Máster en Inteligencia Artificial UMU

Propuestas de Trabajos de Investigación

Aprendizaje por Refuerzo

Máster en Inteligencia Artificial

Universidad de Murcia

"El hombre que ha cometido un error y no lo corrige, comete otro error mayor."

Confucio[a]

[a] Reconocer y corregir los errores propios es una parte fundamental del crecimiento personal y el desarrollo moral. La idea central de esta frase tiene paralelismos interesantes con los algoritmos de aprendizaje por refuerzo.

2 de febrero de 2025

Índice

Resumen

En este documento se presentan una serie de trabajos que puede realizar como proyecto final y que deberán de ser expuestos en público.

1. Introducción

El aprendizaje por refuerzo es una de las ramas más fascinantes de la inteligencia artificial porque está demostrando que cada vez tiene más aplicaciones, desde juegos y finanzas en sus inicios hasta la construcción LLMs (Large Language Models) como DeepSeek. Su enfoque se basa en la interacción de un agente con un entorno para aprender a tomar decisiones óptimas a través de la experiencia y la retroalimentación.

El objetivo de su trabajo final es que profundicen algunos conceptos clave, en los fundamentos y las aplicaciones del aprendizaje por refuerzo a través de actividades teórico-prácticas. Más allá del desarrollo de los ejercicios que se proponen en las prácticas, un aspecto clave en un máster es la búsqueda y selección de bibliografía relevante. Es fundamental que consulten fuentes confiables, como artículos científicos, libros y recursos académicos, y que los integren adecuadamente en sus análisis.

Para ello, tal y como se indica en la guía docente, se trata de que entreguen un trabajo donde muestren creatividad, precisión, rigor técnico y profundidad de análisis, discusión de resultados, justificación de las decisiones adoptadas, dando conclusiones razonadas y avaladas por fuentes acordes al informe realizado.

A continuación, se presentan varias alternativas para que realicen su trabajo final. Deberán seleccionar y desarrollar **dos** de ellas, entregando un informe, para cada parte, con sus análisis, resultados y conclusiones.

Para cada trabajo se indica cuál es la línea de trabajo y se proponen algunas fuentes bibliográficas. Obviamente, y esto es con carácter general, deberán echarles un vistazo y seleccionar las referencias bibliográficas. Es imposible leer y abordarlas todas. No se trata de que tengan que leer y entenderlo todo. Simplemente echen un vistazo y aquello que consideren que mejor entiendan o les atraiga, pues háganlo.

2. Entregas

Todo trabajo constará de dos partes. Una parte experimental que podrá ser reproducible por cualquiera y que lo depositarán en GitHub y una parte documental .pdf que se entregará en el Aula Virtual.

Para la parte experimental seguirán los pasos análogos a lo que se pide en la primera práctica.

Para la parte documental es importante que estructuren cada una de las dos entregas de manera clara, incluyendo:

- **Título y Autores:** Incluir el nombre de la obra y los nombres completos de los autores del trabajo, junto con el correo electrónico um.es para cada autor.
- **Introducción:**
  - Breve descripción del problemas y su relevancia.
  - Motivación del trabajo: ¿Por qué es importante el problema estudiado?
  - Objetivos del informe: ¿Qué se pretende lograr con el desarrollo del trabajo?
  - Organización del documento: breve explicación de la estructura del informe.
- **Desarrollo:**
  - Definición formal del problema o aplicación específica abordada.
  - Describir enfoques existentes en la literatura sobre problemas similares.
  - Contexto y antecedentes necesarios para entender el trabajo.
  - Explicación de los métodos utilizados para abordar el problema.
  - Justificar cómo el presente trabajo se diferencia o mejora enfoques previos.
- **Algoritmos:**
  - Explicar en detalle los pasos de los algoritmos empleados.
  - Incluir pseudocódigo y, opcionalmente, diagramas de flujo si es necesario.
  - Justificar la elección de los algoritmos y su aplicabilidad al problema estudiado.

