)\right)^{-1}\frac{1}{N}\left(\sigma^2\Sigma_0^{-1}\phi_0 + \sum_{i=1}^N v_i b_i\right). \end{aligned} \tag{C.5}$$

As the TDC/GTD2 is a frequentist algorithm, we must derive the Bayesian Bellman operator in the limit $N \to \infty$. Taking the limit $N \to \infty$ of Eq. (C.5) using the strong law of large numbers, we see that the effect of the prior will diminish (observe that it does not scale with increasing N) and the Bayesian Bellman operator converges to the projection operator:

$$\begin{aligned}\lim_{N\to\infty}\mathcal{B}_{\omega,N} &= \lim_{N\to\infty} v^\top \left(\frac{1}{N}\left(\sigma^2\Sigma_0^{-1} + \sum_{i=1}^N v_i v_i^\top\right)\right)^{-1}\frac{1}{N}\left(\sigma^2\Sigma_0^{-1}\phi_0 + \sum_{i=1}^N v_i b_i\right), \\ &= \lim_{N\to\infty} v^\top \left(\frac{1}{N}\sum_{i=1}^N v_i v_i^\top\right)^{-1}\frac{1}{N}\sum_{i=1}^N v_i b_i, \\ &= v^\top \mathbb{E}_{\rho,\pi}\left[vv^\top\right]^{-1}\mathbb{E}_{\rho,\pi}\left[v\mathcal{B}[\hat{Q}_\omega]\right], \\ &= \mathcal{P}\circ\mathcal{B}[\hat{Q}_\omega].\end{aligned}$$

This confirms the consistency results that we established between the mean squared projected Bellman error and MSBBE in the limit $N \to \infty$ in Section 7.2.2 for Gaussian models. The MSPBE has been well studied for linear function approximators [Sutton et al., 2009a,b, Bhatnagar et al., 2009] and can be shown to take the form:

$$\text{MSPBE}(\omega) = \mathbb{E}_{\rho,\pi}\left[v^\top(\mathcal{B}[\hat{Q}_\omega] - \hat{Q}_\omega)\right]\mathbb{E}_{\rho,\pi}\left[vv^\top\right]^{-1}\mathbb{E}_{\rho,\pi}\left[v(\mathcal{B}[\hat{Q}_\omega] - \hat{Q}_\omega)\right]. \quad \text{(C.6)}$$

Taking gradients to minimise the MSPBE via stochastic gradient descent leads to the TDC/GTD2 algorithms, which are derived in full in Sutton et al. [2009a,b] from the above objective. To avoid the costly matrix inversion in Eq. (C.6), a set of weights to approximate the term $\zeta \approx \mathbb{E}_{\rho,\pi}\left[vv^\top\right]^{-1}\mathbb{E}_{\rho,\pi}\left[v(\mathcal{B}[\hat{Q}_\omega] - \hat{Q}_\omega)\right]$ is learnt, which are updated on a slower timescale:

$$\zeta_{k+1} = \zeta_k + \alpha_k^\zeta(r + \gamma\hat{Q}'_{\omega_k} - \hat{Q}_{\omega_k} - v_k^\top\zeta_k)v_k.$$

This approximation is then used to update the function approximator. When a TD estimate is used, this results in two algorithms :

$$\omega_{k+1} = \omega_k + \alpha_k^\omega\left(v_k - \gamma v'_k\right)v_k^\top\zeta_k \quad \text{(GTD2)},$$
$$\omega_{k+1} = \omega_k + \alpha_k^\omega v_k(r + \gamma\hat{Q}'_{\omega_k} - \hat{Q}_{\omega_k}) - \alpha_k^\omega\gamma v'_k v_k^\top\zeta_k \quad \text{(TDC)}.$$

A nonlinear variant of TDC/GTD2 algorithm can be derived for any twice-differentiable function approximation $\hat{Q}_\omega$ by using the basis vector $v = \nabla_\omega\hat{Q}_\omega$ [Maei et al., 2010].

As the TDC/GTD2 algorithm is derived from a frequentist perspective, it does not characterise uncertainty in the MDP. Conversely, our framework allows us to estimate the predictive variance due to state visitation after N updates of GDT2/TDC, which is essential for achieving uncertainty driven, adaptive exploration.

C.4 Linear Policy Evaluation Experiments

We evaluate the linear BBO from Algorithm 5 in a suite of nine policy evaluation tasks, comparing it with six other methods: TD, TDC, GTD2, BRM, and two LSTD variants, which we refer to as LSTD and LSTD+ (see Dann et al. [2014] for details of these

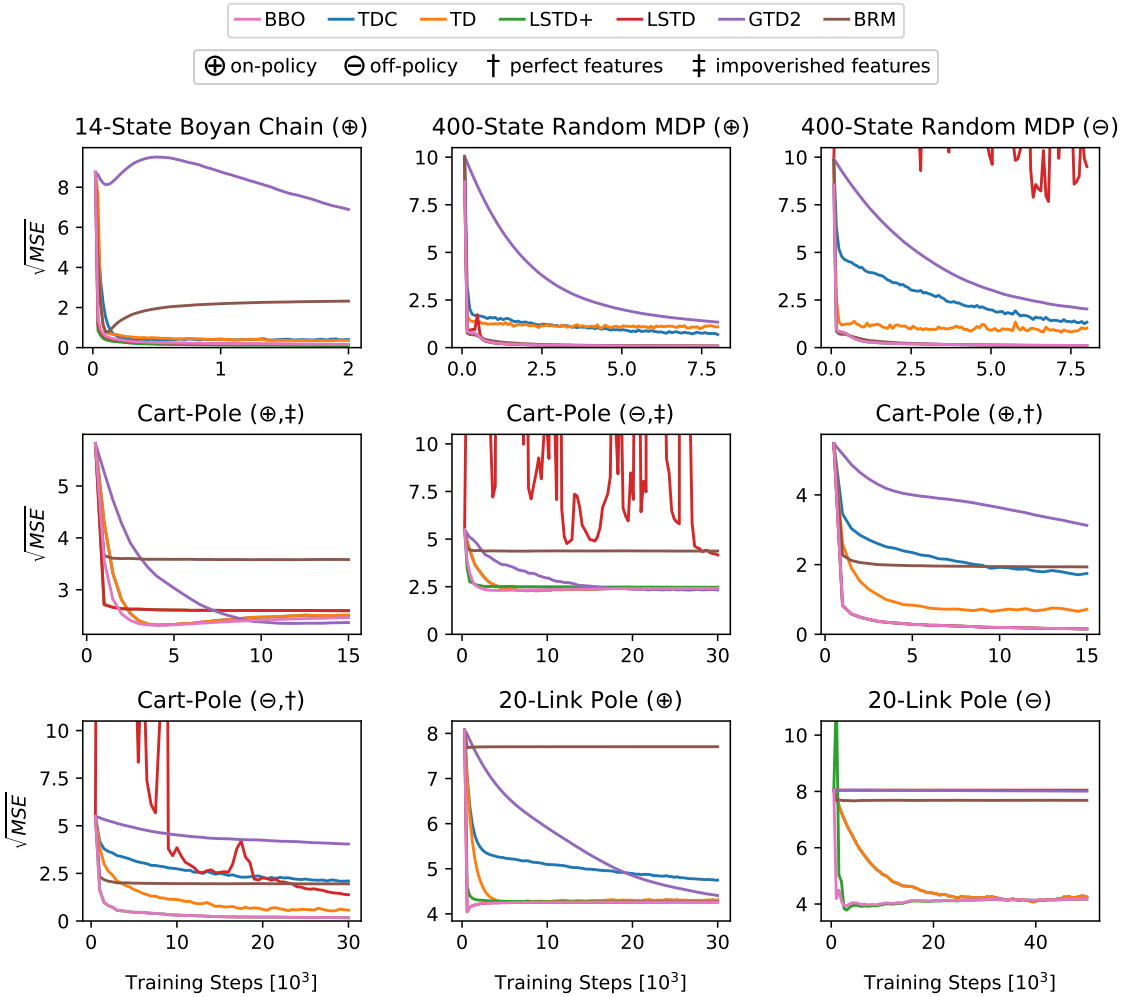


Figure C.1: Linear policy evaluation results.

