Bayesian Reinforcement Learning: A Survey

Mohammad Ghavamzadeh
Adobe Research & INRIA
mohammad.ghavamzadeh@inria.fr

Shie Mannor
Technion
shie@ee.technion.ac.il

Joelle Pineau
McGill University
jpineau@cs.mcgill.ca

Aviv Tamar
University of California, Berkeley
avivt@berkeley.edu

Full text available at: http://dx.doi.org/10.1561/2200000049
Editorial Scope

Topics

Foundations and Trends® in Machine Learning publishes survey and tutorial articles on the theory, algorithms and applications of machine learning, including the following topics:

- Adaptive control and signal processing
- Applications and case studies
- Behavioral, cognitive, and neural learning
- Bayesian learning
- Classification and prediction
- Clustering
- Data mining
- Dimensionality reduction
- Evaluation
- Game theoretic learning
- Graphical models
- Independent component analysis
- Inductive logic programming
- Kernel methods
- Markov chain Monte Carlo
- Model choice
- Nonparametric methods
- Online learning
- Optimization
- Reinforcement learning
- Relational learning
- Robustness
- Spectral methods
- Statistical learning theory
- Variational inference
- Visualization

Information for Librarians

Foundations and Trends® in Machine Learning, 2015, Volume 8, 6 issues. ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.
Bayesian Reinforcement Learning: A Survey

Mohammad Ghavamzadeh
Adobe Research & INRIA
mohammad.ghavamzadeh@inria.fr

Shie Mannor
Technion
shie@ee.technion.ac.il

Joelle Pineau
McGill University
jpineau@cs.mcgill.ca

Aviv Tamar
University of California, Berkeley
avivt@berkeley.edu
# Contents

1 Introduction 2

2 Technical Background 9

2.1 Multi-Armed Bandits 9
2.2 Markov Decision Processes 12
2.3 Partially Observable Markov Decision Processes 16
2.4 Reinforcement Learning 18
2.5 Bayesian Learning 20

3 Bayesian Bandits 27

3.1 Classical Results 28
3.2 Bayes-UCB 31
3.3 Thompson Sampling 31

4 Model-based Bayesian Reinforcement Learning 38

4.1 Models and Representations 38
4.2 Exploration/Exploitation Dilemma 42
4.3 Offline Value Approximation 43
4.4 Online near-myopic value approximation 45
4.5 Online Tree Search Approximation 47
4.6 Methods with Exploration Bonus to Achieve PAC Guarantees 53
4.7 Extensions to Unknown Rewards 60
Abstract

Bayesian methods for machine learning have been widely investigated, yielding principled methods for incorporating prior information into inference algorithms. In this survey, we provide an in-depth review of the role of Bayesian methods for the reinforcement learning (RL) paradigm. The major incentives for incorporating Bayesian reasoning in RL are: 1) it provides an elegant approach to action-selection (exploration/exploitation) as a function of the uncertainty in learning; and 2) it provides a machinery to incorporate prior knowledge into the algorithms. We first discuss models and methods for Bayesian inference in the simple single-step Bandit model. We then review the extensive recent literature on Bayesian methods for model-based RL, where prior information can be expressed on the parameters of the Markov model. We also present Bayesian methods for model-free RL, where priors are expressed over the value function or policy class. The objective of the paper is to provide a comprehensive survey on Bayesian RL algorithms and their theoretical and empirical properties.
A large number of problems in science and engineering, from robotics to game playing, tutoring systems, resource management, financial portfolio management, medical treatment design and beyond, can be characterized as sequential decision-making under uncertainty. Many interesting sequential decision-making tasks can be formulated as reinforcement learning (RL) problems \cite{BertsekasTsitsiklis1996, SuttonBarto1998}. In an RL problem, an agent interacts with a dynamic, stochastic, and incompletely known environment, with the goal of finding an action-selection strategy, or policy, that optimizes some long-term performance measure.

One of the key features of RL is the focus on learning a control policy to optimize the choice of actions over several time steps. This is usually learned from sequences of data. In contrast to supervised learning methods that deal with independently and identically distributed (i.i.d.) samples from the domain, the RL agent learns from the samples that are collected from the trajectories generated by its sequential interaction with the system. Another important aspect is the effect of the agent’s policy on the data collection; different policies naturally yield different distributions of sam-
pled trajectories, and thus, impacting what can be learned from the data.

