Instituto Tecnológico de Costa Rica
Escuela de Ingeniería Electrónica
Trabajo Final de Graduación
Proyecto: Método basado en aprendizaje reforzado para el control automático de una planta no lineal.
Estudiante: Oscar Andrés Rojas Fonseca
I Semestre 2024

Firma del asesor

Bitácora de trabajo

Fecha	Actividad	Anotaciones	Horas dedicadas
01/05/2024	**1**. Reunión de seguimiento con el asesor del proyecto.	*a*) Revisión de avance y errores de forma. *b*) Discusión respecto al método de exploración y explotación utilizado en PPO.	2 horas
01/05/2024	**2**. Entrenamiento del modelo PPO y trabajo con la función $ou\_process()$.	*a*) Entrenamientos del modelo; comportamientos indeseados. *b*) Estudio de la función $ou\_process()$ en Octave facilitada por el profesor asesor.	6 horas
02/05/2024	**3**. Adaptación de la función generación de ruido $ou\_process()$ a Python.	*a*) Creación de la función $ou\_process()$ en $ppo_pahm.py$ al adaptarla desde Octave. *b*) Prueba de funcionamiento. Los valores base son funcionales pero requieren un mayor nivel de ruido para mover el péndulo ($sigma \approx 0.7$ funcional). *c*) Adición de la lógica para la disminución del ruido conforme el tiempo de entrenamiento.	8 horas

03/05/2024	**4**. Suma de más componentes a la observación/entrada de la red y entrenamientos PPO.	*a*) Se agregó la aproximación de la velocidad angular del péndulo al obs_n. *b*) Entrenamiento del modelo con la velocidad como entrada. El desempeño de la red mejoró al dejar de *paralizarse* en el proceso. *c*) Se agregó la aproximación de la aceleración angular del péndulo al obs_n. No se logró importante mejoría.	8 horas
04/05/2024	**5**. Replanteo de la función de recompensas $calculate\_reward()$.	*a*) Cambios de los pesos de los componentes del $reward$. *b*) Se probaron diferentes formas de plantear las ecuaciones (positivas, negativas). Los resultados demuestran el mal desempeño.	8 horas
05/05/2024	**6**. Continuación de los cambios en la función de recompensas y entrenamiento para su comprobación.	*a*) Se utilizaron diferentes métodos y ecuaciones que al entrenar el modelo mantienen el mal desempeño.	6 horas
07/05/2024	**7**. Continuación de los cambios en la función de recompensas y entrenamiento para su comprobación.	*a*) Se consultó al asesor respecto a posibles formas de definición de la función. Se continuaron realizando pruebas con resultados indeseados.	6 horas
		Total de horas de trabajo:	44 horas

Contenidos de actividades

La mayor parte del trabajo se enfocó en la definición de la función de recompensas, esto debido a la previa adición de los componentes de velocidad angular y aceleración del péndulo como entradas a la red neuronal, de manera que ya el agente cuenta con suficiente información para interpretar los comportamientos.

Se probaron funciones para "castigar" la corta duración de los episodios del *env*, dado que el comportamiento recurrente del péndulo es empujar lo suficiente para terminar rápido el episodio y no recibir tanto castigo. De manera que se probó con diferentes versiones de una función exponencial (por ejemplo $f(x) = 1.04^{-x+50}$, Figura 1) para castigar lo suficiente al principio del episodio y conforme avanza disminuir el castigo. El resultado fue el mismo, empujar con la se;al PWM hasta terminar el episodio lo más rápido posible, ejemplificado en la Figura 2

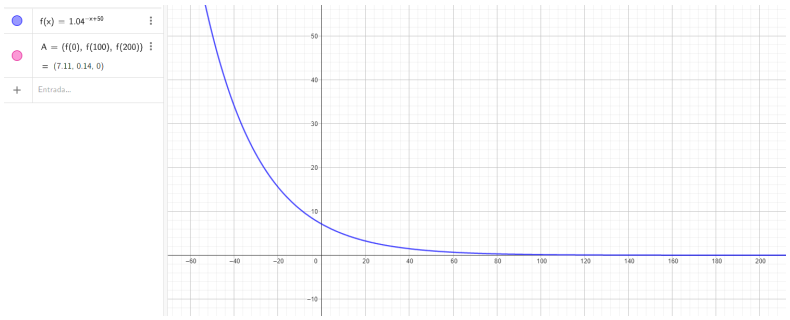

Figure 1: Función utilizada para castigar la corta duración del episodio.

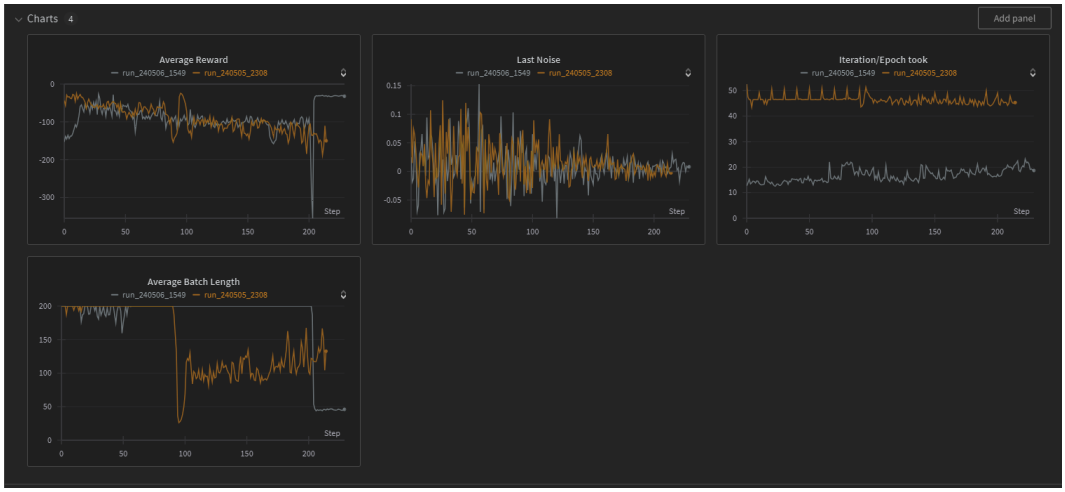

Figure 2: Proceso de entrenamiento del modelo para *PendulumPPO*, resultado insatisfactorio.

Se revisaron algunas de las fuentes ya consultadas como [1] con ningún avance o hallazgo significativo.

Referencias

[1] E. Yang-Yu, "Coding ppo from scratch with pytorch (part 1/4)," *Medium*, 2020.