methods). LSTD corresponds to vanilla version of the algorithm and thus can be seen as the non-regularized frequentist version of BBO, whereas LSTD+ corresponds to LSTD with additional improvements, such as improved off-policy reweighting, eligibility traces, and regularization, as presented in Dann et al. [2014] (Section 3.4). We note that LSTD+ is not directly comparable with BBO, but is rather presented here as a strong baseline. All the environments, evaluated policies, and other training configurations follow exactly those presented by Dann et al. [2014] (Section 3.1). We omit three tasks included in Dann et al. [2014]: the Baird counterexample due to its solution corresponding to 0 weights and thus being unfairly trivial for least-squares methods like BBO and LSTD to solve, and the on- and off-policy Cart-Pole Swingup tasks due to our lack of access to the software required to run the policies.

Table C.1: MSE of final predictions. The values for all methods except for BBO are obtained with code provided by Dann et al. [2014]. ⊕=on-policy, ⊖=off-policy, †=perfect features, ‡=impoverished features.

	BBO	GTD2	TD	TDC	LSTD+	LSTD	BRM
14-State Boyan Chain (⊕)	0.16	6.89	0.36	0.40	**0.10**	0.16	2.32
400-State Random MDP (⊕)	**0.07**	1.34	1.09	0.69	**0.07**	**0.07**	0.08
400-State Random MDP (⊖)	**0.11**	2.03	1.02	1.33	**0.11**	9.50	**0.11**
Cart-Pole (⊕,‡)	2.46	**2.37**	2.51	2.51	2.60	2.60	3.58
Cart-Pole (⊖,‡)	2.42	**2.33**	2.44	2.44	2.47	4.17	4.37
Cart-Pole (⊕, †)	**0.15**	3.13	0.72	1.75	**0.15**	**0.15**	1.93
Cart-Pole (⊖, †)	**0.17**	4.04	0.57	2.10	**0.17**	1.38	1.95
20-Link Pole (⊕)	**4.26**	4.41	4.31	4.75	4.27	**4.26**	7.71
20-Link Pole (⊖)	**4.17**	8.01	4.25	4.25	**4.17**	8.04	7.68

Table C.2: Sum of square root MSE over all timesteps. The values for all methods except for BBO are obtained with code provided by Dann et al. [2014]. ⊕=on-policy, ⊖=off-policy, †=perfect features, ‡=impoverished features.

	BBO	GTD2	TD	TDC	LSTD+	LSTD	BRM
14-State Boyan Chain (⊕)	33.46	841.45	61.33	61.55	**25.06**	32.21	214.35
400-State Random MDP (⊕)	**24.74**	342.56	122.51	119.84	**24.74**	27.75	27.87
400-State Random MDP (⊖)	**29.65**	442.96	113.30	266.36	**29.65**	$> 10^3$	32.13
Cart-Pole (⊕,‡)	**76.99**	89.96	79.65	79.61	81.56	81.56	109.88
Cart-Pole (⊖,‡)	**243.86**	291.02	253.96	253.88	253.53	$> 10^3$	438.94
Cart-Pole (⊕, †)	**13.51**	116.98	31.85	68.75	**13.51**	**13.51**	62.78
Cart-Pole (⊖, †)	24.58	267.92	67.91	159.23	**24.57**	681.19	122.12
20-Link Pole (⊕)	428.98	555.56	441.97	508.58	431.56	**428.97**	770.76
20-Link Pole (⊖)	**415.37**	802.78	470.12	470.20	421.74	804.45	768.59

The training curves are presented in Fig. C.1. We report the cumulative mean squared error in Table C.2 and the final mean squared error in Table C.1 for each algorithm. BBO achieves top final mean squared error in 6 out of 9 tasks, and performs close to the best method with negligible difference in the other tasks. As shown in Table C.2, BBO outperforms other methods in four out of nine tasks in terms of cumulative mean squared error, again performing similarly to top-performing methods across all tasks. This demonstrates the benefits of incorporating priors into policy evaluation models and confirms the converge properties associated with BBO

C.4.1 Hyperparameters

We use the hyperparameters provided by Dann et al. [2014] (Section 3.1.7) for all the baseline algorithms (LSTD, TD, TDC, GTD2, BRM). For BBO, we test the following values for the prior variance: $\{1\mathrm{e}{-1}, 3\mathrm{e}{-1}, 1\mathrm{e}0, 3\mathrm{e}0, 1\mathrm{e}1\}$. In the results reported, we use $1\mathrm{e}{-1}$ for Cart-Pole with impoverished features and $1\mathrm{e}1$ for the rest of the tasks.

D

Experimental Details

D.1 VIREL Continuous Control Experiments

Table D.1: Summary of Experimental Parameter Values

PARAMETER	VALUE
Steps per evaluation	1000
Path Length	999
Discount factor	0.99
Mujoco-v2 Experiments:	
Batch size	128
Net size	300
$\lambda_\beta \approx \frac{1-\gamma}{r_{avg}}$	Humanoid 4e-4 All other 4e-3
Reward scale	Hopper, Half-Cheetah 5 Walker 3 All other 1
Value function learning rate	3e-4
Policy learning rate	3e-4
MLP layout as given in `https://github.com/vitchyr/rlkit`	

D.2 BBO Policy Evaluation Experiments

D.2.1 Tsitsiklis' Triangle Counterexample

Environment and Value Function

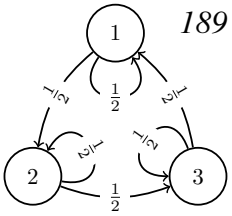


Figure D.1: Tsitsiklis' triangle MDP.

