# A Tutorial on Linear Function Approximators for Dynamic Programming and Reinforcement Learning

**Alborz Geramifard**
Massachusetts Institute of Technology

**Thomas J. Walsh**
Massachusetts Institute of Technology

**Stefanie Tellex**
Massachusetts Institute of Technology

**Girish Chowdhary**
Massachusetts Institute of Technology

**Nicholas Roy**
Massachusetts Institute of Technology

**Jonathan P. How**
Massachusetts Institute of Technology

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

A. Geramifard, T. J. Walsh, S. Tellex, G. Chowdhary, N. Roy, and J. P. How. *A Tutorial on Linear Function Approximators for Dynamic Programming and Reinforcement Learning*. Foundations and Trends® in Machine Learning, vol. 6, no. 4, pp. 375–451, 2013.

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

ISBN: 978-1-60198-761-7

**Foundations and Trends® in Machine Learning**
Volume 6, Issue 4, 2013

# 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, 2013, Volume 6, 4 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. 6, No. 4 (2013) 375–451

DOI: 10.1561/2200000042

# A Tutorial on Linear Function Approximators for Dynamic Programming and Reinforcement Learning

Alborz Geramifard
MIT LIDS
agf@mit.edu

Thomas J. Walsh
MIT LIDS
twalsh@mit.edu

Stefanie Tellex
MIT CSAIL
stefie10@csail.mit.edu

Girish Chowdhary
MIT LIDS
girishc@mit.edu

Nicholas Roy
MIT CSAIL
nickroy@mit.edu

Jonathan P. How
MIT LIDS
jhow@mit.edu

# Contents

## Abstract

A Markov Decision Process (MDP) is a natural framework for formulating sequential decision-making problems under uncertainty. In recent years, researchers have greatly advanced algorithms for learning and acting in MDPs. This article reviews such algorithms, beginning with well-known dynamic programming methods for solving MDPs such as policy iteration and value iteration, then describes approximate dynamic programming methods such as trajectory based value iteration, and finally moves to reinforcement learning methods such as Q-Learning, SARSA, and least-squares policy iteration. We describe algorithms in a unified framework, giving pseudocode together with memory and iteration complexity analysis for each. Empirical evaluations of these techniques with four representations across four domains, provide insight into how these algorithms perform with various feature sets in terms of running time and performance.



A. Geramifard, T. J. Walsh, S. Tellex, G. Chowdhary, N. Roy, and J. P. How. *A Tutorial on Linear Function Approximators for Dynamic Programming and Reinforcement Learning*. Foundations and Trends® in Machine Learning, vol. 6, no. 4, pp. 375–451, 2013.

DOI: 10.1561/2200000042.

# 1

# Introduction

Designing agents to act near-optimally in stochastic sequential domains is a challenging problem that has been studied in a variety of settings. When the domain is known, analytical techniques such as *dynamic programming* (DP) [Bellman, 1957] are often used to find optimal policies for the agent. When the domain is initially unknown, *reinforcement learning* (RL) [Sutton and Barto, 1998] is a popular technique for training agents to act optimally based on their experiences in the world. However, in much of the literature on these topics, small-scale environments were used to verify solutions. For example the famous taxi problem has only 500 states [Dietterich, 2000]. This contrasts with recent success stories in domains previously considered unassailable, such as $9 \times 9$ Go [Silver et al., 2012], a game with approximately $10^{38}$ states. An important factor in creating solutions for such large-scale problems is the use of linear function approximation [Sutton, 1996, Silver et al., 2012, Geramifard et al., 2011]. This approximation technique allows the long-term utility (value) of policies to be represented in a low-dimensional form, dramatically decreasing the number of parameters that need to be learned or stored. This tutorial provides practical guidance for researchers seeking to extend DP and RL techniques to larger domains through linear value function approximation. We introduce DP and RL techniques in a unified framework and conduct experiments in domains with sizes up to $\sim 150$ million state-action pairs.

