Universidad Nacional de San Luis
Facultad de Ingeniería y Ciencias Agropecuarias

DISEÑO DE ALGORITMO DE APRENDIZAJE POR REFUERZOS PARA LA OPTIMIZACION DE SISTEMAS PEER-TO-PEER EN MICROGRIDS

Bruno Boato

Trabajo final de Ingeniería mecatrónica

Director

Nicolas Nehuén Antonelli, 36982195

Villa Mercedes, San Luis
2025

DERECHO DE AUTOR

DEDICATORIA

Le dedico este trabajo principalmente a mis padres, dado gracias a su apoyo moral y económico he podido dedicarme a la ingeniería, y a mi tía por haberme aceptado en su casa por 4 años para que pudiera acceder a una escuela técnica y descubriera mi interés en la robótica que me llevo a estudiar mecatrónica.

AGRADECIMIENTOS

Quisiera expresar mi más profundo agradecimiento a todas las personas que hicieron posible que hoy pueda estar aquí culminando esta etapa.

A mi familia, por su apoyo constante y por ser mi sostén en cada paso de este camino. A mis profesores, en particular, a Dr. Luis Ávila, por haberme abierto las puertas al mundo de la investigación, y junto al Dr. Maximiliano Ascencio, por su generosidad y confianza al escribirme tantas cartas de recomendación que me ayudaron a abrir puertas que sin su apoyo se hubiesen mantenido cerradas. Y a mis compañeros y amigos, por su apoyo y por compartir tantas horas de estudio, que hicieron que el camino fuera mucho más llevadero.

RESUMEN

La estructura de las redes de energía eléctrica ha evolucionado significativamente en los últimos años con la introducción de nuevos actores como las unidades de generación distribuida, el incremento en la capacidad de los sistemas de almacenamiento y la progresiva introducción de la infraestructura requerida para la recarga de vehículos eléctricos. La incorporación de estos nuevos actores ha abierto oportunidades para repensar la forma en que tradicionalmente se ha gestionado la energía, impulsando la búsqueda de soluciones que permitan un uso eficiente, confiable y responsable de la energía, apelando a una concepción optimizada y más flexible de la red bajo el concepto de smart grid. En este contexto, los sistemas Peer-to-Peer (P2P) surgen como una expresión concreta de la evolución hacia redes inteligentes, al facilitar el intercambio directo de energía entre agentes distribuidos, que pueden actuar como consumidores y productores simultáneamente. Sin embargo, la coordinación eficiente de estos sistemas distribuidos plantea desafíos complejos debido a la heterogeneidad de los participantes, la variabilidad de las fuentes renovables y el volumen de información a procesar en tiempo real. Frente a esta complejidad, el uso de técnicas de inteligencia artificial (IA), y en particular el Aprendizaje por Refuerzos (RL), se presenta como una herramienta clave para su desarrollo y éxito futuros. La ventaja de estas técnicas frente a otros enfoques es que permiten identificar patrones o tendencias en los datos y extraer conocimiento crucial para tomar decisiones, hacer predicciones precisas de los estados futuros del sistema e identificar escenarios anómalos que pueden conducir a un comportamiento ineficiente.

Este trabajo tiene como objetivo el diseño e implementación de estrategias de gestión inteligente de energía basadas en RL aplicadas a smart grid con arquitectura P2P. La hipótesis fundamental se centra en que las técnicas de RL permitirían desarrollar estrategias para la gestión de una red eléctrica con el fin de mantenerla en equilibrio, permitiendo reprogramar el consumo de energía mediante predicciones de demanda y energía disponible a escala de red. Eventualmente, el desarrollo de modelos del comportamiento de una smart grid contribuirá no sólo al logro de una mayor racionalidad y eficiencia en el uso de la energía, sino también la gestión de diversas fuentes distribuidas sin perjuicio de la diversidad de origen ni de su tamaño de generación. De esta manera se pretende contribuir a la inserción de este tipo de tecnologías en nuestro país, tanto a través del aporte de tecnologías específicas para el sector, como en la formación de recursos humanos altamente calificados en esta temática.

Palabra claves - Sistemas Inteligentes, Redes Peer-to-Peer, Microrredes

ÍNDICE DE CONTENIDO

CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPITULO 1: Propuesta

1.1 Introducción

El cambio climático en curso está obligando a trasladar la generación eléctrica de las plantas de energía fósil a la generación renovable. En los últimos años, tanto el desarrollo tecnológico como el impulso político llevaron a un incremento relevante en la participación de energías renovables. Aun así, las líneas de transmisión que conectan a las grandes centrales con los consumidores regionales poseen una topología local en forma de estrella, debido al dominio de las grandes centrales eléctricas convencionales. Es probable que la entrada de más fuentes de energía renovable lleve a topologías más descentralizadas y recurrentes debido a su capacidad de generación distribuida. En tal escenario, los consumidores podrán actuar como productores y consumidores al mismo tiempo, ocasionando que el transporte de electricidad ya no sea unidireccional.

En busca de optimizar el funcionamiento de la infraestructura eléctrica y el despliegue eficiente de sistemas de generación distribuidos basados en recursos renovables, los nuevos enfoques conducen de modo ineludible al desarrollo de redes eléctricas inteligentes, más conocidas como smart grids [1]. Una smart grid es una colección de tecnologías, conceptos, topologías y estrategias que facilitan que las etapas de generación, transmisión y distribución de energía sean integradas por un entorno inteligente y con capacidad de toma de decisiones complejas. Así, una smart grid se puede ver como una red que se adapta a distintos modos de operación, permitiendo además a los usuarios interactuar con el sistema de administración de energía para monitorear y ajustar su consumo de energía y reducir sus costos. Cabe aclarar que una smart grid no debería operar de manera aislada, por el contrario, debería coexistir con la red eléctrica existente añadiendo capacidades y funcionalidades a través de un camino evolutivo.

Se espera que el crecimiento orgánico y la evolución de la red inteligente provengan de la integración de algunas estructuras básicas llamadas microrredes inteligentes. Las microrredes generan, distribuyen y regulan el flujo de energía eléctrica a los consumidores, pero lo hacen a nivel local. Por lo tanto, constituyen una manera ideal de integrar recursos renovables a nivel comunitario y de favorecer la participación activa de los clientes en la administración de la energía. La incorporación de nuevos actores como las unidades de generación distribuida, principalmente a partir de parques eólicos y solares, están aportando nuevas posibilidades y flexibilidad en la forma en que tradicionalmente se ha gestionado la energía. Como consecuencia de la emergente participación de los usuarios en la gestión energética, han comenzado a explorarse modelos de intercambio descentralizado como los

sistemas Peer-to-Peer (P2P). Este tipo de arquitectura permite que los usuarios no solo consuman energía, sino que también puedan intercambiarla directamente con otros actores de la red, actuando como prosumidores. Estas transacciones energéticas entre pares promueven el uso eficiente de los recursos distribuidos, incentivan la producción renovable a pequeña escala y fortalecen la participación ciudadana en la transición energética.

Sin embargo, uno de los mayores desafíos de mantener un funcionamiento estable de una red de energía es conseguir un equilibrio entre la oferta y la demanda. Por consiguiente, es indispensable ajustar las fluctuaciones en el suministro descentralizado proveniente de las fuentes renovables a las demandas distribuidas y temporalmente inciertas. Para hacer coincidir la generación y la demanda en una red eléctrica completamente renovable con las características de la demanda actual en cada momento, existen soluciones que son poco rentables como por tal como el uso de grandes instalaciones de almacenamiento energético. Una opción más viable es proponer medidas basadas en incentivos para influir en los patrones de consumo de electricidad, de modo que se haga un uso más eficiente de la energía. Esto a menudo asegura que la infraestructura existente se utilice de manera eficiente, a fin de satisfacer el aumento de la demanda y minimizar las inversiones en capacidad de generación adicional [1]. Otra posibilidad es manejar esa intermitencia mediante sistemas de gestión del almacenamiento, impulsados por la progresiva introducción de la infraestructura requerida para la recarga de vehículos eléctricos [2]. En este sentido, las baterías de iones de litio se han convertido en el dispositivo de suministro de energía estándar para vehículos eléctricos, pero aún deben hacer frente a ciertos desafíos importantes antes de convertirse en una tecnología de consumo masivo [3]. Por ejemplo, aunque la carga a velocidad baja puede ser una buena opción para alargar la vida útil de la batería, un tiempo de carga considerable puede comprometer la comerciabilidad de los vehículos eléctricos. Por lo tanto, es necesario desarrollar políticas de carga optimizadas que eviten tanto los efectos graves de degradación de la batería como las molestias a los usuarios finales.

No obstante, más allá del desarrollo de tecnologías de almacenamiento, persisten desafíos para hacer frente a la creciente complejidad del sistema resultante [4]. Por ejemplo, la naturaleza intermitente y variable de la potencia renovable disponible trae aparejada cierta incertidumbre no programable, que amenaza la confiabilidad y estabilidad de los sistemas energéticos [5]. Por lo tanto, en una smart grid la tarea de monitoreo como herramienta de predicción y detección de situaciones anómalas es un atributo indispensable para determinar la energía que estará disponible en un período próximo, con el fin de mejorar la calidad de la programación operativa y reducir el tamaño de las reservas de capacidad auxiliares. De igual manera, el monitoreo permitirá a los operadores del servicio público tomar mejores decisiones

de gestión en términos de mercado y permitiendo lograr una mejor distribución del servicio. Ante la necesidad de tomar decisiones en entornos dinámicos y con múltiples variables, el Aprendizaje por Refuerzo (RL, por sus siglas en ingles) se posiciona como una herramienta poderosa para la gestión inteligente de las smart grids. A través de esta técnica, un agente de control puede aprender políticas de acción óptimas interactuando directamente con un entorno eléctrico donde las condiciones varían constantemente. El agente maximiza una señal de recompensa asociada a objetivos como el balance oferta-demanda, la reducción de costos o la prolongación de la vida útil de baterías de vehículos eléctricos.

Es importante considerar que el Plan Argentina Innovadora 2020 a través del “Núcleo Socio-Productivo Estratégico (NSPE): Uso Racional y Eficiente de la Energía”, considera que la energía es un recurso estratégico para el desarrollo socio-productivo argentino y que nuestro país puede aprovechar las ventajas de la amplia matriz de fuentes renovables y no renovables, así como de los avances tecnológicos que permitan un consumo cada vez más eficiente. Se pone especial consideración en el desarrollo de redes inteligentes de transmisión y distribución de electricidad, con interconexión a la generación de fuentes renovables y el mejoramiento de la eficiencia de las redes eléctricas existentes. A su vez, el Plan Nacional de Inteligencia Artificial de la Agenda Digital 2030 plantea que la Inteligencia Artificial (IA) aporta valor al sector aumentando la competitividad a través de mejoras en la productividad, optimización de recursos, maximización de la eficiencia, disminución de costos; contribuyendo además a la generación de conocimiento. A pesar de que en Argentina hoy en día no se permite el intercambio de energía entre usuarios particulares, en otros países tales como aquellos pertenecientes a la Unión Europea, ya hay un marco regulatorio para permitir este tipo de transacciones [6], [7], además en países como Suiza ya se han implementado proyectos de este tipo [8]. Esto constituye indudablemente un área de investigación que se encamina al desarrollo de una red de abastecimiento de energía que satisfaga las necesidades de nuestro país a futuro.

1.1.1 Objetivo general

- El objetivo es el diseño de estrategias basadas en IA para la gestión inteligente de los recursos de energía distribuidos disponibles considerando la creciente capacidad de integración de fuentes renovables. Las estrategias basadas en IA permitirán gestionar de forma inteligente una red eléctrica, integrando y balanceando todos los recursos distribuidos, mientras genera una red local sostenible que pueda intercambiar energía con la red principal.

1.1.2 Objetivos específicos

- Modelado de smart grids: el objetivo es implementar modelos dentro de simulaciones aplicados a sistemas complejos caracterizados por la interacción entre componentes sociotécnicos, con el fin de comprender los principales desafíos vinculados a la gestión de redes eléctricas inteligentes. Se investigarán metodologías y enfoques capaces de representar de manera precisa la dinámica y la complejidad proveniente de la interacción entre agentes tecnológicos (como dispositivos de generación, almacenamiento y control) y agentes sociales (como usuarios, prosumidores y reguladores). Se realizarán simulaciones computacionales avanzadas para evaluar el desempeño y la eficacia de los modelos y estrategias propuestos.

- Implementar algoritmos de RL adaptados a contextos energéticos distribuidos, que permitan a los agentes aprender políticas óptimas para la toma de decisiones. Esto abarca tanto escenarios de gestión dinámica de precios para equilibrar los costos entre los distintos participantes del sistema (prosumidores, consumidores y proveedores de servicio), como la optimización de decisiones operativas (compra, venta o almacenamiento de energía) orientadas a la minimización de costos operativos en la red.

1.2 Alcances y limitaciones

El presente proyecto tiene como objetivo principal el desarrollo e implementación de algoritmos de gestión basados en IA aplicados a redes eléctricas modernas que incorporan recursos energéticos distribuidos, tales como generación renovable, almacenamiento de energía y cargas gestionables. La gestión inteligente de estos recursos resulta fundamental para mejorar la eficiencia, la resiliencia y la sostenibilidad del sistema eléctrico en el contexto de la transición energética.

Los algoritmos serán diseñados para operar bajo diferentes condiciones de operación de la red, considerando variaciones en la demanda, la disponibilidad de generación renovable y posibles contingencias del sistema. Con el fin de evaluar su desempeño, dichos algoritmos serán validados experimentalmente en un entorno de simulación que represente una red eléctrica de distribución, permitiendo realizar pruebas bajo escenarios diversos y controlados.

Es importante señalar que el alcance del trabajo se limita al desarrollo teórico y computacional, sin contemplar la implementación física o en campo de los sistemas propuestos. Las pruebas y validaciones se llevarán a cabo exclusivamente en un entorno simulado, utilizando plataformas especializadas que permitan emular el comportamiento dinámico y estocástico de las redes eléctricas reales.

1.3 Marco teórico y estado del arte

1.3.1 Marco teórico

Un informe reciente de la Agencia Internacional de la Energía [9] señala que la falta de actualización de las redes eléctricas podría costar a los países emergentes y en desarrollo 1,3 miles de millones de dólares en pérdidas económicas. El informe enfatiza la necesidad de promover la gestión inteligente de recursos, fomentar la eficiencia en la generación y distribución local de energías renovables, alentar la participación activa de los usuarios, fomentar el uso responsable de la energía y garantizar un suministro estable. La transición de la red eléctrica actual hacia una red sostenible, eficiente y flexible requiere investigar las capacidades para tener un sistema que pueda monitorear, aprender y tomar decisiones sobre la gestión de la red. Estos problemas desafiantes, se relacionan con problemas más fundamentales, como maximizar el rendimiento proveniente de la generación distribuida con fuentes renovables, administrar las capacidades de almacenamiento, integrar los vehículos eléctricos y mejorar la capacidad de previsión de demanda y potencia disponible para garantizar en todo momento la estabilidad de la red. No se trata del futuro de la energía; sino del presente, una convergencia de tecnologías que está reescribiendo la narrativa de la distribución de energía y nuestro país tiene los recursos para liderar el cambio.

1.3.2 Estado del arte

Dado el exponencial crecimiento en la conectividad de los sensores y los sistemas basados en el Internet de las cosas (IoT), las smart grids pueden ser capaces de tomar decisiones de manera autónoma, en tiempo real y a partir de enormes conjuntos de datos heterogéneos [10]. La dinámica compleja del sistema resultante es producto de la demanda variable y la incertidumbre en la disponibilidad de energía renovable. En especial, la predicción de los patrones de consumo y energía disponible son tareas condicionadas a múltiples factores internos y externos intervinientes, como el clima, el rendimiento de los sistemas térmicos y los patrones de ocupación [11]. Hasta la fecha, se han desarrollado varios métodos, incluidos modelos físicos, métodos estadísticos, técnicas de IA y sus híbridos para mejorar la precisión del pronóstico de la energía renovable. Más recientemente, las técnicas de aprendizaje profundo, el cual consiste en el uso de redes neuronales artificiales para aproximar funciones, se han mostrado como enfoque un prometedor dentro del aprendizaje automático (subcampo de IA), estas son capaces de descubrir las características no lineales inherentes y las estructuras invariantes de alto nivel en los datos, exhibiendo capacidades comprobadas para aprender a partir de conjuntos de datos heterogéneos [12]. Las

arquitecturas profundas conducen a representaciones con un mayor poder predictivo, permitiendo además descubrir patrones o tendencias en los datos y extraer conocimiento crucial acerca del estado de un sistema, hacer predicciones de los estados futuros y detectar escenarios de comportamiento subóptimo o de falla. Si bien las redes neuronales se han utilizado para obtener resultados del estado del arte, el mayor foco ha sido puesto en el desarrollo para datos estáticos. Pero debido a que la predicción de energía puede verse como un problema de aprendizaje a partir de series de tiempo [13], debemos centrarnos en aquellas herramientas de representación que permitan la caracterización de patrones de comportamiento a partir de datos secuenciales. Dado que los patrones de demanda y energía disponible muestran dependencias temporales a corto y largo plazo, muchos de los modelos de representación conocidos son incapaces de hacer frente a las dependencias entre los estados del comportamiento del sistema.

Otro factor a considerar en el funcionamiento de la smart grid, es el costo de la energía y como este varía según los mecanismos de fijación de precios diseñados en función de la predicción de energía disponible para los siguientes períodos, ya sea de fuentes renovables o de cargas acumuladas. Es entonces necesario que la microrred tenga una potencia equilibrada entre generadores y consumidores [14]. Incluso, la sensibilidad a los costos y los tiempos de reacción de los productores y consumidores de energía tiene un impacto significativo en la estabilidad de las redes eléctricas distribuidas. [15] proponen un modelo llamado control descentralizado de redes inteligentes (DSGC) para hacer el control del lado de la demanda de las redes eléctricas distribuidas, asociando el precio de la electricidad a las variaciones en la frecuencia de la red un margen de tiempo de unos pocos segundos. Trabajos como el de [16] han utilizado el modelo para simular el consumo y/o producción del lado de la demanda de energía e implementan un algoritmo de aprendizaje automático para prever la estabilidad dinámica de la red para distintos comportamientos por parte de los participantes.

Los vehículos eléctricos (EVs, por sus siglas en ingles) serán, sin lugar a duda, participantes relevantes en la operación de la red inteligente, afectando a la estabilidad del sistema de dos maneras diferentes. Por un lado, podrían ayudar a manejar la intermitencia introducida por las fuentes renovables mediante cargas almacenadas en sus baterías y aumentar así la previsibilidad de operación en la microrred. Por otro lado, la adopción generalizada de EVs podría aumentar el riesgo de sobrecarga de la red eléctrica al inflar los picos de demanda [17]. Por lo tanto, la gestión de carga de EVs es relevante para realizar una suavización de la demanda de electricidad, haciendo que la red sea más económica, eficiente y confiable. Sin embargo, la ausencia de estrategias flexibles que reflejen los intereses propios

de los usuarios de EVs puede reducir su participación en este tipo de iniciativas. Por ejemplo, [18] proponen una estrategia de carga inteligente utilizando herramientas de aprendizaje automático para determinar cuándo cargar la batería del vehículo durante las sesiones de conexión. Esto se logra tomando decisiones de carga en tiempo real basadas en varios datos auxiliares, incluidos la conducción, el entorno, los precios y las series de tiempo de demanda, con el fin de minimizar el costo total de energía del vehículo. El compromiso entre envejecimiento y tiempo de carga de la batería juega un papel importante en los sistemas de gestión de baterías, en los que los algoritmos de carga tienen una fuerte influencia en el rendimiento final [19]. Como tal, se han desarrollado una variedad de estrategias de carga, que van desde las más simples, como la tensión constante de corriente constante (CC/CV) [20], hasta las más creativamente complicadas, como el algoritmo de carga de múltiples etapas [21]. Aunque la literatura se basa principalmente en el problema de carga de tiempo mínimo, algunos trabajos informaron estrategias de carga destinadas a aumentar la velocidad de carga mientras se intenta maximizar la vida útil de la batería [22]. Sin embargo, dado que muchos de estos protocolos incluyen heurísticas para encontrar una estrategia de carga, no existen garantías matemáticas para la optimización de la carga rápida y la satisfacción de las restricciones seguras.

En este contexto, un enfoque clásico de toma de decisiones para el problema de gestión inteligente de una microrred incluiría dos componentes: 1) predecir los estados futuros de las variables ambientales y de comportamiento de los participantes de la red (costo de la electricidad, energía disponible, predicción de la demanda, etc.) y 2) tomar decisiones óptimas dadas estas predicciones. Pero la optimización de estas decisiones solo es posible con un conocimiento perfecto del futuro. En la práctica, es probable que el aspecto de la toma de decisiones encuentre dificultades cuando hay variaciones significativas en los valores predichos, al no poder tener en cuenta la incertidumbre de la predicción, por ejemplo, cuando es probable que se obtengan resultados diferentes. En este sentido, el paradigma de RL podría facilitar el desarrollo de metodologías para la toma de decisiones, haciendo uso eficiente de la experimentación y del aprendizaje profundo para determinar una o más políticas de control, que permitan a los sistemas que operan en contextos de incertidumbre y variabilidad, como las redes de energía inteligentes, obtener un suficiente grado de autonomía y adaptación para aproximarse a una operación óptima. En los últimos años, las técnicas de RL han sufrido una revolución derivada de la combinación con métodos de aprendizaje profundo como métodos de aproximación de funciones. Esta combinación, ha resultado de gran utilidad en problemas con espacios de estados de alta dimensión, resolviendo en parte el problema de los enfoques anteriores.

