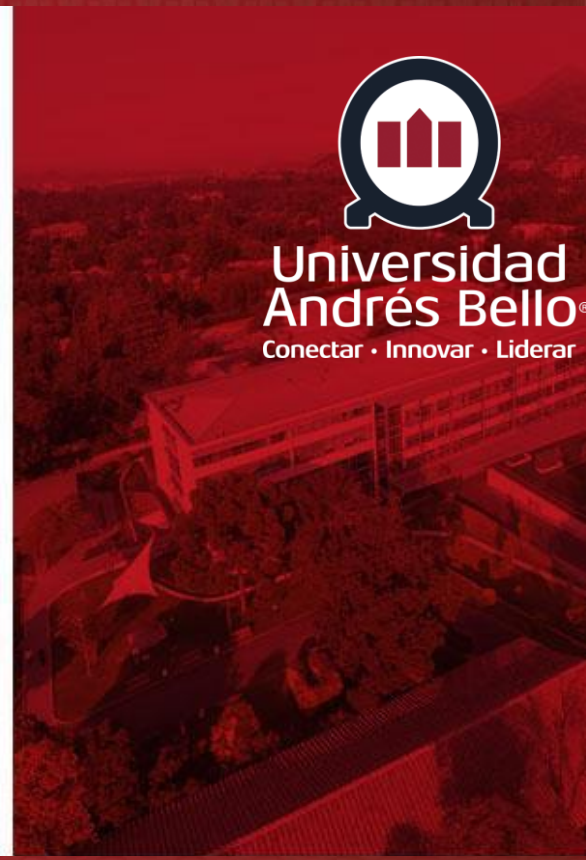
Escuela Internacional de Primavera sobre Entornos Ubicuos y Aplicaciones de Robots Sociales

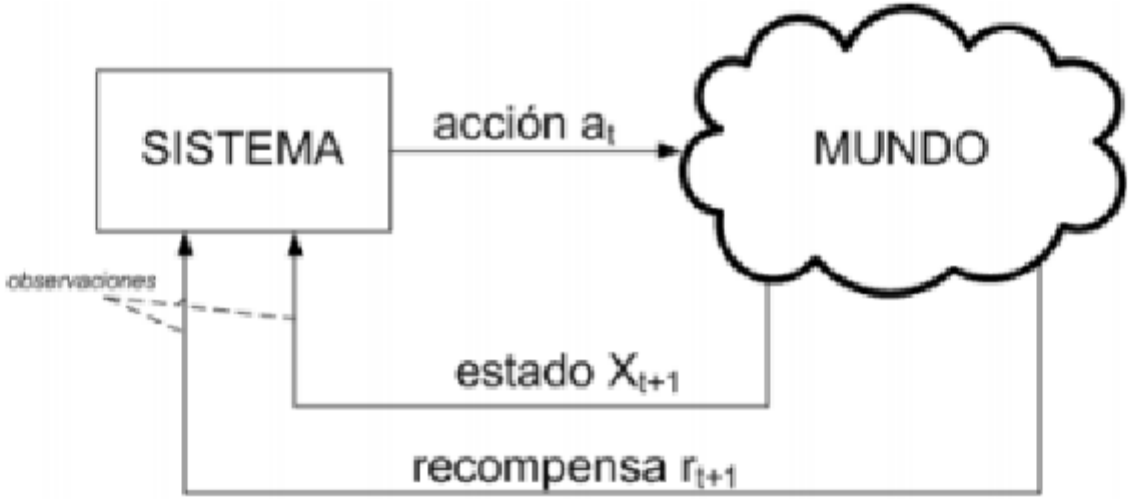


Tutorial #2 – Aprendizaje Reforzado con Gym (material en www.miguelsolis.info)

Dr. Miguel A. Solis
Universidad Andrés Bello
16 de octubre de 2023

Aprendizaje reforzado y conductivismo

El **agente** interactúa con el entorno a través de percepciones y acciones.



- Recibe como entrada (percibe), el estado actual del entorno, *s*.
- Luego, genera una acción (ejecuta) *a* como salida.
- Recibe una señal de refuerzo *r* (recompensa).

Retorno

Función de valor:

$$V^{\pi}(s_k) = r_k + \gamma r_{k+1} + \gamma^2 r_{k+2} + \dots$$
$$= \sum_{i=0}^{\infty} \gamma^i r_{k+i}.$$

Restricciones:

- $0 \leq \gamma < 1$.
- r_k acotado.

