Evaluating and Selecting Deep Reinforcement Learning Models for Optimal Dynamic Pricing: A Systematic Comparison of PPO, DDPG, and SAC

YUCHEN LIU, Xi'an Jiaotong-Liverpool University, China
KA LOK MAN, Xi'an Jiaotong-Liverpool University, China
GANGMIN LI, University of Bedfordshire, United Kingdom
TERRY R. PAYNE, University of Liverpool, United Kingdom
YONG YUE, Xi'an Jiaotong-Liverpool University, China

Given the plethora of available solutions, choosing an appropriate Deep Reinforcement Learning (DRL) model for dynamic pricing poses a significant challenge for practitioners. While many DRL solutions claim superior performance, there lacks a standardized framework for their evaluation. Addressing this gap, we introduce a novel framework and a set of metrics to select and assess DRL models systematically. To validate the utility of our framework, we critically compared three representative DRL models, emphasizing their performance in dynamic pricing tasks. Further ensuring the robustness of our assessment, we benchmarked these models against a well-established human agent policy. The DRL model that emerged as the most effective was rigorously tested on an Amazon dataset, demonstrating a notable performance boost of 5.64%. Our findings underscore the value of our proposed metrics and framework in guiding practitioners towards the most suitable DRL solution for dynamic pricing.

CCS Concepts: • **Information systems** → **Online analytical processing**; *Process control systems*; Association rules.

Additional Key Words and Phrases: Dynamic Pricing, Deep Reinforcement Learning (DRL), Inventory Management, Price Elasticity of Demand, Markov Decision Process, PPO (Proximal Policy Optimization), DDPG (Deep Deterministic Policy Gradient), SAC (Soft Actor-Critic), Model Evaluation, E-commerce

ACM Reference Format:
Yuchen Liu, Ka Lok Man, Gangmin Li, Terry R. Payne, and Yong Yue. 2023. Evaluating and Selecting Deep Reinforcement Learning Models for Optimal Dynamic Pricing: A Systematic Comparison of PPO, DDPG, and SAC. 1, 1 (October 2023), 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Dynamic pricing, a practice of adjusting prices in real-time in response to market conditions, has burgeoned in the digital age, particularly on the Internet. The vast spectrum of online pricing strategies includes traditional methods like direct price modifications and innovative approaches such as Bundle Pricing, Auction, First Come-First Served, Price

Authors' addresses: Yuchen Liu, Yuchen.Liu21@student.xjtlu.edu.cn, Xi'an Jiaotong-Liverpool University, 111 Ren'ai Road, Suzhou, Jiangsu, China, 215123; Ka Lok Man, Xi'an Jiaotong-Liverpool University, 111 Ren'ai Road, Suzhou, China, Ka.Man@xjtlu.edu.cn; Gangmin Li, University of Bedfordshire, Luton, United Kingdom, Gangmin.Li@beds.ac.uk; Terry R. Payne, University of Liverpool, Brownlow Hill, Liverpool, United Kingdom, T.R.Payne@liverpool.ac.uk; Yong Yue, Xi'an Jiaotong-Liverpool University, 111 Ren'ai Road, Suzhou, China, Yong.Yue@xjtlu.edu.cn.

Discrimination, and Cashback [1]. These strategies, albeit diverse, often operate on underlying algorithms governing price adjustments. However, the current landscape faces two significant challenges:

(1) The prevailing heuristic-based pricing mechanisms exhibit substantial lag, impeding real-time adjustments crucial for optimizing revenue and customer satisfaction.
(2) The price adjustments post-dynamic adaptation often fall short of achieving optimal performance, leaving much to be desired in harnessing the full potential of dynamic pricing.

Given these challenges, we propose a robust dynamic pricing mechanism grounded in reinforcement learning to address the latency and optimality issues, paving the way for a more nuanced and effective pricing strategy in the online marketplace. This proposal explores the possibility of leveraging advanced machine learning techniques to refine dynamic pricing models, ensuring they evolve adeptly with the fast-paced e-commerce ecosystem.

Building upon the identified challenges, selecting a suitable algorithm, particularly one employing Deep Reinforcement Learning (DRL), becomes imperative for advancing dynamic pricing. Given the number of DRL models that claim optimal performance, it can be difficult to ascertain which models are most suitable for various contexts. This paper addresses this problem by presenting a novel framework coupled with metrics for evaluating and selecting DRL models tailored explicitly for dynamic pricing applications. To achieve this, we scrutinize three eminent DRL models: Proximal Policy Optimization (PPO), Deep Deterministic Policy Gradients (DDPG), and Soft Actor-Critic (SAC). These models are fundamentally anchored in the Actor-Critic paradigm, a gradient algorithm within deep reinforcement learning. The Actor-Critic architecture bifurcates into two integral components: the actor, tasked with ascertaining actions based on the prevailing states, and the critic, assigned to evaluate and furnish feedback on the actor's decisions. By harnessing this dual structure, the actor's parameters are refined utilizing the critic's gradient on the action.