Sequential decision making problems with full observability of the states are often cast as Markov Decision Processes (MDPs) [Puterman, 1994]. An MDP consists of a set of states, set of actions available to an agent, rewards earned in each state, and a model for transitioning to a new state given the current state and the action taken by the agent. Ignoring computational limitations, an *agent* with full knowledge of the MDP can compute an optimal policy that maximizes some function of its expected cumulative reward (which is often referred to as the expected *return* [Sutton and Barto, 1998]). This process is known as *planning*. In the case where the MDP is unknown, reinforcement learning agents *learn* to take optimal actions over time merely based on interacting with the world.

A central component for many algorithms that plan or learn to act in an MDP is a *value function*, which captures the long term expected return of a policy for every possible state. The construction of a value function is one of the few common components shared by many planners and the many forms of so-called *value-based* RL methods[1]. In the planning context, where the underlying MDP is known to the agent, the value of a state can be expressed recursively based on the value of successor states, enabling dynamic programming algorithms [Bellman, 1957] to iteratively estimate the value function. If the underlying model is unknown, value-based reinforcement learning methods estimate the value function based on observed state transitions and rewards. However, in either case, maintaining and manipulating the value of every state (*i.e.,* a *tabular representation*) is not feasible in large or continuous domains. In order to tackle practical problems with such large state-action spaces, a value function representation is needed that 1) does not require computation or memory proportional to the size of the number of states, and 2) *generalizes* learned values from data across states (*i.e.,* each new piece of data may change the value of more than one state).

One approach that satisfies these goals is to use linear function approximation to estimate the value function. Specifically, the full set of states is

[1]There are other MDP solving techniques not covered here (such as direct policy search) that do not directly estimate a value function and have been used successfully in many applications, including robotics [Williams, 1992, Sutton et al., 2000, Peters and Schaal, 2006, Baxter and Bartlett, 2000].

projected into a lower dimensional space where the value function is represented as a linear function. This representational technique has succeeded at finding good policies for problems with high dimensional state-spaces such as simulated soccer [Stone et al., 2005b] and Go [Silver et al., 2012]. This tutorial reviews the use of linear function approximation algorithms for decision making under uncertainty in DP and RL algorithms. We begin with classical DP methods for exact planning in decision problems, such as policy iteration and value iteration. Next, we describe approximate dynamic programming methods with linear value function approximation and "trajectory based" evaluations for practical planning in large state spaces. Finally, in the RL setting, we discuss learning algorithms that can utilize linear function approximation, namely: SARSA, Q-learning, and Least-Squares policy iteration. Throughout, we highlight the trade-offs between computation, memory complexity, and accuracy that underlie algorithms in these families.

In Chapter 3, we provide a more concrete overview of practical linear function approximation from the literature and discuss several methods for creating linear bases. We then give a thorough empirical comparison of the various algorithms described in the theoretical section paired with each of these representations. The algorithms are evaluated in multiple domains, several of which have state spaces that render tabular representations intractable. For instance, one of the domains we examine, Persistent Search and Track (PST), involves control of multiple unmanned aerial vehicles in a complex environment. The large number of properties for each robot (fuel level, location, etc.) leads to over $150$ million state-action pairs. We show that the linear function approximation techniques described in this tutorial provide tractable solutions for this otherwise unwieldy domain. For our experiments, we used the RLPy framework [Geramifard et al., 2013a] which allows the reproduction of our empirical results.

There are many existing textbooks and reviews of reinforcement learning [Bertsekas and Tsitsiklis, 1996, Szepesvári, 2010, Buşoniu et al., 2010, Gosavi, 2009, Kaelbling et al., 1996, Sutton and Barto, 1998]. This tutorial differentiates itself by providing a narrower focus on the use of linear value function approximation and introducing many DP and RL techniques in a unified framework, where each algorithm is derived from the general concept of policy evaluation/improvement (shown in Figure 2.1). Also, our extensive empirical evaluation covers a wider range of domains, representations, and algorithms than previous studies. The lessons from these experiments provide a guide to practitioners as they apply DP and RL methods to their own large-scale (and perhaps hitherto intractable) domains.

# References

RL competition. http://www.rl-competition.org/, 2012. Accessed: 20/08/2012.

M. Ahmadi, M. E. Taylor, and P. Stone. IFSA: incremental feature-set augmentation for reinforcement learning tasks. In *International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1–8, New York, NY, USA, 2007. ACM. .