Una política π^* es óptima si la función de valor obtenida para esa política es óptima:

$$V^*(s) = V^{\pi^*}(s) \geq V^{\pi}(s) \quad \forall s, \pi$$

Valor Q

- Así como $V(s)$ corresponde a la función que valoriza el estado, $Q(s,a)$ corresponde a la función que valoriza el tomar cierta acción en ese estado.
- Para una política óptima π^*, se cumple

$$Q^{\pi^*}(s,a) \geq Q^{\pi}(s,a) \quad \forall s, a, \pi$$

Q-learning

- Consiste en iterar sobre cada par (estado,acción), para α y γ fijos.
- Regla de actualización:

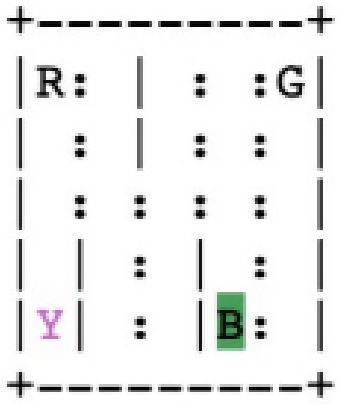
$$Q(s_k, a_k) \leftarrow (1-\alpha)Q(s_k, a_k) + \alpha\left(r_{k+1} + \gamma \max_a Q(s_{k+1,a})\right)$$

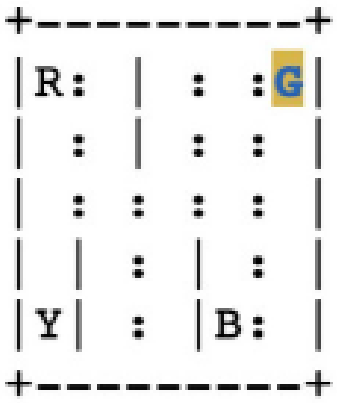
Suponiendo $\hat{Q} = Q^*$, entonces la acción óptima para cada estado se puede obtener maximizando:

$$\pi^*(s_k) = arg\max_{a_k} Q^*(s_k, a_k)$$

Imagen extraída de “Introduction to Q-learning with OpenAI Gym”, Gelana Tostaeva, Medium.

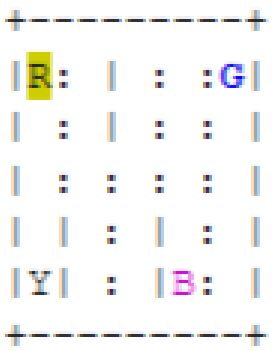
Ejemplo a programar





- R, G, Y, B: ubicaciones
- rectángulo amarillo: taxi vacío
- rectángulo verde: taxi ocupado
- azul: ubicación de pasajero
- morado: ubicación de destino

Acciones



- 0: mover al sur
- 1: mover al norte
- 2: mover al este (derecha)
- 3: mover al oeste (izquierda)
- 4: recoger pasajero
- 5: dejar pasajero

Espacio de estados y acciones

- Acciones: 6
- Estados: 500 (ubicacion_pasajero, ubicacion_taxi, destino)
  - 4 posibles destinos
  - ubicacion_pasajero:
    - 4 ubicaciones para tomar el taxi
    - o pasajero se encuentra en misma ubicación de taxi
  - ubicacion_taxi:
    - 25 ubicaciones posibles según el mapa
- Verificación en Gym:
  - env.action_space
  - env.observation_space