Throughout this investigation, we aim to clarify these models' optimal policy pathways in adjusting prices. More significantly, we propose a methodical approach for practitioners to discern the most suitable DRL model for their dynamic pricing endeavours. Upon rigorous comparison, the gleaned reinforcement learning strategies exhibit promising applicability in the dynamic pricing domain of e-commerce platforms and extend to other price-sensitive realms, such as ride-hailing app pricing, manifesting the proposed framework's expansive utility and underscoring our study's future applications.

2 RELATED WORK

The domain of dynamic pricing, integral to modern commerce, has garnered notable scholarly attention, especially in the context of inventory management. The studies by Elmaghraby et al. [2] and Netessine [3] delve into price adjustments concerning inventory levels, albeit occasionally overlooking the nuanced dynamics of customer demand. This contrasts with the work of Gallego and van Ryzin [4] that introduced a stochastic, periodic-review model, shedding light on the intersection of pricing and inventory control. Concurrently, other studies ventured into demand learning to mitigate uncertainties in price-sensitive demand [5, 6]. The foray of Reinforcement Learning (RL) into pricing has generated interest, with [7] exploring its application in revenue management. [8] and [9] contributed innovative perspectives; however, a comprehensive methodology transcending domain-specific constraints remains elusive.

In this study, we aim to address the problem by assimilating insights from microeconomic principles, emphasising the price elasticity of demand, thereby resulting in a refined simulated environment. Within this work, we evaluate various Deep Reinforcement Learning (DRL) models, with our thorough analysis culminating in the identification of a model that excels in performance and surpasses traditional benchmarks.

3 THREE REINFORCEMENT LEARNING MODELS

3.1 PPO

The traditional Actor-Critic paradigm, while groundbreaking, exhibits fragility in updating parameters via the policy gradient, especially with complex neural networks. This vulnerability is due to significant step lengths, which may deviate from the intended trajectory. To remedy this, the Trust Region Policy Optimization (TRPO) method was introduced [10], although its inclusion introduces further complexity issues. Proximal Policy Optimization (PPO) emerges as a more efficient alternative, moderating the policy update's ambit to bolster stability and accuracy [11]. TRPO and PPO are categorized under stochastic on-policy algorithms, known for their lower sample efficiency.

3.2 DDPG

Despite their merits, on-policy algorithms grapple with the need for abundant samples. Off-policy strategies like Deep Deterministic Policy Gradient (DDPG) address this problem by archiving samples in a replay buffer [12]. Under the Actor-Critic framework, DDPG uniquely employs four neural networks to mitigate overestimation errors: a training and a target network for both the Actor and the Critic.

3.3 SAC

Soft Actor-Critic (SAC), a pinnacle of off-policy algorithms, is deeply rooted in the Actor-Critic philosophy but adds a twist by interweaving entropy into the policy, thereby enriching action diversity [13]. SAC leverages objective functions to balance policy entropy and value amplification.

PPO, DDPG, and SAC are stalwarts in reinforcement learning. While PPO streamlines computations, DDPG strives for accuracy, and SAC aims to balance exploration and exploitation. The suitability of these models is contingent on the specific task and available resources. Their adaptability and versatility continue to be refined, promising even more refined solutions.

4 MARKOV DECISION PROCESS

To rigorously evaluate the proficiency of the three DRL models (PPO, DDPG, SAC) in dynamic pricing, it's imperative to establish a consistent, systematic framework. The Markov Decision Process (MDP) provides a suitable foundation, exploiting the synergies between states, actions, and rewards, which are pivotal in understanding and modelling dynamic pricing scenarios.

(1) **State Set**: States are delineated by the prevailing commodity price and sales volume. Our three models' randomly chosen initial price and sales volume are reference points, ensuring a consistent starting ground.
(2) **Action Set**: Actions pertain to dynamic pricing decisions and span a continuous space within the range $[-1, 1]$. This range signifies that the maximum and minimum actions and correspond to price increments and decrements of 1 unit, respectively.
(3) **Reward Set**: Rewards encapsulate the net profit accrued from each state-action pair. Not only is the net profit per transaction emphasized, but also the sales volume, combining demand elasticity while assessing the volume. The corresponding computation is detailed later.
(4) **Transition Set**: Given the present sales price, volume, and the chosen pricing action, coupled with price elasticity dynamics, we discern the subsequent state for both sales price and volume.

