[1]

# Implementación de nuevas funciones de recompensa en algoritmos de aprendizaje por reforzamiento mediante el framework DREAM-ON-GYM en el contexto de redes ópticas elásticas - Equipo CFM

*Departamento de Electrónica, Universidad Técnica Federico Santa María*

*Carlos Cea - 201730047-7*
*Felipe Garay - 201892003-7*
*Martín Rojas - 201830023-3*

***Resumen* — En el siguiente trabajo se buscará implementar nuevas funciones de recompensa en el contexto del aprendizaje por reforzamiento, esto para mejorar el entrenamiento de agentes en el contexto de las redes elásticas, pudiendo comprobar estas mejoras mediante métricas como la probabilidad de bloqueo, uso de espectro y/o bandas de frecuencia utilizadas, además del desempeño y evolución de las recompensas en el contexto de simulaciones iterativas.**

## I. INTRODUCCIÓN

No es un misterio que cada día las personas utilizan más el internet y buscan una mayor implementación en su vida diaria, no obstante, los recursos no son ilimitados cuando extrapolamos este uso a todo el mundo, dándose a conocer el capacity crunch, una suerte de futuro apocalíptico donde agotamos todos los recursos de la red. Por ello se plantean múltiples maneras para contrarrestar la llegada de ese día, dentro de las más relevantes con nuestro proyecto, tenemos lo que respecta a la utilización de espectro y el ruteo con distintas modulaciones.

## II. DESCRIPCIÓN DEL CONTEXTO

La base del trabajo es la utilización del framework DREAM-ON-GYM, el cual nos permite utilizar distintos agentes de aprendizaje por reforzamiento mediante la importación desde la biblioteca de baseline3. Dada la información correspondiente a otras investigaciones del tema, se decide utilizar los agentes TRPO y PPO2 en nuestro entrenamiento. Situando 1000 pasos de entrenamiento debido a que en el contexto de los rewards, se puede observar la estabilidad de los mismos al alcanzar esta cifra de entrenamiento (decisión tomada en base a la literatura). Se decide iterar el modelo entrenado en el orden de 10^6.
Estas iteraciones se hacen tomando como referencias las topologías de red de Eurocore (Figura 1.) y NSFNet (Figura 2.), cuyas características principales en el apartado de su arquitectura se encuentran en la tabla 1.

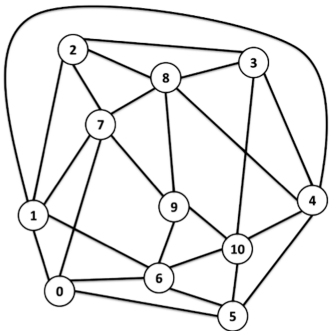


Fig. 1. Topología de la red Eurocore.

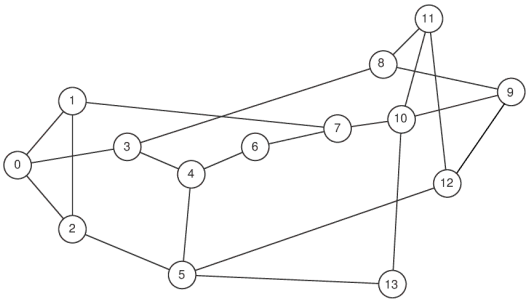


Fig. 2. Topología de la red NSFNet.

| Topología | Nodos | Links | Par de nodos |
|---|---|---|---|
| Eurocore | 11 | 50 | 110 |
| NSFNet | 14 | 42 | 82 |

Tabla 1. Tabla resumen arquitecturas de red

La inclusión de las dos topologías se realiza para situar un precedente en la comparación de topologías con especificaciones distintas, debido a la importancia de implementar un agente que se pueda implementar en distintos ambientes. Esto puede ser importante para encontrar la herramienta más adecuada en base a los distintos escenarios que existen actualmente.

Todo lo mencionado anteriormente corresponden a componentes fundamentales en el problema de RMLSA, el cual hoy en día se busca resolver. El uso de agentes en el contexto del aprendizaje por reforzamiento es un enfoque de investigación distinto para la búsqueda de algoritmos para solucionar el problema antes mencionado. A través de este estudio se buscará trabajar con los agentes para intentar obtener nuevas maneras de guiar su funcionamiento en base a estas recompensas, para así poder entender mayormente el trabajo de estos en un entorno controlado de redes ópticas elásticas.