The three-state Tsitsiklis' Triangle MDP Tsitsiklis and Van Roy [1997] is designed to prove divergence of TD methods with nonlinear function approximators. The purpose of this experiment is to empirically validate the convergence properties of nonlinear BBO algorithms. The environment, illustrated in Fig. D.1, consists of the state space $S = \{1, 2, 3\}$ and the action-independent transition kernel

$$P \coloneqq [p(s' = j|\cdot, s = i)]_{i,j} = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

Let $\hat{V}(\omega) \coloneqq [\hat{V}_\omega(s = 1), \hat{V}_\omega(s = 2), \hat{V}_\omega(s = 3)]^\top$ be the value function vector. It can be shown that any $\hat{V}(\omega)$ parametrised by $\omega \in \mathbb{R}$ and satisfying the linear dynamical system

$$\frac{d\hat{V}(\omega)}{d\omega} = (Q + \epsilon I)\,\hat{V}(\omega), \tag{D.1}$$

with the condition that $\hat{V}(0)^\top 1 = 0$ where $\epsilon > 0$ is a small constant and:

$$Q \coloneqq \begin{bmatrix} 1 & \frac{1}{2} & \frac{3}{2} \\ \frac{3}{2} & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{3}{2} & 1 \end{bmatrix},$$

diverges when updated using the TD algorithm [Tsitsiklis and Van Roy, 1997]. To be consistent with the TDC and GTD2 algorithms presented in Bhatnagar et al. [2009] we choose $\hat{V}(\omega) = \exp(\epsilon\omega)(a\cos(\lambda\omega) - b\sin(\lambda\omega))$ with $a = (-14.9996, -35.0002, 50.0004)$, $b = (-49.0753, 37.5278, 11.5469)$, $\lambda = \sqrt{3}/2$, and $\epsilon = 10^{-2}$ as a solution to Eq. (D.1) (using the values that we received from the authors' implementation), we use the discount $\gamma = 0.9$ and normalize the gradient steps to stabilize the updates. Each update batch includes all 6 environment transitions.

Hyperparameters

We search over the learning rates for all the algorithms. For TDC, and GTD2, we begin with a coarse grid search with values of $\{1e{-}3, 1e{-}2, 1e{-}1, 1e0\}$, followed by finer search of $\{1e{-}1, 2e{-}1, ..., 9e{-}1, 1e0\}$. For TD(0), we manually search around the learning rate used in Bhatnagar et al. [2009]. We do not search over BBO's prior loss weight and set to 1.0. The final hyperparameters are presented in Table D.2.

Table D.2: Hyperparameters for reported Tsitsiklis Triangle experiments.

Method	Parameter	Value
BBO	Lower-level learning rate	8e-1
	Upper-level learning rate	1e-1
TD(0)	Learning rate	2e-3
TDC	Fast timescale learning rate	1e0
	Slow timescale learning rate	1e-1
GTD2	Fast timescale learning rate	8e-1
	Slow timescale learning rate	1e-1

Results

As can be seen in Fig. 7.3, BBO converges to optimal solution similarly to prior convergent nonlinear methods, TDC and GTD2 Bhatnagar et al. [2009], while TD(0), as expected, diverges. These results verify the convergence properties of the proposed nonlinear BBO algorithms.

D.2.2 Neural Network Function Approximators

Environments

We consider three environments commonly used in policy evaluation literature: 20-Link Pendulum [Dann et al., 2014] with 40D continuous observation space, Puddle World [Boyan and Moore, 1995] with 2D continuous observation space, and the continuous variant of Mountain Car [Boyan and Moore, 1995][1] with 2D continuous observation space.

[1] We use the MountainCar-Continuous-v0 implementation from the OpenAI Gym suite [Brockman et al., 2016]

Datasets

The datasets used for evaluating the policies consist of 20000, 20000, and 30000 on-policy transitions for Puddle World, Mountain Car, and 20-Link Pendulum, respectively. For Puddle World and Mountain Car, each transition is sampled independently by resetting the state uniformly at random in the state space after each transition. Puddle World and Mountain Car observations are normalized to range $[-1, 1]$.

Policies. For Puddle World and Mountain Car experiments, we run the policy evaluation using a simple, non-optimal policies. In Puddle World, the policy selects either up or down action uniformly at random. In Mountain Car, the right action is chosen when velocity > 0, and otherwise the left action. The 20-Link Pendulum experiments, on the other hand, use an optimal policy obtained with dynamic programming, similar to the policy used in the corresponding linear experiments. See Dann et al. [2014] for details.

Ground-truth Value Functions. For Puddle World and Mountain Car, the ground-truth value function, for which the mean squared errors are computed, is obtained for a set of 625 evenly-spaced states (25×25 grid) in the 2D state space. We reset the agent to each of the 625 states 1000 times, rolling it out and computing the cumulative sum of rewards for up to 1000 steps, finally averaging over the 1000 resets. The 20-Link Pendulum environment is a linear-quadratic MDP, for which we obtain the exact values for 5000 states, as is done by Dann et al. [2014]. All the value functions are computed with the same discount factor that is used for training the approximate value functions.

Network Structure

All the experiments use a single-layer feedforward network with hidden layer size of 256. TDC uses a $tanh$ activation whereas other algorithms use $relu$.

Training Details

The discount factor is set to 0.98 for all tasks. The optimization is carried out with Adam optimizer [Kingma and Ba, 2014] using randomly sampled mini-batches of size 512. The hyperparameter searches are done using 3 datasets and the final results are reported are averaged over 24 separate datasets and seeds. Each full trial (100k steps) run takes about 15 minutes for BBO variants, 7 minutes for TD(0), and 85 minutes for TDC on standard desktop machine.

Hyperparameters

We use the same grid search for the hyperparameters across all environments. The hyperparameters with their evaluated values for each algorithm are presented in Table D.3 for BBO, Table D.4 for TD(0), and Table D.5 for TDC. For BBO, instead of searching over the full grid at once, we first perform a coarse grid search for learning rates without using a prior, then do another denser search around the best values (with values still given in the Table D.3), and finally perform a grid search over the lower-level gradient steps per training steps and prior values (when applicable). For TDC and TD(0), we perform a full grid search over the listed values.

Table D.3: BBO hyperparameter grid

Hyperparameter	Grid values
Learning rates	{ 1e−6, 3e−6, 1e−5, …, 1e−1, 3e−1, 1e0 }
Weight for the prior loss $(1/\sigma_0^2)^*$	{ 1e−4, 3e−4, 1e−3, …, 1e−1, 3e−1, 1e0 }
Lower-level steps / training step	$\{1, 5, 10, 20\}$

* For 20-Link Pendulum, we also include { 2.5e−1, 5e−1, 7.5e−1 }.

Table D.4: TD(0) hyperparameter grid

Hyperparameter	Grid values
Learning rate*	{ 1e−6, 3e−6, 1e−5, …, 1e−1, 3e−1, 1e0 }

* For 20-Link Pendulum, we also include { 1e−7, 3e−7 }.

Table D.5: TDC hyperparameter grid

Hyperparameter	Grid values
Fast timescale learning rate	{ 1e−6, 3e−6, 1e−5, . . . , 1e−1, 3e−1, 1e0 }
Slow timescale learning rate	{ 1e−6, 3e−6, 1e−5, . . . , 1e−1, 3e−1, 1e0 }

Table D.6: Hyperparameters for reported nonlinear policy evaluation results.

Method	Parameter	Environment		
		20-Link Pendulum	Puddle World	Mountain Car
Gradient BBO w/ prior	Upper-level learning rate	1e-2	3e-2	1e-2
	Lower-level learning rate	1e-3	1e-2	3e-3
	Prior loss weight	5e-1	3e-4	1e-1
	Lower-level steps / training step	20	10	10
Gradient BBO w/o prior	Upper-level learning rate	3e-4	1e-3	1e-2
	Lower-level learning rate	1e-3	1e-2	3e-3
	Lower-level steps / training step	20	10	10
Direct BBO w/ prior	Learning rate	3e-7	1e-3	3e-4
	Prior loss weight	1.0	3e-4	1e-1
TD(0)	Learning rate	3e-7	3e-3	3e-4
TDC	Fast timescale learning rate	3e-3	1e-3	1e-2
	Slow timescale learning rate	1e-4	1e-3	1e-5

The final hyperparameters for each method and task are presented in Table D.6. The values are chosen by manually picking the best-performing set from the hyperparameter search based on the final MSE, and the final results are reported for 24 holdout datasets.