Drawing from microeconomic principles, the price elasticity is articulated as follows:

$$\xi = \frac{(q^{t+1} - q^t)/q^t}{a^t/p^t} \tag{1}$$

Where:

- ξ denotes the product's elasticity, which is assumed to be a constant value.
- q^{t+1} denotes the sales volume at time t+1.
- q^t represents the sales volume at time t.
- a^t symbolises the price change between time t and t+1.
- p^t signifies the price at time t.

Given these parameters, the sales volume at time t+1 can be described by:

$$q^{t+1} = \frac{a^t \times q^t \times \xi}{p^t} + q^t \tag{2}$$

Subsequently, the reward at time t can be represented as follows, where c stands for the fixed cost:

$$r^t = (p^{t+1} - c) \times q^{t+1} = (p^{t+1} - c) \times \left(\frac{a^t \times q^t \times \xi}{p^t} + q^t\right) \tag{3}$$

The overarching aim here is to amplify the aggregate net profit over a predetermined period. PPO, DDPG, and SAC models are individually trained with a uniform random initial price as a reference. Their cumulative returns and temporal efficiency are subsequently evaluated and juxtaposed to discern the most appropriate model for dynamic pricing.

5 EXPERIMENTS WITH THREE MODELS

5.1 Hyperparameters Setting

The different categories' elasticity is hugely diverse, so we set the ξ at different values 0.5, 2, 5. Suppose the price elasticity of demand is greater than 1, such as for fiercely competitive products and daily necessities. In such cases, it leads to a change in price, yielding a significant shift in demand—the more considerable, the greater the difference. Suppose the price elasticity of demand is less than 1, such as with luxury goods, which leads to an insignificant change in demand. Then, we observe how different elasticity affects the result. The cost of the product assumes a fixed number in this experiment case. Those three models are updated from the Actor-Critics framework, and we can keep the main 3-layer neural network structure in the Actor-Critics part. The hidden layer keeps the 128 units all the time.

In the context of our experiments, we have developed and implemented several methods to improve the training of the reinforcement learning models. In the PPO experiment, we propose the PPO-CLIP method to constrain the objective function and maintain the scale, which involves setting ϵ to 0.2 to restrict the range of the function. For the DDPG and SAC experiments, we have utilized the off-policy algorithm for training the models, which requires specifying the reply buffer size in advance. Additionally, due to the characteristics of the Target Q network, we have employed the soft update method to keep it stable during training, which involves setting τ to 0.005. Moreover, in the SAC experiment, introducing entropy adds a layer of complexity, and we need to consider the regularization of the entropy term to improve the model's performance. Therefore, the regularization of the entropy term has also been included in our experimentation process.

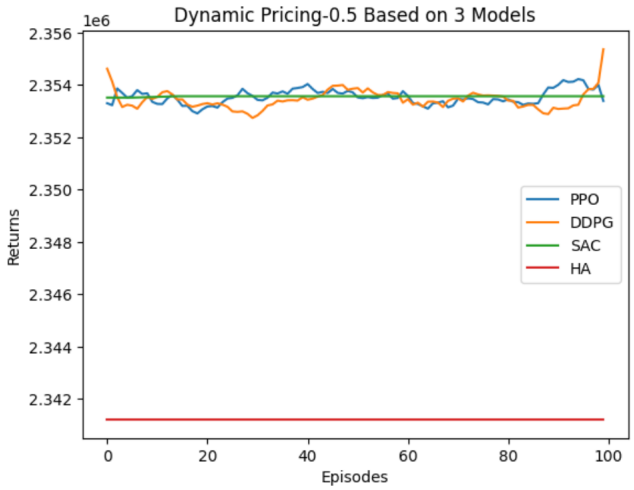

Fig. 1. Sales Reward by $\xi = 0.5$

The proposed methods have been designed to address specific challenges while training reinforcement learning models. The methods' effectiveness has been evaluated through extensive experimentation, and the results demonstrate their ability to improve the performance and stability of the models during training.

5.2 Pre-train the Human Agent Policy

The human-agent policy employed in this study is derived from the Micro-economics Nash Equilibrium price, which has been widely used as a benchmark for evaluating the performance of various pricing models. As such, we have utilized this benchmark as a reference point for comparing and assessing the effectiveness of the other three pricing models under investigation.