- **Evaluación/Experimentos:**
  - Configuración experimental: herramientas, entornos de prueba, datasets empleados, etc.
  - Métricas utilizadas para evaluar el desempeño del método propuesto.
  - Resultados obtenidos, presentados en tablas o gráficos según corresponda.
  - Análisis crítico de los resultados y comparación con otros enfoques.
- **Conclusiones:**
  - Un resumen de los principales resultados obtenidos.
  - Limitaciones del estudio y posibles mejoras futuras.
  - Reflexión sobre la importancia del trabajo y su impacto en el campo del aprendizaje por refuerzo.
  - Líneas *futuras* de estudio.
- **Bibliografía:** Listar todas las referencias utilizadas en el trabajo, siguiendo un formato bibliográfico estándar (APA, IEEE, etc.). Se recomienda utilizar un gestor de referencias como BibTeX para facilitar la organización de las citas.

El alcance del trabajo lo establecerá el grupo, pero la extensión de cada entrega será entre 5 y 10 páginas de acuerdo a la plantilla que tienen en recursos del AV.

3. Trabajos

A continuación se presentan varias líneas de trabajo. Tienen que seleccionar dos; pero tengan en cuenta que algunas de ellas les puede costar más trabajo si aún no hemos dado sus correspondiente fundamentos teóricos.

3.1. Bandidos

Aquí se propone un trabajo que amplíe el estudio realizado sobre el bandido de k-brazos, usando bandidos duelistas, adversos y/o contextuales.

Enfoquen el trabajo como consideren. Por ejemplo pueden abordar estos 3 tipos de bandidos y hacer un estudio comparativo de un algoritmo de cada tipo de bandido. Alternativamente, por ejemplo, pueden profundizar en solo uno de los bandidos y hacer un estudio de 3 o 4 algoritmos para ese tipo. No olviden, si es posible, compararlos con los realizados en la práctica 1.

El trabajo experimental se añadirá a un repositorio GitHub y a ser posible en alguna de las prácticas realizadas para que las complementen. Por ejemplo, si hicieran un trabajo sobre el problema del bandido, parece lógico añadir "lo nuevo.ª la primera práctica con sus correspondientes *notebooks*.

El trabajo documental .pdf se presentará obligatoriamente en la herramienta Tareas del AV (artículo y presentación si la hubiera) y se incluirá, si quieren, en el repositorio GitHub (aunque esto último no lo recomiendo, si quieren mantener la autoría de su trabajo).

Esto que se acaba de indicar, sobre el trabajo experimental y el documental, **se aplica al resto de las propuestas de trabajo**.

3.1.1. Bandidos Duelistas

Los *bandidos duelistas* son una extensión del problema clásico de los *bandidos multi-brazo*. En lugar de recibir una recompensa numérica fija por cada acción, el agente selecciona dos acciones y obtiene información en forma de comparación (*duelo*) sobre cuál es mejor. Este enfoque es particularmente útil en escenarios donde las recompensas absolutas no están disponibles o son difíciles de definir. El enfoque intuitivo es que los humanos rara vez asignan una puntuación exacta a una opción, pero sí pueden decir cuál prefieren entre dos alternativas. El modelo de bandidos duelistas se asemeja más a cómo tomamos decisiones en la vida cotidiana.

Encontramos una aplicación en sistemas de recomendación (como Netflix o Spotify), es difícil cuantificar la

satisfacción del usuario con una única métrica. Comparar dos recomendaciones en términos de preferencia relativa es una solución más efectiva. También tiene aplicación en motores de búsqueda, donde los algoritmos pueden evaluar qué combinación de resultados ofrece una mejor experiencia al usuario sin depender de una métrica fija como los clics. Esto mismo lo habrá podido comprobar en ChatGPT que, de vez en cuando, le ofrece dos posibles respuestas y le pide que seleccione una de ellas.

Citas Bibliográficas

1. Yue, Y., & Joachims, T. (2009). *Beat the mean bandit*. En **Proceedings of the 26th International Conference on Machine Learning (ICML)**. https://www.cs.cornell.edu/people/tj/publications/yue_joachims_11a.pdf Este trabajo introduce el problema del bandido duelista y propone algoritmos para abordarlo.
2. Zoghi, M., Whiteson, S., Munos, R., & de Rijke, M. (2014). *Relative upper confidence bound for the K-armed dueling bandit problem*. En **Proceedings of the 31st International Conference on Machine Learning (ICML)**. https://proceedings.mlr.press/v32/zoghi14.html Los autores proponen el algoritmo Relative Upper Confidence Bound (RUCB) para el problema del bandido duelista con K opciones.
3. Lekang, T., & Lamperski, A. (2019). *Simple Algorithms for Dueling Bandits*. En **arXiv preprint arXiv:1906.07611**. https://arxiv.org/abs/1906.07611 Este artículo presenta algoritmos simples para el problema del bandido duelista, con análisis de sus límites de regret y comparaciones con algoritmos existentes.
4. González González, M. (2020). *Aprendizaje por refuerzo mediante bandidos duelistas*. Trabajo de Fin de Grado, Universidad de Sevilla. https://miguelgg.com/assets/pdf//publications/TFGInfo.pdf Este trabajo ofrece una visión detallada del problema de los bandidos duelistas y analiza diversos algoritmos, incluyendo aquellos de fácil implementación.

