目　次

第 1 章　マルコフ決定過程　1

1.1　本書の表記と前提とする知識　・・・・・・・・・・・・・・・・・・・・・・・・　1
1.2　マルコフ決定過程　・・・・・・・・・・・・・・・・・・・・・・・・・・・・・　2
1.3　価値関数　・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・　7
1.4　MDP を解くための動的計画法　・・・・・・・・・・・・・・・・・・・・・・・　11

第 2 章　価値推定問題　13

2.1　有限な状態空間での TD 学習　・・・・・・・・・・・・・・・・・・・・・・・　13
2.1.1　テーブル TD(0) 法　・・・・・・・・・・・・・・・・・・・・・・・・・・　14
2.1.2　逐一訪問モンテカルロ法　・・・・・・・・・・・・・・・・・・・・・・・　17
2.1.3　TD(λ) 法: モンテカルロ法と TD(0) 法の統一　・・・・・・・・・・・　20
2.2　大規模状態空間でのアルゴリズム　・・・・・・・・・・・・・・・・・・・・　22
2.2.1　関数近似を用いた TD(λ) 法　・・・・・・・・・・・・・・・・・・・・　26
2.2.2　勾配 TD 学習 (gradient temporal difference learning)　・・・・・　30
2.2.3　最小二乗法　・・・・・・・・・・・・・・・・・・・・・・・・・・・・・　32
2.2.4　関数空間の選択　・・・・・・・・・・・・・・・・・・・・・・・・・・・　39

第 3 章　制御　45

3.1　学習問題一覧　・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・　45
3.2　閉ループでの対話型学習　・・・・・・・・・・・・・・・・・・・・・・・・・　47
3.2.1　バンディット問題における探索活用並行学習　・・・・・・・・・・・・　47
3.2.2　バンディット問題における純粋探索学習　・・・・・・・・・・・・・・　49
3.2.3　マルコフ決定過程における純粋探索学習　・・・・・・・・・・・・・・　50
3.2.4　マルコフ決定過程における探索活用並行学習　・・・・・・・・・・・・　52
3.3　直接法　・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・　57
3.3.1　有限 MDP における Q 学習　・・・・・・・・・・・・・・・・・・・・・　57
3.3.2　関数近似器を用いた Q 学習　・・・・・・・・・・・・・・・・・・・・・　60

3.4 Actor-critic 法 ・・・・・・・・・・・・・・・・・・・・・・・・・・・・ 64
3.4.1 Critic の実装 ・・・・・・・・・・・・・・・・・・・・・・・・・ 65
3.4.2 Actor の実装 ・・・・・・・・・・・・・・・・・・・・・・・・・ 67

第 4 章 さらなる勉強のために 75
4.1 参考文献 ・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・ 75
4.2 応用 ・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・ 76
4.3 ソフトウェア ・・・・・・・・・・・・・・・・・・・・・・・・・・・・ 76
4.4 謝辞 ・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・ 77

付録 A 割引マルコフ決定過程の理論 79
A.1 縮小写像とバナッハの不動点定理 ・・・・・・・・・・・・・・・・・・ 79
A.2 MDP への適用 ・・・・・・・・・・・・・・・・・・・・・・・・・・・ 83

付録 B TD(λ) 法の前方観測的な見方と後方観測的な見方について 89

付録 C 深層強化学習を含む最近の発展 93
C.1 深層強化学習のための深層学習 ・・・・・・・・・・・・・・・・・・・ 94
C.1.1 ニューラルネットワークを用いた関数近似 ・・・・・・・・・・・・ 94
C.1.2 CNN (convolutional neural network) ・・・・・・・・・・・・・・ 95
C.2 価値反復に基づく強化学習アルゴリズムにおける発展 ・・・・・・・・・・ 97
C.2.1 DQN (deep Q-network) ・・・・・・・・・・・・・・・・・・・・ 97
C.2.2 Double DQN ・・・・・・・・・・・・・・・・・・・・・・・・・・ 99
C.2.3 デュエリングネットワーク (dueling network) ・・・・・・・・・・ 100
C.2.4 優先順位付き経験再生 (prioritized experience replay) ・・・・・ 101
C.3 方策反復に基づく強化学習アルゴリズムにおける発展 ・・・・・・・・・・ 102
C.3.1 A3C (asynchronous advantage actor-critic) ・・・・・・・・・・ 102
C.3.2 TRPO (trust region policy optimization) ・・・・・・・・・・・・ 104
C.3.3 GAE (generalized advantage estimator) ・・・・・・・・・・・・ 106
C.4 深層強化学習の囲碁 AI への応用: AlphaGo ・・・・・・・・・・・・・ 106
C.4.1 強化学習問題としての囲碁 ・・・・・・・・・・・・・・・・・・・ 107
C.4.2 深層ニューラルネットワークの学習 ・・・・・・・・・・・・・・・ 108
C.4.3 深層ニューラルネットワークを使ったモンテカルロ木探索法による着手の選択 ・・・・・・・・・・・・・・・・・・・・・・・・・・ 109
C.5 おわりに ・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・ 110

参考文献 **111**

索　引 **129**