D.3 BBO Continuous Control Experiments

This section provides further details and analysis of the BBAC algorithm and control experiments. We refer to $\lambda := \frac{1}{\sigma_0^2}$ as the regularisation weight and σ^2 as the prior scale. We first investigate BBAC's behavior in the *MountainCar-Continuous-v0* task, where the environment's simplicity and low-dimensionality allows us to visually analyze the behavior of the randomized prior ensemble. We then analyze BBAC's sensitivity to

randomized prior hyperaparameters in a slightly modified version of DeepMind Control Suite's [Tassa et al., 2018] *cartpole-swingup_sparse* environment.

To improve the exploration capability of SAC in our challenging domains, we also introduce a variant of SAC called SAC* which uses a single Q-function to avoid *pessimistic underexploration* [Ciosek et al., 2019b]. Here, instead of learning two soft Q-function approximators independently and choosing the minimum for the actor and critic gradient updates as specified by SAC, we train a single soft Q-function and use updates (6) and (13) of Haarnoja et al. [2018a] directly. This also reduces the inductive bias that SAC has for solving tasks with dense reward structures, making the comparison against BBAC fairer.

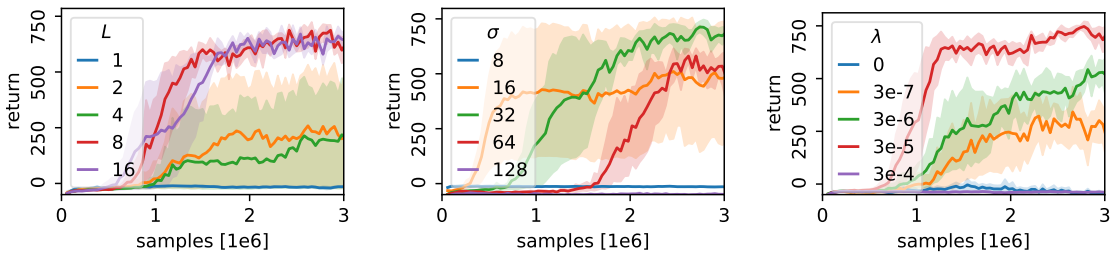
All the experiments in this section use the same hyperparameters from Table D.7. The policies and function approximators are parameterized as fully-connected neural networks.

Table D.7: Common Hyperparameters for BBAC and SAC.

Hyperparameter	Value
Optimizer	Adam
Learning rates	$3\mathrm{e}{-4}$
Discount	0.99
Replay buffer size	$1\mathrm{e}6$
Nonlinearity	ReLU
Hidden layers	2
Hidden units/layer	256
Batch size	256
Target smoothing coefficient	$5\mathrm{e}{-3}$
Training steps per environment step	1

D.3.1 BBAC MountainCar-Continuous-v0

In order to qualitatively measure how the agents explore the state space, we analyze the evolution of critics and coverage of state space throughout learning. We use the standard MountainCar-Continuous-v0 implementation from the OpenAI Gym suite [Brockman et al., 2016], and run each of the algorithms for $1e5$ timesteps, while checkpointing the replay pool and the critics on pre-specified intervals.



(a) Ensemble size L. λ=3e−5, σ=32. **(b)** Prior scale σ. L=8, λ=3e−5. **(c)** Regularization weight λ. L=8, σ=32.

Figure D.2: Evaluation of BBAC's sensitivity to randomized prior hyperparameters. We vary (a) ensemble size L, (b) prior scale σ, and (c) regularization weight λ, while keeping other hyperparameters fixed.

For the final SAC runs, we set the target entropy using the heuristic provided in [Haarnoja et al., 2018b] which is the negative number of action dimensions, i.e. -1 in this case. We also tested target entropy values $\{-8, -4, -2, 0, \frac{1}{2}, 1\}$, all of which behaved similarly as the default target. From the 35 runs with these 7 different entropies, a total of 3 seeds were able to achieve the goal. As our experiments are carried out in a tabula rasa setting where the prior parameters are drawn as randomly initialized neural networks, we find that the choice of prior parameters affects the speed with which BBAC solves the tasks. BBAC is relatively insensitive to the randomized prior hyperparameters in this environment, as long as the ensemble size $L \geq 4$, and for the final results we use ensemble size $L = 8$, prior scale $\sigma = 100$, and regularization weight $\lambda = 3\text{e}{-5}$.

D.3.2 BBAC cartpole

We further investigate BBAC's sensitivity to the choice of randomized prior hyperparameters – ensemble size (L), prior scale (σ), and regularization weight (λ) – in a slightly modified version of DeepMind Control Suite's [Tassa et al., 2018] *cartpole-swingup_sparse* domain, as a continuous analog to the one presented in [Osband et al., 2019].

The original *cartpole-swingup_sparse* is modified such that the reward function includes a control cost of $0.1a_t$ for action a_t on each timestep t, and the agent receives a positive reward of 1 only when both the cart is controllably centered and the pole is controllably upright. That is, the reward function $r(s_t, a_t)$ for action a_t at state $s_t = (\cos(\theta_t), \sin(\theta_t), \dot{\theta}_t, x_t, \dot{x}_t)$ is given by:

$$r(s_t, a_t) = -0.1|a_t| + (\mathbb{1}(|x_t| < 0.1) \cdot \mathbb{1}(0.95 < \cos(\theta_t)) \cdot \mathbb{1}(|\dot{x}_t| < 1) \cdot \mathbb{1}(|\dot{\theta}_t| < 1)).$$

The pole is initialized to a stationary downright position, and the motor is not strong enough to turn the pole upright on a single pass, meaning that in order to reach the goal, the agent has to build momentum by moving the cart and swinging the pole back and forth. This requires executing costly actions for more than a hundred steps, making the exploration problem non-trivial.

The results for BBAC's hyperparameter sensitivity are presented in Fig. D.2. Increasing the ensemble size (L; Fig. D.2a) improves the likelihood of solving the task and, as expected, this effect plateaus and the task can be consistently solved with $L \geq 8$. In the case of both prior scale (σ; Fig. D.2b) and regularization weight (λ; Fig. D.2c), too small values limit the exploration and there is a sweet spot where task performance and exploration are well-balanced. While higher values of prior scale are unable to solve the task within the $3e6$ time steps shown here, we expect them to eventually converge to the optimal solution, whereas higher values of regularization weight are likely to constrain the learning too much to allow convergence.

All our experiments are carried out in a tabula rasa setting where the prior functions are drawn as randomly-initialized neural networks, which is why the effect of the hyperparameter choice to the speed of learning is expected. The range of working hyperparameters is relatively wide and easy to tune, however, and in a real-world scenarios, we might have access to prior knowledge on the task, for example through transfer learning, which would further simplify the choice of BBAC's hyperparameters.

