EXPLORACIÓN DEL DDPG TD3: UN ENFOQUE CONTINUO EN EL APRENDIZAJE POR REFUERZO

"Un Viaje a Través del Twin Delayed Deep Deterministic Policy Gradient y su Aplicación en Ambientes Desafiantes: HumanoidBulletEnv-v0"

José Javier Gutiérrez Gil

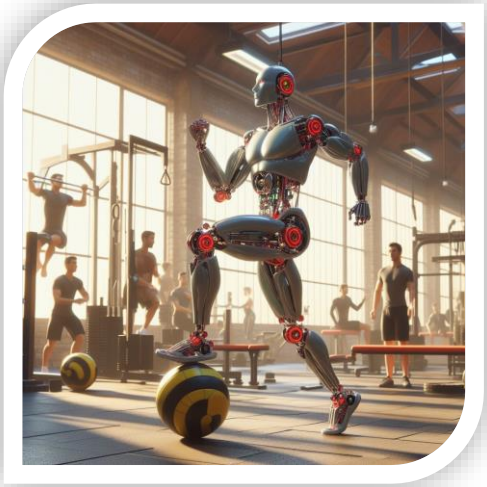
jogugil@alumni.uv.es

VNIVERSITAT DE VALÈNCIA

Grau en **Ciència de Dades**

Escola Tècnica Superior d'Enginyeria (ETSE-UV)

INTRODUCCIÓN

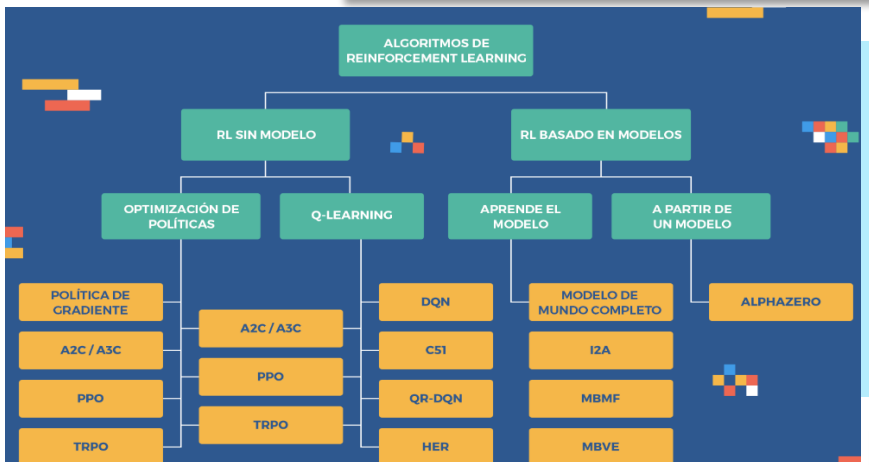


1. *Introducción*
2. *TD3: Twin Delayed DDPG*
3. *Implementación*
4. *Resultados*
5. *Conclusiones y Discusión*

José Javier Gutiérrez Gil (jogugil@alumni.uv.es/ jogugil@gmail.com)

Introducción: RL- GYM

DDPG TD3: Gradiente Política Determinista de doble retardo



Diferencias Temporales (TD)
Vs Programación dinámica (PD):

Actualización valores Q
Optimización Política (∏)
No requiere modelo completo del entorno

❑ Estado

- *Posición, velocidad y orientación del cuerpo del humanoide*

❑ Entorno: *PyBullet y Gym*

- ***HumanoidBulletEnv-v0*** *:Simulador físico 3D*
- *Aprenda a caminar de manera eficiente y sin caerse.*

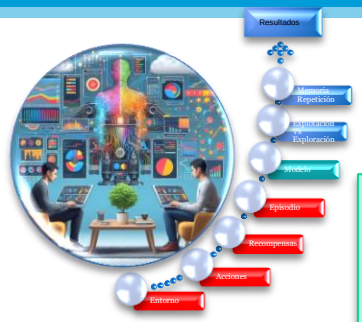
❑ Acciones

- ***Acciones continuas*** *(fuerzas aplicadas a las articulaciones)*

❑ Recompensa:

- *Mantenerse de pie*
- *Progresión hacia delante*
- *Eficiencia energética*

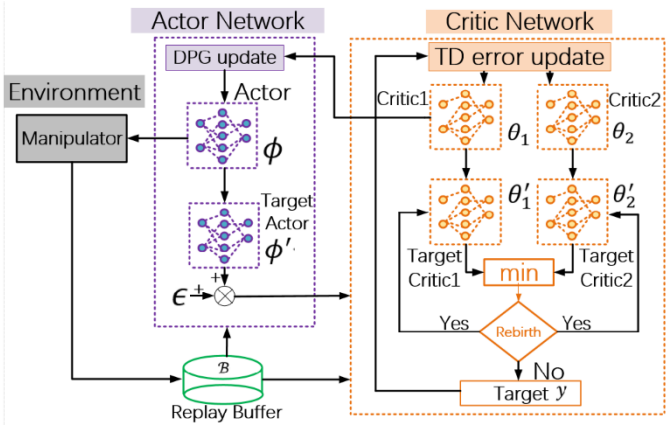
José Javier Gutiérrez Gil (jogugil@alumni.uv.es/ jogugil@gmail.com)



Introducción: RL- GYM

Deep Q-learning no es aplicable en el caso de acciones continúas

DDPG TD3



- ❑ **Actor Target + Actor Modl:**
  - ✓ **2 Capas ocultas + 2 Relu +Tanh**
- ❑ **2 Critic Target + 2 Critic Model:**
  - ❑ **2 Capas Ocultas + 1 Relu**

Addressing Function Approximation Error in Actor-Critic Method.
https://arxiv.org/pdf/1802.09477.pdf

$$R_t = \sum_{i=t}^{T} \gamma^{i-t} r(s_i, a_i)$$

$$J(\phi) = \mathbb{E}_{s_i \sim p_\pi, a_i \sim \pi}[R_0]$$

$$\nabla_\phi J(\phi) = \mathbb{E}_{s \sim p_\pi}\left[\nabla_a Q^\pi(s,a)|_{a=\pi(s)} \nabla_\phi \pi_\phi(s)\right]$$

$$\phi_{t+1} = \phi_t + \alpha \nabla_\phi J(\pi_\phi)|_{\phi_t}$$

ReplayBuffer

(s_1, s'_1, a_1, r_1)	(s_2, s'_2, a_2, r_2)	(s_3, s'_3, a_3, r_3)	(s_4, s'_4, a_4, r_4)	…

TD3: Twin Delayed DDPG

2 Críticos (M Vs T) + 1 Actor (M Vs T)

Algorithm 1 Twin Delayed DDPG

Input: initial policy parameters θ, Q-function parameters ϕ_1, ϕ_2, empty replay buffer $\mathcal{D}$
Set target parameters equal to main parameters $\theta_{\text{targ}} \leftarrow \theta$, $\phi_{\text{targ},1} \leftarrow \phi_1$, $\phi_{\text{targ},2} \leftarrow \phi_2$
repeat
  Observe state s and select action $a = \text{clip}(\mu_\theta(s) + \epsilon, a_{Low}, a_{High})$, where $\epsilon \sim \mathcal{N}$
  Execute a in the environment
  Observe next state s', reward r, and done signal d to indicate whether s' is terminal
  Store (s, a, r, s', d) in replay buffer $\mathcal{D}$
  If s' is terminal, reset environment state.
  if it's time to update **then**
    for j in range(however many updates) **do**
      Randomly sample a batch of transitions, $N = \{(s, a, r, s', d)\}$ from $\mathcal{D}$
      Compute target actions

$$a'(s') = \text{clip}\left(\mu_{\theta_{\text{targ}}}(s') + \text{clip}(\epsilon, -c, c), a_{Low}, a_{High}\right), \quad \epsilon \sim \mathcal{N}(0, \sigma)$$

      Compute targets

$$Q_{\phi}(r, s', d) = r + \gamma(1 - d) \min_{i=1,2} Q_{\phi_{\text{targ},i}}(s', a'(s'))$$

