

Contents

Abstract	i
Acknowledgements	ii
1 Introduction	1
1.1 The focus of this dissertation	2
1.2 Reinforcement Learning	4
1.2.1 A qualitative introduction	4
1.2.2 The distinguishing characteristic	4
1.2.3 Examples of reinforcement learning tasks	5
1.2.4 A brief history of reinforcement learning	7
1.2.5 An illustration of reinforcement learning	9
1.3 The representation problem	15
1.3.1 Features, detectors, and generalized states	16
1.3.2 Feature extraction	21
1.3.3 Cognitive economy and the big picture	24
1.4 Contributions of This Dissertation	29
1.5 A Brief Preview of Subsequent Chapters	30
2 Reinforcement Learning and Feature Extraction	33
2.1 The Elements of Reinforcement Learning	34
2.1.1 The agent and the environment	34
2.1.2 Common assumptions	37

2.2	Q-Learning	40
2.2.1	Policies and value functions	41
2.2.2	An illustration	44
2.2.3	Theoretical convergence	51
2.2.4	Practical considerations	52
2.3	Survey of Feature Extraction	57
2.3.1	Gradient-descent function approximation	58
2.3.2	Targeting “important” regions in state-space	62
2.3.3	Distinguishing states with different action values	70
2.3.4	Good representations	74
3	Formalization of Cognitive Economy for Reinforcement Learning	76
3.1	Introduction	76
3.1.1	Overview	76
3.1.2	State generalization	77
3.1.3	A principled approach	78
3.2	Cognitive Economy	80
3.2.1	Related ideas in reinforcement learning	82
3.2.2	Three aspects of cognitive economy	83
3.3	Preliminaries	84
3.3.1	Representational model	85
3.3.2	Assumptions	90
3.3.3	Definitions	93
3.4	Relevant Features	100
3.4.1	Importance	100

3.4.2	Definitions of importance	102
3.5	Necessary Distinctions	103
3.5.1	Policy distinctions	104
3.5.2	Value distinctions	105
3.5.3	Making sound decisions	106
3.5.4	Criterion for representational adequacy	110
3.5.5	State Compatibility	116
3.6	Feature Extraction	119
3.7	Summary	120
4	Action Values for Generalized States	121
4.1	Introduction	121
4.1.1	Action values depend on the agent's representation	122
4.2	Example: State-Action Values for a Discrete Representation	123
4.3	Region-Action Values	125
4.3.1	Example: Region-action values for a partition representation	129
4.4	Generalized Action Values	135
4.4.1	Method 1: Exploit assumptions of soft state aggregation	136
4.4.2	Method 2: Exploit assumptions of error minimization	137
4.4.3	Example: Generalized action values for a coarse coded representation	143
4.5	The Convex Generalization Property	147
4.6	Effects of Representation on Action Values	154
4.7	Summary	157

5	State Compatibility and Representational Adequacy	158
5.1	Introduction	158
5.1.1	Summary of notation	159
5.1.2	Examples of what can go wrong	162
5.2	Sufficient conditions for ϵ -adequacy	166
5.2.1	Generalization of state separation	166
5.2.2	Common assumptions	168
5.2.3	Proof of the theorem	175
5.3	Discussion	182
5.3.1	Necessary and sufficient conditions	182
5.3.2	The case for partition representations	185
5.4	Appendix: Some Helpful Properties of Convex Combinations	188
6	Case Studies in On-Line Feature Extraction	191
6.1	Introduction	191
6.1.1	Making Relevant Distinctions	192
6.1.2	Methodology	194
6.2	An Algorithm for On-Line Feature Extraction	197
6.2.1	Top Level of the Algorithm	198
6.2.2	Recognizing Surprising States	200
6.2.3	Investigating Surprising States	203
6.2.4	Adding and Merging State-Space Regions	206
6.2.5	Further Considerations	210
6.3	Initial Tests	213
6.4	Case Study: Puck-On-A-Hill Task	214

	vii
6.4.1 Analysis	216
6.4.2 Results	219
6.5 Case Study: Pole Balancing Task	226
6.5.1 Analysis	230
6.5.2 Results	235
6.6 Discussion	236
6.7 Future Work	237
7 Conclusion	238
7.1 Contributions	239
7.2 Future work	241
Afterword	244
Bibliography	247

