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

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Machine Learning

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

M. Ghavamzadeh, S. Mannor, J. Pineau, and A. Tamar. *Bayesian Reinforcement Learning: A Survey*. Foundations and Trends® in Machine Learning, vol. 8, no. 5-6, pp. 359–483, 2015.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-68083-089-7

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

Foundations and Trends® in Machine Learning
Volume 8, Issue 5-6, 2015

Editorial Board

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.

Foundations and Trends® in Machine Learning
Vol. 8, No. 5-6 (2015) 359–483

DOI: 10.1561/2200000049

now
the essence of knowledge

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

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.

M. Ghavamzadeh, S. Mannor, J. Pineau, and A. Tamar. *Bayesian Reinforcement Learning: A Survey*. Foundations and Trends® in Machine Learning, vol. 8, no. 5-6, pp. 359–483, 2015.
DOI: 10.1561/2200000049.

1

Introduction

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 [Bertsekas and Tsitsiklis, 1996, Sutton and Barto, 1998]. 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,

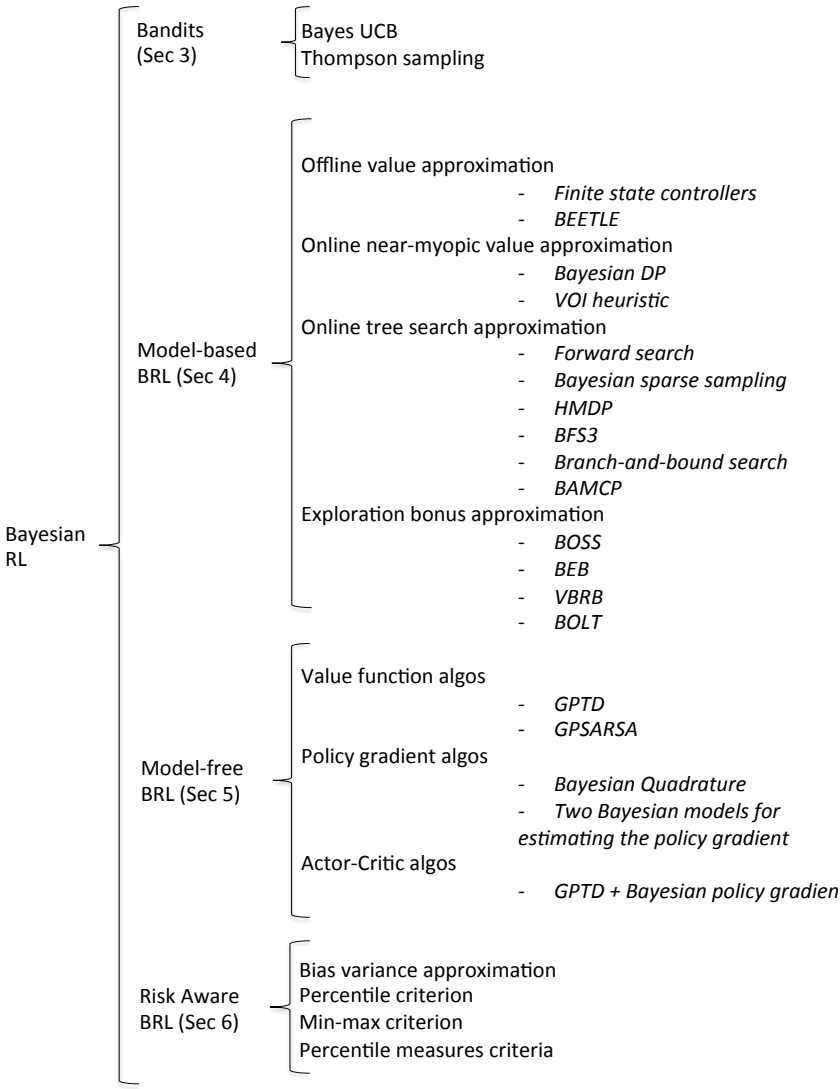

Figure 1.1: Overview of the Bayesian RL approaches covered in this survey.

