A Bibliography of Work Related to Reinforcement Learning

Leslie Pack Kaelbling, Michael L. Littman

Department of Computer Science
Brown University
Providence, Rhode Island 02912

CS-94-39
September 1994

A Bibliography of Work Related to Reinforcement Learning

Leslie Pack Kaelbling
Michael L. Littman
Brown University

August 31, 1994

This bibliography is the compilation of a lot of work by a lot of people. Thanks go to

- Andrew Moore (awm@cs.cmu.edu)
- Frank Weber and Alexander Linden of the reinforcement learning archive at ftp.gmd.de (129.26.8.84) (frank.weber@gmd.de)
- Jude Shavlik (shavlik@cs.wisc.edu)
- Judy Franklin (jfranklin@gte.com)
- Jürgen Schmidhuber (schmidhu@informatik.tu-muenchen.de)
- Sven Koenig (koenig@cs.cmu.edu)
- Matthew McDonald (mafm@cs.uwa.edu.au)
- Engels (engelsg@uni-duesseldorf.de)
- Kevin Markey (markey@tigger.cs.colorado.edu)
- Russell W. Anderson (rwa@fisher.bio.uci.edu)
- John Moody moody@cse.ogi.edu

We are quite happy to accept additions, annotations, and corrections. Please send them to mlittman@cs.brown.edu. Also helpful would be a volunteer to sort this into sections by topic area and someone who can convert this into HTML with pointers to FTPable papers.

References

David H. Ackley and Michael L. Littman. Generalization and scaling in reinforcement learning. In *Advances in Neural Information Processing 2*, San Mateo, CA, 1989. Morgan Kaufmann.

*Introduces "complementary reinforcement backpropogation" (*CRBP*), a connectionist approach to choosing actions that give positive reward.*

Charles W. Anderson. *Learning and Problem Solving with Multilayer Connectionist Systems.* PhD thesis, University of Massachusetts, Amherst, MA, 1986.

Solves "Towers of Hanoi" with a connectionist reinforcement learning approach.

Charles W. Anderson. Strategy learning with multilayer connectionist representations. In *Proceedings of the Fourth International Workshop on Machine Learning*, pages 103–114, Ann Arbor, MI, 1987.

Charles W. Anderson. Learning to control an inverted pendulum using neural networks. *IEEE Control Systems Magazine*, 9(3):31–37, 1989.

Charles W. Anderson. Tower of Hanoi with Connectionist Networks: Learning New Features. In *Proceedings of the Sixth International Workshop on Machine Learning*, pages 345–349. Morgan Kaufmann, June 1989.

R. W. Anderson. *Stochastic Optimization of Neural Networks and Implications for Biological Learning.* PhD thesis, University of California, San Francisco, 1991.

R. W. Anderson and V. Vemuri. Neural networks can be used for open-loop, dynamic control. *International Journal of Neural Networks*, 2(3), 1992.

W. Ross Ashby. *Design for a Brain: The Origin of Adaptive Behaviour.* John Wiley and Sons, New York, NY, second edition, 1960.

C. G. Atkeson and D. J. Reinkensmeyer. Using Associative Content-Addressable Memories to Control Robots. In W. T. Miller, R. S. Sutton, and P. J. Werbos, editors, *Neural Networks for Control.* MIT Press, 1989.

J. R. Bachrach. *Connectionist Modeling and Control of Finite State Environments.* PhD thesis, Graduate School of the University of Massachusetts, Amherst, MA, February 1992.

Uses Rivest and Schapire's diversity representation to learn about hidden state environments.

Andrew G. Barto. Learning by statistical cooperation of self-interested neuron-like computing elements. *Human Neurobiology*, 5:229–256, 1985.

Andrew G. Barto. Connectionist learning for control. Technical Report 89-89, Department of Computer and Information Science, University of Massachusetts, Amherst, MA, 1989.

Andrew G. Barto. Connectionist learning for control. In W. Thomas Miller, III, Richard S. Sutton, and Paul J. Werbos, editors, *Neural networks for control*, chapter 1, pages 5–58. The MIT Press, Cambridge, MA, 1990.

Survey of reinforcement learning and other "neural" approaches to control.

Andrew G. Barto. Reinforcement learning and adaptive critic methods. In David A. White and Donald Sofge, editors, *Handbook of Intelligent Control: Neural, Fuzzy and Adaptive Approaches.* Van Nostrand Reinhold, 1992.

Andrew G. Barto. Artificial intelligence, neural networks, and control. In G. A. Bekey and K. Y. Goldberg, editors, *Neural Networks in Robotics.* Kluwer Academic Press, 1993.

Andrew G. Barto and P. Anandan. Pattern recognizing stochastic learning automata. *IEEE Transactions on Systems, Man, and Cybernetics*, 15:360–374, 1985.

Andrew G. Barto, Steven J. Bradtke, and Satinder Pal Singh. Real-time learning and control using asynchronous dynamic programming. Technical Report 91-57, Department of Computer and Information Science, University of Massachusetts, Amherst, MA, 1991.

Andrew G. Barto, Steven J. Bradtke, and Satinder Pal Singh. Learning to act using real-time dynamic programming. Technical Report 93-02, Department of Computer and Information Science, University of Massachusetts, Amherst, MA, 1993.

To appear in Artificial Intelligence

Andrew G. Barto and Michael I. Jordan. Gradient following without back-propagation in layered networks. In *Proceedings of the IEEE First International Conference on Neural Networks*, volume 2, pages 629–636, San Diego, CA, 1987.

Andrew G. Barto and Richard S. Sutton. Time-derivative models of Pavlovian reinforcement. In *Learning and Computational Neuroscience: Foundations of Adaptive Networks.* The MIT Press, Cambridge, MA, 1990.

Andrew G. Barto, Richard S. Sutton, and Charles W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(5):834–846, 1983.

Andrew G. Barto, Richard S. Sutton, and Peter S. Brouwer. Associative search network: A reinforcement learning associative memory. *Biological Cybernetics*, 40:201–211, 1981.