[1]

**Implementación de nuevas funciones de recompensa en algoritmos de aprendizaje por reforzamiento mediante el framework DREAM-ON-GYM en el contexto de redes ópticas elásticas – Equipo CFM**

## III. OBJETIVOS

El principal objetivo de nuestro trabajo consiste en la generación e implantación de recompensas a utilizar en el entrenamiento de los diversos agentes en nuestro trabajo, esto con el motivo de mejorar el rendimiento en las simulaciones.

## IV. METODOLOGÍA RESOLUCIÓN DESAFÍO

La primera instancia de trabajo se basó en la creación de funciones de recompensa, para esto se estudió las capacidades de DREAM-ON-GYM para ver qué variables estaban disponibles a trabajar. Luego se decidió trabajar con la función de recompensa ya establecida, que denominamos "Caso Base", y crear dos funciones más para comparar el rendimiento de estas. La función "Caso Base" consiste de una recompensa simple donde se recompensa si la última conexión se establece, en caso contrario, se resta un valor de recompensa. Las funciones creadas fueron "Band Select" y "Length Based", la primera recompensa está basada en las bandas (Tabla 2.) que se eligen en la última conexión realizada. La banda que se recompensa negativamente es cuando se elige la banda "E", esto resulta en una recompensa incompleta, es decir, se le otorga 0.8 en vez de 1 al agente. Esto se debe a que la banda "E" la consideramos la peor dentro de las disponibles, por lo tanto, se busca indicarle al agente que esta es una banda no favorable. La segunda función, recompensa positivamente si el agente logró elegir la ruta más corta, dándole la recompensa entera de valor 1, o bien, si es una de las más cortas aparte de la primera, se le entrega un valor de 0.9. Finalmente, en el caso que no fue una de las 3 rutas más cortas, se le da un valor de recompensa de 0.8. El propósito de la segunda función, es indicarle al agente que hay múltiples rutas favorables que además están dentro de las más cercanas para los pares de nodos.

| Band | Frequency (THz) | Bandwidth (BW) | Slots (BW/12.5 GHz) |
|---|---|---|---|
| **L** | 185.7 -191.7 | 6 | 480 |
| **C** | 191.7 - 196 | 4.3 | 344 |
| **S** | 196 - 205.5 | 9.5 | 760 |
| **E** | 205.5 - 219.7 | 14.2 | 1136 |
| **Total** | 185.7 - 219.7 | 34 | 2720 |

Tabla 2. Tabla resumen información bandas

Luego, cada una de estas recompensas (por separado) serán utilizadas para el entrenamiento de los agentes disponibles, para luego culminar con la simulación del modelo una cantidad de 100.000 iteraciones, teniendo como salida la variación de la suma de las recompensas en cada iteración y las probabilidades de bloqueo asociada a la misma.

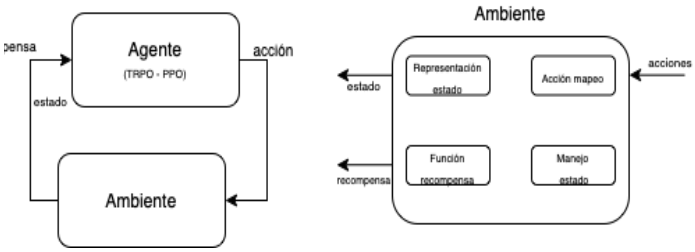


Diagrama 1. Diagrama contexto agente/ambiente

Luego de entrenar estos agentes con cada una de las recompensas, se simula la utilización de la red, esto para buscar como el entrenamiento previo con los rewards inciden en las recompensas al simular la red, además de contrastar en una fase posterior con las probabilidades de bloqueo (gráficamente).

## V. RESULTADOS OBTENIDOS

Para facilitar la interpretación de los resultados, se graficaron de manera comparativa los resultados de cada función de recompensa y se obtuvieron finalmente dos conjuntos de gráficos. Uno correspondiente a los rewards y otro a la probabilidad de bloqueo, ambos sets fueron a lo largo de las iteraciones de sus simulaciones.