3.1.2. Bandidos Adversarios

"Un bandido estocástico con K acciones está completamente determinado por las distribuciones de recompensas, $P_1, \ldots, P_K$, de las respectivas acciones. En particular, en la ronda t, la distribución de la recompensa X_t recibida por un aprendiz que elige la acción $A_t \in [K]$ es P_{A_t}, independientemente de las recompensas y acciones pasadas. Si se observa detenidamente cualquier problema del "mundo real", ya sea el diseño de medicamentos, la recomendación de artículos en la web o cualquier otro, pronto descubriremos que en realidad no existen distribuciones adecuadas. En primer lugar, será difícil argumentar que la recompensa se genera realmente de manera aleatoria y, aun si fuera generada aleatoriamente, las recompensas podrían estar correlacionadas en el tiempo. Tener en cuenta todos estos efectos haría que un modelo estocástico sea potencialmente bastante complicado. Una alternativa es adoptar un enfoque pragmático, en el que casi nada se asuma sobre el mecanismo que genera las recompensas, pero manteniendo el objetivo de competir con la mejor acción en retrospectiva. Esto nos lleva al llamado ***modelo de bandido adversarial***" *https://banditalgs.com/2016/10/01/adversarial-bandits/*.

El problema del *bandido adversarial* es una variante del problema del bandido multibrazo en la que las recompensas asociadas a cada acción pueden ser determinadas por un adversario, en lugar de seguir una distribución fija y desconocida. Este enfoque es relevante en escenarios donde las recompensas pueden cambiar de manera adversa o estratégica

Citas Bibliográficas

1. Auer, P., Cesa-Bianchi, N., Freund, Y., & Schapire, R. E. (2002). *The nonstochastic multi-armed bandit problem*. **SIAM Journal on Computing,** 32(1), 48-77. Disponible en: https://doi.org/10.1137/S0097539701398375 Dispone de una versión libre en https://cseweb.ucsd.edu/~yfreund/papers/bandits.pdf Este artículo introduce el algoritmo EXP3, diseñado para entornos adversariales en el problema del bandido multibrazo, y proporciona análisis teóricos sobre su rendimiento.

2. Lattimore, T., & Szepesvári, C. (2020). Bandit Algorithms. Cambridge University Press. https://doi.org/10.1017/9781108571401. Disponible una versión libre en https://tor-lattimore.com/downloads/book/book.pdf Este libro estudia en la tercera parte el algoritmo EXP3
3. Bubeck, S., & Cesa-Bianchi, N. (2012). *Regret analysis of stochastic and nonstochastic multi-armed bandit problems*. **Foundations and Trends in Machine Learning,** 5(1), 1-122. Disponible en: https://doi.org/10.1561/2200000024 y también en http://sbubeck.com/SurveyBCB12.pdf Este trabajo ofrece una revisión exhaustiva de los análisis de regret en problemas de bandidos tanto estocásticos como adversariales, incluyendo algoritmos clave y sus propiedades.
4. Seldin, Y., & Slivkins, A. (2014). *One practical algorithm for both stochastic and adversarial bandits*. En **Proceedings of the 31st International Conference on Machine Learning (ICML)**. Disponible en: https://proceedings.mlr.press/v32/seldin14.html Los autores presentan un algoritmo práctico que es efectivo tanto en entornos estocásticos como adversales, adaptándose dinámicamente a la naturaleza del entorno.