Bibliography

Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. Number July, pages 30–37. 1995. doi: 10.1.1.48.3256.

J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, May 1997. ISSN 2334-3303. doi: 10.1109/9.580874.

Hado van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2094–2100. AAAI Press, 2016.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. *International Conference on Machine Learning*, abs/1802.09477, 2018a.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018. URL http://incompleteideas.net/book/the-book-2nd.html.

Sergey Levine. Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review. 2018a. URL https://arxiv.org/pdf/1805.00909.pdf.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1861–1870, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018a. PMLR. URL http://proceedings.mlr.press/v80/haarnoja18b.html.

A. W. van der Vaart. *Bayes Procedures*, page 138–152. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. doi: 10.1017/CBO9780511802256.011.

Leonard J. Savage. *The Foundations of Statistics*. Wiley Publications in Statistics, 1954.

Richard Bellman. A problem in the sequential design of experiments. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 16(3/4):221–229, 1956. ISSN 00364452. URL http://www.jstor.org/stable/25048278.

Michael O'Gordon Duff. *Optimal Learning: Computational Procedures for Bayes-Adaptive Markov Decision Processes*. PhD thesis, 2002. AAI3039353.

Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 8617–8629. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/8080-randomized-prior-functions-for-deep-reinforcement-learning.pdf.

Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. Varibad: A very good method for bayes-adaptive deep rl via meta-learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/pdf?id=Hkl9JlBYvr.

Abraham Wald. *Statistical decision functions*. Wiley publications in statistics. John Wiley & Sons, New York : London, 1950.

Chao Yu, Jiming Liu, and Shamim Nemati. Reinforcement learning in healthcare: a survey. *arXiv preprint arXiv:1908.08796*, 2019.

B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A. Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey, 2020.

Y. Deng, F. Bao, Y. Kong, Z. Ren, and Q. Dai. Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems*, 28(3):653–664, 2017.

Zhengyao Jiang, Dixing Xu, and Jinjun Liang. A deep reinforcement learning framework for the financial portfolio management problem. 06 2017.

Thomas Spooner, Rahul Savani, John Fearnley, and Andreas Koukorinis. Market making via reinforcement learning. In *17th International Conference on Autonomous Agents and Multiagent Systems*, 07 2018.

Adam Smith and Janna Anderson. Ai, robotics, and the future of jobs. 2017.

Daron Acemoglu and Pascual Restrepo. Artificial intelligence, automation, and work. In *The Economics of Artificial Intelligence: An Agenda*, pages 197–236. National Bureau of Economic Research, Inc, 2018. URL https://EconPapers.repec.org/RePEc:nbr:nberch:14027.

Nick Srnicek and Alex Williams. *Inventing the future: postcapitalism and a world without work*. Verso, 2015. ISBN 9781784780968.

Adam Greenfield. *Radical Technologies: The Design of Everyday Life*. Verso, 2018. ISBN 1784780456.

Daron Acemoglu and Pascual Restrepo. Unpacking skill bias: Automation and new tasks. *AEA Papers and Proceedings*, 110:356–61, May 2020. doi: 10.1257/pandp.20201063. URL https://www.aeaweb.org/articles?id=10.1257/pandp.20201063.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. 08 2019.

Patrick Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1999. ISBN 0-471-19745-9. A Wiley-Interscience Publication.

K.L. Chung and K. Zhong. *A Course in Probability Theory*. Elsevier Science, 2001. ISBN 9780121741518. URL https://books.google.co.uk/books?id=zu80w7wd40UC.

S.T. Rachev. *Probability Metrics and the Stability of Stochastic Models*. 1991. ISBN 9780471928775.

B.J.K. Kleijn and A.W. van der Vaart. The bernstein-von-mises theorem under misspecification. *Electron. J. Statist.*, 6:354–381, 2012. doi: 10.1214/12-EJS675. URL https://doi.org/10.1214/12-EJS675.

Csaba Szepesvári. Algorithms for Reinforcement Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 4(1):1–103, 2010. ISSN 1939-4608. doi: 10.2200/S00268ED1V01Y201005AIM009. URL http://www.morganclaypool.com/doi/abs/10.2200/S00268ED1V01Y201005AIM009.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Bert Fristedt and Lawrence F Gray. *A modern approach to probability theory*. Probability and its applications. Birkhäuser, Boston, 1997. ISBN 9780817638078.

Cameron Browne, Edward Powley, Daniel Whitehouse, Simon Lucas, Peter I. Cowling, Stephen Tavener, Diego Perez, Spyridon Samothrakis, Simon Colton, and et al. A survey of monte carlo tree search methods. *IEEE TRANSACTIONS ON COMPUTATIONAL INTELLIGENCE AND AI*, 2012.

Maciej Świechowski, Konrad Godlewski, Bartosz Sawicki, and Jacek Mańdziuk. Monte carlo tree search: A review on recent modifications and applications. *arXiv Preprint*, 03 2021.

L. Song, K. Fukumizu, and A. Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013. doi: 10.1109/MSP.2013.2252713.

Harm Vanseijen and Rich Sutton. A deeper look at planning as learning from replay. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2314–2322, Lille, France, 07–09 Jul 2015. PMLR. URL http://proceedings.mlr.press/v37/vanseijen15.html.

Long Ji Lin. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Mach. Learn.*, 8:293–321, 1992. doi: 10.1007/BF00992699. URL https://doi.org/10.1007/BF00992699.

Thomas M. Moerland, Joost Broekens, and Catholijn M. Jonker. Model-based reinforcement learning: A survey. *CoRR*, abs/2006.16712, 2020. URL https://arxiv.org/abs/2006.16712.

Christopher John Cornish Hellaby Watkins. *Learning from Delayed Rewards*. PhD thesis, King's College, University of Cambridge, Cambridge, UK, May 1989. URL http://www.cs.rhul.ac.uk/~chrisw/new_thesis.pdf.

Christopher J. C. H. Watkins and Peter Dayan. Technical note: q -learning. *Mach. Learn.*, 8(3–4):279–292, May 1992. ISSN 0885-6125. doi: 10.1007/BF00992698. URL https://doi.org/10.1007/BF00992698.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, February 2015. ISSN 00280836. URL http://dx.doi.org/10.1038/nature14236.

D. Hochbaum. Complexity and algorithms for nonlinear optimization problems. *Annals of Operations Research*, 153:257–296, 2007.

Ronald J. Williams, Leemon C. Baird, and III. Analysis of some incremental variants of policy iteration: First steps toward understanding actor-critic learning systems, 1993.

Richard S Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. A. Solla, T. K. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 1057–1063. MIT Press, 2000. URL http://papers.nips.cc/paper/1713-policy-gradient-methods-for-reinforcement-learning-with-function-approximation.pdf.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://dblp.uni-trier.de/db/journals/corr/corr1707.html#SchulmanWDRK17.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1889–1897, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/schulman15.html.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1928–1937, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/mniha16.html.

Yuhuai Wu, Elman Mansimov, Roger B. Grosse, Shun Liao, and Jimmy Ba. Scalable trust-region method for deep reinforcement learning using kronecker-factored approximation. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *NIPS*, pages 5279–5288, 2017. URL http://dblp.uni-trier.de/db/conf/nips/nips2017.html#WuMGLB17.

