An improved Soft Actor-Critic Strategy for Optimal Energy Management

Bruno Boato, Carolina Saavedra Sueldo, Luis Avila and Mariano de Paula.

***Abstract*—The transition from the current electrical grid to a smart, sustainable, efficient, and flexible electrical grid requires detecting future capabilities in order to have a system that can monitor, predict, learn, and make decisions on local energy consumption and production in real-time. A microgrid with these characteristics will allow the integration of distributed renewable energy systems efficiently, reducing the demand on power plants. The use of reinforcement learning can help find creative ways to keep the grid balanced; reschedule energy consumption through incentives; make predictions of demand and available energy at the grid scale, and assess the complexity of making these decisions. This work proposes using the novel Soft Actor-Critic (SAC) Deep Reinforcement Learning technique to manage electrical microgrids efficiently. SAC uses an entropy-based objective function that allows it to overcome the problem of convergence brittleness by encouraging exploration without assigning a high probability of occurrence to any part of the range of actions. Results show the benefits of the proposed technique for the coordinated energy management of the microgrid.**

***Index Terms*—Energy management, Distributed resources, Demand response, Deep Reinforcement Learning.**

I. Introduction

Muchos de los desafíos que imponen las microrredes se deben principalmente a la necesidad de mitigar el impacto que tienen las fallas en la red, la naturaleza intermitente de las fuentes renovables y las alteraciones en la calidad de la energía proveída a los consumidores. Teniendo en cuenta este complejo escenario, podemos vislumbrar que la gestión de una microrred impone nuevos desafíos tecnológicos para hacer frente a los requerimientos de sistemas de control, monitoreo y comunicación avanzados; desafíos económicos para desarrollar nuevos modelos de negocio que surjan de una nueva forma de concebir el mercado energético; y desafíos regulatorios para establecer estándares que especifiquen la base de la interoperabilidad requerida para su factibilidad [1].

El aprendizaje por refuerzos (RL) [2] podría ser una de las tecnologías habilitadoras para el futuro desarrollo de sistemas de gestión de microrredes, facilitando el desarrollo de estrategias para toma de decisiones. El RL hace uso eficiente de la experimentación 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. Trabajos recientes, han evaluado el potencial de las técnicas RL para gestionar la demanda mediante incentivos [3], la gestión de fuentes de almacenamiento térmico [4], gestión de fuentes renovables [5]. Otros estudios han comenzado a poner énfasis en los mecanismos de gestión para dar cuenta del cambio de pico de la demanda cuando múltiples agentes toman las mismas decisiones de control [6].

No obstante, algunos de los algoritmos de RL más exitosos de los últimos años sufren de ineficiencia durante la toma de muestras dado que necesitan muestras completamente nuevas después de cada actualización de la política. Esto se debe principalmente a que utilizan una estrategia *on-policy* mediante la cual utilizan la misma política para aprender y para explorar el entorno. Por el contrario, los métodos *off-policy* pueden aprender de manera eficiente de muestras pasadas utilizando búferes de reproducción de experiencia. Sin embargo, el problema con estos métodos es que son muy sensibles a los hiperparámetros y requieren muchos ajustes para que converjan. Soft Actor-Critic [7] utiliza el concepto de aprendizaje de máxima entropía para combatir la fragilidad de la convergencia. La característica más importante de SAC es que utiliza una función objetivo modificada, y busca maximizar la entropía en la política además de la recompensa esperada. Políticas con baja entropía tienen una tendencia a muestrear con mayor regularidad ciertos valores, ya que la masa de probabilidad se distribuye de manera relativamente desigual. Recompensar la política por alta entropía trae varias ventajas conceptuales: primero, fomenta explícitamente la exploración del espacio de estado; segundo, evita la convergencia prematura a malos óptimos locales.

Este trabajo aborda el problema de operar de manera eficiente los componentes de una microrred eléctrica con fuentes renovables, capacidades de almacenamiento y demanda variable [8]. El problema 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. El sistema de gestión de la microrred se basa en un algoritmo de control SAC que permite una mejor convergencia de la política. Los resultados muestran los beneficios proporcionados por la técnica propuesta para la gestión energética coordinada de la microrred. Para fines de comparación, se implementan estrategias de gestión optimizadas de manera heurística.

La organización del trabajo es como sigue. La sección II

Bruno Boato and Luis Avila are with Facultad de Ingeniería y Ciencias Agropecuarias, Universidad Nacional de San Luis, Ruta Prov. Nº 55, D5730EKQ, San Luis, Argentina. Luis Avila is with Laboratorio de Investigación y Desarrollo en Inteligencia Computacional (LIDIC), CONICET-UNSL, Av. Ejército de los Andes 950, D5700HHW San Luis, Argentina e-mail: brunoboato@gmail.com, loavila@unsl.edu.ar.

