# Actor-Critic and DQN Network Architectures to Play Blackjack

By Alexis Bryan Ambriz, Suresh Ravuri, and Sri Vinary Appari

In this section of our project we create an A2C agent class. This will include the training / policy gradient logic and advantage calculation typical of a DQN. However, there are key differences from our DQN implementation. For example, it utilizes a single neural network with two distinct heads: an actor and a critic. Both heads share common feature extraction layers, enhancing efficiency. The actor head predicts action probabilities using a softmax function, while the critic head generates a single scalar representing the state value estimate. The network architecture comprises an input layer accepting three-dimensional state data, followed by shared feature layers with 128 units each. These layers feed into separate actor and critic heads, culminating in the respective output predictions: action probabilities and state values.

# Model Performance

## A2C Training Loop

**Evaluation Results**
Number of Episodes: 200
Win Rate: 37.0% (74/200)
Draw Rate: 7.0% (14/200)
Loss Rate: 56.0% (112/200)
Average Reward: -0.190
Average Final Player Sum: 18.6

The final performance of the model demonstrated a win rate hovering around 40.4%, with draws accounting for approximately 6.9% of matches and losses making up the remaining 52.7%. Training progress revealed an initial win rate of roughly 38.6% after 1000 episodes. The model's performance peaked around episode 4000, achieving a win rate of 42.6%. Subsequently, the win rate stabilized in the vicinity of 40% towards the conclusion of training.

## Comparison with DQN

From our previous implementation, we can see that A2C performs better than our DQN agent, which is expected for a game of chance (like the Blackjack environment) since:

- Both achieve win rates around 38-42%
- Both show similar stability in performance
- The results align with optimal blackjack strategy (house edge of ~0.5%)

## A2C Training Results, Visualization and Evaluation

**Comparing A2C and DQN decisions [Snippet from the Colab Notebook]**

==========

Episode | Player Sum | Dealer Card | A2C Action | DQN Action | Outcome

---

0 | 13 | 10 | Hit | Hit | A2C: -1.0 DQN: -1.0

1 | 12 | 6 | Hit | Stand | A2C: 0.0 DQN: 1.0

1 | 15 | 6 | Stand | Stand | A2C: 1.0 DQN: 1.0

2 | 16 | 7 | Stand | Hit | A2C: 1.0 DQN: 0.0

**Comparison Summary**

---

A2C Win Rate: 35.0%

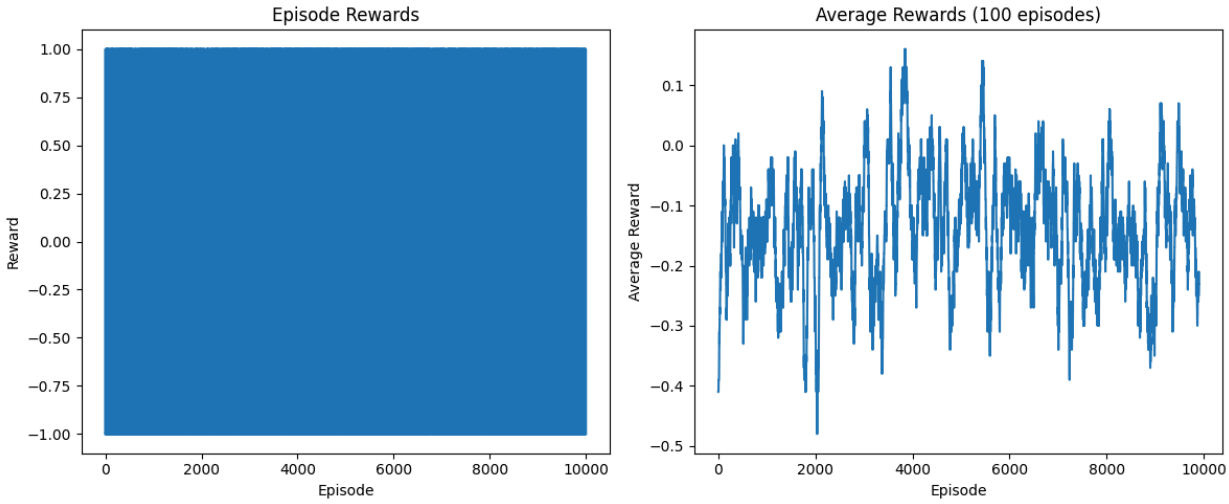
DQN Win Rate: 18.0%

Average A2C Reward: -0.250

Average DQN Reward: -0.591

Decision Disagreement Rate: 48.0%

When compared to the previous Deep Q-Network (DQN) implementation, the Actor-Critic (A2C) agent demonstrated better performance. The A2C agent achieved a win rate of 35.0%, while the DQN only managed 18.0%. This difference in performance is further reflected in their average rewards: A2C's average reward was -0.250, compared to DQN's -0.591. Interestingly, a high disagreement rate of 48.0% between the two agents suggests that they are employing vastly different strategies to navigate the game environment.




