

Truncated Importance-Sampling Emphatic TD3 at Deregulated Electricity Market

Jakson F. S. Araujo¹, Vítor O. Pochmann² and Luís G. P. Meloni³

Abstract—This paper introduces a novel reinforcement learning method, called Truncated Importance-Sampling Emphatic TD3 (TISE-TD3), which is tested in a Home Energy Management Systems (HEMS) framework. Operating in a free-energy market, the model uses real-world household consumption data, weather signals (irradiance and temperature), and price series at 5-minute granularity. By integrating emphatic weightings with importance-sampling truncation into the TD3 architecture, the method addresses off-policy distribution mismatch while ensuring stability. Results demonstrate that TISE-TD3 offers higher performance and lower computational complexity than the standard TD3 and TISE-SAC baselines, making it ideal for real-time energy optimization in energy storage systems.

Keywords—HEMS, TISE-SAC, TD3, Electricity Market.

I. INTRODUCTION

The modern energy sector is transitioning toward decentralized architectures with a growing penetration of renewable energy sources and free energy markets. In this context, Vehicle-to-Everything (V2X) technologies and residential energy storage systems (ESS) emerge as important components capable of transforming electric vehicles (EVs) into distributed storage assets that support grid operation and improve energy efficiency.

The integration of ESS and EVs can mitigate renewable energy curtailment by storing excess photovoltaic (PV) generation and reallocating energy during periods of high demand.

However, coordinating these assets is a complex task due to the stochastic behavior of residential consumption, renewable generation, and dynamic electricity prices. The high dimensionality of the state space and the requirement for real-time decision-making render traditional heuristic-based management ineffective. Without the support of autonomous agents trained via Reinforcement Learning (RL), it is nearly impossible to optimize complex objectives, such as minimizing battery degradation (State of Health - SoH) while maximizing economic arbitrage in dynamic market conditions.

Home Energy Management Systems (HEMS) combined with Reinforcement Learning (RL) algorithms have become a very useful solution for optimizing residential energy consumption and reducing operational costs [1]. To address the scalability limitations of classical RL approaches, Deep Reinforcement Learning (DRL) algorithms have been increasingly adopted in HEMS applications. In [2], TD3 was applied

to residential energy management with PV, EV, ESS, and HVAC devices, achieving cost reductions of up to 12.45%. Likewise, [3] compared TD3, DDPG, SAC, and PPO in a hybrid HEMS scenario, where TD3 achieved the best balance between operational cost reduction and PV self-consumption.

The Truncated Importance-Sampling Emphatic SAC (TISE-SAC) algorithm [4] mitigates the distribution mismatch problem inherent to off-policy reinforcement learning through emphatic weighting combined with truncated importance sampling. By prioritizing updates according to a state-interest measure, the method improves the relevance of sampled transitions during learning. The paper results presented superior performance and enhanced training stability relative to the conventional SAC algorithm.

Among DRL algorithms for continuous control, the Twin Delayed Deep Deterministic Policy Gradient (TD3) stands out due to its dual-critic architecture and delayed policy updates, which reduce value overestimation and improve learning stability [5]. Motivated by these characteristics, this work proposes a TD3-based HEMS framework combined with emphatic weighting and truncated importance sampling to coordinate the operation of stationary ESS and V2X assets. The proposed approach aims to improve learning stability, optimize residential energy management, and reduce costs under dynamic pricing scenarios.

The major paper contributions are:

- Starting from the classical Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm, this work proposes an enhanced version that incorporates two key mechanisms: (a) an interest function driven by both the battery state-of-charge (SoC) and electricity market prices, and (b) truncated importance sampling to mitigate variance and improve numerical stability. In addition, the paper provides a detailed pseudocode description of the proposed method.
- The proposed approach is further evaluated with respect to computational complexity, demonstrating significant advantages over two baseline methods, namely the standard TD3 and the TISE-SAC algorithm.