Carolina Saavedra Sueldo and Mariano de Paula are with Centro de Investigaciones en Física e Ingeniería del Centro -UNICEN - CICpBA - CONICET, INTELYMEC, Olavarría, B7400JWI, Argentina e-mail: carolina.saavedra@fio.unicen.edu.ar, mariano.depaula@fio.unicen.edu.ar

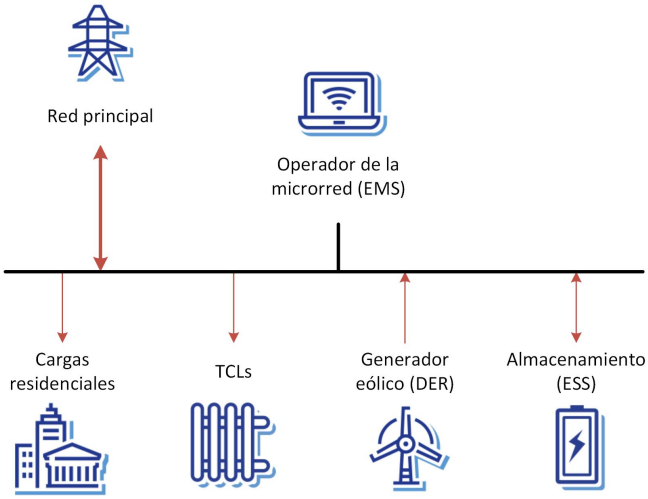


Fig. 1. Diagrama de la arquitectura de la microrred.

describe el modelo de la microrred utilizada para simulación. En la sección III se introducen las técnicas de aprendizaje por refuerzos y se presenta en detalle el algoritmo SAC utilizado para la gestión de la microrred. En la sección IV 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 V se presentan las conclusiones y se deja abierta la discusión para futuros trabajos.

II. Arquitectura de la Microrred

La microrred estará gestionada por un agregador o una empresa de servicios públicos que se encarga de suministrar la electricidad para satisfacer la demanda local [9]. Si bien la microrred tiene sus propios recursos de energía distribuida basados en turbinas eólicas, también está conectada a la red principal a través de la cual compra o vende energía continuamente en los mercados de electricidad. La arquitectura de la microrred que se muestra en la Fig. 1, incluye un generador eólico (DER), un sistema de almacenamiento comunitario (ESS), un clúster de cargas controladas termostáticamente (TCLs) y un grupo de cargas residenciales vinculadas mediante precios dinámicos. Se considera un sistema de comunicación bidireccional entre cada uno de los componentes y el gestor de la microrred que informa sobre los precios de la electricidad, el estado de carga de la batería y la generación de energía. El agente inteligente, que controla la gestión de la energía (EMS), envía señales a los componentes sobre el control de encendido/apagado de los TCLs, el control de carga/descarga del sistema de almacenamiento de energía (ESS) y el control de compra/venta de la red de energía.

A. Agente EMS

El agente EMS debe utilizar la información proporcionada por los diferentes componentes de la red y el entorno para determinar la estrategia óptima de equilibrio entre oferta y demanda. El agente realiza la gestión global de la microrred utilizando cuatro mecanismos de control: control directo de las unidades TCL, control de precios y toma de decisiones frente a la deficiencia y exceso de energía.

Con respecto al control de los TCLs, en cada paso de tiempo t, el agente EMS asigna una cierta cantidad de energía para usar en la operación del clúster. Luego, esta energía se envía a través de un agregador intermedio a los TCLs individuales. El agregador determina las acciones de encendido/apagado de cada TCL en función de su prioridad de estado de carga (SoC). El agregador de TCL también comunica el SoC promedio en tiempo real del clúster de TCLs al agente de EMS.

Por otro lado, el agente EMS debe determinar el nivel de tarifa residencial δ_t a utilizar en cada paso de tiempo. Dado que el administrador de la microrred no tiene el monopolio sobre la demanda, los precios ofrecidos pueden fluctuar alrededor de un valor medio mientras se mantiene el precio promedio diario P_{avg} cerca del precio de mercado P_{market} ofrecido por los minoristas de electricidad. Las variaciones del precio sirven como herramienta para desplazar las cargas de los periodos pico hacia periodos de mayor disponibilidad de potencia.

Cuando los DER locales no pueden satisfacer la demanda, la microrred local puede usar la energía almacenada en el ESS o comprar energía de la red principal y guardar la energía del ESS para su uso posterior. En cada paso de tiempo, el agente EMS establece la prioridad de uso entre estos dos recursos. En consecuencia, 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.

Dado que la energía generada por los DER locales puede exceder la demanda, el exceso de energía puede almacenarse en el ESS o venderse a la red principal. El agente EMS especifica la opción prioritaria para el uso excesivo de energía por adelantado, de manera similar al escenario de deficiencia de energía. Si el ESS es la opción prioritaria y se alcanza la capacidad de la batería, la energía restante se transfiere automáticamente a la red principal.