Traditionally, RL algorithms have been categorized as being either model-based or model-free. In the former category, the agent uses the collected data to first build a model of the domain’s dynamics and then uses this model to optimize its policy. In the latter case, the agent directly learns an optimal (or good) action-selection strategy from the collected data. There is some evidence that the first method provides better results with less data [Atkeson and Santamaria, 1997], and the second method may be more efficient in cases where the solution space (e.g., policy space) exhibits more regularity than the underlying dynamics, though there is some disagreement about this.

A major challenge in RL is in identifying good data collection strategies, that effectively balance between the need to explore the space of all possible policies, and the desire to focus data collection towards trajectories that yield better outcome (e.g., greater chance of reaching a goal, or minimizing a cost function). This is known as the exploration-exploitation tradeoff. This challenge arises in both model-based and model-free RL algorithms.

Bayesian reinforcement learning (BRL) is an approach to RL that leverages methods from Bayesian inference to incorporate information into the learning process. It assumes that the designer of the system can express prior information about the problem in a probabilistic distribution, and that new information can be incorporated using standard rules of Bayesian inference. The information can be encoded and updated using a parametric representation of the system dynamics, in the case of model-based RL, or of the solution space, in the case of model-free RL.

A major advantage of the BRL approach is that it provides a principled way to tackle the exploration-exploitation problem. Indeed, the Bayesian posterior naturally captures the full state of knowledge, subject to the chosen parametric representation, and thus, the agent can select actions that maximize the expected gain with respect to this information state.
Another major advantage of BRL is that it implicitly facilitates regularization. By assuming a prior on the value function, the parameters defining a policy, or the model parameters, we avoid the trap of letting a few data points steer us away from the true parameters. On the other hand, having a prior precludes overly rapid convergence. The role of the prior is therefore to soften the effect of sampling a finite dataset, effectively leading to regularization. We note that regularization in RL has been addressed for the value function [Farahmand et al., 2008b] and for policies [Farahmand et al., 2008a]. A major issue with these regularization schemes is that it is not clear how to select the regularization coefficient. Moreover, it is not clear why an optimal value function (or a policy) should belong to some pre-defined set.

Yet another advantage of adopting a Bayesian view in RL is the principled Bayesian approach for handling parameter uncertainty. Current frequentist approaches for dealing with modelling errors in sequential decision making are either very conservative, or computationally infeasible [Nilim and El Ghaoui, 2005]. By explicitly modelling the distribution over unknown system parameters, Bayesian methods offer a promising approach for solving this difficult problem.

Of course, several challenges arise in applying Bayesian methods to the RL paradigm. First, there is the challenge of selecting the correct representation for expressing prior information in any given domain. Second, defining the decision-making process over the information state is typically computationally more demanding than directly considering the natural state representation. Nonetheless, a large array of models and algorithms have been proposed for the BRL framework, leveraging a variety of structural assumptions and approximations to provide feasible solutions.

The main objective of this paper is to provide a comprehensive survey on BRL algorithms and their theoretical and empirical properties. In Chapter 2, we provide a review of the main mathematical concepts and techniques used throughout this paper. Chapter 3 surveys the Bayesian learning methods for the case of single-step decision-making, using the bandit framework. This section serves both as an exposition of the potential of BRL in a simpler setting that is well understood, but is
also of independent interest, as bandits have widespread applications. The main results presented here are of a theoretical nature, outlining known performance bounds for the regret minimization criteria. Chapter 4 reviews existing methods for model-based BRL, where the posterior is expressed over parameters of the system dynamics model. Chapter 5 focuses on BRL methods that do not explicitly learn a model of the system, but rather the posterior is expressed over the solution space. Chapter 6 focuses on a particular advantage of BRL in dealing with risk due to parameter-uncertainty, and surveys several approaches for incorporating such risk into the decision-making process. Finally, Chapter 7 discusses various extensions of BRL for special classes of problems (PAC-Bayes model selection, inverse RL, multi-agent RL, and multi-task RL). Figure 1.1 outlines the various BRL approaches covered throughout the paper.

An Example Domain

We present an illustrative domain suitable to be solved using the BRL techniques surveyed in this paper. This running example will be used throughout the paper to elucidate the difference between the various BRL approaches and to clarify various BRL concepts.

