

Reinforcement Learning in Continuous State and Action Spaces

DISSERTATION

Zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften (Dr. rer. nat.)
im Fach Informatik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät II
der Humboldt-Universität zu Berlin

von
Herr M.Sc. Víctor Uc-Cetina
geboren am 16.04.1977 in Merida, Mexiko

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Dr. h.c. Christoph Marksches

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät II:
Prof. Dr. Wolfgang Coy

Gutachter:

1.
2.
3.

eingereicht am:

Tag der mündlichen Prüfung:

Contents

1	Introduction	1
1.1	Research Summary	2
1.2	Outline	3
2	Background	7
2.1	Markov Decision Processes	7
2.1.1	Value Iteration and Policy Iteration	9
2.2	Reinforcement Learning	10
2.2.1	Q-Learning	12
2.2.2	Sarsa	12
2.2.3	Actor-Critic Methods	13
2.3	Supervised Learning	15
2.3.1	Multilayer Perceptrons	15
2.3.2	Backpropagation Algorithm	18
2.3.3	Radial Basis Function Networks	18
3	Robot Dribbling Problem	21
3.1	Introduction	21
3.2	RoboCup	21
3.3	Soccer Simulator	22
3.3.1	Server	22
3.3.2	Monitor	23
3.3.3	Client	24
3.4	Players, Ball and Environment	24
3.4.1	Perceptions and Actions	24
3.4.2	Environment	25
3.4.3	Interaction	26
3.5	Definition of the Dribbling Problem	26
3.5.1	Noisy Vision System	27
3.5.2	Noisy Dash Model	27
3.5.3	Noisy Kick Model	29

3.5.4	Parameters of Heterogeneous Players	30
3.6	Desired Characteristics of the Solution	32
3.7	Summary	33
4	Concurrent Reinforcement Learning	35
4.1	Concurrent Reinforcement Learning	35
4.1.1	Teachers	37
4.1.2	Agents	38
4.2	Testbed	38
4.3	Experiments and Results	39
4.4	Related work	40
4.5	Summary	41
5	Advice-Giving	43
5.1	Introduction	43
5.2	Advice-Giving	44
5.3	Optimal Policy Problem	46
5.4	Learning a Path Policy	47
5.4.1	Without Advice	47
5.4.2	Following Always the Imperfect Advice	49
5.4.3	Using Imperfect Advice to Create Reduced Sets of Promising Actions	50
5.5	Learning a Complete Policy	51
5.5.1	Without Advice	52
5.5.2	Following Always the Imperfect Advice	52
5.5.3	Using Imperfect Advice to Create Reduced Sets of Promising Actions	53
5.6	Advice-Giving Architecture	53
5.7	General Advice-Giving Algorithm	55
5.8	Summary	56
6	Advice-Giving with Function Approximation	59
6.1	Introduction	59
6.2	Function Approximation	59
6.2.1	Multilayer Perceptron with Radial Basis Functions	60
6.2.2	Mountain Car Task	62
6.3	Advice-Giving for Dribbling	64
6.3.1	Experimental Scenario	66
6.3.2	Sarsa and Q-learning Experimental Results	69
6.3.3	Actor-Critic Experimental Results	72
6.4	Related Work	76

6.5	Summary	82
7	Sarsa Actor-Actor-Critic	83
7.1	Introduction	83
7.2	Actor-Actor-Critic Architecture	84
7.3	Sarsa Actor-Actor-Critic Algorithm	85
7.4	Experimental Results	86
7.5	Final Policy Performance	92
7.6	Discussion	95
7.7	Related Work	96
7.8	Summary	98
8	Conclusions	99
8.1	Summary	99
8.2	Future Work	101
8.3	Concluding Remarks	101