El Aprendizaje por Refuerzos Profundo (DRL, por sus siglas en ingles) ha tenido éxito en tareas complejas, incluso con menor conocimiento previo, gracias a su capacidad para aprender diferentes niveles de abstracciones en los datos. A pesar de esto, aún quedan varios desafíos por atender, ser capaz de generalizar un buen comportamiento en un entorno real por ejemplo no es un caso sencillo. No obstante, las smart grids proporcionan un amplio campo para el testeo de algoritmos de DRL que comprende: generación distribuida para determinar la ubicación y el tamaño óptimo de las redes inteligentes considerando beneficios económicos y ambientales; conexión de EVs bajo enfoques Vehicle-to-Grid (V2G); programación óptima de recursos energéticos en la red inteligente [23]; gestión del lado de la demanda para determinar perfiles de consumo; mercados basados en incentivos para permitir a los usuarios modificar sus patrones de demanda de acuerdo con sus costos de consumo de energía.

En trabajo reciente, se ha investigado la aplicación de RL en diversas áreas de gestión energética, incluyendo la administración de la demanda mediante incentivos [24], el control de sistemas de almacenamiento térmico [25] y la integración eficiente de fuentes de generación renovable [26].

CAPITULO 2: Introducción al Aprendizaje por Refuerzos

2.1 Proceso de Decisión de Márkov y Aprendizaje por Refuerzos

Los Procesos de Decisión de Márkov (MDPs, por sus siglas en inglés) son una formulación matemática clásica del problema de la toma de decisiones secuenciales cuando hay incertidumbre en el sistema. Se utilizan para modelar situaciones en las que un agente toma decisiones que pueden no solo afectar los resultados inmediatos, sino que también las situaciones futuras.

Los MDPs ofrecen un marco general para describir formalmente como un agente puede aprender a interactuar con el entorno para lograr un objetivo, evaluando las consecuencias de sus acciones a lo largo del tiempo, se puede observar la dinámica de un MDP en la Figura Nº 1, donde π_θ es la política utilizada para seleccionar acciones en base a observaciones y $p(s_{t+1}|s_t, a_t)$ denota las probabilidades de transición del estado s_t al s_{t+1} al tomar una acción a_t.

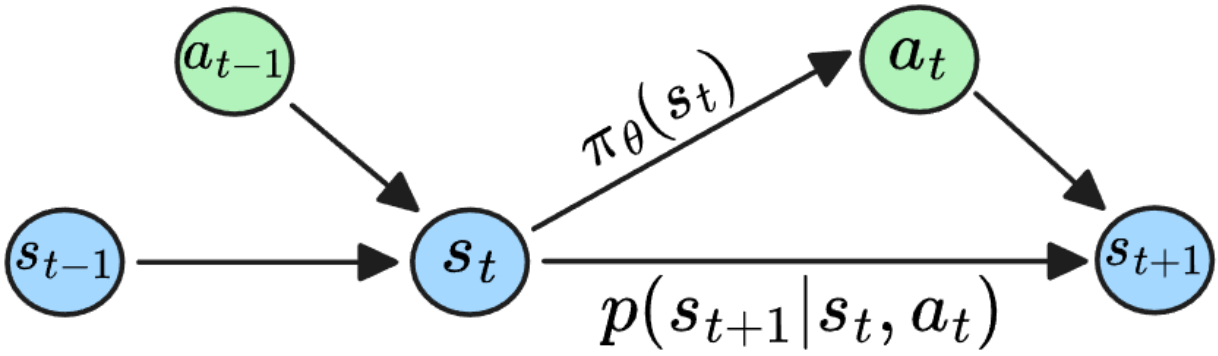


Figura Nº 1: Ilustración de proceso de decisiones de Márkov

La base de un MDP está definida por un conjunto de estados, acciones, probabilidades de transición y recompensas. Este marco que permite formular problemas de toma de decisiones se basa en la aplicación de la propiedad de Márkov, que establece que el estado futuro del sistema solo depende del estado actual y la acción tomada, es decir, sin tener en cuenta el historial de estados y acciones previas, lo que simplifica el modelo, y elimina la necesidad de conocer la secuencia de estados anteriores. Lo que también puede ser interpretado como que cada estado tiene una observabilidad completa.

El funcionamiento de un MDP es el siguiente: el agente comienza en un estado inicial. En cada paso, el agente elige una acción disponible en ese estado. El sistema consecuentemente transiciona a un nuevo estado según una distribución de probabilidad determinada por el estado actual y la acción seleccionada. Al mismo tiempo, el agente recibe una recompensa (o costo) asociada a esa transición. El objetivo del agente es maximizar la recompensa acumulada a lo largo del tiempo, haciendo uso de un factor de descuento que reduce el valor de recompensas futuras (para evitar soluciones de horizonte infinito).

En aplicaciones de ingeniería del mundo real (por ejemplo: control de robots, gestión de redes, sistemas autónomos, etc…), el modelo completo del entorno no siempre es conocido, o puede ser demasiado complejo para ser descrito analíticamente, en estos casos el agente no tiene acceso directo a las reglas del entorno ni a los resultados exactos de sus acciones, sino que debe descubrirlos mediante interacción, como se puede observar en la Figura Nº 2.

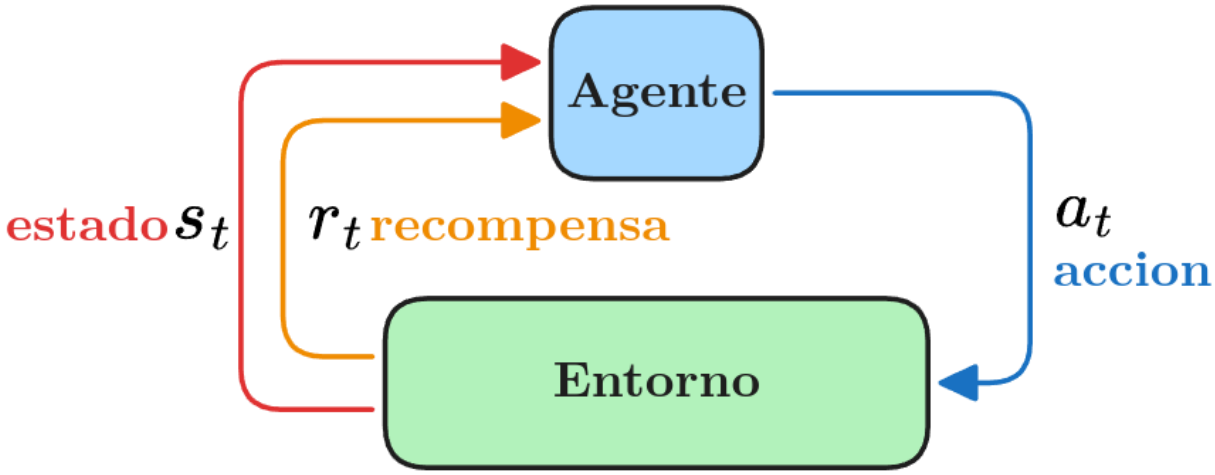


Figura Nº 2: Esquema de interacción entre el agente y el entorno

El RL es un campo dentro del aprendizaje automático que estudia cómo los agentes pueden aprender a tomar decisiones óptimas en entornos estocásticos a partir de su experiencia, sin requerir un modelo del entorno. En este contexto, los MDPs actúan como el marco matemático que describe la estructura del problema, aunque el modelo no sea conocido, se asume que el comportamiento puede ser descrito como un MDP.

El objetivo del Aprendizaje por Refuerzos (RL) es aprender una política óptima $\pi(s)$ la cual selecciona acciones que tomar para maximizar la recompensa esperada acumulada $G_0 = \sum_{t=0}^{T-1} \gamma^t r_t$ donde γ es un factor de descuento que se utiliza para agregar prioridad a soluciones que conduzcan a un mejor estado terminal más rápido.

Una estrategia clásica de RL consiste en aprender funciones de valor, es decir funciones que estiman que tanta recompensa esperada acumulada se espera por estar en un estado, una clase de estas funciones normalmente utilizada es el valor Q_{π}, la función de valor de pares (acción, estado), el cual representa el valor (recompensa esperada acumulada) esperado al tomar una acción a en un estado s y luego seguir una política π.

Para conseguir una política óptima π^*, existe una familia de técnicas conocidas como métodos basados en valor, consisten en aproximar la función $Q^*(s,a)$.

$$Q^*(s,a) = \max_{\pi} E_{\pi}[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s, a_0 = a] \quad (1)$$

Para poder calcular o aproximar esta expectativa en práctica, se explotan dos principios a partir de los cuales se puede obtener la ecuación de Belman de la optimalidad (Ec. 2), basado en el hecho de que la esperanza es un operador lineal, y la propiedad de Márkov que establece que el siguiente estado y su recompensa solo dependen del estado y acciones actuales. Se puede deducir que el valor óptimo para un par estado-acción, es la recompensa inmediata conseguida al tomar dicha acción en el estado dado, más la suma descontada del valor al tomar la mejor acción en el siguiente estado, ponderado por las probabilidades de transición $P(s' \mid s,a)$. Una manera más simple de razonar esta ecuación es el pensar que la función de Q para una acción y un estado, es igual a la recompensa inmediata más el valor descontado del mejor futuro posible.

$$Q^*(s,a) = R(s,a) + \gamma \sum_{s'} P(s' \mid s,a) \max_{a'} Q^*(s',a') \quad (2)$$

Una vez obtenida la función Q, que nos permite evaluar el valor de tomar una acción en un estado, se puede deducir que tomar la política óptima simplemente consiste en tomar la acción que maximice la función de Q en cada estado.

$$\pi^*(s) = \arg\max_{a} Q^*(s,a) \quad (3)$$

Si bien las probabilidades de transición no son conocidas, estas se pueden muestrear mediante interacción con el entorno. Un método clásico para aprender la función Q sin conocimiento del modelo con el que el agente interactúa se conoce como Q-Learning. En este método, se construye una aproximación de la función óptima de Q mediante la ecuación de Bellman de optimalidad (Ec. 2) utilizando muestras, para ello primero se observa una transición compuesta por un estado, una acción, un siguiente estado y una recompensa y luego se puede actualizar iterativamente la estimación utilizando:

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha_t \left[r_t + \gamma \max_{a'} \quad Q_t(s_{t+1}, a') - Q_t(s_t, a_t) \right] \tag{4}$$

Siendo α la tasa de aprendizaje, la cual se usa para regular el tamaño de los pasos a tomar en este método iterativo en dirección a la nueva estimación. El término entre corchetes describe lo que se conoce como diferencia temporal, y representa la señal de error usada para actualizar los valores de Q, la diferencia temporal denota la diferencia entre el valor de Q estimado en el estado actual, y el "target", que es el valor Q obtenido al sumar la recompensa obtenida cuando se toma la acción elegida en el estado actual, y sumarle el valor de Q del estado siguiente asumiendo que se tomara la mejor acción posible, de esta manera denota el error que se tenía en la estimación, y se usa la experiencia de la interacción con el ambiente para mejorar esta estimación. El target también se puede ver como una estimación de una muestra de la Ecuación 2.

En este método iterativo se mantiene registro de los valores de Q para cada acción y estado en una tabla, y si bien esto funciona y tiene garantías de convergencia para espacios de acciones y estados de baja dimensionalidad, a medida que la dimensionalidad aumenta se vuelve imposible almacenar todos los valores en una tabla, y puede que el agente nunca visite el mismo estado y tome la misma acción dos veces. Por lo tanto, se requiere un método que sea capaz de generalizar lo aprendido al tomar una acción en un estado, para utilizarlo en otros estados con nuevas acciones.

2.2 Redes Neuronales

Una red neuronal artificial es un modelo computacional inspirado en el funcionamiento de las neuronas biológicas, donde se hace uso de unidades de procesamiento interconectadas organizadas en capas, para procesar información y aprender relaciones entre los datos de entrada y de salida. La idea básica de una red neuronal [27] es el aprender funciones complejas, a partir de un aproximador universal, cuyos parámetros se ajustan para reducir el error en la aproximación.

Inspiradas por el funcionamiento de neuronas biológicas, las redes neuronales hacen uso de unidades de cómputo (neuronas o perceptrones) las cuales reciben entradas y las transforman en salidas mediante la transformación lineal de las entradas, y una función no lineal aplicada sobre esta transformación, el resultado pasa a la siguiente capa formada por más neuronas.

Al encadenar estas capas de neuronas, la red puede representar relaciones muy complejas entre entradas y salidas.

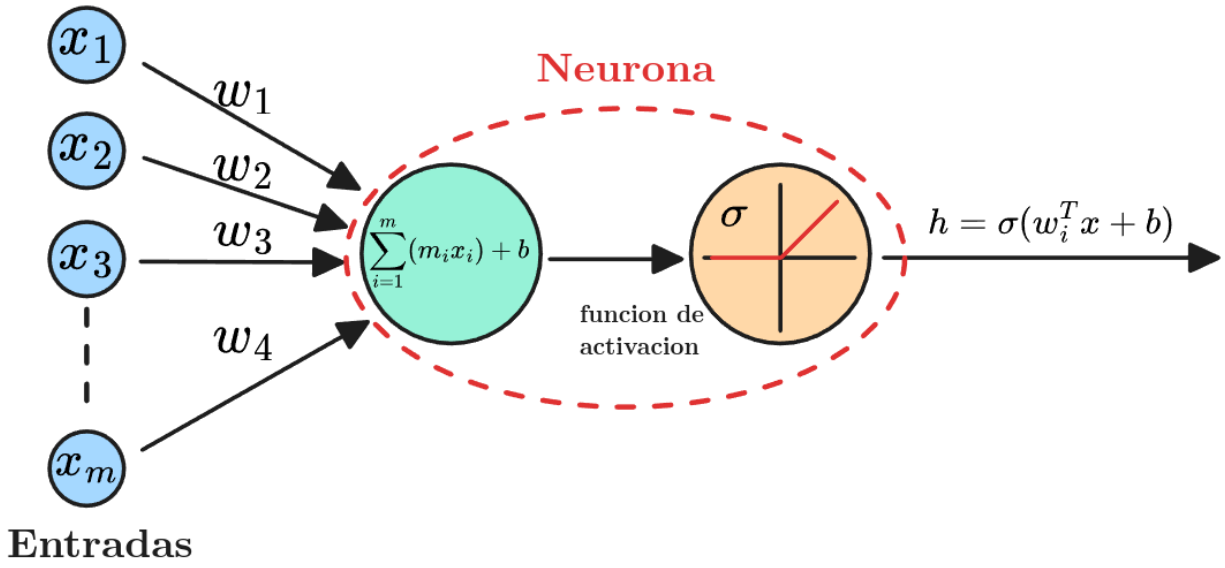


Figura Nº 3: Ilustración mostrando el funcionamiento de una neurona en una red neuronal.

Matemáticamente, cada neurona realiza una operación dada por:

$$h = \sigma(w_i^T x + b_i) \tag{5}$$

Donde x es el vector de entradas, w_i y b_i son los parámetros de la neurona, y definen los pesos aplicados a cada elemento del vector de entrada, y el sesgo a la salida respectivamente. σ es una operación no lineal, normalmente conocida como función de activación (ej: ReLU, Sigmoid, Tanh, …), esto puede ser visualizado en la Figura Nº 3. Para pasar del caso de una única neurona a una capa completa de neuronas para aumentar la capacidad de representación del aproximador de funciones, se agrupan en paralelo todos los vectores de pesos w_i y sesgos b, para formar una matriz W que se utiliza para calcular la combinación lineal de las entradas para cada neurona, a la que luego se aplica una suma del vector de sesgos b, antes de aplicar la operación no lineal de la misma manera que se realiza para una única neurona, por lo que se puede denotar la salida de una capa de neuronas como:

$$h_i = \sigma(Wx + b) \tag{6}$$

A medida que se aplican más capas de neuronas, se obtiene una red más profunda (Fig. 4) y aumenta la capacidad de representación de la red neuronal, permitiendo a la red neuronal extraer de los datos características de más alto nivel, con el costo de agregar parámetros adicionales que deberán ser aprendidos.

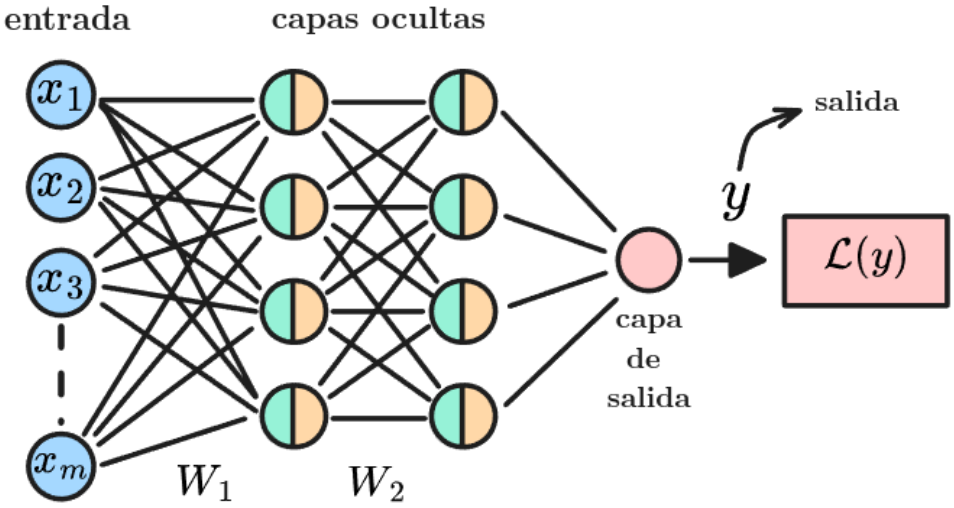


Figura Nº 4: Ilustración de una red neuronal.

Para poder ajustar los parámetros de la red, se define una función de pérdida $\mathcal{L}(\theta)$ la cual mide el error entre la salida de la red neuronal y el valor deseado. A partir de esta función de pérdida, se hace uso de descenso gradiente para ajustar los parámetros de la red neuronal θ de manera que se reduzca la pérdida.

$$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(\theta) \tag{7}$$

Donde η es la tasa de aprendizaje, y $\nabla_\theta \mathcal{L}(\theta)$ se calcula mediante el algoritmo de backpropagation, el cual aplica la regla de la cadena para computar de manera eficiente el gradiente de una red neuronal. En lugar de recomputar el gradiente completo desde cero, backpropagation inicia en la capa de final y calcula el "error local" de cada neurona (cómo cambia la pérdida con su salida) y lo propaga hacia atrás capa a capa ajustando cada peso según su contribución al error final, esto se puede visualizar en la Figura Nº 5.

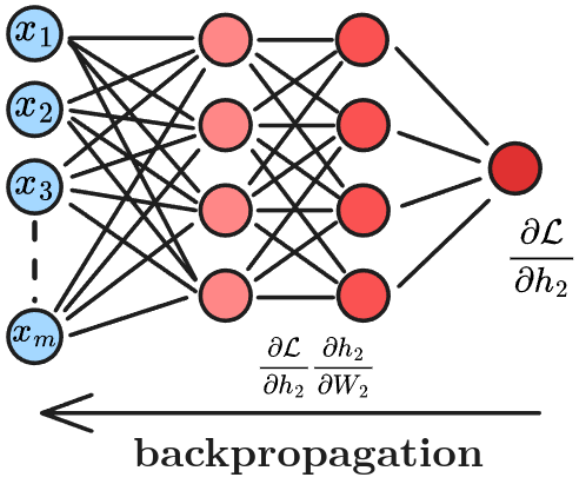


Figura Nº 5: Ilustración del funcionamiento de backpropagation.

De esta manera, las redes neuronales presentan un método de aproximar funciones a partir de datos y una función de pérdida a minimizar.

2.3 Aprendizaje por Refuerzos Profundo

En el contexto de RL, el sustituir la tabla de valores utilizada por métodos como Q-Learning por un parametrizador universal tal como las redes neuronales permite aportar un mayor grado de generalización que permite trabajar con espacios de estados de altas dimensiones, y a generalizar entre estados y acciones no vistos anteriormente. La incorporación de redes neuronales a métodos de RL da lugar al campo de Aprendizaje por Refuerzos Profundo (DRL), donde se aprenden representaciones (aproximadores de funciones) y se ajustan las políticas o valores de manera conjunta.

Los métodos de DRL se pueden clasificar en dos familias principales (Fig. 6). Por un lado, están los métodos basados en valor, donde se estima la función $Q(s,a)$ utilizando redes neuronales, y minimizando la diferencia temporal, mediante una aproximación de la ecuación de optimalidad de Bellman (Ec. 2). Un ejemplo de un método de DRL proveniente de esta familia es el de Deep-Q-Networks (DQN) [28], en el que se hace uso de búferes de repetición que almacenan experiencias para entrenar con una mayor cantidad de datos por cada interacción con el ambiente, y de redes neuronales de objetivo para estabilizar el proceso de entrenamiento.