Ziyu Wang, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Rémi Munos, Koray Kavukcuoglu, and Nando de Freitas. Sample efficient actor-critic with experience replay. *CoRR*, abs/1611.01224, 2016. URL http://dblp.uni-trier.de/db/journals/corr/corr1611.html#WangBHMMKF16.

Chris Nota and Philip S. Thomas. Is the policy gradient a gradient? In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, page 939–947, Richland, SC, 2020. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450375184.

Vijay R. Konda and John N. Tsitsiklis. Actor-critic algorithms. In S. A. Solla, T. K. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 1008–1014. MIT Press, 2000. URL http://papers.nips.cc/paper/1786-actor-critic-algorithms.pdf.

Ian Osband, Benjamin Van Roy, Daniel J. Russo, and Zheng Wen. Deep exploration via randomized value functions. *Journal of Machine Learning Research*, 20(124):1–62, 2019. URL http://jmlr.org/papers/v20/18-339.html.

Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, Aug 1988. ISSN 1573-0565. doi: 10.1007/BF00115009. URL https://doi.org/10.1007/BF00115009.

Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951. doi: 10.1214/aoms/1177729586. URL https://doi.org/10.1214/aoms/1177729586.

Peter Dayan. The convergence of td(λ) for general λ. *Mach. Learn.*, 8(3–4):341–362, May 1992. ISSN 0885-6125. doi: 10.1007/BF00992701. URL https://doi.org/10.1007/BF00992701.

T. E. Harris. *The Existence of Stationary Measures for Certain Markov Processes*, pages 113–124. University of California Press, 1956. doi: doi:10.1525/9780520350670-011. URL https://doi.org/10.1525/9780520350670-011.

W.R. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in Practice*, chapter Introduction to General State-Space Markov Chain Theory. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis, 1995. ISBN 9780412055515.

Sean Meyn, Richard L. Tweedie, and Peter W. Glynn. *Markov Chains and Stochastic Stability*. Cambridge Mathematical Library. Cambridge University Press, 2 edition, 2009. doi: 10.1017/CBO9780511626630.

V. S. Borkar and S. P. Meyn. The o.d. e. method for convergence of stochastic approximation and reinforcement learning. *SIAM J. Control Optim.*, 38(2):447–469, January 2000. ISSN 0363-0129. doi: 10.1137/S0363012997331639. URL https://doi.org/10.1137/S0363012997331639.

Vivek Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. 01 2008. ISBN 978-81-85931-85-2. doi: 10.1007/978-93-86279-38-5.

John N. Tsitsiklis and Benjamin Van Roy. Feature-based methods for large scale dynamic programming. In *Machine Learning*, pages 59–94, 1994.

Geoffrey J. Gordon. Stable function approximation in dynamic programming. In Armand Prieditis and Stuart Russell, editors, *Machine Learning Proceedings 1995*, pages 261–268. Morgan Kaufmann, San Francisco (CA), 1995. ISBN 978-1-55860-377-6. doi: https://doi.org/10.1016/B978-1-55860-377-6.50040-2. URL https://www.sciencedirect.com/science/article/pii/B9781558603776500402.

Justin Boyan and Andrew Moore. Generalization in reinforcement learning: Safely approximating the value function. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7. MIT Press, 1995. URL https://proceedings.neurips.cc/paper/1994/file/ef50c335cca9f340bde656363ebd02fd-Paper.pdf.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1352–1361, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/haarnoja17a.html.

Sergey Levine. Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review. 2018b. URL https://arxiv.org/pdf/1805.00909.pdf.

Allan Birnbaum. On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298):269–306, 1962. doi: 10.1080/01621459.1962.10480660. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1962.10480660.

Greg Gandenberger. A new proof of the likelihood principle. *The British Journal for the Philosophy of Science*, 66(3):475–503, 2015. doi: 10.1093/bjps/axt039. URL https://doi.org/10.1093/bjps/axt039.

D. V. Lindley and L. D. Phillips. Inference for a bernoulli process (a bayesian view). *The American Statistician*, 30(3):112–119, 1976. ISSN 00031305. URL http://www.jstor.org/stable/2683855.

Morris H. DeGroot. *Conjugate Prior Distributions*. John Wiley and Sons, Ltd, 2004. ISBN 9780471729006.

Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*, chapter 7. The MIT Press, 2012. ISBN 0262018020, 9780262018029.

Ward Edwards, Harold Lindman, and Leonard J. Savage. Bayesian statistical inference for psychological research. *Psychological Review*, 70:193—-242, 1963.

Michael I. Jordan, editor. *Learning in Graphical Models*. MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-60032-3.

Matthew James Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, 2003.

Samuel J. Gershman, Matthew D. Hoffman, and David M. Blei. Nonparametric variational inference.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

C F Jeff Wu. On the Convergence Properties of the EM Algorithm'. *Source: The Annals of Statistics The Annals of Statistics*, 11(1):95–103, 1983.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1530–1538. JMLR, 2015.

Matthew Fellows, Anuj Mahajan, Tim G. J. Rudner, and Shimon Whiteson. Virel: A variational inference framework for reinforcement learning. In *Advances in Neural Information Processing Systems 32*, pages 7120–7134. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8934-virel-a-variational-inference-framework-for-reinforcement-learning.pdf.

Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Remi Munos, Nicolas Heess, and Martin Riedmiller. Maximum a posteriori policy optimisation. In *International Conference on Learning Representations*, 2018a. URL https://openreview.net/forum?id=S1ANxQW0b.

T. Haarnoja, Aurick Zhou, Sehoon Ha, J. Tan, G. Tucker, and S. Levine. Learning to walk via deep reinforcement learning. *ArXiv Preprint*, abs/1812.11103, 2019.

Peter Dayan and Geoffrey E. Hinton. Using expectation-maximization for reinforcement learning. *Neural Computation*, 9(2):271–278, 1997. doi: 10.1162/neco.1997.9.2.271. URL https://doi.org/10.1162/neco.1997.9.2.271.

Marc Toussaint and Amos Storkey. Probabilistic inference for solving discrete and continuous state Markov Decision Processes. *Proceedings of the 23rd international conference on Machine learning - ICML '06*, pages 945–952, 2006. doi: 10.1145/1143844.1143963. URL http://portal.acm.org/citation.cfm?doid=1143844.1143963.

Jan Peters and Stefan Schaal. Reinforcement learning by reward-weighted regression for operational space control. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 745–750, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-793-3. doi: 10.1145/1273496.1273590. URL http://doi.acm.org/10.1145/1273496.1273590.

Hirotaka Hachiya, Jan Peters, and Masashi Sugiyama. Efficient sample reuse in em-based policy search. In Wray Buntine, Marko Grobelnik, Dunja Mladenić, and John Shawe-Taylor, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 469–484, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. ISBN 978-3-642-04180-8.

Gerhard Neumann. Variational inference for policy search in changing situations. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 817–824, USA, 2011. Omnipress. ISBN 978-1-4503-0619-5. URL http://dl.acm.org/citation.cfm?id=3104482.3104585.

Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Remi Munos, Nicolas Heess, and Martin Riedmiller. Maximum a posteriori policy optimisation. In *International Conference on Learning Representations*, 2018b. URL https://openreview.net/forum?id=S1ANxQW0b.

Marc Toussaint. Probabilistic inference as a model of planned behavior. *Kunstliche Intelligenz*, 3, 01 2009.