Andrew. G. Barto, Richard. S. Sutton, and Christopher J. C. H. Watkins. Learning and sequential decision making. Technical Report 89-95, Department of Computer and Information Science, University of Massachusetts, Amherst, MA, 1989.

Also published in Learning and Computational Neuroscience: Foundations of Adaptive Networks, Michael Gabriel and John Moore, editors. The MIT Press, Cambridge, MA, 1991.

Richard Bellman. *Dynamic Programming.* Princeton University Press, Princeton, NJ, 1957.

"The" book on dynamic programming by Mr. Dynamic Programming. Discusses Markov Decision Processes.

Hamid Benbrahim, John Doleac, Judy A. Franklin, and Oliver G. Selfridge. Real-time learning: A ball on a beam. In *Proceedings of the 1992 International Joint Conference on Neural Networks*, Baltimore, MD, June 1992.

2-Action, Actor-Critic, Robotics, Ball Balancing, Hardware Implementation, Boxes

Hamid Benbrahim and Judy A. Franklin. Self-scaling reinforcement: A new algorithm for learning continuous control. In *Proceedings of the Eighth Yale Workshop on Adaptation and Learning*, Yale University, New Haven, CT, June 1994.

Real-valued Output, Actor-Critic, Gain Adaptation, Robotics, Ball Balancing, Hardware Implementation, CMAC

Donald A. Berry and Bert Fristedt. *Bandit Problems: Sequential Allocation of Experiments.* Chapman and Hall, London, UK, 1985.

Dmitri P. Bertsekas. *Dynamic Programming: Deterministic and Stochastic Models.* Prentice-Hall, 1987.

*Covers many important results in the theory of Markov decision processes (*MDP*).*

Dmitri P. Bertsekas and John N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods.* Prentice-Hall, Englewood Cliffs, NJ, 1989.

Justin A. Boyan. Modular neural networks for learning context-dependent game strategies. Master's thesis, University of Cambridge, Cambridge, UK, August 1992.

Uses connectionist representations and reinforcement learning to play games include backgammon and tic-tac-toe.

Steven J. Bradtke. Reinforcement learning applied to linear quadratic regulation. In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems 5: Proceedings of the 1992 Conference*, pages 295–302, San Mateo, CA, 1993. Morgan Kaufmann.

H. J. Bremermann and R. W. Anderson. How the brain adjusts synapses—maybe. In R. S. Boyer, editor, *Automated Reasoning: Essays in Honor of Woody Bledsoe.* Kluwer, New York, NY, 1991.

Robert R. Bush and William K. Estes, editors. *Studies in Mathematical Learning Theory.* Stanford University Press, Stanford, CA, 1959.

P. V. C. Caironi and M. Dorigo. Training q-agents. Technical Report IRIDIA/94-14, Université Libre de Bruxelles, Belgium, 1994.

Anthony R. Cassandra, Leslie Pack Kaelbling, and Michael L. Littman. Acting optimally in partially observable stochastic domains. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, Seattle, WA, 1994.

*Discusses the use of partially-observable Markov decision processes (*POMDP*) to represent hidden state planning or reinforcement learning problems.*

David Chapman and Leslie Pack Kaelbling. Input generalization in delayed reinforcement learning: An algorithm and performance comparisons. In *Proceedings of the International Joint Conference on Artificial Intelligence*, Sydney, Australia, 1991.

Hsien-Te Cheng. *Algorithms for Partially Observable Markov Decision Processes.* PhD thesis, University of British Columbia, British Columbia, Canada, 1988.

Includes empirical comparisons with other algorithms. The "linear support" algorithm is arguably the fastest exact algorithm for solving POMDP *problems.*

Lonnie Chrisman. Reinforcement learning with perceptual aliasing: The perceptual distinctions approach. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, pages 183–188, San Jose, CA, 1992. AAAI Press.

Introduces POMDP*'s as a model of reinforcement learning in hidden state environments. Describes an* EM *algorithm for learning a* POMDP *model from data and uses a very rough approximation for optimal policies.*

Margaret E. Connell and Paul E. Utgoff. Learning to control a dynamic physical system. In *Proceedings of the Sixth National Conference on Artificial Intelligence*, volume 2, pages 456–460, Seattle, WA, 1987. Morgan Kaufmann.

Peter Dayan. The convergence of TD(λ) for general λ. *Machine Learning*, 8(3):341–362, 1992.

Peter Dayan and Geoffrey E. Hinton. Feudal reinforcement learning. In *Advances in Neural Information Processing Systems 5*, San Mateo, CA, 1993. Morgan Kaufmann.

A hierarchical model of reinforcement learning.

Peter Dayan and Terrence J. Sejnowski. The variance of covariance rules for associative matrix and reinforcement learning. *Neural Computation*, 5:205–209, 1993.

Peter Dayan and Terrence J. Sejnowski. TD(λ) converges with probability 1. *Machine Learning*, 14(3), 1994.

Thomas Dean, Leslie Pack Kaelbling, Jak Kirman, and Ann Nicholson. Planning with deadlines in stochastic domains. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, Washington, DC, 1993.

Describes the PLEXUS *algorithm for producing acceptable solutions to large* MDP *problems quickly.*

M. Dorigo. Alecsys and the AutonoMouse: Learning to control a real robot by distributed classifier systems. *Machine Learning*, to appear.

M. Dorigo and H. Bersini. A comparison of q-learning and classifier systems. In *From Animals to Animats: Proceedings of the Third International Conference on the Simulation of Adaptive Behavior*, Brighton, UK, 1994.

M. Dorigo and M. Colombetti. Robot shaping: Developing autonomous agents through learning. *Artificial Intelligence*, in press.

Gary L. Drescher. *Made-up Minds: A Constructivist Approach to Artificial Intelligence.* The MIT Press, Cambridge, MA, 1991.

Describes a Piagetian-style development scheme for creating intelligent behavior.

William K. Estes. Toward a statistical theory of learning. *Psychological Review*, 57:94–107, 1950.

Judy A. Franklin. Learning control in a robotic system. In *The Proceedings of the 1987 IEEE International Conference on Systems Man, and Cybernetics*, Alexandria, VA, 1987.