Por otro lado, existen los métodos basados en políticas, métodos basados en gradientes de política, en los que en lugar de modelar una función de valor, se utilizan redes neuronales para parametrizar directamente la política $\pi(a|s)$. En estos se busca optimizar los parámetros de la red neuronal directamente para maximizar la recompensa esperada, sin la necesidad de calcular un Q o V para todas las acciones en cada paso.

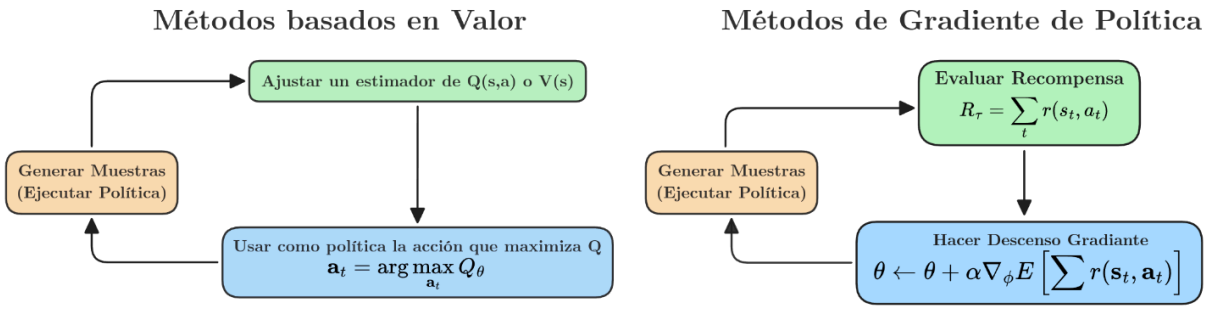


Figura Nº 6: Diagrama simplificado de funcionamiento de métodos basados en valor y métodos basados en gradiente de política

2.3.1 Gradientes de Política

En los métodos basados en política [29], se intenta maximizar la recompensa total esperada al actuar bajo la política π_ϕ, a partir del cual se formula una expresión para su gradiente, el cual se estima mediante muestras, luego será usado en un algoritmo de descenso gradiente para actualizar los parámetros de la red neuronal que representa la política. La idea de estos métodos es que pequeños cambios en los parámetros de la red neuronal provocan pequeños cambios en la forma de la distribución de la salida, por lo que se ajustan estos parámetros suavemente en la dirección que conduce a incrementar la probabilidad de acciones que resultan en mayores recompensas.

La función objetivo se define como:

$$J(\phi) = E_{\tau\sim\pi_\phi}\left[\sum_{t=0}^{T} \gamma^t \, r_t\right] \tag{8}$$

Donde $\tau = (s_0, a_0, r_0, s_1, \dots)$ es una trayectoria generada al seguir la política $\pi_\phi(a|s)$ y se ajustan los parámetros ϕ de la red neuronal para maximizar $J(\phi)$.

Para poder obtener una estimación de la función dada en la ecuación (Ec. 8) sin tener conocimiento del modelo, se estima esta función a través de Monte Carlo, una técnica estadística la cual consiste en la generación de múltiples muestras aleatorias mediante repetidas simulaciones, y en este caso en particular, mediante la ejecución de la política en el entorno, obteniendo trayectorias completas, sobre las cuales se suma para obtener un estimativo del objetivo. Esta estimación nos da una medida del rendimiento de la política, y al repetir múltiples veces este método resulta en una aproximación del objetivo.

$$J(\phi) \approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{T} \gamma^t \, r_{i,t} \tag{9}$$

Pero para poder mejorar la política se necesita la gradiente de esta cantidad, y dado que no se puede diferenciar esta función desconocida, se necesita estimar la gradiente a partir de muestras, lo que se puede realizar a través del teorema de gradientes de política, el cual demuestra que:

$$\nabla_\phi J(\phi) = E_{\tau\sim p_\phi(\tau)}\left[\sum_{t=1}^{T} \nabla_\phi \log \pi_\phi\,(a_t|s_t) G_t\right] \tag{10}$$

donde G_t es la suma descontada de recompensas a lo largo de la trayectoria:

$$G_t = \sum_{k=t}^{T} \gamma^{k-t} r_k \tag{11}$$

Este teorema elimina las dependencias en el modelo para poder estimar el gradiente de la función objetivo desde muestras, estimando como el gradiente del logaritmo de la probabilidad de la acción elegida en cada paso multiplicada por la recompensa obtenida posteriormente.

En la práctica, se reemplaza la esperanza por un promedio de Monte Carlo, produciendo N trayectorias completas y estimando la gradiente a partir de esto.

$$\nabla_\phi J(\phi) \approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{T} \nabla_\phi \log \pi_\phi \left(a_{i,t} \mid s_{i,t} \right) G_{i,t} \tag{12}$$

Finalmente, se actualizan los parámetros en la dirección del estimador de gradiente.

$$\phi \leftarrow \phi + \alpha \nabla_\phi J(\phi) \tag{13}$$

De esta manera para cada muestra se ejecutan dos pasos, primero se evalúa la política mediante la recompensa obtenida en trayectorias completas, y luego se ajusta la política incrementando la probabilidad de acciones que resultaron en recompensas mayores, y disminuyendo la probabilidad de acciones que resultaron en recompensas menores.

Este tipo de método por sí mismos tienen múltiples limitaciones, al depender de experiencias coleccionadas por Monte Carlo, corriendo secuencias completas, las estimaciones de gradiente suelen presentar alta varianza, lo que se traduce en actualizaciones muy ruidosas y un aprendizaje lento. En adición, este tipo de métodos son lo que se conoce como ”on-policy”, es decir, cada pequeña modificación de los parámetros requiere la recolección de nuevas trayectorias, lo que reduce la eficiencia del sistema y desperdicia datos previos. Aunque el uso de baselines (funciones que se restan a la recompensa obtenida, como restar la recompensa promedio), reducen ligeramente la varianza en estos métodos, estos enfoques no aprovechan como el valor de estados individuales influye en los resultados a largo plazo.

2.3.2 Métodos Actor-Críticos

Los métodos Actor-Críticos [30] emergen como una solución para los problemas presentados por los métodos de gradiente de politica, estos métodos en lugar de usar estimadores de Monte Carlo G_t introducidos en la Ecuación 11, para estimar que tan buena es una acción se introduce un crítico, una segunda red neuronal la cual se usa para aproximar funciones de valor, tales como el valor de un estado $V(s)$ o la ventaja de un estado $A(s, a)$. El

actor (o política) en estos algoritmos actualiza sus parámetros utilizando una estimación más refinada del gradiente:

$$\nabla_{\phi} J(\phi) \approx E_{\tau \sim p_{\phi}(\tau)}\left[\nabla_{\phi} \log \pi\left(a_t | s_t; \phi\right) \cdot \hat{A}(s_t, a_t)\right] \tag{14}$$

Donde $\hat{A}(s_t, a_t)$ sirve para cuantificar en cuánto una acción supera el valor promedio de la política en un estado dado. La introducción de esta función de valor disminuye la varianza de los métodos basados en gradientes de política, permitiendo a la política o actor lograr mayor precisión, mientras el crítico refina continuamente sus predicciones de valor reduciendo la diferencia temporal. Se puede observar un diagrama simplificado del funcionamiento de estos métodos en la Figura Nº 7.

Este tipo de métodos combina la flexibilidad de los métodos de gradientes de politica con la estabilidad que aporta la estimación de valor, sin embargo, siguen existiendo riesgos de desplazar la política drásticamente, volviendo inestable el entrenamiento. Otro problema que presentan estos métodos es que el actor y el crítico deben cambiar de manera coordinada, ya que si estos están desalineados el entrenamiento puede volverse inestable.

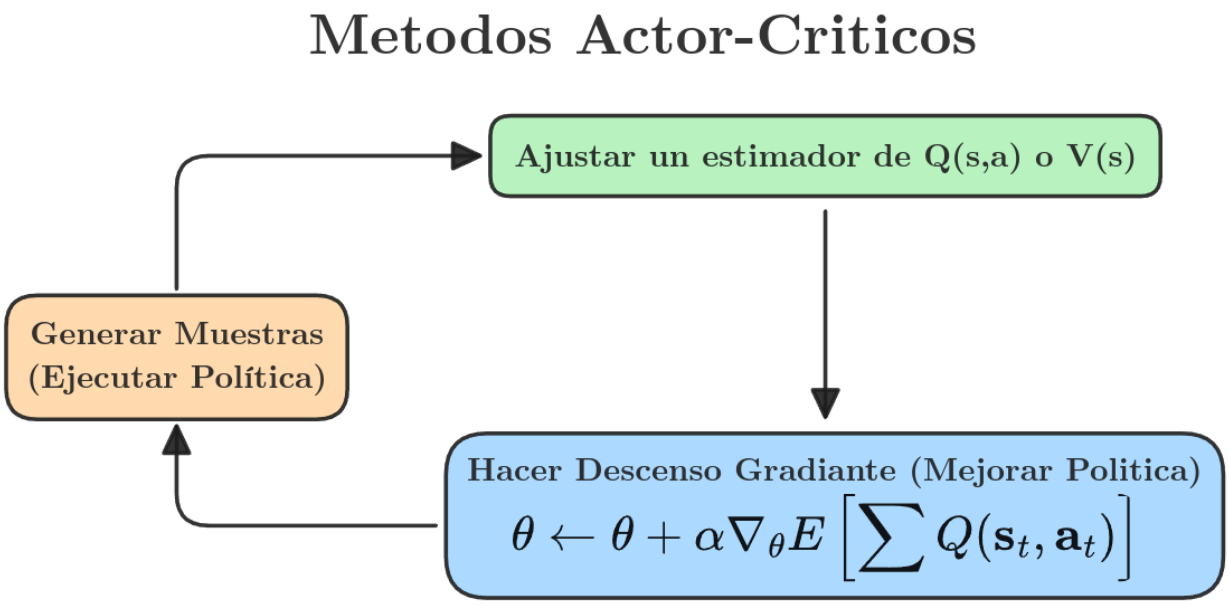


Figura Nº 7: Diagrama simplificado de funcionamiento de los algoritmos Actor-Críticos

CAPITULO 3: Estrategia Basada en Soft Actor-Critic para la Gestión Óptima de la Energía

3.1 Resumen del Capitulo

Las estructuras tradicionales de las redes eléctricas han sido profundamente transformadas por la incorporación de nuevos factores, como las unidades de generación distribuida, la expansión en la capacidad de almacenamiento energético y la implementación gradual de programas de gestión de la demanda. Aunque estos elementos introducen una mayor flexibilidad y abren nuevas oportunidades en la administración de la energía, también plantean desafíos significativos asociados a la creciente complejidad del sistema eléctrico resultante. En este contexto, adquiere relevancia el desarrollo de microrredes inteligentes, concebidas como una respuesta técnica orientada a optimizar el uso de la energía de forma eficiente, confiable y sostenible, promoviendo una visión más dinámica y adaptable del sistema eléctrico [31].

Los principales desafíos que enfrentan las microrredes se derivan de la necesidad de mitigar el impacto de fallas en la red, gestionar la variabilidad nativa de las fuentes renovables, y garantizar la calidad de la energía entregada al usuario final. Frente a este panorama complejo, surgen requerimientos tecnológicos enfocados en el diseño de sistemas de control, monitoreo y comunicación de alto nivel. Al mismo tiempo, aparecen exigencias económicas que impulsan la creación nuevos modelos de negocio que surjan de una nueva forma de concebir el mercado energético. Además, se plantean desafíos que buscan establecer marcos regulatorios capaces de garantizar la interoperabilidad necesaria para que estas soluciones sean viables [32].

A pesar de su diversidad, estos desafíos comparten una serie de características estructurales comunes que permiten abordarlos desde una perspectiva computacional. La gestión eficaz de una microrred exige herramientas capaces de operar en entornos dinámicos, inciertos, variables y complejos. Esto impone un crecimiento en la complejidad del sistema, que requiere experimentar con nuevas arquitecturas que permitan utilizar conocimiento externo o a priori agregar información de características del entorno en el aprendizaje para agilizar los procesos de decisión.

Este capítulo aborda la problemática de operar de manera eficiente los componentes de una microrred eléctrica con fuentes renovables, capacidades de almacenamiento y demanda variable [33]. El mismo se formula como un problema de toma de decisiones secuencial bajo incertidumbre donde, en cada paso de tiempo, la incertidumbre proviene de

la falta de conocimiento sobre el consumo futuro de electricidad y la generación renovable dependiente de las condiciones meteorológicas. En este escenario, la microrred realiza transacciones de compra y venta de energía con la red principal en tiempo real, utilizando precios dinámicos provenientes del mercado eléctrico, por lo que el agente debe adaptarse de forma continua a los precios de mercado vigentes, esta adaptación constituye un paso relevante hacia la integración de sistemas de energía Peer-to-Peer (P2P), donde los participantes intercambian energía directamente de manera descentralizada, ya que el agente debe aprender a optimizar el uso de sus componentes en presencia de un mercado cambiante. El sistema de gestión de la microrred se basa en un algoritmo de control Soft Actor-Critic (SAC), el cual utiliza el concepto de aprendizaje de máxima entropía para combatir la fragilidad de la convergencia común en métodos basados en valor. Para fines de comparación, se implementan estrategias de gestión optimizadas de manera heurística.

La organización del capítulo es como sigue. La sección 3.2 describe el modelo de la microrred utilizada para simulación. En la sección 3.3 se introducen las técnicas de DRL utilizadas y se presenta en detalle el algoritmo SAC utilizado para la gestión de la microrred. En la sección 3.4 se presentan varios experimentos que prueban la capacidad de convergencia del algoritmo SAC y su superioridad frente a técnicas de aprendizaje similares. Por último, en la sección 3.5 se presentan las conclusiones alcanzadas.

3.2 Arquitectura de la Microrred

La microrred opera bajo la coordinación de un agregador o de una empresa de servicios públicos responsable de garantizar el suministro eléctrico para atender la demanda local [34]. A pesar de contar con recursos de generación distribuida basados en turbinas eólicas, la microrred permanece conectada a la red principal, a través de la cual puede comprar o vender energía de forma continua en los mercados eléctricos.

En la Figura Nº 8 se ilustra la arquitectura de este sistema, que integra cinco elementos principales: un Recurso de generación de Energía Distribuido (DER), un sistema de almacenamiento comunitario de energía (ESS, por sus siglas en ingles), un conjunto de cargas controladas termostáticamente (TCLs, por sus siglas en ingles), un grupo de cargas residenciales sujetas a precios dinámicos, y la red de suministro principal. Cada componente mantiene una comunicación bidireccional con el gestor de la microrred, informando sobre los precios de la electricidad, el estado de carga de la batería y la generación de energía.

Figura Nº 8: Ilustración de los componentes del sistema modelado

El agente inteligente del Sistema de Gestión de Energía (EMS, por sus siglas en ingles) procesa la información recibida para emitir señales de control dirigidas a los distintos recursos. En particular, decide el encendido y apagado de los TCLs, regula los procesos de carga y descarga del ESS y determina las operaciones de compra y venta de energía de la red principal.

3.2.1 Agente del Sistema de Gestión de Energia

Para determinar la estrategia óptima y lograr equilibrio entre la oferta y la demanda, el agente EMS debe utilizar la información proporcionada por los componentes de la red para realizar la gestión de la microrred, mediante cuatro mecanismos de control: manejo directo de las unidades TCL (cuando la temperatura se encuentra dentro de un rango permitido), ajuste de precios, gestión de déficit energético y administración de excedentes

En cada intervalo de tiempo t, el agente asigna energía para la operación del clúster de TCLs. A través de un agregador intermedio, se transmiten las órdenes de encendido o apagado a cada unidad en función de su estado de carga (SoC, por sus siglas en ingles). Además, el agregador envía al EMS el SoC promedio del conjunto en tiempo real.

Por otro lado, el agente EMS debe determinar el nivel de tarifa residencial δ_t a utilizar en cada paso de tiempo. Dado que la microrred no monopoliza la demanda, los precios pueden oscilar alrededor de un valor de referencia que mantiene el promedio diario P_{avg} cercano al precio de mercado P_{market} ofrecido por los distribuidores [34]. Estas variaciones incentivan el desplazamiento de la demanda desde los periodos de máxima carga hacia momentos con mayor disponibilidad de potencia.

Frente a situaciones de déficit energético, donde la generación de los DER locales es insuficiente para satisfacer la demanda, la microrred local puede hacer uso de la energía almacenada en el ESS o comprar energía de la red principal. En cada paso de tiempo, el agente EMS establece la prioridad de uso entre estos dos recursos. Por lo que, cuando hay una caída de tensión en la microrred, la energía puede ser suministrada automáticamente desde el recurso prioritario. En caso de que el recurso prioritario sea el ESS y la energía requerida no puede cubrirse en su totalidad, la demanda restante se abastece automáticamente desde la red principal.

Cuando se presenta un excedente de energía generado por los DER, el EMS define anticipadamente cuál es el destino preferente de esta energía: almacenarla en el ESS o venderla a la red principal. Si el ESS ha sido designado como destino prioritario, pero ya ha alcanzado su capacidad máxima, el exceso de energía se desvía automáticamente hacia la red. Esta estrategia permite maximizar la rentabilidad del sistema y evitar desperdicios energéticos, fomentando una operación sostenible y económicamente viable de la microrred.

3.2.2 Almacenamiento de Energía

En lugar de disponer de sistemas de almacenamiento en cada vivienda, la microrred emplea un ESS comunitario con capacidad para satisfacer al menos dos horas de la demanda agregada. En cada paso de tiempo t el estado de carga del ESS es modelado como:

$$B_t = B_{t-1} + \eta_c C_t - \frac{D_t}{\eta_d} \tag{15}$$

Donde $B_t \in [0, B_{max}]$ es la energía almacenada en el ESS en el tiempo t, $\mathrm{B_{max}}$ es la capacidad máxima del ESS y $(\eta_c, \eta_d) \in [0,1]$ son los coeficientes de eficiencia en la carga y descarga. Las variables $\mathrm{C_t} \in [0, \mathrm{C_{max}}]$ y $\mathrm{D_t} \in [0, \mathrm{D_{max}}]$ son las potencias de carga y descarga, las cuales están acotadas por limitaciones en velocidad de carga y descarga del ESS $\mathrm{C_{max}}$ y $\mathrm{D_{max}}$ respectivamente. También se define la variable de estado de carga del ESS como:

$$BSC_t = \frac{B_t}{B_{max}} \quad (16)$$

El comportamiento del ESS en respuesta a las señales de control de carga/descarga está representado por la energía proporcionada y solicitada desde las baterías. En el caso de una señal de carga por parte del EMS, el agente ESS recibe una tasa de energía para el almacenamiento en las baterías, verifica la factibilidad de las operaciones de carga (basadas en la capacidad máxima y la tasa de carga máxima), almacena la energía en cuestión y devuelve la energía restante para ser vendida a la red principal. En el caso del proceso de descarga, el agente ESS recibe una solicitud de energía del EMS, verifica las condiciones de suministro y devuelve la energía disponible. Si el ESS no puede suministrar completamente la potencia solicitada, la diferencia se suministra automáticamente desde la red principal.

3.2.3 Recursos de Energía Distribuida

La generación local corresponde a turbinas eólicas cuyo aporte $\mathrm{G_t}$ fluctúa con las condiciones meteorológicas. En lugar de emplear un modelo sintético, se utilizan datos reales de un parque eólico [35]. En cada instante, el DER comunica al EMS la magnitud de $\mathrm{G_t}$ y vierte directamente esa energía al sistema.

3.2.4 Red de suministro

Debido a la naturaleza intermitente e incontrolable de los recursos energéticos distribuidos, la oferta y la demanda en la microrred no pueden equilibrarse utilizando únicamente estos recursos, por lo tanto, la microrred se mantiene conectada a una red principal que actúa como un ente de regulación. Esta conexión permite que la red principal suministre energía instantáneamente ante déficits o acepte excedentes energéticos en la microrred.

Las transacciones entre la red principal y la microrred se realizan en tiempo real utilizando precios reales al alza y a la baja del mercado de regulación [36], representados como $\left(P_t^u, P_t^d\right)$. Para definir la fuente de suministro prioritaria ante una deficiencia y la fuente de descarga prioritaria ante un exceso, el EMS controla específicamente el interruptor eléctrico hacia la red principal. Después de cada paso de tiempo, el EMS recibe información sobre la energía E_t comprada o vendida a esta red principal, donde los valores positivos indican energía comprada y los negativos energía vendida.