List of Tables

1	Maximum returns	12
2	Tabular representation of the action values $Q(s, a)$	14
3	Action values for up —discrete state representation	124
4	Action values for right —discrete state representation	124
5	Whole path values for up —partition representation	131
6	Whole path values for right —partition representation	131
7	One-step values for up —partition representation	134
8	One-step values for right —partition representation	134
9	Probability of state occurrence under a random policy	144
10	Generalized action values under soft-state aggregation	145
11	Whole-path action values, $v_1^*(s, \text{right})$	147
12	Whole-path action values, $v_1^*(s, \text{up})$	147
13	Generalized action values under minimization of “regional errors”	147
14	Generalized action values under minimization of “global errors”	152
15	Inconsistency of action rankings may prevent ϵ -adequacy ($\delta = 0.08, \epsilon =$ 0.2). Here $\text{pref}_\delta(s) = \{a_1, a_2, a_3, a_4\}$	163
16	Inadequate representation for $\delta > \epsilon/2$ ($\delta = 0.15, \epsilon = 0.2$.)	164
17	State value incompatibilities. ($\delta = 0.1, \epsilon = 0.2$)	164
18	Inadequate representation where Equation 43 is not met ($\delta = 0.1, \epsilon =$ 0.2). Here $\text{pref}_\delta(s) = \{a_1, a_2, a_3, a_4\} \not\subseteq \{a_1, a_2, a_3\} = \text{pref}_\epsilon(s_1)$	165

List of Figures

1	A 5×4 gridworld with start state S (2, 3) and goal state G (4, 2) . . .	9
2	Gridworld partitioned according to preferred actions	19
3	An agent and its environment	36
4	A simple three-node reinforcement learning task	45
5	A larger gridworld, with state generalization	55
6	A two-action gridworld task	63
7	$Q(s, \text{right}) - Q(s, \text{up})$ for the two-action gridworld	66
8	Plot of the state-probability distribution for the two-action gridworld under random exploration	68
9	This state region appears to have different values for the action, depend- ing on whether we enter it from s_1 or s_2 . Thus the Markov property does not hold for the partition region.	72
10	Examples of successful feature extraction for the two-action gridworld task	74
11	The agent's representation simplifies the environment by grouping states of the world into generalized states that share the same action values .	79
12	A possible recognition function for the feature tall	94
13	Several different representations of the gridworld	97
14	Widely-separated action values are characteristic of important features.	101
15	A necessary policy distinction: The state grouping must be split to avoid mistakes in policy.	104
16	An <i>unnecessary</i> distinction: Splitting the generalized state is probably not worth-while.	105

17	A necessary value distinction: The resulting states must be kept separate for the agent to seek the better outcome.	106
18	The representation must make appropriate policy distinctions, which affect its next move, and value distinctions, which allow wise choices from earlier states.	107
19	Incremental regret: Compare R , the expected long-term reward from s when we act according to the policy of the generalized state, with R_s , the return that results from taking the action which is best at s itself. .	108
20	State compatibility: Allow s_1 and s_2 to be grouped together if a one-step look-ahead reveals their overall values to be close and their preference sets similar.	117
21	Three representations of the 4×4 gridworld	123
22	An optimal path through the partitioned gridworld	129
23	$Q(s, \text{right})$ under soft-state aggregation	145
24	$Q(s, \text{right}) - Q(s, \text{up})$ under soft-state aggregation	146
25	$Q(s, \text{right})$ under minimization of “regional errors”	148
26	$Q(s, \text{right}) - Q(s, \text{up})$ under minimization of “regional errors”	149
27	Superposition of the detectors for the counter-example	152
28	$Q(s, \text{right}) - Q(s, \text{up})$ under minimization of “global errors”	153
29	$Q(s, \text{right})$ for the discrete representation	155
30	$Q(s, \text{right}) - Q(s, \text{up})$ for the discrete representation	156
31	A value distinction which may or may not be necessary, depending on the values of r_{zx} and r_{zy}	183
32	Top level of the algorithm.	199

33	Selecting “surprising” states for further investigation.	202
34	Strategy for Active Investigations of Surprising States.	205
35	Feature Extraction Algorithm.	208
36	Judging the Compatibility of Two States.	209
37	The puck-on-a-hill task: balance the puck on the hill to avoid negative reinforcement from hitting the wall.	215
38	Controllable states: states outside this band result in failure.	217
39	An ideal representation: must-push-left states (top curve) and must-push-right states (bottom curve) are separated by the diagonal line. . .	218
40	Representation constructed automatically, from scratch (24 categories).	220
41	Representation constructed from a good seed representation.	221
42	A representation inspired by Variable Resolution Dynamic Programming.	223
43	Enhanced VRDP representation.	224
44	Representation designed to limit the loss of controllability (from Yendo Hu, 1996).	225
45	Averaged performance curves for the original VRDP representation and Yendo Hu’s controllability quantization.	227
46	Averaged performance curves for the four best representations.	228
47	The cart-pole apparatus. The task is to balance the pole by pushing the cart to either the left or the right in each control interval.	229
48	Angular acceleration for the pole, $f = 10.0$	231
49	Acceleration of the cart, for $f = -10.0$ and $\ddot{\theta} = 17.38$	232
50	Acceleration of the cart, for $f = 10.0$ and $\ddot{\theta} = -17.38$	233