Hybrid Control, 2-action, ARP*, Boxes*

Judy A. Franklin. *Compliance and Learning: Control Skills for a Robot Operating in an Uncertain World.* PhD thesis, University of Massachusetts, Amherst, MA, 1988.

Hybrid Control, 2-action, ARP*, Boxes, Conventional Control, Robotics, Hardware Implementation*

Judy A. Franklin. Refinement of robot motor skills through reinforcement learning. In *Proceedings of the* 27^{th} *IEEE Conference on Decision and Control*, Austin, TX, 1988.

Hybrid Control, Real-valued Actions, Boxes

Judy A. Franklin. Historical perspective and state of the art in connectionist learning control. In *The Proceedings of the 28th IEEE Conference on Decision and Control*, Tampa, FL, 1989.

Judy A. Franklin. Input representation for refinement learning control. In *Proceedings of the 4th International IEEE Symposium on Intelligent Control*, Albany, NY, September, 1989.

Hybrid Control, Real-valued Actions, Radial-basis Functions

Judy A. Franklin, M. D. Smith, and J. C. Yun. Learning channel allocation strategies in real time. In *Proceedings of the IEEE Vehicular Technology Conference 1992*, Denver, CO, 1992.

Wireless Communication, Hardware Implementation, Actor-Critic, Boxes

Judy A. Franklin and D. A. White. Artificial neural networks in manufacturing and process control. In D. A. White and D. A. Sofge, editors, *Handbook of Intelligent Control: Neural, Fuzzy, and Adaptive Approaches*, Florence, KY, 1992. Van Nostrand Reinhold.

Reinforcement learning and supervised learning applied to manufacturing, CMAC*, Actor-Critic, robotics*

King-Sun Fu. Learning control systems—review and outlook. *IEEE Transactions on Automatic Control*, 15(2):210–221, April 1970.

D. Gachet, M.A. Salichs, E. Moreno, and J.R. Pimentel. Learning emergent tasks for an autonomous mobile robot. To appear in Proc. IROS 94 Conference, 1994.

John J. Grefenstette. Incremental learning of control strategies with genetic algorithms. In *Proceedings of the Sixth International Workshop on Machine Learning*, pages 340–344, Ithaca, NY, 1989. Morgan Kaufmann.

Vijay Gullapalli. A stochastic reinforcement learning algorithm for learning real-valued functions. *Neural Networks*, 3:671–692, 1990.

Real-Valued Action, Convergence Proof

Vijay Gullapalli. A comparison of supervised and reinforcement learning methods on a reinforcement learning task. In *Proceedings of the 1991 IEEE Symposium on Intelligent Control.* Arlington, VA, 1991.

Vijay Gullapalli. *Reinforcement Learning and its application to control.* PhD thesis, University of Massachusetts, Amherst, MA, 1992.

Real-Valued Action, Convergence Proof, Robotics

Vijay Gullapalli. Learning control under extreme uncertainty. In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems 5: Proceedings of the 1992 Conference*, pages 327–334, San Mateo, CA, 1993. Morgan Kaufmann.

Vijay Gullapalli, Judy A. Franklin, and Hamid Benbrahim. Acquisition of robot skills using reinforcement learning: Two hardware demonstrations. *IEEE Control Systems Magazine*, February, 1993.

Real-Valued Action, Convergence Proof, Robotics, Actor-Critic, Ball Balancing, Hardware Implementation, Boxes, Backpropagation

Vijay Gullapalli, R. A. Grupen, and A. G. Barto. Learning reactive admittance control. In *Proceedings of the 1992 IEEE International Conference on Robotics and Automation*, pages 1475–1480, Nice, France, May 1992.

Real-Valued Action, Robotics, Hardware Implementation

P. D. Heermann. *Neural Network Techniques for Stable Learning Control of Nonlinear Systems.* PhD thesis, University of Texas at Austin, Austin, TX, May 1992.

M. Heger. Consideration of risk in reinforcement learning. Revised submission to the 11th International Machine Learning Conference ML94, 1994.

Points out drawbacks of expected return as a measure of performance for policies and gives dynamic programming algorithms, including Q-Learning, to optimize the worst-case discounted return.

D. Heyman and M. Sobel. *Stochastic Models in Operations Research: Stochastic Optimization*, volume 2. McGraw-Hill, New York, NY, 1984.

Good reference on MDP*'s. Also covers other models that may be of interest to reinforcement learning researchers.*

Ernest R. Hilgard and Gordon H. Bower. *Theories of Learning.* Prentice-Hall, Englewood Cliffs, NJ, fourth edition, 1975.

Ronald A. Howard. *Dynamic Programming and Markov Processes.* The MIT Press, Cambridge, MA, 1960.

Classic reference for the policy iteration method for solving MDP*'s.*

Tommi Jaakkola, Michael I. Jordan, and Satinder P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. Technical Report 9307, MIT Computational Cognitive Science, 1993.

Convergence theorem for a class of dynamic-programming-based stochastic processes including Q-learning and TD*(λ). Submitted to Neural Computation*

T. T. Jervis and F. Fallside. Pole balancing on a real rig using a reinforcement learning controller. Technical Report CUED/F-INFENG/TR 115, Cambridge University Engineering Department, 1992.

Michael Jordan and David E. Rumelhart. Forward models: Supervised learning with a distal teacher. Technical Report Occasional Paper #40, MIT Center for Cognitve Science, Cambridge, MA, 1990.

A revised version has appeared in Cognitive Science, Vol. 16, 1992

Leslie Pack Kaelbling. Learning functions in k-dnf from reinforcement. In *Proceedings of the Seventh International Conference on Machine Learning*, pages 162–169, Austin, TX, 1990. Morgan Kaufmann.

Leslie Pack Kaelbling. Foundations of learning in autonomous agents. *Robotics and Autonomous Systems*, 6(2), 1991.