A. Antos, R. Munos, and C. Szepesvári. Fitted Q-iteration in continuous action-space MDPs. In *Proceedings of Neural Information Processing Systems Conference (NIPS)*, 2007.

A. Antos, C. Szepesvári, and R. Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

J. Asmuth, L. Li, M. Littman, A. Nouri, and D. Wingate. A Bayesian sampling approach to exploration in reinforcement learning. In *International Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 19–26, Arlington, Virginia, United States, 2009. AUAI Press.

L. C. Baird. Residual algorithms: Reinforcement learning with function approximation. In *ICML*, pages 30–37, 1995.

A. d. M. S. Barreto and C. W. Anderson. Restricted gradient-descent algorithm for value-function approximation in reinforcement learning. *Artificial Intelligence*, 172:454 – 482, 2008.

A. Barto and M. Duff. Monte carlo matrix inversion and reinforcement learning. In *Neural Information Processing Systems (NIPS)*, pages 687–694. Morgan Kaufmann, 1994.

A. Barto, S. Bradtke, and S. Singh. Learning to act using real-time dynamic programming. *Artificial Intelligence*, 72:81–138, 1995.

J. Baxter and P. Bartlett. Direct gradient-based reinforcement learning. In *Circuits and Systems, 2000. Proceedings. ISCAS 2000 Geneva. The 2000 IEEE International Symposium on*, volume 3, pages 271–274. IEEE, 2000.

R. E. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, 1957.

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

D. P. Bertsekas. *Dynamic Programming and Optimal Control*. AP, 1976.

D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming (Optimization and Neural Computation Series, 3)*. Athena Scientific, May 1996.

B. Bethke and J. How. Approximate Dynamic Programming Using Bellman Residual Elimination and Gaussian Process Regression. In *American Control Conference (ACC)*, St. Louis, MO, 2009.

S. Bhatnagar, R. S. Sutton, M. Ghavamzadeh, and M. Lee. Incremental natural actor-critic algorithms. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 105–112. MIT Press, 2007.

M. Bowling, A. Geramifard, and D. Wingate. Sigma Point Policy Iteration. In *International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, volume 1, pages 379–386, Richland, SC, 2008.

J. Boyan and A. Moore. Generalization in reinforcement learning: Safely approximating the value function. In G. Tesauro, D. Touretzky, and T. Lee, editors, *Neural Information Processing Systems (NIPS)*, pages 369–376, Cambridge, MA, 1995. The MIT Press.

J. A. Boyan. Least-squares temporal difference learning. In *International Conference on Machine Learning (ICML)*, pages 49–56. Morgan Kaufmann, San Francisco, CA, 1999.

S. J. Bradtke and A. G. Barto. Linear least-squares algorithms for temporal difference learning. *Journal of Machine Learning Research (JMLR)*, 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, 2002.

L. Buşoniu, R. Babuška, B. De Schutter, and D. Ernst. *Reinforcement Learning and Dynamic Programming Using Function Approximators*. CRC Press, Boca Raton, Florida, 2010.

T. G. Dietterich. Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of Artificial Intelligence and Research (JAIR)*, 13(1):227–303, Nov. 2000.

A. Dutech, T. Edmunds, J. Kok, M. Lagoudakis, M. Littman, M. Riedmiller, B. Russell, B. Scherrer, R. Sutton, S. Timmer, et al. Reinforcement learning benchmarks and bake-offs II. In *Advances in Neural Information Processing Systems (NIPS) 17 Workshop*, 2005.

Y. Engel, S. Mannor, and R. Meir. Bayes meets bellman: The gaussian process approach to temporal difference learning. In *International Conference on Machine Learning (ICML)*, pages 154–161, 2003.

A. Farahmand. *Regularities in Sequential Decision-Making Problems*. PhD thesis, Department of Computing Science, University of Alberta, 2009.

A. Farahmand, M. Ghavamzadeh, C. Szepesvári, and S. Mannor. Regularized policy iteration. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 441–448. MIT Press, 2008.