B. Almacenamiento de Energía

La arquitectura de la microrred hace uso de un ESS comunitario en lugar de almacenamientos de batería en cada residencia. El ESS utilizado es capaz de cubrir un mínimo de 2h de la demanda energética de la microrred. En cada paso de tiempo t la dinámica de almacenamiento del ESS es modelada como

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

donde $B_t \in [0, B_{\text{máx}}]$ es la energía almacenada en el ESS en el tiempo t, $B_{\text{máx}}$ 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 $C_t \in [0, C_{\text{máx}}]$ y $D_t \in [0, D_{\text{máx}}]$ son las potencias de carga y descarga que están restringidas por las limitaciones de velocidad de carga y descarga del ESS $C_{\text{máx}}$ y $D_{\text{máx}}$ respectivamente. También se define la variable de estado de carga del ESS como

$$BSC_t = \frac{B_t}{B_{\text{máx}}} \tag{2}$$

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 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.

C. Recursos de Energía Distribuida

Se considera que la microrred está equipada con turbinas eólicas capaces de generar cantidades variables de energía, dependiendo de las condiciones meteorológicas. En lugar de utilizar un modelo para la generación de energía, se utilizan datos reales de producción de energía eólica de un parque eólico [10]. El DER comparte la información sobre la generación de energía actual G_t con el EMS y suministra la energía generada directamente a la red local.

D. Red de Suministro

La microrred está conectada a una red principal que actúa como ente de regulación. La oferta y la demanda en la microrred no se pueden equilibrar utilizando únicamente los DER, debido a la naturaleza intermitente e incontrolable de estos recursos. La red eléctrica principal puede suministrar energía instantáneamente a la microrred en caso de deficiencia de energía o aceptar el exceso de energía en caso de excedente. 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 [11], representados respectivamente como (P_t^u, P_t^d). Para definir la fuente de suministro prioritaria en caso de deficiencia y la fuente de descarga de energía prioritaria en caso de exceso, el EMS controla solo el interruptor eléctrico a la red principal. Después de cada paso de tiempo, el EMS recibe información sobre la energía E_t comprada o vendida a la red principal, donde los valores positivos indican energía comprada y los valores negativos energía vendida.

E. Cargas Controladas Termostáticamente

Un clúster de TCLs puede proporcionar una fuente significativa de flexibilidad debido a su conservación térmica de la energía. Asumimos que la mayoría de los hogares en la microrred estaban equipados con un TCL (como un aire acondicionado, bomba de calor, calentador de agua o refrigerador). Los TCLs se pueden controlar directamente en cada paso de tiempo t mediante señales de control del agregador del clúster. A cada usuario se le cobra el costo de generación de energía C_{gen} para preservar los niveles de comodidad. El controlador de respaldo recibe la acción de encendido/apagado del agregador, verifica las restricciones de temperatura y modifica la acción según

$$a_{b,t}^i = \begin{cases} 0 & if \quad T_t^i > T_{max}^i \\ a_t^i & if \quad T_{min}^i < T_t^i < T_{max}^i \\ 1 & if \quad T_t^i < T_{min}^i \end{cases} \tag{3}$$

donde $a_{b,t}^i$ es la acción final de encendido/apagado después de la decisión del controlador de respaldo; T_t^i es la temperatura operativa de TCLi en el tiempo t; y T_{max}^i y T_{min}^i son los límites de temperatura superior e inferior establecidos por el usuario final, respectivamente. La dinámica de temperatura de cada TCL se modela como

$$\dot{T}_t^i = \frac{1}{C_a^i}\left(T_t^0 - T_t^i\right) + \frac{1}{C_m^i}\left(T_{m,t}^i - T_t^i\right) + L_{tcl}^i u_{b,t}^i + q^i \tag{4}$$

$$\dot{T}_{m,t}^i = \frac{1}{C_m^i}\left(T_t^i - T_{m,t}^i\right) \tag{5}$$

donde T_t^i es la temperatura del aire interior medida; $T_{m,t}^i$ es la temperatura de la masa del edificio no observable; T_t^0 es la temperatura exterior; C_a^i y C_m^i son las masas térmicas del aire y los materiales de construcción, respectivamente. Por su parte, q_i es la calefacción interna del edificio y L_{tcl}^i es la potencia nominal de la TCL. Finalmente, para cada TCL existe una medida de estado de carga SoC_t^i que determina la posición relativa de T_t^i en el rango de temperatura deseado.

F. 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 durante las cuales se devuelven las cargas desplazadas. La carga eléctrica L_t^i del hogar i en el tiempo t se modela como

$$L_t^i = L_{b,t} - SL_t^i + PB_t^i \tag{6}$$

$$SL_t^i = L_{b,t} * \beta_i * \delta_t \tag{7}$$

donde $L_{b,t} > 0$ indica la carga básica que sigue un patrón de consumo diario [12]. 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.

III. SAC PARA GESTIÓN DE LA MICRORRED

El aprendizaje por refuerzo asume que hay un agente ubicado en un entorno que interactúa mediante acciones sobre el entorno en búsqueda de maximizar la recompensa total recibida por el agente, para lo cual el problema de RL debe formalizarse como un Proceso de Decisión de Markov (MDP) [13].