Also published in Toward Learning Robots, W. Van de Velde, Ed., The MIT Press, 1991

Leslie Pack Kaelbling. Hierarchical learning in stochastic domains: Preliminary results. In *Proceedings of the Tenth International Conference on Machine Learning*, Amherst, MA, 1993. Morgan Kaufmann.

Leslie Pack Kaelbling. *Learning in Embedded Systems.* The MIT Press, Cambridge, MA, 1993.

Also available as a PhD Thesis from Stanford University, 1990

Leslie Pack Kaelbling. Learning to achieve goals. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, Chambéry, France, 1993. Morgan Kaufmann.

Leslie Pack Kaelbling. Associative reinforcement learning: A generate and test algorithm. *Machine Learning*, 15(3), 1994.

Leslie Pack Kaelbling. Associative reinforcement learning: Functions in k-DNF. *Machine Learning*, 15(3), 1994.

John G. Kemeny and J. Laurie Snell. *Finite Markov Chains.* Springer-Verlag, New York, NY, 1976.

Sven Koenig and Reid G. Simmons. Complexity analysis of real-time reinforcement learning. In *Proceedings of the 11th National Conference on Artificial Intelligence AAAI-93*, pages 99–105, 1993.

Shows how a very simple scheme based on state values can be used to explore an environment in cubic time.

Sven Koenig and Reid G. Simmons. Complexity analysis of real-time reinforcement learning applied to finding shortest paths in deterministic domains. Technical Report CMU-CS-93-106, School of Computer Science, Carnegie Mellon University, 1993.

Sven Koenig and Reid G. Simmons. How to make reactive planners risk-sensitive. In *Proceedings of the 2nd International Conference on Artificial Intelligence Planning Systems*, pages 293–304, 1994.

Sven Koenig and Reid G. Simmons. Risk-sensitive planning with probabilistic decision graphs. In *Proceedings of the 4th International Conference on Principles of Knowledge Representation and Resoning*, pages 363–373, 1994.

Tze Leung Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, 15(3):1091–1114, 1987.

Gunar E. Liepins, Michael R. Hilliard, Mark Palmer, and Gita Rangarajan. Alternatives for classifier system credit assignment. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence (IJCAI-89)*, pages 756–761, 1989.

Describes bucket brigade, classifer chunking and backwards averaging and compares them on a simple problem. Also compares the resulting algorithm to Sutton's AHC *and* TD *systems.*

Long-Ji Lin. Programming robots using reinforcement learning and teaching. In *Proceedings of the Ninth National Conference on Artificial Intelligence*, 1991.

Long-Ji Lin. Self-improving based on reinforcement learning, planning, and teaching. In *Proceedings of the Eighth International Workshop on Machine Learning*, Evanston, IL, 1991. Morgan Kaufmann.

Long-Ji Lin. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*, 8(3):293–322, 1992.

Long-Ji Lin. Hierachical learning of robot skills by reinforcement. In *Proceedings of the International Conference on Neural Networks*, 1993.

Long-Ji Lin. *Reinforcement Learning for Robots Using Neural Networks.* PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 1993.

Long-Ji Lin and Tom M. Mitchell. Memory approaches to reinforcement learning in non-markovian domains. Technical Report CMU-CS-92-138, Carnegie Mellon University, School of Computer Science, May 1992.

Empirically compares several connectionist approaches to reinforcement learning in hidden state environments include window-based, model-based, and recurrent value function methods

Long-Ji Lin and Stephen Hanson. On-line Learning for Indoor Navigation: Preliminary Results with RatBot". Talk given at NIPS workshop on Robot Learning, 1994.

A. Linden. On discontinuous Q-functions in reinforcement learning. In H. J. Ohlbach, editor, *GWAI-92: Advances in Artificial Intelligence*, pages 199–209. Springer, 1993.

About discontinuous cumulative evaluation functions in reinforcement learning, which are present even in very simple robotic path finding domains. Lecture Notes in Artificial Intelligence, Vol. 671

Michael Littman and Justin Boyan. A distributed reinforcement learning scheme for network routing. Technical Report CMU-CS-93-165, School of Computer Science, Carnegie Mellon University, 1993.

Shows how to route packets efficiently using a network of Q-learners.

Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the Machine Learning Conference*, 1994.

Describes a Q-learning-like algorithm (minimax-Q) for learning to play two-player games. Applies it to a simple game in which optimal actions must sometimes be chosen probabilistically.

Michael L. Littman. Memoryless policies: Theoretical limitations and practical results. In *From Animals to Animats 3*, Brighton, UK, 1994.

Shows that the problem of finding the best mapping from observation to action in hidden state environments is NP*-complete.*

William S. Lovejoy. A survey of algorithmic methods for partially observed markov decision processes. *Annals of Operations Research*, 28(1):47–65, 1991.

Excellent review of algorithms for solving POMDP*'s.*

G. Lukes. Review of Schmidhuber's paper 'Recurrent networks adjusted by adaptive critics'. *Neural Network Reviews*, 4(1):41–42, 1990.

R. Maclin and J. Shavlik. Incorporating advice into agents that learn from reinforcement. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, Seattle, WA, 1994. AAAI Press.

Pattie Maes and Rodney A. Brooks. Learning to coordinate behaviors. In *Proceedings Eighth National Conference on Artificial Intelligence*, pages 796–802. AAAI, Morgan Kaufmann, 1990.

Hardware Implementation, Robotics

Pattie Maes and Robyn Kozierok. Learning interface agents. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, Washington, DC, 1993.

S. Mahadevan. Enhancing Transfer in Reinforcement Learning by Building Stochastic Models of Robot Actions. In *Machine Learning: Proceedings of the Ninth International Workshop*. Morgan Kaufmann, June 1992.

Sridhar Mahadevan and Jonathan Connell. Automatic programming of behavior-based robots using reinforcement learning. In *Proceedings of the Ninth National Conference on Artificial Intelligence*, Anaheim, CA, 1991.

Sridhar Mahadevan and Jonathan Connell. Scaling reinforcement learning to robotics by exploiting the subsumption architecture. In *Proceedings of the Eighth International Workshop on Machine Learning*, pages 328–332, 1991.

