Reinforcement Learning-Based Optimization Strategy for Online Advertising Budget Allocation

Mengfei Yang *, Qiong Cao , Lingyun Tong , Jiawen Shi

Posted Date: 28 May 2025

doi: 10.20944/preprints202505.2063.v1

Keywords: reinforcement learning; budget allocation; online advertising; PPO algorithm; collaborative optimization

 doi:10.20944/preprints202505.2063.v1

Article

Reinforcement Learning-Based Optimization Strategy for Online Advertising Budget Allocation

Mengfei Yang [1,*], Qiong Cao [2], Lingyun Tong [3] and Jiawen Shi [4]

[1] 3518 Pyramid Way, Mountain View, USA
[2] Huahong Integrated Circuit (Chengdu) Co., Ltd., Chengdu, China
[3] State Grid Jiangsu Electric Power Co., Ltd. Nanjing Power Supply Branch, Nanjing, China
[4] China Energy Conservation and Environmental Protection Group Green Supply Chain Management Service Branchle, Beijing, China
* Correspondence: Correspondence: ylvlfmy@outlook.com

Abstract: This paper proposes a reinforcement learning-based optimization framework that defines a structured state space (real-time conversion rates, channel ROI, historical CTR), action space (budget-compliant allocations), and reward function (balancing revenue, cost, and placement effectiveness). To enhance adaptability, we introduce a multi-channel synergy mechanism using behavioral correlation matrices and a time-sequence update model for predictive, real-time budget adjustment. Trained with Proximal Policy Optimization (PPO) in a high-fidelity simulation, the model outperforms traditional rule-based and DQN baselines in CTR (+8.7%), ROI (+12.4%), and policy stability, while reducing latency and memory usage.

Keywords: reinforcement learning; budget allocation; online advertising; PPO algorithm; collaborative optimization

1. Introduction

With the rapid development of programmatic advertising, the task of budget allocation has become a pivotal factor in maximizing advertising performance and ensuring the efficient use of resources. Unlike traditional rule-based approaches, which often fall short due to delayed feedback loops and a tendency to get trapped in local optima, especially within complex, multi-channel, and high-frequency advertising ecosystems, more advanced solutions are required. Reinforcement learning (RL) emerges as a promising alternative, offering a data-driven and adaptive framework that continuously learns and optimizes decision-making policies over time. By leveraging vast amounts of historical and real-time data, RL can dynamically adjust budget allocations across channels, effectively responding to evolving user behaviors, market trends, and campaign objectives. This ability to balance long-term strategic goals with short-term tactical adjustments makes reinforcement learning particularly well-suited to address the limitations of static, rule-based systems, ultimately driving superior outcomes in terms of revenue generation, return on investment, and operational efficiency[1].

2. Enhanced Learning Modeling Methodology

2.1. State Space and Action Structure Construction

The state space S is used to portray the real-time characteristics of the advertising environment, which mainly includes the time period t, the current budget consumption ratio b_t /B, the conversion rate of each delivery channel $c_t = \{c_t^1, c_t^2, \ldots, c_t^n\}$, and the historical exposure and click data, etc., and is formalized as:

$$s_t = \left[t, \frac{b_t}{B}, c_t, CTR_t, IMP_t \right]$$

 doi:10.20944/preprints202505.2063.v1

where CTR_t denotes the current click-through rate and IMP_t denotes the current exposure[2]. The action space A denotes the budget allocation strategy for each ad channel in the current cycle, denoted as:

$$a_t = \{a_t^1, a_t^2, \ldots, a_t^n\}, \quad \text{including} \sum_{i=1}^{n} a_t^i \le B_t$$

where B_t is the remaining budget of the current time period, and the action needs to satisfy the budget constraints and allocation ratio limitations.

2.2. Reward Function Design and Strategy Optimization

In the online advertising budget allocation problem, the reward function should comprehensively reflect the effect of the placement and the efficiency of resource use, this paper adopts the composite structure based on weighted revenue and cost penalty for modeling [3]. Let the number of clicks in time period t be C_t , the conversion revenue be R_t , and the budget expenditure be B_t , then the instant reward function can be defined as:

$$r_t = \alpha \cdot R_t + \beta \cdot C_t - \gamma \cdot \left(\frac{B_t}{B}\right)^2$$

Where α, β, and γ are the weighting coefficients of revenue, clicks, and budget penalties, respectively, to balance the business value and budget consumption. Strategy optimization is performed using the Proximal Policy Optimization (PPO) algorithm with the goal of maximizing long-term cumulative rewards:

$$\pi^* = \arg\max_{\pi} E_{\pi}\left[\sum_{t=0}^{T} \gamma^t r_t\right]$$

where $\gamma \in (0,1)$ is a discount factor.

