

### Introducció

Motivación y descripción de

#### Marco teório

Aprendizaje po Refuerzo

Redes Generativas Adversarias (GANs)

Aprendizaje po Imitación Generativo Adversario

Hibridación Q-learning (HQL)

Entornos GYM OpenA

Conclusiones y vías futuras

## Uso de Modelos Generativos en Aprendizaje por Refuerzo

Silvia Barroso Moreno silviabm98@ugr.es

Directores: Juan Gómez Romero y Miguel Molina Solana Departamento Ciencia de la Computación e Inteligencia Artificial Trabajo Fin de Máster: Ciencia de Datos e Ingeniería de Computadores 2022-2023







## Índice

### Introducción Motivación y

problema

### iviaico teoric

Refuerzo

Redes Generativas Adversarias (GANs)

Imitación Generativo Adversario (GAIL)

Hibridación Q-learning (HQL)

Experimentación
Entornos GYM OpenAl

- Introducción
   Motivación y descripción del problema
- 2 Marco teórico Aprendizaje por Refuerzo Redes Generativas Adversarias (GANs)
- 3 Aprendizaje por Imitación Generativo Adversario (GAIL)
- 4 Hibridación Q-learning (HQL)
- **5** Experimentación Entornos GYM OpenAl Entornos Sinergym
- 6 Conclusiones y vías futuras



## Índice

### Introducción Motivación y

problema

Marco teorio

Refuerzo
Redes Generativas

Adversarias (G

Aprendizaje Imitación Generativo Adversario

Hibridación Q-learning (HQL)

Experimentación Entornos GYM OpenAl Entornos Sinergym

- Introducción Motivación y descripción del problema
- 2 Marco teórico Aprendizaje por Refuerzo Redes Generativas Adversarias (GANs
- 3 Aprendizaje por Imitación Generativo Adversario (GAIL)
- 4 Hibridación Q-learning (HQL)
- Experimentación Entornos GYM OpenAl Entornos Sinergym
- 6 Conclusiones y vías futuras



Motivación y descripción del

problema

Marco teoric

Refuerzo

Redes Generativas Adversarias (GANs)

Aprendizaje pol Imitación Generativo Adversario (GAIL)

Hibridación Q-learning (HQL)

Experimentación

Entornos GYM Opena

Conclusiones y vías futuras  Objetivo: establecer conexión entre modelos generativos y aprendizaje por refuerzo



Motivación y descripción del problema

Widico ccome

Refuerzo

Redes Generativas Adversarias (GANs)

Aprendizaje po Imitación Generativo Adversario

Hibridación Q-learning (HQL)

Experimentación

Entornos Sinergym

- Objetivo: establecer conexión entre modelos generativos y aprendizaje por refuerzo
- Conexión:



- Motivación y descripción del problema
- Marco teorio
- Refuerzo
- Redes Generativas
- Aprendizaje po Imitación Generativo
- Hibridació Q-learning (HQL)
- Experimentació
- Entornos Sinergym
- Conclusiones y vías futuras

- Objetivo: establecer conexión entre modelos generativos y aprendizaje por refuerzo
- Conexión:
  - Modelo generativo ← GANs



- Motivación y descripción del problema
- Marco teório
- Aprendizaje po
- Redes Generativas Adversarias (GANs)
- Aprendizaje po Imitación
- Imitación Generativo Adversario (GAIL)
- Hibridación Q-learning (HQL)
- Experimentació
- Entornos GYM OpenA
- Conclusiones y vías futuras

- Objetivo: establecer conexión entre modelos generativos y aprendizaje por refuerzo
- Conexión:
  - Modelo generativo ← GANs
  - ② Aprendizaje por refuerzo ← Aprendizaje por Imitación



- Motivación y descripción del problema
- Marco teório
- Aprendizaje p
- Redes Generativas Adversarias (GANs)
- Aprendizaje po Imitación Generativo Adversario
- Hibridación Q-learning (HQL)
- Experimentació
- Entornos Sinergym
- Conclusiones y vías futuras

- Objetivo: establecer conexión entre modelos generativos y aprendizaje por refuerzo
- Conexión:
  - Modelo generativo ← GANs
  - 2 Aprendizaje por refuerzo  $\leftarrow$  Aprendizaje por Imitación
- Aprendizaje por Imitación: el agente observa e imita el comportamiento del EXPERTO. NO tiene acceso al entorno NI a la recompensa.