M. J. Mataric. Reward Functions for Accelerated Learning. In W. W. Cohen and H. Hirsh, editors, *Proceedings of the Eleventh International Conference on Machine Learning*. Morgan Kaufmann, 1994.

R. Andrew McCallum. Overcoming incomplete perception with utile distinction memory. In *Proceedings of the Tenth International Conference on Machine Learning*, Amherst, MA, 1993. Morgan Kaufmann.

Gives an alternative to Chrisman's approach to reinforcement learning in hidden state environments where states are distinguished by whether or not they lead to different rewards.

L. Meeden, G. McGraw, and D. Blank. Emergent control and planning in an autonomous vehicle. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, pages 735–740, Hillsdale, NJ, 1993. Erlbaum.

Describes a connectionist network (namely CRBP*) trained with reinforcement to control an autonomous robot vehicle and examines the plan-like behavior that emerges when the robot is given a goal-based problem.*

Bartlett W. Mel. Building and using mental models in a sensory-motor domain: A connectionist approach. In *Proceedings of the Fifth International Conference on Machine Learning*, pages 207–213, Ann Arbor, MI, 1988.

D. Michie and R. A. Chambers. BOXES: An experiment in adaptive control. In E. Dale and D. Michie, editors, *Machine Intelligence 2*. Oliver and Boyd, Edinburgh, UK, 1968.

W. T. Miller. Real-Time Application of Neural Networks for Sensor-Based Control of Robots with Vision. *IEEE Trans. on systems, Man and Cybernetics*, 19(4):825–831, July 1989.

W. T. Miller, F. H. Glanz, and L. G. Kraft. Application of a General Learning Algorithm to the Control of Robotic Manipulators. *International Journal of Robotics Research*, 6(2), 1987.

W. Thomas Miller III. Sensor-based control of robotic manipulators using a general learning algorithm. *IEEE Journal of Robotics and Automation*, RA-3(2):157–165, 1987.

W. Thomas Miller, III, Richard S. Sutton, and Paul J. Werbos, editors. *Neural Networks for Control*. The MIT Press, Cambridge, MA, 1990.

Hisashi Mine and Shunji Osaki. *Markovian Decision Processes*. American Elsevier, New York, NY, 1970.

Marvin L. Minsky. *Theory of Neural-Analog Reinforcement Systems and Its Application to the Brain-Model Problem*. PhD thesis, Princeton University, Princeton, NJ, 1954.

Thomas M. Mitchell and Sebastian B. Thrun. Explanation-based neural network learning for robot control. In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems 5: Proceedings of the 1992 Conference*, pages 287–294, San Mateo, CA, 1993. Morgan Kaufmann.

Yoshiro Miyata. An unsupervised PDP learning model for action planning. In *Proceedings of the 10th Annual Conference of the Cognitive Science Society*, pages 223–229, 1988.

Describes a hierarchical architecture that learns a truck backing task.

George E. Monahan. A survey of partially observable Markov decision processes: Theory, models, and algorithms. *Management Science*, 28(1):1–16, 1982.

The title says it all. Includes an original algorithm erroneously attributed to Sondik that has come to be called "Monahan's algorithm."

John Moody and Volker Tresp. A trivial but fast reinforcement controller. *Neural Computation*, To appear.

A. W. Moore. Fast, Robust Adaptive Control by Learning only Forward Models. In J. E. Moody, S. J. Hanson, and R. P. Lippman, editors, *Advances in Neural Information Processing Systems 4*. Morgan Kaufmann, April 1992.

A. W. Moore. Variable Resolution Reinforcement Learning. In *Proceedings of the Eighth Yale Workshop on Adaptive and Learning Systems*, 1994.

Prioritized sweeping with geometry.

A. W. Moore and C. G. Atkeson. The Parti-game Algorithm for Variable Resolution Reinforcement Learning in Multidimensional State-spaces. Accepted for Publication in Machine Learning, 1994.

Andrew W. Moore. Acquisition of dynamic control knowledge for a robotic manipulator. In *Proceedings of the Seventh International Conference on Machine Learning*, pages 244–252, Austin, TX, 1990. Morgan Kaufmann.

Andrew W. Moore. Variable resolution dynamic programming: Efficiently learning action maps in multivariate real-valued spaces. In *Proc. Eighth International Machine Learning Workshop*, 1991.

Andrew W. Moore. The parti-game algorithm for variable resolution reinforcement learning in multidimensional state spaces. In S. J. Hanson, J. D Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems 6*, San Mateo, CA, 1994. Morgan Kaufman.

Gives a method for function approximation over large state spaces based on k-d trees. Uses a game theoretic approach to deciding when additional resolution is needed in the representation.

Andrew W. Moore and Christopher G. Atkeson. Memory-based reinforcement learning: Efficient computation with prioritized sweeping. In *Advances in Neural Information Processing 5*, San Mateo, CA, 1993. Morgan Kaufmann.

Uses a model of the environment to make value back ups as efficient as possible.

Paul Munro. A dual back-propagation scheme for scalar reward learning. In *Proceedings of the Ninth Conference of the Cognitive Science Society*, pages 165–176, Seattle, WA, 1987.

Kumpati Narendra and M. A. L. Thathachar. *Learning Automata: An Introduction*. Prentice-Hall, Englewood Cliffs, NJ, 1989.

Kumpati S. Narendra and M. A. L. Thathachar. Learning automata—a survey. *IEEE Transactions on Systems, Man, and Cybernetics*, 4(4):323–334, July 1974.

Derrick Nguyen and Bernard Widrow. The truck backer-upper: An example of self-learning in neural networks. In W. Thomas Miller, III, Richard S. Sutton, and Paul J. Werbos, editors, *Neural networks for control*, chapter 12. The MIT Press, Cambridge, MA, 1990.

Backpropagation through trucks. Train a network to model the kinematics of a truck, using randomly generated data. Then run a controller and the truck recurrently until something is hit. Then run BP-*through-time (replacing the truck with the modelling net, because you can't differentiate a truck.)*

Christos H. Papadimitriou and John N. Tsitsiklis. The complexity of Markov decision processes. *Mathematics of Operations Research*, 12(3):441–450, August 1987.