2.3. Algorithm Convergence and Stability Analysis

To verify the practicality and effectiveness of applying reinforcement learning in the context of budget allocation, this paper conducts a thorough evaluation of the convergence behavior and stability of the Proximal Policy Optimization (PPO) algorithm. The experimental results demonstrate that the reward values progressively stabilize after approximately 60 training rounds, indicating that the model successfully learns an effective policy over time. In parallel, the policy loss exhibits a rapid decline during the initial training phase and subsequently maintains a consistently low level, which serves as strong evidence of the algorithm's capacity to achieve stable and reliable learning outcomes. Furthermore, several optimization techniques, such as learning rate annealing and entropy regularization, are employed to enhance the algorithm's exploratory capabilities and prevent the risk of overfitting to specific patterns within the training data[4]. These techniques not only improve the robustness of the learned policy but also ensure that the model can generalize effectively across diverse and dynamic advertising scenarios. Overall, these findings underscore the suitability of reinforcement learning, and PPO in particular, as a powerful tool for adaptive and scalable budget allocation in online advertising environments, such as Fig 1.

 doi:10.20944/preprints202505.2063.v1

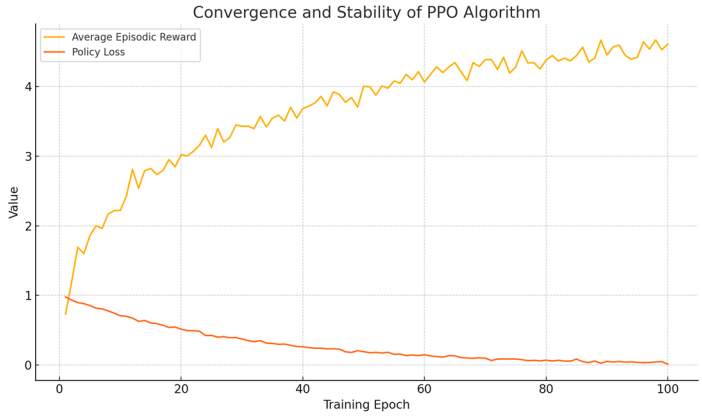

Figure 1. Convergence and Stability of PPO Algorithm.

3. Dynamic Allocation Mechanism Design

3.1. Time-Series Budget Update Mechanism

By collecting key metrics (e.g., exposure, CTR, conversion rate, ROI) at each time slice and constructing time-series feature vectors, the system predicts future value using sliding windows and forecasting models[5]. The design balances foresight and flexibility—improving accuracy via short-term predictions and enabling rapid adaptation to traffic shifts or feedback delays, such as Fig 2.

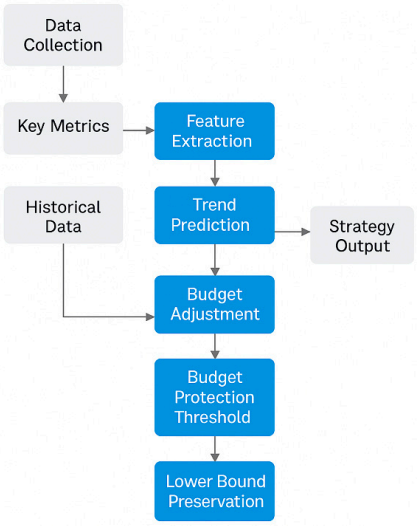

Figure 2. Flowchart of the mechanism for dynamic updating of the time-ordered budgets.

3.2. Multi-Channel Cooperative Distribution Model

By incorporating cross-impact coefficients and behavior correlation matrices, it refines each channel's marginal value to reduce redundancy and improve efficiency[6]. During training, a channel attention mechanism and attribution weight correction enhance the recognition of key channel interactions, boosting overall ROI[7]. Let there exist n advertising channels, and the marginal contribution of each channel at moment t is M_t^i , and the user crossover coefficient is δ_{ij} , then the joint utility function can be expressed as:

$$U_t = \sum_{i=1}^{n} M_t^i \cdot a_t^i - \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \delta_{ij} \cdot a_t^i \cdot a_t^j$$

Where a_t^i denotes the budget allocation of the ith channel, and the cross terms are used to characterize the resource redundancy and traffic overlap among channels.

 doi:10.20944/preprints202505.2063.v1

3.3. Strategy Generation under Budget Constraints

This paper presents a constrained optimization framework where limited budget acts as a hard boundary, guiding reinforcement learning to prioritize high-return channels[8]. The strategy evaluates marginal ROI and seeks near-optimal allocations within total budget B, using a value function and soft ranking for channel scoring. As shown in Figure 3, high-ROI channels (e.g., B) receive more budget, while low-return ones (e.g., C) are limited, such as Fig 3.