### Motivación y descripción del

problema Marco teóric

Aprendizaje por

Refuerzo

Redes Generativas Adversarias (GANs)

Aprendizaje p Imitación Generativo Adversario

Hibridación Q-learning (HQL)

Entornos GYM OpenA

Conclusiones y

- Objetivo: establecer conexión entre modelos generativos y aprendizaje por refuerzo
- Conexión:
  - Modelo generativo ← GANs
  - $oldsymbol{2}$  Aprendizaje por Imitación
- Aprendizaje por Imitación: el agente observa e imita el comportamiento del EXPERTO. NO tiene acceso al entorno NI a la recompensa.
  - Aprendizaje por Imitación Generativo Adversario (GAIL)



- Motivación y descripción del problema
- Marco teóric
- Aprendizaje p
- Redes Generativas Adversarias (GANs)
- Imitación
  Generativo
  Adversario
- Hibridación Q-learning (HQL)
- Entornos GYM OpenA
- Conclusiones y

- Objetivo: establecer conexión entre modelos generativos y aprendizaje por refuerzo
- Conexión:
  - Modelo generativo ← GANs
  - $oldsymbol{2}$  Aprendizaje por Imitación
- Aprendizaje por Imitación: el agente observa e imita el comportamiento del EXPERTO. NO tiene acceso al entorno NI a la recompensa.
  - 1 Aprendizaje por Imitación Generativo Adversario (GAIL)
  - **②** Hibridación Q-Learning (HQL) → nueva propuesta



- Motivación y descripción del problema
- Marco teóric
- Aprendizaje po
- Redes Generativas Adversarias (GANs)
- Imitación Generativo Adversario (GAIL)
- Hibridación Q-learning (HQL)
- Experimentación Entornos GYM OpenAl Entornos Sinergym
- Conclusiones y

- Objetivo: establecer conexión entre modelos generativos y aprendizaje por refuerzo
- Conexión:
  - Modelo generativo ← GANs
  - 2 Aprendizaje por refuerzo  $\leftarrow$  Aprendizaje por Imitación
- Aprendizaje por Imitación: el agente observa e imita el comportamiento del EXPERTO. NO tiene acceso al entorno NI a la recompensa.
  - 1 Aprendizaje por Imitación Generativo Adversario (GAIL)
  - ② Hibridación Q-Learning (HQL) → nueva propuesta
- Experimentación



- Motivación y descripción del problema
- Marco teóric
- Aprendizaje po
- Redes Generativas Adversarias (GANs)
- Imitación
  Generativo
  Adversario
- Hibridación Q-learning (HQL)
- Experimentación Entornos GYM OpenAl Entornos Sinergym
- Conclusiones y

- Objetivo: establecer conexión entre modelos generativos y aprendizaje por refuerzo
- Conexión:
  - ① Modelo generativo ← GANs
  - 2 Aprendizaje por refuerzo  $\leftarrow$  Aprendizaje por Imitación
- Aprendizaje por Imitación: el agente observa e imita el comportamiento del EXPERTO. NO tiene acceso al entorno NI a la recompensa.
  - 1 Aprendizaje por Imitación Generativo Adversario (GAIL)
  - ② Hibridación Q-Learning (HQL) → nueva propuesta
- Experimentación
  - **1** GYM OpenAI → Taxi y CartPole



### Motivación y descripción del

problema Marco teório

Aprendizaje por

Redes Generativas

Adversarias (GANs)