      Update Q-functions by one step of gradient descent using

$$\nabla_{\phi_i} \frac{1}{|N|} \sum_{(s,a,r,s',d) \in N} (Q_{\phi_i}(s, a) - Q_{\phi}(r, s', d))^2 \qquad \text{for } i = 1, 2$$

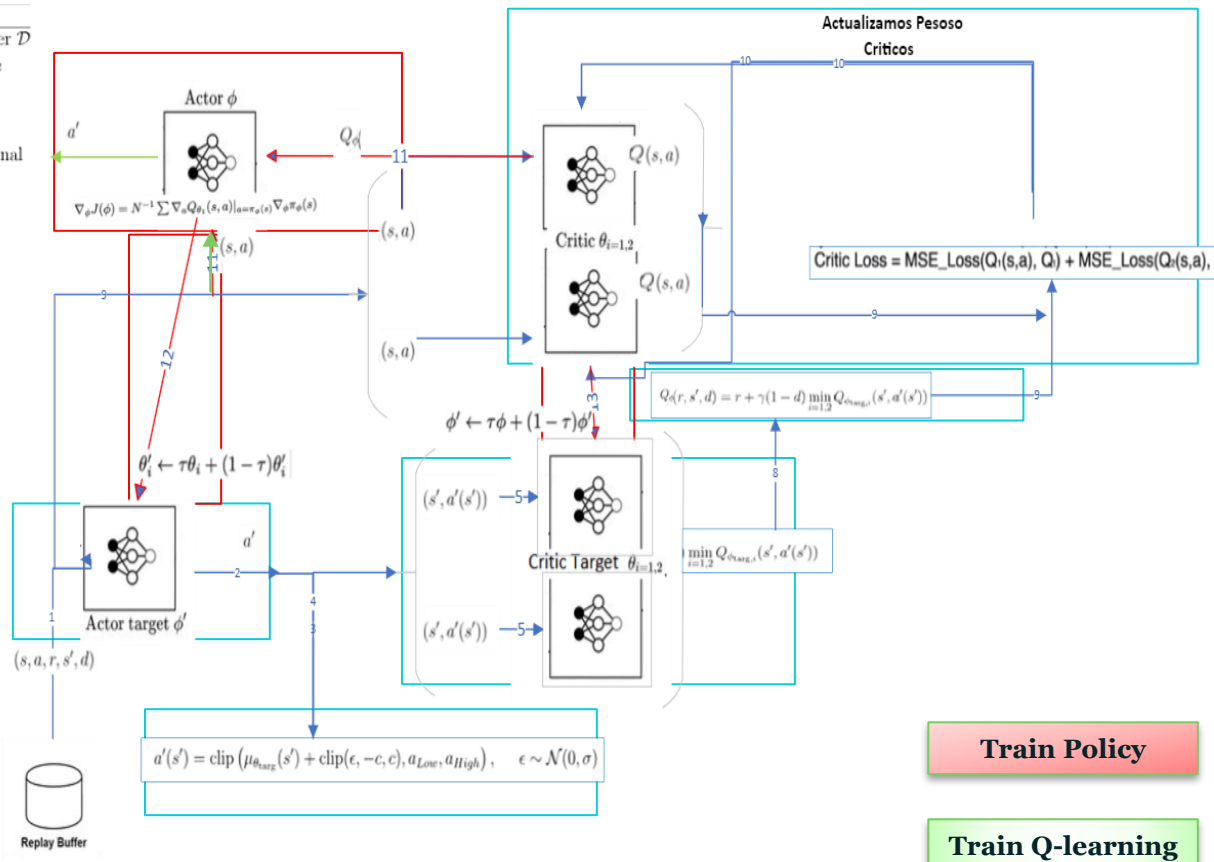
      if j mod `policy_delay` $= 0$ **then**
        Update policy by one step of gradient ascent using

$$\nabla_\phi J(\phi) = N^{-1} \sum \nabla_a Q_{\theta_1}(s, a)|_{a=\pi_\phi(s)} \nabla_\phi \pi_\phi(s)$$