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, $x \to a \in \mathcal{A}$, 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

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

Y. Abbasi-Yadkori and C. Szepesvari. Bayesian optimal control of smoothly parameterized systems. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2015.

P. Abbeel and A. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the 21st International Conference on Machine Learning*, 2004.

S. Agrawal and N. Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT), JMLR W&CP*, volume 23, pages 39.1 – 39.26, 2012.

S. Agrawal and N. Goyal. Further optimal regret bounds for Thompson sampling. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*, pages 99–107, 2013a.

S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 127–135, 2013b.

M. Araya-Lopez, V. Thomas, and O. Buffet. Near-optimal BRL using optimistic local transitions. In *International Conference on Machine Learning*, 2012.

J. Asmuth. *Model-based Bayesian Reinforcement Learning with Generalized Priors*. PhD thesis, Rutgers, 2013.

J. Asmuth and M. Littman. Approaching Bayes-optimality using Monte-Carlo tree search. In *International Conference on Automated Planning and Scheduling (ICAPS)*, 2011.

J. Asmuth, L. Li, M. Littman, A. Nouri, and D. Wingate. A Bayesian sampling approach to exploration in reinforcement learning. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2009.

K. Astrom. Optimal control of Markov decision processes with incomplete state estimation. *Journal of Mathematical Analysis and Applications*, 10: 174–205, 1965.

C. G. Atkeson and J. C. Santamaria. A comparison of direct and model-based reinforcement learning. In *International Conference on Robotics and Automation (ICRA)*, 1997.

A. Atrash and J. Pineau. A Bayesian reinforcement learning approach for customizing human-robot interfaces. In *International Conference on Intelligent User Interfaces*, 2009.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

M. Babes, V. Marivate, K. Subramanian, and M. Littman. Apprenticeship learning about multiple intentions. In *Proceedings of the 28th International Conference on Machine Learning*, pages 897–904, 2011.

A. Barto, R. Sutton, and C. Anderson. Neuron-like elements that can solve difficult learning control problems. *IEEE Transaction on Systems, Man and Cybernetics*, 13:835–846, 1983.

J. Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.

J. Baxter and P. Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

J. Baxter, A. Tridgell, and L. Weaver. Knightcap: A chess program that learns by combining TD(λ) with game-tree search. In *Proceedings of the 15th International Conference on Machine Learning*, pages 28–36, 1998.

J. Baxter, P. Bartlett, and L. Weaver. Experiments with infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15: 351–381, 2001.

R. Bellman. *Dynamic Programming*. Princeton Universty Press, 1957.

D. Bertsekas and J. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

L. Bertuccelli, A. Wu, and J. How. Robust adaptive Markov decision processes: Planning with model uncertainty. *Control Systems, IEEE*, 32(5): 96–109, Oct 2012.

S. Bhatnagar, R. Sutton, M. Ghavamzadeh, and M. Lee. Incremental natural actor-Critic algorithms. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 105–112, 2007.

S. Bhatnagar, R. Sutton, M. Ghavamzadeh, and M. Lee. Natural actor-critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

J. Boyan. Least-squares temporal difference learning. In *Proceedings of the 16th International Conference on Machine Learning*, pages 49–56, 1999.

S. Bradtke and A. Barto. Linear least-squares algorithms for temporal difference learning. *Journal of Machine Learning*, 22:33–57, 1996.

R. Brafman and M. Tennenholtz. R-max - a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research (JMLR)*, 3:213–231, 2003.

S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012. ISSN 1935-8237.

R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

P. Castro and D. Precup. Using linear programming for Bayesian exploration in Markov decision processes. In *International Joint Conference on Artificial Intelligence*, pages 2437–2442, 2007.

P. Castro and D. Precup. Smarter sampling in model-based Bayesian reinforcement learning. In *Machine Learning and Knowledge Discovery in Databases*, 2010.