Imitación Generativo Adversario (GAIL)

Hibridación Q-learning (HQL)

Experimentación Entornos GYM OpenAl Entornos Sinergym

- Objetivo: establecer conexión entre modelos generativos y aprendizaje por refuerzo
- Conexión:
  - ① Modelo generativo ← GANs
  - 2 Aprendizaje por refuerzo  $\leftarrow$  Aprendizaje por Imitación
- Aprendizaje por Imitación: el agente observa e imita el comportamiento del EXPERTO. NO tiene acceso al entorno NI a la recompensa.
  - 1 Aprendizaje por Imitación Generativo Adversario (GAIL)
  - ② Hibridación Q-Learning (HQL) → nueva propuesta
- Experimentación
  - **1** GYM OpenAI → Taxi y CartPole
  - ${f 2}$  Sinergym ightarrow 5Zone, Datacenter y Warehouse



Motivación y descripción del

problema Marco teório

Aprendizaje po

Redes Generativas Adversarias (GANs)

Aprendizaje po

Imitación Generativo Adversario (GAIL)

Hibridación Q-learning (HQL)

Entornos GYM OpenAl Entornos Sinergym

- Objetivo: establecer conexión entre modelos generativos y aprendizaje por refuerzo
- Conexión:
  - Modelo generativo ← GANs
  - 2 Aprendizaje por refuerzo  $\leftarrow$  Aprendizaje por Imitación
- Aprendizaje por Imitación: el agente observa e imita el comportamiento del EXPERTO. NO tiene acceso al entorno NI a la recompensa.
  - 1 Aprendizaje por Imitación Generativo Adversario (GAIL)
  - ② Hibridación Q-Learning (HQL) → nueva propuesta
- Experimentación
  - **1** GYM OpenAI → Taxi y CartPole
  - 2 Sinergym  $\rightarrow$  5Zone, Datacenter y Warehouse
  - 3 Proyecto de investigación → IA4TES



## Índice

## Introducción v

### Marco teórico

Aprendizaje po Refuerzo Redes Generati

Redes Generativ Adversarias (GA

Imitación Generativo Adversario (GAIL)

Hibridación Q-learning (HQL)

Entornos GYM OpenAl Entornos Sinergym

- 1 Introducción Motivación y descripción del problema
- Marco teórico Aprendizaje por Refuerzo Redes Generativas Adversarias (GANs)
- 3 Aprendizaje por Imitación Generativo Adversario (GAIL)
- 4 Hibridación Q-learning (HQL)
- **5** Experimentación Entornos GYM OpenAl Entornos Sinergym
- 6 Conclusiones y vías futuras



• Espacio de estados  ${\cal S}$  y espacio de acciones  ${\cal A}$ ,  $\Pi=\{\pi:{\cal S}\to{\cal A}\}$ 

### Motivación v

descripción del problema

#### Marco teórico

#### Aprendizaje por Refuerzo

Redes Generativas Adversarias (GANs)

Aprendizaje por Imitación Generativo Adversario (GAIL)

Hibridación Q-learning (HQL)

## Entornos GYM OpenA

Entornos Sinergym

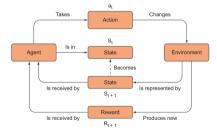


Figura: Funcionamiento de RL



Introducción Motivación v Espacio de estados S y espacio de acciones A, Π = {π : S → A}
 La política se define como el conjunto de reglas que establece el mar

#### Marco teório

### Aprendizaje por Refuerzo

Redes Generativas Adversarias (GANs)

Aprendizaje po Imitación Generativo Adversario (GAIL)

Hibridación Q-learning (HQL)

## Entornos GYM OpenA

Conclusiones y

La política se define como el conjunto de reglas que establece el mapeo de situaciones o estados del entorno a las acciones que el
agente debe tomar con el fin de maximizar las recompensas a lo largo del tiempo.

