©2025 Published in 11th International Symposium on Intelligent Technologies in Engineering and Science 16-18 May 2025 (ISITES2025 Diyarbakır - Türkiye)

Hybrid Importance Sampling for Efficient and Stable Off-Policy Reinforcement Learning

[1]Ahmet KALA, [2,3]Cem ÖZKURT, and [4]Özkan CANAY
*[1]Department of Information Technologies, Sakarya University of Applied Sciences, Türkiye
[2]Department of Computer Engineering, Sakarya University of Applied Sciences, Türkiye
[3]AI and Data Science Research and Application Center, Sakarya University of Applied Sciences, Türkiye
[4]Faculty of Computer and Information Sciences, Department of Information Systems and Technologies Sakarya University, Türkiye

Abstract

Off-policy reinforcement learning improves sample efficiency by reusing data from a behavior policy different from the target policy. This benefits costly domains like robotics and healthcare, enhancing generalization and rare-event learning. However, high variance and instability remain challenges, risking local optima convergence. This study introduces Hybrid Importance Sampling (HIS), combining adaptive clipping and dynamic normalization to stabilize importance weight estimation. Adaptive clipping limits extreme weights, reducing variance, while dynamic normalization balances weight distribution. Compared to standard methods such as Ordinary Importance Sampling (OIS), Weighted Importance Sampling (WIS), and Per-Decision Importance Sampling (PDIS), HIS shows superior stability in high-variance settings, making off-policy learning more reliable for real-world applications.

Key words: Off-Policy Reinforcement Learning, Importance Sampling, Monte Carlo Methods.

1. Introduction

In reinforcement learning, off-policy methods enable agents to learn from data generated by a behavior policy different from the current target policy [1, 2]. Unlike traditional on-policy approaches, off-policy algorithms improve sample efficiency by reusing past experiences or simulation data [3, 4]. This method is particularly advantageous in real-world applications where data collection is costly.

The key benefit of off-policy learning is its ability to enhance data diversity. By learning from experiences generated by different policies, agents develop more generalizable strategies [5, 6]. This strengthens policy adaptation in complex environments. Additionally, experience replay mechanisms stabilize learning by repeatedly incorporating rare events into training [7,8].

However, off-policy methods have significant drawbacks. The primary issue is learning instability due to high variance [9, 10]. Distributional mismatch between behavior and target policies can lead to Q-value overestimation or incorrect convergence [11, 12]. Another challenge is premature convergence, where agents may get trapped in local optima, losing exploration capability [13, 14].

Importance sampling techniques address these issues by reweighting behavior policy data to match the target distribution [15, 16]. For instance, Q-Prop combines importance sampling with actor-

*Corresponding author: Address: Department of Information Technologies, Sakarya University of Applied Sciences, 54187, Sakarya TURKİYE. E-mail address: ahmetkala@subu.edu.tr, Phone: +0264 616 0054

critic architectures for low-variance, fast learning [15]. Similarly, emphatic weightings balance policy updates by prioritizing critical states [9, 13].

This study tackles two core challenges in off-policy RL: high variance from uncontrolled importance weight growth and slow convergence. We propose Hybrid Importance Sampling (HIS), integrating:

1. Adaptive Clipping: Controls weight explosion by thresholding extreme values

2. Dynamic Normalization: Maintains stable weight distributions through adaptive scaling

By synergizing these components, HIS significantly reduces variance, enhances policy evaluation reliability, and preserves environmental adaptability. The method is particularly promising for real-world applications like robotics, autonomous vehicles, and medical decision systems where data collection is costly or risky.

2. Materials and Method

This study compares the policy evaluation performance of different importance sampling techniques within a Reinforcement Learning (RL) problem framework based on Markov Decision Processes (MDPs). The methodology consists of four main components: (1) Problem and simulation dynamics, (2) proposed hybrid importance sampling models, (3) baseline methods for comparison, and (4) performance evaluation metrics.