Addresses the computational complexity of MDP *problems. Shows that solving* MDP*'s is* P*-complete (*MDP*'s are probably inherently sequential) and that solving finite-horizon* POMDP*'s is* PSPACE*-complete (worse than* NP*-complete).*

Jing Peng and Ronald J. Williams. Efficient learning and planning within the dyna framework. *Adaptive Behavior*, 1(4):437–454, 1993.

Like Prioritized-Sweeping, uses a learned model to propogate reward information efficiently.

R.A. Rescorla and A. R. Wagner. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical Conditioning II*. Appleton-Century-Crofts, 1972.

Mark Ring. Two methods for hierarchy learning in reinforcement environments. In I. H. Meyer, H. Roithlat, and S. Wilson, editors, *From Animals to Animats: Proc. Second International Conference on Simulation and Adaptive Behavior*. MIT Press, 1993.

Describes novel algorithms for reinforcement learning in hidden state environments based on learning macro-type sequences.

B. E. Rosen, J. M. Goodwin, and J. J. Vidal. Adaptive range coding. In R. P. Lippmann, J. E. Moody, and D. S. Touretzky, editors, *Advances in Neural Information Processing Systems 3: Proceedings of the 1990 Conference*, pages 486–492, San Mateo, CA, 1991. Morgan Kaufmann.

Sheldon M. Ross. *Introduction to Stochastic Dynamic Programming*. Academic Press, New York, NY, 1983.

A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3:211–229, 1959.

Reprinted in E. A. Feigenbaum and J. Feldman, editors, Computers and Thought, McGraw-Hill, NY 1963

A. L. Samuel. Some studies in machine learning using the game of checkers II—Recent progress. *IBM Journal of Research and Development*, pages 601–617, 1967.

S. Schaal and C. G. Atkeson. Assessing the Quality of Local Linear Models. In *Advances in Neural Information Processing Systems 6*. Morgan Kaufmann, April 1994.

J. H. Schmidhuber and R. Huber. Learning to generate artificial fovea trajectories for target detection. *International Journal of Neural Systems*, 2(1 & 2):135–141, 1991.

The goal is to to find details of certain objects in simple visual scenes. The task is of the "reward-only-at-goal" type. A recurrent model/controller system learns to generate appropriate sequences of retinal movements (translations/rotations).

Jürgen Schmidhuber and Reiner Wahnsiedler. Planning simple trajectories using neural subgoal generators. In *Proceedings of the Second International Conference on Simulation of Adaptive Behavior, 1992, Hawaii*, 1993.

Jürgen H. Schmidhuber. A local learning algorithm for dynamic feedforward and recurrent networks. *Connection Science*, 1(4):403–412, 1989.

Describes the "Neural Bucket Brigade Algorithm": connections that contribute to the activation of an output unit in the moment of external reward become stronger. To perform spatio-temporal credit assignment, connections that contribute to the activation of any unit (perhaps hidden in time or space) have to pay "weight substance" to those connections "setting the stage" one time step earlier. Only action sequences leading to success survive.

Jürgen H. Schmidhuber. Making the world differentiable: On using fully recurrent self-supervised neural networks for dynamic reinforcement learning and planning in non-stationary environments. Technical Report FKI-126-90 (revised), Institut für Informatik, Technische Universität München, November 1990.

Revised and extended version of an earlier report from February

Jürgen H. Schmidhuber. Recurrent networks adjusted by adaptive critics. In *Proc. IEEE/INNS International Joint Conference on Neural Networks*, volume 1, pages 719–722, Washington, DC, 1990.

An attempt to apply adaptive critics to recurrent nets. Can learn simple tasks like delayed XOR *by reinforcement learning.*

Jürgen H. Schmidhuber. Curious model-building control systems. In *Proc. International Joint Conference on Neural Networks*, volume 2, pages 1458–1463. IEEE, 1991.

A method for learning world models of non-deterministic environments faster than by random exploration, by translating improvements of world model quality into reinforcement for a Q-learner.

Jürgen H. Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In J. A. Meyer and S. W. Wilson, editors, *Proceedings of the International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, pages 222–227. The MIT Press, 1991.

Jürgen H. Schmidhuber. Reinforcement learning in markovian and non-markovian environments. In D. S. Lippman, J. E. Moody, and D. S. Touretzky, editors, *Advances in Neural Information Processing Systems 3*, pages 500–506, San Mateo, CA, 1991. Morgan Kaufmann.

Description of interacting recurrent nets for learning to deal with Non-Markovian interfaces to the external world. Can solve simple tasks that require looking back into time (like learning to be a flipflop), by using delayed scalar reinforcement signals only.

Nicol N. Schraudolph, Peter Dayan, and Terrence J. Sejnowski. Using the td(lambda) algorithm to learn an evaluation function for the game of go. In *Advances in Neural Information Processing Systems 6*, San Mateo, CA, 1994. Morgan Kaufman.

Applies reinforcement learning to the problem of learning to play the game of Go with less success than has been attained for backgammon. To appear.

Anton Schwartz. A reinforcement learning method for maximizing undiscounted rewards. In *Proceedings of the Tenth International Conference on Machine Learning*, Amherst, MA, 1993. Morgan Kaufmann.

Introduces R-learning, a method for maximizing average per-step reward. Argues that discounting is often unmotivated and that average reward is often more in line with researchers' needs.

Anton Schwartz. Thinking Locally to Act Globally: A Novel Approach to Reinforcement Learning. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, Hillsdale, NJ, 1993. Lawrence Erlbaum Associates.

Oliver G. Selfridge and Richard S. Sutton. Training and tracking in robotics. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pages 670–672, Los Angeles, CA, 1985. Morgan Kaufmann.

Wei-Min Shen. *Learning from the Environment Based on Percepts and Actions.* PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 1989.

J. Simons, H. V. Brussel, J. D. Schutter, and J. Verhaert. A self-learning automaton with variable resolution for high precision assembly by industrial robots. *IEEE Transactions on Automatic Control*, 27(5):1109–1113, October 1982.