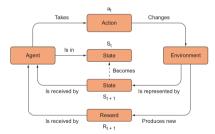


Figura: Funcionamiento de RL



- Espacio de estados  ${\cal S}$  y espacio de acciones  ${\cal A},\,\Pi=\{\pi:{\cal S}\to{\cal A}\}$ 
  - La política se define como el conjunto de reglas que establece el mapeo de situaciones o estados del entorno a las acciones que el
    agente debe tomar con el fin de maximizar las recompensas a lo largo del tiempo.
  - Señal de recompensa en el paso t: R<sub>t</sub>

Takes Changes Action Is in State Environment Agent Becomes State Is received by Is represented by St + 1 Reward Is received by Produces new R<sub>t+1</sub>

Figura: Funcionamiento de RL

### Introducción Motivación v

Marca toório

#### iviarco teorici

Aprendizaje por Refuerzo

Redes Generativas Adversarias (GANs)

Aprendizaje po Imitación Generativo Adversario (GAIL)

Hibridación Q-learning (HQL)

Experimentacion Entornos GYM OpenAl

Conclusiones y



- Motivación v
- Espacio de estados S y espacio de acciones A.  $\Pi = \{\pi : S \to A\}$ La política se define como el conjunto de reglas que establece el mapeo de situaciones o estados del entorno a las acciones que el agente debe tomar con el fin de maximizar las recompensas a lo largo del tiempo.
  - Señal de recompensa en el paso t:  $R_t$ 
    - Función valor:  $V_{\pi}(S) = \mathbb{E}_{\pi}[G_t|S_t = S]$ ,  $G_t = R_1 + R_2 + \ldots + R_t$

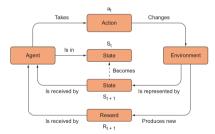


Figura: Funcionamiento de RL

Adversarias (GANs)

Aprendizaje por Refuerzo Redes Generativas



Introducción Motivación y

• Espacio de estados  ${\cal S}$  y espacio de acciones  ${\cal A}$ ,  $\Pi=\{\pi:{\cal S}\to{\cal A}\}$ 

problema

La política se define como el conjunto de reglas que establece el mapeo de situaciones o estados del entorno a las acciones que el
agente debe tomar con el fin de maximizar las recompensas a lo largo del tiempo.

Marco teórico

Señal de recompensa en el paso t: Rt

• Función valor:  $V_{\pi}(S) = \mathbb{E}_{\pi}[G_t|S_t = S]$ ,  $G_t = R_1 + R_2 + \ldots + R_t$ 

Aprendizaje por Refuerzo Redes Generativas Adversarias (GANs) • Función acción-valor:  $Q_{\pi}(S,A) = \mathbb{E}_{\pi}[G_t|S_t = S,A_t = A]$ 

Aprendizaje po Imitación Generativo Adversario (GAIL)

Changes Takes Action Is in State Environment Agent Becomes State Is received by Is represented by St + 1 Reward Is received by Produces new R<sub>1+1</sub>

Q-learning (HQL)

Figura: Funcionamiento de RL

Entornos GYM OpenA

Conclusiones



## Métodos para soluciones tabulares

### Introducción Motivación v

problema

Marco teóric

#### Aprendizaje por Refuerzo

Redes Generativas Adversarias (GANs

Aprendizaje po Imitación Generativo Adversario

Hibridación Q-learning (HQL)

### Experimentació

Entornos Sinergym

Conclusiones y vías futuras

## Métodos para soluciones tabulares: Algoritmo Q-Learning

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma \max_{a} Q(S_{t+1}, a) - Q(S_t, A_t)]$$

Q-Table So S1 S2 S3 S4

$$Q(s, a) \rightarrow Q(3, 1) \rightarrow a_1 +2.53 +7.44 +3.34 +5.31 +6.22 \rightarrow +5.31$$

Figura: Acceso a la tabla Q(S,A)



## Métodos para soluciones aproximadas

# Métodos para soluciones aproximadas: Proximal Policy Optimixation (PPO)

- Incorpora una red neuronal para realizar la aproximación
- Corresponden a métodos que calculan gradientes de políticas