2.1. Problem and Simulation Dynamics

The visual representation in Figure 1 illustrates the action selection and reward acquisition process in each episode. This process was repeated for 2,000 episodes to compute the average reward. The study was conducted over 10 independent runs, each tested with a different random initialization. The available actions are categorized as "up" and "down." The target policy always selects the "up" action, while the behavior policy chooses "up" with 75% probability and "down" with 25% probability.

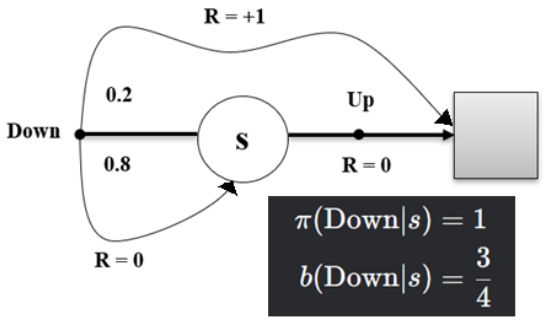

Figure 1. Transition dynamics and action probabilities under the target and behavior policies.

When considering the transition probabilities for the 'up' action, there is an 80% chance of moving correctly and receiving a reward of 1. Conversely, there is a 20% chance of incorrect movement, which results in no reward.

2.2. Hybrid Importance Sampling Model

To mitigate instability caused by large importance weights, we developed a framework that dynamically adjusts weight scaling and clipping based on recent observations. The core models are:

- Adaptive Clipping (HybridClip): This method applies soft thresholding to importance weights, preventing extreme values. Clipping bounds are adjusted using a moving window of recent weights, ensuring stability without introducing excessive bias.

- Dynamic Normalization (HybridNorm): Instead of fixed normalization, this approach standardizes weights based on their recent mean and variance, reducing variance while minimizing bias.

- Full Hybrid Model: Combines both adaptive clipping and dynamic normalization to balance the bias-variance trade-off.

2.3. Baseline Importance Sampling Methods (For Comparison)

To evaluate the effectiveness of the proposed hybrid models, we compared them against three standard techniques:

- Ordinary Importance Sampling (OIS): A basic variance reduction technique in Monte Carlo methods, where expected values under the target distribution are estimated by weighting samples drawn from a proposal distribution [17].

- Weighted Importance Sampling (WIS): Addresses high variance in classical IS by normalizing raw importance weights across all samples, yielding more stable and lower-variance estimates [18].

- Per-Decision Importance Sampling (PDIS): An adaptation of IS for sequential decision-making, where importance ratios are computed per decision step to stabilize return estimation in long-horizon tasks [19]. Unlike traditional IS, which multiplies weights over entire trajectories (leading to exploding variance), PDIS mitigates cumulative variance by step-wise reweighting.

2.4. Performance Evaluation Metrics

To comprehensively assess the effectiveness and robustness of the evaluated methods, multiple quantitative and statistical criteria were employed. Performance was analyzed through dynamic trend visualizations and distribution-based metrics, ensuring a holistic comparison. First, running

mean of value estimates (window size = 500) was used to track the smoothed convergence behavior of each algorithm over time. Complementing this, the running variance of value estimates highlighted stability and consistency in learning, while the running mean squared error (MSE) (window size = 500) quantified deviations from ground truth values. Additionally, running average rewards provided insight into the practical efficacy of each method in maximizing returns. To examine final performance, the distribution of final estimates (last 500 episodes) was visualized via box plots, revealing bias and dispersion, whereas the distribution of value estimates (density plot) illustrated overall estimation patterns. Statistical validation was further conducted through t-tests and effect size analysis to determine significant differences between methods, supported by a summary statistics table reporting mean, variance, MSE, and relative efficiency (1/Var). Together, these metrics ensured a rigorous, multi-faceted evaluation of algorithmic performance.

3. Results and Discussion

In this study, the performance of the proposed Hybrid Importance Sampling (HIS) method was compared with standard techniques including Ordinary Importance Sampling (OIS), Weighted Importance Sampling (WIS), and Per-Decision Importance Sampling (PDIS).