3.2.5 Cargas controladas termostáticamente

Un clúster de TCLs ofrece una fuente relevante de flexibilidad por su capacidad de conservación térmica de energía. Partimos de la premisa de que la mayoría de los hogares en la microrred contaban con un TCL (aire acondicionado, bomba de calor, calentador de agua o refrigerador). En cada instante temporal t, los TCLs reciben señales de control directo desde el agregador del clúster. Para preservar los niveles de confort el controlador de respaldo recibe la acción de encendido/apagado del agregador, valida las restricciones térmicas y ajusta la acción conforme a la siguiente ecuación:

$$a_{b,t}^{i} = \begin{cases} 0 \; si \; T_t^i > T_{max}^i \\ a_t^i \; si \; T_{min}^i < T_t^i < T_{max}^i \\ 1 \; si \; T_t^i < T_{min}^i \end{cases} \tag{17}$$

Donde $a_{b,t}^{i}$ es la acción final de encendido/apagado tras la intervención del controlador de respaldo, T_t^i denota la temperatura operativa del TCL i en t, mientras T_{max}^i y T_{min}^i son los limites térmicos superior e inferior definidos por el usuario. La evolución térmica de cada TCL está dada por:

$$\dot{T_t^i} = \frac{1}{C_a^i}\left(T_t^0 - T_t^i\right) + \frac{1}{C_m^i}\left(T_{m,t}^i - T_t^i\right) + L_{tcl}^i u_{b,t}^i + q^i \tag{18}$$

$$\dot{T_t^l} = \frac{1}{C_m^i}\left(T_t^i - T_{m,t}^i\right) \tag{19}$$

Siendo T_t^i la temperatura interior del aire medida, $T_{m,t}^i$ la temperatura inobservable de la masa del edificio, y T_t^0 la temperatura exterior. Los términos C_a^i y C_m^i representan las masas térmicas del aire y los materiales constructivos respectivamente, q_i es la calefacción interna del edificio, y L_{tcl}^i la potencia nominal del TCL. Adicionalmente, cada TCL posee un estado de carga SoC_t^i que cuantifica la posición relativa de T_t^i dentro del rango térmico deseado.

$$SoC_t^i = \frac{T_t^i - T_{min}^i}{T_{max}^i - T_{min}^i} \tag{20}$$

3.2.6 Cargas Eléctricas

Las cargas residenciales representan la demanda de electricidad de los hogares en la microrred que no se pueden controlar de manera directa. Estas cargas siguen un patrón diario con un componente variable que puede verse afectado por los precios de la electricidad.

Cada hogar i se caracteriza por dos parámetros. El parámetro de sensibilidad $\beta_i \in [0,1]$ es el porcentaje de carga que se puede aumentar o disminuir ante una disminución o aumento respectivamente del precio. El parámetro de paciencia λ_i es el número de horas en las que se compensan los ajustes temporales realizados sobre las cargas. La carga eléctrica L_t^i del hogar i en el tiempo t se modela como:

$$L_t^i = L_{b,t} - SL_t^i + PB_t^i \tag{21}$$

$$SL_t^i = L_{b,t} * \beta_i * \delta_t \tag{22}$$

Donde $L_{b,t} > 0$ indica la carga básica que sigue un patrón de consumo diario [37], SL_t^i es la carga desplazada con δ_t igual al nivel de precios en el tiempo t. Por lo tanto, SL_t^i es positivo para precios altos $\delta_t > 0$ y negativo para precios bajos $\delta_t < 0$. PB_t^i corresponde a las cargas transferidas de períodos de tiempo anteriores para ser reembolsadas. Las cargas desplazadas positivas de una determinada hora deben ejecutarse después de un cierto número de horas, y las cargas desplazadas negativas se retendrán en pasos de tiempo próximos, debido a que se ejecutaron con antelación.

3.3 Algoritmo SAC

Los algoritmos actor-críticos como fueron introducidos en la Sección 2.5 presentan una solución eficiente al problema de DRL, permitiendo soluciones eficientes en términos de muestras. Sin embargo, en su forma clásica, estos algoritmos presentan limitaciones en términos de exploración y estabilidad durante el entrenamiento. El algoritmo Soft Actor-Critic (SAC) [38] surge como una extensión que incorpora principios de entropía para combatir estas limitaciones.

En el campo de DRL, el concepto de entropía se utiliza como en el contexto de teoría de la información, es una medida de incertidumbre en una variable aleatoria, y en este caso se calcula como:

$$\mathcal{H}[\pi_\phi](s) = -\sum_{a \in \mathcal{A}} \pi_\phi(a|s) \log \pi_\phi(a|s) \tag{23}$$

Intuitivamente, este término cuantifica la incertidumbre o diversidad de comportamiento bajo la política. Una política con alta entropía asigna probabilidad de forma más distribuida entre las posibles acciones, mientras que una política con menor entropía toma acciones de una manera más determinista. La incorporación de la entropía en la función objetivo incentiva la exploración activa, evitando la convergencia temprana a una mala

política, ya que el agente busca mantener una mayor amplitud en la variación de acciones que toma mientras trata de maximizar el retorno recibido en interacciones con el entorno.

Soft Actor-Critic modifica la función objetivo de la política, para que, en vez de solo optimizar el retorno esperado, se maximice una versión suavizada de la recompensa acumulada que incorpora la entropía como parte del objetivo.

$$\pi^* = \arg\max_{\pi} E_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t\left(R(s_t,a_t) + \alpha\mathcal{H}\big(\pi(\cdot\,|s_t)\big)\right)\right] \tag{24}$$

Donde $\mathcal{H}\big(\pi(\cdot\,|s)\big) = -E_{a\sim\pi}[\log\pi\,(a|s)]$ representa la entropía de la política, y $\alpha > 0$ es un hiperparámetro que regula la prioridad dada a la entropía. Esta formulación incentiva al agente a mantener variedad en sus acciones durante el proceso de aprendizaje, por lo que la política tiende a ser más robusta y generalizable.

Para incorporar este cambio, las funciones de valor en SAC, se redefinen para incorporar la entropía en su estimación de valor.

$$Q^{\pi}(s,a) = E_{s'\sim P}\big[R(s,a,s') + \gamma E_{a'\sim\pi}[Q^{\pi}(s',a') - \alpha\log\pi\,(a'|s')]\big] \tag{25}$$

Donde el término incluyendo al logaritmo de la política, $\alpha\log\pi\,(a'|s')$, penaliza las políticas que llevan a tomar acciones más determinísticas, creando un equilibrio entre maximizar recompensa y mantener políticas flexibles donde se puede tomar un mayor rango de acciones.

Además de la regularización por entropía, SAC incorpora otras innovaciones importantes para mejorar la estabilidad del algoritmo durante el entrenamiento. Primero, para contrarrestar el sesgo por sobreestimación común en métodos basados en el aprendizaje de funciones Q, se entrenan dos redes neuronales $Q_1(s,a)$ y $Q_2(s,a)$ de forma independiente. Y al momento de actualizar los "targets" usados en la actualización basada en la diferencia temporal, se hace uso del valor mínimo de ambas redes, para evitar un estimador sobre-optimista.

$$\hat{Q}_{target} = R(s,a) + \gamma\left(\min_i Q_i\,(s',a') - \alpha\log\pi\,(a'|s')\right) \tag{26}$$

Para poder optimizar las políticas estocásticas, estas deben ser diferenciables, pero procesos de sampleo no son diferenciables, para resolver este problema se reparametriza la acción como una función determinística de ruido gaussiano, para de esta manera poder optimizar la función a la que se somete el ruido.

$$a = \tanh\left(\mu_\phi(s) + \sigma_\phi(s) \odot \xi\right), \quad \xi \sim \mathcal{N}(0, I) \tag{27}$$

Esto reduce la varianza de los gradientes de política y permite entrenar directamente mediante gradientes sobre la función objetivo, donde μ_ϕ y σ_ϕ provienen de la red neuronal que actúa como la política.

Adicionalmente SAC utiliza un buffer de repetición para actualizar las redes utilizando muestras pasadas, lo que mejora la eficiencia de datos del método.

3.3.1 SAC con Acciones Discretas

Para ajustar el algoritmo a un espacio de acciones discreto, como el provisto por el entorno en el cual se controlan las acciones del agente EMS, se incorporaron las modificaciones introducidas por [39]. Donde la principal diferencia consiste en que la política $\pi_\psi(a|s)$ ahora representa una distribución de probabilidad discreta en lugar de una densidad. Por lo tanto, las funciones objetivo de SAC conservan su forma funcional, pero deben adaptarse al contexto discreto mediante 5 modificaciones principales.

Primero, la reformulación del valor-Q discreto, ya que ahora es más eficiente parametrizar el crítico para que devuelva valores Q de todas las acciones posibles en un solo paso. Por lo tanto, la función $Q: \mathcal{S} \times \mathcal{A} \to R$ se reinterpreta como $Q: \mathcal{S} \to R^{|\mathbb{A}|}$.

Segundo, dada la salida discreta de la política, ésta ya no necesita parametrizar la media y varianza de una distribución continua. En vez de esto, se predice directamente una distribución discreta sobre las acciones $\pi: \mathcal{S} \to [0,1]^{|\mathcal{A}|}$.

Tercero, al ser finito el conjunto de acciones, se puede calcular de forma exacta la expectativa que define la función de valor suave utilizada para estimar que tan bueno es un estado además de la función objetivo del gradiente de política. Ya no es necesario estimar con Monte Carlo.

$$V(s_t) = \sum_{a_t} \pi(a_t|s_t)[Q(s_t, a_t) - \alpha \log \pi(a_t|s_t)] = \pi(s_t)^\top [Q(s_t) - \alpha \log \pi(s_t)] \tag{28}$$

$$J_\pi(\phi) = E_{s_t \sim \mathbb{D}}\left[\pi(s_t)^\top [\alpha \log \pi(s_t) - Q_\theta(s_t)]\right] \tag{29}$$

Cuarto, se realiza una optimización sobre el coeficiente de importancia de la entropía en la función de costo, donde se optimiza el coeficiente α de manera que se minimice la diferencia entre la entropía y una entropía objetivo $\overline{\mathcal{H}}$, ponderada por la política.

$$J(\alpha) = E_{s_t \sim \mathbb{D}}\left[\pi(s_t)^{\top}[-\alpha(\log \pi(s_t) + \bar{\mathcal{H}})]\right] \quad (30)$$

Sin embargo, con estas modificaciones el entrenamiento del algoritmo se vuelve inestable, por lo que fue necesario incorporar las modificaciones hechas por [40] para estabilizar el entrenamiento de la versión discreta de SAC.

El primer cambio introducido, se debe a que en el caso del algoritmo SAC discreto, ya no se presenta el problema de sobre estimación común en el ámbito continuo, por lo tanto, al usar el mínimo de dos valores Q para actualizar el target utilizado para calcular la diferencia temporal, el target termina siendo pesimista, por lo tanto, el primer cambio consiste en utilizar el promedio de ambos valores estimados de Q en lugar del mínimo.

$$y = r + \gamma \sum_{a'} \pi(a'|s') \left[\frac{1}{2}\left(Q_{\theta_1}(s', a') + Q_{\theta_2}(s', a')\right) - \alpha \log \pi(a'|s')\right] \quad (31)$$

$$J(\theta) = E[(Q_i(s, a) - y)^2] \quad (32)$$

El segundo cambio se debe a que, en el algoritmo SAC discreto original, la entropía de la política afecta al valor objetivo del crítico. Por lo tanto, si la política sufre un cambio que cause que tienda a tener mayor confianza en una acción particular (como es común en etapas tempranas del entrenamiento), esto produce una variación abrupta en la entropía, lo que a su vez genera un cambio brusco en el target usado para estimar la diferencia temporal. Dado que la red intenta minimizar esta diferencia, estos cambios hacen que la red persiga un objetivo móvil, desestabilizando el entrenamiento. Para mitigar este problema, se introduce un término de penalización por variación de entropía en el objetivo de la política. Esto permite que la política se vuelva más determinista de manera controlada, estabilizando el proceso de entrenamiento:

$$J_{\pi}(\phi) = E[\alpha \log\left(\pi_{\phi}(a_t|s_t)\right) - Q_{\theta}(a_t|s_t)]] \; + \; 0.5\, \beta(\mathcal{H}_{old} - \mathcal{H}_{\pi})^2 \quad (33)$$

3.4 Experimentos

Para construir el entorno de microrred descrito en la sección 3.2 y poner a prueba el desempeño del algoritmo SAC, se utilizó la plataforma OpenAI Gym [41]. La simulación se desarrolla a lo largo de varios días, considerando cada uno como un episodio independiente. Dado que el paso de tiempo está discretizado por horas, cada día equivale a 24 pasos. Al inicio de cada episodio, se selecciona aleatoriamente uno de los diez días disponibles en el conjunto de datos.

El estado de la microrred en un instante dado está compuesto por un conjunto de variables clave: el estado promedio de carga de los dispositivos de carga térmica (TCL), el nivel de energía almacenada en el sistema de baterías, un contador de precios, la temperatura ambiente, la cantidad de energía renovable generada, los precios del mercado de regulación, la hora del día, y el consumo actual según el patrón diario de carga. Todos estos elementos conforman un vector que describe de forma compacta la situación actual del sistema s_t.

El siguiente paso clave en el diseño del entorno es la formulación de la función de recompensa. Esta tiene como objetivo maximizar el beneficio económico neto, calculado como la diferencia entre ingresos y costos operativos en cada instante:

$$R_t = \text{Ingresos}_t - \text{Costos}_t \tag{34}$$

Los ingresos se obtienen por tres vías principales: la venta de energía a los usuarios, el cobro por consumo controlado de los TCL, y la exportación de energía a la red externa. Este componente se define como:

$$\text{Ingresos}_t = P_t \sum_{cargas} L_t^i - C_{gen} \sum_{TCLs} L_{TCL}^i u_{b,t}^i + P_t^b E_t^V \tag{35}$$

Donde P_t representa el precio de venta al consumidor, C_{gen} es el costo unitario de generación, y $u_{b,t}^i$ indica la acción de encendido de cada TCL. La última parte corresponde al ingreso por energía vendida a la red.

Los costos, en cambio, incluyen tanto la energía comprada al mercado como los cargos asociados al transporte de electricidad:

$$\text{Costos}_t = \left(P_t^a + C_{tr_{imp}}\right) E_t^C + C_{tr_{exp}} E_t^V \tag{36}$$

Donde E_t^C y E_t^V representa la energía generada por fuentes renovables vendida a la red externa y la energía comprada de la red externa respectivamente. Por último, $C_{tr_{imp}}$ y $C_{tr_{exp}}$ son los costos asociados con la transmisión de energía para la importación y exportación a la red externa, respectivamente.

Para evaluar la política aprendida por SAC, se la comparó contra dos estrategias de referencia, i) un controlador óptimo teórico basado en un algoritmo genético con información perfecta de producción, consumo, precios y temperaturas; ii) un minorista (retailer) teórico que compra la cantidad exacta de electricidad en el mercado diario y la vende a la misma base de clientes en nuestra simulación al precio de mercado. También se propusieron dos algoritmos

DRL, deep Q-Network (DQN) [42] y SARSA [43], con el fin de comparar el rendimiento contra otros enfoques de aprendizaje profundo.

Los experimentos se llevaron a cabo en un intervalo de 10 días consecutivos (días 50 al 59 del conjunto de datos). Durante este periodo, se recopilaron recompensas diarias promediadas, cuyos resultados globales se encuentran presentados en Figura Nº 9, donde se puede observar que SAC obtiene una ganancia media superior a la de los algoritmos con la que se los compara, incluyendo la estrategia del minorista teórico.

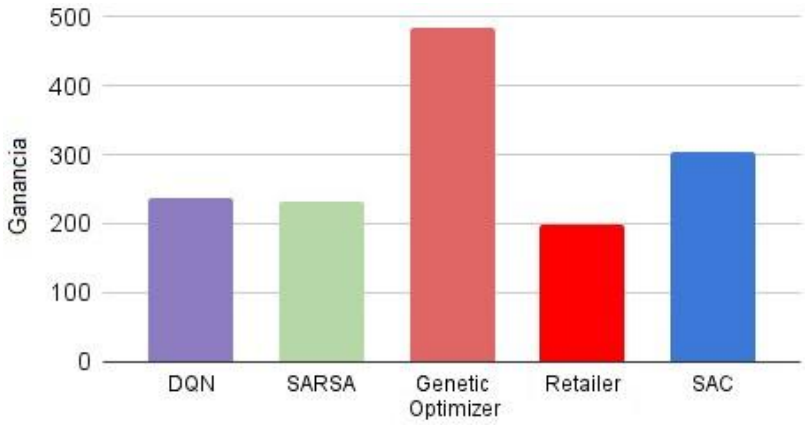


Figura Nº 9: Comparación de ganancia total acumulada por los algoritmos de DRL y el proveedor optimo

Se puede observar en la Figura Nº 10Figura Nº 10 el rendimiento diario, mostrando que el enfoque basado en SAC presenta ventajas competitivas en varias jornadas, incluso bajo condiciones de alta variabilidad en generación y precios. Si bien las estrategias óptima y minorista alcanzan mayores beneficios, debe tenerse presente que ambas operan con conocimiento perfecto, algo irrealizable en entornos reales.

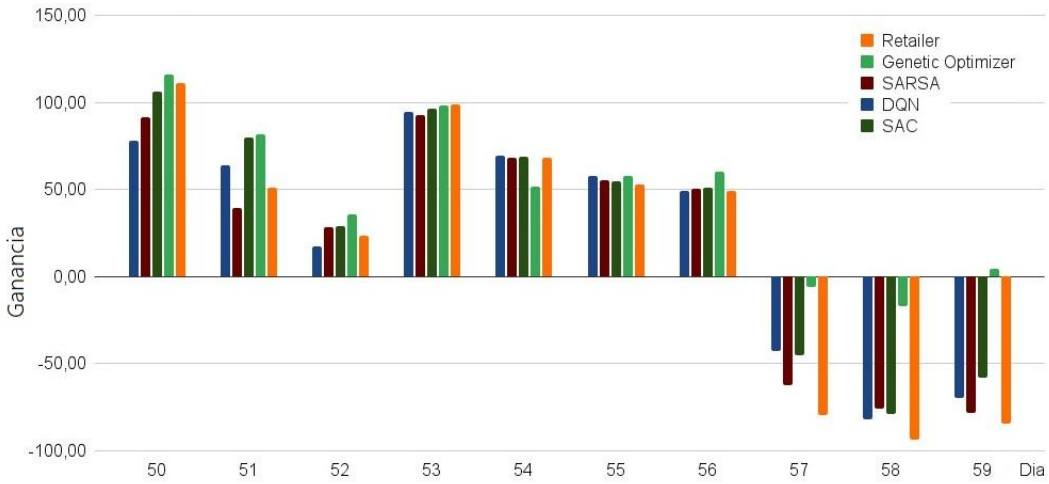
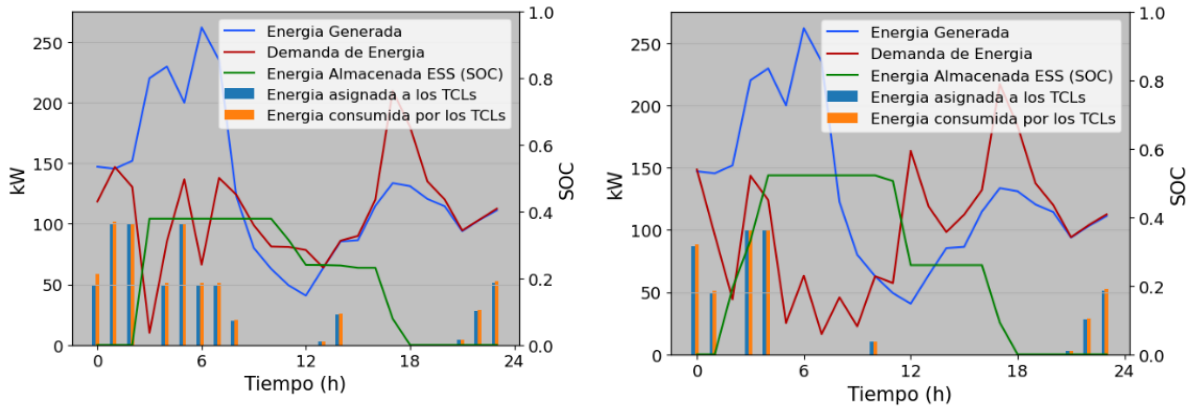


Figura Nº 10: Ganancia diaria obtenida por los algoritmos DRL y el proveedor optimo

Los resultados referentes a la asignación energética entre los TCL y el ESS se muestran en las gráficas mostradas por la Figura 11. Para el día 50, se evidencia un comportamiento similar entre SAC (Fig. 11a) y el controlador óptimo (Fig. 11b). No obstante, el algoritmo SAC tiende a favorecer el almacenamiento en TCLs, mientras que la estrategia óptima distribuye más energía hacia el ESS. Esta diferencia puede atribuirse a la capacidad del controlador óptimo de anticiparse a la evolución completa del entorno.