P. Geibel and F. Wysotzki. Risk-sensitive reinforcement learning applied to chance constrained control. *Journal of Artificial Intelligence and Research (JAIR)*, 24, 2005.

A. Geramifard, F. Doshi, J. Redding, N. Roy, and J. How. Online discovery of feature dependencies. In L. Getoor and T. Scheffer, editors, *International Conference on Machine Learning (ICML)*, pages 881–888. ACM, June 2011.

A. Geramifard, J. Redding, J. Joseph, N. Roy, and J. P. How. Model estimation within planning and learning. In *American Control Conference (ACC)*, June 2012.

A. Geramifard, R. H. Klein, and J. P. How. RLPy: The Reinforcement Learning Library for Education and Research. `http://acl.mit.edu/RLPy`, April 2013a.

A. Geramifard, T. J. Walsh, N. Roy, and J. How. Batch iFDD: A Scalable Matching Pursuit Algorithm for Solving MDPs. In *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, Bellevue, Washington, USA, 2013b. AUAI Press.

S. Girgin and P. Preux. Feature Discovery in Reinforcement Learning using Genetic Programming. Research Report RR-6358, INRIA, 2007.

G. H. Golub and C. F. V. Loan. *Matrix Computations*. The John Hopkins University Press, 1996.

G. Gordon. Stable function approximation in dynamic programming. In *International Conference on Machine Learning (ICML)*, page 261, Tahoe City, California, July 9-12 1995. Morgan Kaufmann.

A. Gosavi. Reinforcement learning: A tutorial survey and recent advances. *INFORMS J. on Computing*, 21(2):178–192, April 2009.

H. Hachiya, T. Akiyama, M. Sugiyama, and J. Peters. Adaptive importance sampling with automatic model selection in value function approximation. In *Association for the Advancement of Artificial Intelligence (AAAI)*, pages 1351–1356, 2008.

S. Haykin. *Neural Networks: A Comprehensive Foundation*. Macmillan, New York, 1994.

R. A. Howard. *Dynamic Programming and Markov Processes*. MIT Press, Cambridge, Massachusetts, 1960.

T. Jaakkola, M. Jordan, and S. Singh. on the convergence of stochastic iterative dynamic programming algorithms. Technical report, Massachusetts Institute of Technology, Cambridge, MA, August 1993.

T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research (JMLR)*, 11:1563–1600, 2010.

W. Josemans. *Generalization in Reinforcement Learning*. PhD thesis, University of Amsterdam, 2009.

T. Jung and P. Stone. Gaussian processes for sample efficient reinforcement learning with RMAX-like exploration. In *European Conference on Machine Learning (ECML)*, September 2010.

L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence and Research (JAIR)*, 4:237–285, 1996.

S. Kalyanakrishnan and P. Stone. Characterizing reinforcement learning methods through parameterized learning problems. *Machine Learning*, 2011.

J. Z. Kolter and A. Y. Ng. Regularization and feature selection in least-squares temporal difference learning. In *International Conference on Machine Learning (ICML)*, pages 521–528, New York, NY, USA, 2009. ACM.

R. Kretchmar and C. Anderson. Comparison of cmacs and radial basis functions for local function approximators in reinforcement learning. In *International Conference on Neural Networks*, volume 2, pages 834–837 vol.2, 1997.

O. Kroemer and J. Peters. A non-parametric approach to dynamic programming. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1719–1727, 2011.

L. Kuvayev and R. Sutton. Model-based reinforcement learning with an approximate, learned model. In *Proceeding of the ninth Yale workshop on adaptive and learning systems*, pages 101–105, 1996.

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

L. Li. Sample complexity bounds of exploration. In M. Wiering and M. van Otterlo, editors, *Reinforcement Learning: State of the Art*. Springer Verlag, 2012.

L. Li, M. L. Littman, and C. R. Mansley. Online exploration in least-squares policy iteration. In *International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 733–739, Richland, SC, 2009a. International Foundation for Autonomous Agents and Multiagent Systems.

L. Li, J. D. Williams, and S. Balakrishnan. Reinforcement learning for dialog management using least-squares policy iteration and fast feature selection. In *New York Academy of Sciences Symposium on Machine Learning*, 2009b.