This paper is organized as follows: Section II presents the background on the photovoltaic and battery storage models, as well as the baseline reinforcement learning methods. Section III introduces the proposed TISE-TD3 algorithm. Section IV describes the reinforcement learning framework and the environment formulation. Section V discusses the simulation setup and experimental results. Finally, Section VI concludes the paper and outlines future research directions.

¹²³Communications Department, School of Electrical and Computer Engineering, University of Campinas, Campinas, SP 13083-852 BRAZIL,

¹e-mail: jaksonfelix@gmail.com;

²e-mail: v.pochmann15@gmail.com;

³e-mail: meloni@unicamp.br

II. BACKGROUND

The proposed reinforcement learning (RL) framework incorporates models for photovoltaic (PV) panels and energy storage systems (ESS). For baseline comparison, the classical TD3 algorithm is also considered.

A. Solar Photovoltaic Model

The power output of the photovoltaic (PV) system, p_t^{PV} , is modeled considering solar irradiance G_t and the thermal influence on cell efficiency [6]:

$$p_t^{\text{PV}} = P_{\text{max}}^{\text{PV}} \frac{G_t}{G_{\text{STC}}} \eta^{\text{PV}} [1 + \zeta(T_t^{\text{cell}} - T^{\text{STC}})], \quad (1)$$

where $P_{\text{max}}^{\text{PV}}$ is the peak power, η^{PV} is the system efficiency, and ζ is the temperature coefficient. The cell temperature T_t^{cell} accounts for ambient temperature T_t^{amb} and irradiance effects as follows:

$$T_t^{\text{cell}} = T_t^{\text{amb}} + \frac{G_t}{G_{\text{NOCTC}}} (T^{\text{NOCT}} - T^{\text{NOCTC}}). \quad (2)$$

Parameters G_{STC} , T^{STC} , T^{NOCT} , and G_{NOCTC} represent standard and normal operating test conditions. All power and temperature units are in Watts and Celsius, respectively.

B. Battery Energy Storage System

The Battery Energy Storage System (BESS) state of charge (SoC) dynamics is modeled as [7]:

$$\text{SoC}_t = \text{SoC}_{t-1} + \frac{\delta}{C_{\text{bat}}} \left(p_t^{\text{ch}} \eta_{\text{ch}} - \frac{p_t^{\text{dch}}}{\eta_{\text{dch}}} \right), \quad (3)$$

where C_{bat} is the nominal capacity, δ is the time step, and $\eta_{\text{ch}}, \eta_{\text{dch}}$ are charging/discharging efficiencies. To ensure physical feasibility, the power command p_t is subject to three constraints: (i) SoC limits within $[\text{SoC}_{\text{min}}, \text{SoC}_{\text{max}}]$, (ii) ramp-rate limits $|p_t - p_{t-1}| \leq R_{\text{max}}$, and (iii) directional clipping to prevent simultaneous charging and discharging.

The final admissible power is determined by the priority of these constraints. In *SoC priority*, the BESS longevity is prioritized by enforcing SoC bounds before ramp-rate limits. Conversely, *ramp priority* emphasizes grid responsiveness by applying ramp-rate constraints first, enabling broader SoC excursions to accommodate rapid power setpoint changes.

C. Twin Delayed Deep Deterministic Policy Gradient (TD3)

The Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm is an off-policy, model-free reinforcement learning framework that extends DDPG to mitigate overestimation bias and enhance stability in continuous control tasks. By utilizing deterministic policies and a replay buffer, the method demonstrates high sample efficiency in high-dimensional action spaces. To combat the positive bias common in single-critic approaches, TD3 introduces clipped double Q-learning, maintaining two independent critics to compute the target value from the minimum estimate between them, as expressed by

$$y = r + \gamma \min_{i=1,2} Q_{\theta'_i}(s', \pi_{\phi'}(s') + \epsilon), \quad (4)$$

where the target network parameters are denoted by their respective indices. Additionally, the algorithm applies target policy smoothing by adding clipped Gaussian noise to the action used in the target value computation, defined as

$$\epsilon \sim \text{clip}(\mathcal{N}(0, \sigma), -c, c), \quad (5)$$

which regularizes the value function and prevents the policy from exploiting erroneous peaks in the estimates. A third essential contribution is the use of delayed policy updates, in which the actor is updated less frequently than the critics. This delay allows the Q-functions to converge toward a more stable target before each policy update, significantly reducing learning variance. By integrating these mechanisms, TD3 overcomes the limitations of DDPG under function approximation and establishes a competitive benchmark against entropy-regularized methods such as SAC. The TD3 algorithm is presented in [5].