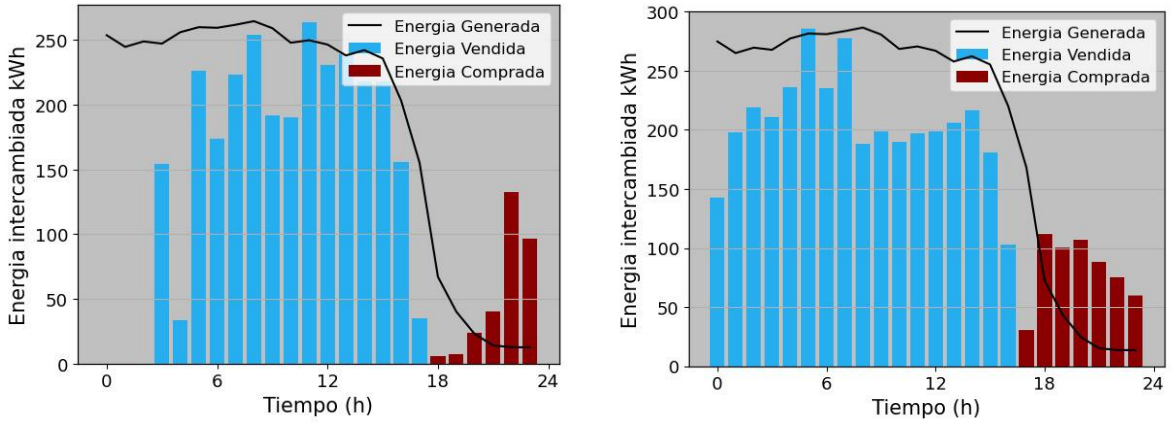


(a) Estrategia SAC

(b) Estrategia Optima

Figura Nº 11: Cantidad de energía generada y almacenada

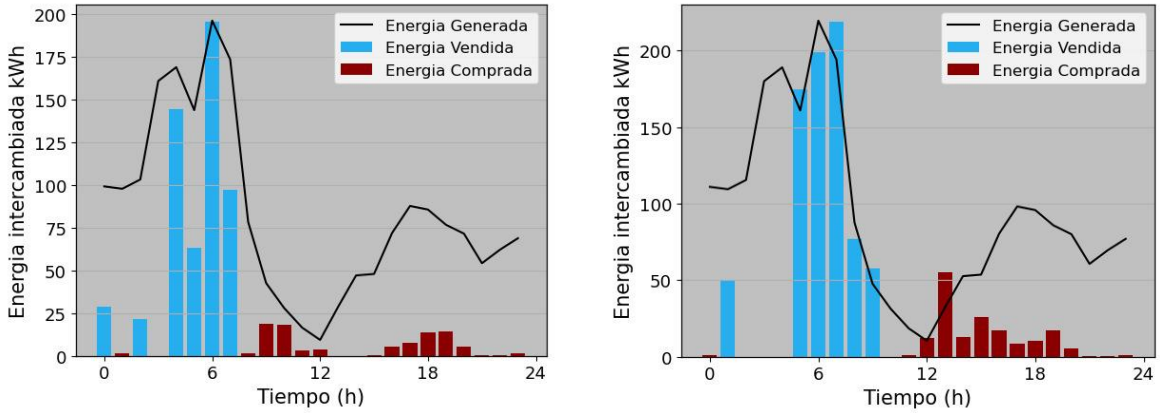
En la Figura Nº 12Figura Nº 12 y la Figura Nº 13 Figura Nº 13 se muestran las curvas de intercambio energético con la red para los días 50 y 56, tanto para el controlador óptimo como para el algoritmo SAC. Las gráficas incluyen la energía comprada, la energía vendida, y la generación renovable disponible. A pesar de las diferencias en disponibilidad y precios entre días, ambos métodos exhiben decisiones similares: exportan excedentes cuando la demanda es baja y realizan compras ante picos de consumo.



(a) Estrategia Optima

(b) Estrategia SAC

Figura Nº 12: Energía intercambiada con la red (Dia 56).



(a) Estrategia Optima

(b) Estrategia SAC

Figura Nº 13: Energía intercambiada con la red (Dia 50).

3.5 Conclusiones

En este capítulo se abordó el diseño e implementación de estrategias de gestión energética para una microrred con generación renovable, almacenamiento y cargas variables, operando bajo precios dinámicos del mercado eléctrico.

Los resultados obtenidos demuestran que el algoritmo SAC es capaz de aprender políticas de gestión que se aproximan al comportamiento óptimo teórico, superando en rentabilidad a otras técnicas de aprendizaje profundo. Si bien el algoritmo óptimo con conocimiento perfecto mantiene un rendimiento superior debido a su acceso a información completa de la dinámica del sistema y precios futuros, el algoritmo SAC logra capturar de manera efectiva los patrones de generación, demanda y precios, permitiendo decisiones de almacenamiento e intercambio energético con la red externa de manera anticipada y eficiente.

El análisis de las curvas de energía comprada y vendida, junto con el comportamiento del almacenamiento en TCLs y ESS, muestra que el algoritmo SAC desarrolla políticas que priorizan la utilización de energía renovable disponible, vendiendo excedentes en momentos de baja demanda y comprando energía en momentos de alta demanda, de forma similar a la estrategia óptima.

CAPITULO 4: Fijación de Precios de Energía en Sistemas Energéticos P2P Usando Aprendizaje por Refuerzo

4.1 Resumen del Capitulo

La adopción masiva de generación renovable a pequeña escala, como los sistemas fotovoltaicos instalados en hogares, ha impulsado el surgimiento de prosumidores, los cuales son individuos que producen, consumen y comercializan energía localmente. Esto ha impulsado la evolución de las redes eléctricas hacia microrredes comunitarias donde el intercambio de energía se realiza de manera Peer-to-Peer (P2P), permitiendo transacciones entre prosumidores y consumidores con menor dependencia de la red convencional [44].

En mercados P2P, los agentes negocian energía mediante plataformas digitales que permiten la fijación de precios dinámicos, lo que favorece la eficiencia energética, mejora la resiliencia local y fomenta la autonomía de los usuarios. Sin embargo, esta descentralización conlleva una serie de desafíos: la intermitencia de las fuentes renovables introduce incertidumbre en la oferta, las cargas varían tanto en espacio como en tiempo y la heterogeneidad de participantes implica intereses a menudo contrapuestos. Además, las estructuras regulatorias aún no se han adaptado completamente a estos nuevos esquemas, lo que intensifica la complejidad del diseño de mecanismos de precio adecuados.

Para poder solucionar este desafío, un área critica es la fijación de precios dinámicos que cumpla múltiples objetivos: garantizar el balance entre oferta y demanda, estimular la participación de prosumidores con incentivos adecuados, y preservar la sostenibilidad financiera del proveedor de servicios energéticos. Tradicionalmente, la estrategia de puesta de precios se ha basado en esquemas estáticos o en tasas reguladas, pero un entorno P2P requiere un enfoque más adaptativo y autorregulado.

En este capítulo, el enfoque está puesto en la superación de este desafío mediante el uso de DRL. Se diseño un entorno de simulación de microrred que incluye prosumidores, consumidores, una batería comunitaria, un proveedor y la conexión a la red tradicional. El objetivo es entrenar un agente DRL que aprenda a fijar precios en tiempo real, maximizando indicadores como eficiencia económica, equidad y estabilidad del sistema.

El entorno de simulación utilizado se basa en la implementación de [45] donde utilizaron las herramientas de la librería Python-Microgrid [46], la cual introduce módulos que permiten realizar simulaciones basadas en datos reales de microrredes, sin embargo, si bien

el entorno propuesto demostró ser utilizable, el entrenamiento sin normalización de costos resultó en gradientes inestables, lo que resulto en políticas simplistas, y restringió la exploración de acciones en el entorno. Además, este entorno no contempla la presencia de vehículos eléctrico, los cuales introducen patrones de demanda móviles y flexibles.

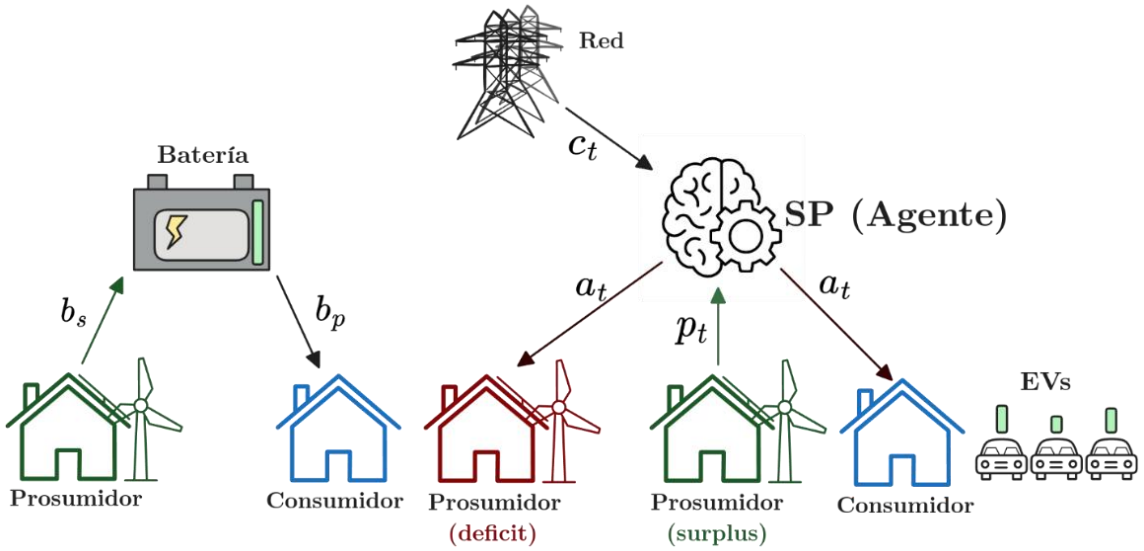


Figura Nº 14: Diagrama ilustrativo del funcionamiento del sistema

La contribución presentada por este trabajo al modelado consiste en dos aspectos clave. Primero, la introducción de normalización en la función de recompensa, lo que estabiliza el aprendizaje y facilita la convergencia. Segundo, la integración de EVs como cargas dinámicas predecibles.

Finalmente, en lugar de Q-learning, se hizo uso de Proximal Policy Optimization (PPO) [47], un algoritmo “on-policy” que ofrece mayor robustez y eficiencia en entornos con alta variabilidad y dimensionalidad.

En las siguientes secciones se presentan en detalle: El modelado de la red eléctrica y sus componentes principales (Sección 4.2), el funcionamiento del algoritmo PPO y cómo se adapta al problema de fijación dinámica de precios (Sección 4.3) y Finalmente, en la Sección 4.4 se presentan los resultados obtenidos

4.2 Modelado de la red eléctrica

En el modelo de simulación utilizado, se hace uso de una microrred compuesta por un proveedor de servicios (SP), un conjunto de prosumidores (P), un conjunto de consumidores (C), una batería comunitaria y se suma la opción de incorporar un conjunto de

EVs. Consideramos una microrred dinámica en el tiempo, el SP se encarga de la determinación de precios, en cada momento t determinando un precio de venta de energía al por menor $a_t \in R^+$, el cual afecta tanto a los prosumidores como a los consumidores, y un precio de compra de energía de los prosumidores $p_t \in R^+$. El SP también puede cubrir la demanda energética comprando energía de la red eléctrica principal a un precio fijo. Un diagrama ilustrativo del sistema puede ser observado en la Figura Nº 14.

Los prosumidores tienen la opción de almacenar su energía excedente en la batería, la cual mantiene precios fijos de compra y venta de energía, o vender su excedente energético al SP por el precio p_t. Las decisiones sobre el uso de la batería se determinan en base a comparaciones de precios entre los precios dinámicos del SP y el precio estático de la batería, lo que introduce un criterio interno de arbitraje energético.

Los EVs, cuando están presentes, se asignan aleatoriamente a algunos participantes en la proporción establecida como parámetro en la simulación, y su impacto en la microrred depende de sus patrones temporales de conexión, sus niveles de carga y sus prioridades de recarga.

Para capturar las variaciones periódicas en la demanda y generación energética, especialmente las asociadas a ciclos diarios, se incorpora características cíclicas para representar el tiempo de una manera cíclica, haciendo uso de funciones armónicas del tiempo: $sin(2\pi h_t/24)$ y $cos(2\pi h_t/24)$, donde h_t representa la hora del día. Esta representación constante permite modelar de manera eficiente la estacionalidad diaria sin introducir las discontinuidades que introduce el uso directo de la hora.

4.2.1 Modelado de la Batería

Se incorpora un sistema de almacenamiento energético compartido, en forma de una batería, que permite gestionar localmente el balance energético dentro de la microrred. Este sistema puede ser cargado utilizando el excedente proveniente de los prosumidores y descargado para abastecer la demanda energética de la comunidad, incluyendo tanto consumidores como prosumidores.

La evolución temporal del SoC de la batería, está sujeta a dos factores principales: la eficiencia energética del proceso de carga P_{BC}^t y descarga P_{BD}^t, y la capacidad nominal de almacenamiento. El SoC se actualiza en cada instante de tiempo en función de la energía neta transferida y considerando una eficiencia bidireccional constante η que refleja las pérdidas internas del sistema. La batería opera dentro de un rango seguro delimitado por

límites mínimos y máximos del SoC, los cuales garantizan la estabilidad operativa y previenen la degradación temprana del sistema.

$$SoC^t = SoC^{t-1} + \frac{(P_{BC}^t * \eta) - (P_{BD}^t / \eta)}{\Lambda} \quad (37)$$

Adicionalmente, existen restricciones sobre las tasas máximas de carga y descarga. Estas limitaciones definen los máximos admisibles de potencia en cada dirección durante un intervalo temporal dado. La capacidad efectiva de carga o descarga disponible en cada instante determinado depende tanto del SoC actual como de estos límites físicos, lo cual introduce una dinámica no lineal en la disponibilidad energética almacenada. Esta se modela de la siguiente manera:

$$SoC_{min} \leq SoC_t \leq SoC_{max} \quad (38)$$

$$0 \leq P_{BC}^t \leq P_{BC,max} \quad (39)$$

$$0 \leq P_{BD}^t \leq P_{BD,max} \quad (40)$$

El intercambio de energía con la batería comunitaria se realiza bajo un esquema de precios fijos. Específicamente, los prosumidores reciben un monto fijo por cada unidad de energía almacenada en la batería, mientras que todos los participantes deben pagar un monto fijo por unidad de energía extraída. Estos precios se mantienen constantes en el tiempo.

Los prosumidores pueden optar por vender energía a la batería en caso de que el precio ofrecido por el proveedor de servicios (SP) sea muy bajo. De la misma manera los consumidores y prosumidores tienen la opción de comprar energía de la batería en caso de que el precio de venta de energía ofrecido por el SP sea muy elevado.

4.2.2 Modelo de Respuesta del Cliente

El modelo de respuesta del cliente describe como los distintos participantes de la microrred adaptan su comportamiento en función de las señales de precio y sus necesidades en cada instante de tiempo.

En cada paso de tiempo, cada cliente presenta una demanda de carga que define el volumen de energía que se necesita proveer. La demanda de carga de cada cliente i en el paso de tiempo t se define como $d_i^t = d_{base,i}^t + d_{EV,i}^t$, donde d_{base}^t es la demanda base obtenida de datos históricos de microrredes, y luego normalizada y escalada por un factor multiplicativo para mantener consistencia en el modelo, y d_{EV}^t es la demanda de EVs, la cual es dinámicamente calculada para cada participante considerando todos los EVs asignados que estén conectados en el momento actual.

Para los prosumidores, la generación fotovoltaica en cada paso de tiempo proviene también de datos de reales normalizados y escalados. Los consumidores tienen generación PV igual a cero por definición.

La lógica de decisión de los participantes se basa en comparaciones de precios dinámicos versus precios fijos de la batería, en el caso de los consumidores, si el precio de venta ofrecido por el SP es menor al precio de venta ofrecido por la batería, el consumidor primero satisface su demanda energética comprando energía de la batería hasta donde sea posible y satisface la energía restante comprando energía del SP.

En el caso de prosumidores, la decisión se basa en el balance energético entre generación y demanda. En caso de un excedente energético, si el precio de compra del SP es menor que el precio de compra de la batería, entonces el prosumidor almacena su excedente en el batería primero y vende cualquier excedente restante a el SP. En el caso de un déficit energético, el prosumidor compensa su déficit de la misma manera que el consumidor.

4.2.3 Costos de Consumidor

El costo asociado al consumidor i en cada instante de tiempo t se define en función de la cantidad de energía adquirida entre dos fuentes: el proveedor de servicios (SP), y la batería comunitaria.

$$\phi_i^t\left(d_{i,sp}^t, d_{i,b}^t\right) = b_p\left(d_{i,b}^t\right) + a^t\left(d_{i,sp}^t\right) \tag{41}$$

Donde $d_{i,sp}^t$ y $d_{i,b}^t$ son la demanda al proveedor de servicios y a la batería comunitaria respectivamente. a_t es el precio de venta de energía establecido por el SP y b_p es el precio de venta de energía establecido por la batería. La decisión sobre cuanta energía demandar de cada fuente es dependiente en los precios ofrecidos por ambas.

4.2.4 Costos de Prosumidor

El costo modelado para los prosumidores presenta comportamientos diferentes según su balance energético. Un prosumidor en cada paso de tiempo t puede experimentar uno de dos escenarios, si su generación energética supera la demanda total entonces este prosumidor presenta un excedente, y en caso de que su generación g_i^t no sea suficiente para cubrir la demanda d_i^t entonces este presenta un déficit energético y deberá comprar energía de una de las fuentes disponibles de la misma manera que un consumidor.

Cuando existe un excedente, ($surplus = g_i^t - d_i^t > 0$), el prosumidor evalúa las opciones de venta de energía disponibles. La decisión de almacenar energía en la batería es dependiente de la relación entre los precios de compra de energía de batería b_c y el precio de compra de energía del proveedor de servicio p^t, donde le prosumidor siempre elige vender la energía al mayor postor. La cantidad de energía que puede ser almacenada en la batería $w_{i,b}^t$ está limitada por las restricciones operativas de la batería, por lo cual en caso de que ya no sea posible almacenar energía en la batería, se vende el excedente sobrante $w_{i,sp}^t$ al SP.

En situaciones de déficit energético ($deficit = d_i^t - g_i^t > 0$), el prosumidor puede satisfacer su demanda insatisfecha mediante descarga de la batería comunitaria, sujeto a la condición de precio $a^t > b_p$, donde a^t es el precio de venta de la energía del SP y b_v es el precio de venta de energía de la batería. La energía obtenida de la batería $d_{i,b}^t$ se puede ver limitada por las restricciones de descarga, por lo cual el déficit restante deberá ser cubierto mediante compra de la energía demandada restante $d_{i,sp}^t$ al SP.

Teniendo en cuenta estas consideraciones se puede formular el costo del prosumidor en cada instante de tiempo como:

$$\phi_i^t\left(d_{i,sp}^t, d_{i,b}^t, w_{i,sp}^t, w_{i,b}^t\right) = b_v \cdot d_{i,b}^t + a^t \cdot d_{i,sp}^t - b_p \cdot w_{i,b}^t - p^t \cdot w_{i,sp}^t \quad (42)$$

4.2.5 Vehículos Eléctricos como Cargas Dinámicas

La incorporación de EVs introduce un nuevo tipo de carga eléctrica dinámica cuyo comportamiento tiene patrones periódicos. En esta sección los EVs son modelados como cargas móviles.

La creación de la flota de vehículos en el modelo se define a partir de una tasa de adopción, a partir de la cual se determina el número de vehículos que se asignan a los participantes del sistema. Las características de cada vehículo son definidas de manera estocástica, y con parámetros fijos definidos individualmente para cada vehículo en el sistema. Cada vehículo se modela con su capacidad de almacenamiento energético, su potencia máxima de carga, y con una eficiencia asignada a su sistema de carga.

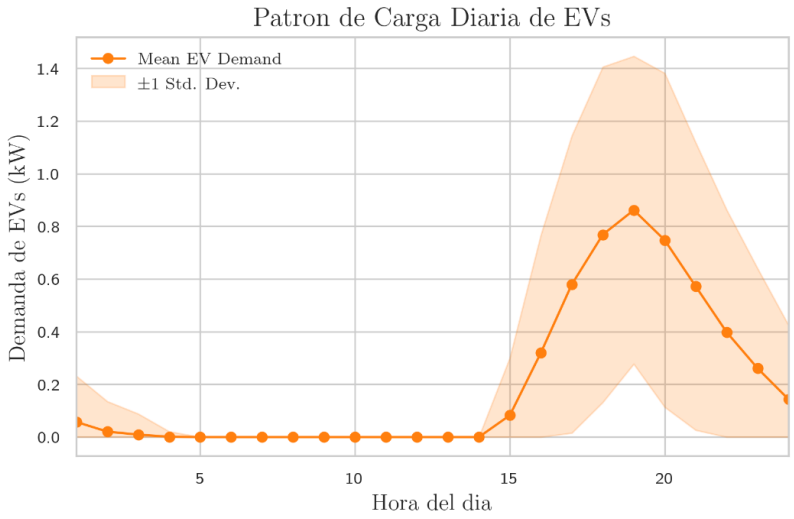


Figura Nº 15: Patrón de carga diaria de vehículos eléctricos

El comportamiento diario de los EVs se modela mediante distribuciones de Poisson, adaptado del modelo utilizado en [48] para su aplicación en zonas residenciales, permitiendo simular de forma realista los momentos de conexión y desconexión de los vehículos a la red doméstica. En este esquema se modelan de forma probabilística los patrones de conexión y desconexión de los vehículos en base a los horarios de salida y regreso de estos, los cuales se modelan a partir de un tiempo base de partida con la incorporación de una duración definida por una distribución de Poisson. Esto se puede observar en la Figura Nº 15.