3.1.3. Bandidos Contextuales

En la mayoría de los problemas de bandidos, es probable que haya alguna información adicional disponible al comienzo de las rondas y, a menudo, esta información puede ayudar potencialmente con las opciones de acción. Por ejemplo, en un sistema de recomendación de artículos web, donde el objetivo es mantener a los visitantes interesados en el sitio web, la información contextual sobre el visitante del sitio web, la hora del día, información sobre lo que está de moda, etc., probablemente puede mejorar la elección del artículo que se colocará en la "portada". Por ejemplo, es más probable que un artículo orientado a la ciencia capte la atención de un fanático de la ciencia, y es posible que a un fanático del béisbol le importe poco el fútbol europeo.

Si utilizáramos un algoritmo estándar de búsqueda de artículos (como UCB), el enfoque único que adoptan estos algoritmos, que solo buscan encontrar el artículo más atractivo, probablemente decepcionará a una parte innecesariamente grande de los visitantes del sitio. En situaciones como esta, dado que el parámetro de referencia que los algoritmos de búsqueda de artículos buscan alcanzar funciona mal al omitir la información disponible, es mejor cambiar el problema y redefinir el parámetro de referencia. Sin embargo, es importante darse cuenta de que aquí hay una dificultad inherente. https://banditalgs.com/2016/10/14/exp4/

El problema del *bandido contextual* es una extensión del problema del bandido multibrazo que incorpora información contextual en el proceso de toma de decisiones. En cada ronda, el agente observa un contexto y debe seleccionar una acción (o brazo) basándose en este contexto, con el objetivo de maximizar la recompensa acumulada a lo largo del tiempo. Este enfoque es especialmente relevante en aplicaciones como sistemas de recomendación, publicidad en línea y personalización de contenido.

Citas Bibliográficas

1. Lattimore, T., & Szepesvári, C. (2020). Bandit Algorithms. Cambridge University Press. https://doi.org/10.1017/9781108571401. Disponible una versión libre en https://tor-lattimore.com/downloads/book/book.pdf Este libro estudia en la quinta parte el algoritmo EXP4
2. Bubeck, S., & Cesa-Bianchi, N. (2012). *Regret analysis of stochastic and nonstochastic multi-armed bandit problems*. **Foundations and Trends in Machine Learning,** 5(1), 1-122. Disponible en: https://doi.org/10.1561/2200000024 y también en http://sbubeck.com/SurveyBCB12.pdf Este trabajo ofrece una revisión exhaustiva de los análisis de regret en problemas de bandidos tanto estocásticos como contextuales, incluyendo algoritmos clave y sus propiedades.
3. Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010). *A Contextual-Bandit Approach to Personalized News Article Recommendation*. En **Proceedings of the 19th International Conference on World Wide Web (WWW)**. Disponible en: https://doi.org/10.1145/1772690.1772758 y https://arxiv.org/abs/1003.0146 Este trabajo presenta un enfoque de bandido contextual para la recomendación personalizada de artículos de noticias, introduciendo el algoritmo **LinUCB** y

demostrando su eficacia en escenarios reales.

4. Zhou, L. (2015). *A Survey on Contextual Multi-armed Bandits*. **arXiv preprint arXiv:1508.03326**. Disponible en: https://arxiv.org/abs/1508.03326 Este artículo presenta una revisión exhaustiva de los algoritmos de bandidos contextuales, abordando tanto enfoques estocásticos como adversariales, y proporcionando un análisis detallado sobre sus límites de arrepentimiento y suposiciones.
5. Chu, W., Li, L., Reyzin, L., & Schapire, R. E. (2011). *Contextual Bandits with Linear Payoffs*. En **Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)**. Disponible en: https://proceedings.mlr.press/v15/chu11a.html Este artículo introduce un enfoque basado en límites superiores de confianza (UCB) para bandidos contextuales con recompensas lineales, ofreciendo un equilibrio entre exploración y explotación y facilitando su implementación práctica.
6. Agrawal, S., & Goyal, N. (2013). *Thompson Sampling for Contextual Bandits with Linear Payoffs*. En **Proceedings of the 30th International Conference on Machine Learning (ICML)**. Disponible en: https://proceedings.mlr.press/v28/agrawal13.html Los autores analizan el uso de Thompson Sampling en el contexto de bandidos con recompensas lineales, proporcionando garantías teóricas sobre su rendimiento y comparándolo con otros algoritmos.

3.2. Métodos Aproximados

Al igual que en el apartado anterior, son libres de centrarse en un tema o tratar varios a la vez. Por ejemplo se pueden centrar solo en los métodos TD(λ) o centrarse solo en los métodos Q(λ), o bien tratar uno o dos de cada tipo y hacer un estudio conjunto.