A. Aprendizaje por Refuerzos

Una formulación de RL se define por cuatro elementos: el espacio de estado S, el espacio de acción A, la probabilidad de transición de estado p y la función de recompensa r. Para el problema de control en el tiempo se toma una acción que corresponde a un valor a_t para la variable manipulada. Durante el proceso de aprendizaje, el agente interactúa con el sistema aplicando una acción $a_t \in A \subseteq \mathbb{R}^{n_A}$ y, después de eso, el sistema evoluciona desde el estado $s_t \in S \subseteq \mathbb{R}^{n_S}$ a un estado sucesor s_{t+1} y el agente recibe una señal numérica r_t que proporciona una medida de qué tan buena (o mala) fue la acción a_t elegida y aplicada en el sistema en el estado s_t en función de la transición de estado ocurrida s_{t+1}. Las recompensas actúan como pistas sobre el logro de objetivos o el comportamiento óptimo. Por lo tanto, el objetivo de los métodos RL es encontrar una política óptima π^* que satisfaga

$$J^* = \max_{\pi} J_\pi = \max_{\pi} \mathrm{E}\left\{R_t | s_t\right\} \tag{8}$$

donde J_π corresponde a la recompensa total esperada dada la política de control π.

Supongamos que dada una política π, la función de valor $V^\pi(s_t)$ durante un cierto intervalo de tiempo, depende de la secuencia de transiciones de estado $T(s, a, s') = p(s_{t+1} = s' | s_t = s, a_t = a)$, lo que da lugar a la suma descontada de recompensas futuras $R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots$, que permite describir la función de valor de estado esperada descontada para una política π en el estado s_t

$$V^\pi(s_t) = \mathrm{E}_\pi\{R_t | s_t\} = \mathrm{E}_\pi\left\{r_t + \gamma V\pi(s_{t+1})\right\} \tag{9}$$

donde $\gamma \in (0, 1]$ es el factor de descuento que pondera las recompensas futuras. $V^*(s_t)$ se usa para denotar la máxima recompensa con descuento obtenida cuando el agente comienza en el estado s_t y ejecuta la política óptima. Por lo tanto se tiene que

$$\pi^* = \operatorname*{argmax}_{\pi} V^\pi(s_t) \tag{10}$$

Por otro lado, la función Q, describe el rendimiento esperado de seleccionar una acción a_t, en el estado s_t, y seguir la política π en adelante

$$Q^\pi(s_t, a_t) = r_t + \gamma V^\pi(s_{t+1}) \tag{11}$$

B. Algoritmo SAC

SAC es un algoritmo *off-policy* basado en el marco RL de máxima entropía, introducido por Haarnoja et al. [7]. SAC es capaz de manejar espacios de acción continuos, mejorando su aplicabilidad a una variedad de problemas de control. SAC utiliza una arquitectura actor-crítico que emplea dos redes neuronales profundas diferentes para aproximar la función Q y la función estado-valor V. El actor mapea el estado actual en función de la acción que estima óptima, mientras que el crítico evalúa la acción calculando la función de valor. Este algoritmo se basa en el marco de aprendizaje de refuerzo de máxima entropía, en el que el objetivo es maximizar tanto la recompensa esperada como la entropía de la siguiente manera

$$J_\pi = \operatorname*{argmax}_{\pi} \mathrm{E}_\pi \left\{ \sum_{t=0}^{T} \gamma^t (r_t + \alpha H_t^\pi) \right\} \tag{12}$$

donde H^π es el término de entropía de Shannon, que expresa la actitud del agente al realizar acciones aleatorias, y α es un coeficiente de regularización que indica la importancia del término de entropía sobre la recompensa. Generalmente, α es cero cuando se consideran algoritmos de aprendizaje por refuerzo convencionales. La maximización de esta función objetivo asegura que el agente sea empujado explícitamente hacia la exploración de nuevas políticas y al mismo tiempo evita que se quede estancado en un comportamiento subóptimo.

Ahora que sabemos para qué estamos optimizando, comprendamos cómo hacemos la optimización. SAC hace uso de tres redes: una función de valor de estado V parametrizada por ψ, una función suave Q parametrizada por θ y una función de política π parametrizada por ϕ. Entonces necesitamos entrenar los tres aproximadores de función.

En primer lugar entrenamos la función de valor minimizando el siguiente error

$$J_V(\psi) = \mathrm{E}_{s_t}\left[\frac{1}{2}\Big(V_\psi(s_t) - \mathrm{E}_{a_t \backsim \pi_\phi}\big(Q_\theta(s_t, a_t) - \log \pi_\phi(a_t|s_t)\big)\Big)^2\right] \tag{13}$$

donde debe disminuirse la diferencia al cuadrado entre la predicción de nuestra función de valor V y la predicción esperada de la función Q más la entropía de la función de política π (medida aquí por el logaritmo negativo de la función de política).