The rolling average graph in Figure 2 (Window Size = 500) shows that HIS exhibited lower variance in mean estimates and demonstrated consistent convergence behavior, particularly between episodes 800-2000. For instance, while HIS maintained stable estimates in the ~25-35 range, OIS showed high fluctuations, and WIS and PDIS displayed more pronounced oscillations.

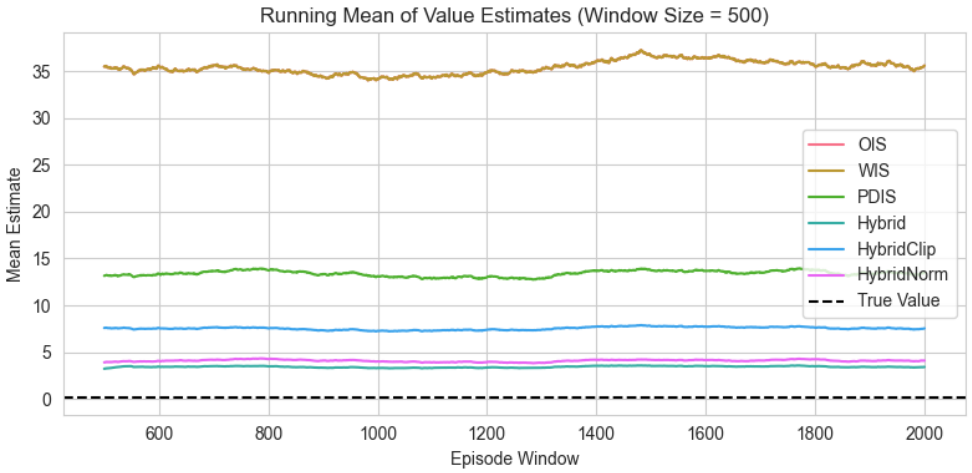

Figure 2. Mean of value estimates

The variance analysis presented in Figure 3 reveals that hybrid methods provided significantly lower and more stable variance values compared to standard techniques. While OIS showed the highest variance values, WIS and PDIS performed better than OIS. However, the hybrid methods, particularly through the combination of dynamic normalization and adaptive clipping techniques, most effectively reduced variance. Furthermore, as the number of episodes increased, the variance

in hybrid methods decreased more consistently.

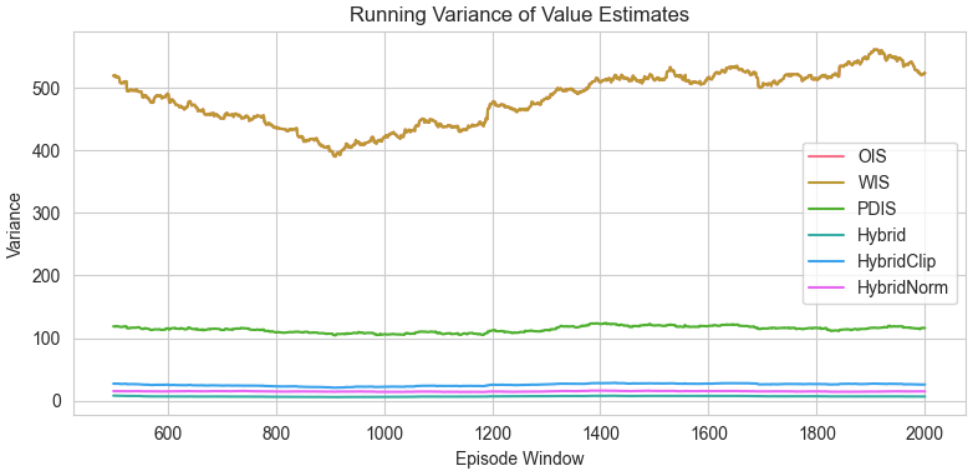

Figure 3. The variance of value estimates

The Moving Average Mean Squared Error (MSE) analyses in Figure 4 confirm that hybrid methods provide a clear performance advantage over standard techniques. The Hybrid method showed the best performance with the lowest MSE values, while OIS performed worst, and WIS and PDIS showed intermediate performance. These findings demonstrate that hybrid methods both improve estimation accuracy and stabilize the learning process.

