DATA SCIENCE. INSTITUTE by Fabian Rappert

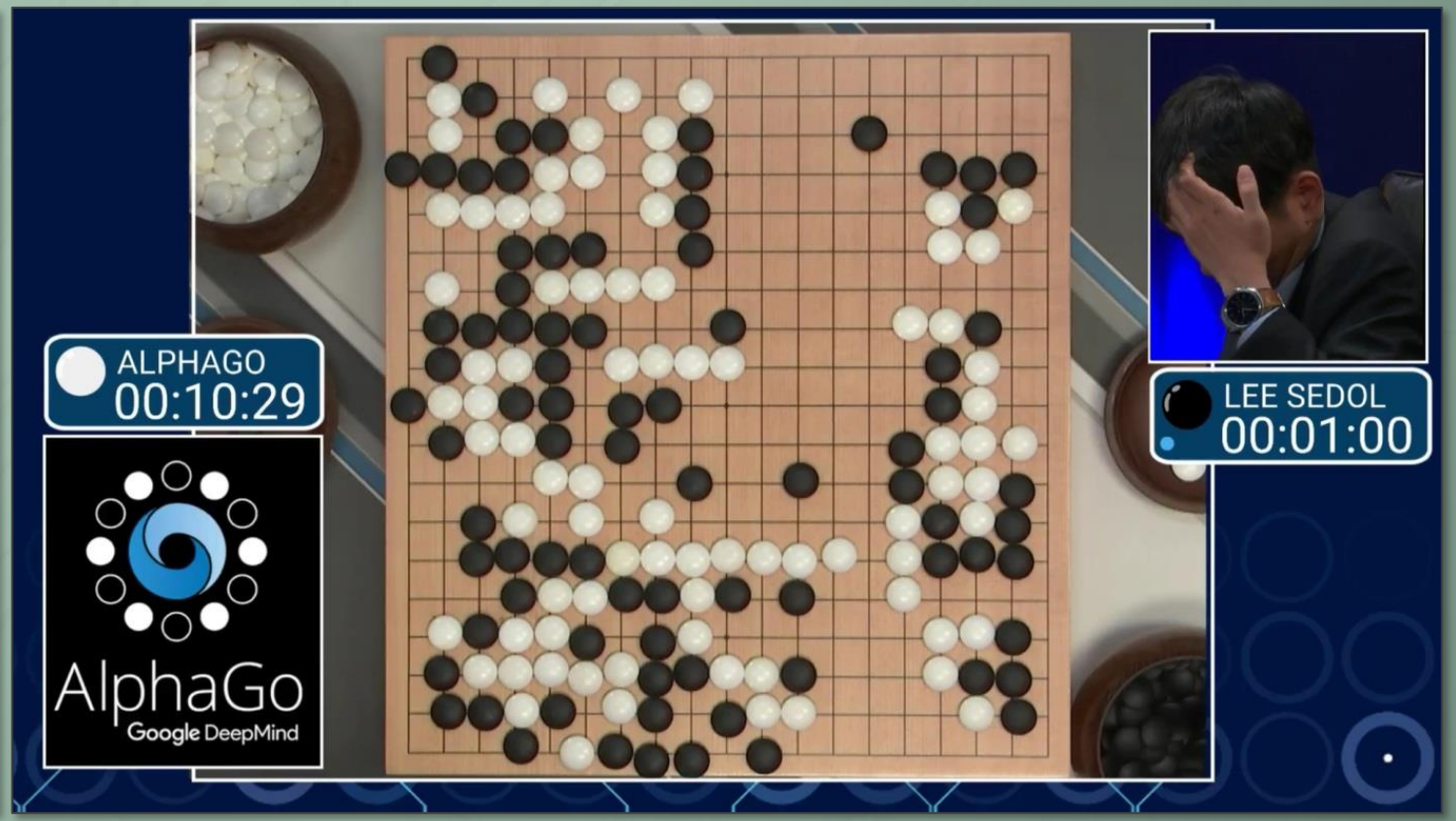
Deep Reinforcement Learning

Advanced Methods

Dr. Harald Stein
Aug 2024