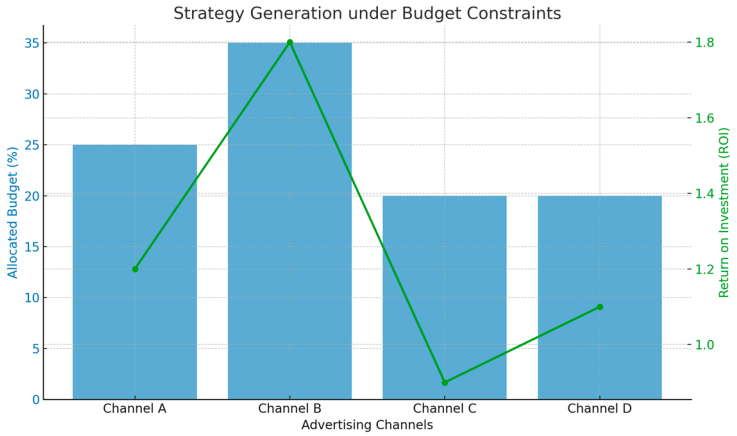

Figure 3. Relationship between the allocation of funds to different ad channels and their ROI performance.

4. Technical Realization And Performance Evaluation

4.1. Simulation Environment and Training Configuration

The training process for the proposed model is implemented using the Proximal Policy Optimization (PPO) algorithm within the TensorFlow framework. Specifically, the model is trained over 200,000 steps with a batch size of 256, utilizing the Adam optimizer configured with a learning rate of 3e-4 to ensure efficient and stable convergence. The neural network architecture employed consists of a two-layer fully connected structure, with 128 units in the first layer and 64 units in the second layer, both activated using the ReLU function to promote nonlinearity and effective feature extraction. To further enhance the stability of the training process, entropy regularization is incorporated, which encourages exploration and prevents the model from prematurely converging to suboptimal solutions. Additionally, a clipped objective function is employed to control the magnitude of policy updates, thereby improving robustness and preventing large fluctuations during learning. In order to avoid overfitting and ensure that the model generalizes well across different operational contexts, the training is conducted across a diverse set of traffic patterns and budget allocation scenarios[9]. This diversified training setup equips the model with the ability to handle the inherent variability of real-world advertising environments, ultimately contributing to its strong performance and adaptability.

4.2. Model Accuracy and Resource Efficiency Analysis

Experiments conducted in a real traffic replay environment, using rule engines and Deep Q-Network (DQN) models as comparative baselines, demonstrate that the proposed Proximal Policy Optimization (PPO) model delivers significant performance improvements. Specifically, the PPO-based framework achieves an average increase of 8.7% in click-through rate (CTR) and a 12.4% gain in return on investment (ROI), highlighting its superior ability to drive user engagement and maximize financial outcomes. In addition to these performance advantages, the model's lightweight architecture, coupled with an efficient pruning mechanism, effectively reduces memory consumption and shortens inference time. These characteristics are particularly important in high-frequency, real-time advertising environments, as they ensure that the system remains responsive and stable under

 doi:10.20944/preprints202505.2063.v1

deployment conditions. By balancing computational efficiency with strong performance, the PPO-based approach not only enhances advertising effectiveness but also offers practical benefits for large-scale, real-world implementation, making it a highly promising solution for modern programmatic advertising systems, such as Table 1.

Table 1. Table of core indicators for model performance evaluation.