Para entrenar la función que aproxima la política π debe minimizarse el siguiente error

$$J_\pi(\phi) = \mathrm{E}_{s_t}\left[\mathrm{D}_{\mathrm{KL}}\left(\pi_\phi(s_t) \middle\| \frac{\exp\left(Q_\theta(s_t)\right)}{Z_\theta(s_t)}\right)\right] \tag{14}$$

donde el término D_{KL} es la Divergencia Kullback-Leibler que mide la distancia entres dos distribuciones. Buscando el mínimo de esta función objetivo, se pretende que la distribución de nuestra función de política se parezca más a la distribución de la exponenciación de nuestra función Q normalizada por otra función Z.

Para minimizar este objetivo, se puede utilizar el truco de la reparametrización haciendo $a_t = f_\phi(\epsilon_t; s_t)$. Este truco se usa para asegurarse de que el muestreo de la política sea un proceso diferenciable. La política parametrizada se escribe de la siguiente manera

$$J_\pi(\phi) = \mathrm{E}_{s_t, \epsilon_t}\left[\log \pi_\phi\left(f_\phi(\epsilon_t, s_t)|s_t\right) - Q_\theta\left(s_t, f_\phi(\epsilon_t, s_t)|s_t\right)\right] \tag{15}$$

La función de normalización Z se descarta ya que no depende del parámetro ϕ. Un estimador insesgado para el gradiente del objetivo anterior es

$$\hat{\nabla}_\phi J_\pi(\phi) = \nabla_\phi \log \pi_\phi(a_t|s_t) + \left(\nabla_{a_t} \log \pi_\phi(a_t|s_t) - \nabla_{a_t} Q(s_t, a_t)\right) \nabla_\phi f_\phi(\epsilon_t, s_t) \tag{16}$$

IV. Experimentos

Para implementar el entorno de la microrred descrito en la sección II y poder evaluar el rendimiento del algoritmo SAC propuesto, se utilizó la herramienta OpenAI Gym [14]. La simulación ejecuta varios días de gestión energética en los

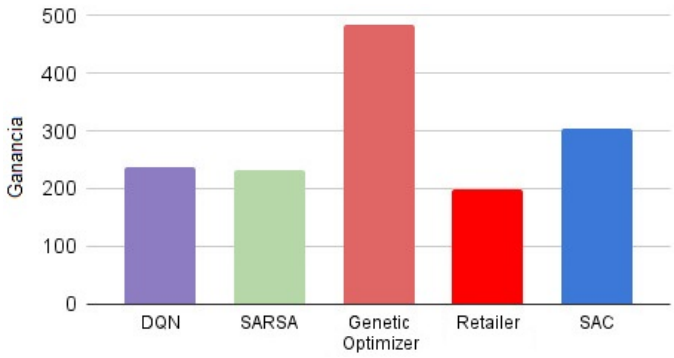


Fig. 2. Ganancia total acumulada por los algoritmos DRL y los proveedores ideales.

que cada episodio corresponde a 1 día. Cada paso de tiempo t representa el paso de una hora, por lo que tenemos 24 pasos en un día. Para cada episodio, se eligió arbitrariamente uno de los 10 primeros días del set de datos.

En primera instancia, el estado actual de la microrred s_t se formuló a partir del estado de carga promedio (SoC) de las TCL, el estado de carga del almacenamiento de energía BSC_t, el contador de precios C_t^b, la temperatura T_t, la generación de energía G_t, los precios de la electricidad en el mercado de regulación P_t^u, la hora del día t y el valor de carga actual del patrón de consumo diario $L_{b,t}$. Estas variables se organizaron mediante la representación vectorial:

$$s_t = [SoC_t, BSC_t, C_t^b, T_t, G_t, P_t^u, L_{b,t}, t] \tag{17}$$

El segundo paso importante, es diseñar una función de recompensa que busque maximizar la ganancia económica. La recompensa R_t se calcula como el margen bruto de las operaciones, es decir, los ingresos obtenidos por la venta de electricidad a la mirrored y a la red externa, menos los costos relacionados con la generación de energía, las compras y la transmisión desde la red externa. Por lo tanto, la función de recompensa R_t fue definida como:

$$R_t = \text{Ingresos}_t - \text{Costos}_t \tag{18}$$

$$\text{Ingresos}_t = P_t \sum_{cargas} L_t^i - C_{gen} \sum_{TCLs} L_{TCL}^i u_{b,t}^i + P_t^b E_t^V \tag{19}$$

$$\text{Costos}_t = (P_t^a + C_{tr_{imp}}) E_t^C + C_{tr_{exp}} E_t^V \tag{20}$$

donde P_t^a y P_t^b son los precios de regulación a la alza y baja respectivamente, es decir, los precios a los que se vende y se compra energía de la red externa. A su vez, G_t, E_t^V y E_t^C representan la energía generada por las fuentes renovables, vendida a la red externa y comprada de la red externa. C_{gen} es el costo de generación de energía que se cobra a los clientes que poseen carga controlada. $u_{b,t}^i$ corresponde a la acción de encendido o apagado de los TCLs. 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.

Par evaluar la política de gestión hallada por el algoritmo SAC, comparamos los resultados con dos estrategias: 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) [15] y un SARSA [16], con el fin de comparar el rendimiento contra otros enfoques de aprendizaje profundo.

