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
24/04/2024	**1**. Reunión de seguimiento con el asesor del proyecto.	*a*) Revisión de avance y errores de forma. *b*) Se acordó contienuar con el desarrollo de los métodos ajustados para el $PAHM$.	2 horas
26/04/2024	**2**. Pruebas de mejora en la función *get_action*() del método PPO.	*a*) Se discute con el asesor respecto al método de exploración y explotación del código base PPO. *b*) Se decide plantear un cambio escalado del valor inicial de la matriz de covarianza $self.cov\_mat$. *c*) Montaje de lógica para cambio de varianza y pruebas con diferentes variación de la desviación estandar.	8 horas
27/04/2024	**3**. Prueba del montaje del método DQN discretizado para el control del env $PAHM$.	*a*) Adaptación del código probado con *Pendulum* al $PAHM$. Ajuste para utilización de la libreria *argparse*. *b*) Pruebas de entrenamiento del modelo fallidas por problemas en la función *optimize_model*().	6 horas

27/04/2024	**4**. Adaptación del código de los métodos para observar los resultados de los entrenamientos mediante la herramienta *Weights & Biases* (W&B).	*a*) Se estudió la forma de enviar la información a W&B. *b*) Ajuste de forma para mantener el registro por fecha y hora.	4 horas
28/04/2024	**5**. Continuación de pruebas de variación de varianza y división estandar para el PPO del $PAHM$.	*a*) Se crean nuevas versiones de *get_action*() y *evaluate*() con uso diferentes de distribución normal. *b*) Entrenamientos para cada caso de variación. Resultados fallidos.	6 horas
29/04/2024	**6**. Pruebas de entrenamiento del modelo PPO del $PAHM$.	*a*) Se probaron diferentes métodos de distribución normal y variaciones de varianza para la etapa de exploración en el entrenamiento. *b*) Entrenamiento de modelos con cada método con aproximadamente $150,000$ *timesteps*. El aprendizaje se estanca luego de unos $100,000$ *timesteps*.	6 horas
30/04/2024	**7**. Continuación de entrenamientos con cambios en el valor de la varianza para el método PPO.	*a*) Se realizan entrenamientos con *timesteps* cercanos a los $500,000$. *b*) Algunos resultados presentan cualidades prometedoras, pero en su mayoría no son aceptables.	6 horas
		Total de horas de trabajo:	38 horas

Contenidos de actividades

Los resultados de los entrenamientos del modelo PPO y sus métodos se compararon con una nueva referencia $RL - Adventure - 2$ y una anteriormente mencionada como [1].

Con la implementación de W&B en el código, ahora es posible acceder a los resultados de las pruebas para cada caso del variación con su respectiva fecha y hora, además de ciertos comentarios adicionados en alguna ocación.

Se cuenta con un proyecto para las pruebas del PPO como se observan en la Figura 1 y un proyecto para las pruebas con DQN discreto en la Figura 2.

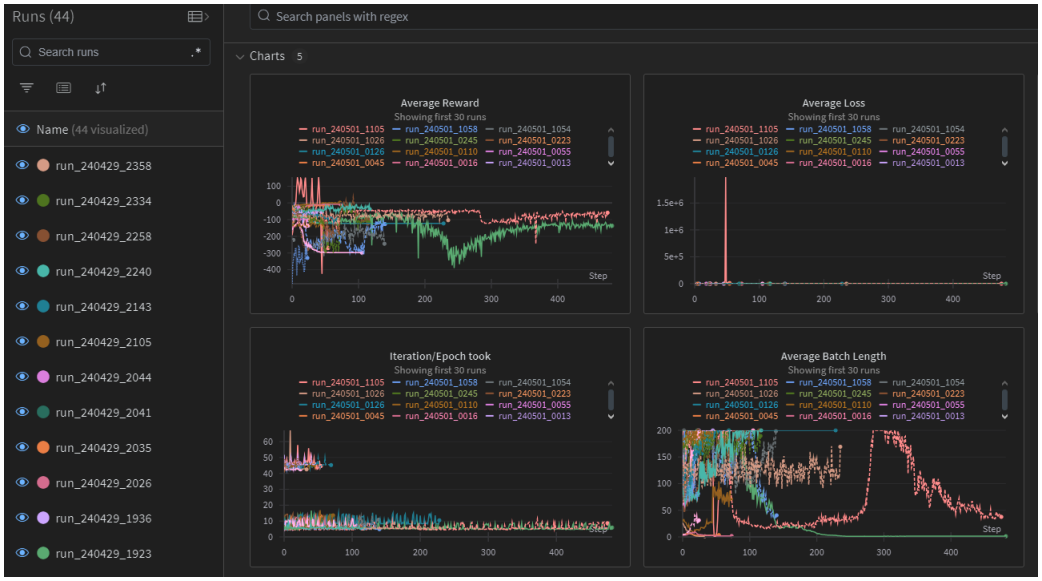

Figure 1: Pruebas de entrenamiento del modelo $PahmPPO$ en $wandb$.

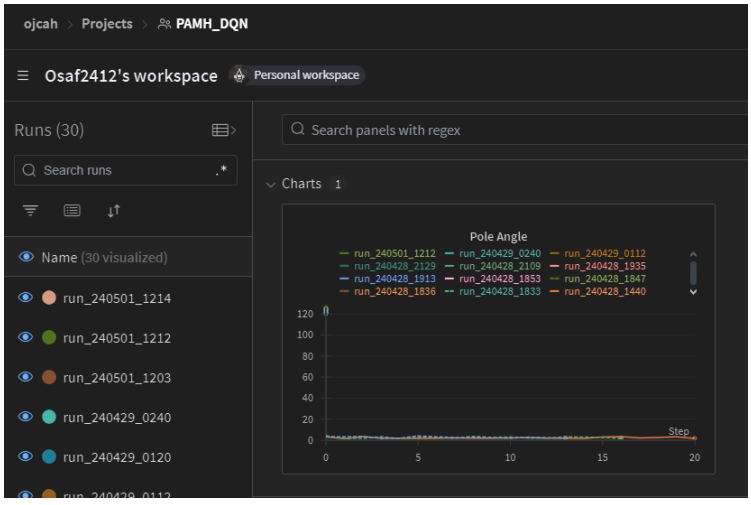

Figure 2: Pruebas de entrenamiento del modelo $PahmDQN$ en $wandb$.

Referencias

[1] S. Huang, R. F. J. Dossa, A. Raffin, A. Kanervisto, and W. Wang, "The 37 implementation details of proximal policy optimization," in *ICLR Blog Track*, 2022, https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/. [Online]. Available: https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/