        Update target networks with

$$\phi' \leftarrow \tau\phi + (1 - \tau)\phi' \qquad \text{for } i = 1, 2$$
$$\theta'_i \leftarrow \tau\theta_i + (1 - \tau)\theta'_i$$

      end if
    end for
  end if
until convergence



José Javier Gutiérrez Gil (jogugil@alumni.uv.es/ jogugil@gmail.com)

Implementación: Train

Métodos de regularización

Discovery Vs Applicaction

- EGreedyPolicy: (AdaptiveEGreedyPolicy)
- AdaptiveBalancedPolicy
- AdaptiveHybridPolicy
- Reward-proximity Exploration

Noise Action

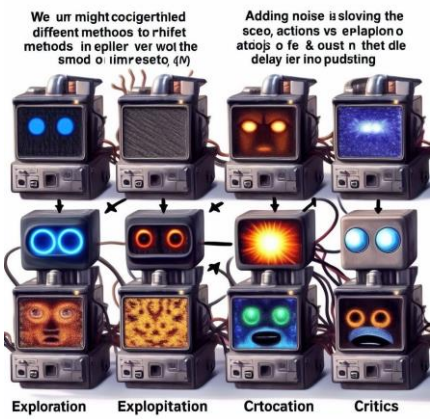
- Exploración
- Explotación (decaimiento)

Replay Memory

- Simple
- Simple PrioritizedReplayBuffer
- PrioritizedReplayBuffer
- ReplayBufferPos

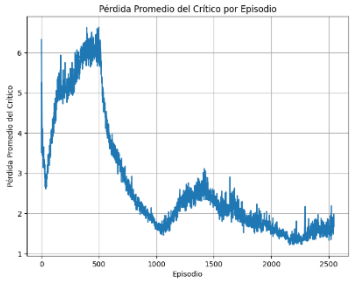
Dual Update Delay

- Cada dos pasos

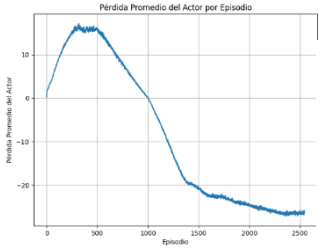


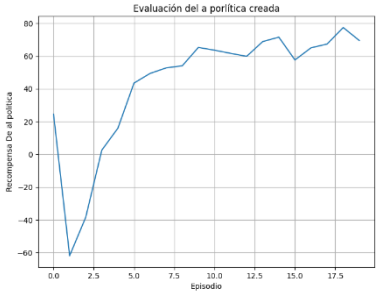
Resultados: Evaluación (I)

