



UNIVERSIDAD DE LA REPÚBLICA
FACULTAD DE INGENIERÍA



Asignación de Recursos en Redes Inalámbricas Mediante Aprendizaje por Refuerzo Profundo con Redes Neuronales de Grafos

TESIS PRESENTADA A LA FACULTAD DE INGENIERÍA DE LA
UNIVERSIDAD DE LA REPÚBLICA POR

Bruno Ottonelli, Nahuel Piñeyro, Mauricio Vieira

EN CUMPLIMIENTO PARCIAL DE LOS REQUERIMIENTOS
PARA LA FINALIZACIÓN DE LA CARRERA DE
GRADO EN INGENIERÍA EN SISTEMAS DE COMUNICACIÓN E
INGENIERÍA ELÉCTRICA.

TUTOR

German Capdehourat	Universidad de la República
Santiago Fernández	Universidad de la República
Federico La Rocca	Universidad de la República

TRIBUNAL

Isabel Amigo	Universidad de la República
Gastón García	Universidad de la República
Claudina Rattaro	Universidad de la República

Montevideo
lunes 3 agosto, 2026

Asignación de Recursos en Redes Inalámbricas Mediante Aprendizaje por Refuerzo Profundo con Redes Neuronales de Grafos, Bruno Ottonelli, Nahuel Piñeyro, Mauricio Vieira.

ISSN 1688-2806

Esta tesis fue preparada en L^AT_EX usando la clase iietesis (v1.2).

Contiene un total de 98 páginas.

Compilada el lunes 3 agosto, 2026.

<http://iie.fing.edu.uy/>

“Schreiben ist gut, Denken ist besser. Klugheit ist gut, Geduld ist besser.”

Escribir es bueno, pensar es mejor. La inteligencia es buena, la paciencia es mejor.

- HERMAN HESSE

Esta página ha sido intencionalmente dejada en blanco.

Agradecimientos

En primer lugar, queremos agradecer a la Facultad de Ingeniería de la Universidad de la República. Estas aulas no solo fueron el lugar de nuestra formación técnica, sino también el espacio donde aprendimos que la ingeniería tiene mucho de constancia, paciencia y trabajo compartido. Gracias a esta institución pública por darnos las herramientas para transformar la curiosidad en conocimiento. También queremos agradecer al Instituto de Ingeniería Eléctrica (IIE). Más que un lugar de trabajo, fue un espacio donde siempre encontramos puertas abiertas, ideas, discusiones y apoyo. Agradecemos especialmente a los docentes y funcionarios que, incluso en los momentos de mayor complejidad técnica, estuvieron para dar una mano y hacer posible que este proyecto avanzara. Un agradecimiento muy especial a nuestros tutores: Federico La Rocca, Germán Capdehourat y Santiago Fernández. Gracias por guiarnos en este proceso, por la paciencia, por las correcciones, por la exigencia y por ayudarnos a ordenar las ideas cuando todo parecía más ruido que señal. Su mirada fue fundamental para que este trabajo llegara a buen puerto. A Ceibal, por confiar en este camino y por facilitar los datos que sostienen empíricamente esta investigación. Trabajar con la realidad de redes desplegadas a escala nacional fue un desafío enorme, pero también una oportunidad poco común para salir de la teoría y acercarnos a problemas concretos. A nuestros compañeros y amigos, por estar. Cuando arrancamos la facultad, una de las primeras cosas que nos dijeron fue que armáramos un grupo de estudio, porque ahí iba a estar buena parte del apoyo para llegar al final. Tenían razón. Con el tiempo, ese consejo se transformó en algo mucho más grande, compañerismo, historias compartidas y, en muchos casos, una segunda familia. A nuestras familias, por el apoyo incondicional durante todo este proceso. Gracias por la paciencia, por la comprensión y por bancarnos también en los momentos más pesados. Su presencia fue un sostén fundamental para que este camino pudiera completarse. Y finalmente, a nosotros mismos, por la constancia. Por no aflojar ante las dificultades, por seguir empujando incluso cuando el panorama no estaba del todo claro, y por llegar hasta acá juntos. Si algo nos deja este proceso, más allá del resultado, es todo lo que aprendimos y crecimos en el camino. A quienes lean estas páginas, gracias por darle sentido a este esfuerzo.

Esta página ha sido intencionalmente dejada en blanco.

Resumen

Las redes inalámbricas densas son una parte central de la conectividad actual. En estos entornos, múltiples transmisores comparten un espectro limitado y deben decidir, de forma coordinada, qué canal utilizar y con qué potencia transmitir. Una mala asignación incrementa la interferencia, degrada la tasa de transmisión y afecta la calidad del servicio. El problema es especialmente desafiante porque el canal inalámbrico varía en el tiempo y el número de configuraciones posibles crece rápidamente con el tamaño de la red.

Esta tesis estudia el uso de aprendizaje por refuerzo profundo para la asignación de recursos en redes inalámbricas reales. La política de decisión se parametriza mediante redes neuronales de grafos, modelos adecuados para representar sistemas donde los nodos interactúan mediante relaciones de interferencia o asociación. Esta elección permite que una misma política opere sobre redes de distinto tamaño y estructura, sin depender de un orden fijo de los nodos ni de una topología particular. El trabajo se apoya en medidas reales de la red de Ceibal, lo que permite evaluar el enfoque bajo condiciones representativas de una infraestructura en operación.

Se consideran dos escenarios complementarios. El primero modela una red de enlaces inalámbricos que interfieren entre sí y estudia la asignación conjunta de canal y potencia bajo restricciones de consumo. El segundo incorpora clientes dinámicos, llegadas y salidas de usuarios, asociaciones entre clientes y puntos de acceso, y períodos de decisión más realistas. Esta extensión transforma el problema en un sistema donde la política de decisión no tiene acceso completo al estado de la red, ya que clientes y asociaciones cambian de forma continua, y cuya dimensión varía en el tiempo, acercándolo a la operación cotidiana de una red inalámbrica administrada.

El trabajo desarrolla las formulaciones matemáticas de ambos escenarios, define estados, acciones y recompensas, y construye simuladores basados en datos reales para entrenar y evaluar políticas de decisión. También se comparan distintas estrategias de aprendizaje por refuerzo sobre una misma representación por grafos, con el objetivo de estudiar tanto la capacidad alcanzada como la estabilidad del entrenamiento y la posibilidad de operar sobre topologías variables.

En conjunto, la tesis plantea una metodología para combinar aprendizaje por refuerzo con redes neuronales de grafos en el diseño de mecanismos de control adaptativos para redes inalámbricas densas. La propuesta enfatiza que este tipo de políticas debe analizarse considerando no solo la capacidad promedio, sino también la estabilidad del aprendizaje, la observabilidad del estado, la dinámica de asociación de usuarios y la diversidad topológica de las redes reales.

Esta página ha sido intencionalmente dejada en blanco.

Tabla de contenidos

Agradecimientos	III
Resumen	V
1. Introducción	1
1.1. Motivación	1
1.2. Definición del problema	3
1.3. Objetivos	3
1.3.1. Objetivos específicos	3
1.4. Origen de los datos	4
1.5. Antecedentes y trabajos relacionados	4
1.6. Contribuciones del Trabajo	5
1.7. Estructura del Documento	6
2. Marco Teórico	7
2.1. Asignación de Recursos en Redes Inalámbricas	7
2.1.1. Canal inalámbrico	7
2.1.2. Tasa de Transmisión y Función de Desempeño Global	8
2.1.3. Gestión de Recursos de Radio	8
2.2. Aprendizaje por Refuerzo	9
2.2.1. ¿Por qué Aprendizaje por Refuerzo?	9
2.2.2. Proceso de Decisión de Markov	9
2.2.3. Aproximación Paramétrica de la Política con GNN	11
2.3. Herramientas de Implementación	13
2.3.1. Gymnasium	13
2.3.2. PyTorch Geometric	13
2.4. Síntesis	14
3. Asignación de Recursos en Redes de Enlaces Punto a Punto	15
3.1. Antecedente directo y contribuciones	15
3.1.1. El trabajo antecedente: Fernández et al. (2024)	15
3.1.2. Elementos heredados y aportes	16
3.2. Definición del problema	16
3.3. Formulación del problema	17
3.3.1. Elección de canal	17
3.3.2. Extensión a canal y potencia	18

Tabla de contenidos

3.4.	Arquitectura GNN-RL	19
3.4.1.	Entrada de la GNN	19
3.4.2.	Estructura de la GNN	21
3.4.3.	Entrenamiento con REINFORCE	22
3.4.4.	Entorno	23
3.5.	Conjunto de datos	24
3.5.1.	Aspectos generales	24
3.5.2.	Procesamiento	25
3.5.3.	Análisis	26
3.6.	Pruebas de funcionamiento	27
3.6.1.	Verificación de la asignación de canales	28
3.7.	Políticas de referencia	31
3.8.	Resultados	32
3.8.1.	Elección de canal	32
3.8.2.	Asignación de canal y potencia	34
3.9.	Síntesis	34
4.	Asignación de Recursos con Clientes Dinámicos	37
4.1.	Definición de variables	37
4.2.	Dinámica del sistema	38
4.3.	Formulación como problema de decisión	40
4.4.	Conjunto de datos y simulación	41
4.4.1.	Descripción del conjunto de datos y procesamiento básico	41
4.4.2.	Simulación de la dinámica temporal y topológica	42
4.4.3.	Construcción del grafo	45
4.5.	Arquitectura del modelo GNN-RL	48
4.5.1.	Codificación posicional	48
4.5.2.	Solución a la invarianza del signo	50
4.5.3.	Flujo global de la arquitectura GNN	51
4.5.4.	Detalle de las capas GNN	52
4.6.	Algoritmos de entrenamiento	54
4.6.1.	REINFORCE	54
4.6.2.	A2C	55
4.6.3.	PPO	56
4.7.	Políticas de referencia	57
4.8.	Resultados	59
4.8.1.	Asignación de canales con PPO	61
4.8.2.	Asignación de canal y potencia con PPO	62
4.8.3.	Comparativa entre REINFORCE, A2C y PPO	63
4.8.4.	Comparación Sticky Client	64
4.8.5.	Impacto del tamaño del período de decisión	65
4.8.6.	Impacto del POMDP	66
4.9.	Síntesis	67

Tabla de contenidos

5. Conclusiones y Trabajo futuro	69
5.1. Discusión final	69
5.2. Trabajo futuro	70
A. Apéndice	73
A.1. Repositorio de Código y Reproducibilidad	73
A.2. Declaración de uso de Inteligencia Artificial	73
Lista de Siglas	75
Referencias	77
Índice de tablas	79
Índice de figuras	80

Esta página ha sido intencionalmente dejada en blanco.

Capítulo 1

Introducción

1.1. Motivación

El espectro radioeléctrico es un recurso finito que múltiples transmisores deben compartir. Cuando varios de ellos operan simultáneamente en la misma banda de frecuencias, sus señales se superponen en el receptor, la interferencia co-canal (CCI, del inglés *Co-Channel Interference*) degrada la calidad de los enlaces y reduce la capacidad efectiva del sistema. Gestionar esa interferencia sin sacrificar cobertura es el problema central de la asignación de recursos en redes inalámbricas densas.

El problema no es nuevo, pero su escala actual lo transforma cualitativamente. Las redes celulares de quinta generación (5G) despliegan cientos de celdas pequeñas en áreas urbanas, donde la CCI entre celdas vecinas limita estructuralmente la capacidad cuando los recursos no se coordinan con precisión. Las constelaciones satelitales de órbita baja enfrentan el mismo desafío a escala global, coordinando el acceso de millones de terminales en tiempo real. En ambos casos, la asignación estática de recursos pierde efectividad a medida que aumenta la densidad de la red, porque las condiciones de interferencia cambian con mayor frecuencia y requieren una coordinación más fina entre transmisores.

La variabilidad temporal de la demanda se observa con claridad en redes Wi-Fi de alta concurrencia, como estadios, recitales o grandes eventos, donde muchos usuarios comparten recursos del espectro en un área reducida. Los entornos educativos no alcanzan necesariamente ese nivel de densidad, aunque también operan con recursos exigidos y con una demanda fuertemente variable. Un edificio escolar puede concentrar decenas de puntos de acceso (AP, del inglés *Access Point*), mientras la demanda se dispara en el horario de entrada y durante los recreos, por el uso intensivo de dispositivos móviles de los estudiantes, y desaparece completamente en vacaciones de julio. Una asignación de recursos dimensionada para estos intervalos de máxima demanda enfrenta un problema de optimización claramente distinto al que correspondería durante los períodos de baja actividad del resto de la jornada.

Además, la ausencia de un modelo matemático explícito de la dinámica del canal provoca que la optimización clásica requiera conocer con anterioridad cómo evoluciona el canal, cuántos usuarios estarán activos y cuándo se conectarán o

Capítulo 1. Introducción

desconectarán. En una red operativa, sin embargo, no es posible prever todos esos escenarios antes de tomar una decisión. El aprendizaje por refuerzo (RL, del inglés *Reinforcement Learning*) permite formular el problema en esos términos, mediante una política que observa el estado de la red, actúa en consecuencia y ajusta su comportamiento a partir de la retroalimentación recibida. Ese estado puede incluir, por ejemplo, el nivel de señal recibida (RSSI, del inglés *Received Signal Strength Indicator*) que cada AP detecta de sus vecinos, la carga actual de la red y la configuración activa. En el contexto de la gestión de recursos de radio (RRM, del inglés *Radio Resource Management*), el agente observa el estado del canal, decide qué canal y qué potencia asignar a los puntos de acceso, y recibe como retroalimentación la tasa de transmisión resultante. A lo largo del entrenamiento, la política aprende a tomar buenas decisiones de canal y potencia sin requerir en ningún momento un modelo explícito de la dinámica del canal.

Además de la selección del canal de operación, la potencia de transmisión constituye uno de los recursos fundamentales que deben administrarse en una red inalámbrica. Incrementar la potencia permite mejorar la intensidad de la señal recibida y ampliar el alcance de un enlace, favoreciendo la comunicación entre nodos más alejados o bajo condiciones de propagación desfavorables. Sin embargo, una mayor potencia también incrementa la interferencia generada sobre otros enlaces que operan en la misma banda de frecuencia o en canales parcialmente solapados, reduciendo la reutilización espacial del espectro y degradando el rendimiento global de la red. Por el contrario, una potencia insuficiente puede provocar enlaces inestables, menor capacidad y un aumento en las retransmisiones.

A su vez, la potencia transmitida está asociada al consumo energético de los equipos inalámbricos. En dispositivos alimentados mediante baterías, como sensores IoT, teléfonos móviles o equipos portátiles, una potencia superior a la necesaria reduce la autonomía disponible. Incluso en infraestructuras con alimentación permanente, minimizar la potencia empleada permite reducir el consumo energético y evitar configuraciones excesivamente conservadoras que generen interferencia innecesaria. Por lo tanto, la gestión de potencia representa un compromiso entre cobertura, calidad del enlace, interferencia, capacidad y eficiencia energética.

Representar el estado de la red como un grafo es la elección consistente con la física del problema, porque captura directamente la estructura de la red y las relaciones de CCI entre transmisores y receptores. Esta representación, además, admite redes de tamaño variable, y se busca que además las políticas aprendan sobre esa representación de forma invariante al orden en que se indexan los nodos [1]. Las redes neuronales de grafos (GNN, del inglés *Graph Neural Networks*) parametrizan la política directamente sobre el grafo y comparten los mismos pesos entre nodos y entre topologías de distinto tamaño, una propiedad relevante para nuestro caso de estudio, donde cada edificio educativo concentra un número distinto de APs. Esto permite entrenar una única política que generaliza a escenarios no observados durante el entrenamiento.

En temas de redes de datos, los algoritmos suelen evaluarse mediante simulación con datos sintéticos del canal pero, en este trabajo, se utilizan datos reales de la red inalámbrica de Ceibal [2], el proveedor de conectividad del programa educativo

uruguayo, que administra de forma centralizada miles de APs distribuidos en más de mil centros educativos en todo el país.

1.2. Definición del problema

Se estudian dos formulaciones de la RRM, cada una abordada en su propio capítulo y bajo supuestos distintos sobre cómo se representa el estado de la red.

La primera formulación, **Asignación de Recursos en redes de enlaces Punto a Punto** (ARPP, capítulo 3), modela el conjunto de enlaces punto a punto. Sobre esa representación se aborda la selección conjunta de canal y nivel de potencia discreta, con el objetivo de coordinar espectralmente los transmisores y estudiar si las políticas aprendidas se transfieren a edificios y períodos temporales no observados durante el entrenamiento. El espacio de configuraciones crece de manera combinatoria con el número de enlaces, canales y niveles de potencia, y la CCI genera dependencias mutuas entre las decisiones de todos los transmisores activos simultáneamente.

La segunda formulación, **Asignación de Recursos con Clientes Dinámicos** (ARCD, capítulo 4), aborda el problema de coordinación espectral sobre APs considerando ahora una red Wi-Fi, con clientes conectados explícitamente a la red. En el modelo, los clientes se conectan y desconectan en el tiempo, su asociación a los APs evoluciona y el agente debe gestionar el espectro para una población variable cuyo tamaño cambia en cada instante de decisión. Esto introduce dos desafíos ausentes en la ARPP. El primero es estructural, ya que el grafo sobre el que opera la política ya no tiene un número fijo de nodos, sino que crece y se decrece con cada conexión o desconexión de un cliente. El segundo es de información, ya que el agente observa quiénes están conectados y con qué señal, pero desconoce el tiempo de sesión restante de cada cliente conectado y cuánto falta para la llegada del próximo usuario. Esa incertidumbre aproxima el problema a las condiciones reales de operación de una red Wi-Fi, pero también lo hace considerablemente más difícil de abordar que la ARPP, donde el estado de la red queda determinado por las decisiones del agente.

1.3. Objetivos

El objetivo general del trabajo es estudiar la aplicación de RL profundo basado en GNN al problema de asignación de recursos en redes inalámbricas densas, utilizando medidas reales de canal tanto para el entrenamiento como para la evaluación de las políticas obtenidas.

1.3.1. Objetivos específicos

Los objetivos específicos son los siguientes:

1. Evaluar la capacidad de generalización de una política basada en aprendizaje

Capítulo 1. Introducción

por refuerzo sobre grafos para la ARPP en edificios y períodos temporales no observados durante el entrenamiento, utilizando datos reales.

2. Desarrollar una formulación de la ARCD con demanda variable, construyendo un modelo de tráfico a partir de medidas reales de la red de Ceibal y evaluar el desempeño de una política operando bajo condiciones de observabilidad total y parcial.
3. Comparar distintas estrategias de formulación y entrenamiento para la ARCD en términos de estabilidad del aprendizaje y tasa de transmisión alcanzada.

1.4. Origen de los datos

Los datos utilizados provienen de la infraestructura de red de Ceibal. Cada edificio educativo cuenta con múltiples APs administrados de forma centralizada por varios Controladores de LAN Inalámbrica (WLC, del inglés *Wireless LAN Controller*), lo que permite recolectar mediciones sistemáticas de la red en operación. Para obtener esas mediciones, cada AP debe abandonar brevemente su canal de operación habitual para enviar, en cada uno de los canales posibles, un paquete de sondeo (NDP, del inglés *Neighbor Discovery Protocol*) que los APs vecinos reciben y reportan al WLC junto con el RSSI medido; este mecanismo propietario de recolección de RSSI no forma parte del estándar 802.11 y su efecto sobre el servicio es imperceptible para los usuarios conectados.

Para la ARPP se dispone de medidas de atenuación del canal entre APs, recogidas cada quince minutos a lo largo del año lectivo 2018 (febrero a diciembre). Estos datos fueron procesados por Capdehourat et al. [2] y constituyen la matriz de canal que el agente observa en cada instante de decisión. La resolución temporal de quince minutos es suficiente para capturar la variabilidad diurna y estacional que caracteriza una red punto a punto.

Para la ARCD se utilizan registros del protocolo simple de gestión de red (SNMP, del inglés *Simple Network Management Protocol*) correspondientes a escaneos periódicos (tres veces al día), de los cuales se extraen medidas de RSSI por cliente. Dado que mediciones separadas por varias horas no permiten reconstruir sesiones continuas de clientes, estos registros se utilizan para estimar distribuciones empíricas de atenuación de canal y parametrizar un modelo sintético de tráfico con granularidad temporal más fina, desarrollado en el capítulo 4.

1.5. Antecedentes y trabajos relacionados

El punto de partida directo del presente trabajo es la propuesta de Fernández et al. [3], que entrena una GNN con REINFORCE para la asignación binaria de potencia sobre los mismos datos de Ceibal. Ese trabajo establece un resultado no trivial. Una política paramétrica basada en grafos puede superar heurísticas estáticas y transferirse a edificios no observados durante el entrenamiento. Lo que

1.6. Contribuciones del Trabajo

permanece abierto es si ese resultado se mantiene cuando el espacio de acciones crece y cuando los clientes aparecen explícitamente en el modelo.

La introducción de GNN como aproximador de política para problemas de CCI fue explorada por Eisen et al [1], quienes demostraron que la invarianza topológica de las GNN es especialmente valiosa cuando los grafos de interferencia varían en tamaño, pues los pesos aprendidos sobre una topología son directamente aplicables a otra sin reentrenamiento. La expresividad de la entrada a la GNN fue analizada por Kanatsoulis et al [4], cuyo resultado es el que motiva la elección de características de entrada de la GNN en el capítulo 3.

La tesis doctoral de Randall Carlevaro [5] es el antecedente más cercano al segundo escenario explorado en este trabajo. Randall Carlevaro formula el problema de asociación justa de usuarios como un proceso de decisión de Markov parcialmente observable en el que los usuarios arriban aleatoriamente y tienen duraciones de sesión finitas, y entrena una política basada en GNN con aprendizaje por refuerzo profundo, validando el enfoque con datos de una red 5G real en París. El trabajo muestra que la combinación GNN + RL produce políticas equitativas y estables bajo dinámica de usuarios, e introduce explícitamente el grafo bipartito estación base–usuario como representación del estado. Los dos elementos centrales de esa propuesta—el grafo bipartito heterogéneo y la formulación POMDP con dinámica de clientes—son también el punto de partida de la ARCD desarrollada en el capítulo 4, que adapta ambas ideas al contexto de la red Wi-Fi de Ceibal con múltiples APs y asignación de canal y potencia.

La tesis parte desde el análisis de Capdehourat et al. [2] sobre la propia red de Ceibal. A partir de más de 750 millones de mediciones de RSSI recopiladas durante diez meses, los autores muestran dos propiedades del canal que los modelos sintéticos convencionales no capturan: la asimetría entre enlaces, con $h_{ij} \neq h_{ji}$ en la mayoría de los pares, y la no estacionariedad de aproximadamente la mitad de las series temporales de RSSI. Ambas propiedades son verificables en los datos utilizados en el capítulo 3 y son la razón concreta por la que el entrenamiento se realiza sobre registros reales en lugar de sobre canal sintético.

El presente trabajo explora dos extensiones respecto a Fernández et al. [3]. La primera amplía el espacio de acciones desde la asignación binaria de potencia hacia la selección conjunta de canal y nivel de potencia discreta, lo que incrementa la complejidad combinatoria del problema. La segunda introduce clientes con dinámica explícita, lo que cambia la representación del estado, que pasa de un grafo de enlaces punto a punto a un grafo bipartito heterogéneo AP–cliente de dimensión variable. Ambas extensiones se evalúan sobre datos de la red operativa de Ceibal, lo que sitúa los resultados en condiciones representativas de despliegue.

1.6. Contribuciones del Trabajo

Los aportes de este trabajo se organizan en tres líneas principales: la extensión del problema de asignación de recursos, la construcción de entornos de evaluación basados en datos reales y el estudio de políticas de aprendizaje por refuerzo sobre grafos.

Capítulo 1. Introducción

En primer lugar, se extendió el problema de asignación de recursos estudiado previamente en la red de Ceibal. Además de la asignación binaria de potencia, se consideró la selección de canal y la asignación conjunta de canal y potencia discreta bajo restricciones de potencia promedio. Esta formulación permite evaluar si las políticas basadas en GNN siguen siendo útiles cuando el espacio de acciones crece y el problema se aproxima a decisiones de RRM más expresivas.

En segundo lugar, se formuló un nuevo escenario con clientes dinámicos. Este modelo incorpora llegadas y salidas de usuarios, asociación AP–cliente, períodos de decisión y observabilidad parcial. De esta forma, el problema pasa de un grafo homogéneo entre APs a un grafo heterogéneo AP–cliente de tamaño variable, más cercano a la operación real de una red Wi-Fi.

En tercer lugar, se construyeron entornos de simulación a partir de mediciones reales de Ceibal extraídas vía SNMP desde las WLCs de Cisco, incluyendo la atenuación del canal entre APs y los valores de RSSI con los que cada AP escucha a los clientes. Estos entornos permiten evaluar políticas en condiciones que conservan propiedades empíricas de la red, como variabilidad temporal, diferencias entre edificios y asimetrías en las relaciones de interferencia.

Finalmente, se implementaron y compararon políticas de aprendizaje por refuerzo profundo basadas en GNN para ambos escenarios. En la ARPP se estudió la transferencia de políticas a edificios no vistos durante el entrenamiento. En la ARCD se compararon REINFORCE, A2C y PPO, observando que todos superan una referencia aleatoria y que PPO ofrece el comportamiento más estable en los experimentos realizados. Además, para la implementación de estos hitos, se utilizó el entorno de trabajo de Gymnasium lo que permitió desarrollar un marco de experimentación modular, reproducible y fácilmente extensible para futuras investigaciones.

1.7. Estructura del Documento

Lo que resta del documento se organiza de la siguiente forma. El capítulo 2 sienta las bases teóricas sobre las que se apoya el resto del trabajo, cubriendo el modelo de canal inalámbrico, los procesos de decisión de Markov, los algoritmos de gradiente de política y el uso de GNN para aproximar políticas sobre estructuras de grafo. El capítulo 3 formula y resuelve la ARPP. Allí se valida el enfoque sobre datos reales de Ceibal y se estudia su generalización a topologías distintas de las usadas en el entrenamiento. La ARCD, donde los clientes se conectan y desconectan dinámicamente y el agente debe coordinar el espectro bajo esa variabilidad adicional, es el tema del capítulo 4, que compara distintos algoritmos de entrenamiento frente a ese desafío. Por último, el capítulo 5 retoma los objetivos planteados al inicio de la tesis para evaluar en qué medida se cumplieron, señala las limitaciones encontradas en el camino y propone posibles líneas de continuidad.

Capítulo 2

Marco Teórico

El problema de asignación de recursos involucra tres elementos que se formalizan por separado: una métrica de desempeño, un marco para la toma de decisiones secuencial y una arquitectura capaz de operar sobre la estructura de grafo de la red. Este capítulo desarrolla esos tres elementos en ese orden.

2.1. Asignación de Recursos en Redes Inalámbricas

2.1.1. Canal inalámbrico

La comunicación inalámbrica entre dos nodos está sujeta a atenuación; la potencia recibida es solo una fracción de la potencia transmitida. Dado un transmisor $t(i)$ del enlace i y un receptor $r(j)$ del enlace j en un tiempo t , definimos la atenuación de canal $h_{ij}(t) \in \mathbb{R}^+$, que resume los efectos de propagación entre ambos en un único escalar.

Cuando múltiples transmisores operan simultáneamente en el mismo canal, las señales de los demás transmisores llegan al receptor como CCI no deseada. La calidad de un enlace deja de depender únicamente de h_{ii} y pasa a depender de toda la configuración activa del sistema.

En todo receptor real, la señal recibida incluye fluctuaciones aleatorias generadas por la electrónica del receptor y por el entorno. Ese efecto se modela como ruido aditivo gaussiano blanco de potencia σ^2 . La relación señal a interferencia más ruido (SINR) combina este ruido con la dependencia de la calidad del enlace respecto a la configuración activa del sistema descrita arriba. Para el enlace i se escribe como

$$\text{SINR}_i = \frac{h_{ii} P_i}{\sigma^2 + \sum_{j \neq i} h_{ij} P_j \mathbf{1}[c_i = c_j]}, \quad (2.1)$$

donde P_i y c_i son la potencia y el canal asignados al enlace i , y el indicador $\mathbf{1}[c_i = c_j]$ restringe la suma a los enlaces j que operan en el mismo canal que el enlace i , porque solo esos generan CCI.

Capítulo 2. Marco Teórico

La ecuación anterior es común a los dos problemas estudiados en este trabajo. Lo que cambia entre ellos es el rol del receptor, que en la ARPP (capítulo 3) es el extremo de un enlace punto a punto, mientras que en la ARCD (capítulo 4) es un cliente asociado a un AP.

2.1.2. Tasa de Transmisión y Función de Desempeño Global

Dado que las redes inalámbricas modernas adaptan su esquema de modulación y codificación a las condiciones del canal, se asume que cada enlace opera a la capacidad de Shannon del canal, evitando así atarse a una implementación particular. La tasa espectral alcanzable en el enlace i , normalizada por el ancho de banda B y expresada en bits/s/Hz, es

$$r_i = \log_2(1 + \text{SINR}_i). \quad (2.2)$$

En este trabajo se adopta la suma directa de tasas como métrica de desempeño global, en línea con la formulación del trabajo previo de referencia [3]:

$$R_{\text{sum}} = \sum_{i \in \mathcal{L}} r_i, \quad (2.3)$$

donde $\mathcal{L}$ es el conjunto de enlaces activos en la red.

2.1.3. Gestión de Recursos de Radio

El problema de RRM consiste en elegir la asignación de canal y potencia para cada transmisor con el objetivo de maximizar R_{sum} . Esa elección no se realiza de forma continua, sino en instantes de decisión separados por un intervalo de tiempo. Entre dos actualizaciones consecutivas, la configuración elegida permanece activa y condiciona el funcionamiento de la red.

La dificultad no es solo espacial, por la interferencia entre transmisores que comparten canal, sino también temporal. Cuando se impone una restricción de potencia promedio, la acción presente consume parte de un presupuesto que debe respetarse a lo largo del tiempo. Transmitir con potencia alta puede aumentar la tasa instantánea, pero también reduce el margen disponible para mantener el consumo promedio bajo el límite permitido en ranuras futuras. Por eso, una política que optimiza únicamente la tasa del instante actual puede violar la restricción o tomar decisiones subóptimas a largo plazo. El aprendizaje por refuerzo resulta adecuado para este escenario porque aprende a elegir acciones considerando sus efectos futuros, sin requerir un modelo explícito de la dinámica de la red.

Se supone que cada enlace puede operar en uno de C canales y a uno de P niveles de potencia discretos. Para una red de M transmisores, el espacio de configuraciones conjuntas es finito pero crece exponencialmente con M , con un total de $(C \times P)^M$ combinaciones posibles. Para una red de veinte transmisores con tres canales y tres niveles de potencia el espacio supera los 10^{19} puntos, lo que descarta cualquier búsqueda exhaustiva. La naturaleza aleatoria del canal descarta las soluciones estáticas, ya que una asignación óptima para un estado particular no lo es necesariamente para otro.