5.3 Train the Model Separately

5.3.1 $\xi = 0.5$. Following the configuration of all hyper-parameters, the model was trained, and the results are presented in Figure 1. In this case, elasticity values were equal to 0.5, indicating that price changes are expected to have a moderate impact on changes in quantity. Among the three models investigated, each can achieve a relatively good return compared to the benchmark human agent performance (HA). Furthermore, all models demonstrated convergence during the training process. However, there are notable differences in the stability of the models during training. The SAC model was more stable and performed well in the initial episodes, whereas the PPO and DDPG models exhibited oscillatory behaviour.

Our analysis suggests that retailers may choose the latter models (PPO and DDPG) if they intend to adopt an aggressive exploration strategy during the lower elasticity pricing process. Overall, the findings provide valuable insights into the performance of different models in the context of pricing strategies and can inform the development of more effective pricing models.

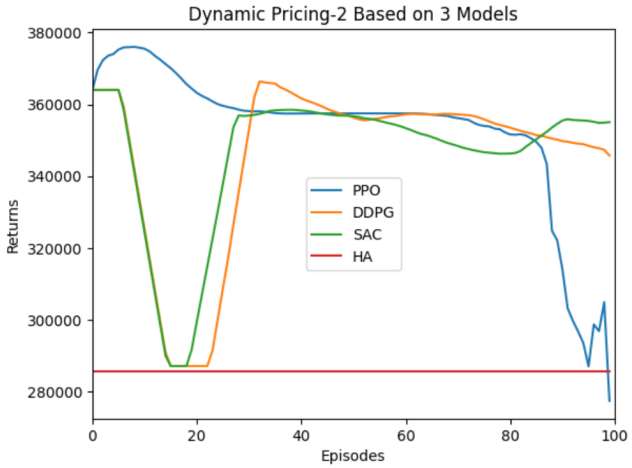

Fig. 2. Sales Reward by $\xi = 2$

5.3.2 $\xi = 2$. Figure 2 presents the elasticity values, which have been calculated to equal 2. This indicates that price changes are likely to have a significant impact on changes in quantity. Interestingly, the three models under evaluation exhibit notable differences in their performance. However, it is worth noting that all three models consistently outperform human agent decision-making in most scenarios. Specifically, the PPO model initially demonstrates superior performance, but its performance deteriorates as the training progresses. On the other hand, the DDPG and SAC models initially yield lower returns, but by fine-tuning their parameters, their returns gradually improve. Eventually, these two models converge to a similar return trajectory, making it difficult to determine which model is superior.

Compared to a scenario with elasticity equal to 0.5, the present case exhibits significant fluctuations in performance. The observed oscillations indicate the models' sensitivity to changes in elasticity values, highlighting the importance of fine-tuning and understanding the impact of different parameters in the pricing models.

5.3.3 $\xi = 5$. Figure 3 depicts the elasticity values, set to 5, whereby any price change would significantly impact the quantity demanded. Compared to the human agent's decisions, the three models do not perform well, owing to the pronounced demand change resulting from price alterations. While the human agent adopted a two-step strategy, the other models incorporated more exploration into their algorithms, compromising their returns. Among the three models, PPO performed the poorest in all episodes. On the other hand, SAC and DDPG had a relatively similar training trajectory, with SAC slightly outperforming DDPG most of the time.

We evaluated the efficacy of three reinforcement learning models, PPO, DDPG, and SAC, against a benchmark human agent model across 100 training episodes. The performance of these models was analyzed at different levels of price elasticity ($\xi = 0.5, 2, 5$) as depicted in Figures 1, 2, and 3, respectively. The data presented in Table 1 were derived from these figures, summarizing the performance metrics of each model at various price-elasticity levels.

As seen in Table 1, PPO exhibited suboptimal performance across all price elasticity levels, with a significant performance dip to -71.12% at a price elasticity of 5. Conversely, DDPG showed improved performance, notably

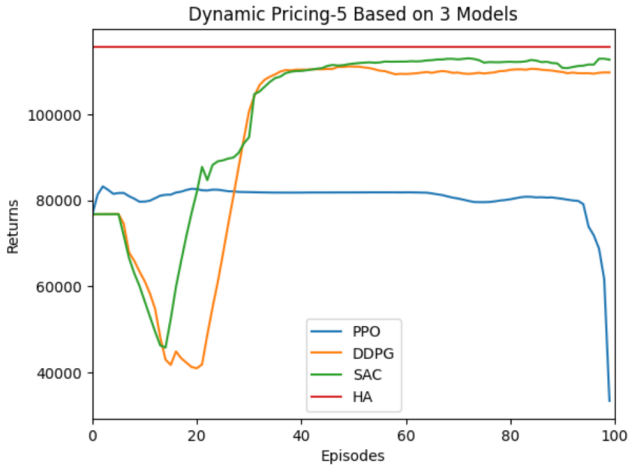

Fig. 3. Sales Reward by $\xi = 5$

Table 1. Three Reinforcement Models' Performance

Price Elasticity	Models Compared with Human Agent		
	PPO	***DDPG***	***SAC***
0.5	0.52%	**0.61**%	0.53%
2	-2.90%	21.03%	**24.17**%
5	-71.12%	-5.20%	**-2.58**%