Figure 4. Mean squared error (MSE)

The average reward graph in Figure 5 shows that the hybrid model produced lower reward values. This can be explained by HIS clipping overly optimistic estimates to reduce variance and learning more conservative strategies. Although the hybrid model sacrificed some reward for stability, its low MSE and variance values prove that its estimates are reliable.

Figure 5. The average rewards.

Figure 6 displays the distribution of value estimates produced by different importance sampling methods in the last 500 episodes. Hybrid methods (Hybrid, HybridClip, HybridNorm) generated much narrower and more concentrated estimates compared to standard methods (OIS, WIS, PDIS). While OIS and WIS showed a wide spread between 0 and 120, the Hybrid method concentrated in the 20-40 range, providing more consistent and stable estimates. This confirms that hybrid methods successfully reduced variance and controlled outliers. PDIS performed better than OIS and WIS but still lagged behind hybrid methods.

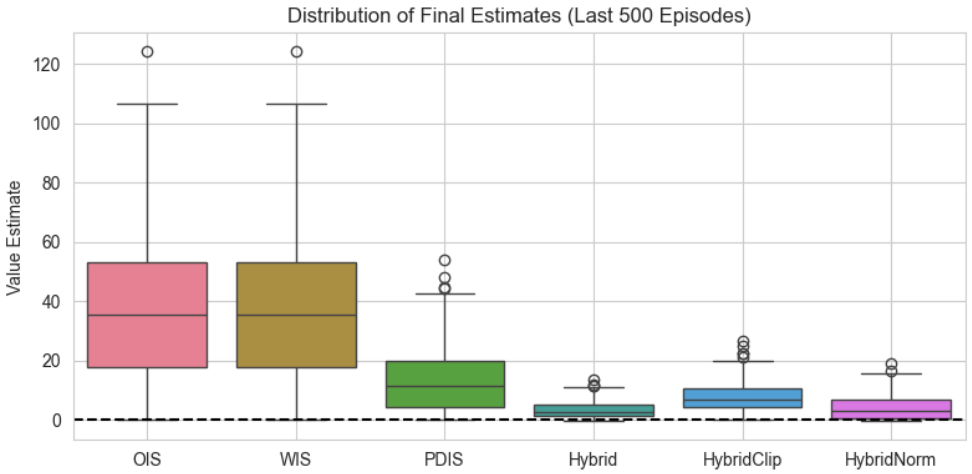

Figure 6. The distribution of final estimates

Figure 6 compares the probability density functions of the estimates. Hybrid methods (Hybrid, HybridClip, HybridNorm) showed a sharp, symmetric distribution close to the true value (0.2), while OIS and WIS had flat, wide distributions. This indicates that OIS and WIS produced high-variance, noisy estimates. HybridNorm achieved the highest density in the 0.1-0.3 range, showing the most stable performance, while HybridClip and Hybrid also concentrated in a narrow band.

PDIS showed lower density than hybrid methods but performed better than OIS and WIS. This graph supports the superiority of hybrid methods in both accuracy and consistency.