El tiempo de llegada y salida se definen como:

$$T_{llegada,j} = T_{llegada,base_{j,}} + \mathrm{e}^{(\lambda_{llegada,j})} \tag{43}$$

$$T_{salida,j} = T_{salida,base_j} + \mathrm{e}^{(\lambda_{salida,j})} \tag{44}$$

Donde los parámetros $\lambda_{llegada,j}$ y $\lambda_{salida,j}$ son definidos aleatoriamente para cada vehículo individual, permitiendo variabilidad en los patrones de movilidad. Al desconectarse del sistema, cada vehículo experimenta una reducción en su estado de carga debido al consumo durante el trayecto, calculado como:

$$SoC_{llegada,j} = \max\left(\mathrm{SoC}_{\mathrm{min,j}}, \mathrm{SoC}_{\mathrm{inicial,j}} - \mathrm{E}_{\mathrm{trayecto,j}}\right) \tag{45}$$

Este proceso cíclico de desconexión, trayecto y reconexión permite una simulación continua y coherente del comportamiento diario de cada vehículo.

Una vez que un vehículo se conecta al sistema, se calcula la energía requerida para alcanzar su estado de carga objetivo antes de su salida. Esta energía se define, en el instante de tiempo t, como:

$$E_{necesaria,j}(t) = \left(\mathrm{SoC}_{target,j} - \mathrm{SoC}_j(t)\right) \cdot C_{EV,j} \quad (46)$$

Donde $\mathrm{SoC}_{target,j}$ es el valor de energía a partir del cual se deja de cargar el vehículo, definido por el usuario, y $C_{EV,j}$ es la capacidad de batería del vehículo. A partir de la energía calculada, y conociendo el tiempo ingresado por el usuario para la siguiente salida, se calcula la potencia de carga requerida como:

$$P_{carga,j}(t) = \min\left(\frac{E_{necesaria,j}(t)}{T_{disponible,j} \cdot \eta_{EV,j}}, P_{carga,max,j}\right) \quad (47)$$

Esta potencia se agrega a la carga base residencial del usuario correspondiente, lo cual permite calcular la demanda total en cada instante de simulación. Esto permite estudiar el impacto agregado del uso de múltiples vehículos eléctricos en los perfiles de consumo.

Dada la variabilidad introducida por la carga de EVs, y para acomodar la posibilidad de una futura incorporación de una función V2G, se implementa un mecanismo de coordinación que asigna prioridades a cada vehículo en función de su urgencia de carga. Esta prioridad se define como la razón entre el tiempo necesario para completar la carga y el tiempo restante hasta la salida:

$$\text{Prioridad}_j(t) = \frac{T_{carga\ necesaria,j}}{T_{disponible,j}} \quad 48)$$

Donde el tiempo de carga necesario depende de la energía necesaria y la máxima potencia que el vehículo puede recibir. En situaciones donde el tiempo de carga necesario es menor al tiempo disponible, se asigna una prioridad crítica, garantizando que estos vehículos empiecen a cargarse de inmediato.

Una vez calculadas las prioridades, los vehículos conectados se ordenan y la energía disponible se distribuye de forma secuencial, priorizando aquellos con mayor urgencia de carga. Este enfoque permite gestionar la demanda de forma eficiente y equitativa, evitando sobrecargas y mejorando la utilización de los recursos energéticos disponibles.

4.3 Algoritmo PPO

Para encontrar los precios óptimos en la microrred propuesta, se optó por utilizar DRL debido a que es capaz de capturar la estocasticidad de la microrred observando los estados y aprendiendo del entorno a través de retroalimentación para refinar sus decisiones. Se caracterizo la problemática de la utilización de precios dinámicos con RL para el comercio

P2P como un problema representado por un MDP, donde se consideran estados $s^t = \left(\mathrm{SoC}^t, d_{sp}^t, sin(2\pi h_t/24), cos(2\pi h_t/24)\right)$, acciones (a^t, p^t) y recompensa $r^t(s^t, a^t, p^t)$.

El estado del sistema incorpora características temporales cíclicas mediante funciones trigonométricas que capturan la periodicidad diaria de los patrones energéticos, además de parámetros de ponderación α y β que permiten al agente adaptar dinámicamente las prioridades del sistema, y están sujetos a restricción $0 \leq \alpha + \beta \leq 1$.

El costo de operación se construye mediante normalización adaptativa de los costos individuales de cada tipo de participante, permitiendo una comparación equitativa entre componentes con rangos de magnitud diferentes. Los costos brutos se definen como:

$$C_{\text{cons}}^t = b_v \cdot d_{\text{batt,cons}}^t + a^t \cdot d_{\text{sp,cons}}^t, \tag{49}$$

$$C_{\text{pros}}^t = \left(b_v \cdot d_{\text{batt,pros}}^t + a^t \cdot d_{\text{sp,pros}}^t\right) - \left(b_c \cdot w_{\text{batt,pros}}^t + p^t \cdot w_{\text{sp,pros}}^t\right) \tag{50}$$

$$C_{\text{prov}}^t = c_t \cdot I_{UG_{net}}^t + p^t \cdot w_{\text{sp,total}}^t - a^t \cdot d_{\text{sp,total}}^t \tag{51}$$

Donde $d_{\text{batt, cons}}^t$ y $d_{\text{sp, cons}}^t$r epresentan la energía suministrada a los consumidores por la batería y por el proveedor de servicios respectivamente. $d_{\text{batt, pros}}^t$, $d_{\text{sp, pros}}^t$, $w_{\text{batt,pros}}^t$, y $w_{\text{sp,pros}}^t$ denotan los déficits y excedentes energéticos de los prosumidores. c_t es el costo de la importación de energía de la red externa $I_{UG_{net}}^t$.

Para asegurar la comparabilidad en la importancia de los costos en los diferentes componentes con sus diferentes escalas, cada costo C_i^t fue normalizado dinámicamente por medio de una estrategia de actualización exponencial adaptativa durante el entrenamiento del agente, manteniendo limites adaptativos

$$C_i^{max}(t) = \max(\delta \cdot C_i^{max}(t-1), \phi_i^t) \tag{52}$$

$$C_i^{min}(t) = \min\left(\delta \cdot C_i^{min}(t-1) + (1-\delta) \cdot \phi_i^t, \phi_i^t\right) \tag{53}$$

Donde ϕ_i^t es el costo bruto observado para cada tipo de participante i en el paso t, y δ es el factor de decaimiento exponencial. Este método permite que los limites se ajusten suavemente a las condiciones del entorno.

Una vez completada una fase de exploración, los limites se fijan como constantes:

$$\tilde{C}_i^t = \frac{\phi_i^t - C_i^{min}}{C_i^{max} - C_i^{min}} \tag{54}$$

La función de costo total del sistema se define como una combinación convexa de los costos normalizados de cada tipo de participante, más un término adicional penalizando las emisiones de carbono.

$$\rho^t(s^t, a^t, p^t) = (1 - \alpha - \beta) \cdot \tilde{C}^t_{prov} + \alpha \cdot \tilde{C}^t_{cons} + \beta \cdot \tilde{C}^t_{pros} + \lambda_{CO_2} \cdot E^t_{CO_2} \quad (55)$$

Donde $\widetilde{C_i^t}$ representa los costos normalizados, $E^t_{CO_2}$ representa las emisiones asociadas a la importación energética desde la red, y λ_{CO2} es el costo de penalización por emisiones de carbono.

La señal de recompensa que guía el aprendizaje del agente se define como el negativo del costo total normalizado:

$$r^t(s^t, a^t, p^t) = -\rho^t(s^t, a^t, p^t). \quad (56)$$

El objetivo del agente es encontrar una política óptima $\pi^*: S \to A$ que maximice la recompensa esperada descontada a lo largo de un episodio completo de duración T (equivalente a 8760 pasos horarios, o un año en simulación):

$$G_t: \quad \max_{\pi: S \to A} E\left[\sum_{t=0}^{T} \gamma^t \cdot r^t\left(s^t, \pi(s^t)\right)\right] \quad (57)$$

Para trabajar en este entorno, se seleccionó el algoritmo de PPO, el cual ha demostrado gran robustez en entornos complejos y con alta variabilidad, este algoritmo es uno de los más ampliamente aplicados en práctica debido a su consistencia entrenando políticas eficientes, sea en espacios continuos o discretos. En este caso la capacidad de PPO para manejar entornos ruidosos lo hace una opción atractiva.

El algoritmo PPO surge como una respuesta a los problemas de inestabilidad de los métodos basados en gradientes de política y los métodos actor-críticos. Este algoritmo limita la magnitud de las actualizaciones de la política para evitar cambios drásticos que puedan degradar el rendimiento de los métodos basados en gradientes de política mientras mantienen su flexibilidad y hacen uso de la eficiencia incorporada por la estimación de valor. El algoritmo general puede ser observado en la Figura Nº 16.

PPO modifica la función objetivo del gradiente de política imponiendo una restricción en la razón de probabilidades entre la nueva política y la política anterior, limitando el cambio en cada iteración. Esta proporción se limita para permanecer en un rango $[1 - \epsilon, 1 + \epsilon]$ la política es definida como:

$$\mathcal{L}^{\text{Clip}}(\phi) = E_t\left[\min\left(r_t(\phi)\cdot \widehat{A_t}, \text{clip}(r_t(\phi), 1-\epsilon, 1+\epsilon)\cdot \widehat{A_t}\right)\right] \quad (58)$$

Donde:

$$r_t(\theta) = \frac{\pi_\phi^t(a|s)}{\pi_\phi^{t-1}(a|s)} \quad (59)$$

Y la función clip restringe el valor de r_t entre los limites $[1-\epsilon, 1+\epsilon]$, lo que permite que se favorezca cambios en direcciones que mejoren la política, pero sin hacer pasos demasiado grandes que puedan causar problemas de estabilidad.

Otro cambio que introduce PPO, es la optimización simultánea del crítico o la función de valor, la cual se entrena minimizando el error cuadrático entre el valor predicho y la estimación de retorno.

$$\mathcal{L}_{\text{valor}} = \frac{1}{2}(V(s_t;\psi) - G_t)^2 \quad (60)$$

Donde G_t es normalmente reemplazada por métodos de estimación de retornos más avanzados.

Finalmente, PPO hace uso de un término que incentiva la variedad en la política, es decir que tome acciones más diversas, mediante la incentivación de la entropía en la política. Al incluir un término involucrando el uso de entropía $\mathcal{H}$, se incentiva la exploración, empujando la política a mantener diversidad en sus acciones durante el aprendizaje.

Formando finalmente la función de pérdida conjunta para optimizar de manera simultánea la política, y el crítico, se combina de forma ponderada el termino de clip o recorte, el del valor, y el de la entropía, quedando expresada la función de perdida general:

$$\mathcal{L}^{\text{PPO}}(\phi,\psi) = E_t\left[\mathcal{L}^{CLIP}(\phi) - c_1\cdot \mathcal{L}_{valor}(\psi) + c_2\cdot \mathcal{H}[\pi_\phi](s_t)\right] \quad (61)$$

Algorithm 1 PPO (Algoritmo simplificado)

input: Parámetros iniciales de política ϕ_0, parámetros iniciales de critico θ_0.
for $k = 0, 1, 2, \ldots$ **do**
    Juntar trayectorias τ usando la política π_{ϕ_k}
    Calcular ventaja estimada $\hat{A}_t$ y retornos $\hat{G}_t$
    Calcular la razón de probabilidades entre políticas:

$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_k}(a_t \mid s_t)}$$

    Definir el objetivo limitado:

$$L_t^{\text{CLIP}}(\phi) = \min\left(r_t(\phi)\hat{A}_t, \text{clip}(r_t(\phi), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)$$

    Actualizar parámetros de la política:

$$\phi_{k+1} \leftarrow \phi_{k+1} + \alpha \nabla_\phi \left[\mathbb{E}_t\left[L_t^{\text{CLIP}}(\phi) + c_2 \mathcal{H}(\pi_\phi(\cdot|s_t))\right]\right]$$

    Actualizar los parámetros del critico:

$$\psi_{k+1} \leftarrow \psi_{k+1} - \beta \nabla_\psi \mathbb{E}_t\left[\tfrac{1}{2}(V_\psi(s_t) - \hat{R}_t)^2\right]$$

end for

Figura Nº 16: Algoritmo PPO

4.4 Evaluación

4.4.1 Configuración del Entorno

El entorno experimental fue implementado mediante una plataforma llamada OpenAI Gym, construido en base al trabajo de [49], diseñado para simular una microrred con comercio energético entre pares. Para el entrenamiento del agente de DRL se simulo una red compuesta por diez participantes, de los cuales un 50% son consumidores sin capacidad de generación de energía. Todos los participantes comparten acceso a un sistema comunitario de almacenamiento en baterías con una capacidad total de 25 kWh, una potencia máxima de carga y descarga de 10.5 kW y una eficiencia energética del 95 %. Como parte de la dinámica de consumo energético, se incorporan EVs como cargas móviles con una tasa de adopción del 50 % distribuidos aleatoriamente entre los participantes. Se pueden visualizar las trazas de generación y demanda del sistema simulado en la Figura Nº 17.

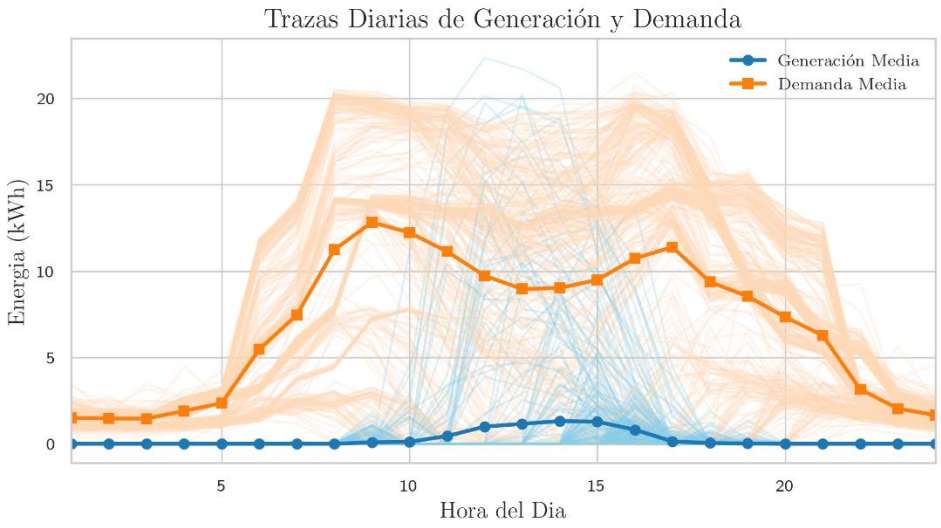


Figura Nº 17: Trazas diarias observadas de generación y demanda a lo largo de un año.

Las emisiones de CO_2 asociadas a la intensidad de carbono de la red (Fig. 18), los perfiles de carga y generación solar utilizados en la simulación se obtienen mediante simulación en el entorno diseñado, el cual genera patrones sintéticos pero representativos del comportamiento energético residencial. Estos perfiles se normalizan aplicando un escalado min-max y posteriormente se multiplican por un factor de seis con el objetivo de ajustar las magnitudes energéticas a un contexto realista de microrred, manteniendo al mismo tiempo las características temporales originales de los datos.

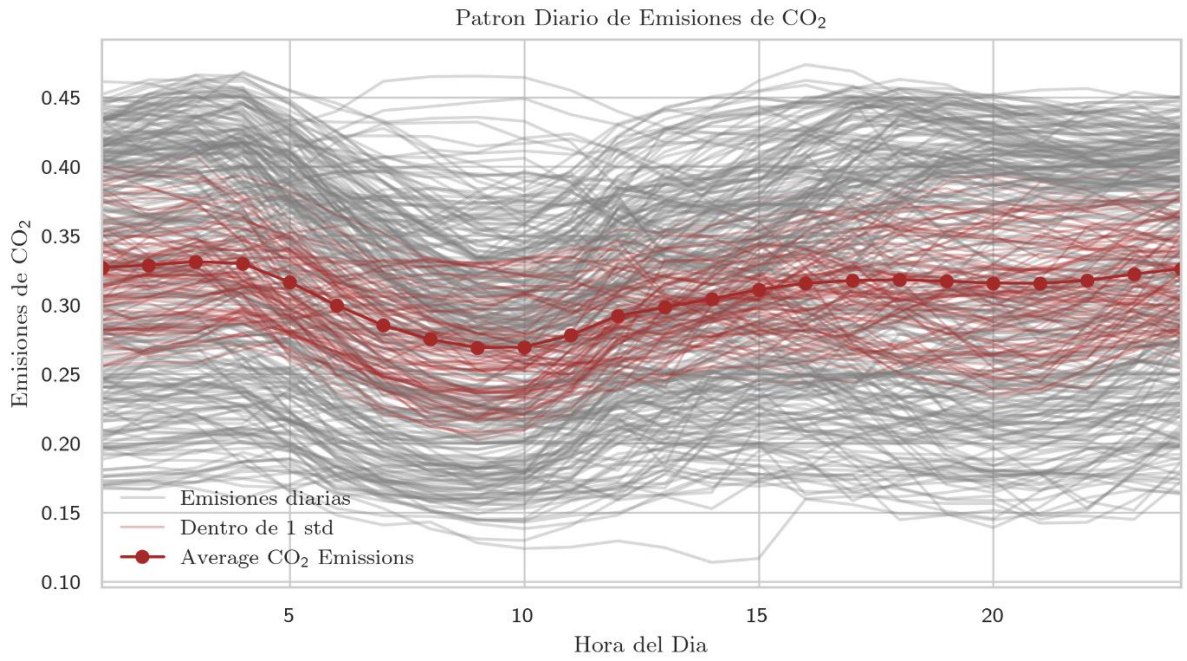


Figura Nº 18: Patrón diario de emisiones de carbono a lo largo de un año.

El algoritmo fue implementado mediante la biblioteca Stable-Baselines3 [50], utilizando redes neuronales del tipo perceptrón.

Para abordar el desafío de la normalización de costos en un contexto de múltiples actores, se adopta un protocolo de entrenamiento en dos fases. En la primera fase, se ejecutan cinco episodios con acciones aleatorias con el objetivo de establecer límites dinámicos de costos para cada tipo de participante (consumidor, prosumidor y proveedor), aplicando promedios móviles exponenciales con un factor de decaimiento de 0.99. En la segunda fase, dedicada al entrenamiento propiamente dicho, estos límites se mantienen fijos, lo que permite una normalización consistente de los costos a lo largo de 1,000,000 pasos de simulación. La función de costo total combina los costos normalizados de los distintos actores, ponderados por los coeficientes asignados a cada uno, e incluye además una penalización por emisiones de carbono proporcional a la energía importada desde la red.

La evaluación del rendimiento del modelo se realiza mediante simulaciones anuales completas que comprenden 8,760 pasos horarios, con un registro detallado de los flujos energéticos, los indicadores económicos y los impactos ambientales. Se incluyen métricas como el consumo de energía desde la red, el uso del sistema de almacenamiento, el comercio de excedentes entre pares, los costos asumidos por cada tipo de agente, y las decisiones de precios.

4.4.2 Resultados

Se analizo el comportamiento del agente y del sistema al entrenar el agente con coeficientes α = 0.33 y β = 0.33. Bajo esta configuración es forzado a explorar políticas de precios que armonicen los incentivos de consumidores, prosumidores y proveedor de la red, buscando un equilibrio en el que la minimización de los costos agregados para cada una de las partes.

El análisis de la evolución temporal del coeficiente de compra a prosumidores p_t (Fig. 19), revela que el agente fija precios elevados durante la franja nocturna, cuando la generación solar es nula, apostando por no ejecutar transacciones que pudieran aumentar sus costos y, al mismo tiempo, señalando la escasez potencial de energía. Justo al amanecer se observa un descenso abrupto del precio, lo que favorece la activación temprana de la batería comunitaria y reduce el coste residual de importación en la posterior baja en la demanda. Durante el pico solar, el agente sitúa p_t al límite mínimo, alentando tanto el consumo directo de excedentes como el almacenamiento de energía en la batería, en el posterior incremento en la demanda el agente ajusta el precio al alza para compensar las menores ganancias del periodo anterior y sostener el incentivo a la inyección de prosumidores.