Agenda

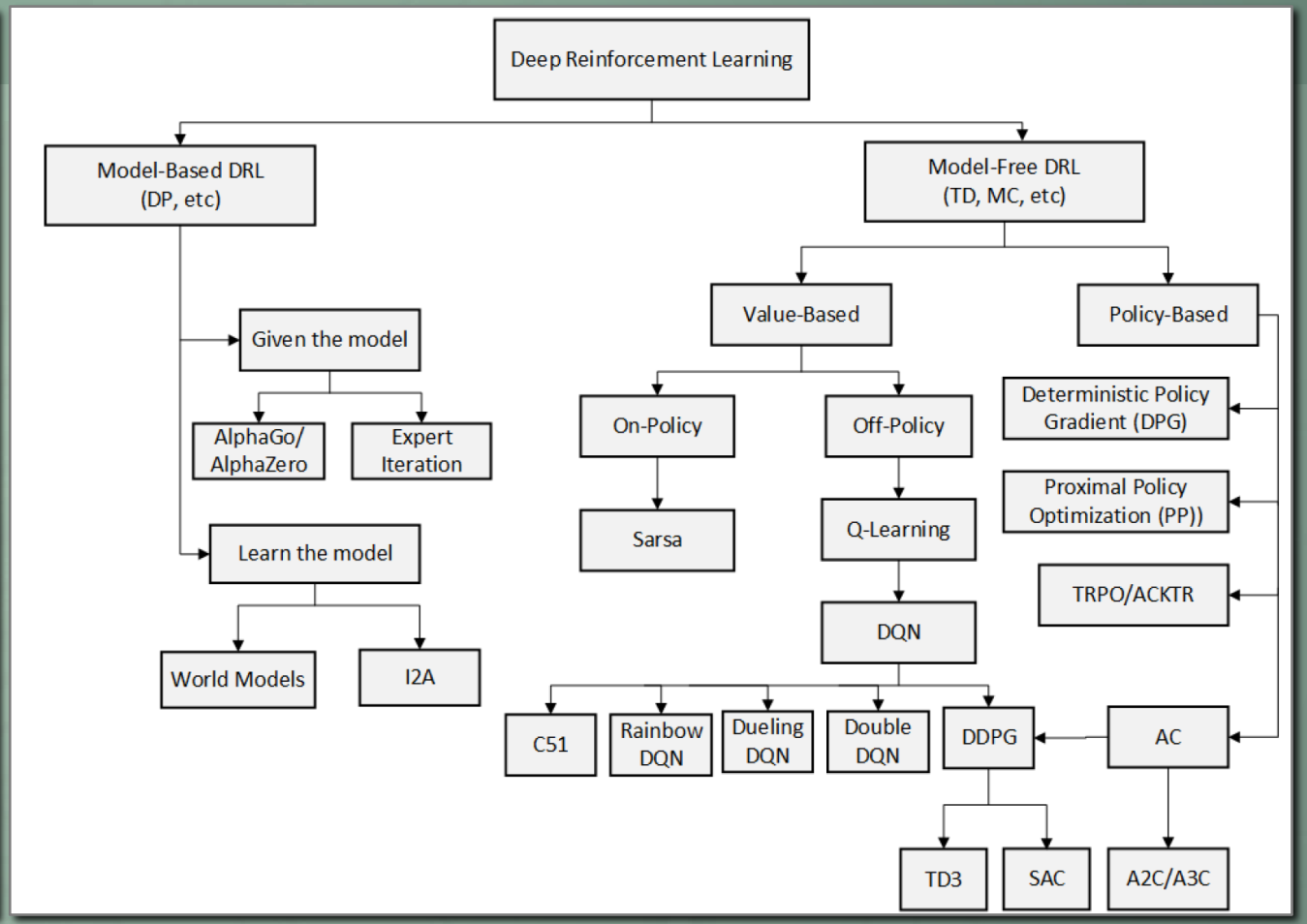
- Taxonomy
- Model-Based DRL
- Model-Free DRL
- Value-Based Methods
- Policy-Based Methods
- Actor-Critic Methods