III. THE TRUNCATED IMPORTANCE-SAMPLING EMPHATIC TD3 METHOD

The truncated emphatic formulation is designed to mitigate the distribution mismatch inherent to off-policy learning, where samples are generated by a behavior policy distinct from the target policy. Emphatic weighting modulates updates through a user-defined interest function, prioritizing states that are more relevant to the control objective. At the same time, truncation of importance-sampling ratios controls variance, leading to improved numerical stability.

We extend the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm by incorporating importance sampling into the temporal-difference updates. Let $i(s) \geq 0$ denote an interest function specifying the relative importance of state s . We consider an off-policy setting where trajectories are collected using a behavior policy b , while the deterministic target policy π_{ϕ} is optimized.

The follow-on trace F_t aggregates discounted future interest values weighted by importance sampling, and satisfies the recursion [8]

$$F_t = i(s_t) + \gamma_t \rho_{t-1} F_{t-1}, \quad (6)$$

with $F_{-1} = 0$ and importance ratio $\rho_t = \frac{\pi(a_t|s_t)}{b(a_t|s_t)}$. In TD3, since π_{ϕ} is deterministic, $\pi(a_t|s_t)$ is interpreted as a Dirac measure at $\pi_{\phi}(s_t)$; in practice, we rely on a stochastic approximation to evaluate the ratio.

To stabilize learning, we adopt a truncated variant of the V-trace weighting scheme [9]. Rather than using multi-step returns, we retain only the clipped importance ratios and apply them to scale one-step actor and critic updates, which is compatible with replay-buffer training.

Given a minibatch $(s_t, a_t, r_t, s_{t+1}, i_t, \log \pi_b(a_t|s_t))$ drawn from replay buffer (RB) $\mathcal{B}$, we approximate the target policy density using Gaussian smoothing (as in TD3 target policy smoothing) and compute

$$\rho_t = \exp(\log \pi_{\phi}(a_t | s_t) - \log \pi_b(a_t | s_t)), \quad (7)$$

where π_{ϕ} denotes the stochastic surrogate of the deterministic actor. We then apply truncation:

$$\rho_t \leftarrow \min(\rho_t, \bar{\rho}), \quad c_i = \min(\rho_i, \bar{c}), \quad (8)$$

although in the present one-step setting, only ρ_t is used. An emphatic weight is constructed by combining the truncated ratio with the interest:

$$F_t = \frac{i_t \rho_t}{\frac{1}{B} \sum_{k=1}^B i_k \rho_k}. \quad (9)$$

This weight is treated as a constant during optimization, i.e., $\nabla_\phi F_t = 0$, avoiding gradient flow through the importance ratios.

The resulting weight scales both critic and actor objectives. Using two critics to reduce overestimation bias, the critic loss is

$$L_Q = \mathbb{E}_t \left[F_t \sum_{i=1}^2 (Q_{\theta_i}(s_t, a_t) - y_t)^2 \right], \quad (10)$$

with target

$$y_t = r_t + \gamma \min_{i=1,2} Q_{\theta_i}(s_{t+1}, \tilde{a}_{t+1}), \quad (11)$$

where

$$\tilde{a}_{t+1} = \pi_\phi(s_{t+1}) + \epsilon, \quad \epsilon \sim \text{clip}(\mathcal{N}(0, \sigma), -c, c).$$

The actor is updated via the deterministic policy gradient

$$L_\pi = \mathbb{E}_t [F_t (-Q_{\theta_1}(s_t, \pi_\phi(s_t)))], \quad (12)$$

where only Q_{θ_1} is used. As in TD3, actor updates are delayed relative to critic updates.