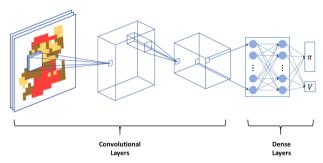


Figura: Ejemplo PPO: aprender a jugar a Mario Bros

### Introducción Motivación v

descripcion dei

#### iviarco teorico

#### Aprendizaje por Refuerzo

Redes Generativas Adversarias (GANs)

Aprendizaje por Imitación Generativo Adversario (GAIL)

Hibridación Q-learning (HQL)

Entornos GYM OpenAl



## Aprendizaje por Refuerzo Profundo (DRL)

Aprendizaje por Refuerzo

Q-learning

La estimación de la función ventaja en el timestep t,  $\bar{A}_t$ , se define como

$$\bar{A}_t = \delta_t + (\gamma \lambda) \delta_{t+1} + \ldots + (\gamma \lambda)^{T-t+1} \delta_{T-1} \text{ donde } \delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$$

Recordemos, la función ventaja se define como  $A_t := Q_t(s, a) - V_t(s)$ 

## Nuestra **función de pérdida**:

$$L^{CPI}(\theta) = \mathbb{E}_t[\min(r_t(\theta)\bar{A}_t, clip(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\bar{A}_t)]$$

## Algoritmo PPO

for 
$$i = 1$$
 to  $M$  do
for  $i = 1$  to  $N$  do

- Ejecutar la política  $\pi_{\theta \, Old}$  en el entorno con T timesteps
- Calcular las estimaciones de la función ventaias  $\bar{A}_1, \bar{A}_2, \dots, \bar{A}_T$

end

- Optimizar el obietivo clipped surrogated  $L^{CPI}$ , con K épocas y tamaño de minibatch M < NT
- $\theta_{OH} \rightarrow \theta$

end



## Redes Generativas Adversarias (GANs)

### Motivación y

problema

### Marco teório

Aprendizaje po Refuerzo

### Redes Generativas Adversarias (GANs)

Aprendizaje por Imitación Generativo Adversario (GAII.)

Hibridación Q-learning (HQL)

### Experimentación

Entornos Sinergym

Conclusiones y vías futuras

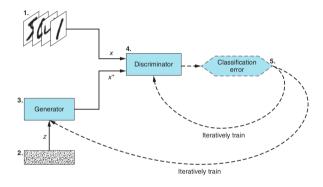


Figura: Funcionamiento de una GANs

TFG: Redes Generativas Adversarias para la creación de deepfakes:

https://github.com/silviabm98/TFG



## Índice

## Motivación v

Marca taária

#### Wareo teorie

Refuerzo

Adversarias (GA

### Aprendizaje por Imitación Generativo Adversario (GAIL)

Hibridación Q-learning (HQL)

## Experimentacio

Entornos Sinergym

- 1 Introducción Motivación y descripción del problema
- 2 Marco teórico Aprendizaje por Refuerzo Redes Generativas Adversarias (GANs
- 3 Aprendizaje por Imitación Generativo Adversario (GAIL)
- 4 Hibridación Q-learning (HQL)
- Experimentación
  Entornos GYM OpenAl
  Entornos Sinergym
- 6 Conclusiones y vías futuras



## Aprendizaje por Imitación Generativo Adversario (GAIL)

### Introducción Motivación y

### problema

#### Marco teorio

Aprendizaje po Refuerzo

Redes Generativ Adversarias (GA

Aprendizaje por Imitación Generativo Adversario

### Hibridación Q-learning (HQL)

(GAIL)

Entornos GYM OpenA

Conclusiones y vías futuras

## Descripción GAIL

- ullet Aprendizaje de una política  $\pi$  (generador) y de un discriminador D
- El generador trata de imitar a la política experta,  $\pi_E$ , de la secuencia [s, a], generando una secuencia falsa  $[s, a]^*$  con la política  $\pi$