Los algoritmos de RL se corrieron en el entorno simulado y registramos las recompensas diarias promediadas durante 10 días, desde el día 50 al 59. Las ganancias totales de 10 días generadas por los algoritmos se muestran en la Fig. 2. Puede observarse que algoritmo SAC supera en rentabilidad media al resto de los algoritmos de aprendizaje profundo, incluso a la obtenida por el minorista. La Fig. 3 muestra la ganancia diaria estimada para las diferentes estrategias durante 10 días consecutivos, en los que la energía generada y los precios de la electricidad del día siguiente difieren significativamente. Puede verse nuevamente como el algoritmo SAC supera en rendimiento en varios días a las demás estrategias de aprendizaje. Si bien las estrategias óptima y minorista demuestran un comportamiento mas rentable, debe recordarse que ambas representan comportamientos óptimos teóricos, con conocimiento perfecto de la dinámica de la red tanto en el estado actual como en los sucesores.

Los resultados relacionados con la asignación de energía a los TCLs y el estado de carga del ESS se presentan en la Fig. 4. Podemos ver un comportamiento similar entre la estrategia óptima y la estrategia derivada del algoritmo SAC para el día 50 de simulación. Una leve distinción, es que el algoritmo SAC opta por acumular más energía en los TCLs mientras que el óptimo asigna más energía al sistema de almacenamiento ESS. Nuevamente, el algoritmo óptimo tiene conocimiento completo de la dinámica de la microrred, y por lo tanto, puede predecir tanto precios de compra y venta de la energía como su disponibilidad. Ambas estrategias deciden almacenar la mayor parte de la energía mientras existe disponibilidad debido a la generación renovable y no se observan picos de demanda.

La energía comprada a la red y la energía vendida a la red se muestran en la Fig. 5, tanto para la estrategia óptima como para la estrategia SAC propuesta. En la imagen también puede verse la curva de energía renovable generada, que está disponible no solo para almacenamiento sino también para ser vendida a la red. Puede verse un comportamiento similar en las curvas de intercambio de energía con la red para las dos estrategias, incluso en días de generación muy variable. Esto significa que la estrategia SAC aprende un comportamiento cercano al óptimo, y puede anticipar las fluctuaciones tanto en precios como en energía. Nótese como ambas toman la decisión de vender el exceso de energía cuando la demanda es baja y optan por comprar ante picos de consumo.

V. Conclusiones

En este trabajo, se evalúa el desarrollo de estrategias de gestión para una microrred con múltiples desafíos técnicos debido a la necesidad de coordinación entre los dispositivos de almacenamiento, la red principal, las cargas termostáticas y las cargas sensibles al precio, con el objetivo de garantizar una gestión óptima de los recursos.

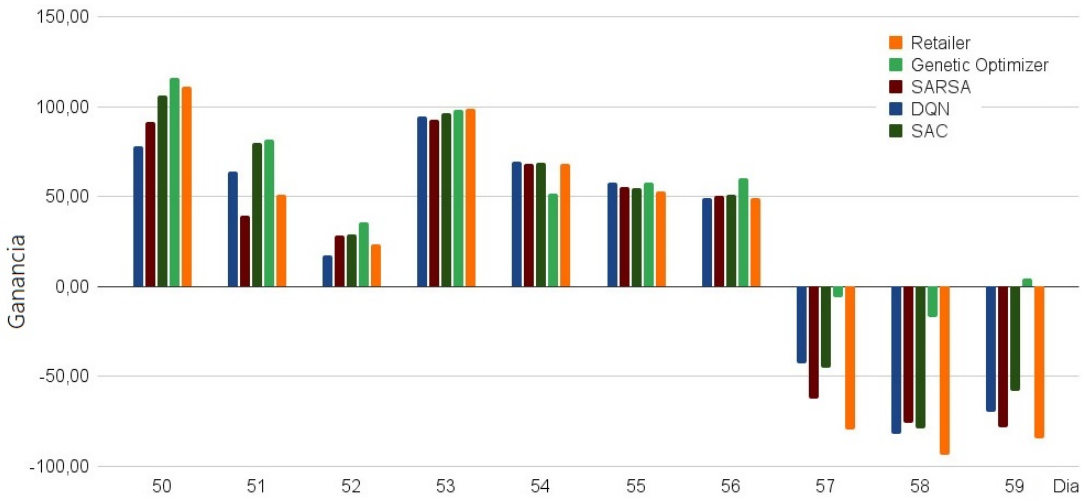
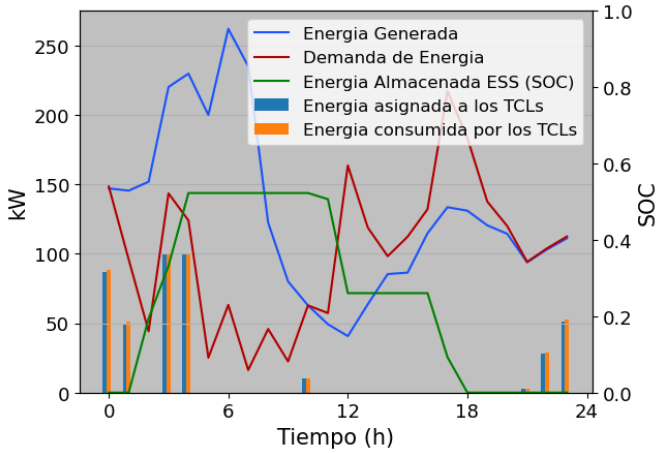
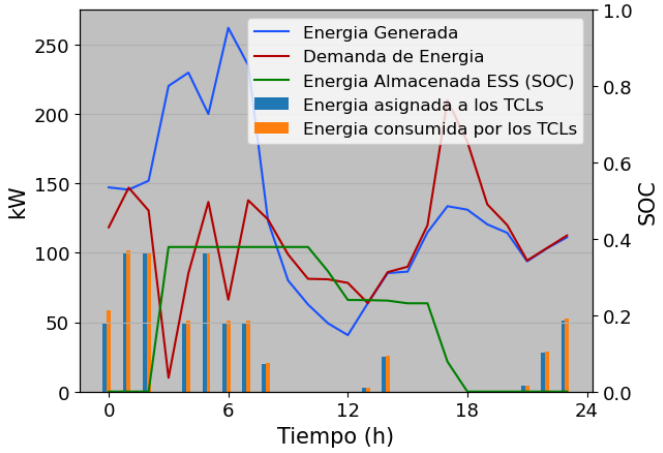