W. Liu, P. Pokharel, and J. Principe. The kernel least-mean-square algorithm. *IEEE Transactions on Signal Processing*, 56(2):543–554, 2008.

W. Liu, J. C. Principe, and S. Haykin. *Kernel Adaptive Filtering: A Comprehensive Introduction*. Wiley, Hoboken, New Jersey, 2010.

H. R. Maei and R. S. Sutton. GQ($\lambda$): A general gradient algorithm for temporal-difference prediction learning with eligibility traces. In *Proceedings of the Third Conference on Artificial General Intelligence (AGI)*, Lugano, Switzerland, 2010.

H. R. Maei, C. Szepesvári, S. Bhatnagar, and R. S. Sutton. Toward off-policy learning control with function approximation. In J. Fürnkranz and T. Joachims, editors, *International Conference on Machine Learning (ICML)*, pages 719–726. Omnipress, 2010.

S. Mahadevan. Representation policy iteration. *International Conference on Uncertainty in Artificial Intelligence (UAI)*, 2005.

Mausam and A. Kolobov. *Planning with Markov Decision Processes: An AI Perspective*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2012.

F. S. Melo, S. P. Meyn, and M. I. Ribeiro. An analysis of reinforcement learning with function approximation. In *International Conference on Machine Learning (ICML)*, pages 664–671, 2008.

O. Mihatsch and R. Neuneier. Risk-sensitive reinforcement learning. *Journal of Machine Learning Research (JMLR)*, 49(2-3):267–290, 2002.

J. Moody and C. J. Darken. Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1(2):281–294, June 1989.

A. W. Moore and C. G. Atkeson. Prioritized sweeping: Reinforcement learning with less data and less time. In *Machine Learning*, pages 103–130, 1993.

A. Nouri and M. L. Littman. Multi-resolution exploration in continuous spaces. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 1209–1216. MIT Press, 2009.

R. Parr, C. Painter-Wakefield, L. Li, and M. Littman. Analyzing feature generation for value-function approximation. In *International Conference on Machine Learning (ICML)*, pages 737–744, New York, NY, USA, 2007. ACM.

R. Parr, L. Li, G. Taylor, C. Painter-Wakefield, and M. L. Littman. An analysis of linear models, linear value-function approximation, and feature selection for reinforcement learning. In *International Conference on Machine Learning (ICML)*, pages 752–759, New York, NY, USA, 2008. ACM.

J. Peters and S. Schaal. Policy gradient methods for robotics. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2219–2225. IEEE, October 2006.

J. Peters and S. Schaal. Natural actor-critic. *Neurocomputing*, 71:1180–1190, March 2008.

M. Petrik, G. Taylor, R. Parr, and S. Zilberstein. Feature selection using regularization in approximate linear programs for Markov decision processes. In *International Conference on Machine Learning (ICML)*, 2010.

M. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 1994.

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

B. Ratitch and D. Precup. Sparse distributed memories for on-line value-based reinforcement learning. In *European Conference on Machine Learning (ECML)*, pages 347–358, 2004.

M. Riedmiller, J. Peters, and S. Schaal. Evaluation of policy gradient methods and variants on the Cart-Pole benchmark. In *IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning (ADPRL)*, pages 254–261, April 2007.

G. A. Rummery and M. Niranjan. Online Q-learning using connectionist systems (tech. rep. no. cued/f-infeng/tr 166). *Cambridge University Engineering Department*, 1994.

S. Sanner. International Probabilistic Planning Competition (IPPC) at International Joint Conference on Artificial Intelligence (IJCAI). `http://users.cecs.anu.edu.au/~ssanner/IPPC_2011/`, 2011. Accessed: 26/09/2012.

B. Scherrer. Should one compute the temporal difference fix point or minimize the bellman residual? the unified oblique projection view. In *International Conference on Machine Learning (ICML)*, 2010.

B. Schölkopf and A. Smola. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computations*, 10(5):1299–1319, 1998.

B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, 2002.

P. Schweitzer and A. Seidman. Generalized polynomial approximation in Markovian decision processes. *Journal of mathematical analysis and applications*, 110:568–582, 1985.

