

Table of Contents

Preface	xxi
Part I: Introduction to RL	1
Chapter 1: What Is Reinforcement Learning?	3
Supervised learning	4
Unsupervised learning	4
Reinforcement learning	5
Complications in RL	6
RL formalisms	7
Reward • 8	
The agent • 9	
The environment • 10	
Actions • 10	
Observations • 11	
The theoretical foundations of RL	14
Markov decision processes • 14	
<i>The Markov process</i> • 14	
<i>Markov reward processes</i> • 19	
<i>Adding actions to MDP</i> • 22	
Policy • 25	
Summary	26
Chapter 2: OpenAI Gym API and Gymnasium	27
The anatomy of the agent	28

Hardware and software requirements	31
The OpenAI Gym API and Gymnasium	33
The action space • 34	
The observation space • 34	
The environment • 36	
Creating an environment • 38	
The CartPole session • 40	
The random CartPole agent	43
Extra Gym API functionality	44
Wrappers • 44	
Rendering the environment • 47	
More wrappers • 49	
Summary	49
 Chapter 3: Deep Learning with PyTorch	 51
Tensors	52
The creation of tensors	52
Scalar tensors • 55	
Tensor operations • 56	
GPU tensors • 56	
Gradients	58
Tensors and gradients • 60	
NN building blocks	63
Custom layers	64
Loss functions and optimizers	67
Loss functions • 68	
Optimizers • 68	
Monitoring with TensorBoard	70
TensorBoard 101 • 71	
Plotting metrics • 72	
GAN on Atari images	74
PyTorch Ignite	80
Ignite concepts • 81	
GAN training on Atari using Ignite • 82	
Summary	85

Chapter 4: The Cross-Entropy Method 87

The taxonomy of RL methods	88
The cross-entropy method in practice	89
The cross-entropy method on CartPole	91
The cross-entropy method on FrozenLake	101
The theoretical background of the cross-entropy method	109
Summary	110

Part II: Value-based methods 111

Chapter 5: Tabular Learning and the Bellman Equation 113

Value, state, and optimality	114
The Bellman equation of optimality	116
The value of the action	118
The value iteration method	121
Value iteration in practice	123
Q-iteration for FrozenLake	130
Summary	132

Chapter 6: Deep Q-Networks 133

Real-life value iteration	134
Tabular Q-learning	135
Deep Q-learning	140
Interaction with the environment • 142	
SGD optimization • 143	
Correlation between steps • 143	
The Markov property • 144	
The final form of DQN training • 144	
DQN on Pong	145
Wrappers • 146	
The DQN model • 151	
Training • 153	
Running and performance • 164	
Your model in action • 167	
Things to try	169

Summary	170
Chapter 7: Higher-Level RL Libraries	171
Why RL libraries?	172
The PTAN library	172
Action selectors • 174	
The agent • 176	
<i>DQN</i> Agent • 177	
<i>Policy</i> Agent • 179	
Experience source • 180	
<i>Toy environment</i> • 181	
<i>The ExperienceSource class</i> • 182	
<i>The ExperienceSourceFirstLast Class</i> • 184	
Experience replay buffers • 186	
The TargetNet class • 188	
Ignite helpers • 190	
The PTAN CartPole solver	190
Other RL libraries	193
Summary	194
Chapter 8: DQN Extensions	195
Basic DQN	196
Common library • 196	
Implementation • 202	
Hyperparameter tuning • 203	
Results with common parameters • 207	
Tuned baseline DQN • 209	
N-step DQN	210
Implementation • 213	
Results • 213	
Hyperparameter tuning • 215	
Double DQN	215
Implementation • 216	
Results • 218	
Hyperparameter tuning • 219	

Noisy networks	219
Implementation • 220	
Results • 223	
Hyperparameter tuning • 225	
Prioritized replay buffer	225
Implementation • 226	
Results • 231	
Hyperparameter tuning • 232	
Dueling DQN	233
Implementation • 235	
Results • 236	
Hyperparameter tuning • 237	
Categorical DQN	237
Implementation • 240	
Results • 246	
Hyperparameter tuning • 248	
Combining everything	248
Results • 249	
Hyperparameter tuning • 251	
Summary	252
Chapter 9: Ways to Speed Up RL	253
Why speed matters	253
Baseline	256
The computation graph in PyTorch	258
Several environments	261
Playing and training in separate processes	264
Tweaking wrappers	268
Benchmark results	271
Summary	272
Chapter 10: Stocks Trading Using RL	273
Why trading?	273
Problem statement and key decisions	274
Data	276

The trading environment	277
Models	285
Training code	287
Results	288
The feed-forward model • 288	
The convolution model • 293	
Things to try	294
Summary	295
Part III: Policy-based methods	297
Chapter 11: Policy Gradients	299
Values and policy	299
Why the policy? • 300	
Policy representation • 301	
Policy gradients • 302	
The REINFORCE method	302
The CartPole example • 304	
Results • 308	
Policy-based versus value-based methods • 310	
REINFORCE issues	310
Full episodes are required • 310	
High gradient variance • 311	
Exploration problems • 312	
High correlation of samples • 312	
Policy gradient methods on CartPole	313
Implementation • 313	
Results • 316	
Policy gradient methods on Pong	319
Implementation • 320	
Results • 321	
Summary	321
Chapter 12: Actor-Critic Method: A2C and A3C	323
Variance reduction	324

