

Comparative Analysis of Parameterized Action Actor-Critic Reinforcement Learning Algorithms for Web Search Match Plan Generation

Ubayd Bapoo and Clement N Nyirenda

University of the Western Cape, Bellville 7535, South Africa,
2934126@myuwc.ac.za, cnyirenda@uwc.ac.za

Abstract. This study evaluates the performance of Soft Actor Critic (SAC), Greedy Actor Critic (GAC), and Truncated Quantile Critics (TQC) in high-dimensional decision-making tasks using fully observable environments. The focus is on parameterized action (PA) spaces, eliminating the need for recurrent networks, with benchmarks Platform-v0 and Goal-v0 testing discrete actions linked to continuous action-parameter spaces. Hyperparameter optimization was performed with Microsoft NNI, ensuring reproducibility by modifying the codebase for GAC and TQC. Results show that Parameterized Action Greedy Actor-Critic (PAGAC) outperformed other algorithms, achieving the fastest training times and highest returns across benchmarks, completing 5,000 episodes in 41:24 for the Platform game and 24:04 for the Robot Soccer Goal game. Its speed and stability provide clear advantages in complex action spaces. Compared to PASAC and PATQC, PAGAC demonstrated superior efficiency and reliability, making it ideal for tasks requiring rapid convergence and robust performance. Future work could explore hybrid strategies combining entropy-regularization with truncation-based methods to enhance stability and expand investigations into generalizability.

Keywords: Match Plan, Parameterised Action, Soft Actor-Critic, Greedy Actor-Critic, Truncated Quantile Critics

1 Introduction

Over the past two decades, text search engine technology has advanced significantly through innovations in index design, storage, and query evaluation (Zobel and Moffat, 2006). However, many contemporary techniques remain underrepresented in textbooks, creating resource gaps. The rapid growth of web documents linked to recognized keywords has made processing them all in short timeframes impractical (Luo et al., 2021), necessitating scalable and efficient search mechanisms. Search engines address this by organizing documents into ranked posting lists and employing match plans—sequential rules balancing result quality and response time—to process billions of pages efficiently (Rosset et al., 2020).

Despite these advances, challenges such as varied match rules, unpredictable data distribution, and rigid designs persist, requiring frequent customization.

Luo et al. (2021) reframed match plan generation as a reinforcement learning (RL) problem, eliminating predefined inputs like stopping quotas. Their approach dynamically parameterizes state signals and match plans for adaptive development. Actions are represented in a hybrid parameterized space using runtime signals and static query features, with sparse, delayed rewards based on result quality and response time. To address these challenges, Luo et al. proposed Parameterized Action Soft Actor-Critic (PASAC), a deep RL algorithm combining discrete actions and continuous parameters for high-level decision-making. Inspired by Bester et al. (2021) and Haarnoja et al. (2018), PASAC integrates parameterized actions with SAC's maximum entropy framework, enabling simultaneous reward optimization and entropy maximization. This enhances exploration while maintaining a stable actor-critic architecture.

While PASAC's SAC introduces a novel approach to match plan generation, emerging Actor-Critic reinforcement learning methods offer strong alternatives. Greedy Actor-Critic (GAC) (Neumann et al., 2023) improves convergence speed and stability, outperforming SAC in some tasks. Truncated Quantile Critics (TQC) (Kuznetsov et al., 2020) mitigates overestimation bias through distributional representations, truncation, and ensembling. These algorithms present diverse strategies for balancing exploration and exploitation in high-dimensional continuous action spaces, necessitating comparison with SAC in parameterized action spaces. This study evaluates SAC, GAC, and TQC in two games over 5,000 episodes to assess policy learning and decision-making effectiveness. Key metrics include evaluation returns, training efficiency, convergence speed, and reward maximization. By comparing these algorithms, the study aims to provide insights into their performance, resilience, and applicability, advancing understanding of parameterized reinforcement learning in high-dimensional settings.