Taxonomy of Reinforcement Learning Algorithms

Hierarchical Overview of Modern Deep RL Techniques and Algorithms

- DP Dynamic Programming
- TD Temporal Difference
- MC Monte Carlo
- I2A Imagination-Augmented Agent
- DQN Deep Q-Network
- TRPO Trust Region Policy Optimization
- ACKTR Actor Critic using Kronecker-Factored Trust Region
- AC Actor-Critic
- A2C Advantage Actor Critic
- A3C Asynchronous Advantage Actor Critic
- DDPG Deep Deterministic Policy Gradient
- TD3 Twin Delayed DDPG
- SAC Soft Actor-Critic



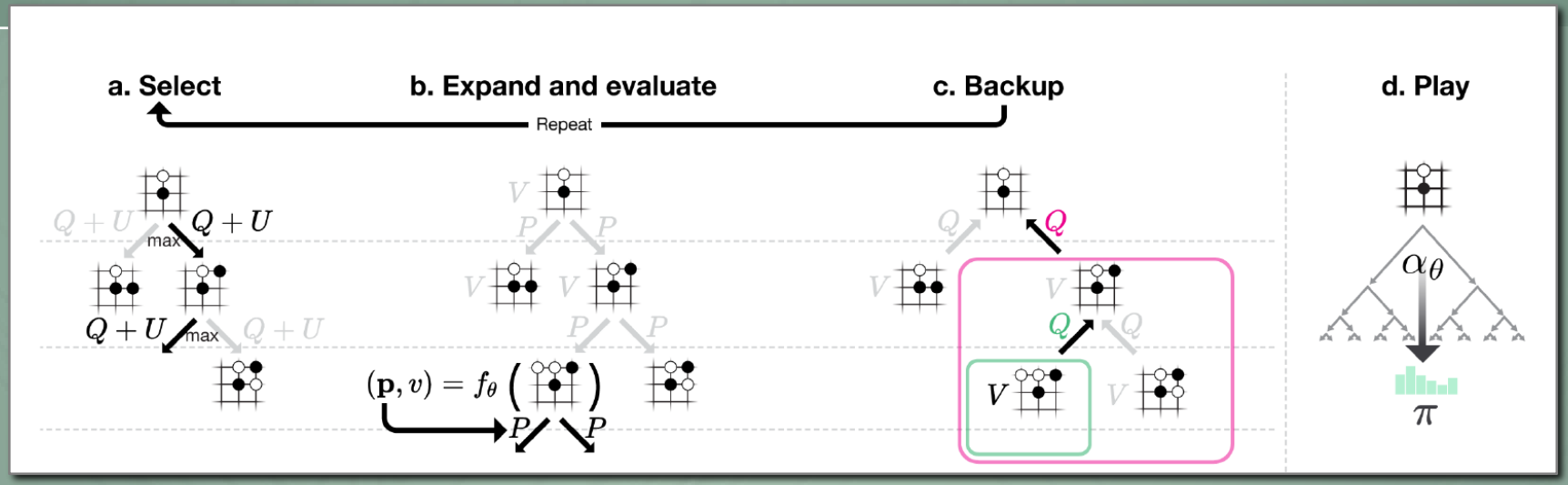
Model-Based Deep Reinforcement Learning

… use or learn a model of the environment to make decisions. They can be more sample-efficient than model-free methods but may struggle with complex environments