2.2. Aprendizaje por Refuerzo

2.2.1. ¿Por qué Aprendizaje por Refuerzo?

El problema de RRM consiste en tomar decisiones sobre un sistema dinámico variante en el tiempo cuya evolución no se conoce en forma de modelo explícito. Los métodos de optimización clásica requieren conocer la función de transición del sistema para calcular la solución óptima, y el aprendizaje supervisado requiere ejemplos de decisiones óptimas para imitar. Ninguna de esas condiciones se cumple en el entorno considerado. El aprendizaje por refuerzo permite optimizar bajo un sistema dinámico que puede modelarse como un proceso de Markov, sin necesidad de conocer explícitamente esa dinámica [6, 7]. El problema de RRM es un caso de esa clase, ya que el sistema evoluciona de acuerdo a una dinámica desconocida, pero la suposición de que cumple la propiedad de Markov es razonable y habilita el uso directo de los algoritmos de RL.

2.2.2. Proceso de Decisión de Markov

Para aplicar aprendizaje por refuerzo es necesario describir el problema como una interacción repetida entre un agente y un entorno. En RRM, el agente observa el estado de la red, elige una configuración de canal y potencia, y luego recibe una recompensa asociada a la tasa obtenida. Esa decisión también modifica el estado desde el cual se tomará la próxima acción. Por lo tanto, no alcanza con optimizar cada instante por separado, sino que la formulación debe representar una secuencia de decisiones acopladas en el tiempo. La figura 2.1 resume este ciclo de interacción.

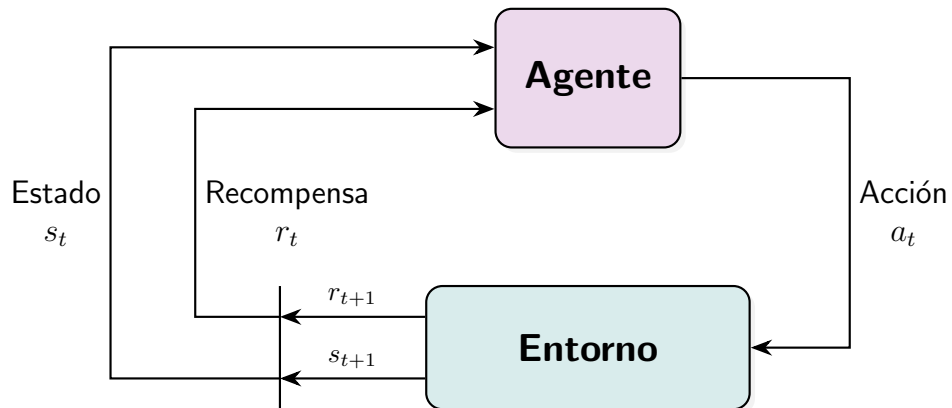


Figura 2.1: Ciclo de interacción entre el agente y el entorno en un problema de aprendizaje por refuerzo. En cada instante t el agente observa el estado s_t , ejecuta una acción a_t y el entorno responde con una recompensa r_{t+1} y un nuevo estado s_{t+1} .

El modelo estándar para formalizar la interacción mostrada en la figura 2.1 es el Proceso de Decisión de Markov (MDP, del inglés *Markov Decision Process*). Un

Capítulo 2. Marco Teórico

MDP se define por la cuádrupla $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$, donde $\mathcal{S}$ es el espacio de estados, $\mathcal{A}$ el espacio de acciones, $\mathcal{P}$ la dinámica de transición y $\mathcal{R}$ la función de recompensa. En cada instante discreto t , el agente observa un estado s_t , ejecuta una acción a_t , recibe una recompensa r_t y el entorno pasa a un nuevo estado s_{t+1} [6, 7].

En este trabajo, la política se representa mediante una familia paramétrica $\pi_\theta(a \mid s)$, con parámetros $\theta \in \mathbb{R}^d$. Dada una observación del estado, la política asigna una distribución de probabilidad sobre las acciones disponibles. El entrenamiento busca ajustar θ para aumentar la recompensa acumulada esperada, en lugar de buscar una acción óptima aislada para cada estado.

El problema se vuelve tratable bajo la suposición de que el sistema cumple la propiedad de Markov:

$$\mathcal{P}(s_{t+1} \mid s_t, a_t, s_{t-1}, a_{t-1}, \dots) = \mathcal{P}(s_{t+1} \mid s_t, a_t). \quad (2.4)$$

Es decir, la distribución del estado siguiente depende únicamente del estado y la acción actuales, no de toda la historia anterior. Esta propiedad es la que permite formular el problema dentro del marco de RL. Si la observación actual contiene información suficiente para predecir la evolución del sistema bajo una acción, el agente puede aprender una política a partir de transiciones estado-acción-recompensa sin modelar explícitamente toda la historia previa.

Optimización de la política

Una vez definido el MDP y parametrizada la política, resta especificar cómo se ajustan sus parámetros a partir de la experiencia recolectada. El criterio de entrenamiento consiste en maximizar el retorno esperado bajo la política π_θ [6, 7]. Para un episodio de horizonte Γ , ese objetivo se escribe como:

$$\theta^* = \arg \max_{\theta \in \mathbb{R}^d} J(\theta) = \arg \max_{\theta \in \mathbb{R}^d} \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^{\Gamma-1} \gamma^t R_{t+1} \right], \quad (2.5)$$

donde $\gamma \in [0, 1]$ es el factor de descuento. El teorema de Policy Gradient [7–9] provee el mecanismo para optimizar este objetivo sin conocer la probabilidad de transición $\mathcal{P}$:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^{\Gamma-1} G_t \nabla_\theta \log \pi_\theta(A_t \mid S_t) \right], \quad (2.6)$$

donde $G_t = \sum_{\tau=t}^{\Gamma-1} \gamma^{\tau-t} R_{\tau+1}$ es el retorno acumulado descontado a partir del paso t siguiendo la política π_θ . Esta estimación no requiere modelo del entorno, lo que la hace directamente aplicable al problema de RRM. En particular, REINFORCE (REward INcrement = Nonnegative Factor $\times$ Offset Reinforcement $\times$ Characteristic Eligibility) [9] es la implementación Monte Carlo de este estimador: se generan trayectorias siguiendo la política estocástica actual, se calcula el retorno observado y se actualizan los parámetros en la dirección $G_t \nabla_\theta \log \pi_\theta(A_t \mid S_t)$, aumentando la probabilidad de las acciones que produjeron mayor retorno [9].

2.2.3. Aproximación Paramétrica de la Política con GNN

Una vez restringida la búsqueda a políticas paramétricas π_θ , queda por elegir la familia paramétrica concreta. En principio, cualquier arquitectura que tome un conjunto de nodos y asigne una distribución de probabilidad sobre el espacio de canales y potencias para cada nodo es una candidata válida. Sin embargo, el problema de RRM tiene una estructura adicional que conviene explotar, ya que la red de enlaces no es un conjunto sin estructura sino un grafo, donde la relación entre dos nodos depende de su posición relativa y no de su índice. Una red neuronal estándar requeriría que el orden en que se listan los enlaces fuera informativo, lo cual no lo es. Las redes neuronales sobre grafos son la familia paramétrica natural para este problema porque operan directamente sobre la estructura de grafo y son independientes del orden de indexación de los nodos [1, 4, 10].

Fundamentos del message passing y transferibilidad

Formalmente, se representa la red mediante un grafo $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ con $N = |\mathcal{V}|$ nodos. Las características de los nodos se agrupan en una matriz $\mathbf{X} \in \mathbb{R}^{N \times F}$, donde la fila $\mathbf{x}_i$ contiene la información disponible para el transmisor i . La conectividad del grafo se codifica mediante una matriz de soporte $\mathbf{S} \in \mathbb{R}^{N \times N}$, por ejemplo la matriz de adyacencia o una versión normalizada de ella; en la aplicación de este trabajo, $\mathbf{S}$ toma el valor de la matriz de canal $\mathbf{H}$ normalizada por su norma espectral, como se detalla en el capítulo 3. El producto $\mathbf{S}\mathbf{X}$ reemplaza, para cada nodo, sus características por una combinación de las características de sus vecinos; de forma análoga, $\mathbf{S}^k\mathbf{X}$ agrega información proveniente de nodos a distancia de hasta k saltos.

Una convolución sobre grafos explota esa operación local mediante un filtro polinomial en $\mathbf{S}$:

$$\mathbf{Z} = \sum_{k=0}^K \mathbf{S}^k \mathbf{X} \mathbf{A}_k, \quad (2.7)$$

donde K define el alcance del filtro y las matrices $\mathbf{A}_k \in \mathbb{R}^{F \times G}$ son parámetros aprendibles. El término $k = 0$ conserva la información propia del nodo, mientras que los términos con $k > 0$ incorporan información de vecinos cada vez más lejanos. Una GNN apila capas de este tipo con una no linealidad $\sigma(\cdot)$:

$$\mathbf{X}^{(\ell+1)} = \sigma \left(\sum_{k=0}^K \mathbf{S}^k \mathbf{X}^{(\ell)} \mathbf{A}_k^{(\ell)} \right), \quad (2.8)$$

con $\mathbf{X}^{(0)} = \mathbf{X}$. Después de varias capas, cada nodo dispone de una representación que combina sus propias características con las de su entorno local en el grafo. Esta construcción se conoce como *message passing* [10], en la que cada nodo agrega información de sus vecinos iterativamente a través de las capas.

La equivarianza al reordenamiento de nodos es, precisamente, la propiedad que hace a la GNN adecuada para parametrizar π_θ . Los mismos filtros $\mathbf{A}_k^{(\ell)}$ se aplican a todos los nodos, por lo que la operación no depende del índice usado para enumerar

Capítulo 2. Marco Teórico

los APs. Si se permutan los nodos, se permutan de la misma forma las salidas. Esa propiedad permite aplicar los pesos aprendidos sobre un grafo de interferencia a grafos de distinto tamaño y topología sin reentrenamiento.

Arquitecturas de agregación y atención

A partir de la convolución sobre grafos, es posible mejorar el mecanismo con el cual se combina la información local. Una primera evolución es el enfoque de muestreo y agregación (SAGE, del inglés *S*Ample and *aggreG*atE) [11], diseñado para permitir el aprendizaje inductivo de representaciones. A diferencia de los métodos que optimizan un vector estático por nodo, este enfoque aprende funciones que aprovechan los atributos de los nodos para generalizar a estructuras no observadas durante el entrenamiento. Desde el punto de vista de un nodo, en cada nivel de profundidad k , el estado de un nodo se actualiza integrando primero la información de su vecindad $\mathcal{N}(v)$ mediante una función de agregación

$$\mathbf{h}_{\mathcal{N}(v)}^k \leftarrow \text{AGGREGATE}_k \left(\mathbf{h}_u^{k-1}, \forall u \in \mathcal{N}(v) \right), \quad (2.9)$$

donde el agregador puede tomar diversas formas, como el promedio o el *pooling* de características transformadas. A diferencia de otros operadores de convolución que promedian la información del nodo con la de su entorno, esta arquitectura emplea una concatenación explícita para generar el nuevo estado del nodo. Esto significa que el vector que representa al nodo central ($\mathbf{h}_v^{k-1}$) no se suma al vector de la vecindad ($\mathbf{h}_v^k$), sino que ambos se mantienen como bloques de información independientes antes de ser procesados por la matriz de pesos:

$$\mathbf{h}_v^k = \sigma \left(\mathbf{W}^k \cdot [\mathbf{h}_v^{k-1} \parallel \mathbf{h}_{\mathcal{N}(v)}^k] \right). \quad (2.10)$$

Por otro lado, también existe el concepto de autoatención, que permite que el modelo aprenda a priorizar la información de ciertos vecinos. Mientras que las convoluciones estándar suelen asignar pesos fijos o basados únicamente en la topología, las redes con mecanismos de atención para grafos (GAT, del inglés *Graph Attention Networks*) [12] calculan coeficientes de importancia α_{ij} para cada par de nodos conectados. Este coeficiente se obtiene aplicando un mecanismo de atención sobre las representaciones transformadas de los nodos i y j , normalizada mediante la función *Softmax* sobre los nodos vecinos:

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(\mathbf{a}^T [\mathbf{W}\mathbf{h}_i \parallel \mathbf{W}\mathbf{h}_j]))}{\sum_{k \in \mathcal{N}_i} \exp(\text{LeakyReLU}(\mathbf{a}^T [\mathbf{W}\mathbf{h}_i \parallel \mathbf{W}\mathbf{h}_k]))}. \quad (2.11)$$

La nueva representación se genera entonces como una combinación lineal ponderada de las características de sus vecinos,

$$\mathbf{h}'_i = \sigma \left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W}\mathbf{h}_j \right), \quad (2.12)$$

lo que permite diferenciar la contribución de cada vecino y mejora la representación de grafos con atributos de aristas heterogéneos.

2.3. Herramientas de Implementación

Estos enfoques teóricos de agregación por concatenación y mecanismos de atención tienen una implementación disponible mediante las capas SAGEConv y GATConv, respectivamente.

2.3. Herramientas de Implementación

2.3.1. Gymnasium

Gymnasium [13] es una biblioteca de Python para implementar y evaluar entornos de aprendizaje por refuerzo. Provee una interfaz estándar para definir esos entornos, separando explícitamente el entorno del agente. La interfaz se reduce a dos métodos. `reset()` inicializa el entorno y devuelve la observación inicial s_0 ; `step(a)` recibe la acción a_t elegida por el agente, avanza la dinámica interna del entorno y devuelve la observación siguiente s_{t+1} , la recompensa r_t y dos indicadores booleanos distintos, `terminated`, que señala si se alcanzó un estado terminal definido por el propio MDP, y `truncated`, que señala si el episodio se cortó por una condición externa al MDP, como un límite de pasos. Esta estructura corresponde directamente a la tupla estado-acción-recompensa del MDP de la sección 2.2.2, de modo que la dinámica de la red queda implementada sin exponer al agente ningún detalle de cómo se calcula esa dinámica internamente. Además de estos dos métodos, cada entorno declara explícitamente su espacio de observación y su espacio de acción.

En este trabajo, Gymnasium se usa principalmente como una convención de implementación, permite escribir los problemas de asignación de recursos con la misma interfaz independientemente de los detalles de cada escenario. Así, el agente no necesita conocer cómo se actualiza internamente el canal, cómo se calcula la recompensa ni cómo se representa la red; solo interactúa con el entorno mediante observaciones, acciones y recompensas. Esta separación es útil porque permite reutilizar el mismo bucle de entrenamiento con distintos algoritmos de RL y modificar la dinámica del entorno sin cambiar la implementación del agente.

Los capítulos siguientes especifican cómo se instancian estos elementos para la ARPP y la ARCD. Allí se definen las observaciones, los espacios de acción, la dinámica temporal y la recompensa de cada problema. En este capítulo basta destacar que Gymnasium provee la estructura común sobre la que se implementan esas variantes.

2.3.2. PyTorch Geometric

PyTorch Geometric (PyG) [14] es una biblioteca construida sobre PyTorch para desarrollar modelos de aprendizaje profundo sobre datos con estructura de grafo. PyG implementa directamente la operación de convolución de la ecuación (2.7), junto con el resto de las operaciones de *message passing* y agregación de vecinos, y provee además una colección de arquitecturas de GNN ya implementadas. Esto permite concentrar el desarrollo en la definición del problema y en la adaptación

Capítulo 2. Marco Teórico

de la arquitectura al caso de uso específico, en lugar de reimplementar la mecánica interna de la convolución sobre grafos.

Un grafo se representa en PyG mediante un par de estructuras. La matriz de características $\mathbf{X}$ de la sección 2.2.3 se almacena como un tensor $\mathbf{x}$ con una fila por nodo, y la conectividad codificada en la matriz de soporte $\mathbf{S}$ se almacena como una lista de aristas `edge_index`, en lugar de la matriz densa $N \times N$ completa. Esta representación dispersa es la que permite a PyG trabajar naturalmente con grafos de distinto tamaño dentro del mismo flujo de entrenamiento; la cantidad de APs y clientes activos puede cambiar entre instantes o entre escenarios, y el grafo correspondiente se construye sin ajustar el tamaño de ninguna estructura de datos a mano.

En un edificio con N APs, esto se traduciría en lo siguiente. Cada AP sería un nodo, de modo que $\mathbf{x}$ reuniría las características de canal y potencia de todos los APs activos en ese instante. Cada par de APs cuya interferencia mutua resulte relevante aportaría una arista a `edge_index`, de modo que el grafo resultante reflejaría únicamente esas relaciones de CCI, y no una malla completa entre todos los APs. Un edificio con pocos APs produciría así un grafo pequeño y disperso, mientras que uno con muchos APs concentrados en poco espacio produciría un grafo más denso, sin que la arquitectura de la GNN deba modificarse entre uno y otro.

2.4. Síntesis

Con los elementos de este capítulo es posible traducir el problema de asignación de recursos a términos operacionales. Las ecuaciones (2.1)–(2.3) definen qué se quiere maximizar y cómo se mide. La sección 2.2.1 justifica el uso de RL en este entorno, y el MDP de la sección 2.2.2 lo formaliza como una secuencia de decisiones; el estimador de Policy Gradient de la ecuación (2.6) provee el mecanismo de aprendizaje sin modelo. La GNN de la sección 2.2.3 aporta la arquitectura que respeta la estructura de grafo de la red, con equivarianza topológica y capacidad de generalizar a topologías nuevas.

El capítulo 3 ensambla estos tres bloques para la ARPP, donde la SINR define la recompensa, el MDP encuadra las decisiones sobre el grafo de interferencia entre enlaces punto a punto, y la GNN parametriza la política. El capítulo 4 reutiliza ese encuadre con dos cambios, ya que el grafo pasa a ser bipartito AP–cliente, y el MDP se reemplaza por un proceso de decisión con observabilidad parcial, que captura precisamente la incertidumbre introducida por la dinámica de clientes.

Capítulo 3

Asignación de Recursos en Redes de Enlaces Punto a Punto

En una red de enlaces punto a punto, la decisión de transmisión de cada enlace afecta a los demás mediante la interferencia cocanal. La asignación de recursos debe, por lo tanto, coordinar las transmisiones para aumentar la tasa agregada y, al mismo tiempo, respetar una restricción de potencia promedio por enlace. Como esta restricción se evalúa a lo largo del tiempo, la política debe equilibrar la recompensa inmediata de transmitir con el consumo de potencia acumulado.

Este capítulo parte del enfoque de aprendizaje por refuerzo parametrizado mediante redes neuronales de grafos (GNN) presentado en [1] y lo extiende a dos espacios de acciones: la selección de canal y la asignación conjunta de canal y potencia discreta. El objetivo es estudiar el desempeño de la política al ampliar el conjunto de acciones y evaluar su transferencia a otras topologías y a períodos no utilizados durante el entrenamiento.

3.1. Antecedente directo y contribuciones

El antecedente directo es el trabajo de Fernández et al. [3], que estudia la asignación de potencia en redes inalámbricas reales mediante GNN y aprendizaje por refuerzo. A continuación se resumen su formulación y sus resultados, y se delimitan los elementos que este proyecto conserva y modifica.

3.1.1. El trabajo antecedente: Fernández et al. (2024)

Fernández et al. [3] consideran un único canal y un espacio de acciones binario: cada enlace transmite a potencia p_0 o permanece apagado. La política, parametrizada por una GNN, se entrena para maximizar la tasa agregada sujeta a una restricción de potencia promedio para toda la red:

$$\max_{\Phi} \quad \mathbb{E} \left[\sum_{i=1}^M \log(1 + \text{SINR}_i) \right] \quad \text{s.t.} \quad \mathbb{E}[\mathbf{1}^T \mathbf{p}(\mathbf{H}, x)] \leq P_{\max}, \quad (3.1)$$

Capítulo 3. Asignación de Recursos en Redes de Enlaces Punto a Punto

donde $\Phi(\mathbf{H})$ denota la política parametrizada por la GNN.

La restricción se incorpora mediante un esquema primal–dual basado en REINFORCE [1], con un único multiplicador dual μ asociado al consumo de la red. El modelo se entrena con mediciones de RSSI de Plan Ceibal correspondientes al edificio 856, una red de 20 nodos en la banda de 2,4 GHz.

Los experimentos de [3] estudian la transferencia de la política tanto entre redes como entre períodos. Para evaluar la transferencia espacial, los pesos entrenados en el edificio 856 se aplican, sin reentrenamiento, a edificios con topologías y cantidades de nodos diferentes. Para evaluar la transferencia temporal, la política entrenada con datos de marzo a agosto se prueba con mediciones de septiembre a diciembre. En ambos casos, la política aprendida supera las referencias consideradas en ese trabajo.

3.1.2. Elementos heredados y aportes

Este proyecto conserva las mediciones de Plan Ceibal de 2018 y el procedimiento de preprocesamiento empleado para construir la matriz de atenuaciones $\mathbf{H}_t$ a partir del RSSI. También mantiene el entrenamiento primal–dual mediante REINFORCE para imponer restricciones promedio de largo plazo y el uso de una GNN para obtener una política equivariante a permutaciones y aplicable a grafos de distintos tamaños. El esquema de evaluación sigue el mismo criterio: la política se entrena en el edificio 856 y se aplica, sin reentrenamiento, a otras topologías y períodos.

Sobre esta base, el primer cambio consiste en ampliar el espacio de acciones. Se consideran dos variantes multicanal: en la primera, cada enlace selecciona uno de tres canales o permanece inactivo; en la segunda, selecciona además uno de dos niveles de potencia. Esta última variante da lugar a siete acciones por enlace, incluidas las seis combinaciones de canal y potencia y la opción de no transmitir.

La formulación también reemplaza la restricción de potencia promedio para toda la red por una restricción independiente para cada enlace. En consecuencia, se utiliza el vector de multiplicadores $\boldsymbol{\mu} = (\mu_1, \dots, \mu_M)^T$, donde cada componente está asociado al presupuesto de potencia de un enlace.

Otra modificación corresponde a las entradas de la GNN. En lugar de emplear características constantes, se incorporan características topológicas deterministas basadas en ciclos cerrados.

Por último, antes de evaluar los modelos con las mediciones reales, se realizan pruebas de funcionamiento en dos topologías sintéticas con soluciones conocidas: interferencia central e interferencia total. Para la comparación en escenarios multicanal se emplean dos políticas de referencia: el acceso aleatorio no coordinado y la asignación estática de canal.

3.2. Definición del problema

Se retoma el modelo de enlaces punto a punto introducido previamente. En este capítulo la unidad de decisión es siempre el enlace, cada enlace elige si transmite, en qué canal lo hace y, en la segunda variante, con qué canal y nivel de potencia. La

3.3. Formulación del problema

topología de enlaces se considera fija durante la evaluación. Además, las decisiones están sujetas a una restricción de potencia promedio por enlace, que acopla las acciones a lo largo del tiempo.

Sean M enlaces indexados por $i = 1, \dots, M$ en cada ranura t , el estado de propagación se representa mediante la matriz $\mathbf{H}_t \in \mathbb{R}^{M \times M}$. Una vez fijada la topología, la entrada $[\mathbf{H}_t]_{ij} = h_{ij}(t)$ mide el acoplamiento desde el transmisor del enlace i hacia el receptor del enlace j ; los términos diagonales representan la atenuación directa y los términos fuera de la diagonal representan interferencia potencial. Con esta matriz se calculan la señal útil de cada enlace y la interferencia generada por los demás enlaces activos en el mismo canal. La calidad de cada enlace se mide mediante la SINR definida en la ecuación (2.1), y la tasa alcanzable se obtiene aplicando la fórmula de Shannon (2.2).

El objetivo es aprender una política que, dado el estado del canal y el estado de la restricción de potencia, elija acciones para cada enlace. Como la política es aleatoria, en cada ranura se muestrea una acción de la cual surge la asignación de canal y potencia $\{(c_i, p_i)\}_{i=1}^M$. Durante el entrenamiento se busca aumentar la tasa agregada definida en la ecuación (2.3) a lo largo del tiempo, sin superar la potencia promedio permitida.

3.3. Formulación del problema

El problema definido en la sección anterior se aborda en dos variantes de complejidad creciente, que difieren en el espacio de acciones disponible para cada enlace. La primera variante fija la potencia y solo elige el canal; la segunda extiende la decisión a canal y nivel de potencia conjuntamente. Ambas comparten la misma función objetivo y la misma restricción de potencia promedio. Esta restricción es el elemento que vincula las decisiones a través del tiempo, transmitir más en una ranura aumenta la tasa instantánea, pero también consume presupuesto y modifica el margen de potencia disponible para ranuras futuras.

3.3.1. Elección de canal

En la primera variante, cada enlace selecciona uno entre C canales disponibles o no transmite. La potencia se fija en p_0 para el canal seleccionado y en cero para los demás:

$$p_{i,c}(\mathbf{H}_t) = \begin{cases} p_0, & \text{si el enlace } i \text{ transmite en el canal } c, \\ 0, & \text{en caso contrario.} \end{cases} \quad (3.2)$$

El problema no se formula como una maximización puramente instantánea de tasa. Además de obtener alta capacidad, se busca controlar el uso de energía de cada enlace a lo largo del tiempo. Esta restricción puede interpretarse como un presupuesto energético asociado a la autonomía de los dispositivos, por ejemplo para prolongar la duración de la batería o limitar el consumo promedio de la red. Por

Capítulo 3. Asignación de Recursos en Redes de Enlaces Punto a Punto

eso se impone una restricción de potencia promedio por enlace:

$$\mathbb{E} \left[\sum_{c=1}^C p_{i,c}(\mathbf{H}) \right] \leq P_{\text{máx},i}, \quad \forall i. \quad (3.3)$$

La forma de incorporar esta restricción sigue la idea lagrangiana de Eisen et al [1]. En el trabajo de Fernández et al [3] la restricción de potencia se representa mediante un único multiplicador dual; aquí se extiende el mismo mecanismo a un multiplicador independiente por enlace. La recompensa dual usada para entrenar la política queda definida como:

$$\mathcal{R}(\mathbf{p}, \boldsymbol{\mu}) = - \sum_{i=1}^M r_i - \sum_{i=1}^M \mu_i \left(\mathbb{E} \left[\sum_{c=1}^C p_{i,c}(\mathbf{H}) \right] - P_{\text{máx},i} \right). \quad (3.4)$$

El multiplicador no es un parámetro fijo de la política, sino una variable del sistema. En cada paso se actualiza siguiendo el gradiente de la violación de potencia,

$$\mu_{i,t+1} = \mu_{i,t} + \epsilon_{\mu} \left(\mathbb{E} \left[\sum_{c=1}^C p_{i,c}(\mathbf{H}) \right] - P_{\text{máx},i} \right). \quad (3.5)$$

Así, μ_i aumenta cuando el consumo promedio supera el presupuesto y disminuye cuando la restricción se satisface con holgura. Esta actualización formaliza la memoria temporal mencionada en la sección 3.2 y a su vez el efecto de las decisiones pasadas queda resumido en $\boldsymbol{\mu}$. Por eso, al incorporar ese vector al estado, no hace falta recordar toda la trayectoria previa; basta conocer el estado aumentado actual para describir su evolución en el siguiente paso. El problema conserva así una estructura markoviana, pero sobre un estado que incluye explícitamente el presupuesto dual acumulado, lo que justifica tratarlo mediante RL.

3.3.2. Extensión a canal y potencia

La elección de canal con potencia fija captura la dimensión espectral del problema pero no la energética. En redes reales, reducir la potencia de un enlace disminuye la CCI que genera sobre sus vecinos, aunque a costa de reducir su propia tasa. La segunda variante extiende la decisión para aprender a explotar ese compromiso de forma adaptativa.

El rango de potencias disponibles se discretiza en dos niveles $[p_{\text{máx}}/2, p_{\text{máx}}]$. El espacio de acciones se divide en una acción de no transmitir más todas las combinaciones posibles de canal y nivel de potencia, con cardinalidad $1 + C \times P$. La tabla 3.1 detalla el espacio resultante para $C = 3$ canales y $P = 2$ niveles. La acción de no transmitir rompe la simetría del espacio, las $C \times P$ combinaciones activas son simétricas entre sí, pero la opción de no transmitir no tiene análogo en la dimensión de canal y potencia. La función objetivo descrita en la ecuación (3.4) y la restricción de potencia promedio (3.3) se mantienen iguales que en la variante anterior. Lo que cambia es la exigencia sobre la política, la GNN debe aprender no solo a separar enlaces en canales distintos sino también a modular la potencia individual para reducir la CCI generada sobre los vecinos.

3.4. Arquitectura GNN-RL

Tabla 3.1: Espacio discreto de acciones por enlace para $C = 3$ canales y $P = 2$ niveles de potencia. La opción 0 corresponde a que ese enlace no transmite; las opciones 1–6 asignan canal y potencia de forma conjunta.

Opción a_i	Canal c_i	Potencia p_i
0	—	0
1	0	$p_{\text{máx}}/2$
2	0	$p_{\text{máx}}$
3	1	$p_{\text{máx}}/2$
4	1	$p_{\text{máx}}$
5	2	$p_{\text{máx}}/2$
6	2	$p_{\text{máx}}$

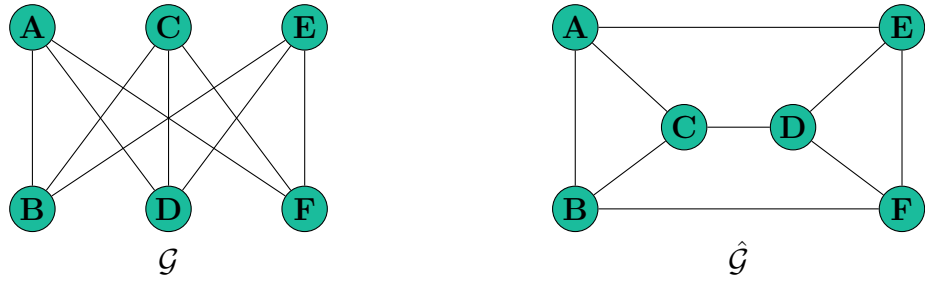


Figura 3.1: Ejemplo esquemático de dos grafos evaluados con una entrada constante. El grafo $\mathcal{G}$ corresponde a la topología de interferencia total usada en la verificación de funcionamiento.

3.4. Arquitectura GNN-RL

3.4.1. Entrada de la GNN

Como se discutió en la sección 2.2.3, se utiliza una GNN como familia paramétrica de políticas por su equivarianza al reordenamiento de nodos y su capacidad de operar sobre grafos de distinto tamaño. Una vez fijada esa familia de funciones, resta decidir qué señal se entrega como entrada a cada nodo. En principio, la formulación del problema solo especifica la estructura del grafo de interferencia, no atributos propios para distinguir a un enlace de otro. Una elección usual en este contexto es usar una entrada constante, por ejemplo el vector $\mathbf{1}$.