The structure of the rest of this paper is as follows: Section 2 provides an overview of match plan generation, while Section 3 introduces actor-critic reinforcement learning. In Section 4, we describe the Parameterized Actor-Critic Algorithms employed in this study. Sections 5 outlines the simulation setup and the ensuing results, respectively. Finally, Section 6 offers the study's conclusions.

2 Match Plan Generation Overview

Rosset et al. (2020) explain that search engines represent documents using fields—descriptions sourced from various inputs. Each field includes an inverted index where keywords are stored with posting lists, organized by document-level properties like static rank. During query processing, the system classifies the query into predefined types and generates a match plan to guide document scanning with specific criteria and stopping conditions, as illustrated in Figure 1. The merged documents are then subject to ranking and filtering to improve the quality of the search results. Match rules define the constraints a document must meet for ranking, while stopping criteria dictate when to end the index scan or proceed to the next rule. Together, they balance fast query responses with high-quality results. Complex or rare queries may require deeper matching

strategies to uncover relevant documents. After executing the match plan, the selected candidates are refined using machine learning models to ensure optimal ranking and relevance.

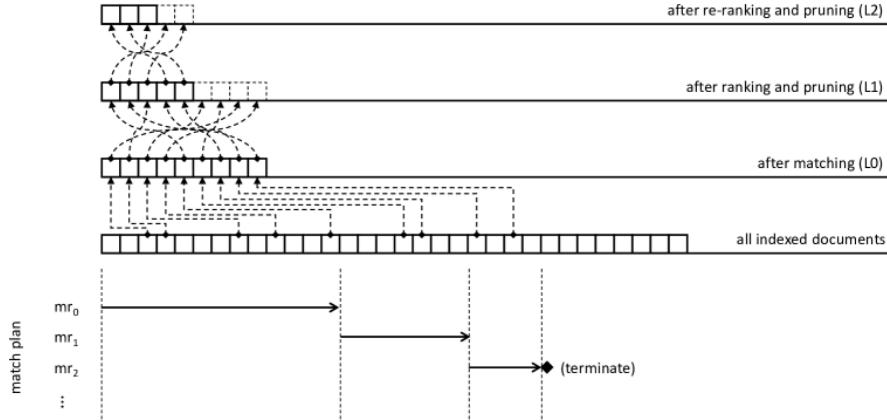


Fig. 1. Illustration of the telescoping design used in Bing’s retrieval mechanism.

3 Overview on Actor-critic Reinforcement Learning

Actor-critic methods are a class of algorithms in Reinforcement Learning (RL) (Sutton and Barto, 2018). Before delving into the details of actor-critic methods, it is important to first provide an overview of the fundamentals of RL.

3.1 RL Fundamentals

RL is a machine learning subfield focused on teaching agents to act in environments to maximize cumulative rewards. Unlike supervised learning, which relies on labeled data, RL involves learning through interactions, enabling agents to explore and exploit their environment to discover optimal strategies (Sutton and Barto, 2018). An agent observes the current state of the environment s_t at time t and chooses an action a_t based on a policy π , which is a mapping from states to actions:

$$\pi : S \rightarrow A, \quad (1)$$

where S represents the state space, and A represents the action space. After the agent takes an action a_t , the environment responds by transitioning to a new state s_{t+1} and providing a scalar reward r_t . The agent’s objective is to maximize the expected sum of rewards over time, typically represented by the return G_t ,

which is defined as the discounted sum of future rewards:

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}, \quad (2)$$

where γ is the discount factor (with $0 \leq \gamma \leq 1$), representing the importance of future rewards relative to immediate rewards. RL has two primary components: the policy and the value function, which are described in more detail as follows:

1. *Policy*: A strategy used by the agent to decide which action to take in a given state. It can be deterministic or stochastic. For a given state s_t , the policy determines the probability of taking action a_t :

$$\pi(a_t|s_t) = P(a_t|s_t) \quad (3)$$

2. *Value Function*: A function that estimates the expected return from a given state (or state-action pair). The state-value function $V(s)$ estimates the expected return starting from state s and following the policy π :

$$V^\pi(s) = \mathbb{E}[G_t|s_t = s] \quad (4)$$

For an action-value function $Q(s, a)$, the expected return is based on both the state and action:

$$Q^\pi(s, a) = \mathbb{E}[G_t|s_t = s, a_t = a] \quad (5)$$

Value-based methods, such as Q-learning (Watkins and Dayan, 1992), SARSA (Rummery and Niranjan, 1994), and Deep Q-learning (Mnih et al., 2015), focus on estimating the value function and derive the policy indirectly by acting greedily with respect to this function. These methods have strengths, such as their ability to learn effectively in discrete action spaces, but struggle in environments with continuous action spaces, as finding the action that maximizes becomes computationally challenging. However, policy-based methods, such as REINFORCE (Williams, 1992), directly optimize the policy by using gradient ascent to maximize expected return. These methods excel in continuous action spaces and naturally handle stochastic policies, which are beneficial in exploration. However, they suffer from high variance in gradient estimates, leading to slow convergence and instability.

3.2 Actor-critic Techniques

Actor-critic methods are a class of RL algorithms that merge value-based and policy-based approaches, providing a robust framework for complex decision-making tasks. These techniques involve two components: the actor and the critic (Konda and Tsitsiklis, 1999; Sutton and Barto, 2018). The actor selects actions based on a policy, while the critic evaluates these actions by estimating the value function. This separation exploits the strengths of both approaches, making actor-critic methods particularly effective in continuous action spaces (Silver et al., 2014) and large state spaces (Mnih et al., 2016). Mathematically, the actor-critic framework is defined by two primary functions:

1. **The Actor (Policy):** The policy function $\pi(a|s; \theta)$ defines the probability of selecting an action a in state s , given parameters θ . The goal of the actor is to adjust its parameters to maximize the expected return $J(\theta)$, which is the cumulative reward that the agent expects to obtain over time:

$$J(\theta) = \mathbb{E} \left[\sum_{t=0}^T \gamma^t r_t \mid s_0, a_0, \dots \right]. \quad (6)$$

2. **The Critic (Value Function):** The critic estimates the value function $V(s)$, which evaluates the expected return for being in state s and following the current policy. The value function is updated based on the temporal difference (TD) error, δ_t , which is the difference between the expected value and the actual reward:

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t). \quad (7)$$

The critic uses the TD error to adjust the value function parameters ϕ , typically using gradient descent methods:

$$\phi \leftarrow \phi + \alpha \delta_t \nabla_\phi V(s_t), \quad (8)$$

where α is the learning rate.

The actor updates the policy parameters θ by utilizing the advantage function $A(s_t, a_t)$, which measures the difference between the actual return and the estimated value of the state-action pair. The advantage function is defined as:

$$A(s_t, a_t) = Q(s_t, a_t) - V(s_t), \quad (9)$$

where $Q(s_t, a_t)$ represents the action-value function. The policy update rule for the actor is typically performed using the policy gradient method:

$$\theta \leftarrow \theta + \beta \nabla_\theta \log \pi(a_t|s_t; \theta) A(s_t, a_t), \quad (10)$$

where β is the learning rate (Williams, 1992). In these methods, the actor-critic interaction is key for efficient learning. The critic guides the actor by indicating actions that yield higher rewards, while the actor refines the policy based on the critic's feedback.

3.3 Emerging Deep Learning based Actor-Critic Techniques

Over the past decade, numerous actor-critic extensions, primarily leveraging deep learning architectures, have been introduced. These advancements have significantly improved both the stability and performance of reinforcement learning models. Notable examples include Deep Deterministic Policy Gradient (DDPG) (Lillicrap et al., 2015), Advantage Actor-Critic (A2C) (Mnih, 2016), Proximal Policy Optimization (PPO) (Schulman et al., 2017), Twin Delayed Deep Deterministic Policy Gradient (TD3) (Fujimoto et al., 2018), Soft Actor-Critic (SAC)

(Haarnoja et al., 2018), Truncated Quantile Critics (TQC) (Kuznetsov et al., 2020), and Greedy Actor-Critic (GAC) (Neumann et al., 2023). This subsection focuses on the three most recent methods in this list, which were implemented and evaluated in this study. These techniques are detailed as follows:

1. **Soft Actor-Critic (SAC)**: SAC maximizes an agent’s interaction with its environment through soft policy iteration, evaluating and improving policies under the maximum entropy framework (Ding et al., 2022). SAC outperforms other RL algorithms in continuous action benchmarks, excelling in learning speed and resilience (Ding et al., 2022). It optimizes the maximum entropy objective, balancing exploration and exploitation (Haarnoja et al., 2018). The entropy target promotes broad exploration, discards unfeasible options, and allows policies to capture multiple near-optimal behaviors, assigning equal probability to equally appealing actions.
2. **Greedy Actor-Critic (GAC)**: Neumann et al. (2023) introduce a cross-entropy method for policy improvement, iteratively selecting the top percentile of actions based on their learned values. The process focuses on the most effective actions for each state-action pair. N actions are sampled using a proposal policy, sorted by magnitude, and the policy is updated to increase the likelihood of the top $\lceil \rho N \rceil$ actions for $\rho \in (0, 1)$. Neumann et al. (2023) highlight GreedyAC’s advantages over Boltzmann greedification, showing it guarantees policy improvement in the original MDP and prevents policy collapse by incorporating entropy into the proposal policy rather than the final actor policy. This design allows the agent to explore potentially optimal actions without bias in decision-making.
3. **Truncated Quantile Critics (TQC)**: Foo et al. (2022) applied CPT Actor and TQC in DRL trading strategies. TQC, a quantile-based distributional RL approach, mitigates overestimation bias in off-policy algorithms by combining three concepts: (i) learning a critic’s distributional representation, (ii) pooling critics, and (iii) truncating pooled critics. It uses an actor-critic architecture, where the truncated critics estimate a risk-sensitive Q-function, guiding action selection and resulting in a stable policy. Kuznetsov et al. (2020) propose truncating continuous distributional quantile critics to reduce overestimation bias by selecting atoms with the highest values and calculating the Q-value from the remaining atoms. This method improves Q-value evaluation accuracy by adjusting the number of atoms, resolving inflated overestimation caused by high return variation.

4 Parameterised Action Actor-Critic Algorithms

This section outlines the implementation of parameterised action (PA) actor-critic algorithms, evaluated in this study. It begins by introducing the PA concept and its relevance to match plan generation. Next, it describes the formulation of the match plan generation problem as a reinforcement learning (RL) task. Finally, it details the integration of PA with the actor-critic techniques discussed in Section 3, providing a generalized framework for implementation.

Masson et al. (2016) describe a parameterized action as discrete but accompanied by a real-valued vector serving as its parameter. This modeling approach introduces structure by differentiating between continuous actions. Each step requires the agent to select an action and execute it with specified parameters. For instance, consider a football-playing robot capable of kicking, passing, and running. Each action can be associated with a continuous parameter vector, such as kicking the ball with a specified force, passing to a precise location, or running at a particular speed. These actions are parameterized uniquely based on their respective requirements. Using reinforcement learning (RL) for match plan generation in search engines presents several challenges:

1. *Hybrid Action Space*: The environment for match plan creation combines discrete and continuous actions, challenging conventional RL algorithms that handle only one action type at a time.
2. *Complex Environment*: Search engine settings are complex, with diverse queries and dynamic state variables. Queries require high-dimensional representations, and intermediate state signals span wide continuous value ranges, increasing complexity.
3. *Sparse Rewards*: Sparse rewards hinder RL training, as early policies offer limited guidance, leading to infrequent positive feedback and slower convergence.

Addressing these challenges requires advanced techniques capable of handling hybrid action spaces, managing high-dimensional data, and mitigating the limitations of sparse reward signals. The Parameterised Action (PA) Actor-Critic Framework handles hybrid action spaces, combining discrete and continuous actions. In the original SAC, an actor network estimates the mean and variance of a Gaussian distribution for continuous actions, while a critic network evaluates their Q values. PASAC extends this by using separate actor branches to generate both action types simultaneously. These branches share initial layers to encode state information but differ in configurations. PASAC uses a single critic network to estimate Q values for both types. Unlike traditional actor-critic methods, the PA framework integrates parameterised actions through significant modifications.