M. Castronovo. Bbrl a c++ open-source library used to compare bayesian reinforcement learning algorithms. https://github.com/mcastron/BBRL/, 2015.

M. Castronovo, D. Ernst, and R. Fonteneau A. Couetoux. Benchmarking for bayesian reinforcement learning. Working paper, Inst. Montefiore, http://hdl.handle.net/2268/185881, 2015.

G. Chalkiadakis and C. Boutilier. Coordination in multiagent reinforcement learning: A Bayesian approach. In *Proceedings of the 2nd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2013.

G. Chalkiadakis, E. Elkinda, E. Markakis, M. Polukarov, and N. Jennings. Cooperative games with overlapping coalitions. *Journal of Artificial Intelligence Research*, 39(1):179–216, 2010.

O. Chapelle and L. Li. An empirical evaluation of Thompson sampling. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 2249–2257, 2011.

K. Chen and M. Bowling. Tractable objectives for robust policy optimization. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 2078–2086, 2012.

J. Choi and K. Kim. Map inference for Bayesian inverse reinforcement learning. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 1989–1997, 2011.

J. Choi and K. Kim. Nonparametric Bayesian inverse reinforcement learning for multiple reward functions. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 305–313, 2012.

R. Crites and A. Barto. Elevator group control using multiple reinforcement learning agents. *Machine Learning*, 33:235–262, 1998.

P. Dallaire, C. Besse, S. Ross, and B. Chaib-draa. Bayesian reinforcement learning in continuous POMDPs with Gaussian processes. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009.

R. Dearden, N. Friedman, and S. J. Russell. Bayesian Q-learning. In *AAAI Conference on Artificial Intelligence*, pages 761–768, 1998.

R. Dearden, N. Friedman, and D. Andre. Model based Bayesian exploration. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 1999.

E. Delage and S. Mannor. Percentile optimization for Markov decision processes with parameter uncertainty. *Operations Research*, 58(1):203–213, 2010.

C. Dimitrakakis and C. Rothkopf. Bayesian multi-task inverse reinforcement learning. In *Proceedings of the European Workshop on Reinforcement Learning*, 2011.

F. Doshi, J. Pineau, and N. Roy. Reinforcement learning with limited reinforcement: using Bayes risk for active learning in POMDPs. In *International Conference on Machine Learning*, 2008.

F. Doshi-Velez. The infinite partially observable Markov decision process. In *Proceedings of the Advances in Neural Information Processing Systems*, 2009.

F. Doshi-Velez, D. Wingate, N. Roy, and J. Tenenbaum. Nonparametric Bayesian policy priors for reinforcement learning. In *Proceedings of the Advances in Neural Information Processing Systems*, 2010.

F. Doshi-Velez, J. Pineau, and N. Roy. Reinforcement learning with limited reinforcement: Using Bayes risk for active learning in POMDPs. *Artificial Intelligence*, 2011.

M. Duff. Monte-Carlo algorithms for the improvement of finite-state stochastic controllers: Application to Bayes-adaptive Markov decision processes. In *International Workshop on Artificial Intelligence and Statistics (AISTATS)*, 2001.

M. Duff. *Optimal Learning: Computational Procedures for Bayes-Adaptive Markov Decision Processes*. PhD thesis, University of Massachusetts Amherst, Amherst, MA, 2002.

B. Efron. *Large-Scale Inference: Empirical Bayes Methods fro Estimation, Testing, and Prediction*. IMS Statistics Monographs. Cambridge University Press, 2010.

Y. Engel. *Algorithms and Representations for Reinforcement Learning*. PhD thesis, The Hebrew University of Jerusalem, Israel, 2005.

Y. Engel, S. Mannor, and R. Meir. Sparse online greedy support vector regression. In *Proceedings of the 13th European Conference on Machine Learning*, pages 84–96, 2002.

Y. Engel, S. Mannor, and R. Meir. Bayes meets Bellman: The Gaussian process approach to temporal difference learning. In *Proceedings of the 20th International Conference on Machine Learning*, pages 154–161, 2003.