Konrad Rawlik, Marc Toussaint, and Sethu Vijayakumar. On stochastic optimal control and reinforcement learning by approximate inference. In *Robotics: Science and Systems*, 2012.

Xinyue Chen, Che Wang, Zijian Zhou, and Keith W. Ross. Randomized ensembled double q-learning: Learning fast without a model. *The Ninth International Conference on Learning Representations (ICLR)*, abs/2101.05982, 2021.

Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, AAAI'08, pages 1433–1438. AAAI Press, 2008. ISBN 978-1-57735-368-3. URL http://dl.acm.org/citation.cfm?id=1620270.1620297.

Brian D. Ziebart, J. Andrew Bagnell, and Anind K. Dey. Modeling interaction via the principle of maximum causal entropy. In *Proceedings of the 27th International*

Conference on International Conference on Machine Learning, ICML'10, pages 1255–1262, USA, 2010. Omnipress. ISBN 978-1-60558-907-7. URL http://dl.acm.org/citation.cfm?id=3104322.3104481.

Brian D Ziebart. *Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy*. PhD thesis, 2010. URL http://www.cs.cmu.edu/{~}bziebart/publications/thesis-bziebart.pdf.

Roy Fox, Ari Pakman, and Naftali Tishby. Taming the noise in reinforcement learning via soft updates. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, pages 202–211, Arlington, Virginia, United States, 2016. AUAI Press. ISBN 978-0-9966431-1-5. URL http://dl.acm.org/citation.cfm?id=3020948.3020970.

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Soft actor-critic algorithms and applications. *CoRR*, abs/1812.05905, 2018b.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Konrad Rawlik, Marc Toussaint, and Sethu Vijayakumar. Approximate inference and stochastic optimal control. *CoRR*, abs/1009.3958, 2010. URL http://arxiv.org/abs/1009.3958.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1587–1596, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018b. PMLR. URL http://proceedings.mlr.press/v80/fujimoto18a.html.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2014. URL http://dblp.uni-trier.de/db/conf/iclr/iclr2014.html#KingmaW13.

Nino Vieillard, Tadashi Kozuno, B. Scherrer, O. Pietquin, Rémi Munos, and M. Geist. Leverage the average: an analysis of regularization in rl. *Advances in Neural Information Processing Systems*, 33, 2020.

Jose M Bernardo and Adrian Smith. *Bayesian Theory*. Wiley series in probability and statistics. John Wiley & Sons Ltd., Chichester, 2000. URL https://cds.cern.ch/record/1319894.

Abraham Wald. An Essentially Complete Class of Admissible Decision Functions. *The Annals of Mathematical Statistics*, 18(4):549 – 555, 1947. doi: 10.1214/aoms/1177730345. URL https://doi.org/10.1214/aoms/1177730345.

Thomas S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209 – 230, 1973. doi: 10.1214/aos/1176342360. URL https://doi.org/10.1214/aos/1176342360.

Alexander Terenin and David Draper. A noninformative prior on a space of distribution functions. *Entropy*, 19:391, 2017.

Yee Whye Teh. *Dirichlet Process*, pages 280–287. Springer US, Boston, MA, 2010. ISBN 978-0-387-30164-8. doi: 10.1007/978-0-387-30164-8_219. URL https://doi.org/10.1007/978-0-387-30164-8_219.

Steffen Grünewälder, Guy Lever, Luca Baldassarre, Massimilano Pontil, and Arthur Gretton. Modelling transition dynamics in mdps with rkhs embeddings. In *Proceedings of the 29th International Coference on International Conference on Machine*

Learning, ICML'12, page 1603–1610, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Larry Wasserman. *All of Nonparametric Statistics (Springer Texts in Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387251456.

Richard Dearden, Nir Friedman, and David Andre. Model based bayesian exploration. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, page 150–159, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1558606149.

Nikos Vlassis, Mohammad Ghavamzadeh, Shie Mannor, and Pascal Poupart. *Bayesian Reinforcement Learning*, pages 359–386. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-27645-3. doi: 10.1007/978-3-642-27645-3_11. URL https://doi.org/10.1007/978-3-642-27645-3_11.

M. Ghavamzadeh, S. Mannor, J. Pineau, and A. Tamar. *Bayesian Reinforcement Learning: A Survey*. now, 2015. ISBN null. URL https://ieeexplore.ieee.org/document/8187142.

John Asmuth and Michael Littman. Learning is planning: near bayes-optimal reinforcement learning via monte-carlo tree search. pages 19–26, 01 2011.

Arthur Guez, David Silver, and Peter Dayan. Efficient bayes-adaptive reinforcement learning using sample-based search. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/35051070e572e47d2c26c241ab88307f-Paper.pdf.

Arthur Guez, David Silver, and Peter Dayan. Scalable and efficient bayes-adaptive reinforcement learning based on monte-carlo tree search. *Journal of Artificial Intelligence Research*, 48:841–883, 10 2013. doi: 10.1613/jair.4117.

Pascal Poupart, Nikos Vlassis, Jesse Hoey, and Kevin Regan. An analytic solution to discrete bayesian reinforcement learning. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 697–704, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143932. URL https://doi.org/10.1145/1143844.1143932.

Emma Brunskill. Bayes-optimal reinforcement learning for discrete uncertainty domains. In *In AAMAS '12*, 2012.

William R Thomson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 12 1933. ISSN 0006-3444. doi: 10.1093/biomet/25.3-4.285. URL https://doi.org/10.1093/biomet/25.3-4.285.

Malcolm Strens. A bayesian framework for reinforcement learning. In *In Proceedings of the Seventeenth International Conference on Machine Learning*, pages 943–950. ICML, 2000.

Ian Osband, Benjamin Van Roy, and Daniel Russo. (more) efficient reinforcement learning via posterior sampling. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3003–3011, Red Hook, NY, USA, 2013. Curran Associates Inc.

Leslie Pack Kaelbling. Associative reinforcement learning: Functions in k-dnf. *Mach. Learn.*, 15(3):279–298, June 1994. ISSN 0885-6125. doi: 10.1023/A:1022689909846. URL https://doi.org/10.1023/A:1022689909846.

Tao Wang, Daniel Lizotte, Michael Bowling, and Dale Schuurmans. Bayesian sparse sampling for on-line reward optimization. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 956–963, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595931805. doi: 10.1145/1102351.1102472. URL https://doi.org/10.1145/1102351.1102472.

J. Kolter and Andrew Ng. Near-bayesian exploration in polynomial time. volume 382, page 65, 01 2009. doi: 10.1145/1553374.1553441.

Kamil Ciosek, Quan Vuong, Robert Loftin, and Katja Hofmann. Better exploration with optimistic actor critic. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a. URL https://proceedings.neurips.cc/paper/2019/file/a34bacf839b923770b2c360eefa26748-Paper.pdf.

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2701–2710. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/osband17a.html.

Yaakov Engel, Shie Mannor, and Ron Meir. Reinforcement learning with gaussian processes. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 201–208, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595931805. doi: 10.1145/1102351.1102377. URL https://doi.org/10.1145/1102351.1102377.

Mohammad Ghavamzadeh, Yaakov Engel, and Michal Valko. Bayesian policy gradient and actor-critic algorithms. *Journal of Machine Learning Research*, 17(66):1–53, 2016. URL http://jmlr.org/papers/v17/10-245.html.