Aspect	Given the Model	Learn the Model
Definition	Uses a pre-defined model of the environment	Learns the model of the environment from experience
Examples	▪ AlphaGo/AlphaZero ▪ Expert Iteration	▪ World Models ▪ I2A (Imagination-Augmented Agents)
Advantage	▪ Can leverage expert knowledge ▪ potentially faster initial performance	More flexible, can adapt to changing environments
Disadvantage	May be less adaptable to changes in the environment	Requires more data and time to build an accurate model
Typical Use Case	Environments with well-understood dynamics (e.g., game rules)	Complex or partially observable environments
Planning	Can plan effectively using the given model	Planning improves as the learned model becomes more accurate

AlphaGo

… is AI program developed by DeepMind that uses deep neural networks, tree search algorithms to play complex board game Go at superhuman level



March 2016
Lee Sedol vs.
AlphaGo

Core Components of AlphaGo

- Monte Carlo Tree Search (MCTS):
  Plans sequences of moves by simulating many possible future game states
- Policy Network:
  Predicts probable next moves based on current board state
- Value Network:
  Evaluates the likelihood of winning from a given board position

World Models

… are learned representations of the environment that an agent can use for planning and decision-making

Components

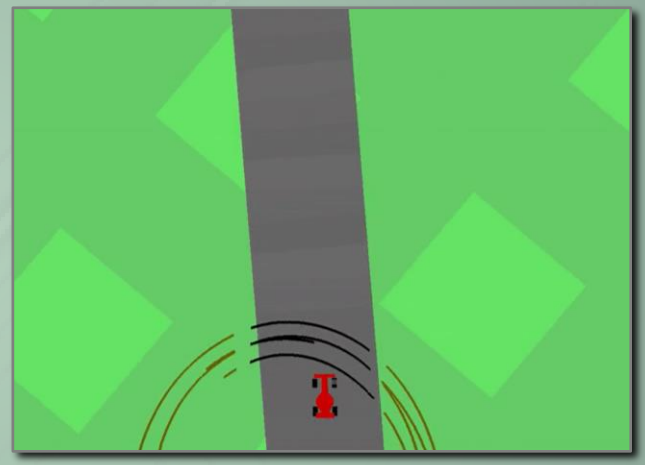
- V: Visual Encoder (VAE)
- M: Memory RNN (LSTM)
- C: Controller (Simple NN)

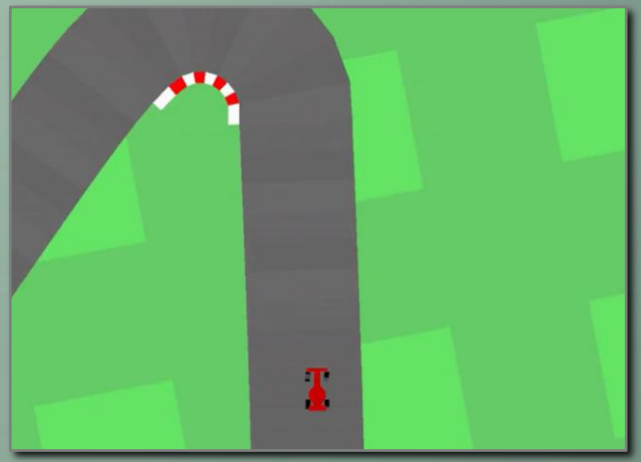
Training Process

- Train V to encode observations
- Train M to predict future
- Train C for desired task

Key Insights

- Agents can learn tasks in their own dream environments
- Compact world model can capture essential aspects of the environment
- Enables imaginative planning and reasoning





Value-Based Methods

... are a class of algorithms that focus on estimating value of states or state-action pairs to guide an agent's decision-making process.