Y. Engel, S. Mannor, and R. Meir. Reinforcement learning with Gaussian processes. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 201–208, 2005a.

Y. Engel, P. Szabó, and D. Volkinshtein. Learning to control an octopus arm with gaussian process temporal difference methods. In *Proceedings of the Advances in Neural Information Processing Systems*, 2005b.

D. Ernst, P. Geurts, and L. Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005.

A. Farahmand, M. Ghavamzadeh, C., and Shie Mannor. Regularized policy iteration. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 441–448, 2008a.

A. Farahmand, M. Ghavamzadeh, C. Szepesvári, and S. Mannor. Regularized fitted q-iteration: Application to planning. In *Recent Advances in Reinforcement Learning, 8th European Workshop, EWRL*, pages 55–68, 2008b.

M. M. Fard and J. Pineau. PAC-Bayesian model selection for reinforcement learning. In *Proceedings of the Advances in Neural Information Processing Systems*, 2010.

M. M. Fard, J. Pineau, and C. Szepesvari. PAC-Bayesian policy evaluation for reinforcement learning. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2011.

A. Feldbaum. Dual control theory, parts i and ii. *Automation and Remote Control*, 21:874–880 and 1033–1039, 1961.

N. M. Filatov and H. Unbehauen. Survey of adaptive dual control methods. In *IEEE Control Theory and Applications*, volume 147, pages 118–128, 2000.

N. Friedman and Y. Singer. Efficient Bayesian parameter estimation in large discrete domains. In *Proceedings of the Advances in Neural Information Processing Systems*, 1999.

S. Gelly, L. Kocsis, M. Schoenauer, M. Sebag, D. Silver, C. Szepesvari, and O. Teytaud. The grand challenge of computer Go: Monte Carlo tree search and extensions. *Communications of the ACM*, 55(3):106–113, 2012.

M. Ghavamzadeh and Y. Engel. Bayesian policy gradient algorithms. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 457–464, 2006.

M. Ghavamzadeh and Y. Engel. Bayesian Actor-Critic algorithms. In *Proceedings of the 24th International Conference on Machine Learning*, pages 297–304, 2007.

M. Ghavamzadeh, Y. Engel, and M. Valko. Bayesian policy gradient and actor-critic algorithms. Technical Report 00776608, INRIA, 2013.

J. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 148–177, 1979.

P. Glynn. Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM*, 33:75–84, 1990.

A. Gopalan and S. Mannor. Thompson sampling for learning parameterized markov decision processes. In *Proceedings of the 28th Conference on Learning Theory (COLT)*, pages 861–898, 2015.

A. Gopalan, S. Mannor, and Y. Mansour. Thompson sampling for complex online problems. In *Proceedings of the 31st International Conference on Machine Learning*, pages 100–108, 2014.

T. Graepel, J.Q. Candela, T. Borchert, and R. Herbrich. Web-scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft's Bing search engine. In *Proceedings of the 27th International Conference on Machine Learning*, pages 13–20, 2010.

A. Greenfield and A. Brockwell. Adaptive control of nonlinear stochastic systems by particle filtering. In *International Conference on Control and Automation*, 2003.

A. Guez, D. Silver, and P. Dayan. Efficient Bayes-adaptive reinforcement learning using sample-based search. In *Proceedings of the Advances in Neural Information Processing Systems*, 2012.

S. Guha and K. Munagala. Stochastic regret minimization via Thompson sampling. In *Proceedings of The 27th Conference on Learning Theory*, pages 317–338, 2014.

T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *Proceedings of the Advances in Neural Information Processing Systems*, 1999.

R. Jaulmes, J. Pineau, and J. Precup. Active learning in partially observable Markov decision processes. In *European Conference on Machine Learning*, 2005.

L. Kaelbling, M. Littman, and A. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1998.

E. Kaufmann, O. Cappé, and A. Garivier. On Bayesian upper confidence bounds for bandit problems. In *International Conference on Artificial Intelligence and Statistics*, pages 592–600, 2012a.