Robotics, Hardware Implementation, Compliance, Historical

S. P. Singh. Robust Reinforcement Learning in Motion Planning. In S. J. Hanson, J. D Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems 6*. Morgan Kaufmann, April 1994.

Shows how safe 'high-level' actions can be used in a traditional reinforcement learning framework.

Satinder P. Singh. The efficient learning of multiple task sequences. In *Advances in Neural Information Processing Systems 5: Proceedings of the 1992 Conference*, 1992.

Gives a hierarchical method for representing policies based on decomposing the Q-function.

Satinder Pal Singh. Reinforcement learning with a hierarchy of abstract models. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, pages 202–207, San Jose, CA, 1992. AAAI Press.

Satinder Pal Singh. Scaling reinforcement learning algorithms by learning variable temporal resolution models. In *Proceedings of the Ninth International Conference on Machine Learning*, pages 406–415, Aberdeen, Scotland, UK, 1992. Morgan Kaufmann.

Satinder Pal Singh. Transfer of learning by composing solutions of elemental sequential tasks. *Machine Learning*, 8(3):323–340, 1992.

Satinder Pal Singh, Tommi Jaakkola, and Michael I. Jordan. Model-free reinforcement learning for non-Markovian decision problems. In *Proceedings of the Machine Learning Conference*, 1994.

Sets up the basic problem of learning probabilistic memoryless policies for hidden state environments. Shows why existing algorithms solve the wrong problem.

Satinder Pal Singh and Richard C. Yee. An upper bound on the loss from approximate optimal-value functions. *Machine Learning*, To appear.

Shows that if we can bound the maximal error in value functions introduced by function approximation, then we can bound the worst-case performance of the greedy policy.

Richard D. Smallwood and Edward J. Sondik. The optimal control of partially observable markov processes over a finite horizon. *Operations Research*, 21:1071–1088, 1973.

Describes the first exact method for solving POMDP*'s. Mostly of mathematical and historical interest.*

Edward J. Sondik. *The Optimal Control of Partially Observable Markov Processes.* PhD thesis, Stanford University, Stanford, CA, 1971.

Introduces most of the fundamental results in the theory of POMDP*'s with regards to the representation of optimal policies.*

Edward J. Sondik. The optimal control of partially observable markov processes over the infinite horizon: Discounted costs. *Operations Research*, 26(2):282–304, 1978.

Gives a policy iteration type algorithm that gives provably near-optimal policies for infinite-horizon POMDP *problems.*

E. D. Sontag. Some topics in neural networks and control. Technical Report LS93-02, Rutgers University, New Brunswick, NJ, July 1993.

Richard S. Sutton. *Temporal Credit Assignment in Reinforcement Learning.* PhD thesis, University of Massachusetts, Amherst, MA, 1984.

Richard S. Sutton. Learning to predict by the method of temporal differences. *Machine Learning*, 3(1):9–44, 1988.

Richard S. Sutton. Implementation details of the TD(λ) procedure for the case of vector predictions and backpropagation. Technical Report TN87-509.1, GTE Laboratories Inc., August 1989.

Richard S. Sutton. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Proceedings of the Seventh International Conference on Machine Learning*, Austin, TX, 1990. Morgan Kaufmann.

Discusses the DYNA *model for using a learned model to plan effectively.*

Richard S. Sutton. Planning by incremental dynamic programming. In *Proceedings of the Ninth International Workshop on Machine Learning.* Morgan Kaufmann, 1991.

Richard S. Sutton. Reinforcement learning architectures for animats. In *Proceedings of the International Workshop on the Simulation of Adaptive Behavior: From Animals to Animats*, pages 288–296, Cambridge, MA, 1991. The MIT Press.

Richard S. Sutton. Introduction: The challenge of reinforcement learning. *Machine Learning*, 8:225–227, 1992.

My favourite quote – it is in a sense the whole AI problem in a microcosm

Richard S. Sutton and Andrew G. Barto. Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, 2:135–170, 1981.

Richard S. Sutton and Brian Pinette. The learning of world models by connectionist networks. In *Proceedings of the Seventh Annual Conference of the Cognitive Science Society*, pages 54–64, 1985.

Richard S. Sutton and Steven D. Whitehead. Online learning with random representations. In *Machine Learning: Proceedings of the Tenth International Conference*, pages 314–321, San Mateo, CA, 1993. Morgan Kaufmann.

Perceptrons with randomly constructed representations. (They do point out that the original perceptrons had random representations too.) Performance is compared to nearest-neighbour techniques and back-prop. The random-representation task works well, especially when there are irrelevant features.

M. Tan. Multi-agent reinforcement learning: independent vs. cooperative agents. In *Proceedings of the Tenth International Conference on Machine Learning*, Amherst, MA, 1993. Morgan Kaufmann.

Describes several methods for training multiple agents and compares them on a simulated search task.

Ming Tan. Cost-sensitive reinforcement learning for adaptive classification and control. In *Proceedings of the Ninth National Conference on Artificial Intelligence*, Anaheim, CA, 1991.

Gerald Tesauro. Practical issues in temporal difference learning. *Machine Learning*, 8:257–277, 1992.

Uses TD *learning to teach a net to play backgammon. One version of the net is apparently the strongest computer player in the world. The empirical success is theoretically suprising.*

Gerald Tesauro. TD-Gammon, a self-teaching backgammon program, achieves master-level play. *Neural Computation*, To appear.

Probably the best known and most successful application of reinforcement learning techniques.

M. A. L. Thathachar and P. S. Sastry. A new approach to the design of reinforcement schemes for learning automata. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-15(1):168–175, 1985.

Sebastian Thrun and K. Möller. Active exploration in dynamic environments. In J. E. Moody, S. J. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems 4: Proceedings of the 1991 Conference*, pages 531–538, San Mateo, CA, 1992. Morgan Kaufmann.

Sebastian Thrun and Anton Schwartz. Issues in using function approximation for reinforcement learning. In M. Mozer, P. Smolensky, D. Touretzky, J. Elman, and A. Weigend, editors, *Proceedings of the 1993 Connectionist Models Summer School*, Hillsdale, NJ, 1993. Lawrence Erlbaum.