### *A. Gráficos de Rewards v/s Iteraciones*

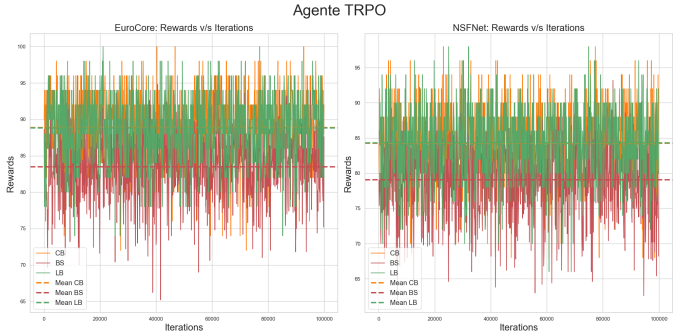


Fig. 3. Gráfico Rewards vs iteraciones agente TRPO.


Fig. 4. Gráfico Rewards vs iteraciones agente PPO.

### A. *Gráficos de Probabilidad de Bloqueo v/s Iteraciones*

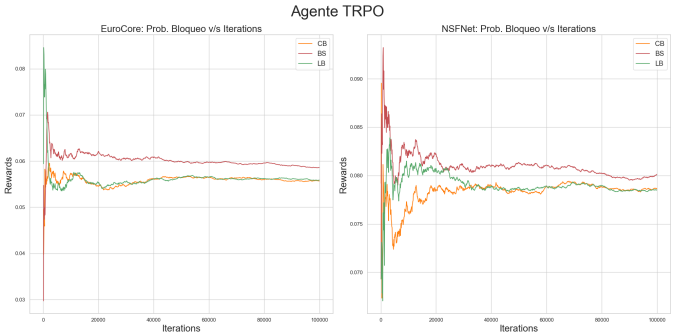


Fig. 5. Gráfico Probabilidad de bloqueo vs iteraciones agente TRPO.

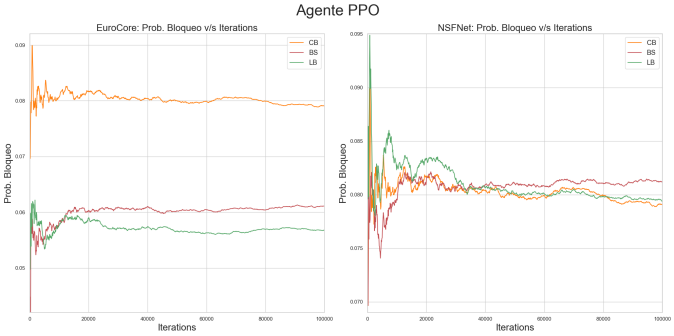


Fig. 6. Gráfico Probabilidad de bloqueo vs iteraciones agente PPO.

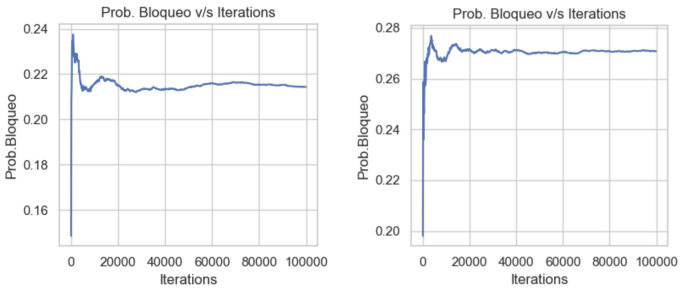


Fig. 7-8. Gráfico Probabilidad de bloqueo vs iteraciones utilizando FistFit en topologías elegidas, Eurocore y NSFNet respectivamente.

Los gráficos permitieron un análisis más profundo de los resultados de cada función, donde se puede observar que la función de "Band Select" tuvo en promedio rewards más bajos que "Caso Base" y "Length Based", además en la mayoría de casos, las probabilidades de bloqueo de esta función de recompensa tendieron a ser las peores de las tres. El "Caso Base" y "Length Based" obtuvieron resultados similares y consistentes para todos los casos, menos donde el agente PPO operó en la topología de EuroCore, este caso en particular el peor fue "Caso Base". Como se observó, la función "Band Select" fue la que terminó con peores resultados en promedio, puede que esto se deba a la mayor dificultad que supone manejar estos eventos por parte del agente, por lo tanto, termina en recompensas con una alta variabilidad (sobre todo en los máximos y mínimos de la gráfica), también probabilidades de bloqueo más altas que las demás funciones. Esta es una función que a futuro se podría trabajar de mayor manera para que el agente llegue a recompensas más estables y a las métricas deseadas y por ende, más óptimas. En el caso de "Length Based", se tienen resultados similares al "Caso Base", pero sí tiene el beneficio de que trabajó particularmente de mejor manera con el agente PPO, por lo que se concluye que sería una función importante a considerar si se trabaja con este agente en un futuro.