E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory*, volume 7568 of *Lecture Notes in Computer Science*, pages 199–213, 2012b.

K. Kawaguchi and M. Araya-Lopez. A greedy approximation of Bayesian reinforcement learning with probably optimistic transition model. In *Adaptive Learning Agents 2013 (a workshop of AAAMAS)*, 2013.

M. Kearns and S. Singh. Near-optimal reinforcement learning in polynomial time. In *International Conference on Machine Learning*, pages 260–268, 1998.

M. Kearns, Y. Mansour, and A. Ng. A sparse sampling algorithm for near-optimal planning in large Markov decision processes. In *International Joint Conference on Artificial Intelligence*, pages 1324–1331, 1999.

J. Kober, D. Bagnell, and J. Peters. Reinforcement learning in robotics: A survey. *International Journal of Robotics Research (IJRR)*, 2013.

L. Kocsis and C. Szepesvari. Bandit based Monte-Carlo planning. In *Proceedings of the European Conference on Machine Learning (ECML)*, 2006.

N. Kohl and P. Stone. Policy gradient reinforcement learning for fast quadrupedal locomotion. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 2619–2624, 2004.

J. Kolter and A. Ng. Near-Bayesian exploration in polynomial time. In *International Conference on Machine Learning*, 2009.

V. Konda and J. Tsitsiklis. Actor-Critic algorithms. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 1008–1014, 2000.

M. Lagoudakis and R. Parr. Least-squares policy iteration. *Journal of Machine Learning Research*, 4:1107–1149, 2003.

T. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

A. Lazaric and M. Ghavamzadeh. Bayesian multi-task reinforcement learning. In *Proceedings of the 27th International Conference on Machine Learning*, pages 599–606, 2010.

L. Li. *A unifying framework for computational reinforcement learning theory.* PhD thesis, Rutgers, 2009.

C. Liu and L. Li. On the prior sensitivity of Thompson sampling. *CoRR*, abs/1506.03378, 2015.

S. Mannor and J. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *The Journal of Machine Learning Research*, 5:623–648, 2004.

S. Mannor, R. Rubinstein, and Y. Gat. The cross entropy method for fast policy search. In *International Conference on Machine Learning*, 2003.

S. Mannor, D. Simester, P. Sun, and J.N. Tsitsiklis. Bias and variance approximation in value function estimates. *Management Science*, 53(2):308–322, 2007.

P. Marbach. *Simulated-Based Methods for Markov Decision Processes.* PhD thesis, Massachusetts Institute of Technology, 1998.

D. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37, 1999.

N. Mehta, S. Natarajan, P. Tadepalli, and A. Fern. Transfer in variable-reward hierarchical reinforcement learning. *Machine Learning*, 73(3):289–312, 2008.

B. Michini and J. How. Bayesian nonparametric inverse reinforcement learning. In *Proceedings of the European Conference on Machine Learning*, 2012a.

B. Michini and J. How. Improving the efficiency of Bayesian inverse reinforcement learning. In *IEEE International Conference on Robotics and Automation*, pages 3651–3656, 2012b.

A. Moore and C. Atkeson. Prioritized sweeping: Reinforcement learning with less data and less real time. *Machine Learning*, 13:103–130, 1993.

A. Neu and C. Szepesvári. Apprenticeship learning using inverse reinforcement learning and gradient methods. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2007.

A. Ng and S. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning*, pages 663–670, 2000.

A. Ng, H. Kim, M. Jordan, and S. Sastry. Autonomous helicopter flight via reinforcement learning. In *Proceedings of the Advances in Neural Information Processing Systems*. MIT Press, 2004.

A. Nilim and L. El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

J. Niño-Mora. Computing a classic index for finite-horizon bandits. *INFORMS Journal on Computing*, 23(2):254–267, 2011.

A. O'Hagan. Bayes-Hermite quadrature. *Journal of Statistical Planning and Inference*, 29:245–260, 1991.

I. Osband, D. Russo, and B. Van Roy. (More) efficient reinforcement learning via posterior sampling. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2013.