Gives bounds on the allowable approximation error and gamma if Q-learning is to be expected to learn optimal policies.

Sebastian B. Thrun. Efficient exploration in reinforcement learning. Technical Report CMU-CS-92-102, Carnegie Mellon University, Pittsburgh, PA, January 1992.

Sebastian B. Thrun. The role of exploration in learning control. In David A. White and Donald A. Sofge, editors, *Handbook of Intelligent Control: Neural, Fuzzy, and Adaptive Approaches*. Van Nostrand Reinhold, New York, NY, 1992.

Sebastian B. Thrun. Exploration and model building in mobile robot domains. In *Proceedings of the International Conference on Neural Networks*, 1993.

M. L. Tsetlin. *Automaton Theory and Modeling of Biological Systems.* Academic Press, New York, NY, 1973.

Christopher J. C. H. Watkins. *Learning from Delayed Rewards.* PhD thesis, King's College, Cambridge, UK, 1989.

Introduces Q-learning.

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

Paul J. Werbos. Generalization of backpropagation with application to a recurrent gas market model. *Neural Networks*, 1:339–356, 1988.

Paul J. Werbos. A menu of designs for reinforcement learning over time. In W. Thomas Miller, III, Richard S. Sutton, and Paul J. Werbos, editors, *Neural networks for control*, chapter 3, pages 67–95. The MIT Press, Cambridge, MA, 1990.

Has a nice diagram of Dynamic Programming. Mainly about an adaptive critic algorithm. The model is explicitly derived as an approximation to DP.

Steven D. Whitehead. Scaling reinforcement learning systems. Technical Report 304, University of Rochester, Department of Computer Science, Rochester, NY, 1989.

Steven D. Whitehead. A complexity analysis of cooperative mechanisms in reinforcement learning. In *Proc. Ninth National Conference on Artificial Intelligence*, 1991.

Steven D. Whitehead. Complexity and cooperation in Q-learning. In *Proceedings of the Eighth International Workshop on Machine Learning*, Evanston, IL, 1991. Morgan Kaufmann.

Steven D. Whitehead and Dana H. Ballard. A role for anticipation in reactive systems that learn. In *Proceedings of the Sixth International Workshop on Machine Learning*, pages 354–357, Ithaca, NY, 1989. Morgan Kaufmann.

Steven D. Whitehead and Dana H. Ballard. Active perception and reinforcement learning. In *Proc. Seventh International Conference on Machine Learning*, 1990.

Describes the LION *algorithm for doing robust reinforcement learning in a restricted class of hidden state environments. Applies the algorithm to an active perception task.*

Steven D. Whitehead and Dana H. Ballard. Learning to perceive and act by trial and error. *Machine Learning*, 7(1):45–83, 1991.

Steven D. Whitehead, Richard S. Sutton, and Dana H. Ballard. Advances in reinforcement learning and their implications for intelligent control. In *Proc. Fifth IEEE Int. Symp. on Intelligent Control*, 1990.

Bernard Widrow, Narendra K. Gupta, and Sidhartha Maitra. Punish/reward: Learning with a critic in adaptive threshold systems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(5):455–465, 1973.

Bernard Widrow and Marcian E. Hoff. Adaptive switching circuits. In *IRE WESCON Convention Record*, New York, NY, 1960.

Reprinted in Neurocomputing: Foundations of Research, James A. Anderson and Edward Rosenfeld, editors, The MIT Press, Cambridge, MA, 1988

Bernard Widrow and S. D. Stearns. *Adaptive Signal Processing.* Prentice Hall, Englewood Cliffs, NJ, 1985.

Ronald J. Williams. Reinforcement learning in connectionist networks: A mathematical analysis. Technical Report ICS-8605, Institute for Cognitive Science, University of California, San Diego, La Jolla, CA, 1986.

Ronald J. Williams. A class of gradient-estimating algorithms for reinforcement learning in neural networks. In *Proceedings of the IEEE First International Conference on Neural Networks*, San Diego, CA, 1987.

Develops the REINFORCE *class of reinforcement learning algorithms for maximizing instantaneous reward.*

Ronald J. Williams. On the use of backpropagation in associative reinforcement learning. In *Proceedings of the IEEE International Conference on Neural Networks*, San Diego, CA, 1988.

Ronald J. Williams. Adaptive state representation and estimation using recurrent connectionist networks. In W. Thomas Miller, III, Richard S. Sutton, and Paul J. Werbos, editors, *Neural networks for control*, chapter 4, pages 97–114. The MIT Press, Cambridge, MA, 1990.

How the state of a plant can be modelled and how (recurrent) networks allow new (to many people) types of models of state.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, 1992.

Ronald J. Williams and Leemon C. Baird, III. Analysis of some incremental variants of policy iteration: First steps toward understanding actor-critic learning systems. Technical Report NU-CCS-93-11, Northeastern University, College of Computer Science, September 1993.

Mathematical results on general update sequences based on policy iteration.

Ronald J. Williams and Leemon C. Baird, III. Tight performance bounds on greedy policies based on imperfect value functions. Technical Report NU-CCS-93-14, Northeastern University College of Computer Science, 1993.

Tight lower bound on cumulative reward obtained from policies based on value functions where Bellman equation is not satisfied exactly.

Ian H. Witten. An adaptive optimal controller for discrete-time markov environments. *Information and Control*, 34:286–295, 1977.

Lambert E. Wixson. Scaling reinforcement learning techniques via modularity. In *Proc. Eighth International Machine Learning Workshop*, 1991.

H. Zhu. *Neural Networks and Adaptive Computers: Theory and Methods of Stochastic Adaptive Computation.* PhD thesis, University of Liverpool, Liverpool, UK, June 1993.

Foundation of stochastic adaptive computation based on neural networks. Simulated annealing learning rule superior to backpropagation and Boltzmann machine learning rules. Reinforcement learning for combinatorial state space and action space. (Mathematics with simulation results plus philosophy.)