Finalmente, los últimos gráficos corresponden a la probabilidad de bloqueo utilizando el algoritmo First Fit a secas, donde podemos ver resultados peores a los obtenidos por los agentes y las diversas funciones de recompensa presentes. Este resultado no corresponde a lo esperado en las investigaciones previas, pudiendo aseverar que hay una anomalía en la generación de los datos, esto se intuye a que al momento de correr el algoritmo, utilizamos el modelo base proporciona DREAM-ON-GYM, el que presenta una tendencia estricta a la utilización de la banda C, lo que explica la mayor cantidad de bloqueos y por ende en la probabilidad de la misma.

## VI. TRABAJO FUTURO

Uno de los trabajos futuros posibles puede ser introducir mejores funciones de recompensa en nuestro entrenamiento, dado que actualmente se premian decisiones basadas en métricas sencillas, tales como la distancia y/o el uso de bandas de forma arbitraria. Estas nuevas funciones de recompensa deben basarse en la búsqueda de mejorar fenómenos tales como la fragmentación de espectro y el agotamiento de los recursos en los diversos canales de comunicación presentes.

Dentro de los trabajos futuros también se cuenta con la posibilidad de añadir variabilidad al apartado de los pasos de entrenamiento, esto para ver la injerencia de esta variable en los resultados tras su implementación, además de ejecutar la simulación una mayor cantidad de veces.

## VII. APRECIACIONES PERSONALES

Durante el trabajo nos encontramos con dificultades provenientes del desconocimiento del framework DREAM-ON-GYM, este desconocimiento principalmente se basaba en el manejo de la información topológica y la representación de la misma en un proceso de asignación de los recursos de esta. Pese a esto, se trabajó arduamente en la superación de esta curva de aprendizaje, concluyendo en resultados certeros a lo que respecta a la generación de funciones de recompensa y comparación con probabilidades de bloqueo dentro de las funciones generadas y las elecciones de topologías/agentes. No obstante, se tiene que tomar en cuenta que estas funciones de recompensa están generadas desde nosotros y en algunas ocasiones estas recompensas pueden inducir a una acción correcta de manera general, pero situándose en el contexto del estado actual, esta puede ser la

no más óptima, además de la corrección de la probabilidad de bloqueo al utilizar First Fit, ya que esta debería ser mejor que las calculadas previamente (dada la literatura investigada).

## VIII. Referencias

**[1]** P. Morales et al., "Multi-band Environments for Optical Reinforcement Learning Gym for Resource Allocation in Elastic Optical Networks," 2021 International Conference on Optical Network Design and Modeling (ONDM), Gothenburg, Sweden, 2021, pp. 1-6, doi: 10.23919/ONDM51796.2021.9492435.

[2] C. Natalino and P. Monti, "The Optical RL-Gym: An open-source toolkit for applying reinforcement learning in optical networks," 2020 22nd International Conference on Transparent Optical Networks (ICTON), Bari, Italy, 2020, pp. 1-5, doi: 10.1109/ICTON51198.2020.9203239.

[3] B. Tang, Y. -C. Huang, Y. Xue and W. Zhou, "Heuristic Reward Design for Deep Reinforcement Learning-Based Routing, Modulation and Spectrum Assignment of Elastic Optical Networks," in IEEE Communications Letters, vol. 26, no. 11, pp. 2675-2679, Nov. 2022, doi: 10.1109/LCOMM.2022.3195778.

[4] A. Beghelli et al., "Approaches to dynamic provisioning in multiband elastic optical networks," 2023 International Conference on Optical Network Design and Modeling (ONDM), Coimbra, Portugal, 2023, pp. 1-6.

## IX. Anexos

Repositorio con los códigos de cada entrenamiento, ejecución del modelo y resultados en formato visual (gráficos).

https://github.com/fevgaray/RL-Implementation-Reward-Functions-Design-DoG