S. Paquet, L. Tobin, and B. Chaib-draa. An online POMDP algorithm for complex multiagent environments. In *International Joint Conference on Autonomous Agents and Multi Agent Systems (AAMAS)*, pages 970–977, 2005.

J. Peters and S. Schaal. Natural actor-critic. *Neurocomputing*, 71(7-9):1180–1190, 2008.

J. Pineau, G. Gordon, and S. Thrun. Point-based value iteration: an anytime algorithm for POMDPs. In *International Joint Conference on Artificial Intelligence*, pages 1025–1032, 2003.

S Png. *Bayesian Reinforcement Learning for POMDP-based Dialogue Systems*. Master's thesis, McGill University, 2011.

S. Png and J. Pineau. Bayesian reinforcement learning for POMDP-based dialogue systems. In *ICASSP*, 2011.

H. Poincaré. *Calcul des Probabilités*. Georges Carré, Paris, 1896.

J. Porta, N. Vlassis, M. Spaan, and P. Poupart. Point-based value iteration for continuous POMDPs. *Journal of Machine Learning Research*, 7, 2006.

P. Poupart, N. Vlassis, J. Hoey, and K. Regan. An analytic solution to discrete Bayesian reinforcement learning. In *International Conference on Machine learning*, pages 697–704, 2006.

W. B. Powell. *Approximate Dynamic Programming: Solving the curses of dimensionality (2nd Edition)*. John Wiley & Sons, 2011.

M. Puterman. *Markov Decision Processes*. Wiley Interscience, 1994.

R. Munos R. Fonteneau, L. Busoniu. Optimistic planning for belief-augmented Markov Decision Processes. In *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, 2013.

D. Ramachandran and E. Amir. Bayesian inverse reinforcement learning. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 2586–2591, 2007.

C. Rasmussen and Z. Ghahramani. Bayesian Monte Carlo. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 489–496. MIT Press, 2003.

C. Rasmussen and M. Kuss. Gaussian processes in reinforcement learning. In *Proceedings of the Advances in Neural Information Processing Systems*. MIT Press, 2004.

C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

N. Ratliff, A. Bagnell, and M. Zinkevich. Maximum margin planning. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

R. Ravikanth, S. Meyn, and L. Brown. Bayesian adaptive control of time varying systems. In *IEEE Conference on Decision and Control*, 1992.

J. Reisinger, P. Stone, and R. Miikkulainen. Online kernel selection for Bayesian reinforcement learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 816–823, 2008.

S. Ross and J. Pineau. Model-based Bayesian reinforcement learning in large structured domains. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2008.

S. Ross, B. Chaib-draa, and J. Pineau. Bayes-adaptive POMDPs. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 20, pages 1225–1232, 2008a.

S. Ross, B. Chaib-draa, and J. Pineau. Bayesian reinforcement learning in continuous POMDPs with application to robot navigation. In *IEEE International Conference on Robotics and Automation*, 2008b.

S. Ross, J. Pineau, S. Paquet, and B. Chaib-draa. Online POMDPs. *Journal of Artificial Intelligence Research (JAIR)*, 32:663–704, 2008c.

S. Ross, J. Pineau, B. Chaib-draa, and P. Kreitmann. A Bayesian approach for learning and planning in partially observable Markov decision processes. *Journal of Machine Learning Research*, 12, 2011.

G. Rummery and M. Niranjan. On-line Q-learning using connectionist systems. Technical Report CUED/F-INFENG/TR 166, Engineering Department, Cambridge University, 1994.

P. Rusmevichientong and J. N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

I. Rusnak. Optimal adaptive control of uncertain stochastic discrete linear systems. In *IEEE International Conference on Systems, Man and Cybernetics*, 1995.