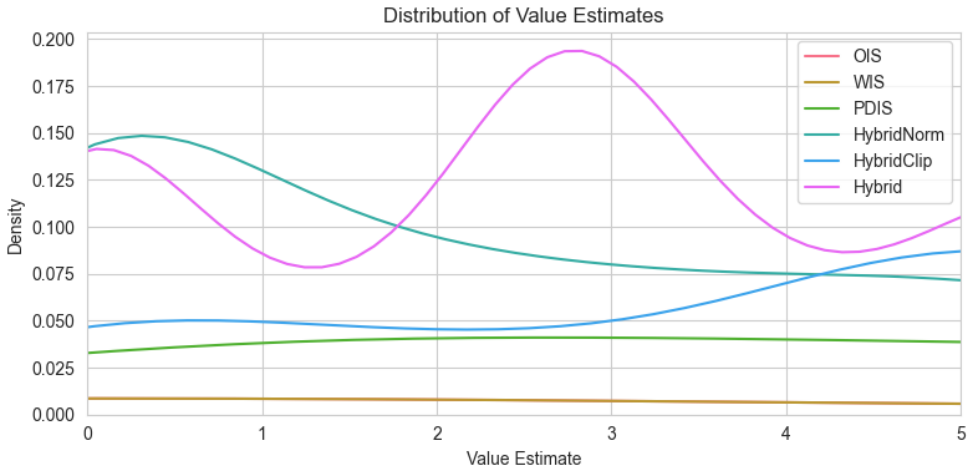

Figure 7. Markov decision processes (MDPs)

The statistical test results in Table 1 demonstrate that hybrid methods show statistically significant differences compared to standard techniques. The Hybrid method in particular has the highest Cohen's d values when compared to OIS, WIS, and PDIS, indicating large effect sizes. Additionally, all comparisons show p-values below the significance threshold (0.05), supporting the reliability of the results.

Table 1. The statistical test results.

Id	Comparison	t-statistic	p-value	Cohen's d	Significant
0	OIS vs WIS	0.000000	1.000000e+00	0.000000	False
1	OIS vs PDIS	40.000379	1.361312e-294	1.265239	True
2	OIS vs Hybrid	64.032996	0.000000e+00	2.025408	True
3	OIS vs HybridClip	54.714520	0.000000e+00	1.730658	True
4	OIS vs HybridNorm	62.197067	0.000000e+00	1.967336	True
5	WIS vs PDIS	40.000379	1.361312e-294	1.265239	True
6	WIS vs Hybrid	64.032996	0.000000e+00	2.025408	True
7	WIS vs HybridClip	54.714520	0.000000e+00	1.730658	True
8	WIS vs HybridNorm	62.197067	0.000000e+00	1.967336	True
9	PDIS vs Hybrid	40.304937	2.222933e-298	1.274873	True
10	PDIS vs HybridClip	21.893717	1.909373e-100	0.692513	True
11	PDIS vs HybridNorm	36.425168	4.045903e-251	1.152153	True
12	Hybrid vs HybridClip	-32.323057	7.740366e-204	-1.022400	True
13	Hybrid vs HybridNorm	-6.384568	1.915271e-10	-0.201948	True
14	HybridClip vs HybridNorm	24.341063	3.739859e-122	0.769925	True

The performance metrics summary in Table 2 confirms that hybrid methods outperform standard methods in terms of mean, variance, and MSE. The Hybrid method stands out with the lowest variance (7.288) and MSE (17.439) values, while OIS and WIS have the highest variance (494.604) and MSE (1738.651) values. These results show that hybrid methods provide significant advantages in stability and efficiency for off-policy learning.

Table 2. The performance metrics summary

Id	Method	Mean	Variance	MSE	Relative Efficiency (1/Var)
3	Hybrid	3.386018	7.288191	17.438903	0.137208
5	HybridNorm	4.062562	15.157929	30.077313	0.065972
4	HybridClip	7.551323	25.907486	79.949436	0.038599
2	PDIS	13.373487	115.457750	288.998501	0.008661
0	OIS	35.471059	494.603887	1738.651486	0.002022
1	WIS	35.471059	494.603887	1738.651486	0.002022

Conclusions

This study proposes a Hybrid Importance Sampling (HIS) method to address the high variance and instability issues in off-policy reinforcement learning. By combining adaptive clipping and dynamic normalization techniques, HIS effectively controls the unbounded growth of importance weights and enhances the stability of estimates. Experimental results demonstrate that HIS achieves lower variance and MSE values compared to standard methods (OIS, WIS, PDIS), while maintaining balanced learning dynamics.

However, the hybrid methods' relatively lower average reward performance indicates a trade-off between stability and reward maximization. While this characteristic may be advantageous in risk-sensitive applications, it requires careful consideration in scenarios prioritizing high rewards. Future work will focus on refining hybrid methods to better optimize the reward-variance trade-off.

In conclusion, the HIS method presents a promising approach for enhancing the efficiency and reliability of off-policy learning in real-world applications such as robotics, autonomous driving, and medical decision support systems.