CartPole variance	325
Advantage actor-critic (A2C)	328
A2C on Pong • 331	
Results • 338	
Asynchronous Advantage Actor-Critic (A3C)	342
Correlation and sample efficiency • 342	
Adding an extra “A” to A2C • 343	
A3C with data parallelism • 346	
Results • 346	
A3C with gradient parallelism • 347	
Implementation • 347	
Results • 353	
Summary	353
Chapter 13: The TextWorld Environment	355
Interactive fiction	355
The environment	359
Installation • 360	
Game generation • 360	
Observation and action spaces • 361	
Extra game information • 362	
The deep NLP basics	364
Recurrent Neural Networks (RNNs) • 365	
Word embedding • 367	
The Encoder-Decoder architecture • 368	
Transformers • 369	
Baseline DQN	369
Observation preprocessing • 371	
Embeddings and encoders • 376	
The DQN model and the agent • 378	
Training code • 379	
Training results • 380	
Tweaking observations	382
Tracking visited rooms • 382	
Relative actions • 384	

Objective in observation • 386	
Transformers	387
ChatGPT	389
Setup • 389	
Interactive mode • 389	
ChatGPT API • 392	
Summary	395
Chapter 14: Web Navigation	397
The evolution of web navigation	397
Browser automation and RL	398
Challenges in browser automation	399
The MiniWoB benchmark	400
MiniWoB++	401
Installation • 401	
Actions and observations • 402	
Simple example • 402	
The simple clicking approach	406
Grid actions • 406	
The RL part of our implementation • 409	
The model and training code • 410	
Training results • 410	
Simple clicking limitations • 412	
Adding text description	414
Implementation • 414	
Results • 419	
Human demonstrations	420
Recording the demonstrations • 421	
Training with demonstrations • 423	
Results • 425	
Things to try	427
Summary	427

Part IV: Advanced RL 429

Chapter 15: Continuous Action Space 431

Why a continuous space?	432
The action space • 432	
Environments • 433	
The A2C method	435
Implementation • 436	
Results • 440	
Using models and recording videos • 441	
Deep deterministic policy gradients	442
Exploration • 444	
Implementation • 445	
Results and video • 450	
Distributional policy gradients	453
Architecture • 453	
Implementation • 454	
Results • 457	
Things to try	459
Summary	459

Chapter 16: Trust Region Methods 461

Environments	462
The A2C baseline	463
Implementation • 463	
Results • 465	
Video recording • 468	
PPO	469
Implementation • 470	
Results • 474	
TRPO	476
Implementation • 477	
Results • 479	
ACKTR	481
Implementation • 481	

Results • 482	
SAC	483
Implementation • 484	
Results • 486	
Overall results	489
Summary	489
Chapter 17: Black-Box Optimizations in RL	491
Black-box methods	491
Evolution strategies	492
Implementing ES on CartPole • 493	
CartPole results • 499	
ES on HalfCheetah • 500	
Implementing ES on HalfCheetah • 501	
HalfCheetah results • 505	
Genetic algorithms	507
GA on CartPole • 508	
GA tweaks • 510	
<i>Deep GA</i> • 510	
<i>Novelty search</i> • 510	
GA on HalfCheetah • 511	
<i>Implementation</i> • 511	
<i>Results</i> • 514	
Summary	515
Chapter 18: Advanced Exploration	517
Why exploration is important	518
What's wrong with ϵ -greedy?	518
Alternative ways of exploration	522
Noisy networks • 522	
Count-based methods • 523	
Prediction-based methods • 524	
MountainCar experiments	524
DQN + ϵ -greedy • 525	
DQN + noisy networks • 527	

DQN + state counts • 528	
PPO method • 530	
PPO + Noisy Networks • 532	
PPO + state counts • 533	
PPO + network distillation • 534	
Comparison of methods • 537	
Atari experiments	537
DQN + ϵ -greedy • 538	
DQN + noisy networks • 539	
PPO • 539	
Summary	539
Chapter 19: Reinforcement Learning with Human Feedback	541
Reward functions in complex environments	542
Theoretical background	544
Method overview • 544	
RLHF and LLMs • 546	
RLHF experiments	547
Initial training using A2C • 548	
Labeling process • 552	
Reward model training • 556	
Combining A2C with the reward model • 560	
Fine-tuning with 100 labels • 563	
The second round of the experiment • 564	
The third round of the experiment • 565	
Overall results • 567	
Summary	567
Chapter 20: AlphaGo Zero and MuZero	569
Comparing model-based and model-free methods	570
Model-based methods for board games	571
The AlphaGo Zero method	572
Overview • 572	
MCTS • 573	
Self-play • 575	

Training and evaluation • 576	
Connect 4 with AlphaGo Zero	576
The game model • 578	
Implementing MCTS • 580	
The model • 585	
Training • 588	
Testing and comparison • 588	
Results • 588	
MuZero	591
High-level model • 592	
Training process • 593	
Connect 4 with MuZero	594
Hyperparameters and MCTS tree nodes • 594	
Models • 598	
MCTS search • 601	
Training data and gameplay • 604	
MuZero results	609
MuZero and Atari	610
Summary	611
Chapter 21: RL in Discrete Optimization	613
The Rubik's cube and discrete optimization	614
Optimality and God's number	615
Approaches to cube solving	616
Actions • 617	
States • 618	
The training process	621
The NN architecture • 622	
The training • 623	
The model application	624
Results	626
The code outline	628
Cube environments • 629	
Training • 633	
The search process • 634	

Preface

The experiment results	635
The 2 × 2 cube • 637	
The 3 × 3 cube • 639	
Further improvements and experiments	641
Summary	641
Chapter 22: Multi-Agent RL	643
What is multi-agent RL?	644
Getting started with the environment	645
An overview of MAgent • 645	
Installing MAgent • 646	
Setting up a random environment • 646	
Deep Q-network for tigers	652
Understanding the code • 652	
Training and results • 656	
Collaboration by the tigers	659
Training both tigers and deer	661
The battle environment	662
Summary	663
Other Books You May Enjoy	671
Index	675