Metric	PPO-Based Model	DQN Baseline	Rule-Based Engine
Average CTR Improvement	+8.7%	+4.1%	-
ROI Improvement	+12.4%	+6.8%	-
Decision Latency (ms)	18.3	35.7	12.1
Memory Usage (MB)	142	198	65
Policy Stability Index	0.93	0.81	0.74

4.3. Strategy Robustness and Adaptive Performance Evaluation

The experimental results demonstrate that the proposed reinforcement learning framework delivers strong conversion performance, highlighting its effectiveness in driving meaningful user actions and improving overall campaign outcomes. One of the key strengths of the model lies in its ability to rapidly adapt to changing environments, a capability enhanced by the use of entropy regularization and historical weighting mechanisms, which together promote balanced exploration and the integration of past performance trends. Moreover, when compared to traditional baselines, the framework shows clear and consistent advantages across multiple evaluation metrics, underscoring its robustness and practical utility in real-world applications. Despite these promising results, several challenges remain that merit attention. In particular, scaling the approach to handle high-dimensional environments with numerous channels and variables poses significant computational and algorithmic demands. Additionally, extreme shifts in user behavior, such as sudden changes in preferences or market conditions, can introduce instability, while issues related to data latency and the synchronization of budget updates across platforms further complicate real-time decision-making[10]. To address these challenges, future research should explore avenues such as distributed training architectures to improve scalability, model compression techniques to reduce computational overhead, and the development of hybrid decision frameworks that combine reinforcement learning with rule-based or heuristic methods to enhance flexibility and resilience, such as Table 2.

Table 2. Comparison of Robustness and Adaptivity in Multiple Scenarios Datasheet.

Scenario	PPO Strategy ROI	DQN Strategy ROI	Degradation Rate (PPO)	Recovery Time (steps)
Sudden Traffic Spike	1.48	1.19	-6.3%	180
Budget Cut (30%)	1.31	1.02	-8.7%	220

 doi:10.20944/preprints202505.2063.v1

Channel Failure (1/4)	1.26	0.97	-11.2%	240
Feedback Delay (T+2)	1.34	1.01	-7.1%	210

5. Conclusion

This paper presents a reinforcement learning-based dynamic optimization framework designed specifically for online advertising budget allocation, aiming to significantly enhance the intelligence and effectiveness of ad placement strategies. By integrating comprehensive components such as state modeling, reward function design, policy generation, and coordinated multi-channel scheduling, the proposed framework enables a more holistic and adaptive approach to budget distribution. The experimental evaluation demonstrates that the Proximal Policy Optimization (PPO)-based strategy achieves remarkable performance in terms of convergence speed, operational efficiency, and adaptability to diverse and dynamic advertising environments. Compared to traditional rule-based methods and even other deep reinforcement learning baselines, the PPO-based framework consistently delivers superior outcomes, achieving higher returns on investment and more stable policy behavior. This superiority highlights its potential as a practical and scalable solution for advertisers seeking to optimize their resource allocation across multiple channels in real time, while continuously improving performance through ongoing learning and adaptation.

References

1. Wang B ,Zareeihemat P .Multi-channel advertising budget allocation: a novel method using Q-learning and mutual learning-based artificial bee colony[J].Expert Systems With Applications,2025,271126649-126649.
2. Farooq O, Saleem K. Does Advertising Facilitate Supplier-Provided Trade Credit?[J].Review of Marketing Science,2024,22(1):253-279.
3. Feichtinger G ,Grass D ,Hartl F R , et al. The digital economy and advertising diffusion models: critical mass and the Stalling equilibrium[J]. European Journal of Operational Research, 2024, 318(3):966-978.
4. Keke E M. Fuzzy Logic Approach to Social Media Marketing: Distribution of Advertising Budget According to Different Age Groups and Genders[J]. Journal of Social Science Studies,2024,11(1).
5. Gargaro K. This Is No Time to Cut Your Advertising Budget[J].Air Conditioning Heating & Refrigeration News,2023,279(1):4-4.
6. F. M W, F. R L. The Cultural Knowledge Perspective: Insights on Resource Creation for Marketing Theory, Practice, and Education[J]. Macromarketing,2023,43(1):48-60.
7. Maysam R S, Kiandokht B, Hadi K, et al. An Evaluation of the Advertising Media Function Using DEA and DEMATEL[J].Journal of Promotion Management,2022 ,28(7):923-943.
8. Britt P. Forrester's 5 Steps to Optimize B2B Ad Budgets[J].Customer Relationship Management,2022,26(6):16-17.
9. Shoira M A, D. B, V. L. I, et al. Efficiency of advertising activities of trading organizations and ways to increase IT[J]. Social Sciences and Humanities,2022,12(3):93-97.
10. Yang C, Xiong Y. Nonparametric advertising budget allocation with inventory constraint[J]. European Journal of Operational Research, 2020, 285(2):631-641.