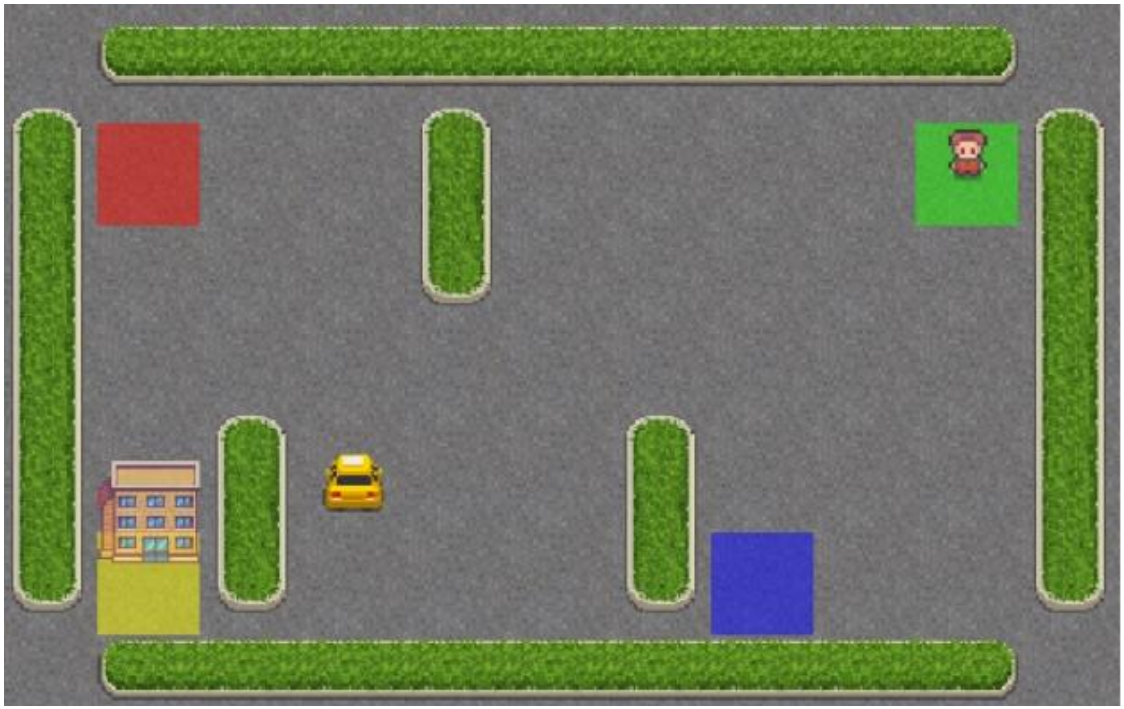
Recompensas

- dejar a pasajero en ubicación correcta: 20 puntos
- descuento de 1 punto cada vez que se mueve con pasajero y no llega a destino
- descuento de 10 puntos por dejar a pasajero en ubicación incorrecta/ilegal

Python Notebook

Universidad Andrés Bello

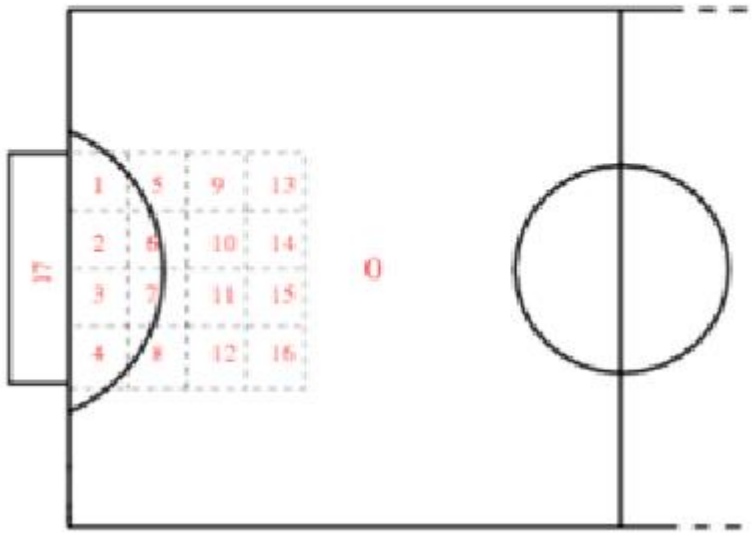
16102023_RL_Parte2.ipynb



Abrir con visor de Python Notebook o colab.research.google.com

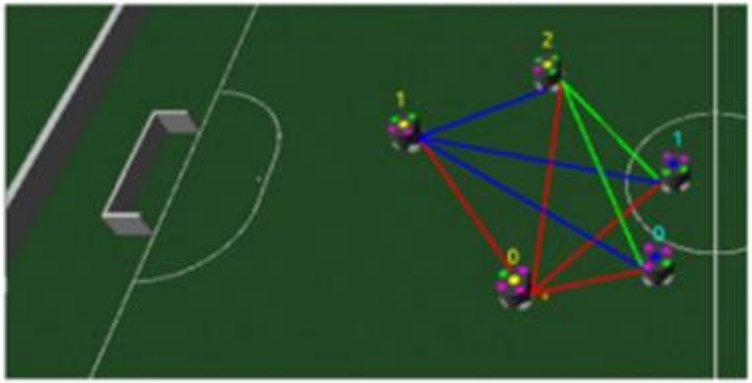
Ejemplos

- G.A. Ahumada, C.J. Nettle and M.A. Solis, 'Accelerating Q-learning through Kalman Filter Estimations applied in a RoboCup SSL Simulation', **Proceedings** of the 10th IEEE Latin American Robotics Symposium, 2013.