3.2.1. TD(λ)

El algoritmo TD(λ) es una técnica de *Aprendizaje por Refuerzo* que combina los métodos de *Diferencia Temporal* (TD) y *Monte Carlo*, utilizando un parámetro de trazas de elegibilidad, λ, para equilibrar entre ambos enfoques. Fue introducido por Richard S. Sutton en 1988 y se ha convertido en una herramienta fundamental en el aprendizaje automático.

La actualización del valor de un estado s_t en TD(λ) se define como: $V(s_t) \leftarrow V(s_t) + \alpha\delta_t e_t$ donde

- α es la tasa de aprendizaje,
- δ_t es el error de TD, calculado como: $\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$
- e_t es la traza de elegibilidad, que se actualiza según: $e_t = \gamma\lambda e_{t-1} + 1$

Aquí, γ es el factor de descuento y λ es el parámetro que controla la longitud de las trazas de elegibilidad.

TD(λ) es un método de aproximación porque no calcula los valores exactos de los estados, sino que estima la función de valor utilizando las trazas de elegibilidad para interpolar entre los métodos TD(0) y Monte Carlo. Ajustando el parámetro λ, se puede controlar el equilibrio entre sesgo y varianza en las estimaciones.

Citas Bibliográficas

1. Sutton, R. S. (1988). *Learning to predict by the methods of temporal differences*. Machine Learning, 3(1), 9-44. Disponible en: https://link.springer.com/article/10.1007/BF00115009 Este es el artículo seminal donde Richard Sutton introduce el método de Diferencia Temporal (TD) y presenta TD(λ) como una interpolación entre métodos TD y Monte Carlo.
2. Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press. Disponible en: http://incompleteideas.net/book/the-book-2nd.html Este libro es una referencia fundamental sobre aprendizaje por refuerzo. Contiene explicaciones detalladas de TD(λ), trazas de elegibilidad y su aplicación en diferentes problemas de optimización.
3. Seijen, H. V., & Sutton, R. S. (2014). *True Online TD(λ)*. En Proceedings of the 31st International

Conference on Machine Learning (ICML-14), 692-700. Disponible en: http://proceedings.mlr.press/v32/seijen14.pdf Este artículo introduce la versión *True Online TD(λ)*, una mejora sobre TD(λ) que hace que la actualización de valores sea más estable y eficiente en tiempo real.
4. van Seijen, H., & Sutton, R. S. (2014). *True Online TD(λ)*. En *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 692-700. Disponible en: https://proceedings.mlr.press/v32/seijen14.pdf Este artículo presenta el algoritmo True Online TD(λ), una variante que mejora la correspondencia entre la vista hacia adelante y la implementación en línea de TD(λ), ofreciendo actualizaciones más precisas y estables.
5. van Seijen, H., Mahmood, A. R., Pilarski, P. M., Machado, M. C., & Sutton, R. S. (2016). *True Online Temporal-Difference Learning*. *Journal of Machine Learning Research*, 17(145), 1-40. Disponible en: https://jmlr.org/papers/volume17/15-599/15-599.pdf En este trabajo, los autores amplían el concepto de True Online TD(λ) a métodos de control, introduciendo True Online Sarsa(λ) y demostrando su eficacia en diversos dominios.

3.2.2. Q(λ)

El algoritmo Q(λ) es una extensión del Q-learning que introduce el concepto de trazas de elegibilidad mediante el parámetro λ. Este mecanismo permite que el aprendizaje no solo dependa de la recompensa inmediata, sino también de múltiples pasos anteriores.

El parámetro λ regula la contribución de estos pasos anteriores:

- $\lambda = 0 \rightarrow$ Se reduce a Q-learning tradicional (actualización basada solo en la recompensa inmediata).
- $\lambda = 1 \rightarrow$ Se comporta como Monte Carlo (utiliza la recompensa total del episodio).
- $0 < \lambda < 1 \rightarrow$ Realiza una interpolación entre ambos enfoques.