### **Definimos**

$$RL \circ IRL_{\psi}(\pi_E) = \operatorname{arg\,min}_{\pi \in \Pi}(-H(\pi) + \psi^*(
ho_{\pi} - 
ho_{\pi_E}))$$

- **1** siendo  $\rho_{\pi}$  la **medida de ocupación** de la política  $\pi \in \Pi$ , definida como  $\rho_{\pi}(s, a) = \pi(a|s) \sum_{t=0}^{\infty} \gamma^{t} P(s_{t} = s|\pi)$
- 3  $\psi^*$  es la conjugada convexa de  $\psi$



## Aprendizaje por Imitación Generativo Adversario (GAIL)

### Introducción Motivación v

### problema

#### Marco teorio

Aprendizaje po

Redes Generativas Adversarias (GANs)

### Aprendizaje por Imitación Generativo Adversario (GAIL)

Q-learning (HQL)

## Entornos GYM OpenA

Entornos Sinergym

Conclusiones y vías futuras

## Conexión entre aprendizaje por imitación y GANs

$$\psi_{\textit{GA}} \triangleq \left\{ egin{array}{ll} \mathbb{E}_{\pi_{\textit{E}}}[g(c(s,a))] & \textit{si} & c \leq 0 \\ +\infty & \text{en otro caso} \end{array} 
ight.$$

donde

$$g(x) = \begin{cases} -x - \log(1 - \exp x) & si & x \le 0 \\ +\infty & \text{en otro caso} \end{cases}$$
 (2)



**OBJETIVO**: Encontrar un punto de silla  $(\pi, D)$  en la siguiente expresión

$$\mathbb{E}_{\pi}[\log(D(s,a))] + \mathbb{E}_{\pi_{E}}[\log(1-D(s,a))] - \lambda H(\pi)$$



## **Algoritmo GAIL**

Introducció

Motivación y

Marco teório

Aprendizaje po

Redes Generativas Adversarias (GANs)

Aprendizaje por Imitación Generativo Adversario (GAIL)

Hibridación Q-learning (HQL)

Experimentación Entornos GYM OpenAl Entornos Sinergym

Conclusiones y vías futuras

### Algoritmo GAIL:

Para cada trayectoria *i*:

for i = 1 to N do

**1** Actualizamos los parámetros del Discriminador de  $\omega_i$  a  $\omega_{i+1}$  con el gradiente:

$$\mathbb{E}_{\pi_i}[
abla_{\omega}\log(D_{\omega}(s,a))] + \mathbb{E}_{\pi_E}[
abla_{\omega}\log(1-D_{\omega}(s,a))]$$

2 Tomamos la política  $\pi_{\theta_i}$  y actualizamos la política  $\pi_{\theta_{i+1}}$  utilizando PPO con su función de coste. Realizamos la actualización del gradiente:

$$\mathbb{E}_{\pi_i}[\nabla_{\theta}\log \pi_{\theta}(a|s)\mathcal{Q}(s,a)] - \lambda \nabla_{\theta}H(\pi_{\theta})$$

donde 
$$\mathcal{Q}(ar{s},ar{a})=\mathbb{E}_{ heta_i}[\log(D_{\omega_{i+1}}(s,a))|s_0=ar{s},a_0=ar{a}]$$

end



## Índice

## Motivación v

Marco toório

TVILLICO CCOTTO

Refuerzo Redes Generativas

Adversarias (GANs)

Imitación Generativo Adversario

Hibridación Q-learning (HQL)

Entornos GYM OpenAl Entornos Sinergym

- 1 Introducción Motivación y descripción del problema
- 2 Marco teórico Aprendizaje por Refuerzo Redes Generativas Adversarias (GANs
- 3 Aprendizaje por Imitación Generativo Adversario (GAIL)
- 4 Hibridación Q-learning (HQL)
- Entornos GYM OpenAl
- 6 Conclusiones y vías futuras