Fig. 3. Ganancia diaria obtenida por los algoritmos DRL y los proveedores ideales.



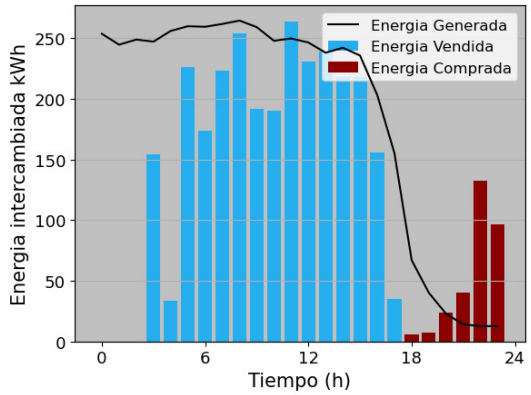
(a) Estrategia óptima



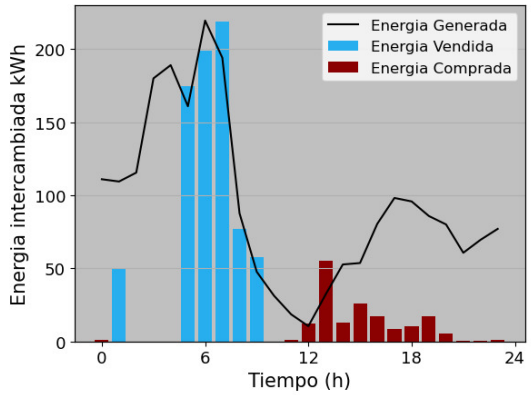
(b) Estrategia SAC

Fig. 4. Cantidad de energía generada y almacenada.

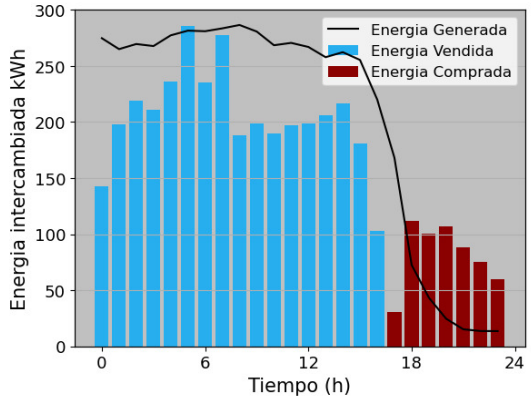
Dada la variabilidad e incertidumbre en la operación de los componentes de la microrred y la alta dimensionalidad de sus variables propusimos un algoritmo de aprendizaje por refuerzos SAC. Los resultados obtenidos por el algoritmo SAC se compararon con otras estrategias de gestión de la microrred, con un controlador óptimo teórico con perfecto conocimiento de la dinámica del sistema y con un minorista que compra electricidad en el mercado diario. La estrategia basada en SAC supera a las otras técnicas de aprendizaje, al minorista y obtiene más del 50 % del beneficio teórico óptimo. Los resultados del trabajo indican que podrían estudiarse distintos criterios para asignar precios a la energía demandada con el objetivo de influenciar a los consumidores a mover sus cargas a periodos sin picos, una posibilidad es emplear técnicas de aprendizaje basadas en multiagentes.



