Balancing A Stick On A Moving Cart

Problem Statement

Balancing a stick on a moving cart is a classical control problem. It simulates real-world challenges in robotics, autonomous systems, and artificial intelligence. The objective is to design an agent that learns to stabilize an inverted pendulum using reinforcement learning techniques.

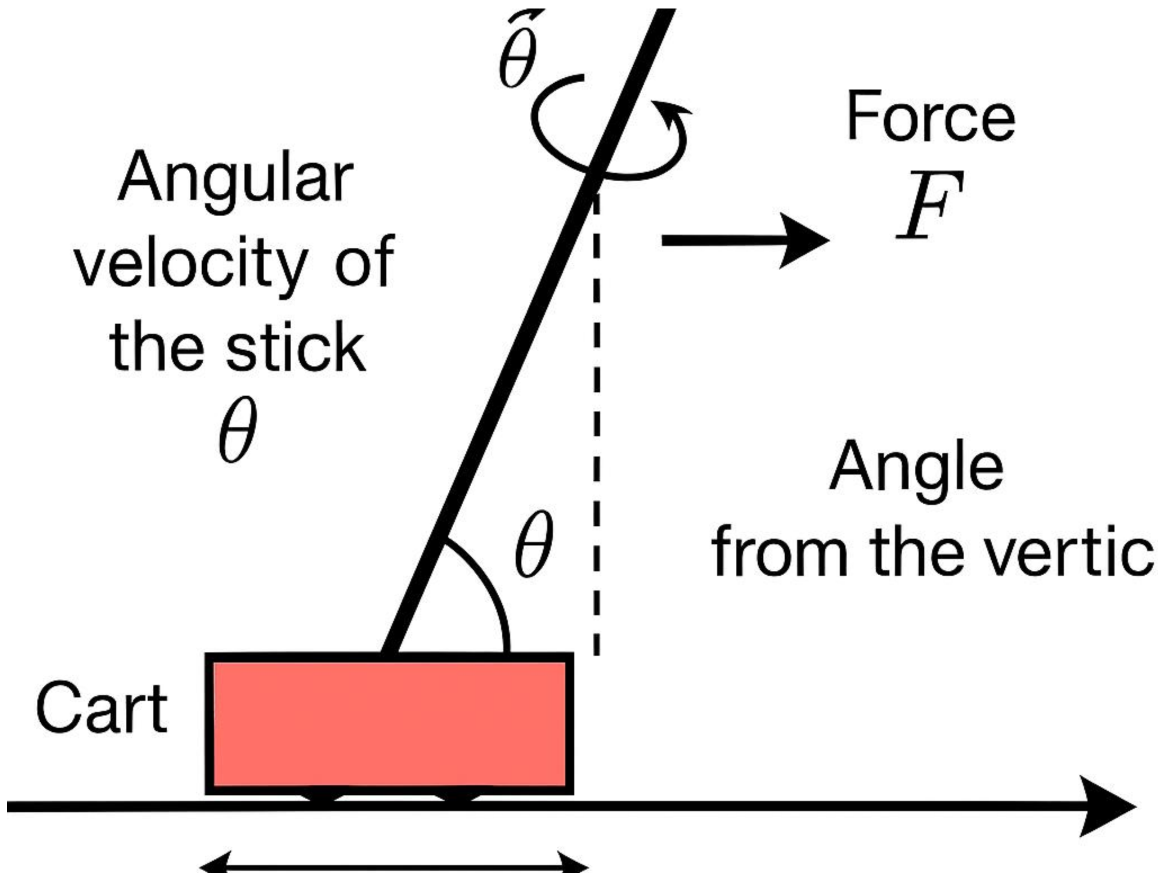

Illustration of the Cart Pole (Inverted Pendulum) Problem

Existing Solutions

- Rule-based Controllers: PID, LQR, etc.
- Hill Climbing: Simple linear policy, struggles with complex states.
- Q-Learning: Limited to discrete spaces.
- Deep Q-Networks (DQN): Generalizes well to large state spaces.
- DDPG: Continuous control using actor-critic architecture.

Our Modification / Contribution

- Hill Climbing with adaptive noise scaling for better exploration.
- Integrated DQN for stable and faster convergence.
- Visualized learning process using custom animations.
- Custom reward shaping for speed and stability.
- Laid foundation for real-world robotic implementation.

<u>LITERATURE REVIEW</u>

Balancing a Stick on a moving cart

Title	**Year**	**Authors**	**Methodology (RL type)**	**Advantages**	**Disadvantages**
Interpretable Control by Reinforcement Learning	2020	Hein *et al.*	Fuzzy and symbolic RL (interpretable policies)	Automatically generates compact, human-readable controllers (fuzzy rules or algebraic equations) for cart-pole, with performance comparable to black-box RL; demonstrated successful balancing on real hardware from human demonstration data.	Offline batch learning (no online RL); requires pre-collected data and expert demonstrations; complexity of controller may still exceed simple PID; limited to the tested scenarios.
A Parametric Study of a Deep RL Control System for Swing-Up Cart-Pole	2020	Escobar *et al.*	DDPG (Deep Deterministic Policy Gradient)	Achieves robust performance under extreme parameter changes: the DDPG agent handled up to 90% change in pole mass and 100% change in cart mass. Post-training adaptation overcame dry-friction effects in only 39 episodes.	Performance degrades severely with unmodeled friction or noise (dry friction "greatly affects" performance); requires extensive simulation and tuning of DDPG; purely simulated study (no real-world tests) and assumes known dynamics.
Double Deep Q Network with Huber Reward for Cart-Pole	2022	Mishra & Arora	DQN and Double DQN (off-policy Q-learning) with Huber loss	Demonstrates that using a Huber loss reward speeds up convergence: the Double DQN agent achieved lower loss and much faster learning than standard DQN. Effective at stabilizing the pole with reduced oscillations.	Relies on careful reward design (Huber vs MSE) and hyperparameter tuning; tested only in simulation; double DQN adds complexity over vanilla DQN; may not generalize to continuous action versions or real hardware without retraining.