Citas Bibliográficas

1. Peng, J., & Williams, R. J. (1996). *Incremental Multi-Step Q-Learning*. En **Machine Learning, 22**(1-3), 283-290. Disponible en: https://link.springer.com/article/10.1007/BF00114731 Este artículo introduce una versión de Q-learning con actualizaciones de múltiples pasos, precursora directa del algoritmo Q(λ).
2. Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press. Disponible en: http://incompleteideas.net/book/the-book-2nd.html Este libro es una referencia fundamental en Aprendizaje por Refuerzo y proporciona una explicación detallada de Q(λ), sus propiedades y su relación con otros algoritmos.
3. Wiering, M. & Schmidhub, J. (s.f.). *Fast Online Q(λ)*. IDSIA, Corso Elvezia 36, 6. http://www.idsia.ch/~juergen/fast_online_q_lambda.pdf Este trabajo presenta una versión rápida y en línea"del algoritmo Q(λ) para el aprendizaje por refuerzo, permitiendo actualizaciones eficientes en tiempo real. (La fecha de publicación no se encuentra especificada.)
4. Mousavi, S. S., Schukat, M., & Howley, E. (2017). *Applying Q(λ)-Learning in Deep Reinforcement Learning to Play Atari Games*. En **Proceedings of the Adaptive and Learning Agents Workshop (ALA 2017)**. Disponible en: https://ala2017.cs.universityofgalway.ie/papers/ALA2017_Mousavi.pdf Este trabajo explora la implementación de Q(λ) en un entorno de Aprendizaje Profundo aplicado a juegos de Atari, mostrando mejoras en la eficiencia del aprendizaje.
5. Sutton, R. S., Mahmood, A. R., Precup, D., & van Hasselt, H. (2014). *A new Q(λ) with interim forward view and Monte Carlo equivalence*. En *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 568-576. Disponible en: https://proceedings.mlr.press/v32/sutton14.pdf Este artículo presenta una nueva variante de Q(λ) que logra una equivalencia exacta con el enfoque de Monte Carlo, mejorando la precisión y estabilidad en el aprendizaje por refuerzo.
6. Harutyunyan, A., Bellemare, M. G., Stepleton, T., & Munos, R. (2016). *Q(λ) with Off-Policy Corrections*. Disponible en: https://arxiv.org/abs/1602.04951 Los autores proponen y analizan un enfoque alternativo para el aprendizaje temporal-diferencial de múltiples pasos off-policy, introdu-

ciendo correcciones que mejoran la convergencia y eficiencia del algoritmo Q(λ).

7. Wiering, M. & Schmidhuber, J. (1998). *Fast Online Q(λ)*. **Machine Learning** Kluwer Academic Publishers, 33, 105–115. (1998) https://link.springer.com/content/pdf/10.1023/A:1007562800292.pdf Este artículo presenta una versión rápida y en línea del algoritmo Q(λ) para el aprendizaje por refuerzo, permitiendo actualizaciones eficientes en tiempo real.

3.2.3. Actor Crítico

Las técnicas de actor-crítico combinan dos componentes principales: el *actor*, que se encarga de seleccionar las acciones (política), y el *crítico*, que evalúa la calidad de las acciones o estados (función de valor). Existen numerosas variantes y mejoras en estos métodos. Cada uno de estos enfoques tiene sus propias ventajas y desafíos, y la elección del método adecuado puede depender de la naturaleza del problema, el tipo de espacio de acción (discreto o continuo) y las características específicas del entorno en el que se aplica. Estas técnicas han sido fundamentales para avanzar en el campo del aprendizaje por refuerzo, permitiendo entrenar agentes que pueden aprender a tomar decisiones complejas en entornos dinámicos y de alta dimensión. Hay una de ellas que parece ser, por ahora, casi insuperable (TD3) hasta el punto de que OpenAI le dedica algunas páginas a su explicación https://spinningup.openai.com/en/latest/algorithms/td3.html