D. Silver, R. S. Sutton, and M. Müller. Sample-based learning and search with permanent and transient memories. In *International Conference on Machine Learning (ICML)*, pages 968–975, New York, NY, USA, 2008. ACM.

D. Silver, R. S. Sutton, and M. Müller. Temporal-difference search in computer go. *Machine Learning*, 87(2):183–219, 2012.

S. P. Singh. Reinforcement learning with a hierarchy of abstract models. In *Proceeding of the Tenth National Conference on Artificial Intelligence*, pages 202–207. MIT/AAAI Press, 1992.

S. P. Singh, T. Jaakkola, M. L. Littman, and C. Szepesvári. Convergence results for single-step on-policy reinforcement-learning algorithms. *Journal of Machine Learning Research (JMLR)*, 38:287–308, 2000.

P. Stone, R. S. Sutton, and G. Kuhlmann. Reinforcement learning for RoboCup-soccer keepaway. *International Society for Adaptive Behavior*, 13(3):165–188, 2005a.

P. Stone, R. S. Sutton, and G. Kuhlmann. Reinforcement learning for RoboCup soccer keepaway. *Adaptive Behavior*, 13(3):165–188, September 2005b.

A. L. Strehl, L. Li, and M. L. Littman. Reinforcement learning in finite mdps: Pac analysis. *Journal of Machine Learning Research (JMLR)*, 10:2413–2444, Dec. 2009.

R. S. Sutton. Generalization in reinforcement learning: Successful examples using sparse coarse coding. In *Neural Information Processing Systems (NIPS)*, pages 1038–1044. The MIT Press, 1996.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems (NIPS)*, 12(22):1057–1063, 2000.

R. S. Sutton, H. R. Maei, D. Precup, S. Bhatnagar, D. Silver, C. Szepesvári, and E. Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *International Conference on Machine Learning (ICML)*, pages 993–1000, New York, NY, USA, 2009. ACM.

C. Szepesvári. *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2010.

I. Szita and C. Szepesvári. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *International Conference on Machine Learning (ICML)*, pages 1031–1038, 2010.

G. Taylor and R. Parr. Kernelized value function approximation for reinforcement learning. In *International Conference on Machine Learning (ICML)*, pages 1017–1024, New York, NY, USA, 2009. ACM.

J. N. Tsitsiklis and B. V. Roy. An analysis of temporal difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, May 1997.

J. N. Tsitsiklis and B. V. Roy. Average cost temporal-difference learning. *Automatica*, 35(11):1799 – 1808, 1999.

N. K. Ure, A. Geramifard, G. Chowdhary, and J. P. How. Adaptive Planning for Markov Decision Processes with Uncertain Transition Models via Incremental Feature Dependency Discovery. In *European Conference on Machine Learning (ECML)*, 2012.

C. J. Watkins. *Models of Delayed Reinforcement Learning*. PhD thesis, Cambridge Univ., 1989.

C. J. Watkins. Q-learning. *Machine Learning*, 8(3):279–292, 1992.

C. J. C. H. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8(3):279–292, May 1992.

S. D. Whitehead. A complexity analysis of cooperative mechanisms in reinforcement learning. In *Association for the Advancement of Artificial Intelligence (AAAI)*, pages 607–613, 1991.

S. Whiteson and M. Littman. Introduction to the special issue on empirical evaluations in reinforcement learning. *Machine Learning*, pages 1–6, 2011.

B. Widrow and F. Smith. Pattern-recognizing control systems. In *Computer and Information Sciences: Collected Papers on Learning, Adaptation and Control in Information Systems, COINS symposium proceedings*, volume 12, pages 288–317, Washington DC, 1964.

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

T. Winograd. Procedures as a representation for data in a computer program for understanding natural language. Technical Report 235, Massachusetts Institute of Technology, 1971.

Y. Ye. The simplex and policy-iteration methods are strongly polynomial for the markov decision problem with a fixed discount rate. *Math. Oper. Res.*, 36(4): 593–603, 2011.

H. Yu and D. P. Bertsekas. Error bounds for approximations from projected linear equations. *Math. Oper. Res.*, 35(2):306–329, 2010.