Girish Chowdhary, Miao Liu, Robert Grande, Thomas Walsh, and Jonathan How. Off-policy reinforcement learning with gaussian processes. volume 1, 01 2013. doi: 10.1109/JAS.2014.7004680.

Martin Riedmiller. Neural fitted q iteration – first experiences with a data efficient neural reinforcement learning method. In João Gama, Rui Camacho, Pavel B. Brazdil, Alípio Mário Jorge, and Luís Torgo, editors, *Machine Learning: ECML 2005*, pages 317–328, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-31692-3.

Kamyar Azizzadenesheli, Emma Brunskill, and Animashree Anandkumar. Efficient exploration through bayesian deep q-networks. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–9, 2018. doi: 10.1109/ITA.2018.8503252.

David Janz, Jiri Hron, Przemysław Mazur, Katja Hofmann, José Miguel Hernández-Lobato, and Sebastian Tschiatschek. *Successor Uncertainties: Exploration and Uncertainty in Temporal Difference Learning*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1050–1059. JMLR.org, 2016.

Yunhao Tang and Alp Kucukelbir. Variational deep q network. *ArXiv Preprint*, 11 2017.

Zachary Lipton, Xiujun Li, Jianfeng Gao, Lihong Li, Faisal Ahmed, and li Deng. Bbq-networks: Efficient exploration in deep reinforcement learning for task-oriented dialogue systems. 11 2018.

Ahmed Touati, Harsh Satija, Joshua Romoff, Joelle Pineau, and Pascal Vincent. Randomized value functions via multiplicative normalizing flows. In Amir Globerson and Ricardo Silva, editors, *UAI*, page 156. AUAI Press, 2019.

Brendan O'Donoghue, Ian Osband, Remi Munos, and Vlad Mnih. The uncertainty Bellman equation and exploration. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3839–3848, Stockholmsmässan, Stockholm

Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/odonoghue18a.html.

Matthew Fellows, Kristian Hartikainen, and Shimon Whiteson. Bayesian bellman operators. In *Advances in Neural Information Processing Systems 34*. Curran Associates, Inc., 2021.

Richard S Sutton, Hamid R. Maei, and Csaba Szepesvári. A convergent o(n) temporal-difference algorithm for off-policy learning with linear function approximation. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1609–1616. Curran Associates, Inc., 2009a.

Richard S. Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 993–1000, New York, NY, USA, 2009b. ACM. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553501. URL http://doi.acm.org/10.1145/1553374.1553501.

Shalabh Bhatnagar, Doina Precup, David Silver, Richard S Sutton, Hamid R. Maei, and Csaba Szepesvári. Convergent temporal-difference learning with arbitrary smooth function approximation. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1204–1212. Curran Associates, Inc., 2009.

Yarin Gal. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

J. F. Bard. Some properties of the bilevel programming problem. *J. Optim. Theory Appl.*, 68(2):371–378, February 1991. ISSN 0022-3239.

Hamid Reza Maei, Csaba Szepesvári, Shalabh Bhatnagar, and Richard S. Sutton. Toward off-policy learning control with function approximation. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 719–726, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

Tim Pearce, Mohamed Zaki, Alexandra Brintrup, and Andy Neely. Uncertainty in neural networks: Bayesian ensembling. *ArXiv Preprint*, abs/1810.05546, 10 2019.

Kamil Ciosek, Vincent Fortuin, Ryota Tomioka, Katja Hofmann, and Richard Turner. Conservative uncertainty estimation by fitting prior networks. In *Eighth International Conference on Learning Representations*, 04 2020.

Vivek S. Borkar. Stochastic approximation with two time scales. *Syst. Control Lett.*, 29 (5):291–294, February 1997. ISSN 0167-6911. doi: 10.1016/S0167-6911(97)90015-3. URL https://doi.org/10.1016/S0167-6911(97)90015-3.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6626–6637. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7240-gans-trained-by-a-two-time-scale-update-rule-converge-to-a-local-nash-equilibrium.pdf.

Alexander Shapiro. Directional differentiability of metric projections onto moving sets at boundary points. *Journal of Mathematical Analysis and Applications*, 131(2):392–403, 1988. ISSN 0022-247X. doi: https://doi.org/10.1016/0022-247X(88)90213-2. URL https://www.sciencedirect.com/science/article/pii/0022247X88902132.

Eduardo H Zarantonello. Projections on convex sets in hilbert space and spectral theory: Part i. projections on convex sets: Part ii. spectral theory. In Eduardo H. Zarantonello, editor, *Contributions to Nonlinear Functional Analysis*, pages 237–424.

Academic Press, 1971. ISBN 978-0-12-775850-3. doi: https://doi.org/10.1016/B978-0-12-775850-3.50013-3. URL https://www.sciencedirect.com/science/article/pii/B9780127758503500133.

Alexander Shapiro. On differentiability of metric projections in rn, 1: Boundary case. *Proceedings of the American Mathematical Society*, 99(1):123–128, 1987. ISSN 00029939, 10886826. URL http://www.jstor.org/stable/2046282.

Wesley Chung, Somjit Nath, Ajin Joseph, and Martha White. Two-timescale networks for nonlinear value function approximation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=rJleN20qK7.

Christoph Dann, Gerhard Neumann, Jan Peters, et al. Policy evaluation with temporal differences: A survey and comparison. *Journal of Machine Learning Research*, 15: 809–883, 2014.

Simon S. Du, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou. Stochastic variance reduction methods for policy evaluation. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1049–1058, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

Kamil Ciosek, Quan Vuong, Robert Loftin, and Katja Hofmann. Better exploration with optimistic actor-critic. *arXiv preprint arXiv:1910.12807*, 2019b.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6402–6413. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7219-simple-and-scalable-predictive-uncertainty-estimation-using-deep-ensembles.pdf.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4026–4034. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6501-deep-exploration-via-bootstrapped-dqn.pdf.

Zhuora Yang, Yuchen Xie, and Zhaoran Wang. A theoretical analysis of deep q-learning, 2019.

David Williams. *Probability with Martingales*. Cambridge mathematical textbooks. Cambridge University Press, 1991. ISBN 978-0-521-40605-5.

A.N. Kolmogorov, N. Morrison, and A.T. Bharucha-Reid. *Foundations of the Theory of Probability*. AMS Chelsea Publishing Series. Chelsea Publishing Company, 1956. URL https://books.google.co.uk/books?id=EmetAAAAMAAJ.

R.F. Bass. *Real Analysis for Graduate Students*, chapter 21. Createspace Ind Pub, 2013. ISBN 9781481869140. URL https://books.google.co.uk/books?id=s6mVlgEACAAJ.

Donald Andrews. Generic uniform convergence. *Econometric Theory*, 8(2):241–257, 1992.

R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Springer Verlag, Heidelberg, Berlin, New York, 1998.

Prasenjit Karmakar and Shalabh Bhatnagar. Two time-scale stochastic approximation with controlled markov noise and off-policy temporal-difference learning. *Math. Oper.*

Res., 43(1):130–151, February 2018. ISSN 0364-765X. doi: 10.1287/moor.2017.0855. URL https://doi.org/10.1287/moor.2017.0855.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.