Citas Bibliográficas

1. Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., ..., & Kavukcuoglu, K. (2016). *Asynchronous Methods for Deep Reinforcement Learning*. https://arxiv.org/abs/1602.01783 Este artículo introduce A3C, cuyo modo sincrónico se conoce como A2C, utilizando múltiples agentes para actualizar la política y reducir la varianza en las actualizaciones.
2. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ..., & Hassabis, D. (2015). *Human-level Control through Deep Reinforcement Learning*. https://www.nature.com/articles/nature14236 Aunque se centra en DQN, este trabajo sienta las bases para los métodos actor-crítico al demostrar la eficacia de las aproximaciones basadas en redes neuronales profundas en RL.
3. Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., ... & Wierstra, D. (2015). *Continuous Control with Deep Reinforcement Learning*. https://arxiv.org/abs/1509.02971 Este artículo introduce DDPG, un método off-policy para control continuo que combina un actor determinista con un crítico basado en Q.
4. Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2017). *Deep Reinforcement Learning That Matters*. https://arxiv.org/abs/1709.06560 Este trabajo analiza comparativamente varios algoritmos de RL, incluyendo DDPG, y destaca buenas prácticas y desafíos en su implementación.
5. Fujimoto, S., van Hoof, H., & Meger, D. (2018). *Addressing Function Approximation Error in Actor-Critic Methods*. https://arxiv.org/abs/1802.09477 Este artículo introduce TD3, que mejora DDPG mitigando el error de aproximación de funciones mediante el uso de dos críticos y actualizaciones retrasadas del actor.
6. Fujimoto, S., van Hoof, H., & Meger, D. (2018). *TD3: Twin Delayed Deep Deterministic Policy Gradient*. https://openreview.net/forum?id=rJk00ekbF Versión publicada de TD3, que detalla las modificaciones implementadas para mejorar la estabilidad y el rendimiento en tareas de control continuo.
7. Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). *Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning*. https://arxiv.org/abs/1801.01290 Este artículo introduce SAC, un método off-policy que utiliza un objetivo de entropía máxima para fomentar la exploración y mejorar la estabilidad en el entrenamiento.
8. Haarnoja, T., Zhou, A., Tang, H., & Levine, S. (2018). *Soft Actor-Critic Algorithms and Applications*. https://arxiv.org/abs/1812.05905 Este trabajo extiende SAC y explora diversas aplicaciones, proporcionando una visión más amplia de su desempeño en entornos de control continuo.

9. Wang, Z., Schaul, T., Hessel, M., Hasselt, H. V., Lanctot, M., & Freitas, N. (2017). *Sample Efficient Actor-Critic with Experience Replay*. https://arxiv.org/abs/1611.01224 Este artículo introduce ACER, que integra el replay de experiencia en el marco actor-crítico para mejorar la eficiencia muestral y la estabilidad del entrenamiento.
10. Kapturowski, S., Ostrovski, G., Quan, J., Munos, R., & Dabney, W. (2018). *Recurrent Experience Replay in Distributed Reinforcement Learning*. https://arxiv.org/abs/1803.07240 Este trabajo propone mejoras al replay de experiencia en entornos distribuidos, complementando y extendiendo ideas presentes en ACER.

3.2.4. Regiones de Confianza

El objetivo de los métodos que se quieren analizar aquí es mejorar la estabilidad de la actualización de la política durante el entrenamiento. Existe un dilema: por un lado, nos gustaría entrenar lo más rápido posible, realizando grandes pasos durante la actualización mediante descenso de gradiente estocástico (SGD). Por otro lado, una actualización grande de la política generalmente es una mala idea. La política es algo muy no lineal, por lo que una actualización grande podría arruinar la política que acabamos de aprender.

La situación puede empeorar aún más en el ámbito del aprendizaje por refuerzo (RL), ya que no se puede recuperar de una mala actualización de la política mediante actualizaciones posteriores. En su lugar, la política incorrecta proporcionará muestras de experiencia deficientes que utilizaremos en pasos de entrenamiento posteriores, lo que podría romper nuestra política por completo. Por lo tanto, queremos evitar realizar actualizaciones grandes a toda costa. Una de las soluciones ingenuas sería utilizar una tasa de aprendizaje pequeña para dar pasos muy reducidos durante el SGD, pero esto ralentizaría significativamente la convergencia.

Para romper este círculo vicioso, varios investigadores han intentado estimar el efecto que tendrá nuestra actualización de la política en términos de resultados futuros. Constituyen los métodos de regiones de confianza (Trust Region Methods) en el aprendizaje por refuerzo. Son técnicas que se utilizan para garantizar que las actualizaciones de la política sean estables y no se produzcan cambios demasiado drásticos de una iteración a la siguiente. La idea central es limitar el "tamaño" del cambio de la política en cada actualización, de modo que la nueva política se mantenga lo suficientemente cercana a la anterior para no deteriorar el rendimiento aprendido.