Types of Value Functions

- V(s): State-Value Function
- Q(s,a): Action-Value Function
  - movement arrows that need to be optimized
  - All movement arrows: Policy

Advantages

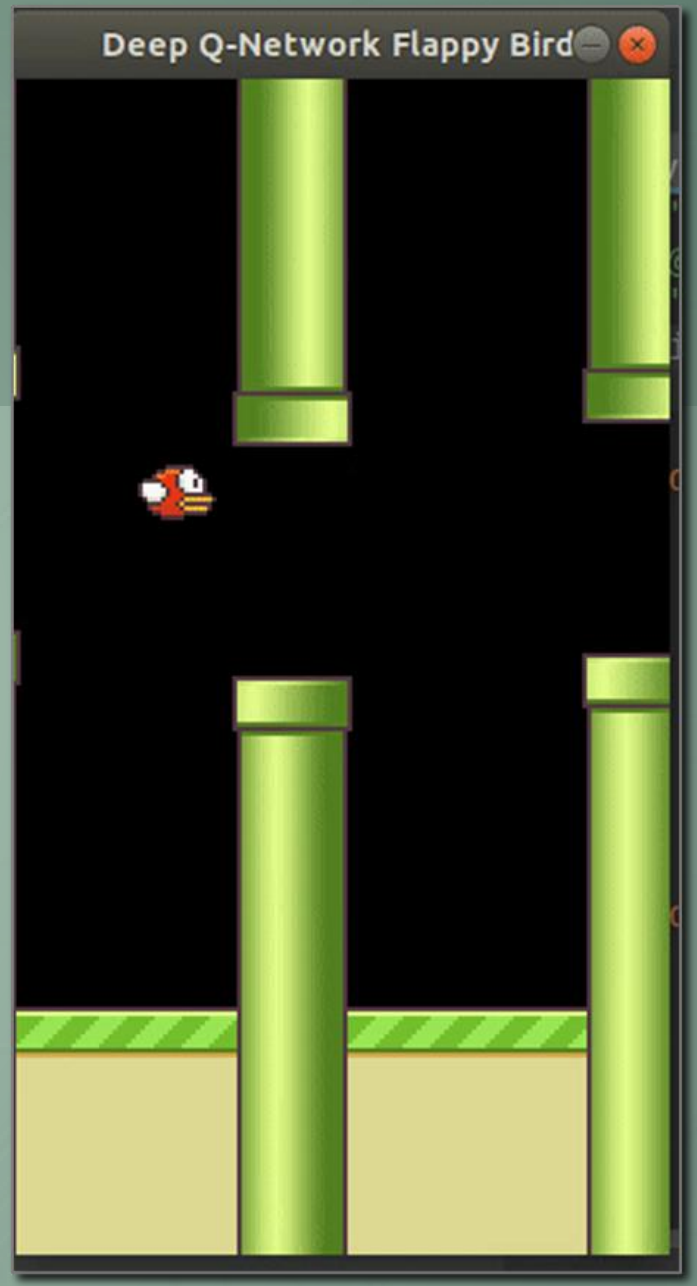
- Sample efficient
- Effective for discrete actions
- Clear state/action value measure

Key Algorithms

- Q-Learning
- SARSA
- DQN (Deep Q-Network)

Challenges

- Continuous actions
- stochastic environments
- Approximations



Policy-Based Methods

… directly learn policy function that maps states to actions, optimizing it to maximize expected rewards.