La limitación de esa elección se ilustra en la figura 3.1. El grafo $\mathcal{G}$ coincide con la topología de interferencia total que se usa luego como prueba de funcionamiento en la sección 3.6.1, donde todos los pares de enlaces se interfieren mutuamente. El segundo grafo, $\hat{\mathcal{G}}$, tiene una conectividad distinta, pero sirve para mostrar que una entrada uniforme no aporta identidad inicial a los nodos. El problema de expresividad se vuelve evidente al analizar los dos grafos de la figura 3.1. Como todos los nodos de $\mathcal{G}$ y $\hat{\mathcal{G}}$ tienen el mismo grado, el vector constante $\mathbf{1}$ es un vector propio en ambos grafos y, por lo tanto, es ortogonal a todos los vectores propios restantes. Como consecuencia, si la entrada de la GNN se inicializa de

Capítulo 3. Asignación de Recursos en Redes de Enlaces Punto a Punto

forma constante con este vector, los grados de los nodos son la única información que las capas de *message passing* son capaces de procesar. En este caso, la GNN no puede distinguir nodos con el mismo grado, por lo que puede asignarles la misma representación y, por tanto, la misma política. Esto es indeseable, porque enlaces ubicados en posiciones distintas del grafo podrían requerir decisiones de canal o potencia diferentes.

Una forma de evitar este problema es usar ruido blanco como entrada, combinado con activación cuadrática y normalización, lo que cuenta con respaldo teórico [4]. El ruido rompe la simetría inicial porque cada nodo recibe una señal distinta. Sin embargo, esa solución también agrega variabilidad a las observaciones del agente y puede dificultar la convergencia del entrenamiento.

La alternativa adoptada es determinista. Para cada nodo i se toman las diagonales de las sucesivas potencias de la matriz de canal normalizada $\mathbf{S}$, esto es, los valores $[\mathbf{S}^k]_{ii}$ para $k = 0, 1, \dots, D - 1$. Cuando $\mathbf{S}$ es una matriz de adyacencia binaria, $[\mathbf{S}^k]_{ii}$ cuenta exactamente cuántos caminos cerrados de longitud k pasan por el nodo i , de ahí que Kanatsoulis y Ribeiro [4] denominen a estas características *k-hop closed loops*. Nodos en posiciones estructuralmente distintas del grafo producen vectores de características distintos para algún k , lo que rompe la indistinguibilidad que afecta a la entrada constante. A diferencia del ruido, estos valores son una función determinista de la topología del grafo y no introducen varianza adicional al estimador del gradiente.

A partir de la matriz de adyacencia $\mathbf{S}$, se construye la matriz de características $\mathbf{X} \in \mathbb{N}^{N \times D}$ como:

$$\mathbf{X} = [\text{diag}(\mathbf{S}^0), \text{diag}(\mathbf{S}^1), \dots, \text{diag}(\mathbf{S}^{D-1})], \quad (3.6)$$

donde la k -ésima columna incorpora los conteos de caminos cerrados de longitud k .

Volviendo al ejemplo de la figura 3.1, si fijamos $D = 4$, las matrices de características $\mathbf{X} \in \mathbb{N}^{6 \times 4}$ para el grafo $\mathcal{G}$ y $\hat{\mathbf{X}}$ para el grafo $\hat{\mathcal{G}}$ quedan determinadas como:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 3 & 0 \\ 1 & 0 & 3 & 0 \\ 1 & 0 & 3 & 0 \\ 1 & 0 & 3 & 0 \\ 1 & 0 & 3 & 0 \\ 1 & 0 & 3 & 0 \end{bmatrix}, \quad y \quad \hat{\mathbf{X}} = \begin{bmatrix} 1 & 0 & 3 & 2 \\ 1 & 0 & 3 & 2 \\ 1 & 0 & 3 & 2 \\ 1 & 0 & 3 & 2 \\ 1 & 0 & 3 & 2 \\ 1 & 0 & 3 & 2 \end{bmatrix}. \quad (3.7)$$

Como se puede observar explícitamente en las matrices, todos los nodos comparten las mismas características debido a la simetría de cada grafo.

Cada columna de estas matrices corresponde a $[\mathbf{S}^k]_{ii}$ para $k = 0, 1, 2, 3$:

- La primera columna ($k = 0$) es la diagonal de la identidad $\mathbf{S}^0 = \mathbf{I}$, resultando en un vector unitario para todos los nodos.
- La segunda columna ($k = 1$) es la diagonal de la matriz de adyacencia $\mathbf{S}$, cuyos valores son nulos dado que no existen bucles propios en los nodos.

3.4. Arquitectura GNN-RL

- La tercera columna ($k = 2$) corresponde al grado de cada nodo. Como ambos grafos son 3-regulares (todos los nodos tienen exactamente 3 vecinos), esta columna es constante e igual a 3. Note que si tomáramos $D \leq 3$, los vectores de características para ambos grafos serían idénticos e indistinguibles.
- La cuarta columna ($k = 3$) cuenta la cantidad de caminos cerrados de longitud 3 (es decir, el doble de la cantidad de triángulos en los que participa cada nodo). En $\mathcal{G}$, al tratarse de un grafo bipartito completo ($K_{3,3}$), no existen ciclos de longitud impar y la columna es nula. En cambio, en $\hat{\mathcal{G}}$ (el prisma triangular), cada nodo participa exactamente en un triángulo (recorrible en dos sentidos, horario y antihorario), dando como resultado un valor de 2 para todos los nodos.

Por lo tanto, se concluye que al tomar $D \geq 4$, la GNN puede romper la simetría y distinguir estructuralmente ambos grafos, superando la limitación de la entrada constante inicial.

Es importante notar que, si bien dentro de un grafo perfectamente regular todos los nodos conservan vectores idénticos (y por ende la GNN carece de información para diferenciarlos internamente), las características k -hop resuelven el problema fundamental de la transferibilidad al lograr distinguir exitosamente entre topologías globales distintas (como $\mathcal{G}$ y $\hat{\mathcal{G}}$).

3.4.2. Estructura de la GNN

La arquitectura de la GNN implementada en PyTorch Geometric (2.3.2) consiste en una secuencia de transformaciones que operan sobre las características antes mencionadas $\mathbf{X}$ y la matriz de canal $\mathbf{H}$. En primer lugar, se emplea un codificador lineal inicial, que denominamos *codificador estructural*, para proyectar el vector de características de cada enlace desde su dimensión original D hacia un espacio de dimensión constante M , formalmente:

$$\mathbf{h}_i^{(0)} = \mathbf{W}\mathbf{x}_i + \mathbf{b}, \quad (3.8)$$

donde $\mathbf{x}_i \in \mathbb{N}^D$ (con $D = K + 1$) es el vector de características iniciales del nodo i , $\mathbf{W} \in \mathbb{R}^{M \times D}$ y $\mathbf{b} \in \mathbb{R}^M$ son los parámetros aprendibles de la capa lineal, y $\mathbf{h}_i^{(0)} \in \mathbb{R}^M$ es la representación inicial proyectada del nodo i .

A continuación, la información de cada nodo se actualiza mediante un bloque de L capas de convolución en grafos. Para ello se utiliza TAGConv (*Topology Adaptive Graph Convolutional Networks*) [15], siguiendo la forma general de la convolución sobre grafos definida en la ecuación (2.7).

Como se observa en la figura 3.2, las primeras $L - 1$ capas TAGConv mantienen la dimensionalidad de las representaciones en M y aplican una función de activación no lineal LeakyReLU. El uso de LeakyReLU evita el problema de los gradientes desvanecidos y asegura estabilidad durante la propagación hacia atrás. La última capa TAGConv proyecta las representaciones ocultas hacia la dimensión de salida requerida, que equivale a la cardinalidad del espacio de acciones del enlace. En esta

Capítulo 3. Asignación de Recursos en Redes de Enlaces Punto a Punto

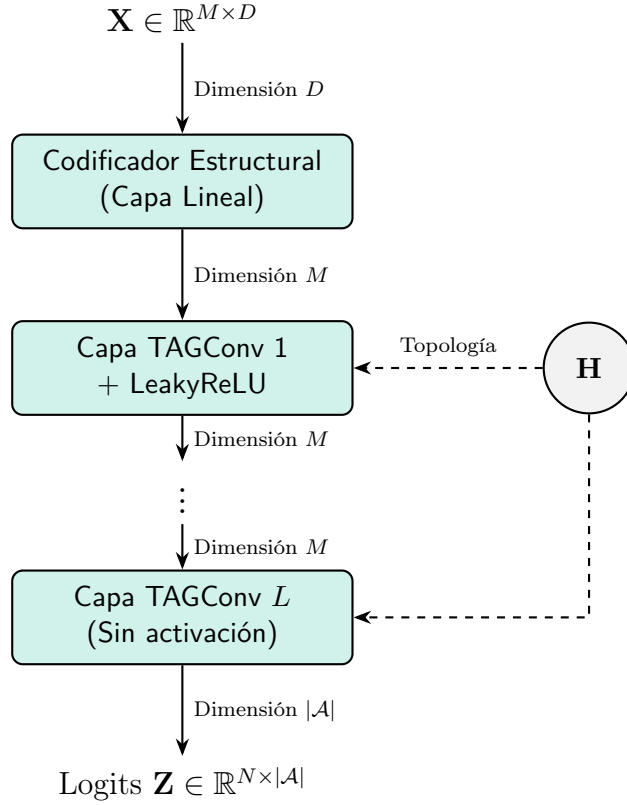


Figura 3.2: Diagrama en bloques de la arquitectura de la GNN. Las características iniciales $\mathbf{X}$ pasan por una capa lineal para homogeneizar su dimensionalidad, seguido de múltiples capas convolucionales TAGConv que agregan información de los vecinos ponderadas por la matriz de canal $\mathbf{H}$. La red finaliza arrojando los *logits* que determinan la preferencia del agente por cada acción en el espacio $\mathcal{A}$.

última etapa no se aplica ninguna función de activación, obteniendo de forma directa los *logits* (puntuaciones numéricas no normalizadas que reflejan la preferencia del agente por cada acción antes de ser transformadas en probabilidades).

3.4.3. Entrenamiento con REINFORCE

Con la entrada definida en la ecuación (3.6), la política queda completamente especificada, la GNN recibe $\mathbf{X}$ y $\mathbf{H}$, produce para cada nodo un vector de valores reales *logits* con una componente por cada acción discreta disponible, y la función Softmax definida como:

$$[\text{Softmax}(\mathbf{z})]_k = \frac{e^{z_k}}{\sum_j e^{z_j}}, \quad (3.9)$$

transforma ese vector en una distribución de probabilidad válida sobre las acciones del enlace. La elección de Softmax no es arbitraria, garantiza que la política π_θ sea diferenciable respecto a los parámetros θ , condición necesaria para aplicar el estimador de gradiente de Policy Gradient. De esa distribución se muestrea la

acción ejecutada en cada instante de decisión, de modo que la política es estocástica durante el entrenamiento y puede volverse aproximadamente determinista a medida que los logits se polarizan.

El entrenamiento sigue el algoritmo REINFORCE [9], definido en el marco teórico como una implementación Monte Carlo de Policy Gradient. En este capítulo solo se particulariza ese procedimiento a la recompensa dual y al espacio de acciones definido para el problema.

Como se detalló en la formulación del problema (ecuaciones (3.4) y (3.5)), la restricción de potencia promedio se incorpora directamente en el costo mediante multiplicadores duales μ_i por cada enlace. Durante el entrenamiento, estos multiplicadores se actualizan, guiando a la política para que aprenda a equilibrar la tasa de transmisión instantánea con el consumo energético a largo plazo.

Adicionalmente, para mitigar la convergencia prematura a políticas deterministas subóptimas y fomentar la exploración del agente, la función de pérdida incorpora una regularización por entropía ponderada por el parámetro β [16]. Para una política evaluada en el estado $(\mathbf{X}, \mathbf{H}_t)$, la entropía total sobre los M enlaces se define como

$$\mathcal{H}(\pi_\theta(\cdot \mid \mathbf{X}, \mathbf{H}_t)) = - \sum_{i=1}^M \sum_{a \in \mathcal{A}} \pi_\theta(a \mid \mathbf{X}, \mathbf{H}_t)_i \log \pi_\theta(a \mid \mathbf{X}, \mathbf{H}_t)_i, \quad (3.10)$$

donde $\mathcal{A}$ es el conjunto de acciones disponibles para cada enlace y $\pi_\theta(a \mid \mathbf{X}, \mathbf{H}_t)_i$ es la probabilidad asignada a la acción a en el enlace i . En la pérdida se resta el término $\beta \mathcal{H}$, de modo que las políticas de baja entropía resultan menos favorables durante el entrenamiento. Esto evita que la política se vuelva determinista demasiado pronto y mantiene exploración sobre el espacio de acciones.

Para la actualización de los parámetros θ , se emplea el optimizador Adam [17], el cual adapta dinámicamente la tasa de aprendizaje para cada peso utilizando estimaciones de los momentos de primer y segundo orden del gradiente. El algoritmo 1 detalla el procedimiento de optimización completo, que produce los pesos θ^* utilizados en los experimentos de la sección 3.8.

3.4.4. Entorno

Para entrenar la política mediante el algoritmo 1, la interacción entre el agente y la red inalámbrica se formaliza a través de un entorno que respeta la interfaz de Gymnasium, tal como se describió en la sección 2.3.1. Esta abstracción oculta la complejidad del cálculo de interferencias y encapsula la dinámica del problema en los métodos estándar `reset()` y `step()`.

- **Inicialización (`reset`):** Al inicio de cada episodio, el entorno carga una secuencia temporal de matrices de canal correspondientes a un horizonte temporal. En este paso se calcula por única vez la matriz de características estructurales $\mathbf{X}$ mediante el conteo de caminos cerrados detallado en la ecuación (3.6). El método devuelve la observación inicial, compuesta por $\mathbf{X}$ y la primera matriz de canal $\mathbf{H}_0$.

Algoritmo 1 Entrenamiento de la GNN con REINFORCE y multiplicadores duales.

Entrada: Pesos θ_1 , multiplicadores $\mu_1 = \mathbf{0}$, tasas $\eta_\theta, \epsilon_\mu$, peso de entropía β , características estructurales $\mathbf{X}$

- 1: **Para** época $e = 1, \dots, E$:
- 2: **Para** cada *batch* $\mathcal{B}$:
- 3: **Para** cada instante $t \in \mathcal{B}$:
- 4: $\mathbf{Z}_t \leftarrow \text{GNN}(\mathbf{X}, \mathbf{H}_t; \theta_e)$ ▷ Cálculo de *logits*
- 5: $\pi_\theta(\cdot \mid \mathbf{X}, \mathbf{H}_t) \leftarrow \text{Softmax}(\mathbf{Z}_t)$
- 6: $a_{i,t} \sim \text{Categórica}(\pi_\theta(\cdot \mid \mathbf{X}, \mathbf{H}_t)) \quad \forall i$
- 7: $p_{i,t} \leftarrow \text{potencia asociada a } a_{i,t} \quad \forall i$
- 8: $R_t \leftarrow \text{ENTORNO}(\mathbf{H}_t, \mathbf{a}_t)$ ▷ Suma de tasas
- 9: $\xi_{i,t} \leftarrow p_{i,t} - P_{\text{máx},i} \quad \forall i$ ▷ Apartamiento de potencia
- 10: **Fin**
- 11: $\mu_i \leftarrow \max\left(0, \mu_i + \frac{\epsilon_\mu}{|\mathcal{B}|} \sum_{t \in \mathcal{B}} \xi_{i,t}\right) \quad \forall i$ ▷ Actualización dual
- 12: $J_t \leftarrow -R_t + \sum_{i=1}^M \mu_i \xi_{i,t} \quad \forall t \in \mathcal{B}$ ▷ Costo penalizado
- 13: $\tilde{J}_t \leftarrow J_t - \frac{1}{|\mathcal{B}|} \sum_{t' \in \mathcal{B}} J_{t'} \quad \forall t \in \mathcal{B}$ ▷ Costo centrado
- 14: $\mathcal{L}(\theta_e) \leftarrow \frac{1}{|\mathcal{B}|} \sum_{t \in \mathcal{B}} \sum_{i=1}^M \left[\tilde{J}_t \log \pi_\theta(a_{i,t} \mid \mathbf{X}, \mathbf{H}_t) - \beta \mathcal{H}(\pi_\theta(\cdot \mid \mathbf{X}, \mathbf{H}_t)) \right]$
- 15: $\theta_{e+1} \leftarrow \text{ADAM}(\theta_e, \nabla_{\theta} \mathcal{L}(\theta_e), \eta_\theta)$ ▷ Actualización de la GNN
- 16: **Fin**
- 17: **Fin**

- **Transición y Recompensa (step):** En cada instante t , el entorno recibe la acción conjunta $\mathbf{a}_t$, donde cada componente $a_{i,t}$ determina el canal $c_{i,t}$ o el canal y la potencia $p_{i,t}$ del enlace i . Con esta configuración, el entorno calcula la SINR _{i} para cada enlace utilizando la ecuación (2.1). A partir de la SINR, se computa la tasa espectral r_i según la ecuación (2.2). El método devuelve la suma de tasas R_{sum} definida en la ecuación (2.3).

De esta forma, el entorno concentra la evaluación física de cada acción: calcula la interferencia, la SINR y la tasa resultante. El bucle de entrenamiento, en cambio, solo utiliza esa recompensa para actualizar la política de la GNN y los multiplicadores lagrangianos.

3.5. Conjunto de datos

3.5.1. Aspectos generales

Los datos provienen de la infraestructura de red de Ceibal, descrita en la sección 1.4, pero aquí se utilizan para modelar los canales de una red de enlaces punto a punto. Las mediciones fueron recopiladas cada quince minutos a lo largo del

3.5. Conjunto de datos

año lectivo 2018, correspondientes a la banda de 2,4 GHz o 5 GHz. Las columnas relevantes para ese procesamiento son las siguientes:

- **Timestamp**: marca temporal de la medición.
- **MAC_AP**: identificador único del nodo medido.
- **Banda**: banda de operación (0 para 2,4 GHz, 1 para 5 GHz).
- **Tx_power**: potencia de transmisión en dBm.
- **RSSI_vecino**: potencia recibida desde un nodo vecino, en dBm.

3.5.2. Procesamiento

El primer paso del procesamiento consiste en ordenar esas mediciones para obtener, en cada edificio e instante, una descripción coherente de la red. La columna **MAC_AP** permite identificar los nodos que pertenecen a cada edificio, mientras que **Timestamp** y **Banda** separan las muestras por tiempo y frecuencia. Sobre esas muestras se estiman luego los coeficientes de atenuación que se usarán como atenuaciones de canal entre enlaces.

Para una medición entre un nodo transmisor a y un nodo receptor b , la atenuación se calcula como $h_{ab}[\text{dBm}] = \text{RSSI}_{ab}[\text{dBm}] - P_a[\text{dBm}]$, donde RSSI_{ab} es la potencia recibida en b desde a y P_a es la potencia de transmisión usada por a en esa medición. El valor resultante se convierte a escala lineal. Tras filtrar los datos con campos ausentes y excluir las mediciones fuera del período lectivo, el conjunto final comprende 1.424 edificios, con redes de hasta 33 nodos.

Siguiendo la formulación de Fernández et al. [3], las mediciones entre nodos se transforman en una red de enlaces punto a punto. Para construir la matriz de canal $\mathbf{H}_t$, es necesario definir para cada transmisor i su receptor asociado $r(i)$. El criterio utilizado consiste en asignar a cada transmisor el receptor con mejor condición de canal, medida a partir del promedio temporal de las atenuaciones observadas entre ambos nodos.

Una vez definida esta asociación, la topología de enlaces queda fija para todo el período analizado. Lo que cambia con el tiempo son las atenuaciones de canal, en cada instante t se toman los coeficientes de atenuación correspondientes y se arma la matriz $\mathbf{H}_t$ a nivel de enlaces. En esa matriz, $h_{ii}(t)$ representa la atenuación directa del enlace i , mientras que $h_{ij}(t)$, con $i \neq j$, representa el acoplamiento desde el transmisor del enlace i hacia el receptor del enlace j . Así, la conectividad del grafo se mantiene fija, pero sus pesos reflejan la variación temporal observada en las mediciones.

Para ilustrar la variabilidad temporal de los canales, la figura 3.3 compara mapas de calor de la matriz $\mathbf{H}$ del edificio 856 en marzo (izquierda) y en julio (derecha). Las diferencias entre ambos meses muestran que las atenuaciones de canal cambian en el tiempo, por lo que los pesos del grafo no pueden tratarse como constantes. Además, la falta de simetría entre entradas opuestas de cada matriz refleja que, en general, $h_{ij} \neq h_{ji}$.

Capítulo 3. Asignación de Recursos en Redes de Enlaces Punto a Punto

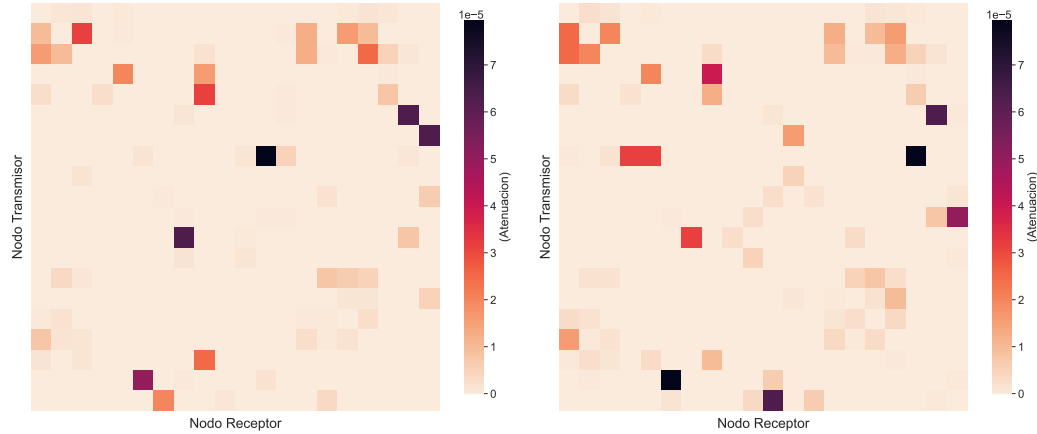


Figura 3.3: Mapas de calor de la matriz de canal $\mathbf{H}$ para el edificio 856 en dos snapshots del año lectivo 2018: marzo (izquierda) y julio (derecha), con escala de color compartida. El patrón de entradas intensas cambia sustancialmente entre meses pares. Un ejemplo se puede ver en la relación (8, 17), la cual domina en julio y no tiene contraparte visible en marzo.

3.5.3. Análisis

El edificio 856 se adopta como caso principal de estudio por dos razones complementarias. En primer lugar, su tamaño de 20 enlaces punto a punto permite construir patrones de interferencia no triviales. En segundo lugar, la densidad de su grafo en 2,4 GHz lo hace representativo de los escenarios más exigentes del conjunto de datos.

La figura 3.4 muestra la distribución de tamaños del conjunto de datos, la mediana se ubica en seis enlace punto a punto por edificio y el tercer cuartil en nueve, lo que confirma que la mayor parte de los edificios son de pequeña escala. Por eso, además de entrenar la política en el edificio 856, los experimentos evalúan su transferencia hacia edificios más pequeños del conjunto. Esta comparación permite medir si la arquitectura aprende reglas de asignación dependientes de la topología, y no solo del tamaño del escenario usado durante el entrenamiento. La banda de operación también incide en la dificultad del problema. La figura 3.5 compara la topología del edificio 856 en 2,4 GHz y 5 GHz para un instante particular de Marzo. Se observa una clara diferencia en el grado de los nodos de la banda baja respecto a los de la banda alta. Un mayor grado por nodo se traduce en una mayor cantidad de información interactuando simultáneamente, lo que incrementa la complejidad del problema. Por esta razón, se justifica concentrar el entrenamiento en la banda de 2,4 GHz, al representar el escenario más desafiante.

La figura 3.6 extiende esta comparación a los edificios analizados en la sección 3.8. Allí se observa que la mayor conectividad de la banda de 2,4 GHz no es un caso aislado del edificio 856. En la mayoría de los edificios esta banda induce más aristas en el grafo de interferencia y, por lo tanto, un mayor grado promedio.

3.6. Pruebas de funcionamiento

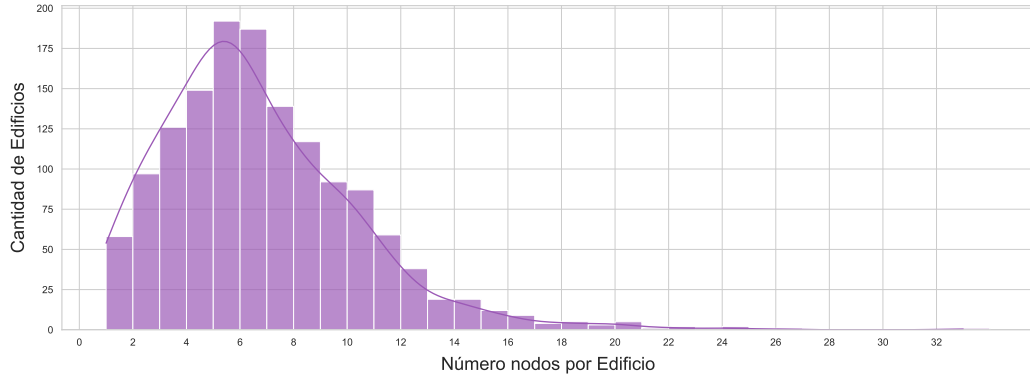


Figura 3.4: Distribución del número de enlaces punto a punto por edificio en el conjunto de datos de Ceibal (1424 edificios). La mediana se ubica en 6 enlaces punto a punto ($Q1=4$, $Q3=9$) y la moda en 5 enlaces; la mayoría de las redes son de pequeña escala. El edificio 856, utilizado como caso principal de entrenamiento con 20 enlaces punto a punto, se ubica en el percentil 95 de la distribución.

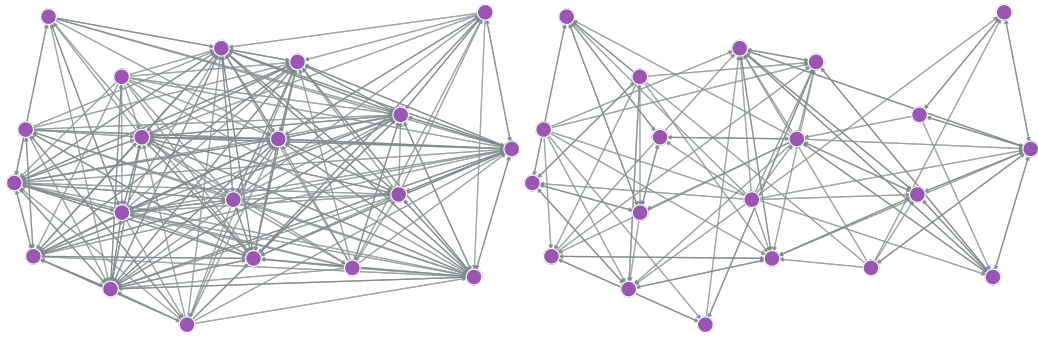


Figura 3.5: Topología del grafo de interferencia del edificio 856 en marzo, para las bandas de 2,4 GHz (izquierda, 241 enlaces) y 5 GHz (derecha, 101 enlaces). Cada círculo representa un nodo de la red; el **tamaño** es proporcional al grado y el **color** codifica la centralidad de intermediación (*betweenness centrality*), los nodos más claros son los que aparecen con mayor frecuencia en los caminos del grafo. Los tres nodos de mayor centralidad se destacan en rojo.

3.6. Pruebas de funcionamiento

Las pruebas de funcionamiento buscan verificar que la política aprendida converge a la solución óptima en escenarios donde esa solución se conoce de antemano. Para ello se diseñan dos topologías controladas, ilustradas en la figura 3.7, con tres enlaces primarios: $A-B$, $C-D$ y $E-F$.

El primer escenario, denominado interferencia central, ubica al enlace $C-D$ entre los otros dos. En esta topología, los enlaces de los extremos no se interfieren entre sí, sus atenuaciones cruzadas se consideran nulas. El segundo escenario, denominado interferencia total, corresponde al caso ya usado en la sección 3.4.1, los enlaces de los extremos también se interfieren entre sí. Ambas topologías sirven, además, como casos simples para presentar la estructura del problema antes de evaluar redes reales.

Capítulo 3. Asignación de Recursos en Redes de Enlaces Punto a Punto

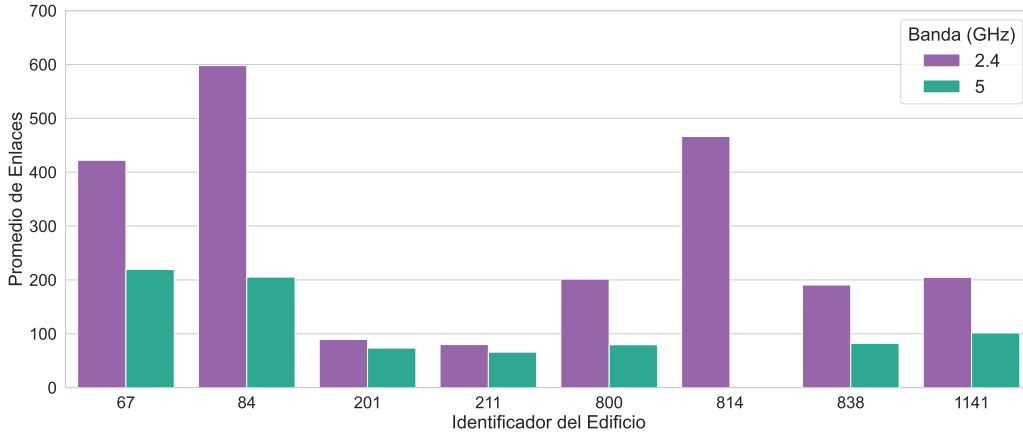


Figura 3.6: Promedio de aristas del grafo de interferencia por instante de medición en marzo para los ocho edificios utilizados en los experimentos de transferibilidad, por banda. En todos los casos la banda de 2,4 GHz (violeta) supera a la de 5 GHz (verde), con ratios que van desde $1,2\times$ en el edificio 201 hasta $2,9\times$ en el edificio 84. El edificio 856, caso de entrenamiento con 241 aristas promedio en 2,4 GHz, no aparece aquí porque estos edificios son exclusivamente los destinos de transferencia.