1. *Dual Actor Networks*: The actor network generates discrete and continuous actions simultaneously, with shared state-encoding layers that diverge into specialized structures for each action type.
2. *Unified Critic Network*: A single critic network evaluates the combined Q-values of discrete and continuous actions, maintaining a consistent hybrid action space representation during training.
3. *State Representation Sharing*: Sharing parameters in the early actor network layers enables efficient state representation and flexibility in creating discrete and continuous actions.

This extension of the actor-critic paradigm allows for successful decision-making in contexts where actions are parameterised, providing a formal answer to the issues presented by hybrid action spaces.

5 Simulation Setup and Results

This study evaluates the adaptability and efficacy of Soft Actor Critic (SAC), Greedy Actor Critic (GAC), and Truncated Quantile Critics (TQC) in high-dimensional decision-making with fully observable environments, removing the need for recurrent networks. The benchmarks, Platform-v0 and Goal-v0, test agents on discrete actions linked to continuous action-parameter spaces. Hyperparameter optimization was performed using Microsoft NNI, with the same codebase modified for GAC and TQC to ensure reproducibility. Ten tests were conducted to assess algorithm stability, efficiency, and performance, based on mean training incentives, final evaluation scores (averaged over 10 repetitions), and time to complete 5,000 episodes.

5.1 Platform Game Description and Results

The Platform domain features three actions: run, hop, and leap, each with a continuous parameter for horizontal displacement. To reach the target state, the agent must hop over opponents and leap across gaps. Contact with an opponent or falling through a gap results in failure. The 9-dimensional state space includes the agent’s and local opponent’s positions and velocities, along with platform characteristics such as length (Bester et al., 2021). Table 1 presents the results for the Platform Game. PAGAC achieves the fastest training times in the Platform game, averaging 41:24 for the benchmark. It completes 5,000 episodes quicker than the other three algorithms while achieving the highest average evaluation return. With a maximum return of 0.1705, PAGAC consistently outperforms its counterparts.

Table 1. Performance Results for Platform Game

Algorithm	Average Evaluation Return	Average Time (hh:mm:ss)
PAGAC	0.170	0:41:24
PASAC	0.143	0:56:35
PATQC	0.145	0:51:38

5.2 Robot Soccer Goal Game Description and Results

The Robot Soccer Goal game simplifies RoboCup 2D, tasking an agent to score against a goalie intercepting the ball. The three parameterised actions—kick-to, shoot-goal-left, and shoot-goal-right—require kicking the ball, with the agent automatically approaching it until close enough to kick again. The state space includes 14 continuous features describing the positions, velocities, and orientations of the agent, goalie, and ball, along with the ball’s distance from the keeper and goal (Bester et al., 2021). Table 2 presents the results for the Robot Soccer

Goal Game. PAGAC achieves the fastest training time in the Goal game, averaging 24:04 for 5,000 episodes. It outperforms the other algorithms. PATQC slightly surpasses PASAC in average return, scoring -8.1552 compared to PASAC's -8.2003.

Table 2. Performance Results for Soccer Goal Game

Algorithm	Average Evaluation Return	Average Time (hh:mm:ss)
PAGAC	-8.552	0:24:04
PASAC	-8.200	0:26:55
PATQC	-8.155	0:27:19

In general, these results show that PAGAC consistently achieved the fastest training times: 41:24 for Platform and 24:04 for Goal. PATQC followed with 51:38 for Platform and 27:19 for Goal, while PASAC was slowest at 56:35 and 26:55, respectively. Although PATQC slightly outperformed PASAC in returns, PAGAC excelled in speed and performance stability, proving its advantages in high-dimensional continuous action spaces. These findings underscore PAGAC's efficiency for rapid convergence and reliable decision-making. Alongside PATQC, it provides efficient, robust solutions, reaffirming prior research.

