

目次

序文	<i>i</i>
訳者序文	<i>v</i>
I 強化学習問題	1
1 序章	2
1.1 強化学習	2
1.2 例	5
1.3 強化学習の構成要素	7
1.4 拡張された例：三日並べ	10
1.5 まとめ	16
1.6 強化学習の歴史	16
1.7 補足 (文献と歴史)	24
2 評価フィードバック	26
2.1 n 本腕バンディット問題	27
2.2 行動価値手法	29
2.3 ソフトマックス行動選択	32
2.4 評価と教示*	33
2.5 漸進的手法による実装	38
2.6 非定常問題への追従	40
2.7 オプティミスティック初期値	42
2.8 強化比較*	44
2.9 追跡手法*	46
2.10 連想探索*	48
2.11 結論	50
2.12 補足 (文献と歴史)	52

3 強化学習問題	55
3.1 エージェントと環境間のインタフェース	55
3.2 目標と報酬	60
3.3 収益	62
3.4 エピソード的タスクと連続タスクの統一的記述	65
3.5 マルコフ性*	66
3.6 マルコフ決定過程	71
3.7 価値関数	74
3.8 最適価値関数	81
3.9 最適性と近似	86
3.10 要約	88
3.11 補足 (文献と歴史)	89
II 基本的な解法群	93
4 動的計画法	94
4.1 方策評価	95
4.2 方策改善	98
4.3 方策反復	102
4.4 価値反復	106
4.5 非同期動的計画法	109
4.6 一般化方策反復	111
4.7 動的計画法の効率	113
4.8 まとめ	114
4.9 補足 (文献と歴史)	116
5 モンテカルロ法	119
5.1 モンテカルロ法による方策評価	120
5.2 モンテカルロ法による行動価値推定	125
5.3 モンテカルロ法による制御	127
5.4 方策オン型モンテカルロ法による制御	131
5.5 他の方策に追従する方策評価	133
5.6 方策オフ型モンテカルロ法による制御	135
5.7 漸進的実装	138
5.8 まとめ	139

5.9 補足 (文献と歴史)	141
6 TD 学習	142
6.1 TD 予測	142
6.2 TD 予測法の利点	147
6.3 TD(0) の最適性	151
6.4 Sarsa: 方策オン型 TD 制御	155
6.5 Q 学習: 方策オフ型 TD 制御	159
6.6 アクター・クリティック手法*	161
6.7 R 学習: 割引のない連続タスクのための学習法*	164
6.8 ゲーム, 事後状態, および他の特殊なケース	167
6.9 まとめ	169
6.10 補足 (文献と歴史)	170
<hr/>	
III 統一された見方	173
<hr/>	
7 適格度トレース	174
7.1 n ステップ TD 予測	175
7.2 TD(λ) の前方観測的な見方	180
7.3 TD(λ) の後方観測的な見方	184
7.4 前方観測的な見方と後方観測的な見方の等価性	188
7.5 Sarsa(λ)	191
7.6 Q(λ)	193
7.7 アクター・クリティック手法における適格度トレース*	199
7.8 人替え更新トレース	200
7.9 実装上の問題	203
7.10 λ 可変更新*	204
7.11 結論	205
7.12 補足 (文献と歴史)	206
8 一般化と関数近似	209
8.1 関数近似による価値予測	209
8.2 最急降下法	213
8.3 線形手法	217
8.4 関数近似を用いた制御	227

8.5 方策オフ型ブートストラップ	233
8.6 ブートストラップを行うべきか	238
8.7 まとめ	239
8.8 補足 (文献と歴史)	241
9 プランニングと学習	247
9.1 モデルとプランニング	247
9.2 プランニング, 行動, 学習の統合	250
9.3 モデルに誤りがある場合	256
9.4 優先度スイープ	260
9.5 完全バックアップとサンプルバックアップ	264
9.6 遷移軌跡サンプリング	269
9.7 ヒューリスティック探索	273
9.8 まとめ	276
9.9 補足 (文献と歴史)	277
10 強化学習の特徴軸	279
10.1 統一された見方	279
10.2 その他の先端的特徴軸	282
11 ケーススタディ	286
11.1 TD-Gammon	286
11.2 Samuel のチェッカー・プレイヤー	292
11.3 Acrobot	297
11.4 エレベータ・ディスパッチ問題	300
11.5 動的チャネル割り当て	306
11.6 ジョブショップ・スケジューリング	311
参考文献	319
記号の説明	337
索引	338