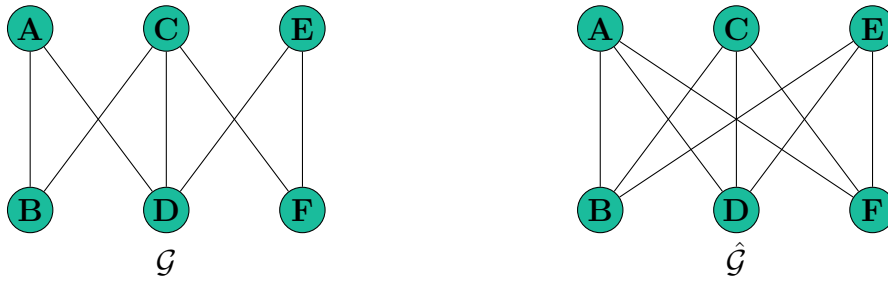


Figura 3.7: Ejemplo esquemático de los dos grafos evaluados en las pruebas de funcionamiento. El grafo $\mathcal{G}$ corresponde a la topología de interferencia central mientras que el grafo $\hat{\mathcal{G}}$ corresponde a la topología de interferencia total.

3.6.1. Verificación de la asignación de canales

Escenario de interferencia central

Para ilustrar la asignación óptima en este escenario, se considera un análisis numérico simplificado sobre el grafo $\mathcal{G}$. Se asume tres canales disponibles, atenuación directa normalizada $h_{ii} = 1,0$ para cada enlace y atenuación de interferencia $h_{ij} = 0,5$ hacia los vecinos adyacentes. Como los enlaces extremos están separados, su interferencia cruzada es nula ($h_{A-F} = h_{E-B} = 0$). Con potencia de transmisión $P_0 = 1,0$ y piso de ruido $\sigma^2 = 0,1$, se evalúa la capacidad resultante para distintas estrategias de asignación.

La tabla 3.2 contrasta la colisión total, en la que todos los enlaces operan en el mismo canal, con la solución óptima, en la que el enlace central se aísla en un canal ortogonal. Como evidencia la tabla, si todos los enlaces operan en el mismo canal, el enlace central ($C-D$) sufre el doble de interferencia que los extremos, degradando

3.6. Pruebas de funcionamiento

Tabla 3.2: Comparativa de estrategias de asignación para la topología de interferencia central.

Métrica	Colisión Total (Mismo Canal)	Solución Óptima (Aislamiento Central)
Canal Asignado		
Enlace $A-B$ (Extremo)	Canal 0	Canal 0
Enlace $C-D$ (Centro)	Canal 0	Canal 1
Enlace $E-F$ (Extremo)	Canal 0	Canal 0
Interferencia (I_i)		
Enlace $A-B$	0.5 (desde C)	0.0
Enlace $C-D$	1.0 (desde A y E)	0.0
Enlace $E-F$	0.5 (desde C)	0.0
Tasa Individual (bits/s/Hz)		
Enlace $A-B$	$\log_2(1 + 1,0/0,6) \approx 1,42$	$\log_2(1 + 10,0) \approx 3,46$
Enlace $C-D$	$\log_2(1 + 1,0/1,1) \approx 0,93$	$\log_2(1 + 10,0) \approx 3,46$
Enlace $E-F$	$\log_2(1 + 1,0/0,6) \approx 1,42$	$\log_2(1 + 10,0) \approx 3,46$
Tasa de Suma (R_{sum})	3.76	10.38

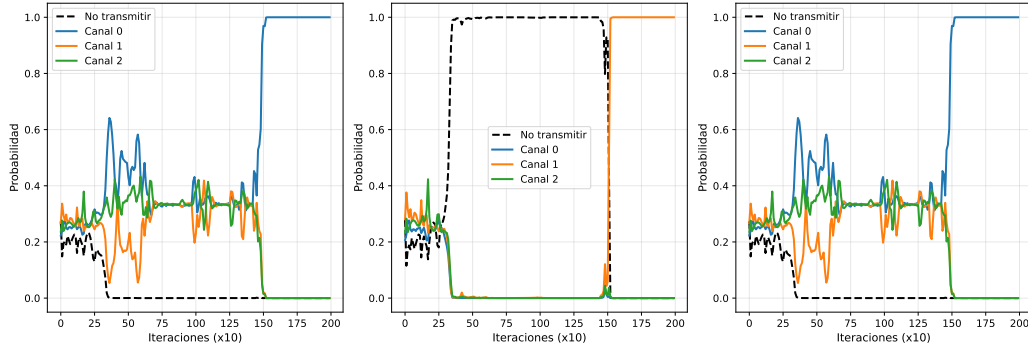


Figura 3.8: Evolución de la distribución de probabilidad sobre acciones en el escenario de interferencia central. Enlace 0 y enlace 2 convergen al canal 0; enlace 1 converge al canal 2. La probabilidad de la acción No TX se va a cero en los tres enlaces.

drásticamente su relación señal a ruido y penalizando el rendimiento global de la red ($R_{sum} \approx 3,76$). Al aislar el enlace central en un canal ortogonal, la interferencia se anula por completo para todos los nodos; los enlaces extremos pueden reutilizar el canal 0 de forma segura ya que no se ven entre sí. Esta configuración casi triplica la capacidad total del sistema ($R_{sum} \approx 10,38$).

La figura 3.8 muestra que la política converge a la solución óptima, el enlace $C-D$ concentra su probabilidad en el canal 1, luego de pasar transitoriamente por la acción No transmitir, mientras que los enlaces extremos comparten el canal 0. El orden de convergencia también es informativo, el enlace $C-D$ estabiliza su elección alrededor de la iteración 100 y los extremos lo siguen aproximadamente 50 iteraciones después, cuando la separación espectral del enlace central les permite identificar el canal libre sin ambigüedad.

Escenario de interferencia total

Para la topología $\hat{\mathcal{G}}$ se considera un segundo escenario, en el que todos los nodos pueden interferir entre sí en función de su distancia física. Para este análisis,

Capítulo 3. Asignación de Recursos en Redes de Enlaces Punto a Punto

Tabla 3.3: Comparativa de estrategias de asignación para la topología de interferencia total con 3 canales disponibles.

Métrica	Colisión Total	Reúso en Extremos	Aislamiento Total (Óptima)
Canal Asignado			
Enlace $A-B$	Canal 0	Canal 0	Canal 0
Enlace $C-D$ (Centro)	Canal 0	Canal 1	Canal 1
Enlace $E-F$	Canal 0	Canal 0	Canal 2
Interferencia (I_i)			
Enlace $A-B$	0.75 (desde C y E)	0.25 (desde E)	0.00
Enlace $C-D$	1.00 (desde A y E)	0.00	0.00
Enlace $E-F$	0.75 (desde A y C)	0.25 (desde A)	0.00
Tasa Individual (bits/s/Hz)			
Enlace $A-B$	$\log_2(1 + 1,0/0,85) \approx 1,12$	$\log_2(1 + 1,0/0,35) \approx 1,95$	$\log_2(1 + 10,0) \approx 3,46$
Enlace $C-D$	$\log_2(1 + 1,0/1,10) \approx 0,93$	$\log_2(1 + 10,0) \approx 3,46$	$\log_2(1 + 10,0) \approx 3,46$
Enlace $E-F$	$\log_2(1 + 1,0/0,85) \approx 1,12$	$\log_2(1 + 1,0/0,35) \approx 1,95$	$\log_2(1 + 10,0) \approx 3,46$
Tasa de Suma (R_{sum})	3.18	7.35	10.38

se asume nuevamente una atenuación directa de $h_{ii} = 1,0$ y una interferencia hacia los vecinos inmediatos de $h_{ij} = 0,5$. La diferencia fundamental radica en que se introduce una atenuación de interferencia no nula entre los extremos lejanos A y F (y simétricamente E y B). Dada la mayor distancia espacial, se asigna un valor representativo de $h_{A-F} = h_{E-B} = 0,25$.

A diferencia del escenario anterior, ahora se dispone de **tres canales ortogonales**. La tabla 3.3 compara tres estrategias: colisión total, con todos los enlaces en el mismo canal; reúso espacial en los extremos, donde los enlaces más alejados comparten espectro; y aislamiento total, donde cada enlace utiliza un recurso exclusivo. Se conservan los parámetros $P_0 = 1,0$ y $\sigma^2 = 0,1$.

Como evidencia el análisis numérico, la colisión total penaliza a toda la red severamente ($R_{sum} \approx 3,18$). Si la política intentara ser eficiente reutilizando el espectro en los nodos más alejados ($A-B$ y $E-F$), la capacidad mejoraría notablemente frente a la colisión ($R_{sum} \approx 7,35$), pero los extremos aún sufrirían una auto-interferencia residual de 0.25 producto de la topología de interferencia total.

Como la cantidad de recursos espectrales coincide con la cantidad de enlaces y cualquier reúso introduce interferencia, la estrategia que maximiza la capacidad global es ortogonalizar por completo las transmisiones. El aislamiento total eleva la tasa de suma al máximo de $R_{sum} \approx 10,38$. En esta topología, la GNN debe reconocer que ya no existen pares de enlaces libres de interferencia cruzada y asignar cada transmisión a un canal exclusivo para alcanzar el óptimo global. Como se discutió en la sección 3.4.1, en un grafo con simetría especular perfecta la GNN tendría dificultades para diferenciar los enlaces extremos. Para emular las condiciones de los datos reales de redes (donde el desvanecimiento y la topología irregular rompen las simetrías matemáticas), se introdujo una leve asimetría en las atenuaciones de interferencia cruzada de los extremos ($h_{A-F} \neq h_{E-B}$).

La figura 3.9 muestra que, aprovechando esta asimetría y utilizando las características estructurales k -hop propuestas, la política converge exitosamente a la solución óptima de aislamiento total. El enlace central ($C-D$) concentra su probabilidad en un canal, mientras que los extremos logran identificarse unívocamente y eligen los dos canales ortogonales restantes.

3.7. Políticas de referencia

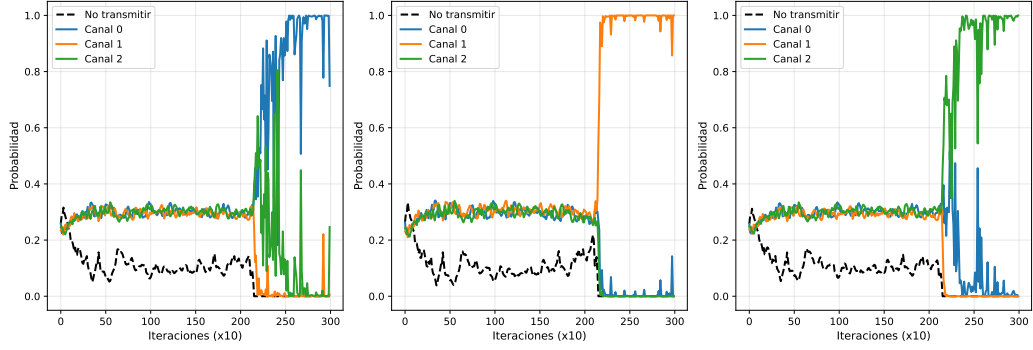


Figura 3.9: Evolución de la distribución de probabilidad sobre acciones en el escenario de interferencia total. La asimetría introducida permite que la GNN distinga los extremos mediante sus características k -hop, convergiendo a tres canales ortogonales distintos (Aislamiento Total) y anulando las colisiones.

3.7. Políticas de referencia

Se definen dos políticas de referencia para cuantificar la mejora que aporta la GNN frente a esquemas que no aprenden a partir de la estructura instantánea de la red. La primera representa el acceso no coordinado. En ella, cada enlace decide de forma independiente si transmite o no, con una probabilidad elegida para satisfacer en promedio la restricción de potencia (3.3).

La probabilidad de transmisión del enlace i se fija como

$$p_i = \frac{P_{\text{máx},i}}{p_0}, \quad (3.11)$$

donde $P_{\text{máx},i}$ es la potencia promedio permitida para el enlace i y p_0 es la potencia máxima de transmisión. Sea a_i la variable aleatoria que representa la acción elegida por el enlace i . La distribución resultante sobre el espacio de acciones es

$$\mathbb{P}(a_i = 0) = 1 - p_i, \quad \mathbb{P}(a_i = c) = \frac{p_i}{C}, \quad c \in \{1, \dots, C\} \quad (3.12)$$

Cuando transmite, el enlace lo hace en el canal seleccionado a potencia p_0 ; cuando no transmite, permanece inactivo. Para la variante de canal y potencia, la misma distribución se aplica sobre las $C \times P$ acciones activas, distribuyendo p_i de forma uniforme entre ellas.

Esta política modela un escenario de acceso no coordinado, en el que cada enlace actúa sin tener en cuenta la interferencia que genera sobre sus vecinos. Superar esta referencia de forma consistente y sobre edificios no vistos es la condición mínima para afirmar que la GNN aprendió representaciones estructuralmente informativas.

La segunda referencia corresponde a una política estática de canal fijo. Al inicio de la evaluación se sortea una configuración completa de canales para todos los enlaces y esa configuración se conserva para todo tiempo t . A diferencia de la referencia no coordinada, el canal asignado a cada enlace no se vuelve a muestrear en cada ranura. La única aleatoriedad que permanece es la decisión de transmitir

Capítulo 3. Asignación de Recursos en Redes de Enlaces Punto a Punto

o no transmitir, tomada con la misma probabilidad p_i para respetar la restricción promedio de potencia.

Si $\bar{c}_i$ denota el canal asignado inicialmente al enlace i , la política mantiene $c_i(t) = \bar{c}_i$ para todo t . Cuando el enlace está activo, transmite en ese canal fijo; cuando no lo está, permanece apagado. En la variante conjunta de canal y potencia, el canal también queda fijado desde el inicio y la política solo sortea el estado de actividad y el nivel de potencia entre las opciones disponibles. Esta referencia permite separar el efecto de contar con una configuración espectral inicial del efecto de adaptarse a las variaciones temporales de la matriz de canal observadas en los datos.

3.8. Resultados

Antes de analizar el desempeño de cada variante, es necesario definir la configuración del entorno y la arquitectura base. Para ambas variantes, el entrenamiento se llevó a cabo sobre los datos del edificio 856 (20 enlaces) en la banda de 2,4 GHz con 3 canales disponibles. Se fijó una potencia máxima de antena (P_{ant}) de 6 dBm y un piso de ruido (σ^2) de -90 dBm. Respecto a la arquitectura de la red neuronal y el algoritmo de aprendizaje, ambos modelos comparten una dimensión oculta (H) de 16, lotes de entrenamiento (batch size) de 32 y una tasa de aprendizaje base para la GNN (η_θ) de 9×10^{-4} . Por otro lado, las diferencias puntuales en los hiperparámetros de cada variante se detallan en la tabla 3.4.

Tabla 3.4: Hiperparámetros específicos para cada variante del modelo.

Parámetro	Elección de canal	Canal y potencia
Restricción de potencia prom. ($P_{\text{máx},i}$)	80 % de P_{ant}	60 % de P_{ant}
Tasa de aprendizaje dual (ϵ_μ)	3×10^{-5}	3×10^{-3}
Épocas de entrenamiento	300	700
Regularización por entropía (β)	0	0,3

3.8.1. Elección de canal

El escenario de entrenamiento fija una potencia máxima de 4 mW por enlace y una restricción de potencia promedio del 80 % (3,2 mW). El modelo se entrena sobre datos del edificio 856 correspondientes a marzo–agosto y se evalúa sobre períodos temporales posteriores y sobre otros edificios, combinando en un mismo experimento la transferibilidad temporal y espacial de la política aprendida.

La figura 3.10 muestra la evolución de la suma promedio de tasas durante el entrenamiento. Al comienzo, la política aprendida se mantiene cerca de la referencia e incluso puede quedar por debajo de ella, pero la supera luego de un período prolongado de exploración y ajuste de los multiplicadores duales. A partir de esa etapa, la ventaja se mantiene de forma estable. La figura 3.11 presenta los resultados

3.8. Resultados

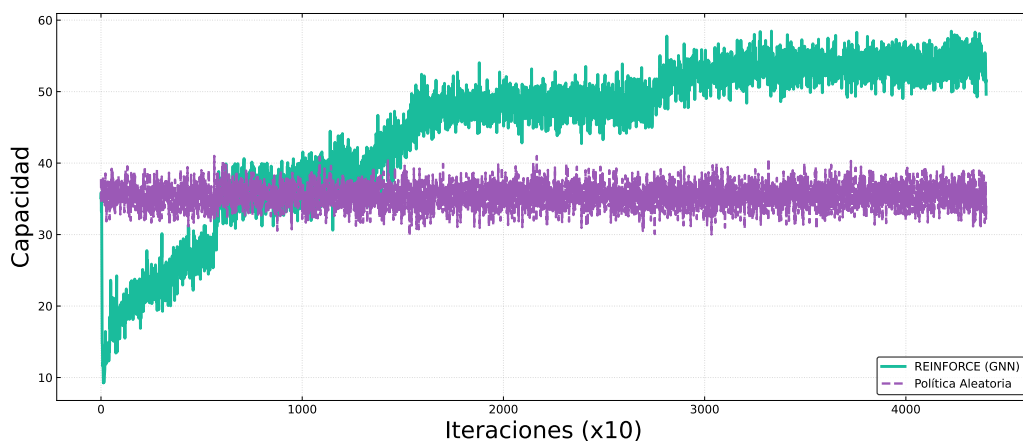


Figura 3.10: Evolución de la suma de tasas durante el entrenamiento sobre el edificio 856. La política aprendida con REINFORCE requiere un período inicial de exploración antes de superar a la política de referencia; luego mantiene esa ventaja de forma consistente.

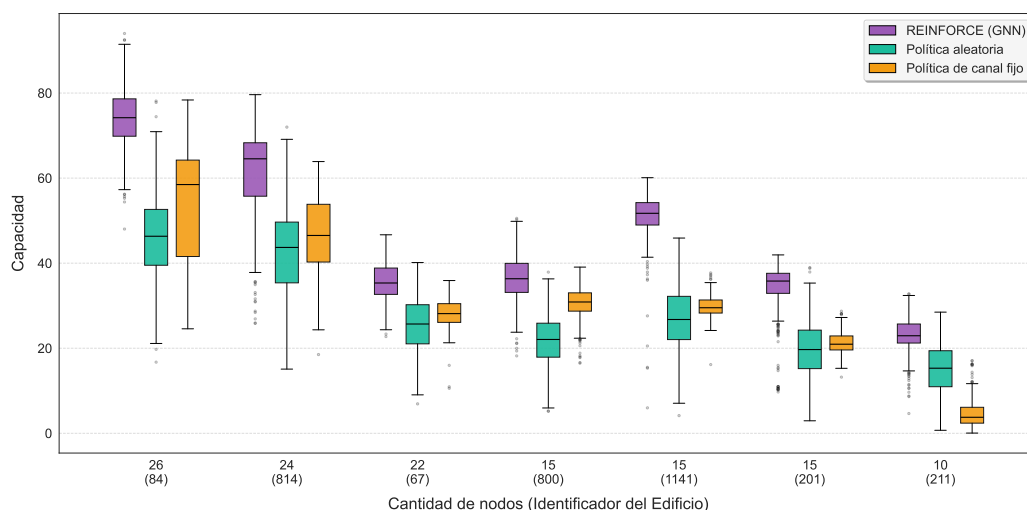


Figura 3.11: Distribución de la suma de tasas al transferir los pesos entrenados en el edificio 856 a siete edificios no vistos durante el entrenamiento. REINFORCE-GNN (violeta) se compara contra acceso aleatorio no coordinado (verde) y potencia fija en canal estático (naranja); esta última asigna a cada enlace un canal sorteado una sola vez y transmite siempre a $0,8 \cdot P_{\text{máx}}$ sin reconfigurar.

de transferibilidad espacial. Los pesos aprendidos sobre el edificio 856 se aplican sin reentrenamiento a los edificios 84 (26 nodos), 814 (24 nodos), 67 (22 nodos), 800, 838 y 1141 (15 nodos cada uno), y 201 y 211 (10 nodos cada uno). La ventaja sobre la política de referencia se mantiene en todos los casos, incluyendo los edificios más pequeños donde la topología difiere del escenario de entrenamiento.

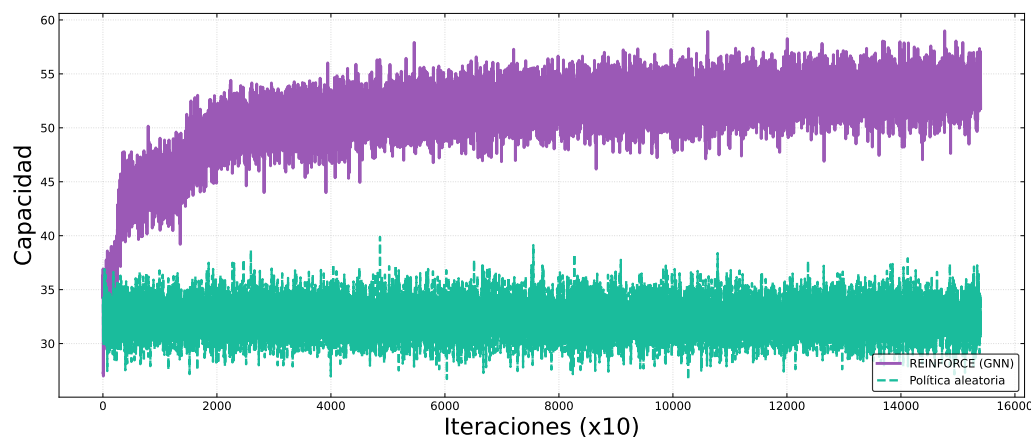


Figura 3.12: Evolución de la suma de tasas promedio durante el entrenamiento sobre el edificio 856 con asignación conjunta de canal y potencia, bajo una restricción del 60 % de la potencia máxima. REINFORCE (violeta) supera a la referencia aleatoria (verde) desde las primeras iteraciones y la brecha sigue creciendo hasta el final del entrenamiento.

3.8.2. Asignación de canal y potencia

La extensión a canal y potencia se evalúa bajo una restricción más exigente, fijada en el 60 % (2,4 mW) de la potencia máxima, frente al 80 % usado en la variante anterior. Con este presupuesto más ajustado, la política ya no puede apoyarse únicamente en transmisiones a potencia máxima, sino que debe aprender cuándo conviene reducir la potencia o no transmitir para respetar la restricción promedio.

La figura 3.12 muestra la evolución de la suma de tasas durante el entrenamiento sobre el edificio 856. A diferencia del experimento de elección de canal, donde la separación respecto a la referencia se establece temprano y se mantiene estable, aquí la brecha crece progresivamente a lo largo de las 8 000 iteraciones sin mostrar signos de saturación. Esto sugiere que el espacio de acciones ampliado, que combina canal y nivel de potencia, requiere más tiempo de exploración para que la política identifique los compromisos entre tasa propia e interferencia generada. La figura 3.13 confirma que la transferibilidad espacial se mantiene en esta variante. Los pesos entrenados sobre el edificio 856 superan a la referencia aleatoria en los ocho edificios evaluados, incluyendo el propio edificio de entrenamiento.

3.9. Síntesis

Los resultados de este capítulo muestran que ampliar el espacio de acciones desde la asignación binaria de potencia hacia la selección conjunta no destruye las propiedades que hacen útil al enfoque. La política aprendida sigue superando a la referencia de acceso no coordinado, y los pesos entrenados sobre el edificio 856 siguen siendo aplicables a otras topologías durante la evaluación del modelo.

Hay dos resultados que merecen destacarse de forma separada. Primero, que la transferibilidad espacial se mantiene incluso hacia edificios significativamente más

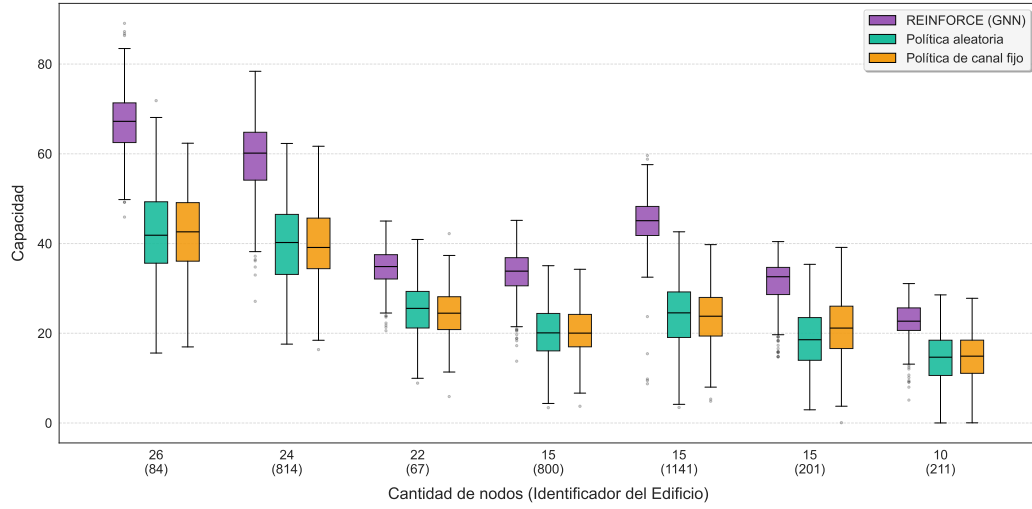


Figura 3.13: Al igual que en el caso anterior se ve la suma de tasas al transferir los pesos entrenados en el edificio 856 a siete edificios no vistos durante el entrenamiento. REINFORCE-GNN (violeta) se compara contra acceso aleatorio no coordinado (verde) y potencia fija en canal estático (naranja); esta última asigna a cada enlace un canal sorteado una sola vez y transmite siempre a $0,6 \cdot P_{\text{máx}}$ sin reconfigurar.

pequeños (10 nodos, contra los 20 del edificio de entrenamiento), donde la topología del grafo difiere, lo cual no es evidente ya que una política podría aprender reglas específicas del tamaño de la red en la que se entrenó. El hecho de que no ocurra es evidencia de que los *k-hop closed loops* de la sección 3.4.1 efectivamente capturan estructura topológica transferible, no patrones de una red particular.

Segundo, las pruebas de funcionamiento sobre topologías controladas muestran convergencia a la solución óptima conocida tanto para la asignación de canal en los escenarios de interferencia central y total. Este resultado da confianza en que la implementación es correcta con independencia de los datos reales, y permite separar posibles errores de implementación de las limitaciones propias del método.

Esta página ha sido intencionalmente dejada en blanco.

Capítulo 4

Asignación de Recursos con Clientes Dinámicos

El capítulo anterior consideró un escenario estático, con una topología de tamaño fijo y decisiones tomadas sobre un conjunto constante de nodos. En este capítulo se estudia un escenario dinámico, en el que la cantidad de clientes activos varía en el tiempo. El foco está puesto en redes Wi-Fi con topologías de tipo estrella, donde conviven puntos de acceso y clientes que ingresan, egresan y modifican su asociación según la dinámica del sistema y las decisiones de control adoptadas. Además, las acciones no se actualizan necesariamente en cada *slot*, sino cada cierto período de decisión T . Esta consideración responde tanto a la manera en que funcionan las controladoras de Ceibal, como al hecho de que los cambios de canal o potencia tienen un costo de señalización no despreciable en redes reales y realizarlos cada slot sería impracticable.

En el problema anterior, la restricción de potencia introducía una dependencia intertemporal, el estado actual dependía de la potencia consumida en instantes previos. En este nuevo problema no se considera una restricción acumulada de potencia. La memoria del sistema queda determinada por la permanencia de los clientes y por sus asociaciones con los APs, que pueden depender de decisiones tomadas anteriormente.

Para abordar el problema, primero se definen las variables que caracterizan el estado de la red en cada instante (sección 4.1); luego se especifica la dinámica que las rige (sección 4.2) y, finalmente, se formula el problema de decisión secuencial.

4.1. Definición de variables

Se considera un tiempo ranurado (*slotted*), indexado por $t \in \mathbb{N}$. Hay N APs indexados por $n \in \{1, \dots, N\}$ y un flujo de clientes que ingresan y egresan del sistema. El sistema modela un escenario de transferencia de bajada (downlink), por lo que el interés principal radica en la potencia de transmisión de los APs y en el impacto de la interferencia generada desde estos hacia los clientes.

El índice $u \in \mathbb{N}$ enumera los clientes en orden de llegada, siendo t_u el instante

Capítulo 4. Asignación de Recursos con Clientes Dinámicos

de arribo y Δ_u su duración de conexión. Durante su sesión, el cliente u experimenta una ganancia de canal $\mathbf{h}_{u,t} \in \mathbb{R}^N$ desde cada AP. La colección de atenuaciones de todos los U_t clientes activos en el instante t forma la matriz $\mathbf{H}_t \in \mathbb{R}^{U_t \times N}$.

La fila v de $\mathbf{H}_t$ corresponde al v -ésimo cliente entre los conectados en el instante t . El índice v no guarda relación directa con el índice de arribo u , que identifica al cliente a lo largo de toda su existencia en el sistema.

La matriz $\mathbf{H}_t$ admite una interpretación natural como grafo bipartito pesado entre clientes y APs, cuya matriz de adyacencia es

$$\begin{pmatrix} 0 & \mathbf{H}_t \\ \mathbf{H}_t^\top & 0 \end{pmatrix}.$$

Cada AP n tiene asignado en tiempo t un canal $c_{n,t} \in \{1, \dots, C\}$ y una potencia de transmisión $P_{n,t} \in \{P_{\min}, \dots, P_{\max}\}$ (conjunto discreto). Estas dos variables constituyen la acción de control del sistema y se codifican como una señal sobre el grafo completo:

$$\mathbf{X}_t \in \mathbb{R}^{(U_t+N) \times 2},$$

donde las filas correspondientes a los N APs contienen $(c_{n,t}, P_{n,t})$ y, en el instante inicial $t = 0$, se inicializan de forma aleatoria dentro del rango. Dado que no se ejerce ninguna acción de control sobre los usuarios, las filas asociadas a los U_t clientes se completan con la tupla constante $(-1, -1)$ para actuar como indicador de esta condición. Definir $\mathbf{X}_t$ sobre el grafo completo y no únicamente sobre los nodos AP garantiza que las operaciones de agregación de vecinos estén bien definidas independientemente de la cantidad de clientes activos $\mathbf{U}$.

La conexión efectiva entre clientes y APs en tiempo t se representa mediante el grafo bipartito binario $\mathbf{A}_t \in \{0, 1\}^{U_t \times N}$, con la restricción de que cada cliente está asociado a exactamente un AP (cada fila de $\mathbf{A}_t$ tiene exactamente un uno). La variable $\mathbf{A}_t$ no es una variable que controle el agente, sino que es una consecuencia de la dinámica del sistema y de las decisiones pasadas, según las reglas de *sticky client* descritas en la sección siguiente. Su matriz de adyacencia es

$$\begin{pmatrix} 0 & \mathbf{A}_t \\ \mathbf{A}_t^\top & 0 \end{pmatrix}.$$

4.2. Dinámica del sistema

La evolución conjunta de la tupla $(\mathbf{H}_t, \mathbf{X}_t, \mathbf{A}_t)$ no es, en general, markoviana. Para conocer el estado en $t + 1$ es necesario saber cuántos clientes hay en t , cuánto tiempo le resta a cada uno y cuándo llegará el próximo. Ninguna de esas dos cantidades está codificada en $(\mathbf{H}_t, \mathbf{X}_t, \mathbf{A}_t)$, de modo que la distribución del estado futuro no queda completamente determinada por el estado presente. Esta sección define qué variables adicionales son necesarias para recuperar la propiedad de Markov, bajo qué reglas de reasignación de AP y qué supuestos sobre la evolución del canal opera el sistema.