Ejemplos

Generación de estrategia defensiva



estado compuesto por:

- dist(K_i,pelota), dist(T_j,pelota)
- dist(K_i,K_j)
- dist(K_i,T_j)
- angle(K_i,T_i)

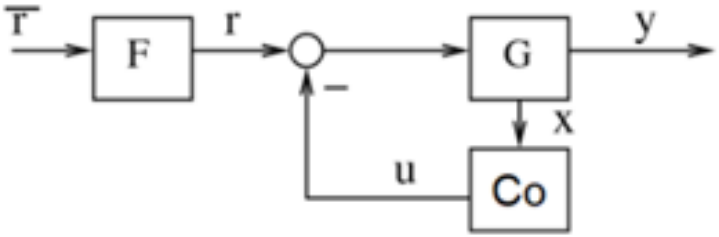
Ollino, F., Solis, M. A., & Allende, H. (2018). Batch reinforcement learning on a RoboCup Small Size League keepaway strategy learning problem. In 4th Congress on Robotics and Neuroscience, CRoNe 2018. CEUR-WS.

Ejemplos

Considere una planta/proceso modelado por:

$$x[k+1] = Ax[k] + Bu[k] + v[k],$$
$$y[k] = Cx[k] + w[k],$$

donde w y v corresponden a ruido de medición y ruido de proceso respectivamente, y considere el siguiente lazo de control



donde el controlador tiene la siguiente estructura:

$$x_c[k+1] = A_c x_c[k] + B_c x[k],$$
$$u[k] = C_c x_c[k] + D_c x[k].$$

Ejemplos

Para estabilizar el sistema, finalmente se determina que la siguiente matriz aumentada debe tener autovalores dentro del círculo unitario:

$$\bar{A} = \begin{bmatrix} A - BD_c & -BC_c \\ B_c & A_c \end{bmatrix}.$$

Con el desempeño J calculado como:

$$J[k] = \xi\{\sum_{i=k}^{\infty} \gamma^{i-k} \left(z^T[i] Q_a z[i] + u^T[i] R u[i]\right)\},$$

donde $z[k] = \begin{bmatrix} (r[k] - y[k]) \\ x_c[k] \end{bmatrix}$, $Q_a = \begin{bmatrix} Q_1 & 0 \\ 0 & Q_2 \end{bmatrix}$,

y Q_1 y Q_2 son matrices de penalización.

M.A. Solis, M. Olivares and H. Allende. **Stabilizing Dynamic State Feedback Controller Synthesis: A Reinforcement Learning Approach**, Studies in Informatics and Control, 2016.