S. Russell. Learning agents for uncertain environments (extended abstract). In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 101–103, 1998.

S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach (2nd Edition)*. Prentice Hall, 2002.

D. Russo and B. Van Roy. An information-theoretic analysis of Thompson sampling. *CoRR*, abs/1403.5341, 2014a.

D. Russo and B. Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014b.

L. Scharf. *Statistical Signal Processing*. Addison-Wesley, 1991.

B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.

S. Scott. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010.

Y. Seldin, P. Auer, F. Laviolette, J. Shawe-Taylor, and R. Ortner. PAC-Bayesian analysis of contextual bandits. In *Proceedings of the Advances in Neural Information Processing Systems*, 2011a.

Y. Seldin, N. Cesa-Bianchi, F. Laviolette, P. Auer, J. Shawe-Taylor, and J. Peters. PAC-Bayesian analysis of the exploration-exploitation trade-off. In *ICML Workshop on online trading of exploration and exploitation*, 2011b.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

R. Smallwood and E. Sondik. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, Sep/Oct 1973.

V. Sorg, S. Sing, and R. Lewis. Variance-based rewards for approximate Bayesian reinforcement learning. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2010.

M. Spaan and N. Vlassis. Perseus: randomized point-based value iteration for POMDPs. *Journal of Artificial Intelligence Research (JAIR)*, 24:195–220, 2005.

A. Strehl and M. Littman. A theoretical analysis of model-based interval estimation. In *International Conference on Machine learning*, pages 856–863, 2005.

A. Strehl and M. Littman. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74: 1209–1331, 2008.

M. Strens. A Bayesian framework for reinforcement learning. In *International Conference on Machine Learning*, 2000.

R. Sutton. *Temporal credit assignment in reinforcement learning*. PhD thesis, University of Massachusetts Amherst, 1984.

R. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988.

R. Sutton. DYNA, an integrated architecture for learning, planning, and reacting. *SIGART Bulletin*, 2:160–163, 1991.

R. Sutton and A. Barto. *An Introduction to Reinforcement Learning*. MIT Press, 1998.

R. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 1057–1063, 2000.

U. Syed and R. Schapire. A game-theoretic approach to apprenticeship learning. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 1449–1456, 2008.

L. Tang, R. Rosales, A. Singh, and D. Agarwal. Automatic ad format selection via contextual bandits. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1587–1594. ACM, 2013.

G. Tesauro. TD-Gammon, a self-teaching backgammon program, achieves master-level play. *Neural Computation*, 6:215–219, 1994.

W. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, pages 285–294, 1933.

J. N. Tsitsiklis. A short proof of the gittins index theorem. *The Annals of Applied Probability*, pages 194–199, 1994.

N. Vien, H. Yu, and T. Chung. Hessian matrix distribution for Bayesian policy gradient reinforcement learning. *Information Sciences*, 181(9):1671–1685, 2011.

T. Walsh, S. Goschin, and M. Littman. Integrating sample-based planning and model-based reinforcement learning. In *Association for the Advancement of Artificial Intelligence*, 2010.

T. Wang, D. Lizotte, M. Bowling, and D. Schuurmans. Bayesian sparse sampling for on-line reward optimization. In *International Conference on Machine learning*, pages 956–963, 2005.

C. Watkins. *Learning from Delayed Rewards*. PhD thesis, Kings College, Cambridge, England, 1989.

R. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

A. Wilson, A. Fern, S. Ray, and P. Tadepalli. Multi-task reinforcement learning: A hierarchical Bayesian approach. In *Proceedings of the International Conference on Machine Learning*, pages 1015–1022, 2007.

B. Wittenmark. Adaptive dual control methods: An overview. In *5th IFAC symposium on Adaptive Systems in Control and Signal Processing*, 1995.

O. Zane. Discrete-time Bayesian adaptive control problems with complete information. In *IEEE Conference on Decision and Control*, 1992.

B. Ziebart, A. Maas, A. Bagnell, and A. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, pages 1433–1438, 2008.