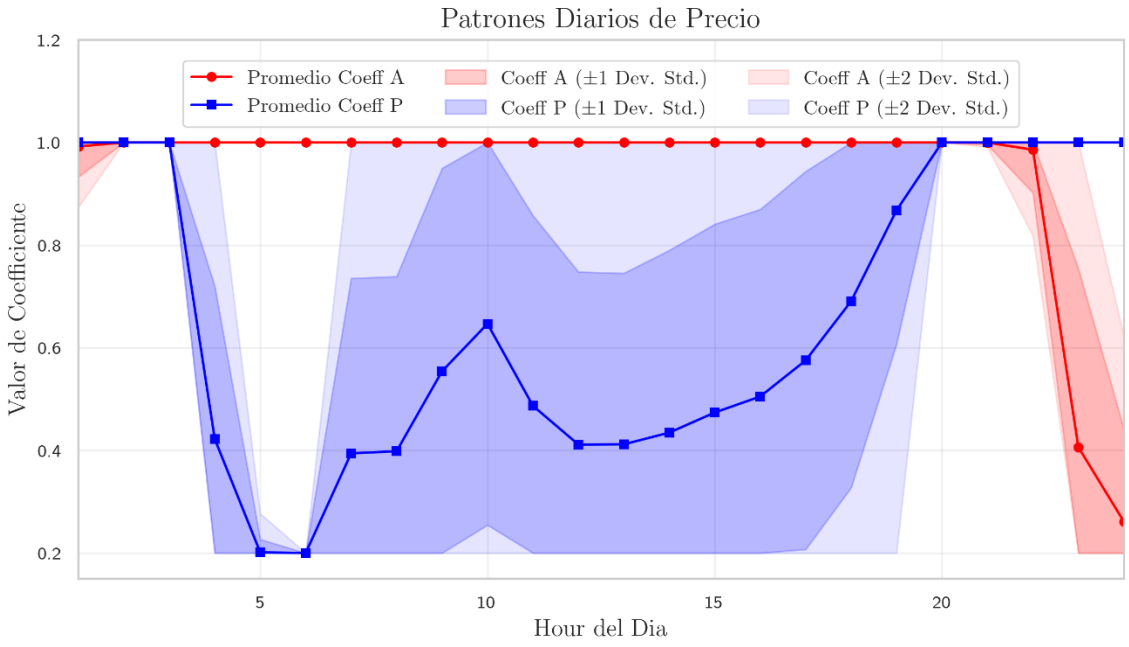


Figura Nº 19: Patrones diarios de precio (Precios de venta de prosumidores (Coeff P) y Precio Minorista (Coeff A).

El coeficiente de venta minorista a_t mantiene valores elevados a lo largo de la jornada de mayor actividad de los consumidores. Solamente reduciendo los precios en las últimas

horas de la noche, donde el impacto sobre el ingreso es marginal, y se prioriza el alivio del coste para el consumidor sin comprometer la rentabilidad del sistema. Este sesgo hacia precios altos en horas pico desacelera el consumo directo de la red, e incentiva el uso de la energía almacenada en la batería.

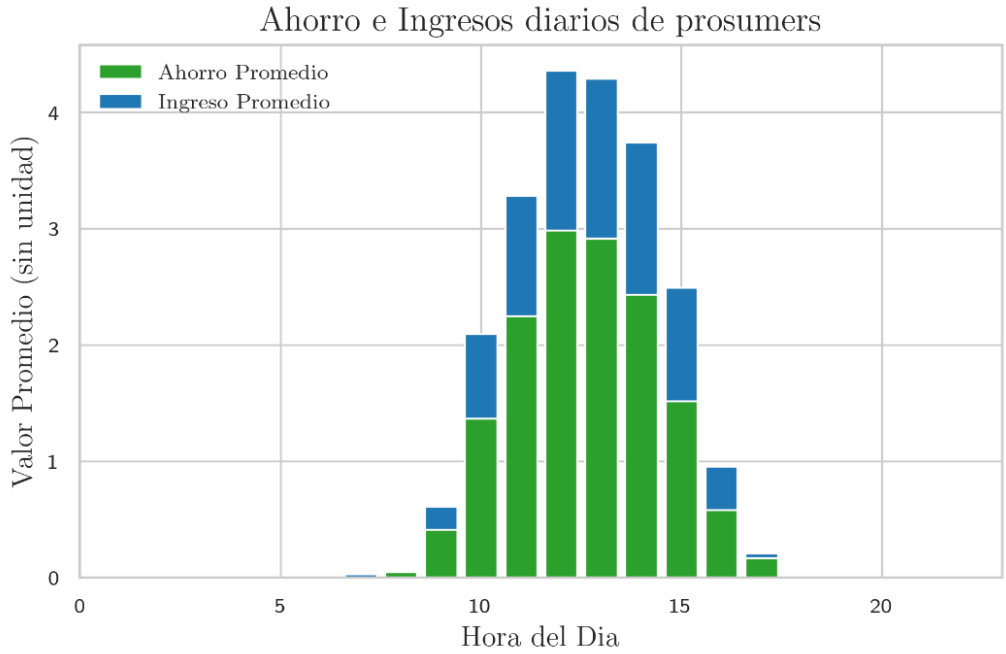


Figura Nº 20: Ahorro e Ingresos diarios de prosumidores a lo largo del día

La integración de la penalización por emisiones de carbono introduce una presión adicional en la política de precios. La mayor intensidad de CO_2 de la red durante la noche eleva el coste efectivo de las importaciones, reforzando la preferencia del agente por maximizar el uso de las fuentes de generación distribuidas y minimizar la dependencia nocturna de la red.

En la Figura Nº 21 se puede observar que el comportamiento de la batería comunitaria refleja una curva de carga que crece desde el amanecer hasta el mediodía, coincidiendo con la puesta de precios bajos de compra a prosumidores, y un vaciado progresivo llegando a la noche, cuando los precios minoristas permiten un arbitraje rentable.

Para evaluar como varían los perfiles de costo dependiendo de las diferentes prioridades del agente, se evalúa al agente frente a diferentes prioridades (α y β), y el costo para cada una de las partes, los resultados pueden ser observados en la Tabla 1. Los costos son evaluados sin unidad monetaria.

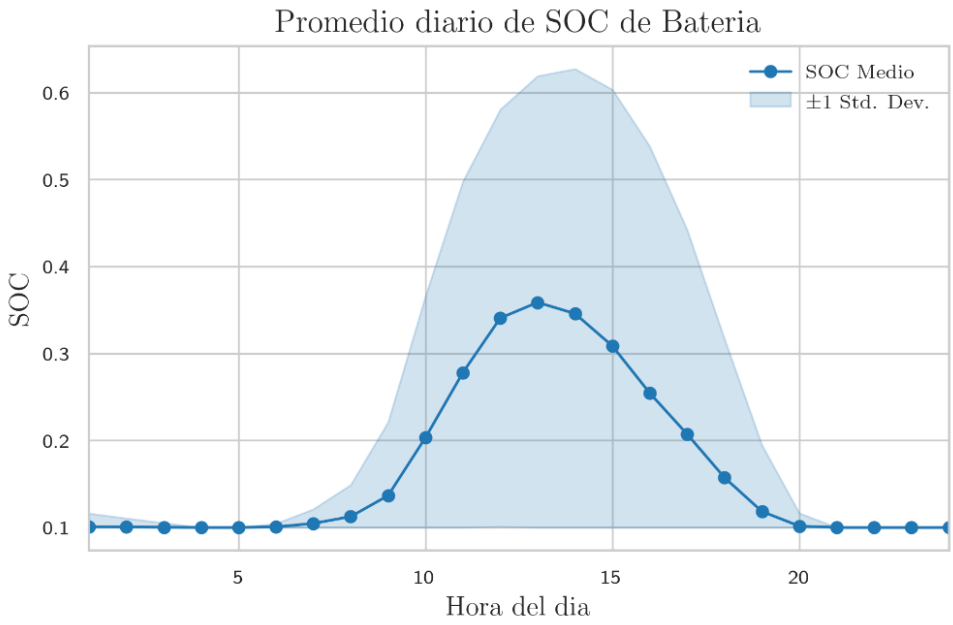


Figura Nº 21: Promedio diario de estado de carga de batería con el agente de DRL en operación

Tabla 1: Comparación de costos para diferentes prioridades (α y β)

α	β	$(1-\alpha-\beta)$	Costo de Consumidor	Costo de Prosumidores	Costo de SP
0.1	0.1	0.8	7.411800	3.769765	-7.865986
0.8	0.1	0.1	1.565005	-0.576144	2.380379
0.1	0.8	0.1	1.556857	-0.581599	2.421298
0.33	0.33	0.34	4.028117	1.072352	-1.812816

Al analizar los resultados de la Tabla 1, se puede observar que en el caso donde se otorga una prioridad mínima a minimizar el costo para el consumidor o prosumidor, nos encontramos en el escenario donde el proveedor de servicios (SP) genera la mayor cantidad de ganancias promedio, recibiendo altos costos para el consumidor y prosumidor. Al asignar $\alpha = 0.8$ la mayor prioridad del agente consiste en disminuir los costos para el consumidor, sin embargo esta configuración también logro costos bajos para los prosumidores, llegando a generar un patrón de costos similar al conseguido al utilizar un coeficiente $\beta = 0.8$, lo que muestra que hay una sinergia entre minimizar los costos para los prosumidores y minimizar los costos para los consumidores cuando se asigna baja importancia al proveedor de servicios, el cual en ambos casos asume perdidas moderadas, mientras la única ganancia es generada por los prosumidores.

En la ponderar de manera equilibrada la importancia de disminuir costos para cada una de las partes involucradas ($\alpha = 0.33, \beta = 0.33$), como se puede observar en la Figura Nº 22, el proveedor llega a obtener un beneficio moderado, mientras que los prosumidores disminuyen sus costos sin llegar a generar un ingreso neto, y los consumidores reciben costos menores a los obtenidos al priorizar los ingresos del proveedor de servicios.

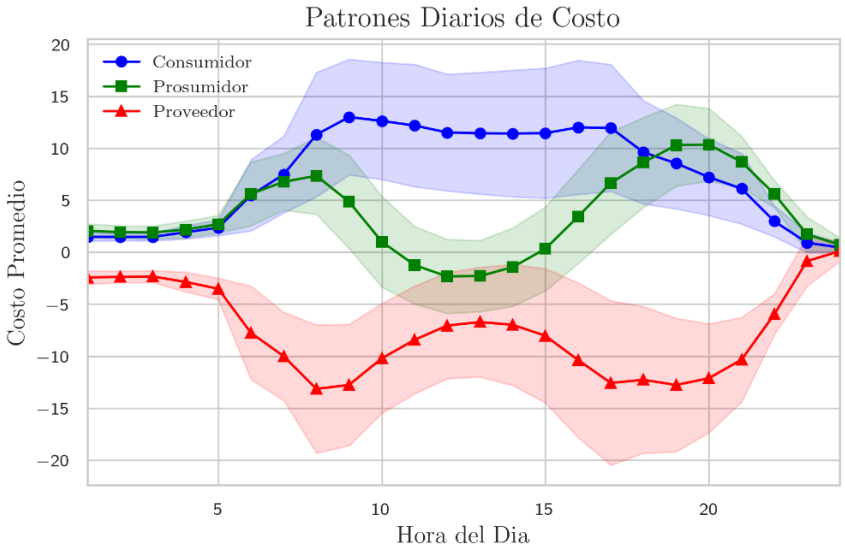


Figura Nº 22: Patrones diarios de costo.

En la Figura Nº 23 se puede observar cómo se reflejan las diferentes prioridades aplicadas en las distribuciones diarias de costos asumidos por cada tipo de participante, en el caso donde se prioriza minimizar los costos para el proveedor de servicios ($\alpha = 0.1, \beta = 0.1$), Se adoptan precios elevados en las horas de mayor demanda, maximizando los ingresos del proveedor, lo que resulta en altos costos para los consumidores y prosumidores en estos picos, también se puede observar que en este caso la ganancia del prosumidor solo disminuye en periodos de menor consumo o menor generación energética.

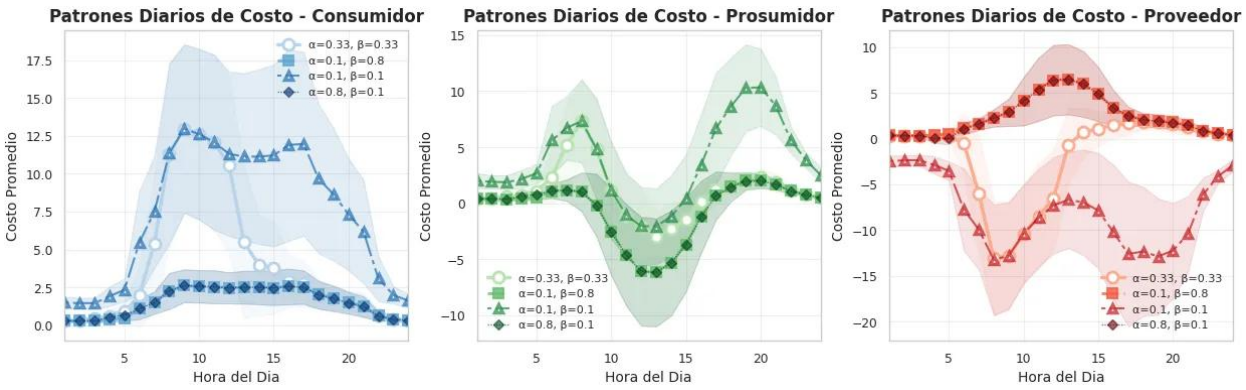


Figura Nº 23: Patrones de costo para diferentes prioridades (α y β)

En el caso orientado a disminuir los costos del consumidor ($\alpha = 0.8, \beta = 0.1$) o prosumidor ($\alpha = 0.1, \beta = 0.8$), el patrón de costos cambia, observando costos reducidos a lo largo del día para los consumidores, donde la suba de costos durante los picos de demanda es mínima y donde en los picos de generación los prosumidores generan ingresos elevados (costo negativo), mientras que el proveedor de servicios asume costos y no genera ganancia.

Al hacer uso de una ponderación equilibrada ($\alpha = 0.33, \beta = 0.33$), se puede observar un comportamiento híbrido. Durante la primera parte del día, el agente explota los momentos de mayor demanda para generar ingresos, y luego transiciona hacia un comportamiento más similar al observado en las políticas donde se priorizo el reducir costos para los consumidores o prosumidores, resultando en costos menores para estos mientras que el sistema se mantiene rentable para el proveedor de servicios.

4.5 Conclusiones

Este trabajo ha demostrado el potencial de DRL, como herramienta para la gestión dinámica de precios en redes P2P de energía, con el fin de descubrir estrategias de asignación de precios para balancear los costos para cada uno de los participantes.

Al ajustar los coeficientes de prioridad para los diferentes participantes, el sistema muestra comportamientos que disminuyen los costos basados en las prioridades dadas, siendo capaz de encontrar sinergias en las estrategias de puesta de precios, como se puede observar en el comportamiento del agente disminuyendo los costos de consumidores para disminuir el costo para consumidores.

La implementación de RL en estos sistemas representa un paso hacia la creación de mercados locales inteligentes y adaptables, donde se pueden aplicar estrategias de ajuste de precios que reflejen las prioridades de minimización de costos deseados mientras incentiven el uso eficiente de recursos del sistema.

CAPITULO 5: Conclusiones

La transición energética actual impulsa la adopción de redes eléctricas inteligentes capaces de integrar de manera eficiente recursos energéticos distribuidos y renovables. Sin embargo, esta transformación introduce una complejidad operativa significante, debido a la naturaleza variable de la generación renovable, la incorporación de sistemas de almacenamiento y la presencia de demandas flexibles como la recarga de EVs. Estos elementos convierten a la red eléctrica en un ecosistema dinámico que requiere estrategias de gestión que superen las limitaciones de enfoques tradicionales. En este contexto, las arquitecturas P2P emergen como una respuesta natural para fomentar la participación activa de los usuarios, además de habilitar intercambios energéticos flexibles y descentralizados. Sin embargo, esta transición hacia redes inteligentes distribuidas plantea problemas complejos de coordinación, predicción y control bajo incertidumbre. DRL se presenta como una herramienta para diseñar estrategias de gestión energética adaptativas que permitan la toma de decisiones eficientes a partir de patrones aprendidos. En este trabajo se desarrollan dos problemáticas, primero, se abordó el problema de operar microrredes con generación renovable, almacenamiento y demanda variable en un entorno dinámico, explorando cómo estrategias basadas en DRL pueden optimizar decisiones locales de compra, venta y almacenamiento de energía, como paso hacia la integración de esquemas P2P en redes inteligentes. En segundo lugar, se estudió el diseño de mecanismos de puesta de precios dinámicos en mercados P2P, utilizando DRL para determinar precios en tiempo real de manera que se minimicen los costos de manera eficiente para los participantes deseados, incentivando la adopción de estos sistemas.

En el capítulo 3 exploró el diseño e implementación de estrategias de gestión energética para microrredes con generación renovable, almacenamiento y cargas variables operando bajo precios dinámicos, demostrando que los algoritmos de DRL, como SAC, pueden aprender políticas de gestión que se aproximan al comportamiento óptimo teórico. El algoritmo SAC logro capturar de forma efectiva los patrones de generación, demanda y precios, permitiendo decisiones anticipadas y eficientes de almacenamiento e intercambio energético. El análisis de las curvas de energía comprada y vendida, junto con el comportamiento de los sistemas de almacenamiento, demostraron que el agente prioriza la utilización de la energía renovable disponible, vendiendo excedentes en momentos de baja demanda y comprando energía en momentos de alta demanda de manera consistente con estrategias óptimas.

En el capítulo 4 se demostró el potencial del DRL como herramienta para la gestión dinámica de precios en redes P2P, permitiendo descubrir estrategias de asignación de precios

que equilibren costos entre distintos participantes mientras se penalizan las emisiones de carbono. Al ajustar los coeficientes de prioridad de cada agente, el sistema logra reducir costos de acuerdo con las prioridades establecidas y encontrar sinergias en las estrategias de pricing, observándose que el agente es capaz de disminuir costos de consumidores con el objetivo de optimizar el costo global del sistema.

Los resultados obtenidos en este trabajo muestran que las técnicas de DRL pueden identificar patrones de comportamiento en sistemas energéticos complejos y utilizarlos para tomar decisiones en sistemas eléctricos con fuentes de generación energética distribuidos. Se observó que estos algoritmos tienen la capacidad de adaptarse a señales dinámicas del sistema y de tomar decisiones basadas en el comportamiento del entorno observado.

Sin embargo, estos algoritmos también presentan limitaciones que deben ser consideradas. Dado que el entrenamiento se realizó únicamente en simulación, para una transición a utilizar hardware real será necesario considerar posibles discrepancias (lo que se conoce como el problema Sim2Real) para implementar estos controladores en sistemas reales de manera segura y efectiva. Se pueden identificar, además, limitaciones en términos de escalabilidad, dado que los algoritmos fueron evaluados en sistemas de tamaño moderado y considerando el entrenamiento de agentes individuales, sin abordar escenarios multiagente en los que múltiples partes aprenden y actúan de manera simultánea. Finalmente, otro aspecto limitante es la dependencia en la definición de funciones de recompensa, ya que esta influye críticamente en el comportamiento aprendido por el agente.

Futuros trabajos pueden hacer uso de métodos de DRL multiagente para entrenar agentes inteligentes capaces de cooperar y operar en entornos compartidos con otros agentes, para lograr un manejo más eficiente y coordinado de la energía en redes P2P. El problema de la dependencia en simulaciones puede ser afrontado mediante el uso de técnicas de DRL Offline, donde se entrena una política a partir de datos históricos, sin interactuación constante con el entorno. Además, el uso de métodos de RL Federado puede contribuir a la aplicación de DRL en una mayor escala, mediante sistemas distribuidos, donde cada agente puede entrenar políticas localmente utilizando sus propios datos de consumo y generación de energía, y compartir de manera federada los parámetros aprendidos para construir colectivamente políticas más robustas sin necesidad de centralizar datos, preservando privacidad y reduciendo el costo de comunicación. Finalmente, la integración de batería como un agente el cual autorregula los precios de compra y venta energética en función de su estado de carga podría brindar una mayor eficiencia al uso de la batería.

En conclusión, este trabajo explora la aplicación de DRL en redes eléctricas inteligentes con arquitectura P2P, demostrando su potencial para mejorar la gestión de energía en sistemas distribuidos de manera eficiente y adaptativa, abriendo oportunidades para el avance hacia sistemas energéticos más sostenibles y adaptativos.

Glosario

P2P (Peer to Peer)	Sistema de intercambio descentralizado donde las entidades participan de manera descentralizada en transacciones.
IA (Inteligencia Artifical)	Campo de la informática que se centra en la creación de sistemas capaces de realizar tareas que normalmente requieren inteligencia humana. Esto incluye el aprendizaje, el razonamiento, la percepción y la toma de decisiones.
RL (Aprendizaje por Refuerzos)	Método de aprendizaje automático que aprende a tomar decisiones secuenciales utilizando señales de recompensa provenientes de interacciones con su entorno.
DRL (Aprendizaje por Refuerzos Profundo)	Extensión de RL que utiliza redes neuronales como aproximadores de funciones para manejar espacios de estado o acción de altas dimensiones.
V2G (Vehicle to Grid)	Tecnología que permite el flujo bidireccional de energía entre vehículos eléctricos y la red eléctrica.
Smart Grid	Red eléctrica avanzada que incorpora sensores, comunicaciones y sistemas de control para monitorear, predecir y gestionar flujos de energía de manera eficiente.
MDPs (Proceso de Decisión de Márkov)	Procesos de Decisión de Márkov, son un tipo de modelo matemático para toma secuencial de decisiones bajo incertidumbre.
EMS (Energy Management System)	Energy Management System, son sistemas encargado de monitorear, controlar y optimizar la generación, almacenamiento y consumo de energía en redes eléctricas, microredes o instalaciones industriales.
SoC (State of Charge)	Estado de carga de un sistema de almacenamiento de energía, generalmente expresado como un porcentaje de su capacidad nominal total.
TCL (Thermostatically Controlled Load)	Carga eléctrica controlada por un termostato, como sistemas de calefacción, ventilación, aire acondicionado y calentadores de agua.
SP (Proveedor de Servicios)	Entidad que provee servicios energéticos como suministro, balanceo, agregación y servicios auxiliares dentro del mercado eléctrico.
DER (Distributed Energy Resource)	Recurso energético descentralizado conectado localmente a la red, como paneles solares, turbinas eólicas, sistemas de almacenamiento y generadores de respaldo.
ESS (Energy Storage System)	Sistema capaz de almacenar energía y liberarla cuando es necesario, como baterías, sistemas de aire comprimido o volantes de inercia.
Prosumidor	Usuario que simultáneamente produce y consume energía, típicamente utilizando generación distribuida como paneles solares.