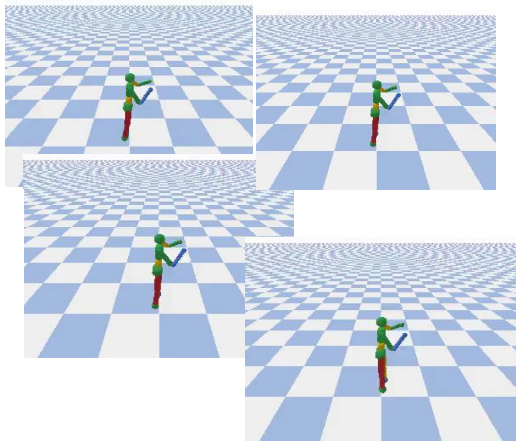
Explotación - Explotación



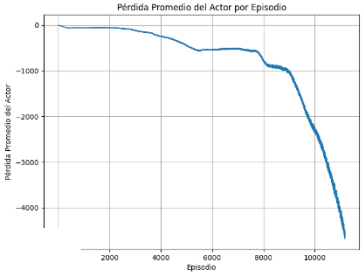
AdaptiveBalancedPolicy

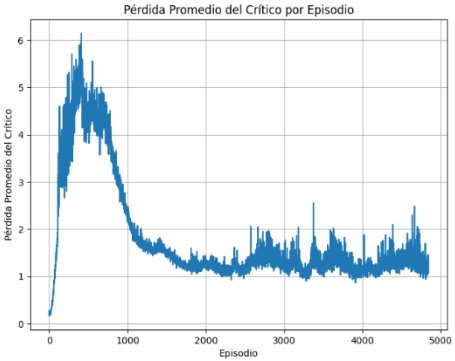


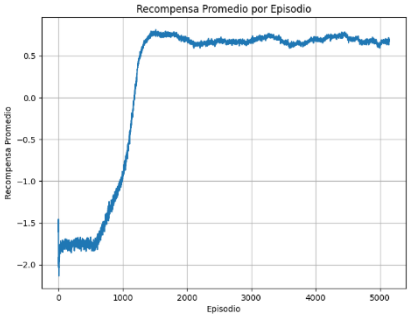


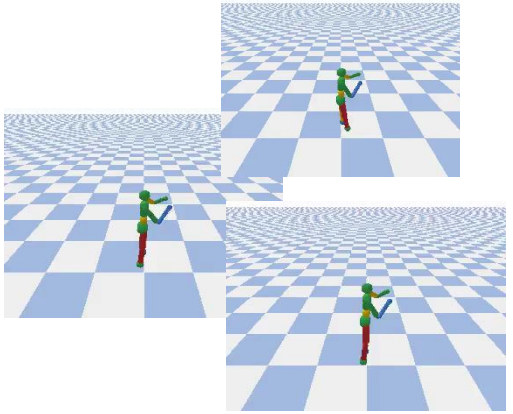


Reward-proximity Exploration



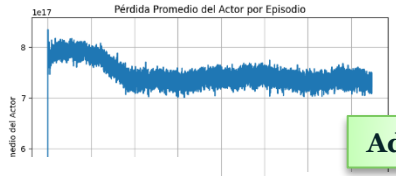




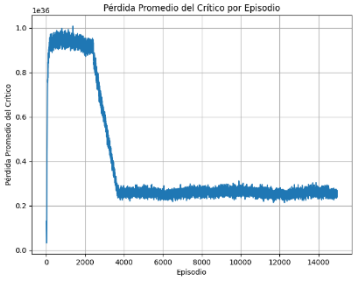


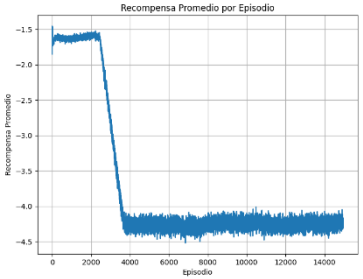
Resultados: Evaluación (II)