Example 1.1 (The Online Shop). In the online shop domain, a retailer aims to maximize profit by sequentially suggesting products to online shopping customers. Formally, the domain is characterized by the following model:

- A set of possible customer states, $\mathcal{X}$. States can represent intrinsic features of the customer such as gender and age, but also dynamic quantities such as the items in his shopping cart, or his willingness to shop;
- A set of possible product suggestions and advertisements, $\mathcal{A}$;
- A probability kernel, $P$, defined below.

An episode in the online shop domain begins at time $t = 0$, when a customer with features $x_0 \in \mathcal{X}$ enters the online shop. Then, a sequential interaction between the customer and the online shop begins,
Introduction

Bayesian RL

- Bandits (Sec 3)
  - Bayes UCB
  - Thompson sampling

- Model-based BRL (Sec 4)
  - Offline value approximation
    - Finite state controllers
    - BEETLE
  - Online near-myopic value approximation
    - Bayesian DP
    - VOI heuristic
  - Online tree search approximation
    - Forward search
    - Bayesian sparse sampling
    - HMDP
    - BFS3
    - Branch-and-bound search
    - BAMCP
  - Exploration bonus approximation
    - BOSS
    - BEB
    - VBRB
    - BOLT

- Model-free BRL (Sec 5)
  - Value function algos
    - GPTD
    - GPSARSA
  - Policy gradient algos
    - Bayesian Quadrature
    - Two Bayesian models for estimating the policy gradient
  - Actor-Critic algos
    - GPTD + Bayesian policy gradient

- Risk Aware BRL (Sec 6)
  - Bias variance approximation
  - Percentile criterion
  - Min-max criterion
  - Percentile measures criteria

Figure 1.1: Overview of the Bayesian RL approaches covered in this survey.
where at each step $t = 0, 1, 2, \ldots$, an advertisement $a_t \in \mathcal{A}$ is shown to the customer, and following that the customer makes a decision to either (i) add a product to his shopping cart; (ii) not buy the product, but continue to shop; (iii) stop shopping and check out. Following the customers decision, his state changes to $x_{t+1}$ (reflecting the change in the shopping cart, willingness to continue shopping, etc.). We assume that this change is captured by a probability kernel $P(x_{t+1} | x_t, a_t)$.

When the customer decides to check out, the episode ends, and a profit is obtained according to the items he had added to his cart. The goal is to find a product suggestion policy, $x \rightarrow a \in \mathcal{A}$, that maximizes the expected total profit.

When the probabilities of customer responses $P$ are known in advance, calculating an optimal policy for the online shop domain is basically a planning problem, which may be solved using traditional methods for resource allocation [Powell, 2011]. A more challenging, but realistic, scenario is when $P$ is not completely known beforehand, but has to be learned while interacting with customers. The BRL framework employs Bayesian methods for learning $P$, and for learning an optimal product suggestion policy.

There are several advantages for choosing a Bayesian approach for the online shop domain. First, it is likely that some prior knowledge about $P$ is available. For example, once a customer adds a product of a particular brand to his cart, it is likely that he prefers additional products of the same brand over those of a different one. Taking into account such knowledge is natural in the Bayesian method, by virtue of the prior distribution over $P$. As we shall see, the Bayesian approach also naturally extends to more general forms of structure in the problem.

A second advantage concerns what is known as the exploitation–exploration dilemma: should the decision-maker display only the most profitable product suggestions according to his current knowledge about $P$, or rather take exploratory actions that may turn out to be less profitable, but provide useful information for future decisions? The Bayesian method offers a principled approach to dealing with this difficult problem by explicitly quantifying the value
Introduction

of exploration, made possible by maintaining a distribution over $P$.

The various parameter configurations in the online shop domain lead to the different learning problems surveyed in this paper. In particular:

- For a single-step interaction, i.e., when the episode terminates after a single product suggestion, the problem is captured by the multi-armed bandit model of Chapter 3.

- For small-scale problems, i.e., a small number of products and customer types, $P$ may be learnt explicitly. This is the model-based approach of Chapter 4.

- For large problems, a near-optimal policy may be obtained without representing $P$ explicitly. This is the model-free approach of Chapter 5.

- When the customer state is not fully observed by the decision-maker, we require models that incorporate partial observability; see §2.3 and §4.9.

Throughout the paper, we revisit the online shop domain, and specify explicit configurations that are relevant to the surveyed methods.


References


References


References


References


119 References


References


References


References