Key Characteristics

- Learn stochastic policies
- Policy gradient theorem
- On-policy learning

Advantages

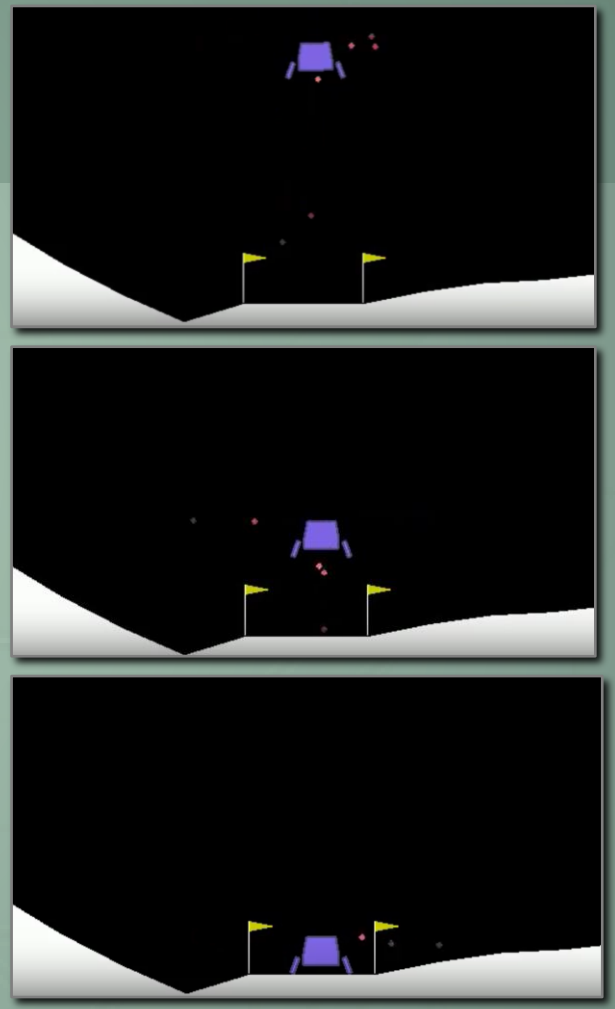
- Effective for continuous action spaces
- Can learn stochastic policies
- Often more stable learning dynamics

Key Algorithms

- REINFORCE
- PPO (Proximal Policy Optimization)
- TRPO (Trust Region Policy Optimization)

Challenges

- Typically high variance in gradient estimates
- Often less sample efficient than value-based methods
- Can be sensitive to hyperparameter choices