Overall, the method inherits the variance-reduction properties of truncated importance sampling while retaining the stability advantages of TD3. The complete Truncated Importance-Sampling Emphatic TD3 (TISE-TD3) algorithm is summarized in **Algorithm 1**.

IV. REINFORCEMENT LEARNING FRAMEWORK

In the agent–environment framework, the system evolves in discrete time steps, with sampling interval $\delta = 5$ minutes. At each time step t , the agent observes the environment state S_t , defined as a vector with

$$S_t = [P_t^{\text{load}}, P_t^{\text{PV}}, \text{SoC}_t, \lambda_t], \quad (13)$$

where P_t^{load} denotes the residential load demand, P_t^{PV} is the photovoltaic generation, SoC_t is the ESS state of charge, and λ_t represents the electricity price. After observing it, the agent applies an action, $a_t^{\text{agent}} \in [-1, 1]$, representing the fraction of the ESS rated power to be used. This action is mapped into the power applied to the environment as

$$a_t^{\text{env}} = a_t^{\text{agent}} P^{\text{ESS}_{\text{max}}}, \quad (14)$$

where $P^{\text{ESS}_{\text{max}}}$ is the maximum rated ESS power that satisfies the ESS operational constraints. After the action is applied, the environment updates its dynamics, computes the reward r_t , and provides the next state S_{t+1} . This process is repeated until a terminal state is reached.

The reward function is designed to reflect the economic objectives of the optimization problem while encouraging

operation around an intermediate SoC level. Therefore, a bonus term is added when the ESS SoC remains close to 50%:

$$r_t = P_t^{\text{ESS}} \lambda_t \delta + \mu \left[1 - 4 \left(\text{SoC}_t - \frac{1}{2} \right)^2 \right], \quad (15)$$

where μ is a weighting constant.

The proposed interest function $i_t(s)$ considers both the electricity price and the SoC of ESS, aiming to prioritize states with higher economic potential while discouraging deep charge and discharge cycles. Then, the price-related component is modeled using a shifted and scaled logistic sigmoid function:

$$i(\lambda_t) = \frac{1}{1 + \exp[-k(|\lambda_t - \bar{\lambda}| - p_0)]}, \quad (16)$$

where $\bar{\lambda} = \mathbb{E}[\lambda]$ is the average electricity price, k controls the curve steepness, and p_0 defines the inflection point. The SoC-related component is defined as

$$i(\text{SoC}_t) = 1 - 4 \left[\frac{\text{SoC}_t - \text{SoC}_{\text{min}}}{\text{SoC}_{\text{max}} - \text{SoC}_{\text{min}}} - \frac{1}{2} \right]^2. \quad (17)$$

Finally, the global interest function is obtained as the weighted combination of both components with weights ν_1, ν_2 and ν_3 :

$$i(\lambda_t, \text{SoC}_t) = \nu_1 i(\text{SoC}_t) + \nu_2 i(\lambda_t) + \nu_3. \quad (18)$$

V. SIMULATION ANALYSIS

The DRL models are evaluated in a synthetic stochastic environment developed using multiple real-world data sources. Residential electricity consumption is modeled using the RE-FIT dataset [10] (House 10) spanning a nine-month period. For the same temporal horizon, electricity market prices were retrieved from the Pennsylvania–New Jersey–Maryland Interconnection (PJM) [11], specifically from the COMED node. To complement the model, solar power generation was estimated using meteorological data (irradiance and temperature) from the National Renewable Energy Laboratory (NREL) [12] for Campinas, Brazil. These variables were combined to compute the final PV power output using (1) and (2).

The DRL models share a common neural network architecture, differing only in a subset of algorithm-specific hyperparameters. The complete set of hyperparameters is summarized in Table I. The interest-function weights were selected based on the sensitivity analysis reported in [4]. All models were trained for 350 episodes using a training partition fully independent from the test partition. At the end of an episode, the battery State of Charge (SoC) is preserved as the initial SoC for the subsequent episode using *ramp priority*, while the starting week is randomly selected to increase scenario diversity and improve policy generalization. In contrast, the test partition is evaluated over a sequential 12-week period, that is, non-randomized week selection.