An Investigation of Pendulum Control: Comparison of Different Agents	2024	Demircioğlu *et al.*	Various (DDPG, SAC, TD3, PPO, A2C)	Systematic comparison in simulation shows DDPG yielding the most stable and best overall performance. Both SAC and TD3 improved over time, indicating adaptability. Insights on how reward weighting (a/b ratio) affects each agent.	Only simulation (OpenAI Gym) and single inverted pendulum (not double or real cart-pole); newer algorithms (e.g. SAC/TD3) were slower than DDPG; lacks real-world validation; focused on standard continuous-state version, not swing-up.
RL Approach for Inverted Pendulum: Educational Framework	2023	Israilov *et al.*	Q-Learning (tabular) and DQN (deep) in simulation and on a real cart-pole	Demonstrated successful training on real hardware: a few hours of Q-Learning or DQN training sufficed to swing up and balance the pole with high accuracy. Provides both simulation and hardware results, offering practical insights for students.	Basic Q-learning is limited (requires discretization and shows “limitations for this system”); DQN requires more computation and tuning; the study is largely educational (small scale) and does not push state-of-art RL performance.
RL Algorithms in CartPole (Unity ML-Agents)	2024	Jo & Kim	DQN, PPO, A2C (comparison study)	Provides a side-by-side comparison under one framework: found DQN to outperform PPO and A2C in stability and efficiency on CartPole. This benchmarking helps identify which classic methods work best in practice for this task.	Only tested in a simple game environment (Unity-based CartPole); results may vary with implementation and hyperparameters; no new algorithmic contribution—merely a performance comparison.
Spiking Neural Networks for DRL in Robotic Tasks	2024	Zanatta *et al.*	Spiking Neural Networks (SNN) trained with PPO (policy gradient)	Introduces **SpikeGym** framework for training SNNs on control tasks. Shows SNNs can indeed balance the pole with PPO (though not as well as ANNs). Promotes	SNNs underperform conventional ANNs significantly on CartPole; training is less efficient (longer) and network depth is limited; specialized hardware (or

				neuromorphic computing for RL by achieving tasks with biologically inspired networks.	simulation) required, and extra encoding/decoding is needed.
Quantum Advantage Actor-Critic for RL	2024	Kölle *et al*.	Hybrid quantum-classical A2C (variational quantum circuits)	Proposes quantum-enhanced A2C: hybrid models (quantum actor or critic) significantly improve learning efficiency and final performance on CartPole versus purely classical A2C with the same parameter count. Demonstrates a "substantial performance increase" for hybrid over classical.	Current quantum hardware limitations: noisy qubits and small scale restrict achievable circuit depth. The approach is computationally intensive and preliminary; performance gains are modest and do not yet justify practical use on current NISQ devices.
Quantum vs Classical DQN in Dynamic Control	2025	Zare & Boroushaki	Quantum DQN (variational ansatz) vs classical DQN	First head-to-head study of quantum DQN (with various ansatz circuits) on CartPole: found that a "RealAmplitudes" quantum ansatz converges faster and yields robust control even under disturbances. Shows quantum RL agents can match the performance of classical DQN and remain stable under perturbations.	In practice, classical DQN still converges faster in some cases. Quantum circuits are limited to toy problems due to qubit count and noise; overhead of quantum training outweighs benefits currently; results are mostly proof-of-concept on simulated quantum hardware.
Benchmarking Robust RL: Disturbance Injection in CartPole	2022	Glossop *et al*.	Standard vs Robust RL (PPO, TRPO, WCPG, etc.)	Large-scale evaluation shows that vanilla RL agents often perform comparably to robustified versions under disturbances.	Robust RL methods provided only modest improvements in this benchmark. The study is empirical (no new algorithm) and focused

				Important insight: RL controls are surprisingly robust to action noise, and even basic algorithms achieve near state-of-art performance in this setting.	on disturbance injection; real-world deployment still challenging. Limited to simulation (benchmark suite); no direct RL method recommendations (beyond broad conclusions).
Cart-Pole as a Neuromorphic Benchmark	2025	Plank *et al.*	Spiking Neural Networks (SNN) with evolutionary training	Proposes cart-pole as a scalable neuromorphic benchmark. Demonstrates that very small SNNs (≤12 neurons, no leak) trained by evolutionary algorithms can control the pole across increasing difficulty levels. Highlights advantages for low-power hardware implementations (e.g. fast, efficient spiking controllers).	The approach relies on computationally expensive genetic training (no on-chip learning rule). Spiking networks here solve a simplified version of the task; real-time learning or noise robustness is not addressed. Specialized neuromorphic hardware is required for full benefit, limiting generality.
Sim-to-Real RL for Double-Inverted Pendulum	2025	Ju *et al.*	Distributional RL (Truncated Quantile Critics, TQC)	Trains a single policy that transitions a rotary double-inverted pendulum among four equilibria. Achieved zero-shot transfer: the TQC-trained controller worked on the real hardware “without any additional tuning or calibration”.	Requires an accurate mathematical model and complex training (distributional RL with many critics). The method is tailored to a specific hardware setup (rotary pendulum); scalability to other systems is unclear. High computational cost and reliance on good simulation fidelity.