La asociación $a_{u,t}$, correspondiente a donde $\mathbf{A}_t$ tiene un uno en la fila del cliente u , determina cómo evoluciona $\mathbf{A}_t$ en el tiempo. La lógica de reasignación opera como

4.2. Dinámica del sistema

un mecanismo separado e independiente del agente RL, ya que el agente decide $\mathbf{X}_t$ pero no controla directamente $\mathbf{A}_t$.

Se consideran tres variantes, que difieren en cuándo un cliente reevalúa su AP asignado:

- **Full sticky:** el cliente elige el AP de mayor atenuación en el instante de conexión t_u y luego no cambia. Para $t > t_u$ se cumple $a_{u,t+1} \leftarrow a_{u,t}$ independientemente de cualquier cambio en $\mathbf{X}_t$ o $\mathbf{H}_t$. Este escenario modela redes donde la señalización de traspaso (*handover*) está deshabilitada o es costosa.
- **Sticky:** el cliente permanece en su AP asignado salvo que la señal recibida caiga por debajo del umbral de conexión θ_{conn} , momento en que vuelve a elegir el mejor AP disponible. Este modo es representativo de cómo suelen comportarse los clientes en muchas redes Wi-Fi reales, aunque cabe destacar que el criterio exacto de reasociación (el valor del umbral, la métrica de señal considerada, e incluso si se contempla algún margen de histéresis) depende del fabricante y de la configuración particular de cada controladora. Para los tiempos posteriores a su nacimiento, la decisión de cambio $a_{u,t}$ depende de $\mathbf{X}_t$, $\mathbf{H}_t$ y $a_{u,t-1}$.
- **Sticky lite:** el cliente se reasigna si algún otro AP ofrece mayor señal que el actualmente asignado. La lógica de decisión es análoga al caso anterior pero considera todas las atenuaciones $\mathbf{H}_t$ y no únicamente la del AP propio. Las dependencias funcionales resultan ser las mismas que en el modo *sticky*.

Los tres modos comparten la propiedad de que a_{u,t_u} depende de $\mathbf{X}_{t_u}$ y $\mathbf{H}_{t_u}$, y que para $t > t_u$ la dependencia pasa a ser además de $a_{u,t-1}$. Las fluctuaciones de canal entre $t-1$ y t se tratan como independientes. Esta hipótesis preserva la variabilidad temporal del canal sin introducir dependencias adicionales no observadas, y simplifica la formulación del MDP frente a un modelo de canal con correlación temporal explícita.

Sin embargo, hasta ahora el proceso no es estrictamente Markoviano debido a la dinámica de los clientes en el tiempo. Salvo que las llegadas sean un proceso de Poisson y las duraciones sigan una distribución exponencial, el tamaño y la evolución del grafo en el instante $t+1$ no pueden determinarse únicamente con la información de las matrices actuales $(\mathbf{H}_t, \mathbf{X}_t, \mathbf{A}_t)$. Se requeriría conocer el historial pasado del sistema para saber exactamente cuándo llegó cada cliente y cuánto tiempo falta para un nuevo arribo.

Esta necesidad es consistente con la formulación clásica de un proceso de decisión de Markov. Como se discute en la tesis de doctorado de Martín Randall Carlevaro [5], para que un problema pueda modelarse como un MDP es necesario que la evolución del sistema satisfaga la propiedad de Markov, es decir, que la distribución del siguiente estado dependa únicamente del estado y la acción actuales. Cuando el estado no contiene toda la información necesaria para describir la evolución temporal del sistema, resulta necesario incorporar variables que capturen dicha información. En particular, Randall motiva la inclusión de descriptores asociados al tiempo restante de conexión de los usuarios precisamente para recuperar

Capítulo 4. Asignación de Recursos con Clientes Dinámicos

una representación markoviana del sistema. Esta misma idea se aplica en el modelo propuesto en esta tesis.

Para recuperar la propiedad de Markov se introducen entonces dos variables. La primera, $\epsilon_{t,u} \in \mathbb{N}^{U_t}$, representa cuántas ranuras le restan a cada usuario activo u en el tiempo t . La segunda, $\delta_t \in \mathbb{N}$, indica las ranuras restantes hasta la llegada del próximo cliente. Esta variable se decrementa en cada slot y se sorteja nuevamente cuando llega a cero, momento en que también se determinan las atenuaciones y la duración del cliente entrante.

Proposición 1 (Proceso de Markov). *El proceso $(\mathbf{H}_t, \mathbf{X}_t, \mathbf{A}_t, \epsilon_t, \delta_t)$ es markoviano: la distribución condicional del estado en $t + 1$ dado el estado en t no depende de estados anteriores.*

Dado el estado completo $(\mathbf{H}_t, \mathbf{X}_t, \mathbf{A}_t, \epsilon_t, \delta_t)$, toda la información necesaria para generar la distribución del estado en $t + 1$ se encuentra contenida en el propio estado. En particular, las duraciones residuales ϵ_t determinan qué clientes permanecen activos y cuáles abandonan el sistema, mientras que δ_t determina si ocurre la llegada de un nuevo cliente y, en tal caso, cuándo sucede. Las nuevas duraciones y atenuaciones se sortean de manera independiente según el modelo de llegadas. Para los clientes que permanecen, el canal en $t + 1$ se obtiene a partir de su atenuación promedio y de una nueva fluctuación independiente. La asociación $\mathbf{A}_{t+1}$ se actualiza a partir de $\mathbf{A}_t$, $\mathbf{H}_{t+1}$ y $\mathbf{X}_t$ según las reglas de *sticky client*. Finalmente, $\mathbf{X}_{t+1}$ queda determinado por la acción aplicada en el siguiente instante de decisión. Por lo tanto, la distribución del estado futuro queda determinada por el estado actual y la acción correspondiente, sin requerir estados anteriores.

En un escenario real, estas últimas dos variables típicamente no se conocen en tiempo t . Si no se observan, el problema inducido para el agente es un proceso de decisión de Markov parcialmente observable (POMDP, del inglés *Partially Observable Markov Decision Process*). En la sección 4.8.6 se estudia el impacto de no observar ϵ_t y δ_t .

4.3. Formulación como problema de decisión

Establecida la estructura markoviana del sistema, se formula el problema de optimización que el agente debe resolver. El desempeño del sistema en el instante t se mide por la tasa total agregada de los clientes activos:

$$r_t = F(\mathbf{H}_t, \mathbf{X}_t, \mathbf{A}_t) = \sum_{v=1}^{U_t} r_v(\mathbf{H}_t, \mathbf{X}_t, \mathbf{A}_t),$$

donde $r_v(\cdot)$ es la tasa obtenida por el cliente v dado el estado actual, calculada a partir de la SINR según la ecuación (2.2).

Indexando las decisiones por τ de modo que $t = \tau T$, el objetivo de las decisiones es

$$\arg \max_{\{\mathbf{X}_{(0)}, \mathbf{X}_{(T)}, \dots, \mathbf{X}_{T(\Gamma-1)}\}} \sum_{\tau=0}^{\Gamma-1} \sum_{\ell=0}^{T-1} F(\mathbf{H}_{\tau T+\ell}, \mathbf{X}_{\tau T}, \mathbf{A}_{\tau T+\ell}).$$

4.4. Conjunto de datos y simulación

La recompensa acumulada en el τ -ésimo período de decisión es

$$R_\tau = \sum_{\ell=0}^{T-1} F(\mathbf{H}_{\tau T+\ell}, \mathbf{X}_{\tau T}, \mathbf{A}_{\tau T+\ell}).$$

La proposición 1 establece que el proceso $(\mathbf{H}_t, \mathbf{X}_t, \mathbf{A}_t, \epsilon_t, \delta_t)$ es markoviano en la escala fina de las ranuras t . Como las decisiones solo se toman cada T ranuras, conviene definir el estado del MDP en la escala de decisión τ como

$$s_\tau = (\mathbf{H}_{\tau T}, \mathbf{X}_{\tau T}, \mathbf{A}_{\tau T}, \epsilon_{\tau T}, \delta_{\tau T}).$$

El proceso $\{s_\tau\}$ hereda la propiedad de Markov de $\{s_t\}$: dado s_τ , la transición a $s_{\tau+1}$ queda determinada por la misma dinámica de la sección 4.2 aplicada recursivamente T veces, sin que se requiera información de instantes anteriores a τT . Submuestrear el proceso a intervalos de T ranuras no rompe, entonces, la estructura markoviana, solo cambia la escala temporal en la que se observa.

En un escenario real, las duraciones residuales $\epsilon_{\tau T}$ y las ranuras restantes hasta la próxima llegada $\delta_{\tau T}$ no se conocen de antemano, por lo que el agente no observa s_τ completo. Cuando estas variables no se observan, el problema inducido es un POMDP.

El objetivo de aprendizaje sigue la misma formulación general introducida en el capítulo 2, donde se restringe la búsqueda a una política parametrizada π_θ , igual que en el problema de la ARPP (capítulo 3), y el entrenamiento consiste en encontrar el vector θ que maximiza el retorno esperado, con el factor de descuento $\gamma \in [0, 1)$ ya introducido en la ecuación (2.5). La diferencia respecto a la ARPP es únicamente la escala temporal sobre la que se define el retorno, aquí se acumula sobre los períodos de decisión τ en lugar de a cada ranura t . La implementación concreta de π_θ mediante una GNN se describe en la sección 4.5.

4.4. Conjunto de datos y simulación

4.4.1. Descripción del conjunto de datos y procesamiento básico

Para este problema las mediciones se realizaron mediante sondeo periódico de la red inalámbrica por los WLC, con una frecuencia de tres capturas diarias en los horarios de 11:00, 15:00 y 19:00 horas. Cada captura registra, para cada cliente activo en ese momento, el RSSI medido desde el AP al que está asociado y desde los APs vecinos visibles. El formato de cada registro sigue el estándar SNMP:

```
SNMPv2-SMI::enterprises.14179.2.1.11.1.5.0.8.34.140.205.251
.136.29.252.119.72.160.0.1 = INTEGER: -94
```

De este registro se extraen: la dirección MAC del cliente, la dirección MAC del AP que reporta, la banda de operación, el índice de antena receptora y el valor de RSSI en dBm.

Los campos que lo componen son los siguientes, y se ilustran usando el ejemplo anterior:

Capítulo 4. Asignación de Recursos con Clientes Dinámicos

- **OID:** `enterprises.14179.2.1.11.1.5`
(numéricamente `1.3.6.1.4.1.14179.2.1.11.1.5`, correspondiente a *bsn-MobileStationRssiData*).
- **mac_cliente:** `0.8.34.140.205.251` (en hexadecimal: `00:08:22:8C:CD:FB`).
- **mac_ap:** `136.29.252.119.72.160` (en hexadecimal: `88:1D:FC:77:48:A0`).
- **banda:** 0 (*Slot ID* del AP; el *slot* 0 corresponde a la radio de 2,4 GHz / 802.11b/g/n).
- **antena:** 1 (índice de la antena física receptora).
- **RSSI:** -94 (indicador de intensidad de señal recibida, en dBm).

A partir de estos datos, se extrajeron los registros conservando únicamente las variables relevantes para el problema, como las direcciones MAC del cliente y del AP vecino, y el valor de RSSI.

Finalmente, dado que la tarea involucra la asignación de potencias, es necesario obtener los valores de atenuación del canal y no de los RSSIs, ya que es a partir de estos y de las decisiones del agente que se calcularía la potencia de recepción durante el entrenamiento. Como el conjunto de datos se registró desde el punto de vista de los APs, no se cuenta con el valor de potencia de transmisión de los clientes. Por este motivo, se adopta una potencia constante para todos ellos, fijada en el valor estándar de 14 dBm.

4.4.2. Simulación de la dinámica temporal y topológica

Con el fin de capturar la naturaleza estocástica de las redes inalámbricas reales y la evolución dinámica de su entorno, diseñamos un simulador específico. Este entorno se encarga de gestionar tanto el ciclo de vida de los usuarios (es decir, sus arribos y partidas) como la construcción de la topología de red subyacente.

Trayectorias y bloques de sensado

Antes de modelar la dinámica temporal de los usuarios, es necesario caracterizar la atenuación del canal que experimentan en el espacio. Como fue mencionado en la sección 4.4.1 los datos se separan en registros que se agrupan en bloques que caracterizan la atenuación espacial del canal de un cliente respecto a todos los APs visibles en las bandas de 2.4 GHz y 5 GHz. En la figura 4.1 se ilustra la distribución de potencia recibida (RSSI) presente en dichos registros, lo cual evidencia la diferencia en los niveles de atenuación según la frecuencia de operación.

Para simular la posición y el comportamiento del canal de un nuevo cliente u sin recurrir a modelos analíticos, el simulador utiliza estos bloques empíricos como distribuciones de probabilidad discretas. El proceso sigue la siguiente secuencia:

4.4. Conjunto de datos y simulación

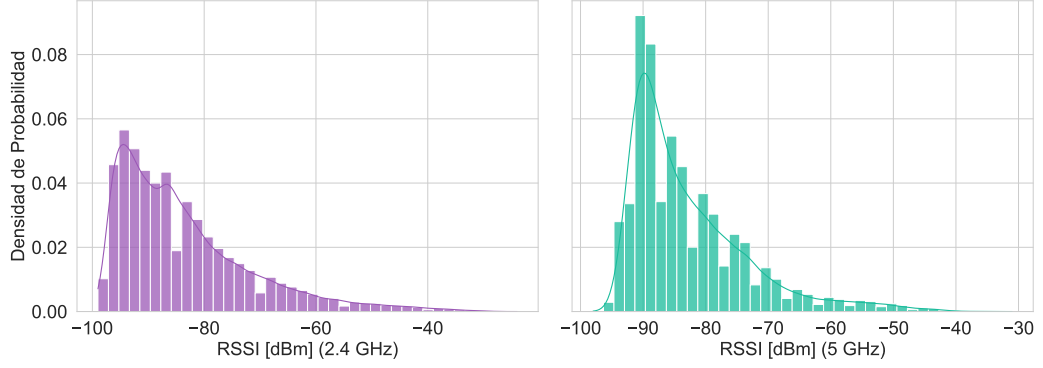


Figura 4.1: Distribución empírica de las intensidades de señal recibida (RSSI) presentes en el dataset de Cisco Ceibal para las bandas de 2.4 GHz y 5 GHz.

1. Para cada cliente u que ingresa al sistema, se selecciona aleatoriamente una trayectoria histórica de mediciones a partir de los datos de la red. Esta trayectoria agrupa todos los registros capturados para una misma dirección MAC real a lo largo del tiempo.
2. De dicha trayectoria, se extrae un bloque específico de medición de forma aleatoria. Este bloque contiene las atenuaciones nominales que ese cliente percibe desde cada AP n en las bandas de frecuencia disponibles.
3. Durante la permanencia del usuario (Δ_u), se asume una posición estacionaria promedio, asignando directamente estos valores nominales como la atenuación base $\mathbf{h}_u^{\text{base}} \in \mathbb{R}^{N \times 2}$.
4. Para simular pequeñas fluctuaciones físicas del canal, se superpone un componente estocástico gaussiano de media cero sobre la atenuación a lo largo de cada *slot* t :

$$\mathbf{h}_{u,t} = \mathbf{h}_u^{\text{base}} + \mathbf{z}_{u,t}, \quad \mathbf{z}_{u,t} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}) \quad (4.1)$$

donde σ representa la desviación estándar del ruido configurada en el entorno.

Este enfoque se adopta debido a la baja resolución temporal del conjunto de datos original, donde mediciones espaciadas cada 8 horas resultan insuficientes para simular la sesión continua de un cliente. Al utilizar una atenuación base empírica e inyectar ruido, es posible generar trayectorias con mayor granularidad temporal manteniendo la correlación espacial de la topología original.

Por otro lado, aunque la figura 4.2 muestra que la fluctuación del canal ya sigue una distribución normal, el parámetro σ utilizado en la simulación es menor a la desviación estándar real observada. El uso de la varianza original introduciría fluctuaciones demasiado abruptas en el RSSI, lo que dificulta el aprendizaje en diferentes modos sticky (*sticky modes*). Por lo tanto, se utiliza un nivel de ruido que proporciona aleatoriedad al entorno sin comprometer la convergencia del algoritmo.

Capítulo 4. Asignación de Recursos con Clientes Dinámicos

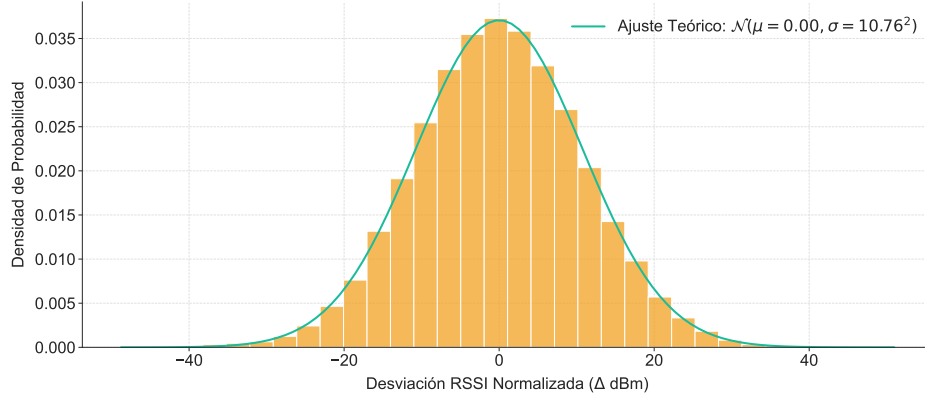


Figura 4.2: Distribución empírica de la fluctuación del canal (Δ RSSI) para la población general de clientes conectados al nodo de mayor tráfico de la red 211. Los datos fueron centrados respecto a la media individual de cada cliente. El ajuste a una función de densidad de probabilidad Normal (curva sólida) evidencia que el ruido de propagación en la red inalámbrica sigue un comportamiento Gaussiano.

Dinámica poblacional y modelo de tráfico

Como se detalló en la sección 4.2, la dinámica temporal del sistema queda determinada por las llegadas de nuevos clientes y la duración de sus sesiones (Δ_u), así como por el control temporal a través de las variables δ_t y $\epsilon_{t,u}$. A nivel de simulación, estos procesos temporales se generan siguiendo un modelo estocástico de tiempo discreto.

El valor original de las ranuras (*slots*) de espera entre arribos, denotado como δ , se muestrea a partir de una distribución normal, truncada inferiormente a la unidad y aproximada al entero más cercano:

$$\delta \sim \text{máx} (1, \lfloor \mathcal{N}(\mu_\delta, \sigma_\delta^2) \rfloor), \quad (4.2)$$

donde μ_δ representa el tiempo de espera promedio y la desviación estándar empírica se establece en $\sigma_\delta = 0,3\mu_\delta$. Esta variable δ inicializa el contador regresivo δ_t .

Una vez que δ_t alcanza el valor cero, se produce el instante de llegada t_u y se determina la cantidad de clientes que ingresan simultáneamente al sistema (N_{arr}) extrayéndola de una distribución exponencial truncada:

$$N_{\text{arr}} \sim \text{máx} (1, \lfloor \text{Exp}(1/\mu_{\text{arr}}) \rfloor), \quad (4.3)$$

siendo μ_{arr} la tasa media de llegadas. A cada nuevo cliente u se le asigna aleatoriamente un bloque de atenuaciones proveniente del conjunto de datos, junto con su tiempo de conexión al sistema (Δ_u). Este tiempo de permanencia también sigue un modelo exponencial discretizado:

$$\Delta_u \sim \text{máx} (1, \lfloor \text{Exp}(1/\mu_{\text{dur}}) \rfloor), \quad (4.4)$$

con μ_{dur} representando la duración media de la conexión. A partir de este valor se instancian directamente los contadores individuales $\epsilon_{t,u}$ para conformar el estado

4.4. Conjunto de datos y simulación

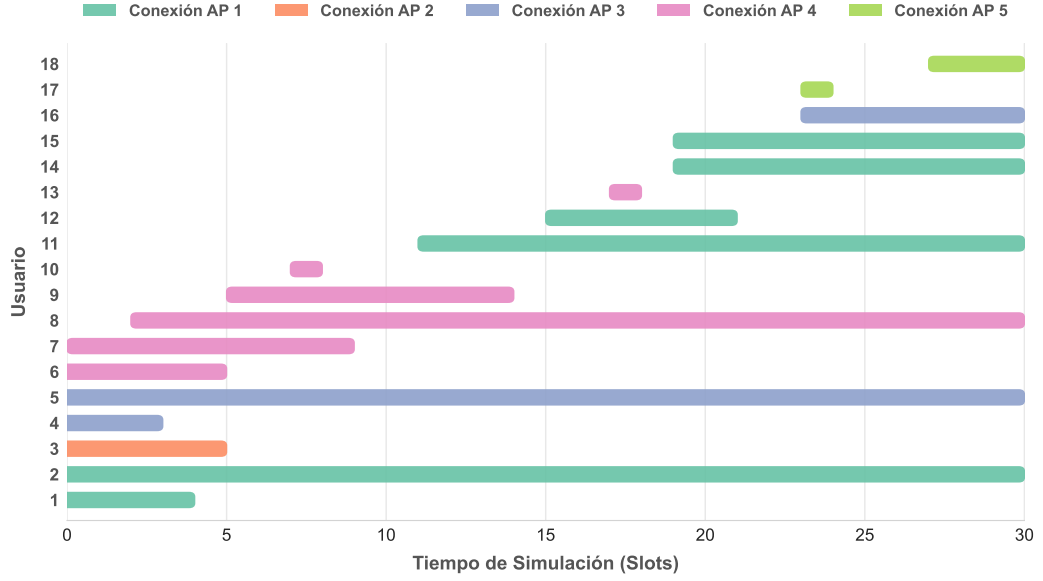


Figura 4.3: Dinámica temporal de los usuarios en el simulador. Se observa el ciclo de vida de los clientes en la red durante un episodio, donde cada barra representa el intervalo de permanencia de un usuario Δ_u . El color de cada bloque identifica el Punto de Acceso (AP) al cual el cliente fue asignado basándose en la señal empírica óptima.

markoviano del agente. La evolución de esta dinámica y la sincronización de las variables temporales a lo largo de un episodio de simulación se ejemplifica en la figura 4.3.

4.4.3. Construcción del grafo

La implementación de la matriz de conexiones $\mathbf{H}_t$ y el estado topológico general toman la forma de un grafo heterogéneo, es decir, una única estructura de grafo con distintas categorías de nodos, como lo son en nuestro caso los nodos tipo AP y nodos tipo cliente, que a su vez presentan distintos tipos de relaciones entre ellos.

Relaciones entre nodos

La interacción entre puntos de acceso y receptores se materializa mediante dos familias de aristas dirigidas. Para el enlace de bajada (*downlink*, de AP n a cliente u), la conectividad se delimita aplicando una máscara basada en un umbral de visibilidad θ_{vis} sobre la potencia de señal ($\text{RSSI}_{n,u}$). En los experimentos se fija $\theta_{\text{vis}} = -85$ dBm, de modo que solo se trazan aristas cuando la señal recibida supera ese umbral. Este valor se eligió a partir de la distribución empírica mostrada en la figura 4.1, ya que permite descartar una fracción importante de enlaces con RSSI muy bajo, asociados principalmente a señales marginales o ruido de medición. Esta decisión reduce la densidad del grafo y evita que APs prácticamente no visibles para un cliente contribuyan como vecinos relevantes en la GNN o como fuentes de interferencia efectiva. En consecuencia, la interferencia considerada se concentra en

Capítulo 4. Asignación de Recursos con Clientes Dinámicos

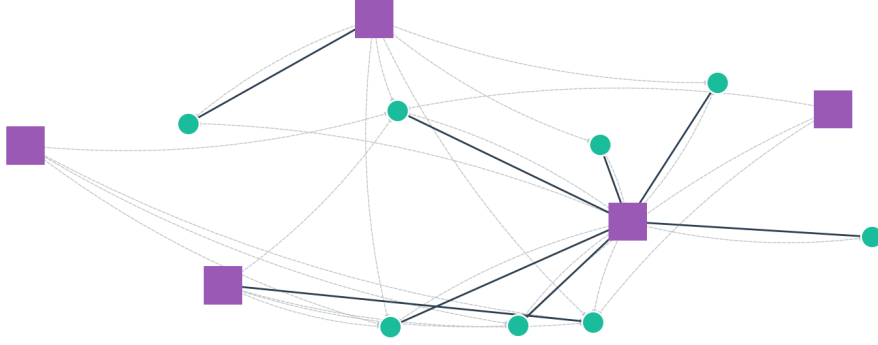


Figura 4.4: Estructura topológica del grafo heterogéneo $\mathcal{G}_t$ construido en un instante t . La representación abstrae el entorno de red diferenciando nodos de Puntos de Acceso (cuadrados) y nodos de clientes activos (círculos). Para mayor claridad visual, se ilustran únicamente las relaciones AP–cliente: las líneas negras sólidas indican las asociaciones activas, mientras que las líneas punteadas representan las atenuaciones medidas desde los APs hacia los clientes.

los APs con potencia recibida suficientemente alta como para afectar la decisión de asignación. Los atributos de dichas aristas se normalizan al rango escalar $[0, 1]$ de la siguiente manera:

$$e_{n \rightarrow u}^{\text{downlink}} = \text{clip} \left(\frac{\text{RSSI}_{n,u} [\text{dBm}] + 100}{70}, 0, 1 \right). \quad (4.5)$$

Donde la operación $\text{clip}(x, 0, 1)$ acota el valor normalizado al intervalo $[0, 1]$: devuelve 0 si $x < 0$, devuelve 1 si $x > 1$ y mantiene x sin cambios cuando $0 \leq x \leq 1$. La normalización lineal utilizada está diseñada para mapear el rango operativo útil del RSSI en redes Wi-Fi (que típicamente va desde -100 dBm, cercano al límite de sensibilidad del receptor, hasta -30 dBm, que representa una señal de máxima intensidad por proximidad) al intervalo $[0, 1]$. La diferencia de 70 dB define la pendiente de la escala, asegurando que señales extremadamente débiles se anulen en 0 y las señales que saturan el receptor se establezcan en 1.

En el sentido opuesto (enlace de asociación), los clientes forman una única conexión con el AP al que está asignado de acuerdo con la lógica (*sticky client*) determinada. Esta arista carece de atributos adicionales al actuar simplemente como indicador binario de conexión. Dadas estas reglas, la topología resultante en un instante dado del simulador se detalla gráficamente en la figura 4.4, donde se puede ver además como estas relaciones modelan un grafo bipartito.

Para facilitar el aprendizaje de la noción de posibles interferencias, se crea además una tercera relación entre nodos de tipo AP, completamente conectada, en la que la magnitud del conflicto entre dos APs, se almacena en la matriz de interferencia entre ellos $\mathbf{W}$, se define a partir de la interferencia que uno le impone a los clientes del otro. Si denotamos como $P_{n,u}^{\text{lineal}}$ a la señal convertida en dominio lineal, calculada como:

$$P_{n,u}^{\text{lineal}} = \text{máx} \left(10^{-10}, 10^{\text{RSSI}_{n,u}/10} \right). \quad (4.6)$$

4.4. Conjunto de datos y simulación

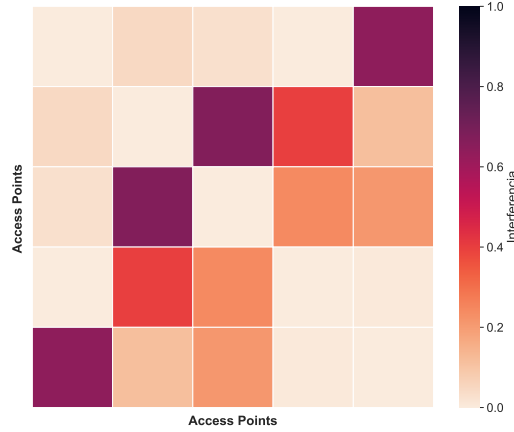


Figura 4.5: Mapa de calor representativo de la matriz de interferencia mutua normalizada $\mathbf{W}$ para las celdas de la red analizada. Los pesos cercanos a 1.0 reflejan un fuerte solapamiento percibido, determinando así la fuerza de los pesos de las aristas w_{ij} .

El peso de dicha asociación, w_{ij} , producida por las transmisiones del AP j sobre el grupo de clientes $\mathcal{C}_i$ vinculados al AP i , se calcula como el valor de señal promedio percibido:

$$w_{ij} = \begin{cases} \frac{1}{\max(1, |\mathcal{C}_i|)} \sum_{u \in \mathcal{C}_i} P_{j,u}^{\text{lineal}} & \text{si } j \neq i \\ 0 & \text{si } j = i \end{cases} \quad (4.7)$$

Dado que el objetivo es dar la noción de interferencia mutua, se simetriza la matriz $\mathbf{W}$ para los pesos entre celdas vecinas:

$$\mathbf{W} = \frac{\mathbf{W} + \mathbf{W}^\top}{2}. \quad (4.8)$$

Los valores de interferencia mutua $w_{i,j}$ son nuevamente convertidos a escala logarítmica (dBm) y normalizados para definir los pesos finales de las aristas entre APs. Esta normalización se realiza con el objetivo de trabajar en un rango acotado y evitar que magnitudes de distintos órdenes afecten la estabilidad numérica de los gradientes durante el entrenamiento. Un ejemplo del resultado de procesar esta interferencia mutua se expone a modo de mapa de calor en la figura 4.5.

Características de los nodos y del contexto global

El grafo contiene, además, otras características que capturan el estado dinámico y topológico del sistema en cierto instante t , y que luego serán útiles para alimentar la arquitectura GNN que utilizaremos para entrenar el modelo de predicción. Estas características varían según el tipo de nodo y el contexto global del grafo, y quedan definidas de la siguiente manera:

- **Nodos AP:**

Capítulo 4. Asignación de Recursos con Clientes Dinámicos

- **Decisión previa:** Dos identificadores que representan la decisión de canal y potencia discreta dentro de las disponibles.
 - **Positional Encoding:** Dota a la red neuronal de una noción geométrica de la ubicación de los AP en la red. La justificación teórica y el cálculo detallado de este embedding espectral basado en el Laplaciano se desarrollan en profundidad en la sección 4.5.1.
 - **Información de interferencia:** Un vector de estadísticas $\mathbf{F}$ sobre la matriz de interferencias $\mathbf{W}$, a modo de resumen de la misma. En concreto, el vector incluye la interferencia mutua normalizada máxima y la interferencia media que los clientes asociados a este AP sufren producto de las transmisiones de vecinos.
- **Nodos Cliente:**
- ϵ_t : Tiempo restante estimado hasta que el cliente finalice su sesión y abandone la red.
- **Variables Globales (Contexto del Grafo):**
- δ_t : Tiempo esperado hasta que ocurra el próximo evento de llegada de nuevos usuarios a la red.