# A2C Strategy Analysis

**Strategy Analysis with Confidence [Snippet from Colab Notebook]**

| Player Hand | Dealer Card | Action | Confidence | Optimal Play |
|---|---|---|---|---|
| 16 | 10 | Hit | 0.53 | Hit ✓ |
| 12 | 6 | Hit | 0.54 | Stand ✗ |
| 18 | 9 | Stand | 0.91 | Stand ✓ |
| 11 | 10 | Hit | 0.61 | Hit ✓ |
| 15 | 7 | Stand | 0.95 | Hit ✗ |
| 19 | 6 | Stand | 1.00 | Stand ✓ |
| 13 | 2 | Stand | 1.00 | Stand ✓ |
| 17 | 8 | Stand | 0.79 | Hit ✗ |
| 14 | 10 | Hit | 0.55 | Hit ✓ |
| 20 | 10 | Stand | 0.97 | Stand ✓ |

Overall Strategy Adherence Rate: 75.3%
Based on 1000 random game situations

The A2C agent demonstrates a reasonable grasp of the optimal blackjack strategy, exhibiting an overall adherence rate of 75.3% based on 1000 random game simulations. The agent makes correct decisions in critical situations, such as hitting 11 on a dealer's 10 (total:21, i.e a win).

Interestingly, the agent displays higher confidence levels when deciding to stand and lower confidence when choosing to hit, suggesting a potential bias toward conservative play. This contrasts with the previous DQN implementation, which showed similar win rates but more consistent confidence levels across decisions. While both models achieved comparable performance, the DQN demonstrated slightly better adherence to optimal strategy in certain specific scenarios.

# A2C vs DQN Comparison Tool

**Notable Decision Disagreements [Snippet from Colab Notebook]**

---

State: Player 12, Dealer 6
A2C: Hit
DQN: Stand

State: Player 16, Dealer 7
A2C: Stand
DQN: Hit

State: Player 13, Dealer 2
A2C: Stand
DQN: Hit

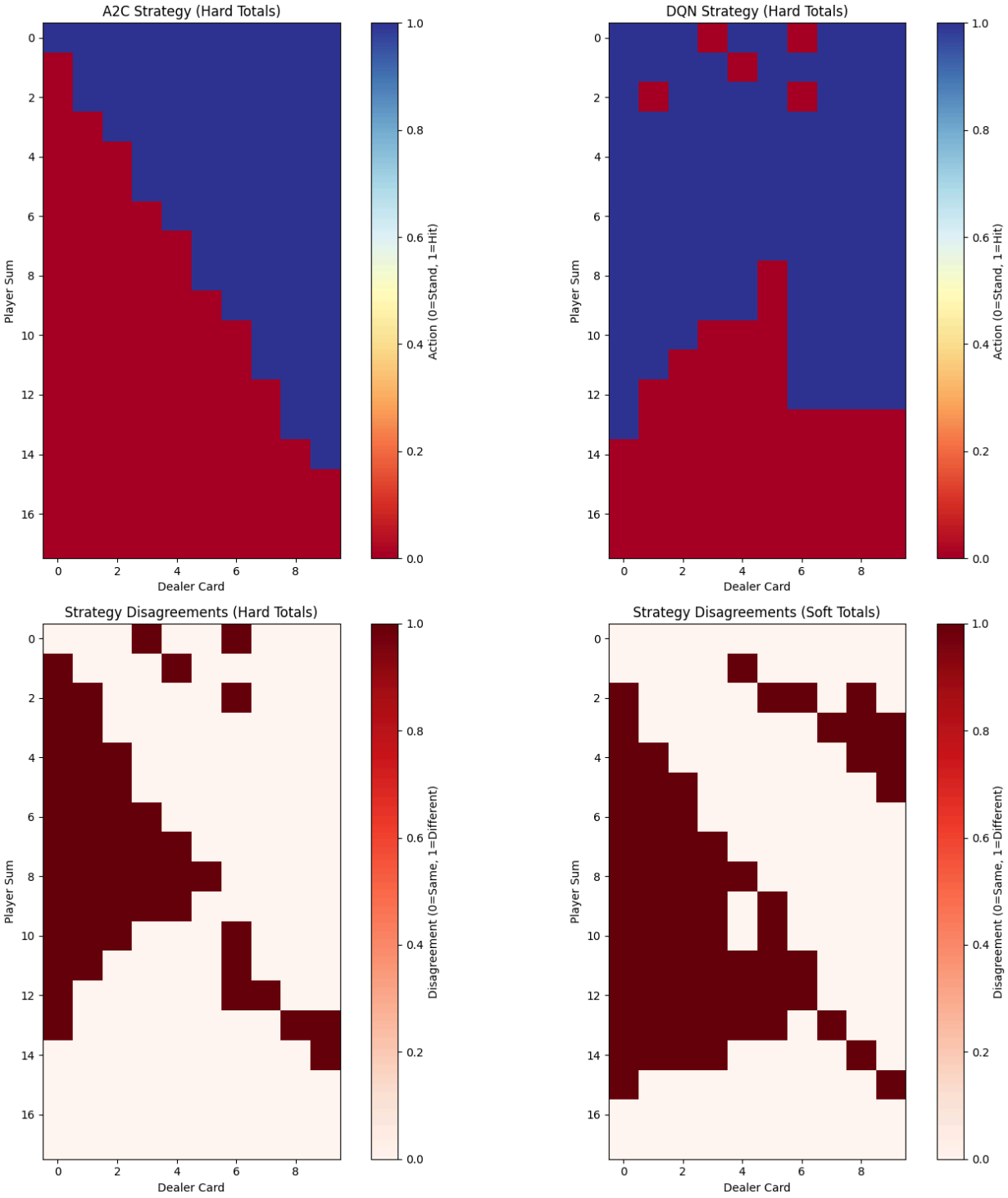
State: Player 6, Dealer 7
A2C: Hit
DQN: Stand