Referencias Bibliográficas

[1] R. Lu y S. H. Hong, “Incentive-based demand response for smart grid with reinforcement learning and deep neural network”, *Appl. Energy*, vol. 236, pp. 937–949, 2019.

[2] A. Ahmadian, B. Mohammadi-Ivatloo, y A. Elkamel, *Electric vehicles in energy systems: Modelling, integration, analysis, and optimization*. Springer, 2020.

[3] S. S. Madani *et al.*, “A Comprehensive Review on Lithium-Ion Battery Lifetime Prediction and Aging Mechanism Analysis”, *Batteries*, vol. 11, núm. 4, Art. núm. 4, abr. 2025, doi: 10.3390/batteries11040127.

[4] R. Kappagantu y S. A. Daniel, “Challenges and issues of smart grid implementation: A case of Indian scenario”, *J. Electr. Syst. Inf. Technol.*, vol. 5, núm. 3, pp. 453–467, dic. 2018, doi: 10.1016/j.jesit.2018.01.002.

[5] M. Anvari *et al.*, “Short term fluctuations of wind and solar power systems”, *New J. Phys.*, vol. 18, núm. 6, p. 063027, jun. 2016, doi: 10.1088/1367-2630/18/6/063027.

[6] *Directiva (UE) 2018/2001 del Parlamento Europeo y del Consejo, de 11 de diciembre de 2018, relativa al fomento del uso de energía procedente de fuentes renovables (versión refundida) (Texto pertinente a efectos del EEE.)*, vol. 328. 2018. Consultado: el 22 de julio de 2025. [En línea]. Disponible en: http://data.europa.eu/eli/dir/2018/2001/oj/spa

[7] *Directiva (UE) 2019/944 del Parlamento Europeo y del Consejo, de 5 de junio de 2019, sobre normas comunes para el mercado interior de la electricidad y por la que se modifica la Directiva 2012/27/UE (versión refundida) (Texto pertinente a efectos del EEE.)*, vol. 158. 2019. Consultado: el 22 de julio de 2025. [En línea]. Disponible en: http://data.europa.eu/eli/dir/2019/944/oj/spa

[8] L. Ableitner, A. Meeuw, S. Schopfer, V. Tiefenbeck, F. Wortmann, y A. Wörner, “Quartierstrom -- Implementation of a real world prosumer centric local energy market in Walenstadt, Switzerland”, el 29 de julio de 2019, *arXiv*: arXiv:1905.07242. doi: 10.48550/arXiv.1905.07242.

[9] International Energy Agency, “Energy Technology Perspectives 2023”, International Energy Agency, Paris, France, 2023. [En línea]. Disponible en: https://www.iea.org/reports/energy-technology-perspectives-2023

[10] W. Hou, Z. Ning, L. Guo, y X. Zhang, “Temporal, Functional and Spatial Big Data Computing Framework for Large-Scale Smart Grid”, *IEEE Trans. Emerg. Top. Comput.*, vol. 7, núm. 3, pp. 369–379, jul. 2019, doi: 10.1109/TETC.2017.2681113.

[11] E. Mocanu, P. H. Nguyen, M. Gibescu, y W. L. Kling, “Deep learning for estimating building energy consumption”, *Sustain. Energy Grids Netw.*, vol. 6, pp. 91–99, jun. 2016, doi: 10.1016/j.segan.2016.02.005.

[12] Y. LeCun, Y. Bengio, y G. Hinton, “Deep learning”, *nature*, vol. 521, núm. 7553, pp. 436–444, 2015.

[13] K. Amasyali y N. M. El-Gohary, “A review of data-driven building energy consumption prediction studies”, *Renew. Sustain. Energy Rev.*, vol. 81, pp. 1192–1205, ene. 2018, doi: 10.1016/j.rser.2017.04.095.

[14] B. Schäfer, C. Grabow, S. Auer, J. Kurths, D. Witthaut, y M. Timme, “Taming Instabilities in Power Grid Networks by Decentralized Control”, *Eur. Phys. J. Spec. Top.*, vol. 225, núm. 3, pp. 569–582, may 2016, doi: 10.1140/epjst/e2015-50136-y.

[15] B. Schäfer, M. Matthiae, M. Timme, y D. Witthaut, “Decentral smart grid control”, *New J. Phys.*, vol. 17, núm. 1, p. 015002, 2015.

[16] V. Arzamasov, K. Böhm, y P. Jochem, “Towards concise models of grid stability”, en *2018 IEEE international conference on communications, control, and computing technologies for smart grids (SmartGridComm)*, IEEE, 2018, pp. 1–6. Consultado: el 29 de junio de 2025. [En línea]. Disponible en: https://ieeexplore.ieee.org/abstract/document/8587498/

[17] A. Tavakoli *et al.*, “Self-scheduling of a generating company with an EV load aggregator under an energy exchange strategy”, *IEEE Trans. Smart Grid*, vol. 10, núm. 4, pp. 4253–4264, 2018.

[18] K. L. Lopez, C. Gagné, y M.-A. Gardner, “Demand-Side Management Using Deep Learning for Smart Charging of Electric Vehicles”, *IEEE Trans. Smart Grid*, vol. PP, pp. 1–1, feb. 2018, doi: 10.1109/TSG.2018.2808247.
[19] Y. Wang, L. Wang, M. Li, y Z. Chen, “A review of key issues for control and management in battery and ultra-capacitor hybrid energy storage systems”, *eTransportation*, vol. 4, p. 100064, may 2020, doi: 10.1016/j.etran.2020.100064.
[20] L. Patnaik, J. a V, y S. Williamson, “A Closed-Loop Constant-Temperature Constant-Voltage Charging Technique to Reduce Charge Time of Lithium-Ion Batteries”, *IEEE Trans. Ind. Electron.*, vol. PP, pp. 1–1, may 2018, doi: 10.1109/TIE.2018.2833038.
[21] Y. Li, K. Li, Y. Xie, J. Liu, C. Fu, y B. Liu, “Optimized charging of lithium-ion battery for electric vehicles: Adaptive multistage constant current–constant voltage charging strategy”, *Renew. Energy*, vol. 146, pp. 2688–2699, feb. 2020, doi: 10.1016/j.renene.2019.08.077.
[22] C. Liu, Y. Wang, y Z. Chen, “Degradation model and cycle life prediction for lithium-ion battery used in hybrid energy storage system”, *Energy*, vol. 166, pp. 796–806, ene. 2019, doi: 10.1016/j.energy.2018.10.131.
[23] W. Yin y X. Qin, “Cooperative optimization strategy for large-scale electric vehicle charging and discharging”, *Energy*, vol. 258, p. 124969, nov. 2022, doi: 10.1016/j.energy.2022.124969.
[24] J. R. Vazquez-Canteli, G. Henze, y Z. Nagy, “MARLISA: Multi-agent reinforcement learning with iterative sequential action selection for load shaping of grid-interactive connected buildings”, en *Proceedings of the 7th ACM international conference on systems for energy-efficient buildings, cities, and transportation*, 2020, pp. 170–179.
[25] S. Liu y G. P. Henze, “Evaluation of Reinforcement Learning for Optimal Control of Building Active and Passive Thermal Storage Inventory”, presentado en ASME 2005 International Solar Energy Conference, American Society of Mechanical Engineers Digital Collection, oct. 2008, pp. 301–311. doi: 10.1115/ISEC2005-76085.
[26] M. Trimboli, L. Avila, y M. Rahmani-Andebili, “Reinforcement Learning Techniques for MPPT Control of PV System Under Climatic Changes”, en *Applications of Artificial Intelligence in Planning and Operation of Smart Grids*, Springer, 2022, pp. 31–73.
[27] J. Heaton, “Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning”, *Genet. Program. Evolvable Mach.*, vol. 19, núm. 1, pp. 305–307, jun. 2018, doi: 10.1007/s10710-017-9314-z.
[28] V. Mnih *et al.*, “Playing Atari with Deep Reinforcement Learning”, el 19 de diciembre de 2013, *arXiv*: arXiv:1312.5602. doi: 10.48550/arXiv.1312.5602.
[29] M. Lehmann, “The Definitive Guide to Policy Gradients in Deep Reinforcement Learning: Theory, Algorithms and Implementations”, el 1 de marzo de 2024, *arXiv*: arXiv:2401.13662. doi: 10.48550/arXiv.2401.13662.
[30] Z. Wang *et al.*, “Sample Efficient Actor-Critic with Experience Replay”, el 10 de julio de 2017, *arXiv*: arXiv:1611.01224. doi: 10.48550/arXiv.1611.01224.
[31] H. Minor-Popocatl, O. Aguilar-Mejía, F. D. Santillán-Lemus, A. Valderrabano-Gonzalez, y R.-I. Samper-Torres, “Economic dispatch in micro-grids with alternative energy sources and batteries”, *IEEE Lat. Am. Trans.*, vol. 100, núm. XXX, 2022.
[32] O. A. Omitaomu y H. Niu, “Artificial intelligence techniques in smart grid: A survey”, *Smart Cities*, vol. 4, núm. 2, pp. 548–568, 2021.
[33] C. Hu, Z. Cai, Y. Zhang, R. Yan, Y. Cai, y B. Cen, “A soft actor-critic deep reinforcement learning method for multi-timescale coordinated operation of microgrids”, *Prot. Control Mod. Power Syst.*, vol. 7, núm. 1, pp. 1–10, 2022.
[34] T. A. Nakabi y P. Toivanen, “Deep reinforcement learning for energy management in a microgrid with flexible demand”, *Sustain. Energy Grids Netw.*, vol. 25, p. 100413, 2021.
[35] “Wind Farm data”, *Fortum Oy Finl.*, 2018.
[36] “Fingrid Open Datasets”, 2018, [En línea]. Disponible en: https://data.fingrid.fi/open-dataforms

[37] T. A. Nakabi y P. Toivanen, “An ANN-based model for learning individual customer behavior in response to electricity prices”, *Sustain. Energy Grids Netw.*, vol. 18, p. 100212, 2019.
[38] T. Haarnoja *et al.*, “Soft Actor-Critic Algorithms and Applications”, el 29 de enero de 2019, *arXiv*: arXiv:1812.05905. doi: 10.48550/arXiv.1812.05905.
[39] P. Christodoulou, “Soft Actor-Critic for Discrete Action Settings”, el 18 de octubre de 2019, *arXiv*: arXiv:1910.07207. doi: 10.48550/arXiv.1910.07207.
[40] H. Zhou *et al.*, “Revisiting Discrete Soft Actor-Critic”, el 20 de noviembre de 2024, *arXiv*: arXiv:2209.10081. doi: 10.48550/arXiv.2209.10081.
[41] G. Brockman *et al.*, “OpenAI gym”, *ArXiv Prepr. ArXiv160601540*, 2016.
[42] L. Yu, S. Qin, M. Zhang, C. Shen, T. Jiang, y X. Guan, “A review of deep reinforcement learning for smart building energy management”, *IEEE Internet Things J.*, vol. 8, núm. 15, pp. 12046–12063, 2021.
[43] D. Zhao, H. Wang, K. Shao, y Y. Zhu, “Deep reinforcement learning with experience replay based on SARSA”, en *2016 IEEE symposium series on computational intelligence (SSCI)*, IEEE, 2016, pp. 1–6.
[44] Irena, “Innovation landscape brief: Peer-to-peer electricity trading”, *Peer--Peer Electr. Trading Innov. Landsc. Brief*, 2020.
[45] N. Avila, S. Hardan, E. Zhalieva, M. Aloqaily, y M. Guizani, “Energy Pricing in P2P Energy Systems Using Reinforcement Learning”, el 24 de octubre de 2022, *arXiv*: arXiv:2210.13555. doi: 10.48550/arXiv.2210.13555.
[46] G. Henri, T. Levent, A. Halev, R. Alami, y P. Cordier, “pymgrid: An Open-Source Python Microgrid Simulator for Applied Artificial Intelligence Research”, el 11 de noviembre de 2020, *arXiv*: arXiv:2011.08004. doi: 10.48550/arXiv.2011.08004.
[47] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, y O. Klimov, “Proximal Policy Optimization Algorithms”, el 28 de agosto de 2017, *arXiv*: arXiv:1707.06347. doi: 10.48550/arXiv.1707.06347.
[48] K.-Y. Lo, J. H. Yeoh, y I.-Y. L. Hsieh, “Towards Nearly Zero-Energy Buildings: Smart Energy Management of Vehicle-to-Building (V2B) Strategy and Renewable Energy Sources”, *Sustain. Cities Soc.*, vol. 99, p. 104941, dic. 2023, doi: 10.1016/j.scs.2023.104941.
[49] N. Avila, S. Hardan, E. Zhalieva, M. Aloqaily, y M. Guizani, “Energy Pricing in P2P Energy Systems Using Reinforcement Learning”, el 24 de octubre de 2022, *arXiv*: arXiv:2210.13555. doi: 10.48550/arXiv.2210.13555.
[50] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, y N. Dormann, “Stable-Baselines3: Reliable Reinforcement Learning Implementations”.

Anexo: Configuración de Algoritmos y Código

Este anexo contiene los detalles de los parámetros y configuraciones utilizadas para entrenar los agentes mostrados en esta tesis. El anexo A.1 contiene enlaces a los repositorios que contienen el código utilizado. En el anexo A.2 se presentan los hiperparametros del algoritmo SAC utilizado para el manejo de microrredes. En el anexo A.3 se presentan los hiperparametros del algoritmo PPO y la simulación utilizada para el manejo de precios en la segunda parte del trabajo. Finalmente, en el anexo A.4 se muestran las curvas de entrenamiento de los algoritmos utilizados.

A.1: Enlaces a Implementaciones

Los códigos utilizados en los experimentos están disponibles en los siguientes repositorios:

- Estrategia basada en Soft-Actor-Critic para la gestión optima de energía: https://github.com/ga5am3/microgrid-energy-managment-model

- Fijación de precios de Energía en sistemas P2P utilizando PPO: https://github.com/ga5am3/RL_p2p_env_python_microgrid

A.2: Hiperparametros de entrenamiento SAC

Los experimentos con SAC usaron la siguiente configuración de hiperparámetros, mostrados en la Tabla 2:

Tabla 2: Hiperparametros SAC

Hiperparametro	**Valor**
Learning Rate	5×10^{-4}
Factor de Descuento (γ)	0.99
Factor de interpolación (τ)	1×10^{-3}
Tamaño de capas ocultas	256
Entropía objetivo	$-dim(\mathcal{A})$

Donde $\mathcal{A}$ es el espacio de acciones del entorno.

A.3: Hiperparametros de PPO y simulación de manejo de precios

Los experimentos con PPO se realizaron usando la biblioteca Stable Baselines3 durante 1,000,001 pasos de entrenamiento. La Tabla 3 muestra los hiperparámetros principales:

Tabla 3: Hiperparametros PPO

Hiperparámetro	Valor
Learning Rate	0.0003
Factor de Descuento (γ)	0.99
Lambda GAE (λ)	0.95
Rango de Clipeo	0.2
Coeficiente de Entropía	0.03
Coeficiente de Valor	0.25

Para la configuración del entorno de simulación utilizado para el entrenamiento del agente, se utilizaron los parámetros generales mostrados en la Tabla 4 y los parámetros de batería en la Tabla 5.

Tabla 4: Parámetros generales del entorno de simulación

Parámetro	**Valor**
Participantes	10
Tasa de consumidores	0.5
Coeficiente de costo de energía	0.3
Penalización CO2	0.01

Donde la cantidad de participantes se refiere a la cantidad conjunta de proveedores y consumidores considerados, y la tasa de consumidores se refiere al porcentaje de participantes representado por consumidores.

Tabla 5: Parámetros de Batería

Parámetro	Valor
Estado de carga Inicial	50%
Capacidad	25.0 kWh
Máximo estado de carga seguro	90%
Mínimo estado de carga seguro	10%
Máxima potencia de carga	10.5 kW
Máxima potencia de descarga	10.5 kW
Eficiencia	90%
Precio compra	0.6
Precio venta	0.6

A.4: Curvas de Entrenamiento

En esta sección presenta las curvas de entrenamiento obtenidas durante los experimentos con los algoritmos SAC y PPO utilizados en esta tesis. En la Figura 24 se muestra la curva de recompensa media móvil obtenido durante el entrenamiento del agente SAC, donde se observa el comportamiento de convergencia del agente y la estabilidad alcanzada en la fase final de entrenamiento. La variación inicial se debe al alto nivel de entropía inicial en la política, la cual lentamente disminuye a medida que el agente explora y descubre mejores políticas.

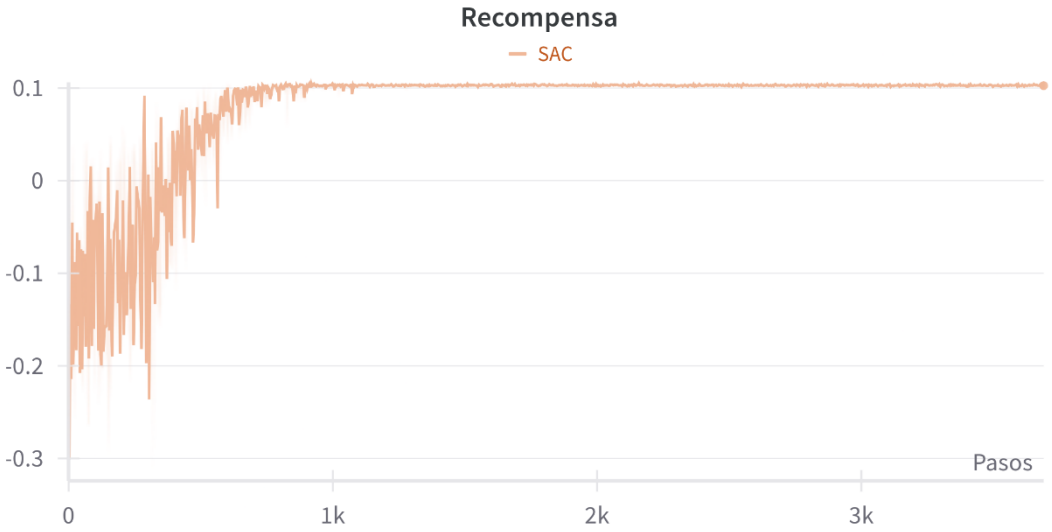


Figura Nº 24: Recompensa obtenida por el agente SAC en entorno del Capítulo 2

En la Figura Nº 25 se muestra la curva de recompensa media obtenida durante el entrenamiento del agente PPO para la simulación de ajuste de precios en el sistema P2P, donde se puede observar la estabilidad del algoritmo ya que mejora monótonamente mejora, debido a los pasos seguiros de avance que caracterizan al algoritmo.

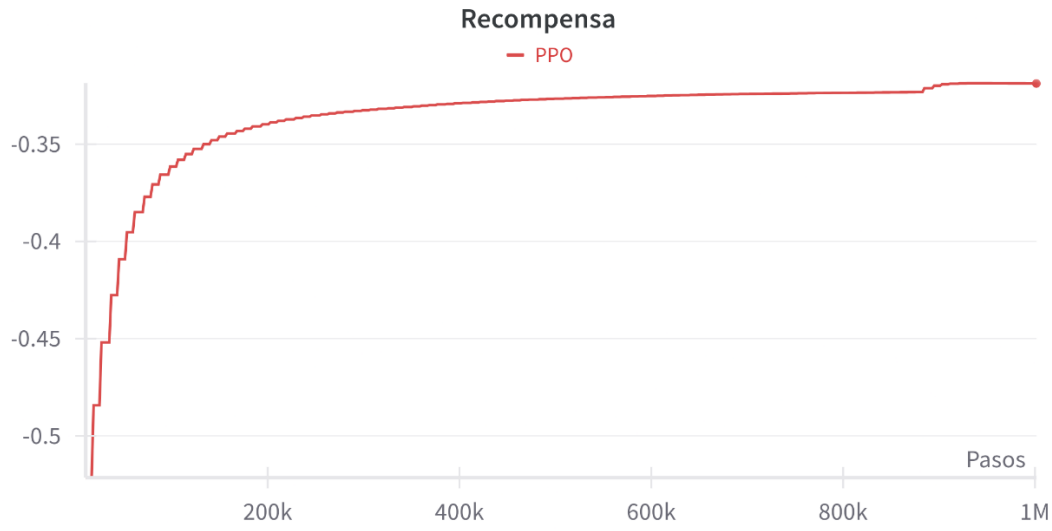


Figura Nº 25: Recompensa obtenida por agente PPO en el entorno de regulación de precios