The evaluation framework considers residential energy consumption as the reference load profile. The analyzed scenario consists of an aggregated household demand integrated with a rooftop photovoltaic (PV) generation system. The PV installation is dimensioned to achieve near self-sufficiency for

TABLE I
THE HYPERPARAMETER SETTINGS FOR THE PROPOSED MODEL.

Parameter	Symbol	Value
2 Hidden-Layers of size	–	256
Reward weight constant	μ	1/30
Discount Factor	γ	0.99
Learning Rate	α	3×10^{-4}
Soft Update	τ	0.005
Exploration noise	σ	0.10
Noise clipping	c	0.2
Policy Delay	–	2
Activation functions	–	ReLU & tanh
Optimizer	–	Adam
Batch Size	–	256
IS truncation	$\bar{\rho}$	1.0
Reward Weights	$[\nu_1, \nu_2, \nu_3]^T$	$[0.5, 0.3, 0.2]^T$
# Episodes in RB	–	500

the aggregated residence, accounting for both local energy consumption and PV power generation dynamics. The electrical characteristics and operational parameters of all modeled devices follow the specifications adopted in [4].

Figure 1 presents the average undiscounted returns throughout training, considering 10 independent runs with different random seeds. The transient regime in Fig. 1(a) illustrates the convergence behavior during the first 100 training episodes, while the asymptotic regime in Fig. 1(b) highlights the long-term stability between episodes 250 and 350.

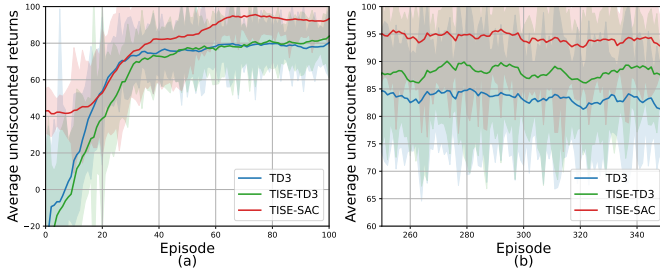


Fig. 1. Average undiscounted returns during training: (a) transient regime and (b) convergence regime.

Figure 2 presents the weekly energy balance achieved by each evaluated control strategy over the test horizon, which is analyzed along with Table II, reporting the average net economic gains achieved by each method across the entire test period. The results — showing the gains of the DRL models relative to home consumption — indicate that TISE-SAC consistently achieves superior economic performance across most evaluation weeks, outperforming both conventional TD3 and TISE-SAC in the proposed HEMS environment. TISE-TD3 outperforms vanilla TD3, suggesting that its integration enhances the agent’s ability to optimize long-term energy management decisions under dynamic operating conditions.

TABLE II
AVERAGE WEEKLY NET GAIN COMPARISON.

Home	Home + PV	TD3	TISE-SAC	TISE-TD3
100.00%	3.47%	38.47%	47.83%	42.68%

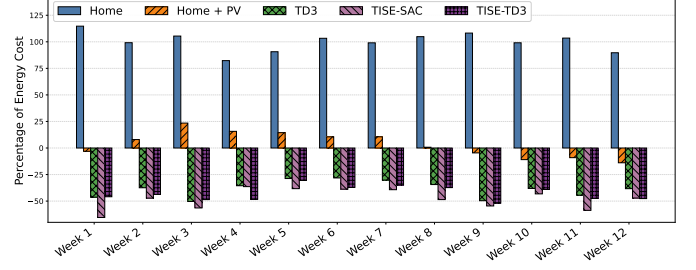


Fig. 2. Weekly net gain comparison among the evaluated reinforcement learning methods.