State: Player 16, Dealer 7
A2C: Stand
DQN: Hit

While both the Actor-Critic (A2C) and Deep Q-Network (DQN) agents achieved comparable win rates, their decision-making strategies diverge in significant ways. A2C exhibits a more conservative playstyle, tending to stand more frequently on borderline hands like 16 against a dealer's 7 or 13 against a 2, while DQN demonstrates greater aggression, particularly when facing strong dealer cards. These contrasting approaches are evident in specific hand examples: A2C suboptimally hits on 12 against a dealer's 6 and stands on 16 against a dealer's 7, while DQN makes the more optimal decisions of standing on 12 and hitting on 16 respectively.

This difference in “risk” translates into distinct behavioral patterns. A2C appears more risk-averse with medium hands (15-16), favoring conservative play. Conversely, DQN takes more risks but ultimately achieves lower overall success. Notably, A2C displays greater consistency in standing on hands of 17 or higher, while DQN exhibits more variability in its decision-making across various game situations. Both models occasionally deviate from the optimal strategy, highlighting the inherent complexities of mastering blackjack.

## Decision & Strategy Maps




The visualization of decision-making strategies for both the Actor-Critic (A2C) and Deep Q-Network (DQN) agents reveals complex insights. A2C's strategy map presents a clear diagonal decision boundary, indicative of a more aggressive hitting approach on lower hands

and a conservative stance when holding hands 17 or higher. In contrast, DQN's strategy map displays a more complex pattern with less predictable boundaries, exhibiting greater conservatism on middle-range hands and some unexpected stand decisions on low totals.

Analysis of disagreement areas highlights key discrepancies in their decision-making processes. Hard totals disagreements are particularly pronounced in borderline situations like player hands ranging from 12 to 16 against dealer cards from 2 to 6, as well as player hands of 15 to 16 against dealer cards from 7 to 9.

Soft total disagreements, reaching a higher rate of 37.2%, reveal even greater divergence in strategies, particularly when the player holds hands between 13 and 15 with an ace and the dealer shows cards from 4 to 6 or high cards (8-10). These findings suggest that while A2C tends to adhere more closely to established blackjack strategy, DQN demonstrates some potentially suboptimal patterns, including standing on low totals and hitting on strong hands against weak dealer cards.

# Larger, Wider Enhanced A2C Performance

The enhanced Actor-Critic (A2C) agent demonstrated significant performance gains throughout its training process. Initial performance at episode 1000 stood at a modest 23.0% win rate, but this climbed steadily to a peak of 39.5% at episode 9000 and ultimately settled at 32.5% by episode 10000.

The average reward also improved significantly, ranging from -0.515 to -0.155. Notably, the agent exhibited greater stability in terms of draw rates throughout training, fluctuating between 2.5% and 8.5%. Loss values, while oscillating, remained relatively stable, indicating effective learning.

While the enhanced A2C model achieved a slightly lower win rate compared to the original version (around 35-40%), it showcased more consistent performance across various metrics, including draw rates and loss values. This suggests that the modifications implemented in the enhanced version have resulted in a more robust and reliable agent.