4.5. Arquitectura del modelo GNN-RL

La aproximación de la política π_θ requiere una arquitectura capaz de operar sobre grafos bipartitos heterogéneos de dimensión variable. Para ello, al igual que en el problema ARPP, se diseña una arquitectura basada en GNN que recibe como entrada los grafos definidos en la sección anterior.

4.5.1. Codificación posicional

Las GNN basadas en *message passing*, presentan ciertas limitaciones en cuanto a su poder expresivo. En arquitecturas donde los nodos carecen de atributos distintivos fuertes, la red tiende a generar representaciones idénticas para ellos (colapso) y dificultando la tarea de predicción. Si bien una solución trivial a este problema de simetría es la asignación de identificadores únicos (IDs) escalares a cada nodo, ya sea con un índice o mediante *One-hot encoding*, este enfoque destruye la invarianza a las permutaciones del grafo e impide la generalización del modelo, ya que la dimensión del espacio de características dependería fuertemente de la cantidad de nodos observados durante el entrenamiento. Por consiguiente, se hace necesaria la búsqueda de soluciones que aporten una noción de posición o contexto espacial sin depender del índice asignado a cada nodo. Una posibilidad sería contar con información geométrica explícita de la red, pero el conjunto de datos no incluye dicha información espacial de APs y clientes. Por eso se optó por implementar un

4.5. Arquitectura del modelo GNN-RL

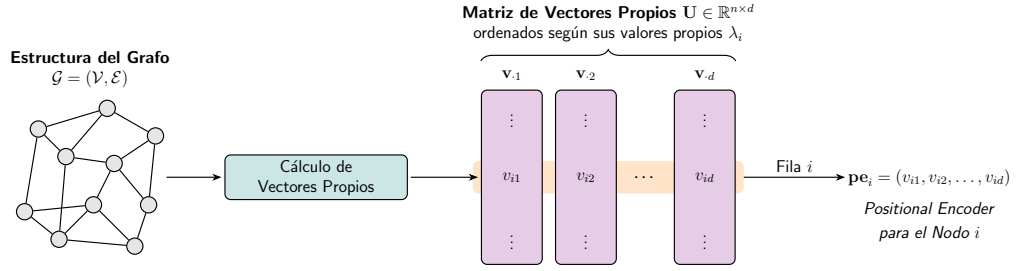


Figura 4.6: Diagrama de flujo del procedimiento utilizado para generar el *Positional Encoding* estructural a partir de un grafo de entrada. El proceso parte de una representación matricial del grafo, sobre la cual se realiza una descomposición espectral para obtener valores y vectores propios. Luego se seleccionan los vectores asociados a las componentes de menor frecuencia, se construye con ellos una matriz de coordenadas posicionales para los nodos.

mecanismo de codificación posicional (Positional Encoding o PE) basado en la estructura del grafo, específicamente mediante un embedding espectral, que provee una noción de posición relativa transferible entre grafos de distinto tamaño.

Este enfoque de codificación posicional está diseñado para extraer información sobre la posición relativa o global de un nodo dentro de la topología de la red. Esto permitiría que nodos localmente simétricos pero globalmente distantes posean representaciones iniciales distintas y posibilita la generalización del modelo a grafos con distintas cantidades de nodos. La extracción de las coordenadas espectrales puede realizarse a partir de cualquier representación del grafo, como lo pueden ser su matriz de adyacencia o su matriz del laplaciano, definida como $\mathbf{L} = \mathbf{D} - \mathbf{A}$, donde $\mathbf{D}$ es la matriz diagonal cuya diagonal contiene los grados de los nodos y $\mathbf{A}$ la matriz de adyacencia.

El proceso para generar los identificadores con este método, ilustrado en la figura 4.6 comienza con la descomposición espectral de la matriz seleccionada, resolviendo la ecuación de estado $\mathbf{M}\mathbf{v} = \lambda\mathbf{v}$ para obtener sus valores y vectores propios, para luego ordenar dichos vectores de forma creciente según sus valores propios asociados. A continuación, se define un hiper-parámetro d que determina la dimensión deseada para el vector identificador de cada nodo. Se seleccionan los d vectores propios correspondientes a las frecuencias más bajas y se colocan como columnas de una nueva matriz de encodings $\mathbf{U} \in \mathbb{R}^{N \times d}$. De este modo, el identificador posicional asignado a un nodo específico i corresponde exactamente a la i -ésima fila de la matriz $\mathbf{U}$. En caso de que la dimensión elegida d sea mayor que la cantidad total de vectores propios disponibles, las entradas faltantes del vector se rellenan mediante relleno con ceros (zero-padding) para asegurar la transferibilidad de la arquitectura. Una vez calculado este identificador posicional para cada nodo, el vector resultante de dimensión d se proyecta comúnmente mediante una capa lineal para ser posteriormente concatenado o sumado al resto de las características de los nodos, actuando como un embedding de entrada para las capas convolucionales de la GNN.

Para implementar este mecanismo, es necesario partir de un estado inicial del

Capítulo 4. Asignación de Recursos con Clientes Dinámicos

grafo con la suficiente información para poder lograr una buena representación. Sin embargo, en el problema planteado en esta parte del trabajo, al principio de la simulación la topología del grafo consiste únicamente de APs que ni siquiera cuentan aún con pesos en sus aristas. Para sortear este obstáculo, se propone aprovechar los datos de atenuaciones presentados en el problema ARPP. Si bien esta información se encuentra desvinculada temporalmente del estado de asociación de los clientes nuevos, resulta altamente representativa de la geometría de la red, aprovechando que los nodos de tipo AP no cambian su ubicación. La utilización de esta matriz de atenuación entre nodos fijos funciona en la práctica como el resultado de una fase de sensado corta y realizada únicamente durante la etapa inicial de configuración del sistema.

Finalmente, la integración de esta técnica introduce ciertos desafíos y limitaciones propias del cálculo espectral que debieron ser gestionados durante el diseño del modelo. El primero de ellos es la ambigüedad o inversión de signo, ya que si un vector $\mathbf{v}$ es vector propio de $\mathbf{M}$, entonces su opuesto $-\mathbf{v}$ también lo es, lo que significa que el solver numérico puede devolver cualquiera de las dos orientaciones de manera arbitraria según la ejecución. Esto requiere que la red neuronal posterior sea capaz de aprender representaciones invariantes a dichos cambios de signo. El segundo fenómeno corresponde a las rotaciones, las cuales ocurren cuando existen valores propios repetidos debido a simetrías perfectas en el grafo. No obstante, al modelar un entorno físico real mediante matrices de atenuaciones que incorporan ruido y atenuaciones, se entiende que la probabilidad de encontrarnos en uno de estos escenarios es improbable o al menos poco frecuente, por lo que los esfuerzos se concentran en lidiar con el problema del signo.

4.5.2. Solución a la invarianza del signo

Para abordar los problemas de la inversión de signo y las rotaciones, Lim et al. [18] presentan las arquitecturas SignNet y BasisNet como solución a estos problemas, respectivamente. Debido a las razones denotadas en la sección anterior, resulta de especial interés enfocarse en el caso de SignNet, donde el principio fundamental de esta arquitectura radica en su capacidad para procesar un conjunto de vectores de forma tal que las representaciones resultantes de los mismos sean independientes de la orientación individual que se haya recibido a su entrada.

Tal como se detalla en el diagrama de flujo de la figura 4.7 diseñado para esta implementación, la arquitectura toma como entrada el conjunto de vectores definidos en la matriz $\mathbf{U} = (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n)$. Para cada vector individual $\mathbf{v}_i$ perteneciente a este conjunto, el sistema realiza un cálculo en paralelo a través de dos ramificaciones: una para el vector directo, $\mathbf{v}_i$, y otro para su opuesto, $-\mathbf{v}_i$.

Ambas ramas pasan simultáneamente por una red de perceptrón multicapa (MLP) ϕ con capas lineales y activaciones ReLU que comparten pesos, para obtener las proyecciones $\phi(\mathbf{v}_i)$ y $\phi(-\mathbf{v}_i)$, para posteriormente, realizar suma algebraica, definida como $\Omega(\mathbf{v}_i) = \phi(\mathbf{v}_i) + \phi(-\mathbf{v}_i)$. Gracias a la propiedad conmutativa de la suma, cualquier inversión en el signo de un vector de entrada producirá exactamente el mismo resultado final en $\Omega(\mathbf{v}_i)$, logrando así la invarianza buscada sin sacrificar

4.5. Arquitectura del modelo GNN-RL

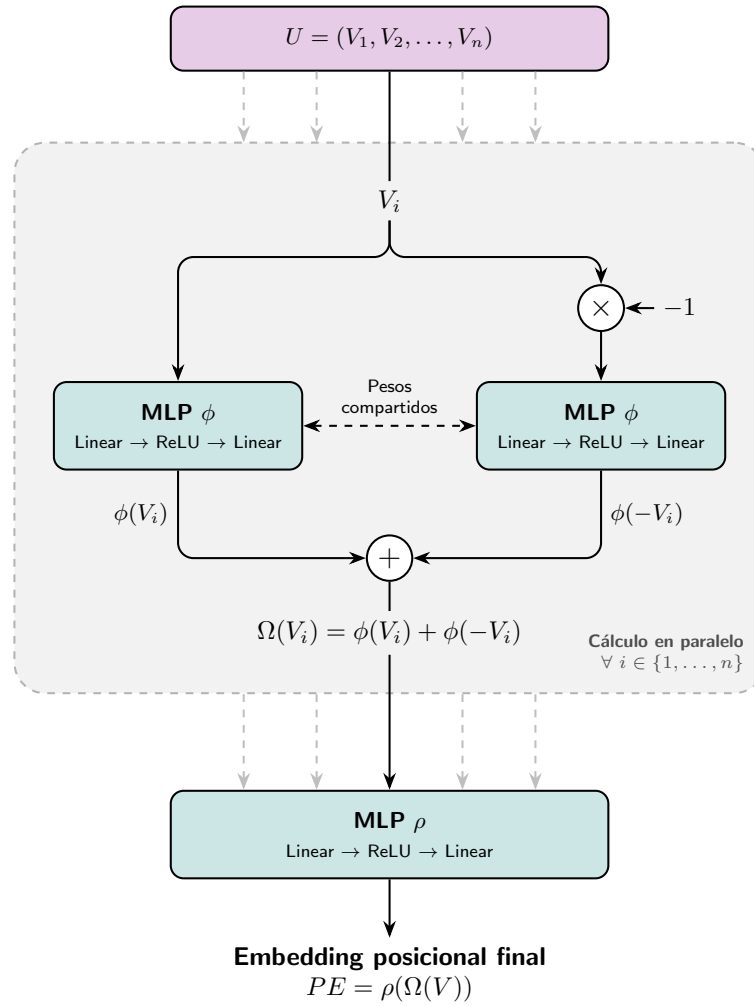


Figura 4.7: Diagrama de la arquitectura de la capa SignNet presentada por Lim et al. [18].

la información de dirección ni las propiedades geométricas originales de los vectores.

Tras realizar este cálculo simultáneo para la totalidad de los vectores del conjunto, las representaciones intermedias resultantes se unifican y se hacen pasar por una segunda arquitectura neuronal común denotada como ρ , formada como otro MLP similar a ϕ , pero con pesos independientes de este a modo de embedding. Su salida, expresada como $\mathbf{PE} = \rho(\Omega(\mathbf{v}))$, es la que se usa finalmente como identificadores de los nodos, ganando además una mayor expresividad.

4.5.3. Flujo global de la arquitectura GNN

El flujo general de procesamiento de la información dentro del modelo propuesto busca integrar las distintas señales provenientes tanto de los nodos AP como de

Capítulo 4. Asignación de Recursos con Clientes Dinámicos

los nodos cliente del grafo de manera adecuada e interconectada. Dicho flujo puede verse representado en la figura 4.8.

El procesamiento se inicia con la etapa de codificación de entrada, donde la arquitectura procesa y proyecta las distintas señales de manera paralela e independiente en un espacio latente con igual dimensión para cada una. Para la rama de los APs, se toma el vector de estadísticas de la matriz de interferencias $\mathbf{W}$ junto con el vector de asignaciones del instante anterior $\mathbf{X}_{t-1}$; ambos vectores se concatenan a lo largo de su dimensión de características y sirven como entrada de codificador lineal. Mientras tanto, en paralelo, los $\mathbf{PE}$ calculados se procesan en un codificador SignNet, cuyo resultado luego será sumado a la codificación anterior, obteniéndose así la representación inicial para los APs en este nuevo espacio:

$$y_{APs}^{(0)} = \text{Encoder}_{\text{lineal}}([\mathbf{W}, \mathbf{X}_{t-1}]) + \text{Encoder}_{\text{SignNet}}(\mathbf{PE}). \quad (4.9)$$

Finalmente, en la rama de los clientes, cada cliente conectado aporta su tiempo restante $\epsilon_{u,t}$ que también es proyectado utilizando nuevamente un codificador lineal:

$$y_{\text{clientes}}^{(0)} = \text{Encoder}_{\text{clientes}}(\epsilon_t). \quad (4.10)$$

Posteriormente, los estados iniciales $y_{APs}^{(0)}$ e $y_{\text{clientes}}^{(0)}$ alimentan el bloque principal de capas GNN. A final de L iteraciones donde cada nodo actualizó su representación mediante paso de mensajes, estos poseen una representación con el contexto de la red. Antes de proceder con la toma de decisiones, dado que la red debe aprender a tomar decisiones sobre los nodos de tipo AP, se descartan las salidas correspondientes a clientes, $y_{\text{clientes}}^{(L)}$. Luego, a la representación de los APs se la enriquece concatenando el contexto temporal δ_t , el cual representa el tiempo restante hasta la próxima llegada de un cliente al sistema:

$$y_{APs}^{\text{final}} = [y_{APs}^{(L)}, \delta_t]. \quad (4.11)$$

Al realizar la concatenación en este punto, las capas GNN se enfocan únicamente en resolver las relaciones espaciales y aquellas que pertenecen a cada nodo, dejando esta información temporal (δ_t) y global a todo el grafo actuar en una segunda etapa, evitando también inyectar la señal en cada nodo de la red.

Finalmente, en el caso de algoritmos de tipo Actor-Crítico, el vector resultante y_{APs}^{final} se bifurca según el rol. En la rama del actor, se utiliza una capa lineal para mapear de manera individual el embedding final de cada AP a la nueva acción de asignación $\mathbf{X}_t$. En la rama del crítico, para obtener una estimación del valor del estado global del sistema $V(s)$, se realiza una agregación global de la información de todos los APs mediante un operador *Global Mean Pooling*. El vector resultante del promedio de la red es proyectado a un único escalar mediante una capa lineal que representa el valor crítico.

4.5.4. Detalle de las capas GNN

El paso de mensajes dentro de una sola capa GNN $\ell \in \{1, \dots, L\}$, representado en la figura 4.9, se diseñó para respetar la forma en la que fluye la información en

4.5. Arquitectura del modelo GNN-RL

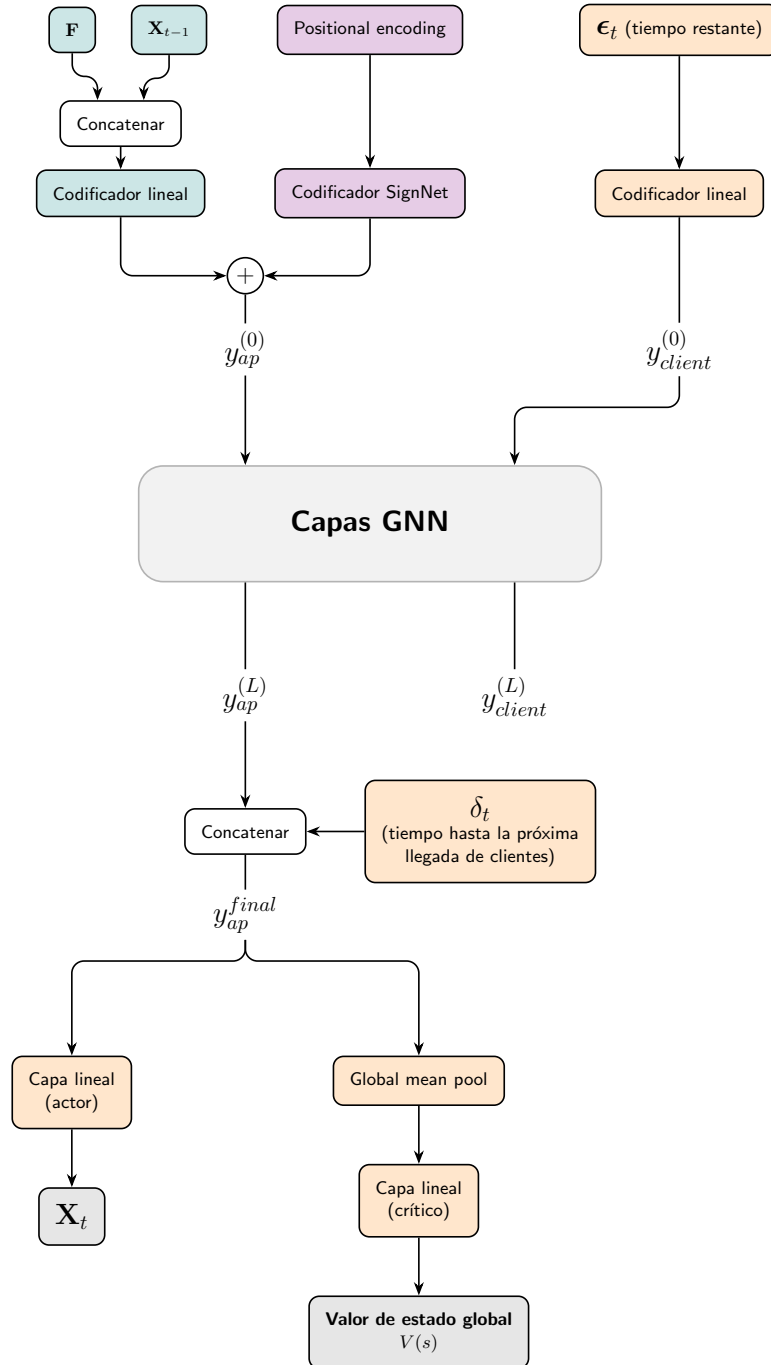


Figura 4.8: Diagrama de la arquitectura de la red GNN completa diseñada para el problema ARCD, particularmente para una estructura de tipo Actor-Critico.

Capítulo 4. Asignación de Recursos con Clientes Dinámicos

las distintas partes del grafo heterogéneo. Para actualizar los embeddings de los APs y los clientes en la capa ℓ , la información se propaga en paralelo segun sus tres relaciones diferenciadas.

En primer lugar, la relación de interferencia de AP a AP ($AP \rightarrow AP$) se procesa mediante la capa convolucional de atención GATConv, que toma como entrada los embeddings $y_{APs}^{(\ell-1)}$ de la capa previa. Dado que las aristas que representan esta relación están pesadas por un valor que representa interferencia, el uso de GATConv permite modular la señal haciendo uso de estos valores, logrando que cada AP logre una mejor representación del impacto de los vecinos más críticos.

En segundo lugar, la relación de asociación de clientes a APs ($Client \rightarrow AP$) se modela con una capa SAGEConv, que recibe las representaciones $y_{APs}^{(\ell-1)}$ e $y_{clientes}^{(\ell-1)}$. Debido a que esta asociación es de carácter binario y carece de pesos continuos en las aristas, el operador SAGEConv permite agregar la información de demanda de todos los clientes conectados de manera uniforme y eficiente.

En tercer lugar, la calidad del enlace físico entre los APs y los clientes ($AP \rightarrow Client$), al tener también pesos en sus aristas, esta vez por los RSSIs, se calcula empleando nuevamente una capa GATConv. Esto permite lograr una representación adecuada del impacto negativo de los APs vecinos basándose en las características del canal de cada uno.

Finalmente, el estado actualizado de un AP debe reflejar tanto el estado de interferencia de sus vecinos ($AP \rightarrow AP$) como la demanda acumulada de sus clientes conectados ($Cliente \rightarrow AP$). Por ello, el nuevo embedding se obtiene sumando aditivamente las salidas de ambas operaciones:

$$y_{APs}^{(\ell)} = \text{GATConv}_{AP \rightarrow AP} \left(y_{APs}^{(\ell-1)} \right) + \text{SAGEConv}_{clientes \rightarrow APs} \left(y_{APs}^{(\ell-1)}, y_{clientes}^{(\ell-1)} \right) \quad (4.12)$$

Por otro lado, la actualización de los clientes ($y_{clientes}^{(\ell)}$) depende directamente del flujo de calidad de canal proveniente de los APs vecinos, por lo que su salida es mapeada directamente desde el operador correspondiente:

$$y_{clientes}^{(\ell)} = \text{GATConv}_{APs \rightarrow Clientes} \left(y_{APs}^{(\ell-1)}, y_{clientes}^{(\ell-1)} \right) \quad (4.13)$$

4.6. Algoritmos de entrenamiento

Los tres algoritmos implementados maximizan la recompensa esperada ajustando θ en la dirección del gradiente estimado de $J(\theta) = \mathbb{E}_{\pi}[\sum_{\tau} \gamma^{\tau} R_{\tau}]$. Todos utilizan la misma GNN; la diferencia entre ellos radica en cómo reducen la varianza del estimador y en cuánto permiten modificar la política en cada actualización.

4.6.1. REINFORCE

El estimador de Policy Gradient es el mismo introducido en la ecuación (2.6) del capítulo 2 y aplicado en su forma de Monte Carlo (REINFORCE) en el capítulo 3

4.6. Algoritmos de entrenamiento

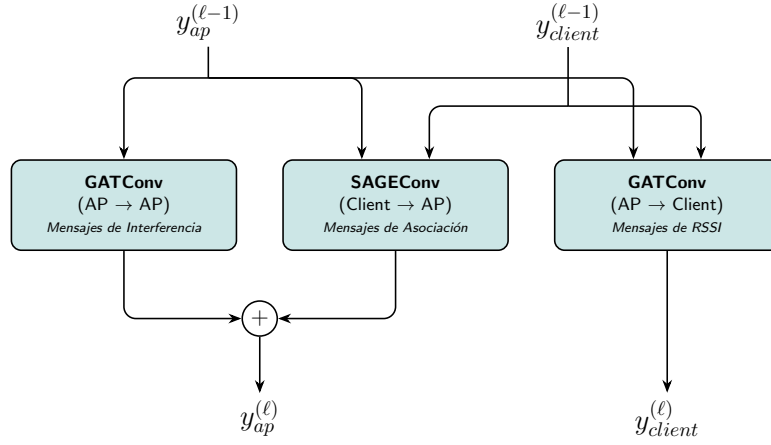


Figura 4.9: Diagrama de la arquitectura de una capa de la red GNN diseñada para el problema ARCD.

para la ARPP; por lo tanto, no se repite aquí su formulación. En este problema solo cambia la escala temporal: en la ARCD las decisiones y recompensas se agrupan por período de decisión τ en lugar de por slot t . Así, el retorno descontado se calcula sobre la secuencia agregada $\{R_\tau\}$, usando el estado s_τ definido en la sección 4.3 y la acción aplicada al inicio del período correspondiente, $X_{\tau T}$.

4.6.2. A2C

A2C (Advantage Actor-Critic) reemplaza la estimación puramente Monte Carlo de REINFORCE por un crítico aprendido V_ϕ , que aproxima el valor esperado del retorno desde el estado actual. En el caso de una actualización de un paso, la ventaja se estima como

$$\hat{A}_\tau = R_\tau + \gamma V_\phi(s_{\tau+1}) - V_\phi(s_\tau). \quad (4.14)$$

Si el período de decisión coincide con el final del episodio, se toma $V_\phi(s_{\tau+1}) = 0$. La diferencia respecto a REINFORCE es que el crítico aprende a predecir cuánto vale cada estado, de modo que $\hat{A}_\tau$ mide cuánto mejor o peor fue la acción tomada respecto a lo esperado en ese estado concreto, no respecto a la media del lote.

La función de pérdida total combina la pérdida de política, el error cuadrático del crítico y un término de regularización por entropía:

$$\mathcal{L} = \mathcal{L}_\pi + c_V \mathcal{L}_V - c_H \mathcal{H}[\pi_\theta], \quad (4.15)$$

donde $\mathcal{L}_\pi$ es la pérdida de Policy Gradient ponderada por $\hat{A}_\tau$ y $\mathcal{L}_V = (G_\tau - V_\phi(s_\tau))^2$ es el error cuadrático del crítico. El término $-c_H \mathcal{H}[\pi_\theta]$ cumple el mismo rol que la regularización por entropía introducida en la sección 3.4.3, mantener exploración durante el entrenamiento y evitar que la política se vuelva determinista demasiado pronto. En este capítulo se conserva el mismo criterio, pero con el peso c_H dentro

Capítulo 4. Asignación de Recursos con Clientes Dinámicos

de la pérdida actor-crítico. El término no es específico de A2C y se reutiliza sin cambios en PPO.

En este caso, el actor y el crítico comparten la GNN y se entrenan con AdamW, una variante de Adam que desacopla el decaimiento de pesos (*weight decay*) de la actualización adaptativa del gradiente [19]. En la práctica, esto permite incorporar una penalización directa sobre la magnitud de los parámetros durante el entrenamiento.

4.6.3. PPO

PPO (del inglés, Proximal Policy Optimization) [20] extiende el esquema actor-crítico limitando cuánto puede cambiar la política en cada actualización. Este control reduce el efecto de estimaciones de ventaja ruidosas y se implementa mediante una pérdida recortada:

$$\mathcal{L}^{\text{CLIP}}(\theta) = \mathbb{E} \left[\min \left(r_\tau \hat{A}_\tau, \text{clip}(r_\tau, 1-\varepsilon, 1+\varepsilon) \hat{A}_\tau \right) \right], \quad (4.16)$$

donde $r_\tau = \pi_\theta(X_\tau | s_\tau) / \pi_{\theta_{\text{ant}}}(X_\tau | s_\tau)$ compara la probabilidad de la acción X_τ bajo la política actual con la probabilidad asignada por la política anterior, es decir, la política usada para recolectar las muestras del lote. El parámetro ε define el margen permitido para ese cociente. La operación $\text{clip}(r_\tau, 1-\varepsilon, 1+\varepsilon)$, en el mismo sentido que la operación de recorte definida previamente, restringe r_τ al intervalo $[1-\varepsilon, 1+\varepsilon]$. De esta forma, si la nueva política intenta alejarse demasiado de la anterior, la contribución de ese término queda acotada y la actualización se vuelve más conservadora.

Para estimar la ventaja, PPO usa el estimador GAE (del inglés, Generalized Advantage Estimation) [21]:

$$\hat{A}_\tau^{\text{GAE}} = \sum_{l \geq 0} (\gamma \lambda)^l \delta_{\tau+l}^{TD}, \quad \delta_\tau = R_\tau + \gamma V_\phi(s_{\tau+1}) - V_\phi(s_\tau), \quad (4.17)$$

donde $\lambda \in [0, 1]$ controla el balance entre sesgo y varianza. Con $\lambda = 1$ el estimador recupera el retorno de Monte Carlo; con $\lambda = 0$ colapsa al error temporal de un paso.

A diferencia de REINFORCE y A2C, que consumen cada lote en una sola pasada, PPO reutiliza el mismo conjunto de muestras durante varias épocas de entrenamiento. Eso mejora la eficiencia en el uso de muestras a costa de requerir el seguimiento de la política antigua $\pi_{\theta_{\text{ant}}}$. La hipótesis de trabajo es que este control del tamaño de paso producirá políticas más estables que REINFORCE y A2C en la ARCD. En este entorno, la cantidad variable de clientes activos U_τ modifica la dimensión efectiva del grafo y aumenta la varianza de las estimaciones de ventaja, por lo que limitar el cociente de probabilidades reduce la magnitud de actualizaciones.

La pérdida total tiene la misma estructura que en A2C (ecuación (4.15)), reemplazando $\mathcal{L}_\pi$ por $-\mathcal{L}^{\text{CLIP}}$; el término de entropía $-c_H \mathcal{H}[\pi_\theta]$ se mantiene sin cambios. El algoritmo 2 resume el flujo de cálculo para una actualización completa de los parámetros.

Algoritmo 2 Actualización de la GNN con PPO.

Entrada: Pesos iniciales θ_1 , tasa η_θ , factor de descuento γ , parámetro GAE λ , pesos c_V, c_H , margen ε

- 1: **Para** actualización $u = 1, \dots, U$:
- 2: $\theta_{\text{ant}} \leftarrow \theta_u$ ▷ Fijar parámetros anteriores
- 3: $\mathcal{B} \leftarrow \emptyset$ ▷ Inicializar *batch*
- 4: **Para** cada episodio $e = 1, \dots, E$:
- 5: **Para** cada instante de decisión $\tau = 0, \dots, \Gamma - 1$:
- 6: $X_\tau \sim \pi_{\theta_{\text{ant}}}(\cdot \mid s_\tau)$ ▷ Muestreo de acción
- 7: $R_\tau, s_{\tau+1} \leftarrow \text{ENTORNO}(X_\tau)$ ▷ Transición
- 8: $V_\tau \leftarrow V_{\theta_{\text{ant}}}(s_\tau)$ ▷ Estimación de valor
- 9: $\mathcal{B} \leftarrow \mathcal{B} \cup \{(s_\tau, X_\tau, R_\tau, V_\tau, \pi_{\theta_{\text{ant}}}(X_\tau \mid s_\tau))\}$
- 10: **Fin**
- 11: Calcular ventaja GAE $\hat{A}_\tau^{\text{GAE}}$ para toda la trayectoria
- 12: **Fin**
- 13: $\hat{A}_\tau \leftarrow \frac{\hat{A}_\tau - \mu(\hat{A})}{\sigma(\hat{A})} \quad \forall \tau \in \mathcal{B}$ ▷ Normalización global
- 14: **Para** época $k = 1, \dots, K$:
- 15: $\mathcal{L}(\theta) \leftarrow 0$
- 16: **Para** cada transición $\tau \in \mathcal{B}$:
- 17: $r_\tau(\theta) \leftarrow \frac{\pi_\theta(X_\tau \mid s_\tau)}{\pi_{\theta_{\text{ant}}}(X_\tau \mid s_\tau)}$ ▷ Cociente de probabilidad
- 18: $\mathcal{L}_\tau^{\text{CLIP}} \leftarrow \min(r_\tau(\theta)\hat{A}_\tau, \text{clip}(r_\tau(\theta), 1 - \varepsilon, 1 + \varepsilon)\hat{A}_\tau)$
- 19: $\mathcal{L}_{V,\tau} \leftarrow (G_\tau - V_\theta(s_\tau))^2$ ▷ Error del crítico
- 20: $\mathcal{L}(\theta) \leftarrow \mathcal{L}(\theta) + [-\mathcal{L}_\tau^{\text{CLIP}} + c_V \mathcal{L}_{V,\tau} - c_H \mathcal{H}(\pi_\theta(\cdot \mid s_\tau))]$
- 21: **Fin**
- 22: $\mathcal{L}(\theta) \leftarrow \frac{1}{|\mathcal{B}|} \mathcal{L}(\theta)$ ▷ Pérdida media
- 23: $\theta \leftarrow \text{ADAMW}(\theta, \nabla_\theta \mathcal{L}(\theta), \eta_\theta)$ ▷ Actualización del modelo
- 24: (Actor-Crítico)
- 25: **Fin**
- 26: $\theta_{u+1} \leftarrow \theta$
- 27: **Fin**

4.7. Políticas de referencia

Siguiendo la línea de evaluación introducida en el capítulo 3, el desempeño de los algoritmos de aprendizaje por refuerzo se contrasta contra políticas de referencia que no utilizan aprendizaje para tomar sus decisiones.

