References

Amari, S. 1998. Natural gradient works efficiently in learning. *Neural Computation* 10(2):251–276.

Bartlett, P. L., and Baxter, J. 2000. Estimation and approximation bounds for gradient-based reinforcement learning. In *Annual Conference on Computational Learning Theory*, 133–141.

Baxter, J., and Bartlett, P. L. 2000. Reinforcement learning in POMDP's via direct gradient ascent. In *International Conference on Machine Learning*, 41–48.

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

Bertsekas, D. P. 1995. *Dynamic Programming and Optimal Control, Volumes 1 and 2*. Athena Scientific.

Boyan, J. A. 2002. Technical update: Least-squares temporal difference learning. *Machine Learning* 49(2-3):233–246.

Bradtke, S. J., and Barto, A. G. 1996. Linear least-squares algorithms for temporal difference learning. *Machine Learning* 22(1-3):33–57.

Gullapalli, V. 1990. A stochastic reinforcement learning algorithm for learning real-valued functions. *Neural Networks* 3(6):671–692.

Kakade, S. 2002. A natural policy gradient. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

Kakade, S. 2003. *On the Sample Complexity of Reinforcement Learning*. Ph.D. thesis, University College London.

Kearns, M., and Singh, S. 2002. Near-optimal reinforcement learning in polynomial time. *Machine Learning* 49(2-3):209–232.

Levin, D.; Peres, Y.; and Wilmer, E. 2008. *Markov Chains and Mixing Times*. American Mathematical Society.

Lovász, L., and Winkler, P. 1998. Mixing times. In Aldous, D., and Propp, J., eds., *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 85–133.

Morimura, T.; Uchibe, E.; Yoshimoto, J.; and Doya, K. 2008. A new natural policy gradient by stationary distribution metric. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*.

Morimura, T.; Uchibe, E.; Yoshimoto, J.; and Doya, K. 2009. A generalized natural actor-critic algorithm. In *Advances in Neural Information Processing Systems*, volume 22.

Morimura, T.; Osogami, T.; and Shirai, T. 2014. Mixing-time regularized policy gradient. In *Technical Report*. IBM Research, RT0961.

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

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

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