See Original Codes : [14]

achieving 21.03% at a price elasticity of 2, although its performance dwindled to -5.20% at a price elasticity of 5. Among the models, SAC demonstrated the most stable and superior performance across all evaluated price elasticity levels, particularly excelling at a price elasticity of 2 with 24.17%.

Subsequently, using the SAC model, we tested its applicability on an empirical dataset as mentioned in [15]. Instead of relying on a random initial price, we employed transfer learning to determine the starting price. With this approach, we observed a notable performance improvement, marking an increase of approximately 5.64%.

6 CONCLUSION

This study evaluated three prominent DRL models for dynamic pricing: PPO, DDPG, and SAC. While PPO faced stability challenges, DDPG and SAC demonstrated robust performance, notably in moderate-elasticity scenarios. The real-world application of the SAC model further affirmed its practical viability, marking a significant performance improvement, especially in simplistic e-commerce pricing scenarios.

Further research can enhance accuracy by integrating a more comprehensive range of product attributes and employing a deep-learning dynamic model for sales prediction. Additionally, exploring multi-agent dynamic game scenarios could provide further insights. This study lays a solid groundwork for practitioners, emphasizing the potential of DRL in dynamic pricing, and beckons further exploration to fully harness this potential in more complex, real-world scenarios.

ACKNOWLEDGMENT

This work is partially supported by the XJTLU AI University Research Centre and Jiangsu Province Engineering Research Centre of Data Science and Cognitive Computation at XJTLU. Also, it is partially funded by the Suzhou Municipal Key Laboratory for Intelligent Virtual Engineering (SZS2022004) as well as funding: XJTLU-REF-21-01-002 and XJTLU Key Program Special Fund (KSF-A-17).

REFERENCES

[1] Yuchen Liu, Ka Lok Man, Gangmin Li, Terry R. Payne, and Yong Yue. Dynamic pricing strategies on the internet. In Proceedings of International Conference on Digital Contents: AICo (AI, IoT and Contents) Technology, 2022.

[2] Wedad Elmaghraby and Pınar Keskinocak. Dynamic pricing in the presence of inventory considerations: Research overview, current practices, and future directions. *Management science*, 49(10):1287–1309, 2003.

[3] Serguei Netessine. Dynamic pricing of inventory/capacity with infrequent price changes. *European Journal of Operational Research*, 174(1):553–580, 2006.

[4] Guillermo Gallego and Garrett Van Ryzin. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management science*, 40(8):999–1020, 1994.

[5] Byung Do Chung, Jiahan Li, Tao Yao, Changhyun Kwon, and Terry L Friesz. Demand learning and dynamic pricing under competition in a state-space framework. *IEEE Transactions on Engineering Management*, 59(2):240–249, 2011.

[6] Rajan Gupta and Chaitanya Pathak. A machine learning framework for predicting purchase by online customers based on dynamic pricing. *Procedia Computer Science*, 36:599–605, 2014.

[7] Nicolas Bondoux, Anh Quan Nguyen, Thomas Fiig, and Rodrigo Acuna-Agost. Reinforcement learning applied to airline revenue management. *Journal of Revenue and Pricing Management*, 19(5):332–348, 2020.

[8] Jiaxi Liu, Yidong Zhang, Xiaoqing Wang, Yuming Deng, and Xingyu Wu. Dynamic pricing on e-commerce platform with deep reinforcement learning: A field experiment. *arXiv preprint arXiv:1912.02572*, 2019.

[9] Rainer Schlosser and Martin Boissier. Dynamic pricing under competition on online marketplaces: A data-driven approach. In *KDD*, pages 705–714, 2018.

[10] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

[11] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[12] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International conference on machine learning*, pages 387–395. Pmlr, 2014.

[13] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.

[14] Yuchen Liu. Dynamic pricing algorithm. https://github.com/Larry-Liu02/Dynamic-Pricing-Algorithm, 2023.

[15] Yuchen Liu, Ka Lok Man, Gangmin Li, Terry R. Payne, and Yong Yue. Enhancing sparse data performance in e-commerce dynamic pricing with reinforcement learning and pre-trained learning. In *2023 International Conference on Platform Technology and Service (PlatCon)*. IEEE, 2023.