Table III presents an ablation analysis of the proposed TISE-TD3 method, isolating the contributions of the truncated importance-sampling factor (ρ_t) and the interest function (i_t). The configuration $\rho_t = i_t = 1$ corresponds to the standard TD3 algorithm—see (9). The results indicate that both mechanisms contribute to the overall performance, while the truncated importance-sampling component plays a central role in improving the agent’s decision-making and economic gains.

TABLE III
ABLATION ANALYSIS OF TISE-TD3.

Method	Gains
TISE-TD3	42.68%
TISE-TD3, $\rho_t = 1$	37.15%
TISE-TD3, $i_t = 1$	33.27%
TISE-TD3, $\rho_t = i_t = 1$	38.47%

In order to evaluate the computational efficiency of the evaluated DRL methods, the training cost of each model was normalized with respect to the TISE-SAC baseline (wall clock). Table IV summarizes the relative computational costs obtained under identical experimental conditions. Standard TD3 achieved a relative computational cost of 67.61% compared to TISE-SAC, while the proposed TISE-TD3 presents a relative computational cost of 69.05%. These results indicate that the proposed method achieves superior computational efficiency than TISE-SAC while preserving the benefits associated with emphatic weighting and truncated importance sampling.

TABLE IV
RELATIVE TRAINING COMPUTATIONAL COST NORMALIZED TO TISE-SAC.

Method	Relative Computational Cost (%)
TISE-SAC	100.00
TD3	67.61
TISE-TD3	69.05

VI. CONCLUSIONS

This paper introduced TISE-TD3, a novel reinforcement learning method that integrates truncated importance sampling and emphatic weighting into the TD3 framework for home energy management systems operating in deregulated electricity markets. The proposed method was evaluated using real-world residential consumption, electricity price, and

photovoltaic generation data. Experimental results showed that TISE-TD3 achieved superior economic performance compared to standard TD3 and inferior performance to TISE-SAC, while presenting lower computational cost than the latter. The ablation analysis demonstrated the relevance of the proposed components, particularly the truncated importance-sampling mechanism, in improving decision-making and training stability. Future work includes extending the method to multi-agent energy systems and vehicle-to-grid applications.

ACKNOWLEDGEMENTS

This work was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior-Brasil (CAPES) under Grant 001.

REFERENCES

- [1] W. Pinthurat, T. Surinkaew, and B. Hredzak, "An overview of reinforcement learning-based approaches for smart home energy management systems with energy storages," *Renewable and Sustainable Energy Reviews*, vol. 202, p. 114648, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1364032124003745>
- [2] Y. Ye, D. Qiu, H. Wang, Y. Tang, and G. Strbac, "Real-time autonomous residential demand response management based on twin delayed deep deterministic policy gradient learning," *Energies*, 2021. [Online]. Available: <https://www.mdpi.com/1996-1073/14/3/531>
- [3] Y. Xu, Y. Li, and W. Gao, "Comparative analysis of reinforcement learning approaches for multi-objective optimization in residential hybrid energy systems," *Buildings*, 2024. [Online]. Available: <https://www.mdpi.com/2075-5309/14/9/2645>
- [4] V. O. Pochmann and L. G. P. Meloni, "Deep reinforcement learning in hems at deregulated electricity market," *IEEE Access*, vol. 14, pp. 1–14, 2026.
- [5] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 1587–1596, 2018.
- [6] H. Fontenot and B. Dong, "Modeling and control of building-integrated microgrids for optimal energy management: A review," *Applied Energy*, vol. 254, p. 113689, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0306261919313765>
- [7] S. Lee and D. H. Choi, "Reinforcement learning-based energy management of smart home with rooftop solar photovoltaic system, energy storage system, and home appliances," *Sensors*, 2019.
- [8] R. S. Sutton, A. R. Mahmood, and M. White, "An emphatic approach to the problem of off-policy temporal-difference learning," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2603–2631, 2016.
- [9] L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, D. Yotam, V. Firoiu, T. Harley, I. Dunning, S. Legg, and K. Kavukcuoglu, "IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures," *arXiv preprint arXiv:1802.01561*, 2018. [Online]. Available: <https://arxiv.org/abs/1802.01561>
- [10] V. S. David Murray, Lina Stankovic, "REFIT: Electrical load measurements," 2016. [Online]. Available: <https://pureportal.strath.ac.uk/en/datasets/refit-electrical-load-measurements-cleaned>
- [11] PJM, "PJM - data miner 2, public data," 2024. [Online]. Available: <https://dataminer2.pjm.com/list>
- [12] NREL, "Data and tools, NREL develops data sets, maps, models, tools, and software for the analysis and development of renewable energy and energy efficiency technologies," 2024. [Online]. Available: <https://www.nrel.gov/research/data-tools.html>