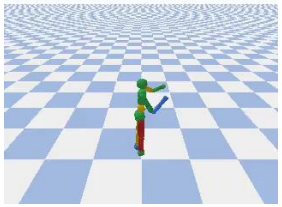
Explotación - Explotación

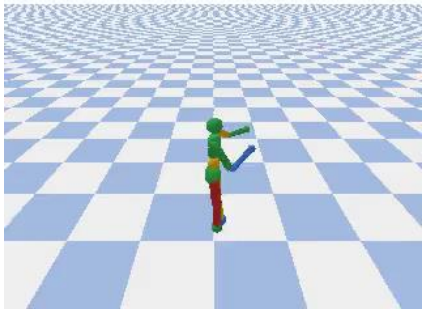


AdaptiveHybridPolicy

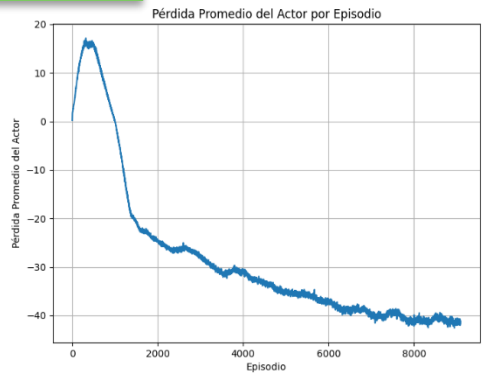


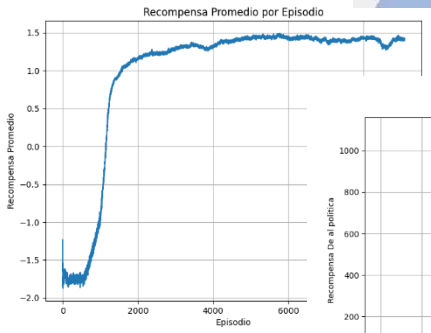


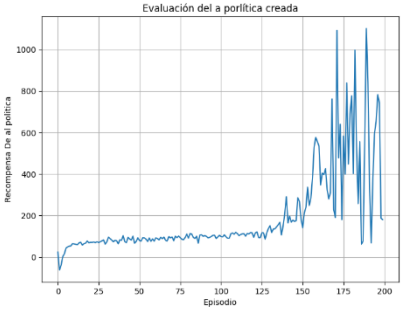


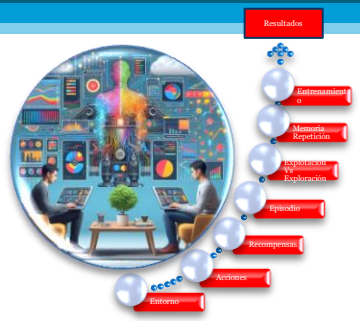


EGreedyPolicy





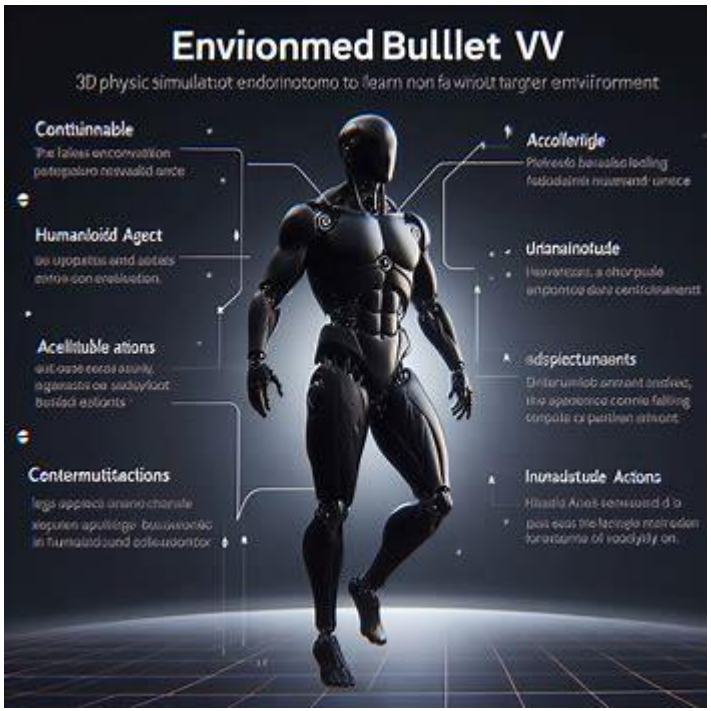




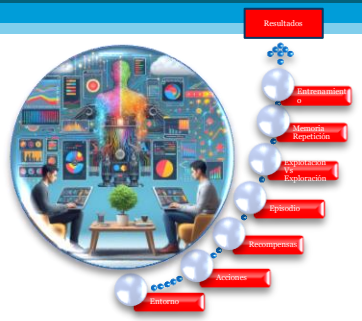
Conclusión: Modelo RL

Elevada Cantidad Pasos // Inestabilidad

Al evaluar la política: Se experimentan episodios en los que parece que el agente no está entrenado y su rendimiento es pobre:



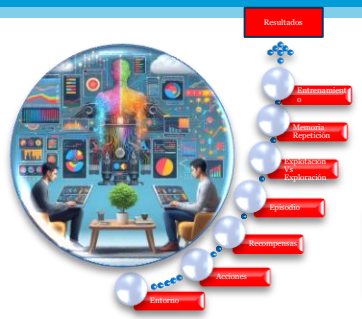
- ✓ Convergencia a acciones subóptimas
- ✓ Exploración insuficiente
- ✓ Convergencia insuficiente
- ✓ Hiperparámetros incorrectos
- ✓ Problema intrínseco de la tarea