Se retoma la **política de acceso aleatorio no coordinado**. A diferencia del problema anterior, donde la aleatoriedad estaba condicionada por una probabilidad de transmisión diseñada para satisfacer una restricción de potencia promedio, en este escenario no existe una limitante de consumo a largo plazo. Por lo tanto, en cada período de decisión τ , cada AP selecciona una configuración de forma unifor-

Capítulo 4. Asignación de Recursos con Clientes Dinámicos

me e independiente sobre todo su espacio de acciones disponible. Si el problema aborda únicamente la asignación espectral, el AP elige el canal $c \in \{1, \dots, C\}$ con probabilidad $1/C$, operando a máxima potencia. En la variante conjunta, se elige uniformemente una de las $|C| \times |P|$ combinaciones de canal y nivel de potencia. Esta política no toma en cuenta la interferencia ni la topología del momento, sirviendo como la cota inferior estricta; superarla de forma consistente es el requisito mínimo para afirmar que la GNN extrae representaciones útiles.

Por otro lado, para establecer una cota de desempeño superior mucho más desafiante, se introduce una política heurística. A diferencia de la aproximación aleatoria que toma decisiones en cada τ , esta heurística se implementa como una asignación estática calculada previamente. Para capturar lo estocástico del escenario sin depender de la topología instantánea, se ejecuta primero un episodio de prueba completo bajo una política aleatoria. Esto permite acumular empíricamente las interacciones espaciales y obtener una versión promediada temporalmente de la matriz de interferencia mutua $\mathbf{W}$.

Sobre esta matriz representativa promedio $\mathbf{W}$, el algoritmo opera en dos fases determinísticas para encontrar una configuración global estática, eludiendo así el potencial costo computacional que conllevaría probar todas las combinaciones.

La primera fase utiliza el grado ponderado $\mathbf{D}_i$ para determinar el orden de asignación. El grado D_i representa la suma de toda la interferencia que el AP i sufre o genera:

$$D_i = \sum_{j \neq i} \mathbf{W}_{i,j}. \quad (4.18)$$

Al ordenar los APs de mayor a menor $\mathbf{D}_i$, el algoritmo establece una prioridad directa; los APs más problemáticos tienen prioridad absoluta para elegir configuración. Por el contrario, los APs más aislados, cuyas decisiones no afectan gravemente al resto, quedan relegados al final del proceso.

La segunda fase utiliza la siguiente función de costo para decidir la acción específica. Una vez que es el turno del AP i (según su orden $\mathbf{D}_i$), el algoritmo observa a todos los APs que ya han elegido su configuración antes que él. A continuación, simula el impacto de cada alternativa disponible, calcula cuánta penalización se sumaría a la función de costo global si el AP se ubicara en el primer canal, cuánta si se ubicara en el segundo, y así sucesivamente. Finalmente, el AP adopta la configuración que agregue el menor valor posible a la función de costo global en ese momento exacto:

$$\mathcal{L}(\mathbf{c}, \mathbf{p}) = \sum_{i=1}^N \sum_{j=i+1}^N \mathbb{I}(c_i = c_j) \cdot \mathbf{W}_{i,j} \cdot (p_i + p_j + 1). \quad (4.19)$$

Donde el indicador $\mathbb{I}(c_i = c_j)$ vale uno cuando los APs i y j operan en el mismo canal, y cero en caso contrario. El término $\mathbf{W}_{i,j}$ es el peso de interferencia mutua entre ambos APs, calculado a partir de la matriz de interferencia promedio, mientras que p_i y p_j son las potencias de transmisión asignadas a los APs i y j . El factor $(p_i + p_j + 1)$ aumenta la penalización cuando alguno de los dos APs utiliza mayor potencia. La constante 1 evita que la penalización desaparezca por completo cuando ambos niveles de potencia toman su valor mínimo codificado. La heurística se

ejecuta una única vez antes de la evaluación. La configuración resultante de canales y potencias queda fija para todos los episodios, sin recalcularse ante las variaciones dinámicas del sistema. Si bien esta política sí hace uso de la información topológica promedio obtenida en el episodio de prueba para encontrar un diseño espectral y de potencia inicialmente robusto, carece por completo de adaptabilidad en tiempo real.

4.8. Resultados

Esta sección evalúa el desempeño de las políticas aprendidas durante el entrenamiento. Los experimentos comparan la capacidad obtenida frente a las políticas de referencia y analizan si la GNN logra adaptar las decisiones de canal y potencia a los cambios en la cantidad de clientes, sus asociaciones y la topología efectiva observada en cada período de decisión.

Antes de presentar los resultados obtenidos para cada variante del problema, se define la configuración del entorno, la arquitectura de la red y los hiperparámetros que rigen el entrenamiento del agente. Salvo indicación en contrario, las evaluaciones y entrenamientos de las siguientes secciones se realizan bajo el supuesto de observabilidad completa (MDP) y con un período de decisión unitario ($T = 1$), postergando el análisis del impacto del período de decisión para la sección 4.8.5 y el análisis del escenario parcialmente observable (POMDP) para la sección 4.8.6. En todos los experimentos de este capítulo, el entrenamiento principal se llevó a cabo utilizando exclusivamente los datos y la topología correspondientes a la red 990, que contiene 5 APs. Esta elección busca mantener acotado el costo computacional y la complejidad inicial del entrenamiento, en lugar de entrenar directamente sobre escenarios más grandes, como las redes de 15 o 20 APs. De esta forma, la evaluación posterior en topologías no vistas permite estudiar una forma exigente de transferencia, una política entrenada en un escenario pequeño debe operar en redes de mayor tamaño y con niveles de interferencia distintos. En ese contexto, resulta especialmente relevante comparar si el agente sigue superando a la política aleatoria incluso en edificios más grandes que el utilizado para entrenar. Se configuró un espacio de acciones compuesto por 3 canales disponibles y, para el problema conjunto, 3 niveles de potencia de transmisión discreta (5, 15 y 20 dBm).

Respecto a la arquitectura de la red neuronal y el proceso de aprendizaje, se definió una dimensión oculta (H) de 128 para los embeddings de los nodos, procesados a través de 2 capas de convolución y 3 capas densas. La elección de 2 capas de convolución está fundamentada en la física del problema: en el grafo bipartito AP–cliente, un camino de dos saltos permite conectar elementos del tipo AP–cliente–AP, lo cual es necesario para capturar la interferencia co-canal entre celdas vecinas a través de los clientes que comparten el espectro. Un número mayor de capas convolucionales no es requerido y podría inducir sobresuavizado (*over-smoothing*) en las representaciones locales. Por otro lado, la dimensión oculta de 128 se justifica por la mayor complejidad de la estructura heterogénea del grafo. Finalmente, las 3 capas densas proporcionan la expresividad no lineal necesaria para mapear estos embeddings a las probabilidades de acción del actor y al valor

Capítulo 4. Asignación de Recursos con Clientes Dinámicos

de estado del crítico. Para mantener un registro estructurado de la configuración, los principales hiperparámetros del entrenamiento se resumen en la tabla 4.1.

Tabla 4.1: Hiperparámetros de entrenamiento y configuración de la arquitectura.

Parámetro	Valor
Tasa de aprendizaje base (η)	4×10^{-4}
Episodios de entrenamiento	6000
Pasos de tiempo por episodio (t)	600
Dimensión del <i>Positional Encoding</i> (d)	4
Coefficiente de regularización por entropía (c_H)	0,005
Coefficiente de pérdida del crítico (c_V)	0,3

Antes de analizar el desempeño en evaluación de las políticas aprendidas, la figura 4.10 muestra la evolución de la recompensa acumulada por episodio durante la fase de entrenamiento en la red 990 para el problema conjunto de asignación de canal y potencia. Se incluye la comparación de los tres algoritmos evaluados: REINFORCE, A2C y PPO. Todos los entrenamientos se ejecutaron en una computadora portátil MacBook Pro con procesador Apple M1. En cuanto a los tiempos de cómputo requeridos, REINFORCE insumió aproximadamente 3 h 16 min, A2C tomó 3 h 20 min y PPO demandó 4 h 02 min. Todos los métodos exhiben un nivel de varianza similar durante el proceso de aprendizaje; sin embargo, se destaca que PPO logra desmarcarse rápidamente de A2C y REINFORCE en torno a los 100 episodios, alcanzando y sosteniendo valores de recompensa más altos durante el resto del entrenamiento.

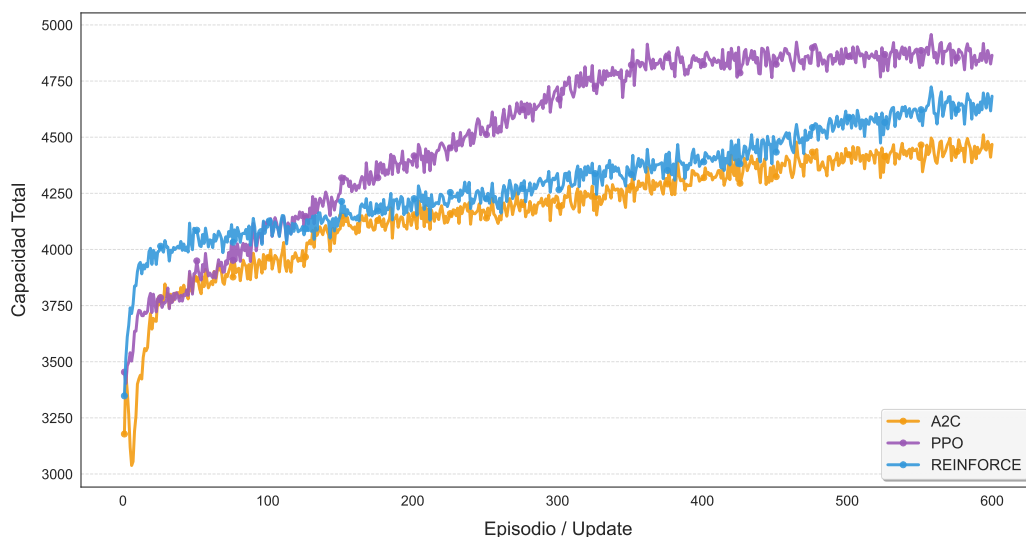


Figura 4.10: Evolución de la recompensa acumulada por episodio durante el entrenamiento de REINFORCE, A2C y PPO en la red 990 para el problema conjunto de asignación de canal y potencia.

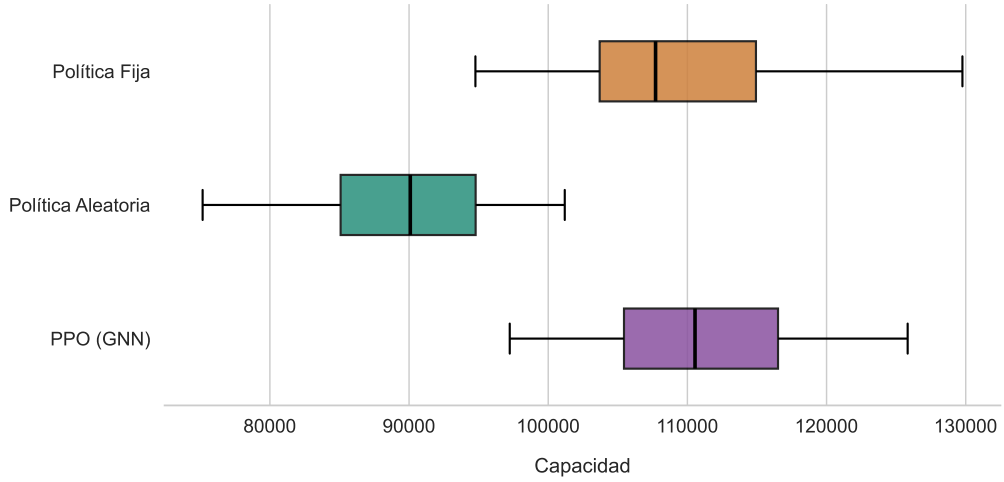


Figura 4.11: Distribución de la capacidad total obtenida en la asignación de canales para la red 990, utilizada durante el entrenamiento del agente. Se comparan la política fija heurística, la política aleatoria y la política aprendida mediante PPO con GNN.

Cabe destacar que las curvas de entrenamiento para las restantes configuraciones (variaciones en las reglas de asociación *sticky*, extensión del período de decisión T y escenario POMDP) mostraron una dinámica de convergencia cualitativamente análoga a la presentada en la figura 4.10. Por este motivo, y a fin de mantener una presentación concisa, las secciones posteriores enfocan su análisis directamente en las distribuciones de capacidad alcanzadas durante la etapa de evaluación.

4.8.1. Asignación de canales con PPO

En la red 990, utilizada durante el entrenamiento, la figura 4.11 muestra que PPO (GNN) supera a la política aleatoria y alcanza una mediana apenas superior a la política fija heurística en evaluación. La diferencia respecto a la heurística es reducida, por lo que no puede interpretarse como una mejora marcada. Más bien, el resultado indica que la política aprendida logra un desempeño comparable al de una regla estática basada en la topología promedio. Esto sugiere que, cuando la acción se limita a la selección de canal, el margen disponible para mejorar mediante decisiones dinámicas es acotado.

La figura 4.12 extiende la evaluación al resto de los escenarios para medir la transferibilidad del modelo. La política basada en PPO (GNN) supera de forma consistente a la política aleatoria en la mediana de capacidad para todas las redes. Esto demuestra que el algoritmo está logrando asimilar las características del entorno, en lugar de tomar decisiones al azar. El desempeño del modelo varía marcadamente según el escenario y la cantidad de nodos. En las topologías más grandes (redes 856 y 838, con 20 y 15 nodos respectivamente), la política fija logra una mayor capacidad y mantiene el liderazgo.

En las redes 211, 405 y 990, el enfoque PPO (GNN) logra superar en mediana a

Capítulo 4. Asignación de Recursos con Clientes Dinámicos

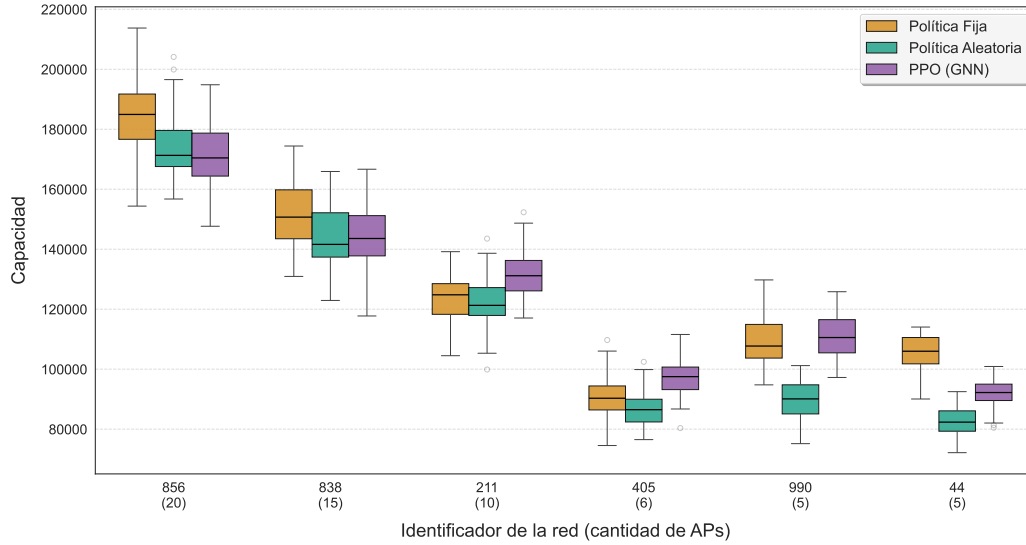


Figura 4.12: Distribución de la capacidad total obtenida al resolver la asignación de canal en múltiples redes. Se comparan la política fija, la política aleatoria y la política aprendida mediante PPO con GNN, entrenada en la red 990 y evaluada en el resto de las topologías. La comparación permite analizar la capacidad de transferencia del agente cuando cambian el tamaño y la estructura de la red.

la política fija. En estas configuraciones particulares, la arquitectura de propagación de mensajes sobre el grafo logra encontrar asignaciones conjuntas que optimizan la capacidad de manera más eficiente que una regla estática, demostrando el valor de este enfoque para ciertas topologías.

El rango entre cuartiles de PPO tiende a ser más compacto en redes con menos nodos, como el 405 y el 44. Esto indica que el modelo logra converger hacia políticas más estables cuando el espacio de acciones se reduce, aunque en el caso particular de la red 44 la política fija vuelve a tener una mejor mediana.

Sin embargo, el hecho de que la política fija mantenga el liderazgo en los escenarios más densos (como el 856 y el 838) evidencia un límite superior en la capacidad de optimización del modelo bajo su configuración actual. Cuando la red crece y la interferencia se vuelve más compleja, la estrategia de asignación actual no es suficiente para aislar las transmisiones de manera óptima. Para evaluar esta limitación y otorgarle al agente la flexibilidad necesaria para escalar a cualquier topología, es de interés incorporar la asignación de potencia discreta.

4.8.2. Asignación de canal y potencia con PPO

En esta sección se evalúa el desempeño de la política entrenada con PPO sobre la arquitectura GNN descrita en la sección 4.5 para el problema conjunto de asignación de canal y potencia. La evaluación se realiza frente a las dos políticas de referencia definidas en la sección 4.7, que corresponden a la política fija y la política de acceso aleatorio no coordinado.

La figura 4.13 presenta la distribución de la capacidad total alcanzada por cada

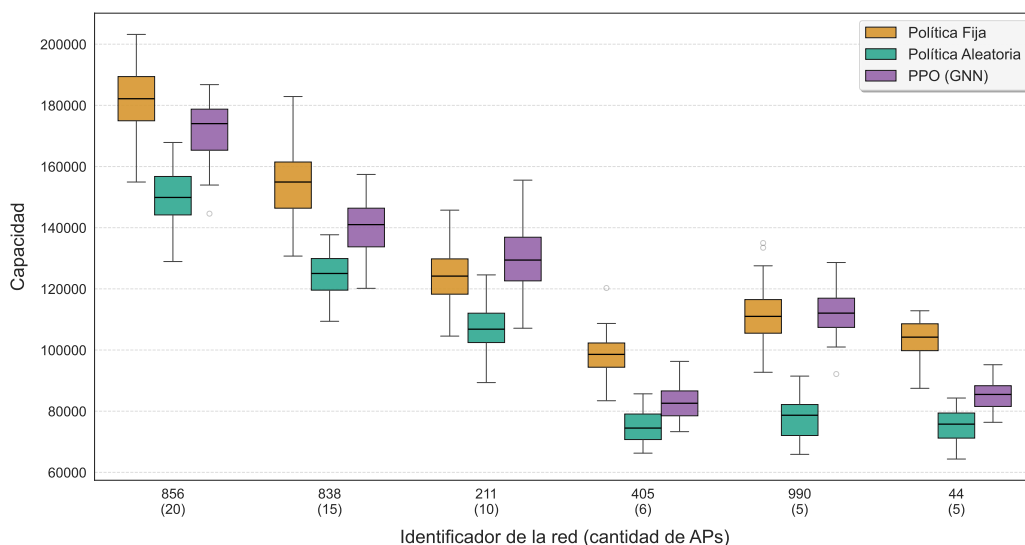


Figura 4.13: Capacidad total de la red alcanzada en el problema conjunto de asignación de canal y potencia por la política heurística estricta (política fija), la política de acceso aleatorio no coordinado (Política Aleatoria) y la política PPO (GNN), en seis redes representativas del conjunto de datos.

alternativa en las seis redes evaluadas. En todos los escenarios, la política PPO obtiene medianas superiores a las de la política aleatoria. Este resultado indica que la política aprendida incorpora información útil del estado de la red para la asignación conjunta de canal y potencia, aunque no implica por sí solo optimalidad ni dominancia frente a la heurística fija.

La comparación con la política fija muestra un comportamiento dependiente del escenario. En particular, el desempeño relativo de PPO no queda explicado únicamente por la cantidad de puntos de acceso. Las redes 990 y 44 cuentan ambas con 5 APs, pero presentan resultados distintos: en la red 990, PPO alcanza una mediana cercana a la de la heurística, mientras que en la red 44 queda por debajo de la política fija. Además, aunque se incorporó la potencia como variable de decisión para darle mayor flexibilidad al agente, esto no fue suficiente para que PPO superara a la heurística fija en los escenarios de mayor tamaño, como las redes 856 y 838. Por lo tanto, la asignación conjunta de canal y potencia mejora el espacio de decisión disponible, pero no elimina la dificultad asociada a las topologías más densas.

4.8.3. Comparativa entre REINFORCE, A2C y PPO

La figura 4.14 compara la capacidad total obtenida por la política aleatoria y por los tres algoritmos de aprendizaje por refuerzo profundo evaluados: REINFORCE, A2C y PPO. En las seis redes consideradas, las políticas aprendidas alcanzan medianas superiores a las de la política aleatoria, lo que indica que los tres métodos extraen información útil del estado del entorno y no se limitan a reproducir decisiones no coordinadas.

Capítulo 4. Asignación de Recursos con Clientes Dinámicos

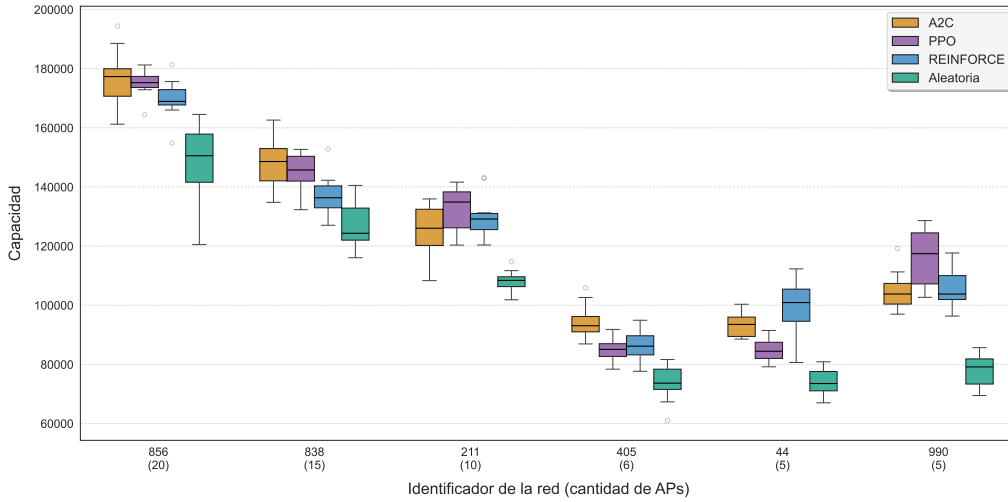


Figura 4.14: Capacidad total de la red alcanzada por la política aleatoria y los algoritmos de aprendizaje A2C, PPO y REINFORCE. Los resultados se presentan para seis redes representativas.

Las diferencias entre algoritmos aparecen principalmente en la dispersión y en el desempeño relativo según la topología. A2C alcanza los valores más altos en algunos escenarios de mayor tamaño, como las redes 856 y 838, aunque con una variabilidad más marcada. PPO presenta distribuciones más compactas, por lo que ofrece un comportamiento más estable, aún cuando no siempre alcanza los máximos observados con A2C. REINFORCE, por su parte, se mantiene por encima de la política aleatoria en los escenarios evaluados, pero no aparece como la mejor alternativa en términos de mediana ni de valores máximos. En conjunto, estos resultados muestran que la elección del algoritmo modifica el compromiso entre capacidad media, valores extremos y estabilidad de la política aprendida.

4.8.4. Comparación Sticky Client

La figura 4.15 compara el desempeño de PPO bajo las tres reglas de asociación de clientes consideradas: *full sticky*, *sticky* y *sticky lite*. En cada caso, tanto el entrenamiento como la evaluación en las distintas redes se realizan usando la misma regla de asociación correspondiente, de modo que la transferencia entre topologías se analiza sin cambiar la dinámica de clientes definida para ese experimento. A priori, permitir que los clientes se reasocien durante el episodio podría volver más difícil el problema de aprendizaje, la acción del agente no solo afecta la capacidad instantánea, sino también la evolución futura de la matriz de asociación $\mathbf{A}_t$. En consecuencia, el entorno observado por la política se vuelve más variable que en el caso *full sticky*, donde las asociaciones quedan fijas luego del ingreso de cada cliente.

Los resultados no muestran una degradación marcada al pasar de *full sticky* a las variantes con reasociación. Las distribuciones de capacidad son similares en la

4.8. Resultados

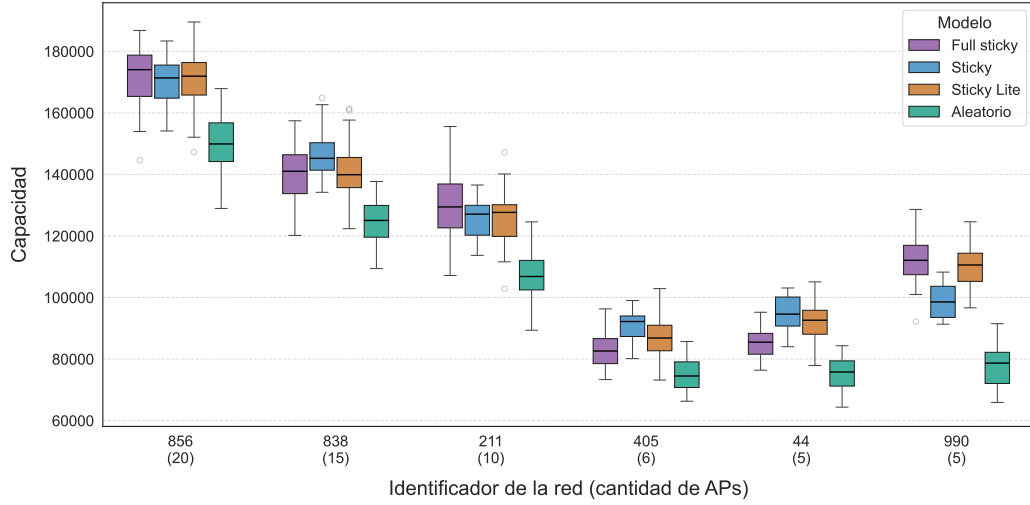


Figura 4.15: Boxplot comparativo de la capacidad obtenida con PPO bajo las tres reglas de asociación de clientes *full sticky*, *sticky* y *sticky lite*, incluyendo la política aleatoria como referencia.

mayoría de las redes, sin un patrón sistemático que indique que *sticky* o *sticky lite* dificulten sustancialmente el entrenamiento en las condiciones evaluadas. Además, en todos los modos de asociación considerados, PPO obtiene capacidades superiores a las de la política aleatoria, lo que indica que la política aprendida conserva una ventaja frente a decisiones no coordinadas aún cuando cambia la dinámica de clientes.

Por lo tanto, la figura se interpreta principalmente como una prueba de robustez frente a distintas dinámicas de asociación, más que como una comparación destinada a identificar una regla universalmente superior. Las diferencias puntuales entre reglas pueden depender de la topología, de la distribución espacial de clientes y APs, y de la estructura de interferencia, pero no alcanzan para concluir que una dinámica complique de forma consistente el aprendizaje respecto a las demás.

4.8.5. Impacto del tamaño del período de decisión

Para evaluar el efecto del período de decisión T , la figura 4.16 muestra la mejora relativa porcentual de PPO respecto a la política aleatoria de cada entrenamiento. El análisis se realiza en el escenario *full sticky*, con asignación conjunta de canal y potencia, considerando períodos de decisión $T = 1$, $T = 3$ y $T = 5$. Para cada valor de T fue necesario entrenar una política distinta, ya que el período de decisión modifica la frecuencia con que el agente actúa y, por lo tanto, la trayectoria de estados y recompensas observada durante el aprendizaje. Por este motivo se evaluó un conjunto acotado de valores de T .

En las tres configuraciones, la distribución de la mejora se ubica por encima de la referencia nula. Esto indica que, para este escenario, la política PPO obtiene capacidades superiores a las de la política aleatoria para los períodos de decisión

Capítulo 4. Asignación de Recursos con Clientes Dinámicos

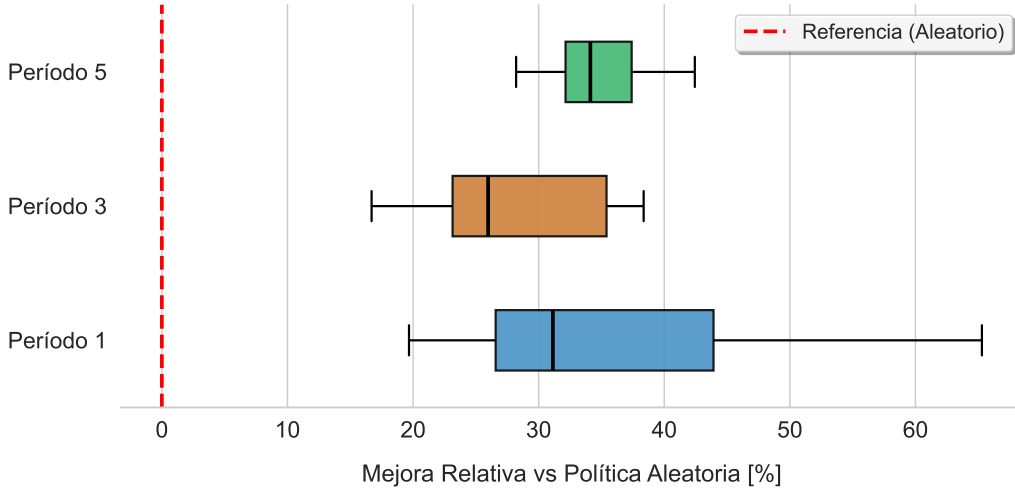


Figura 4.16: Distribución de la mejora relativa porcentual del algoritmo PPO en comparación con una política de asignación aleatoria. Se evalúa el impacto de la longitud del período de decisión (Períodos 1, 3 y 5) en la estabilidad y el rendimiento de la asignación conjunta de canal y potencia para el escenario sticky full (red 990). La línea roja discontinua representa el rendimiento base (0 %) de la política aleatoria.

evaluados. Sin embargo, la magnitud y la dispersión de esa mejora dependen de T .