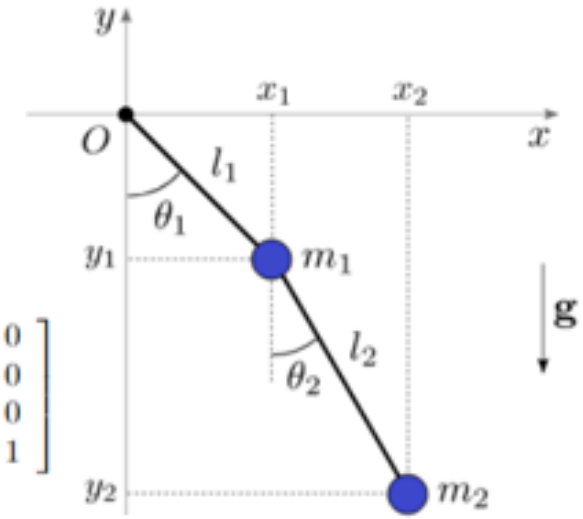
Ejemplos

Universidad Andrés Bello

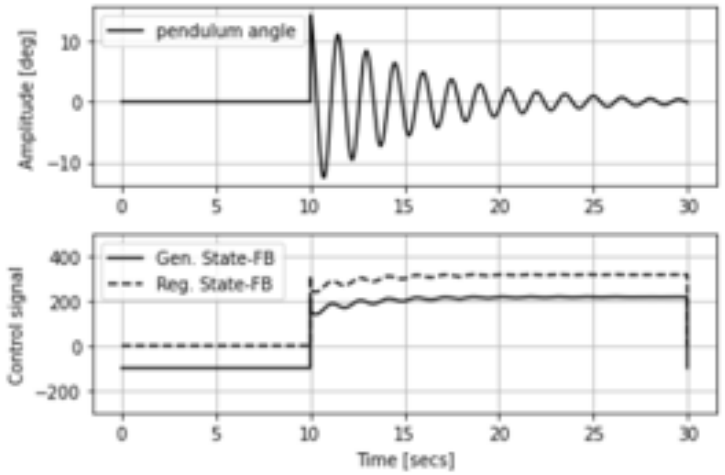
Doble péndulo invertido (simulación)

$$x[k+1] = Ax[k] + Bu[k] + v[k],$$
$$y[k] = Cx[k] + w[k],$$

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -39.69 & -24.5 & 0 & 0 \\ -44.1 & -49 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 4.5 & 2.5 \\ 5 & 5 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$



$$A_c = -1.29,$$
$$B_c = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix},$$
$$C_c = -0.5,$$
$$D_c = \begin{bmatrix} 60.01 & 22.84 & 0.53 & 94.06 \end{bmatrix}.$$

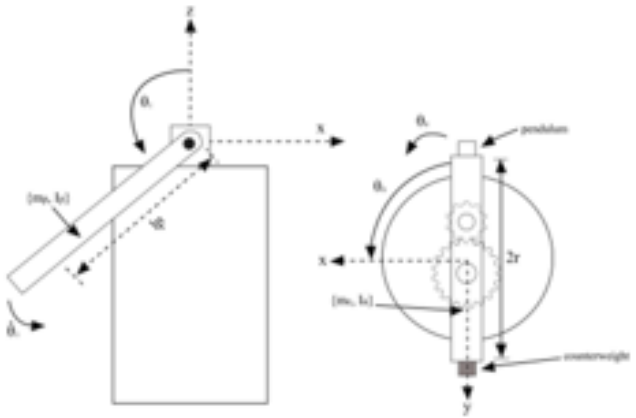


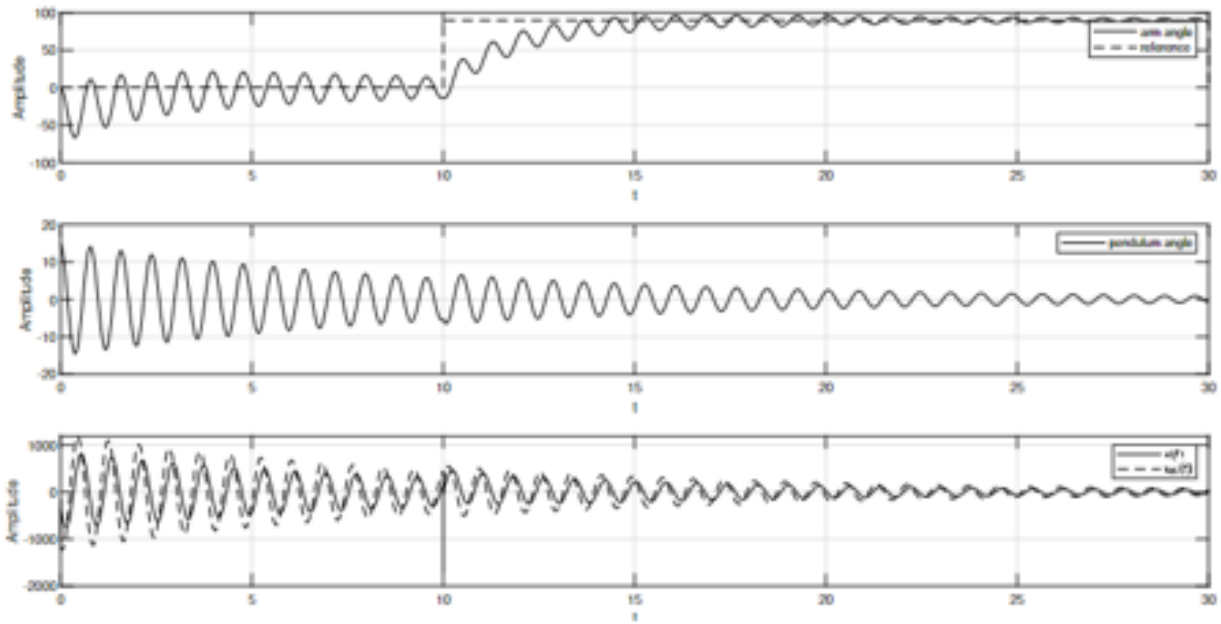
Ejemplos

Universidad Andrés Bello

Péndulo invertido rotatorio (simulación)