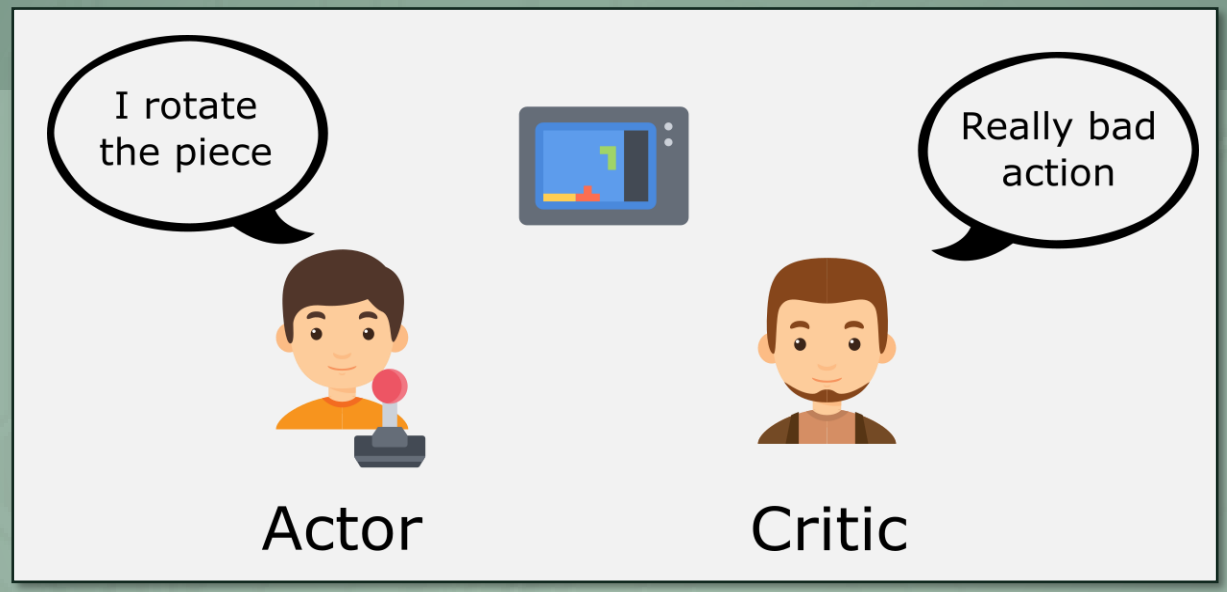


Comparison: Policy-Based Methods

Aspect	REINFORCE	TRPO	PPO
Year Published	1992	2015	2017
Core Principle	Basic policy gradient	KL-divergence constraint	Clipped surrogate objective
Update Mechanism	Direct gradient ascent	Natural gradient descent	Clipped probability ratio
Sample Efficiency	Low	Moderate	Moderate to High
Stability	Low	High	High
Computational Complexity	Low	High	Moderate
Hyperparameter Sensitivity	High	Low	Moderate
Key Advantage	Simplicity	Guaranteed monotonic improvement	Balance of simplicity and performance
Use Cases	▪ Simple environments ▪ Educational purposes ▪ Baseline for comparison	▪ Complex robotics tasks ▪ Game playing (e.g., Atari games) ▪ Continuous control problems	▪ LLM Finetuning/Alignment ▪ Robotic locomotion ▪ Complex, high-dimensional tasks ▪ When stability is crucial

Actor-Critic Methods

… combine policy-based (Actor) and value-based (Critic) approaches to leverage strengths of both methods.



Components

- Learn stochastic policies
- Policy gradient theorem
- On-policy learning

Key Algorithms

- A2C/A3C (Advantage Actor-Critic)
- DDPG (Deep Deterministic Policy Gradient)
- SAC (Soft Actor-Critic)

Advantages

- Reduced variance in policy updates
- Improved sample efficiency
- Effective in continuous action spaces

Challenges

- Balancing Actor and Critic learning
- Potential instability in function approximation
- Complexity in implementation and tuning

Comparison of Actor-Critic RL Algorithms

Aspect	DDPG	A2C/A3C	SAC
Full Name	Deep Deterministic Policy Gradient	Advantage Actor-Critic / Asynchronous Advantage Actor-Critic	Soft Actor-Critic
Year Published	2015	2016	2018
Policy Type	Deterministic	Stochastic	Stochastic
Action Space	Continuous	Discrete or Continuous	Continuous
Key Feature	Combines DPG with ideas from DQN	Uses advantage function to reduce variance	Incorporates entropy regularization
Off-Policy?	Yes	No (On-policy)	Yes
Sample Efficiency	High	Moderate	Very High
Stability	Can be unstable	Moderate	High
Typical Use Cases	▪ Atari games ▪ Simple robotic tasks ▪ When parallelization is beneficial	Robotic manipulation Continuous control tasks When sample efficiency is crucial	▪ Complex robotic tasks ▪ Continuous control with exploration ▪ When stability and efficiency are both important

Comparison of Deep RL Approaches

Aspect	Model-Based	Value-Based	Policy-Based	Actor-Critic
Core Idea	Learn environment model	Estimate value function	Directly optimize policy	Combine value and policy optimization
Sample Efficiency	High	Moderate	Low	Moderate to High
Stability	Can be unstable	Generally stable	Can be unstable	More stable than policy-based
Action Space	Any	Better for discrete	Any, good for continuous	Any, good for continuous
Exploration	Can use model for efficient exploration	Often uses ε-greedy	Can learn stochastic policies	Can balance exploration and exploitation
Example Algorithms	World Models, I2A	DQN, Double DQN	REINFORCE, PPO	A2C, SAC

DATA SCIENCE. INSTITUTE by Fabian Rappert

About me

Dr. Harald Stein

- Data Scientist ~ 8 years experience
- Algotrader ~ 4 years experience
- Ph.D. in Economics, Game Theory

- LinkedIn: https://www.linkedin.com/in/harald-stein-phd-1648b51a
- ResearchGate: https://www.researchgate.net/profile/Harald-Stein