El caso $T = 1$ presenta la mayor variabilidad, con mejoras distribuidas en un rango amplio. Al aumentar el período de decisión, la dispersión se reduce, especialmente para $T = 5$. Este comportamiento es consistente con la hipótesis de que decisiones menos frecuentes disminuyen la variabilidad inducida por cambios sucesivos de canal y potencia. A la vez, períodos más largos reducen la capacidad del agente de reaccionar ante cambios rápidos del entorno, por lo que el efecto de T no debe interpretarse únicamente como una mejora monótona de estabilidad.

La mediana de la mejora tampoco varía de forma estrictamente monótona: disminuye al pasar de $T = 1$ a $T = 3$ y aumenta nuevamente para $T = 5$. En este conjunto de experimentos, $T = 5$ presenta la mayor mediana y la menor dispersión entre las configuraciones analizadas. Por lo tanto, dentro de los valores evaluados, este período ofrece el mejor compromiso observado entre mejora relativa y variabilidad, sin que ello implique que sea óptimo para otras topologías, reglas de asociación o condiciones de tráfico.

4.8.6. Impacto del POMDP

Hasta el momento, los resultados se han desarrollado bajo la hipótesis de que el sistema opera como un Proceso de Decisión de Markov completamente observable. Sin embargo, en despliegues reales, rara vez se dispone de la totalidad de los datos que definen el estado perfecto. En este caso de estudio, las variables invariables de extraer del entorno son aquellas vinculadas a eventos futuros, específicamente ϵ_t y δ_t . Al prescindir de ellas, el agente deja de observar el estado completo del sistema, por lo que el problema pasa a formularse como un Proceso de Decisión de Markov

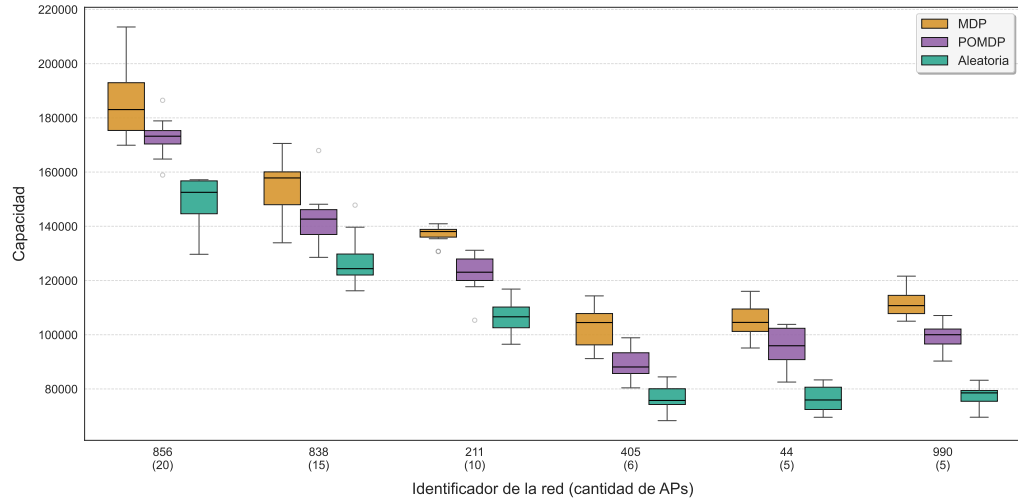


Figura 4.17: Comparación del desempeño del modelo cuando se abandonan las idealidades del marco teórico MDP y se pasa a trabajar con un POMDP.

Parcialmente Observable (POMDP). No obstante, para no privar a la red neuronal de contexto temporal, se propone sustituir estas variables por análogos basados en información empíricamente medible. En el caso de ϵ_t , en lugar de representar el tiempo restante de un cliente en la red, se define ϵ'_t , que simboliza el tiempo que dicho cliente lleva conectado. De forma análoga, δ'_t se define como el tiempo transcurrido desde el último evento de llegada de un cliente al sistema.

En la figura 4.17, se puede observar el desempeño del modelo POMDP cuando se abandonan las idealidades del problema MDP como se describió anteriormente, comparado con el desempeño del caso MDP ideal y de la política de referencia aleatoria. En él, como es esperable, al no conocer estrictamente el futuro, el modelo entrenado bajo el entorno POMDP sufre un claro descenso en su rendimiento, pero aún así logrando superar a la política aleatoria implicando que aún bajo un entorno más realista, sigue siendo posible entrenar un modelo para cumplir los objetivos de este problema, incluso transfiriendo a estructuras mucho más complejas y densas que las vistas durante el entrenamiento.

4.9. Síntesis

En este capítulo se extendió el problema de asignación de recursos a un escenario dinámico, donde la cantidad de clientes activos, sus atenuaciones de canal y sus asociaciones con los APs cambian en el tiempo. A diferencia del caso estático del capítulo anterior, la memoria del sistema no proviene de una restricción acumulada de potencia, sino de la permanencia de los usuarios y de las reglas de asociación *sticky client*. Se mostró que las variables $(\mathbf{H}_t, \mathbf{X}_t, \mathbf{A}_t)$ no bastan para describir una dinámica markoviana, y que es necesario incorporar las duraciones residuales ϵ_t y el tiempo restante hasta la próxima llegada δ_t . Con estas variables, el problema

Capítulo 4. Asignación de Recursos con Clientes Dinámicos

puede formularse como un MDP en la escala de decisión definida por el período T .

Los resultados muestran que las políticas aprendidas superan consistentemente a la política aleatoria, confirmando que la GNN extrae información útil de la topología y del estado dinámico. Entre los algoritmos evaluados, PPO ofrece el comportamiento más estable; A2C alcanza picos mayores, aunque con más varianza; y REINFORCE presenta un rendimiento uniforme, pero sin destacarse como la mejor alternativa.

También se observó que el desempeño no depende únicamente del número de APs. En la asignación conjunta de canal y potencia, PPO mejora de forma consistente a la política aleatoria y alcanza resultados cercanos a la heurística fija en varios escenarios, aunque no la supera de manera sistemática. Esto indica que el enfoque aprendido es prometedor, pero todavía no reemplaza de forma concluyente a una regla heurística bien ajustada. Las diferencias entre escenarios sugieren que la estructura del grafo de interferencia y la distribución espacial son factores determinantes. Además, la comparación entre *full sticky*, *sticky* y *sticky lite* indica que no existe una regla de asociación universalmente dominante.

Finalmente, se analizó el impacto del período de decisión y de la observabilidad parcial. Aumentar T redujo la dispersión de los resultados y, en el entorno evaluado, $T = 5$ ofreció el mejor compromiso observado entre estabilidad y rendimiento. En el caso POMDP, al sustituir variables futuras ideales por observaciones medibles, el desempeño cayó respecto al MDP completamente observable, pero continuó superando a la política aleatoria. En conjunto, los resultados sugieren que la combinación de representaciones basadas en grafos, GNNs y aprendizaje por refuerzo profundo constituye una línea prometedora para la asignación de canales y potencias en redes Wi-Fi dinámicas, aunque su ventaja frente a heurísticas fijas depende del escenario evaluado.

Capítulo 5

Conclusiones y Trabajo futuro

5.1. Discusión final

El objetivo general de esta tesis fue estudiar la aplicación de aprendizaje por refuerzo profundo, parametrizado mediante redes neuronales de grafos, a problemas de asignación de recursos en redes inalámbricas densas a partir de medidas reales de la red de Ceibal. La discusión final retoma los tres objetivos específicos planteados al inicio: evaluar la generalización en la ARPP, formular la ARCD con demanda variable y observabilidad parcial, y comparar estrategias de entrenamiento para este escenario dinámico.

Respecto al primer objetivo, los resultados del problema de enlaces punto a punto muestran que una política basada en GNN puede generalizar a edificios y períodos temporales no observados durante el entrenamiento. En la asignación de canal, la política entrenada con REINFORCE sobre el edificio 856 supera a las políticas de referencia luego de una etapa inicial de exploración y mantiene esa ventaja durante el resto del entrenamiento. Más importante aún, los mismos pesos se aplican sin reentrenamiento a edificios no vistos, con tamaños y estructuras diferentes, y conservan una mejora consistente frente al acceso no coordinado.

La extensión de la ARPP hacia la asignación conjunta de canal y potencia muestra que el enfoque sigue siendo aplicable cuando aumenta la complejidad del espacio de acciones. Bajo una restricción de potencia más exigente, la política aprendida vuelve a superar a la referencia aleatoria en los edificios evaluados. Sin embargo, la convergencia resulta más lenta y la brecha continúa creciendo hasta el final del entrenamiento. Esto indica que el aumento del espacio de acciones no invalida la formulación, pero vuelve más demandante la exploración y puede requerir mayores horizontes de entrenamiento o algoritmos de menor varianza.

En relación con el segundo objetivo, se formuló la ARCD como un problema con clientes dinámicos, demanda variable y grafos AP-cliente de tamaño cambiante. A diferencia del caso punto a punto, el desempeño queda condicionado no solo por la topología, sino también por la evolución temporal de los clientes, las reglas de asociación y la información disponible para el agente. El análisis del estado permitió distinguir entre una formulación MDP ideal, que incorpora variables no observables

Capítulo 5. Conclusiones y Trabajo futuro

en una red real, y una formulación parcialmente observable basada en información medible.

En este escenario dinámico, las políticas aprendidas también superan de forma consistente a la política aleatoria. No obstante, la comparación contra la política fija depende más de la topología. En algunos edificios PPO iguala o supera la heurística, mientras que en otros captura solo una parte de la ganancia posible entre la política aleatoria y la política fija. Aun así, el modelo de RL se entrenó exclusivamente en un edificio de 5 APs y se transfirió al resto de las topologías evaluadas sin reentrenamiento, mientras que la heurística fija depende directamente de las condiciones de cada red.

Respecto al tercer objetivo, la comparación entre REINFORCE, A2C y PPO muestra que el algoritmo de entrenamiento tiene un efecto importante sobre la estabilidad del aprendizaje y la tasa alcanzada. A2C obtiene los picos de capacidad más altos en algunos escenarios complejos, especialmente en edificios grandes, pero con mayor varianza. REINFORCE se comporta de manera estable, aunque sin ubicarse como la alternativa dominante. PPO mantiene una varianza menor y resultados más consistentes entre escenarios, por lo que aparece como la opción más robusta dentro de las alternativas evaluadas.

Las variantes de asociación *sticky client* y los distintos períodos de decisión también afectan la dificultad de entrenamiento. La regla de asociación modifica cómo evolucionan las conexiones entre clientes y APs, mientras que el período T determina con qué frecuencia la política puede actualizar canales y potencias. En los experimentos, aumentar T redujo la dispersión de los resultados, aunque también limita la capacidad de reacción frente a cambios rápidos del entorno. Por lo tanto, la asociación de clientes y la escala temporal de decisión forman parte de la definición efectiva del problema de control.

En conjunto, los resultados muestran que las GNN son una herramienta adecuada para parametrizar políticas de asignación de recursos en redes inalámbricas reales, especialmente cuando la topología cambia y no es viable diseñar una política independiente para cada edificio. Al mismo tiempo, el desempeño no depende únicamente de la capacidad expresiva de la GNN, sino también de la dificultad del espacio de acciones, la estabilidad del algoritmo de RL, la información disponible para el agente y la estructura particular de cada red. En escenarios dinámicos, además, intervienen la evolución temporal de los clientes, la regla de asociación, el período de decisión y el grado de observabilidad del estado.

5.2. Trabajo futuro

Una primera línea de continuidad consiste en estudiar con mayor profundidad qué propiedades topológicas explican cuándo PPO supera a la política fija y cuándo queda por debajo. Los resultados sugieren que el número de APs no es un predictor suficiente. Sería útil incorporar métricas del grafo de interferencia, como centralidad, modularidad, distribución de grados, espectro del Laplaciano o presencia de comunidades, para correlacionarlas con la fracción de ganancia que captura el agente.

5.2. Trabajo futuro

Una segunda línea consiste en mejorar la formulación parcialmente observable. En esta tesis, el agente opera con información limitada sobre duraciones residuales y próximas llegadas. Una extensión natural es incorporar memoria en la política mediante redes recurrentes, filtros de estado o mecanismos de atención temporal, de modo que el agente pueda inferir variables ocultas a partir de la historia de observaciones. Esto permitiría aproximar mejor el estado latente del sistema sin asumir que la red conoce información que, en la práctica, no estaría disponible.

También resulta necesario profundizar en la estabilidad del entrenamiento. La comparación entre REINFORCE, A2C y PPO muestra que el algoritmo de optimización puede ser tan determinante como la arquitectura. Futuras extensiones podrían evaluar variantes de PPO con normalización más robusta de ventajas, métodos off-policy, aprendizaje por currículum, regularización explícita de entropía o entrenamiento multi-edificio, donde cada episodio provenga de una topología distinta para favorecer una generalización más uniforme.

Otra dirección relevante es ampliar el espacio de acciones y las restricciones del modelo. En redes reales, la asignación de canal y potencia convive con restricciones regulatorias, costos de cambio, límites de calidad por cliente, criterios de equidad y objetivos de estabilidad. Incorporar penalizaciones por reconfiguración, métricas de tasa mínima o funciones de utilidad proporcionales permitiría pasar de maximizar capacidad agregada a optimizar desempeño de red con garantías más cercanas a la operación práctica.

Asimismo, sería valioso reemplazar o complementar algunas reglas heurísticas del simulador por modelos aprendidos desde datos. En particular, la dinámica de asociación de clientes, los umbrales de traspaso y la evolución temporal del canal podrían calibrarse con registros operativos más detallados. Esto reduciría la brecha entre simulación y despliegue, y permitiría evaluar si las políticas aprendidas conservan su ventaja bajo modelos de movilidad y tráfico más ricos.

Esta página ha sido intencionalmente dejada en blanco.

Apéndice A

Apéndice

A.1. Repositorio de Código y Reproducibilidad

Todo el código para simulación del entorno, entrenamiento y evaluación de los modelos presentados y representación de los resultados, se encuentran disponibles en este repositorio de GitLab

A.2. Declaración de uso de Inteligencia Artificial

Siguiendo los lineamientos de la *Guía para el uso ético y crítico de inteligencia artificial en las unidades curriculares de la Facultad de Ingeniería* [22], se notifica que durante el desarrollo de este documento se utilizaron las herramientas de inteligencia artificial generativa Claude (Anthropic) y Gemini (Google). Estas tecnologías fueron empleadas como asistentes para la revisión de la redacción y corrección de estilo.

Cumpliendo con dicha guía, todo el contenido, sugerencias y desarrollos de software derivados de estas herramientas fueron verificados, corregidos y validados por los autores.

Esta página ha sido intencionalmente dejada en blanco.

Lista de Siglas

5G	Fifth Generation (quinta generación de redes celulares)
A2C	Advantage Actor-Critic (Ventaja Actor-Crítico)
AP	Access Point (Punto de Acceso)
ARCD	Asignación de Recursos con Clientes Dinámicos
ARPP	Asignación de Recursos en Redes de enlaces Punto a Punto
CCI	Co-Channel Interference (Interferencia Cocanal)
GAT	Graph Attention Network (Red de Atención en Grafos)
GNN	Graph Neural Network (Red Neuronal de Grafos)
MDP	Markov Decision Process (Proceso de Decisión de Markov)
MLP	Multi-Layer Perceptron (Perceptrón Multicapa)
NDP	Neighbor Discovery Protocol (Protocolo de Descubrimiento de Vecinos)
PE	Positional Encoding (Codificación Posicional)
POMDP	Partially Observable Markov Decision Process (Proceso de Decisión de Markov Parcialmente Observable)
PPO	Proximal Policy Optimization (Optimización de Política Proximal)
PyG	PyTorch Geometric
REINFORCE	$REward Increment = Nonnegative Factor \times Offset$ Reinforcement $\times$ Characteristic Eligibility
RL	Reinforcement Learning (Aprendizaje por Refuerzo)
RRM	Radio Resource Management (Gestión de Recursos de Radio)
RSSI	Received Signal Strength Indicator (Indicador de Fuerza de la Señal Recibida)

Apéndice A. Apéndice

SAGE	SAmple and aggreGatE
SINR	Signal to Interference plus Noise Ratio (Relación Señal a Interferencia más Ruido)
SNMP	Simple Network Management Protocol (Protocolo Simple de Administración de Red)
WLC	Wireless LAN Controller (Controlador de LAN Inalámbrica)

Referencias

- [1] M. Eisen and A. Ribeiro, “Optimal wireless resource allocation with random edge graph neural networks,” *IEEE Transactions on Signal Processing*, vol. 68, pp. 2977–2991, 2020.
- [2] G. Capdehourat, F. Larroca, and G. Morales, “A nation-wide Wi-Fi RSSI dataset: Statistical analysis and resulting insights,” in *2020 IFIP Networking Conference*, 2020, pp. 370–378.
- [3] S. Fernández, R. G. Camargo, M. Eisen, A. Ribeiro, and F. Larroca, “On the transferability of graph neural networks for resource allocation in wireless networks,” in *2024 IEEE URUCON*, 2024, pp. 1–5.
- [4] C. I. Kanatsoulis and A. Ribeiro, “Graph neural networks are more powerful than we think,” in *ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 7550–7554.
- [5] M. Randall Carlevaro, “Learning-based resource allocation for fair and efficient mobile networks,” Ph.D. dissertation, Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay, 2026.
- [6] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA: MIT Press, 2018.
- [7] I. Amigo, J. A. Bazerque, and M. del Castillo, “Taller de aprendizaje por refuerzo,” Curso de posgrado, Facultad de Ingeniería, Universidad de la República, 2026, disponible en: <https://eva.fing.edu.uy/course/view.php?id=1908> [Accedido: 20 de junio de 2026].
- [8] R. S. Sutton, D. A. McAllester, S. Singh, and Y. Mansour, “Policy gradient methods for reinforcement learning with function approximation,” in *Advances in Neural Information Processing Systems 12*. MIT Press, 2000, pp. 1057–1063.
- [9] R. J. Williams, “Simple statistical gradient-following algorithms for connectionist reinforcement learning,” *Machine Learning*, vol. 8, pp. 229–256, 1992.
- [10] P. Bermolen, M. Fiori, F. La Rocca, B. Marengo, S. Pérez, and G. Mateos, “Aprendizaje automático para datos en grafos,” Curso semestral, Facultad de Ingeniería, Universidad de la República, 2025, disponible en: <https://eva.fing.edu.uy/course/view.php?id=1626> [Accedido: 20 de junio de 2026].

Referencias

- [11] W. L. Hamilton, R. Ying, and J. Leskovec, “Inductive representation learning on large graphs,” 2018. [Online]. Available: <https://arxiv.org/abs/1706.02216>
- [12] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph attention networks,” 2018. [Online]. Available: <https://arxiv.org/abs/1710.10903>
- [13] M. Towers, A. Kwiatkowski, J. K. Terry, J. U. Balis, G. De Cola, T. Deleu, M. Goulão, A. Kallinteris, M. Krimmel, A. KG, R. Perez-Vicente, A. Pierré, S. Schulhoff, J. J. Tai, H. Tan, and O. G. Younis, “Gymnasium: A standard interface for reinforcement learning environments,” 2024.
- [14] M. Fey and J. E. Lenssen, “Fast graph representation learning with PyTorch Geometric,” in *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.
- [15] J. Du, S. Zhang, G. Wu, J. M. F. Moura, and S. Kar, “Topology adaptive graph convolutional networks,” 2018. [Online]. Available: <https://arxiv.org/abs/1710.10370>
- [16] R. J. Williams and J. Peng, “Function optimization using connectionist reinforcement learning algorithms,” *Connection Science*, vol. 3, no. 3, pp. 241–268, 1991.
- [17] D. P. Kingma and J. L. Ba, “Adam: A method for stochastic optimization,” 2015. [Online]. Available: <https://arxiv.org/pdf/1412.6980>
- [18] D. Lim, J. Robinson, L. Zhao, T. Smidt, S. Sra, H. Maron, and S. Jegelka, “Sign and basis invariant networks for spectral graph representation learning,” 2022. [Online]. Available: <https://arxiv.org/abs/2202.13013>
- [19] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.
- [20] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” 2017.
- [21] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, “High-dimensional continuous control using generalized advantage estimation,” in *International Conference on Learning Representations (ICLR)*, 2016.
- [22] Facultad de Ingeniería, Universidad de la República, “Guía para el uso ético y crítico de inteligencia artificial en las unidades curriculares de la Facultad de Ingeniería,” 2026, disponible en: https://www.fing.edu.uy/sites/default/files/2026-02/guia-de-etica-fing_2026.pdf [Accedido: 30 de junio de 2026].

Índice de tablas

3.1.	Espacio discreto de acciones por enlace para $C = 3$ canales y $P = 2$ niveles de potencia. La opción 0 corresponde a que ese enlace no transmite; las opciones 1–6 asignan canal y potencia de forma conjunta.	19
3.2.	Comparativa de estrategias de asignación para la topología de interferencia central.	29
3.3.	Comparativa de estrategias de asignación para la topología de interferencia total con 3 canales disponibles.	30
3.4.	Hiperparámetros específicos para cada variante del modelo.	32
4.1.	Hiperparámetros de entrenamiento y configuración de la arquitectura.	60

Esta página ha sido intencionalmente dejada en blanco.

Índice de figuras

2.1.	Ciclo de interacción entre el agente y el entorno en un problema de aprendizaje por refuerzo. En cada instante t el agente observa el estado s_t , ejecuta una acción a_t y el entorno responde con una recompensa r_{t+1} y un nuevo estado s_{t+1}	9
3.1.	Ejemplo esquemático de dos grafos evaluados con una entrada constante. El grafo $\mathcal{G}$ corresponde a la topología de interferencia total usada en la verificación de funcionamiento.	19
3.2.	Diagrama en bloques de la arquitectura de la GNN. Las características iniciales $\mathbf{X}$ pasan por una capa lineal para homogeneizar su dimensionalidad, seguido de múltiples capas convolucionales TAG-Conv que agregan información de los vecinos ponderadas por la matriz de canal $\mathbf{H}$. La red finaliza arrojando los <i>logits</i> que determinan la preferencia del agente por cada acción en el espacio $\mathcal{A}$	22
3.3.	Mapas de calor de la matriz de canal $\mathbf{H}$ para el edificio 856 en dos snapshots del año lectivo 2018: marzo (izquierda) y julio (derecha), con escala de color compartida. El patrón de entradas intensas cambia sustancialmente entre meses pares. Un ejemplo se puede ver en la relación (8,17), la cual domina en julio y no tiene contraparte visible en marzo.	26
3.4.	Distribución del número de enlaces punto a punto por edificio en el conjunto de datos de Ceibal (1 424 edificios). La mediana se ubica en 6 enlaces punto a punto ($Q1 = 4$, $Q3 = 9$) y la moda en 5 enlaces; la mayoría de las redes son de pequeña escala. El edificio 856, utilizado como caso principal de entrenamiento con 20 enlaces punto a punto, se ubica en el percentil 95 de la distribución.	27
3.5.	Topología del grafo de interferencia del edificio 856 en marzo, para las bandas de 2,4 GHz (izquierda, 241 enlaces) y 5 GHz (derecha, 101 enlaces). Cada círculo representa un nodo de la red; el tamaño es proporcional al grado y el color codifica la centralidad de intermediación (<i>betweenness centrality</i>), los nodos más claros son los que aparecen con mayor frecuencia en los caminos del grafo. Los tres nodos de mayor centralidad se destacan en rojo.	27

Índice de figuras

3.6.	Promedio de aristas del grafo de interferencia por instante de medición en marzo para los ocho edificios utilizados en los experimentos de transferibilidad, por banda. En todos los casos la banda de 2,4 GHz (violeta) supera a la de 5 GHz (verde), con ratios que van desde $1,2\times$ en el edificio 201 hasta $2,9\times$ en el edificio 84. El edificio 856, caso de entrenamiento con 241 aristas promedio en 2,4 GHz, no aparece aquí porque estos edificios son exclusivamente los destinos de transferencia.	28
3.7.	Ejemplo esquemático de los dos grafos evaluados en las pruebas de funcionamiento. El grafo $\mathcal{G}$ corresponde a la topología de interferencia central mientras que el grafo $\hat{\mathcal{G}}$ corresponde a la topología de interferencia total.	28
3.8.	Evolución de la distribución de probabilidad sobre acciones en el escenario de interferencia central. Enlace 0 y enlace 2 convergen al canal 0; enlace 1 converge al canal 2. La probabilidad de la acción No TX se va a cero en los tres enlaces.	29
3.9.	Evolución de la distribución de probabilidad sobre acciones en el escenario de interferencia total. La asimetría introducida permite que la GNN distinga los extremos mediante sus características k -hop, convergiendo a tres canales ortogonales distintos (Aislamiento Total) y anulando las colisiones.	31
3.10.	Evolución de la suma de tasas durante el entrenamiento sobre el edificio 856. La política aprendida con REINFORCE requiere un período inicial de exploración antes de superar a la política de referencia; luego mantiene esa ventaja de forma consistente.	33
3.11.	Distribución de la suma de tasas al transferir los pesos entrenados en el edificio 856 a siete edificios no vistos durante el entrenamiento. REINFORCE-GNN (violeta) se compara contra acceso aleatorio no coordinado (verde) y potencia fija en canal estático (naranja); esta última asigna a cada enlace un canal sorteado una sola vez y transmite siempre a $0,8 \cdot P_{\text{máx}}$ sin reconfigurar.	33
3.12.	Evolución de la suma de tasas promedio durante el entrenamiento sobre el edificio 856 con asignación conjunta de canal y potencia, bajo una restricción del 60 % de la potencia máxima. REINFORCE (violeta) supera a la referencia aleatoria (verde) desde las primeras iteraciones y la brecha sigue creciendo hasta el final del entrenamiento.	34
3.13.	Al igual que en el caso anterior se ve la suma de tasas al transferir los pesos entrenados en el edificio 856 a siete edificios no vistos durante el entrenamiento. REINFORCE-GNN (violeta) se compara contra acceso aleatorio no coordinado (verde) y potencia fija en canal estático (naranja); esta última asigna a cada enlace un canal sorteado una sola vez y transmite siempre a $0,6 \cdot P_{\text{máx}}$ sin reconfigurar.	35
4.1.	Distribución empírica de las intensidades de señal recibida (RSSI) presentes en el dataset de Cisco Ceibal para las bandas de 2.4 GHz y 5 GHz.	43

4.2.	Distribución empírica de la fluctuación del canal (Δ RSSI) para la población general de clientes conectados al nodo de mayor tráfico de la red 211. Los datos fueron centrados respecto a la media individual de cada cliente. El ajuste a una función de densidad de probabilidad Normal (curva sólida) evidencia que el ruido de propagación en la red inalámbrica sigue un comportamiento Gaussiano.	44
4.3.	Dinámica temporal de los usuarios en el simulador. Se observa el ciclo de vida de los clientes en la red durante un episodio, donde cada barra representa el intervalo de permanencia de un usuario Δ_u . El color de cada bloque identifica el Punto de Acceso (AP) al cual el cliente fue asignado basándose en la señal empírica óptima.	45
4.4.	Estructura topológica del grafo heterogéneo $\mathcal{G}_t$ construido en un instante t . La representación abstrae el entorno de red diferenciando nodos de Puntos de Acceso (cuadrados) y nodos de clientes activos (círculos). Para mayor claridad visual, se ilustran únicamente las relaciones AP–cliente: las líneas negras sólidas indican las asociaciones activas, mientras que las líneas punteadas representan las atenuaciones medidas desde los APs hacia los clientes.	46
4.5.	Mapa de calor representativo de la matriz de interferencia mutua normalizada $\mathbf{W}$ para las celdas de la red analizada. Los pesos cercanos a 1.0 reflejan un fuerte solapamiento percibido, determinando así la fuerza de los pesos de las aristas w_{ij}	47
4.6.	Diagrama de flujo del procedimiento utilizado para generar el <i>Positional Encoding</i> estructural a partir de un grafo de entrada. El proceso parte de una representación matricial del grafo, sobre la cual se realiza una descomposición espectral para obtener valores y vectores propios. Luego se seleccionan los vectores asociados a las componentes de menor frecuencia, se construye con ellos una matriz de coordenadas posicionales para los nodos.	49
4.7.	Diagrama de la arquitectura de la capa SignNet presentada por Lim et al. [18].	51
4.8.	Diagrama de la arquitectura de la red GNN completa diseñada para el problema ARCD, particularmente para una estructura de tipo Actor-Critico.	53
4.9.	Diagrama de la arquitectura de una capa de la red GNN diseñada para el problema ARCD.	55
4.10.	Evolución de la recompensa acumulada por episodio durante el entrenamiento de REINFORCE, A2C y PPO en la red 990 para el problema conjunto de asignación de canal y potencia.	60
4.11.	Distribución de la capacidad total obtenida en la asignación de canales para la red 990, utilizada durante el entrenamiento del agente. Se comparan la política fija heurística, la política aleatoria y la política aprendida mediante PPO con GNN.	61

Índice de figuras

4.12. Distribución de la capacidad total obtenida al resolver la asignación de canal en múltiples redes. Se comparan la política fija, la política aleatoria y la política aprendida mediante PPO con GNN, entrenada en la red 990 y evaluada en el resto de las topologías. La comparación permite analizar la capacidad de transferencia del agente cuando cambian el tamaño y la estructura de la red.	62
4.13. Capacidad total de la red alcanzada en el problema conjunto de asignación de canal y potencia por la política heurística estricta (política fija), la política de acceso aleatorio no coordinado (Política Aleatoria) y la política PPO (GNN), en seis redes representativas del conjunto de datos.	63
4.14. Capacidad total de la red alcanzada por la política aleatoria y los algoritmos de aprendizaje A2C, PPO y REINFORCE. Los resultados se presentan para seis redes representativas.	64
4.15. Boxplot comparativo de la capacidad obtenida con PPO bajo las tres reglas de asociación de clientes <i>full sticky</i> , <i>sticky</i> y <i>sticky lite</i> , incluyendo la política aleatoria como referencia.	65
4.16. Distribución de la mejora relativa porcentual del algoritmo PPO en comparación con una política de asignación aleatoria. Se evalúa el impacto de la longitud del período de decisión (Períodos 1, 3 y 5) en la estabilidad y el rendimiento de la asignación conjunta de canal y potencia para el escenario <i>sticky full</i> (red 990). La línea roja discontinua representa el rendimiento base (0 %) de la política aleatoria.	66
4.17. Comparación del desempeño del modelo cuando se abandonan las idealidades del marco teórico MDP y se pasa a trabajar con un POMDP.	67

Esta es la última página.
Compilado el lunes 3 agosto, 2026.
<http://iie.fing.edu.uy/>