$$A_c = -100,$$
$$B_c = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$
$$C_c = -0.5,$$
$$D_c = \begin{bmatrix} -20.96 & -39.76 & 72.74 & 92.61 & -0.58 \end{bmatrix}.$$



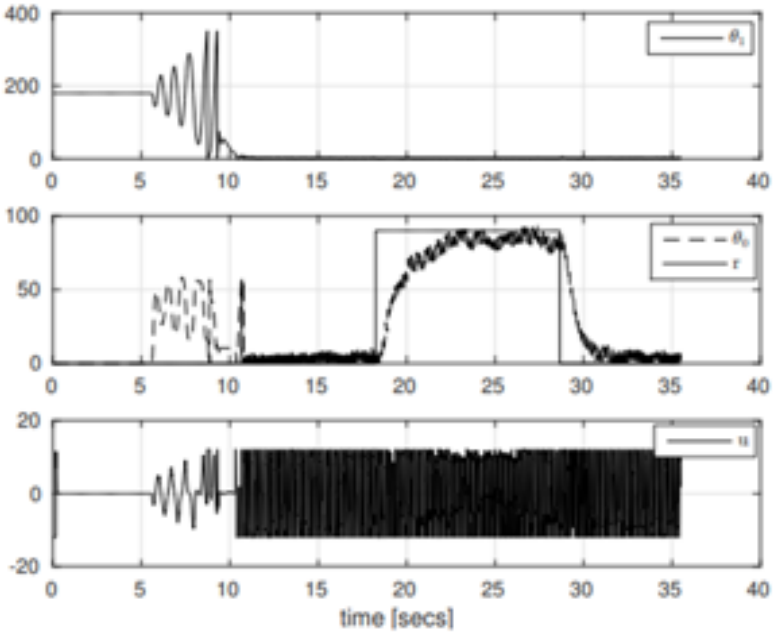


Ejemplos

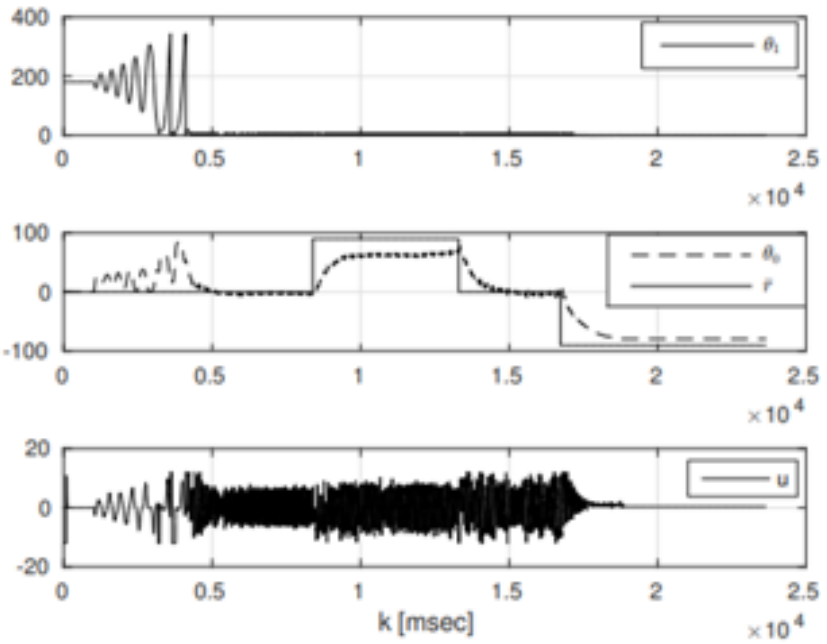
Universidad Andrés Bello

Péndulo invertido rotatorio (real)

Controlador tradicional:

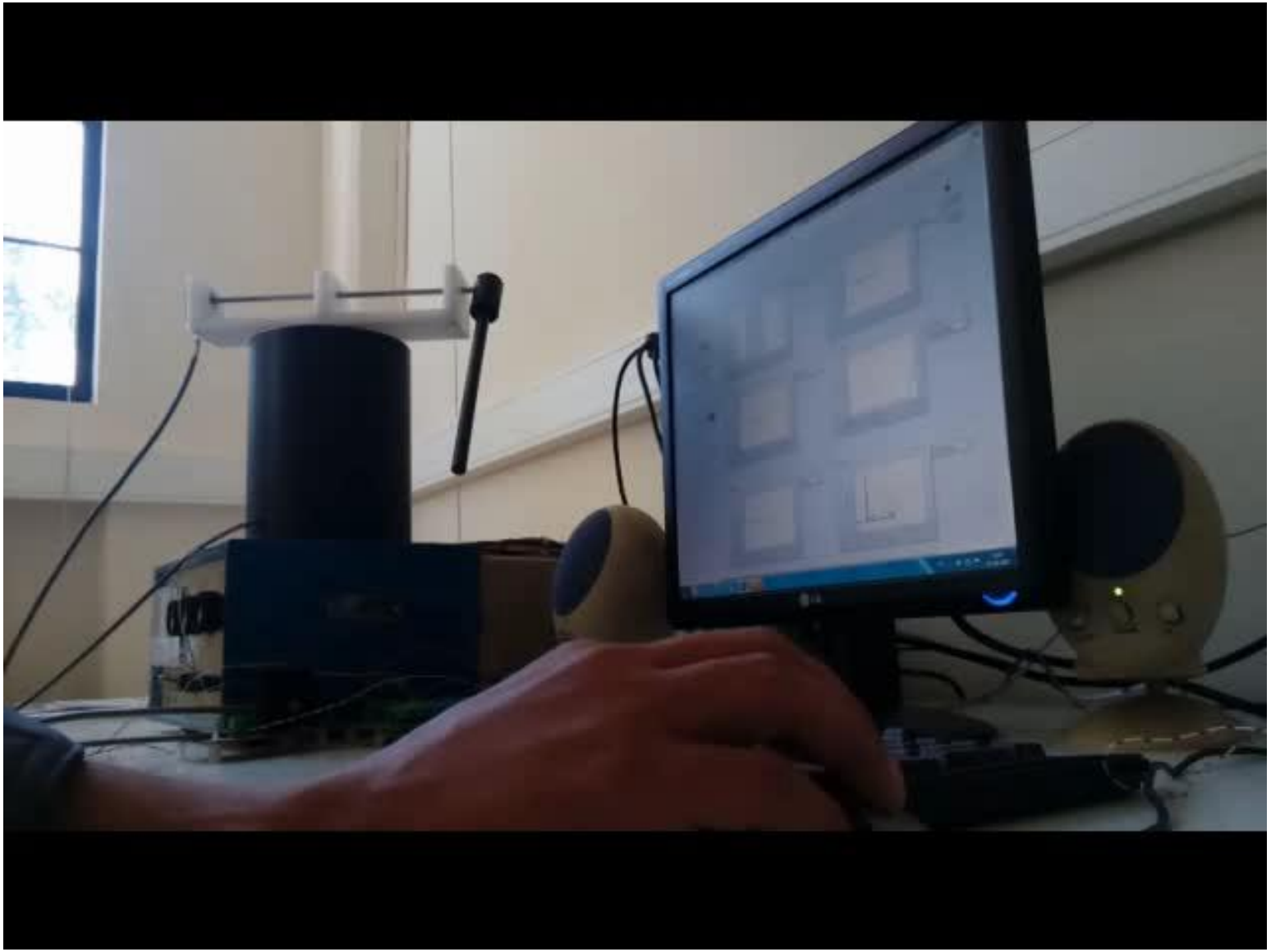


Controlador propuesto:



M.A. Solis, M. Olivares and H. Allende. **A Switched Control Strategy for Swing-Up and State Regulation for the Rotary Inverted Pendulum**, Studies in Informatics and Control, 2019.

Ejemplos



Algunos desafíos

- diseño de recompensas.
- retardo en la ejecución de las acciones.
- representación tabular.

Algunos temas actuales

- affordances

(Objeto, Acción, Efecto)

Dado un objeto y cierta acción, ¿cuál es el efecto?

Dado un objeto y cierto efecto deseado, ¿cuál es la acción requerida?

- continual reinforcement learning (open-ended)

¿Preguntas finales?

Material disponible en www.miguelsolis.info