Discusión: Modelo RL

Elevada Cantidad Pasos // Inestabilidad

- Llegar a entrenar bien el actor - 2 críticos para que el humanoide ande decentemente es un dolor de cabeza.
- Necesidad de regulación para evitar estancamientos y sobrestimación (los críticos)
- Muy sensible a los hiperparámetros y las acciones a realizar .
- La adición de ruido N(0,σ') mejora el algoritmo frente a búsqueda aleatoria entorno real



Aplicación: Modelo RL

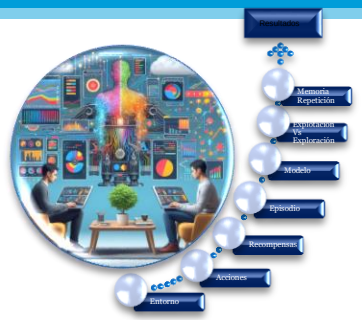
DDPG TD3 4.670 resultados

Multiagent Copilot Approach for Shared Autonomy between Human EEG and TD3 Deep Reinforcement Learning
https://www.semanticscholar.org/reader/836ccd5130525103c150336ae699bb36ceb86e94

The Control Method of Twin Delayed Deep Deterministic Policy Gradient with Rebirth Mechanism to Multi-DOF Manipulator
https://www.mdpi.com/2079-9292/10/7/870

R. Dubey, R. Loka and A. M. Parimi, "Maintaining the Frequency of AI-based Power System Model using Twin Delayed DDPG(TD3) Implementation," *2022 2nd International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC)*, Mathura, India, 2022, pp. 1-4, doi: 10.1109/PARC52418.2022.9726615

T. Tiong, I. Saad, K. T. K. Teo and H. b. Lago, "Deep Reinforcement Learning with Robust Deep Deterministic Policy Gradient," 2020 2nd International Conference on Electrical, Control and Instrumentation Engineering (ICECIE), Kuala Lumpur, Malaysia, 2020, pp. 1-5, doi: 10.1109/ICECIE50279.2020.9309539.



THE-END

Scott Fujimoto and Herke van Hoof and David Meger. (3028). Addressing Function Approximation Error in Actor-Critic Methods https://doi.org/10.48550/arXiv.1802.09477

Kim, Myeongseop & Han, Dong-Ki & Park, Jae-Han & Kim, Jung-Su. (2020). Motion Planning of Robot Manipulators for a Smoother Path Using a Twin Delayed Deep Deterministic Policy Gradient with Hindsight Experience Replay

Applied Sciences. 10. 575. 10.3390/app10020575.

Kim, Myeongseop & Han, Dong-Ki & Park, Jae-Han & Kim, Jung-Su. (2020). Motion Planning of Robot Manipulators for a Smoother Path Using a Twin Delayed Deep Deterministic Policy Gradient with Hindsight Experience Replay

Applied Sciences. 10. 575. 10.3390/app10020575.

Xinrui Shen . Comparison of DDPG and TD3 Algorithms in a Walker2D Scenario. Proceedings of the 2023 International Conference on Data Science, Advanced Algorithm and Intelligent Computing (DAI 2023). Atlantis Press.D 2024.8 2024-02-13T23:00:00.000Z.P 148-155.. .https://doi.org/10.2991/978-94-6463-370-2_17 10.2991/978-94-6463-370-2_17

Pastor, J.M., Díaz, H., Armesto, L., Sala, A. Aprendizaje por refuerzo con búsqueda de políticas: simulación y aplicación a un sistema electromecánico. 7, 8 y 9 de septiembre de 2016, Madrid (pp. 710-717). DOI capítulo: https://doi.org/10.17979/spudc.9788497498081.0710 DOI libro: https://doi.org/10.17979/spudc.9788497498081

José Javier Gutiérrez Gil (jogugil@alumni.uv.es/ jogugil@gmail.com)