## Hibridación Q-learning (HQL)

### Introducció

Motivación y descripción del

### Marco teori

Aprendizaje po Refuerzo

Redes Generativas Adversarias (GANs)

Imitación Generativo Adversario

Hibridación Q-learning (HQL)

Experimentación Entornos GYM OpenAl Entornos Sinergym

Conclusiones y vías futuras

### Descripción HQL

- **1** Aprendizaje de una política  $\pi$  (generador) y de un discriminador D
- 2 El generador trata de imitar a la política experta,  $\pi_E$ , de la tabla Q[s,a], generando una tabla falsa  $Q[s,a]^*$  con la política  $\pi$

### Base de datos experta:

$$Q(S,A) = \{Q(S,A,1), Q(S,A,2), \dots Q(S,A,n)\}, \ \forall (S,A) \in \mathcal{S} \times \mathcal{A}$$

### Vanilla GAN

 $\min_{G} \max_{D} V(D,G) = \min_{Q^*} \max_{D} V(D,Q^*)$ 

$$= \min_{Q^*} \max_{D} (\mathbb{E}_{Q(S,A) \sim P_Q}[\log D(Q(S,A))] + \mathbb{E}_{Q^*(S,A) \sim P_{Q^*}}[\log(1 - D(Q^*(S,A)))])$$
(3)



## Hibridación Q-learning (HQL)

### Introducción Motivación v

descripción del problema

#### Marco teórico

Aprendizaje por

Redes Generativas Adversarias (GANs)

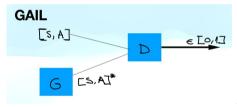
Aprendizaje por Imitación Generativo Adversario

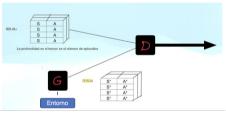
### Hibridación Q-learning (HQL)

Experimentación

Entornos GYM OpenAl

Entornos Sinergym







## Índice

## Introducción v

### Maura tafula

#### Marco teorio

Aprendizaje por Refuerzo Redes Generativas

Adversarias (G

Imitación Generativo Adversario

Hibridación Q-learning (HQL)

### Experimentación

Entornos Sinergym

- 1 Introducción Motivación y descripción del problema
- Marco teórico Aprendizaje por Refuerzo Redes Generativas Adversarias (GANs
- 3 Aprendizaje por Imitación Generativo Adversario (GAIL)
- 4 Hibridación Q-learning (HQL)
- **5** Experimentación Entornos GYM OpenAl Entornos Sinergym
- 6 Conclusiones y vías futuras



## Experimentación

#### Introducció

Motivación y descripción del problema

#### Marco teório

Aprendizaje por Refuerzo

Redes Generativas Adversarias (GANs)

Aprendizaje por Imitación Generativo

Generativo Adversario (GAIL)

Hibridación Q-learning (HQL)

Experimentacio

Entornos GYM OpenAl Entornos Sinergym

	Observacciones		Acciones		$oxed{Algoritmo}$	
Entornos	Discreto	Continuo	Discreto	Continuo	GAIL	$_{ m HQL}$
Taxi - Gym	×		×		×	×
CartPole - Gym		×	×		×	
5Zone - Sinergym		×	×		×	
Datacenter - Sinergym		×	×		×	
Warehouse - Sinergym		×	×		×	



## **Entornos GYM OpenAl**

#### Introducción

Motivación y descripción del

### Marco teório

Aprendizaje por

Redes Generativas Adversarias (GANs)

Aprendizaje por Imitación Generativo Adversario

Hibridación Q-learning (HQL)

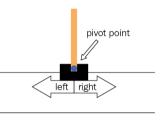
#### Experimentació

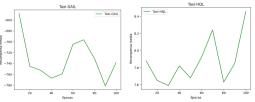
Entornos GYM OpenAl

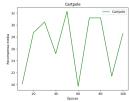
Conclusiones y vías futuras



## **CartPole**









## **Entornos Sinergym**

#### Introducción

Motivación y descripción del problema

#### Marco teórico

Aprendizaje por

Redes Generativas Adversarias (GANs)