Citas Bibliográficas

1. Schulman, J., Levine, S., Moritz, P., Jordan, M. I., & Abbeel, P. (2015). *Trust Region Policy Optimization*. En **Proceedings of the 32nd International Conference on Machine Learning (ICML)**. https://arxiv.org/abs/1502.05477. Este artículo introduce TRPO, un algoritmo que mejora la estabilidad de las actualizaciones de la política al restringir la divergencia de Kullback-Leibler (KL) entre la política antigua y la nueva, garantizando pasos de actualización "seguros".
2. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). *Proximal Policy Optimization Algorithms*. https://arxiv.org/abs/1707.06347 PPO simplifica la complejidad computacional de TRPO utilizando un mecanismo de *clipping* en la función objetivo para limitar la magnitud de las actualizaciones de la política, facilitando su implementación sin perder estabilidad.
3. Schulman, J., Moritz, P., Levine, S., Jordan, M., & Abbeel, P. (2015). *High-Dimensional Continuous Control Using Generalized Advantage Estimation*. https://arxiv.org/abs/1506.02438 Este artículo introduce GAE (Generalized Advantage Estimation), que mejora la estimación de la ventaja reduciendo la varianza sin incurrir en un sesgo excesivo, facilitando actualizaciones de política más estables.
4. Schulman, J., Levine, S., Abbeel, P., Jordan, M., & Moritz, P. (2015). *Trust Region Policy Optimization*. https://arxiv.org/abs/1502.05477 Aunque este artículo se centra en TRPO, incluye una discusión sobre GAE como parte integral del método para lograr actualizaciones de política confiables en entornos complejos.

5. Wu, Y., Mansimov, E., Liao, S., Grosse, R., & Ba, J. (2017). *Scalable Trust-Region Method for Deep Reinforcement Learning Using Kronecker-Factored Approximation*. En **Proceedings of the 34th International Conference on Machine Learning (ICML)**. https://arxiv.org/abs/1708.05144 ACKTR (Actor-Critic using Kronecker-Factored Trust Region) extiende la idea de las regiones de confianza a redes neuronales profundas, utilizando una aproximación factorizada de Kronecker para estimar la curvatura del espacio de parámetros (matriz de Fisher) y, de esta forma, realizar actualizaciones eficientes y estables en entornos de alta dimensión (redes neuronales profundas).
6. Martens, J., & Grosse, R. (2015). *Optimizing Neural Networks with Kronecker-Factored Approximation*.
https://arxiv.org/abs/1503.05671 Aunque no es específico de RL, este trabajo proporciona la base teórica para la aproximación de la curvatura utilizada en ACKTR.
7. Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., ... & Kavukcuoglu, K. (2016). *Asynchronous Methods for Deep Reinforcement Learning*. En **Proceedings of the 33rd International Conference on Machine Learning (ICML)**. https://arxiv.org/abs/1602.01783 El A2C con linea base es una variante sincrónica del método Asynchronous Advantage Actor-Critic (A3C) que sirve como base estable y sencilla para métodos actor-crítico, facilitando la comparación con otros algoritmos porque ofrece un equilibrio adecuado entre exploración y explotación en entornos de acción continua.
8. Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). *Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning*. En **Proceedings of the 35th International Conference on Machine Learning (ICML)**. **URL:** https://arxiv.org/abs/1801.01290 El método SAC (Soft Actor-Critic) no se clasifica (como A2C) como un método basado en regiones de confianza, aunque SAC comparte con estos métodos el objetivo de lograr estabilidad durante la actualización de la política, lo hace mediante mecanismos diferentes. SAC se basa en un objetivo de entropía máxima que incentiva la exploración y la robustez en la toma de decisiones. La estabilidad en SAC se logra a través de la incorporación de un término de entropía en la función de objetivo, lo que suaviza la política y fomenta actualizaciones más estables. No utiliza restricciones explícitas en la forma de una región de confianza como PPO o TRPO; pero como çomparten el mismo objetivo"son candidatos a ser comparables.

3.3. Otros trabajos por iniciativa propia

Alternativamente puede desarrollar un trabajo más aplicado. Por ejemplo, el uso del aprendizaje por refuerzo en el mundo de la robótica, videojuegos, etc ...

Si le interesa trabajar sobre esta otra línea, diferente a las anteriores, remita al profesor la siguiente información: objetivo del trabajo, referencias documentales, librerías a utilizar y encaje y *dificultad* con el resto de los contenidos de la asignatura.

Hasta que no tenga el visto bueno para su realización, mejor no empiece. Su propuesta puede quedar descartada.