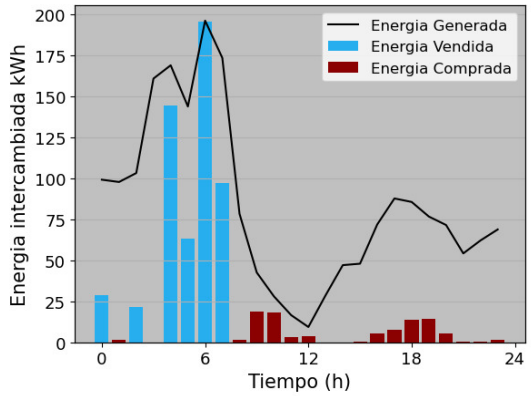
(a) Estrategia óptima para el día 50



(b) Estrategia óptima para el día 56



(c) Estrategia SAC para el día 50



(d) Estrategia SAC para el día 56

Fig. 5. Energía intercambiada con la red.

REFERENCIAS

[1] O. A. Omitaomu and H. Niu, "Artificial intelligence techniques in smart grid: A survey," *Smart Cities*, vol. 4, no. 2, pp. 548–568, 2021.

[2] R. S. Sutton, A. G. Barto, *et al.*, "Introduction to reinforcement learning," 1998.

[3] J. R. Vázquez-Canteli and Z. Nagy, "Reinforcement learning for demand response: A review of algorithms and modeling techniques," *Applied energy*, vol. 235, pp. 1072–1089, 2019.

[4] O. De Somer, A. Soares, K. Vanthournout, F. Spiessens, T. Kuijpers, and K. Vossen, "Using reinforcement learning for demand response of domestic hot water buffers," in *2017 IEEE PES Innovative Smart Grid Technologies Conference Europe*, pp. 1–7, 2017.

[5] M. Trimboli, L. Avila, and M. Rahmani-Andebili, "Reinforcement learning techniques for mppt control of pv system under climatic changes," in *Applications of Artificial Intelligence in Planning and Operation of Smart Grids*, pp. 31–73, Springer, 2022.

[6] J. R. Vazquez-Canteli, G. Henze, and Z. Nagy, "Marlisa: Multi-agent reinforcement learning with iterative sequential action selection for load shaping of grid-interactive connected buildings," in *Proceedings of the 7th ACM international conference on systems for energy-efficient buildings, cities, and transportation*, pp. 170–179, 2020.

[7] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*, pp. 1861–1870, 2018.

[8] C. Hu, Z. Cai, Y. Zhang, R. Yan, Y. Cai, and B. Cen, "A soft actor-critic deep reinforcement learning method for multi-timescale coordinated operation of microgrids," *Protection and Control of Modern Power Systems*, vol. 7, no. 1, pp. 1–10, 2022.

[9] T. A. Nakabi and P. Toivanen, "Deep reinforcement learning for energy management in a microgrid with flexible demand," *Sustainable Energy, Grids and Networks*, vol. 25, p. 100413, 2021.

[10] "Wind farm data," *Fortum Oy, Finland*, 2018.

[11] "Fingrid open datasets," 2018.

[12] T. A. Nakabi and P. Toivanen, "An ann-based model for learning individual customer behavior in response to electricity prices," *Sustainable Energy, Grids and Networks*, vol. 18, p. 100212, 2019.

[13] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[14] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," *arXiv preprint arXiv:1606.01540*, 2016.

[15] L. Yu, S. Qin, M. Zhang, C. Shen, T. Jiang, and X. Guan, "A review of deep reinforcement learning for smart building energy management," *IEEE Internet of Things Journal*, vol. 8, no. 15, pp. 12046–12063, 2021.

[16] D. Zhao, H. Wang, K. Shao, and Y. Zhu, "Deep reinforcement learning with experience replay based on sarsa," in *2016 IEEE symposium series on computational intelligence (SSCI)*, pp. 1–6, IEEE, 2016.

Bruno Boato is an advanced Mechatronic Engineering student at the National University of San Luis (UNSL), Argentina. He participates as a research student in the Computational Intelligence Research and Development Laboratory (LIDIC), in the area of intelligent systems for decision making.

Carolina Saavedra Sueldo is an Industrial Engineer since 2014 from the National University of the Center of the Province of Buenos Aires. Since 2019 she has been a PhD student in Engineering at the same University. As a CICpBA doctoral fellow, her research focuses on Industry 4.0 technologies combining simulation techniques and artificial intelligence for developing digital twins.

Luis Avila is an Electronic Engineer graduated from the National University of San Luis (UNSL), Argentina. He received his PhD in Engineering from the National Technological University (UTN-FRSF), Argentina. He is a researcher at the National Scientific and Technical Research Council of Argentina (CONICET) at the Computational Intelligence Research and Development Laboratory (LIDIC). He is a Professor at UNSL.

Mariano de Paula is an Industrial Engineer graduated from the National University of the Center of the Province of Buenos Aires, Argentina. He received a PhD in Engineering from the National Technological University (UTN-FRSF), Argentina. He is a researcher at the National Scientific and Technical Research Council of Argentina (CONICET) and carries out his activity in the INTELYMEC-UNCPBA. In addition, he is an Adjunct Professor at the UNCPBA Faculty of Engineering.