Algorithm 1 Truncated Importance-Sampling Emphatic TD3 (TISE-TD3)

```

1: Initialize critic  $Q_{\theta_1}, Q_{\theta_2}$  and actor  $\pi_\phi$  networks
2: Initialize target networks  $\bar{\theta}_1 \leftarrow \theta_1, \bar{\theta}_2 \leftarrow \theta_2, \bar{\phi} \leftarrow \phi$ 
3: Initialize replay buffer  $\mathcal{B}$ 
4: Set truncation threshold  $\bar{\rho}$  and interest function  $i(s)$ 
5: Set discount factor  $\gamma$ , policy delay  $d$ 
6: Set exploration noise  $\sigma_{\text{expl}}$ , target noise  $\sigma$ , noise clip  $c$ 
7: Set target update rate  $\tau$ 
8: for each iteration do
9:   Observe state  $s_t$ 
10:  Select action  $a_t = \pi_\phi(s_t) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_{\text{expl}})$ 
11:  Execute  $a_t$ , observe  $r_t, s_{t+1}$ 
12:  Store  $(s_t, a_t, r_t, s_{t+1}, i_t, \log \pi_b(a_t|s_t))$  in  $\mathcal{B}$ 
13:  for each gradient step do
14:    Sample minibatch  $\{(s_t, a_t, r_t, s_{t+1}, i_t, \log \pi_b)\}_{t=1}^B$ 
    from  $\mathcal{B}$ 
15:    Compute  $\log \pi_\phi(a_t|s_t)$  via stochastic approxima-
    tion of  $\pi_\phi$ 
16:     $\rho_t \leftarrow \exp(\log \pi_\phi(a_t|s_t) - \log \pi_b(a_t|s_t))$ 
17:     $\bar{\rho}_t \leftarrow \min(\rho_t, \bar{\rho})$ 
18:    Compute normalized emphatic weights:
        
$$F_t = \frac{i_t \rho_t}{\frac{1}{B} \sum_{k=1}^B i_k \rho_k}$$

19:    Detach  $F_t$  from computational graph
20:    Sample target noise  $\epsilon \sim \text{clip}(\mathcal{N}(0, \sigma), -c, c)$ 
21:     $\tilde{a}_{t+1} \leftarrow \pi_{\bar{\phi}}(s_{t+1}) + \epsilon$ 
22:    Compute target:
        
$$y_t = r_t + \gamma \min_{i=1,2} Q_{\bar{\theta}_i}(s_{t+1}, \tilde{a}_{t+1})$$

23:    Update critics by minimizing:
        
$$L_Q = \frac{1}{B} \sum_{t=1}^B F_t \sum_{i=1}^2 (Q_{\theta_i}(s_t, a_t) - y_t)^2$$

24:    if iteration mod  $d = 0$  then
25:      Update actor using:
        
$$L_\pi = -\frac{1}{B} \sum_{t=1}^B F_t Q_{\theta_1}(s_t, \pi_\phi(s_t))$$

26:      Update target networks:
        
$$\bar{\theta}_i \leftarrow \tau \theta_i + (1 - \tau) \bar{\theta}_i, \quad i = 1, 2$$

        
$$\bar{\phi} \leftarrow \tau \phi + (1 - \tau) \bar{\phi}$$

27:    end if
28:  end for
29: end for

```