Acknowledgements

Not applicable.

References

[1] Lillicrap T., Hunt J., Pritzel A., Heess N., Erez T., Tassa Y. et al.. Continuous control with deep reinforcement learning. 2015. https://doi.org/10.48550/arxiv.1509.02971

[2] Haarnoja T., Zhou A., Hartikainen K., Tucker G., Ha S., Tan J. et al.. Soft actor-critic algorithms and applications. 2018. https://doi.org/10.48550/arxiv.1812.05905

[3] Wang R., Foster D., & Kakade S.. What are the statistical limits of offline rl with linear function approximation?. 2020. https://doi.org/10.48550/arxiv.2010.11895

[4] Fujimoto S., Meger D., & Precup D.. Off-policy deep reinforcement learning without exploration. 2018. https://doi.org/10.48550/arxiv.1812.02900

[5] Silver D., Schrittwieser J., Simonyan K., Antonoglou I., Huang A., Guez A. et al.. Mastering the game of go without human knowledge. Nature 2017;550(7676):354-359. https://doi.org/10.1038/nature24270

[6] Mandyam A., Jones A., Laudański K., & Engelhardt B.. Nested policy reinforcement learning. 2021. https://doi.org/10.48550/arxiv.2110.02879

[7] Fujimoto S., Conti E., Ghavamzadeh M., & Pineau J.. Benchmarking batch deep reinforcement learning algorithms. 2019. https://doi.org/10.48550/arxiv.1910.01708

[8] Mnih V., Kavukcuoglu K., Silver D., Rusu A., Veness J., Bellemare M. et al.. Human-level control through deep reinforcement learning. Nature 2015;518(7540):529-533. https://doi.org/10.1038/nature14236

[9] Chen Z.. A unified lyapunov framework for finite-sample analysis of reinforcement learning algorithms. ACM SIGMETRICS Performance Evaluation Review 2022;50(3):12-15. https://doi.org/10.1145/3579342.3579346

[10] Shi L., Li S., Cao L, Long Y., & Pan G.. Tbq(σ): improving efficiency of trace utilization for off-policy reinforcement learning. 2019. https://doi.org/10.48550/arxiv.1905.07237

[11] Kumar A., Fu J., Tucker G., & Levine S.. Stabilizing off-policy q-learning via bootstrapping error reduction. 2019. https://doi.org/10.48550/arxiv.1906.00949

[12] Touati A., Zhang A., Pineau J., & Vincent P.. Stable policy optimization via off-policy divergence regularization. 2020. https://doi.org/10.48550/arxiv.2003.04108

[13] Imani E., Graves E., & White M.. An off-policy policy gradient theorem using emphatic weightings. 2018. https://doi.org/10.48550/arxiv.1811.09013

[14] Munos R., Stepleton T., Harutyunyan A., & Bellemare M.. Safe and efficient off-policy reinforcement learning. 2016. https://doi.org/10.48550/arxiv.1606.02647

[15] Gu S., Lillicrap T., Ghahramani Z., Turner R., & Levine S.. Q-prop: sample-efficient policy gradient with an off-policy critic. 2016. https://doi.org/10.48550/arxiv.1611.02247

[16] Kallus N. and Uehara M.. Intrinsically efficient, stable, and bounded off-policy evaluation for reinforcement learning. 2019. https://doi.org/10.48550/arxiv.1906.03735

[17] Tokdar S. and Kass R.. Importance sampling: a review. WIREs Computational Statistics 2009;2(1):54-60. https://doi.org/10.1002/wics.56

[18] Yu T., Lu L., & Li J.. A weight-bounded importance sampling method for variance reduction. International Journal for Uncertainty Quantification 2019;9(3):311-319. https://doi.org/10.1615/int.j.uncertaintyquantification.2019029511

[19] Liu Y., Bacon P., & Brunskill E.. Understanding the curse of horizon in off-policy evaluation via conditional importance sampling. 2019. https://doi.org/10.48550/arxiv.1910.06508