Aprendizaje por Imitación Generativo Adversario (GAIL)

Hibridación Q-learning (HQL)

### Experimentació

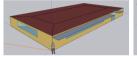
Entornos GYM Open Entornos Sinergym

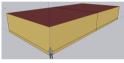
Conclusiones y vías futuras

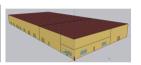
### 5Zone

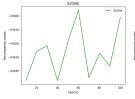
## Datacenter

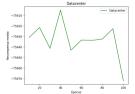
### Warehouse

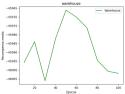














## Comparativa

Introducción

Motivación y descripción del problema

Marco teori Aprendizaje por Refuerzo

Redes Generativas Adversarias (GANs)

Aprendizaje por Imitación Generativo Adversario (GAIL)

Hibridación Q-learning (HQL)

Entornos GYM Open

Entornos Sinergym

	GAIL	GAIL	GAIL	GAIL	GAIL	$_{ m HQL}$
Épocas	5ZONE	DATACENTER	WAREHOUSE	CARTPOLE	TAXI	TAXI
10	-43646.9	-75631.07	-45995.8	20.1	-648.0	7.88
20	-43625.9	-75621.6	-45984.0	28.7	-745.4	7.65
30	-43621.8	-75641.4	-46006.1	30.5	-752.6	7.59
40	-43647.3	-75605	-45982.3	25.2	-767.0	7.82
50	-43618.7	-75643.09	-45966.1	32.3	-759.5	7.68
60	-43595.8	-75633.5	-45970.1	19.7	-705.0	7.92
70	-43645.1	-75633.8	-45976.10	31.2	-696.5	8.24
80	-43627.4	-75632.5	-45994.6	31.2	-731.0	7.63
90	-43636.5	-75622.7	-46000.7	21.4	-781.4	7.85
100	-43601.01	-75672.2	-46001.9	28.6	-738.2	8.46



## Índice

## Introducción v

Marco teório

. . .

Refuerzo Redes Generativas

Adversarias (G

Imitación Generativo Adversario

(GAIL)
Hibridación
Q-learning

(HQL)

Experimentación

Enternos GYM OpenAl

Conclusiones y vías futuras

1 Introducción

Motivación v descripción del problema

2 Marco teórico

Aprendizaje por Refuerzo

Redes Generativas Adversarias (GANs

Aprendizaie por Imitación Generativo Adversario (GAIL)

A Hibridación O learning (HOL)

4 Hibridación Q-learning (HQL)

6 Experimentación

Entornos GYM OpenAl

Entornos Sinergym



## Conclusiones y vías futuras

Adversarias (GANs)

Conclusiones v vías futuras

Conexión entre RL y los modelos generativos 

GANs y Aprendizaje por Imitación

Distintos entornos: CartPole, Taxi, 5Zone, Datacenter, Warehouse

LINEA FUTURA: Establecer nueva conexión entre RL y modelos generativo

-Nuevo modelo generativo, por ejemplo Decision Difusser

-Mejorar los experimentos realizados con CartPole, Taxi, 5Zone, Datacenter, Warehouse -Nuevos entornos distintos a CartPole, Taxi, 5Zone, Datacenter, Warehouse...

-Seguir investigando sobre la nueva propuesta Hibridación Q-Learning (HQL)



## Introducción v

Motivación y descripción de problema

#### Marco teórico

Aprendizaje po Refuerzo

Redes Generativas Adversarias (GANs)

### Aprendizaje por Imitación Generativo Adversario

Hibridación Q-learning (HQL)

### Experimentación

Entornos GYM OpenAl Entornos Sinergym

Conclusiones y vías futuras



# ¡GRACIAS POR SU ATENCIÓN!