6 Conclusion and Future Works

This work compared the soft actor critic (SAC) with the greedy actor critic (GAC) and the truncated quantile critic (TQC) in parametrized action (PA) spaces. Using fully observable environments, it focused on optimizing parameterized action spaces without recurrent networks. Benchmarks Platform-v0 and Goal-v0 tested agents with discrete actions tied to continuous parameters, challenging their ability to handle complex decisions. The results revealed that parametrized action greedy actor-critic (PAGAC) outperformed its counterparts. PAGAC consistently trained faster and provided stable, reliable decision-making, excelling in efficiency and performance. Truncated Quantile Critics (TQC) and Soft Actor Critic (SAC) delivered competitive outcomes but lagged in efficiency, with SAC being the slowest. Future research could explore hybrid strategies combining entropy-regularization methods (Haarnoja et al., 2018) with truncation-based approaches (Kuznetsov et al., 2021) to enhance stability and performance. Expanding the scope of the investigation may also provide valuable insights into the generalizability of the methodologies.

References

1. Bester, C.J., James, S.D., Konidaris, G.D.: Multi-Pass Q-Networks for Deep Reinforcement Learning with Parameterised Action Spaces. University of the Witwatersrand, Johannesburg; Brown University, Providence RI (2021).

2. Fan, J., Wang, Z., Xie, Y., & Yang, Z. (2020, July). A theoretical analysis of deep Q-learning. In Learning for dynamics and control (pp. 486-489). PMLR.
3. Foo, M.J.R., Lesmana, N.S., Pun, C.S.: DRL Trading with CPT Actor and Truncated Quantile Critics. School of Computer Science and Engineering & Nanyang Business School, Nanyang Technological University, Singapore (2022).
4. Fujimoto, S., Hoof, H., & Meger, D. (2018, July). Addressing function approximation error in actor-critic methods. In International conference on machine learning (pp. 1587-1596). PMLR.
5. Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018, July). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. July 10th-15th, Stockholm, Sweden. Vol. 1870. 1861.
6. Konda, V., & Tsitsiklis, J. (1999). Actor-critic algorithms. Advances in neural information processing systems, 12.
7. Kuznetsov, A., Shvechikov, P., Grishin, A., Vetrov, D.: Controlling Overestimation Bias with Truncated Mixture of Continuous Distributional Quantile Critics. In: Proceedings of the 37th International Conference on Machine Learning (ICML), pp. 5500–5510 (2020).
8. Luo, Z., Zhao, L., Cheng, W., Chen, S., Chen, Q., Xue, H., & Zhang, L. (2021, April). Match plan generation in web search with parameterized action reinforcement learning. In Proceedings of the Web Conference 2021 (pp. 1040-1052).
9. Masson, W., Ranchod, P., Konidaris, G.: Reinforcement Learning with Parameterized Actions. In: School of Computer Science and Applied Mathematics, University of Witwatersrand, Johannesburg, South Africa; Department of Computer Science, Duke University, Durham, North Carolina (2016).
10. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *nature*, 518(7540), 529-533.
11. Mnih, V. (2016). Asynchronous Methods for Deep Reinforcement Learning. arXiv preprint arXiv:1602.01783.
12. Neumann, S., Lim, S., Joseph, A., Pan, Y., White, A., & White, M. (2023). Greedy actor-critic: A new conditional cross-entropy method for policy improvement. In proceedings of ICLR 2023.
13. Rosset, C., Jose, D., Ghosh, G., Mitra, B., Tiwary, S.: Optimizing Query Evaluations Using Reinforcement Learning for Web Search. Microsoft AI & Research (2020).
14. Rummery, G. A., & Niranjan, M. (1994). On-line Q-learning using connectionist systems (Vol. 37, p. 14). Cambridge, UK: University of Cambridge, Department of Engineering.
15. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.
16. Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., & Riedmiller, M. (2014, January). Deterministic policy gradient algorithms. In International conference on machine learning (pp. 387-395). Pmlr.
17. Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. *MIT Press*.
18. Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*, 8, 279-292.
19. Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8, 229-256.
20. Zobel, J., Moffat, A.: Inverted files for text search engines. *ACM Comput. Surv.* 38(2), 6 (2006). <https://doi.org/10.1145/1132956.1132959>