MPPT for PV Systems Using Deep Reinforcement Learning Algorithms

L. Avila, M. De Paula, I. Carlucho, and C. Sanchez

***Abstract*—This work proposes the use of reinforcement learning (RL) techniques with deep-learning models to address the maximum power point tracking (MPPT) control problem of a photovoltaic (PV) array. We implemented the deep deterministic policy gradient (DDPG) method, the inverted gradient (IGDDPG) method and the delayed twins (TD3) method to solve the MPPT control problem. Several simulation experiments were performed in the OpenAI Gym platform aiming to evaluate the performance of the proposed control strategies, under different operating conditions in terms of temperature and solar irradiance. The obtained results show that the use of deep reinforcement learning (DRL) achieves a successful performance for the MPPT control problem with a fast response and a stable behavior. Moreover, the algorithms do not require any previous knowledge about the dynamic behavior of the photovoltaic array.**

***Index Terms*—MPPT, Deep RL, PV systems, OpenAI Gym.**

I. Introducción

LOS arreglos fotovoltaicos (PV por sus siglas en inglés) constituyen el medio para convertir la energía solar en electricidad y una manera significativa de generar energía renovable y limpia. Para ser eficiente, un arreglo debe generar constantemente la máxima potencia posible bajo diferentes condiciones ambientales, problema conocido como seguimiento del punto de máxima potencia (MPPT por sus siglas en inglés).

Con el fin de optimizar el rendimiento de las aplicaciones fotovoltaicas, numerosos enfoques de control MPPT han sido propuestos en la literatura científica. Por un lado, los enfoques indirectos se basan en datos de curvas precalculadas de Potencia-Voltaje (*P-V*) para diferentes condiciones ambientales o en modelos matemáticos obtenidos de manera experimental. Entre los métodos indirectos más utilizados están los basados en el voltaje de circuito abierto (V_{oc}) y la corriente de cortocircuito (I_{SC}), donde los cálculos para obtener el punto de máxima potencia (MPP) se basan en los valores de estas variables [1], [2]. Como parte del mismo enfoque, los métodos basados en tablas de búsqueda comparan los valores de voltaje y corriente medidos con MPPs almacenados para determinadas condiciones ambientales, haciendo uso de modelos numéricos para aproximar el comportamiento de la fuente PV [3], [4]. La ventaja de estos métodos es su simpleza, pero no pueden adaptarse fácilmente a ningún cambio externo de la fuente PV.

Por otro lado, los métodos directos se basan en mediciones de corriente y voltaje y tienen la ventaja de ser independientes de la fuente empleada. Dentro de los métodos populares en esta categoría están los métodos de perturbar y observar (P&O) [5]; los métodos de conductancia incremental (IC) [6], [7]; los basados en lógica difusa [8], [9]; y los métodos que usan redes neuronales [10], [11]. Los métodos P&O e IC tienen la ventaja de ser de baja complejidad para su implementación a costa de fluctuaciones en los puntos de operación alrededor del MPP durante el estado estacionario y falta de robustez frente a cambios en las condiciones ambientales. Los enfoques basados en lógica difusa y redes neuronales son más robustos pero dependen del conocimiento a priori, lo que incrementa la complejidad de la implementación.

Recientemente, algunos trabajos han hecho uso de técnicas de aprendizaje por refuerzo (RL por su sigla en inglés) buscando mitigar los problemas mencionados en el control MPPT [12], [13]. El aprendizaje por refuerzo reúne numerosos algoritmos que aprenden mediante la interacción con el entorno cómo lograr un objetivo complejo o maximizar a lo largo de una dimensión particular el valor acumulado de las recompensas obtenidas (*rewards*) [14].

Con el fin de reducir el costo computacional de los algoritmos de RL, a menudo, se emplea un espacio de acciones discretas que se pueden aplicar a una fuente PV para generar un cambio en la operación del sistema. En el caso de un problema de control de MPPT, una acción se considera como un cambio en el valor de tensión para afectar la potencia fotovoltaica entregada en la salida. Por ejemplo, Hsu *et al.* [13] propone un esquema basado en solo cuatro estados, que se definen de acuerdo al lado y la dirección del movimiento del punto de operación con respecto al MPP. Del mismo modo, se propusieron cuatro acciones posibles, dos de ellas se aplican para cambios positivos y las otras dos para cambios negativos del ciclo de operación del arreglo. Para mantener el método computacionalmente eficiente, en Kofinas *et al.* [12] se define una lista de acciones finitas y discretas que deben incluir cambios positivos y negativos en búsqueda de garantizar resolución suficiente en el espacio de acciones como para lograr la máxima potencia.

No obstante, para obtener una alta precisión en el valor de voltaje de salida y maximizar la eficiencia del control MPPT se debe trabajar con una lista de acciones que sea continua. Sin embargo, uno de los principales obstáculos para las formulaciones de RL reside en el manejo de aplicaciones en espacios continuos de estado/acción cuando se requiere el uso de aproximadores de función para estimar la política

Luis Avila pertenece al Laboratorio de Investigación y Desarrollo en Inteligencia Computacional, CONICET-UNSL, Av. Ejército de los Andes 950, San Luis, Argentina.

Mariano De Paula e Ignacio Carlucho pertenecen al grupo INTELYMEC, Centro de Investigaciones en Fìsica e Ingeniería del Centro CIFICEN – UNICEN – CICpBA – CONICET, 7400 Olavarría, Argentina.

Carlos Sanchez Reinoso pertenece al CONICET y al grupo de Investigación y Desarrollo en Modelado, Optimización y Control (MOC), Facultad de Tecnología y Ciencias Aplicadas, Universidad Nacional de Catamarca, Maximio Victoria 55, Catamarca, Argentina.

de control y las funciones de valor [15], [16]. A menudo, los aproximadores lineales no son adecuados para sistemas complejos y por lo tanto se requieren aproximadores de funciones no lineales como son, por ejemplo, las redes neuronales artificiales (ANN). Sin embargo, la falta de linealidad en los modelos ANN puede causar inestabilidades en los algoritmos RL o incluso llevar todo el sistema a divergir. A partir de la creciente popularidad de los algoritmos de entrenamiento para redes neuronales profundas (DNN) [17], [18], Mnih et al. [19] introdujo la técnica deep Q-Network (DQN) que emplea redes neuronales convolucionales (CNN), para aproximar la función de valor para las acciones, logrando estabilizar el proceso de entrenamiento de la red. A partir de esta contribución, las técnicas de aprendizaje por refuerzos profundo (DRL) han emergido como un campo de investigación moderno y se han convertido en un foco atractivo y prometedor para el desarrollo de estrategias de control adaptativo en tiempo real para sistemas autónomos. Sin embargo, el algoritmo DQN solo puede aplicarse a sistemas discretos, es decir, sistemas con espacios de estados/acciones finitos y discretos. Recientemente, Lillicrap et al. [20] extendieron las formulaciones de DRL a dominios continuos, por medio del algoritmo de gradiente de política determinista profundo (DDPG) que incorpora las ideas de normalización por lotes [21] y repetición de experiencias [19]. Siguiendo esta línea, nuevos enfoques al problema de control en espacios continuos han sido propuestos [22], [23].

Este trabajo propone evaluar la factibilidad en la aplicación de técnicas de DRL con espacios estado/acción continuos al problema de control de MPPT en sistemas PV. Con este fin, se emplea la plataforma OpenAI Gym que proporciona un entorno computacional estandarizado para la prueba de técnicas de RL. OpenAI Gym tiene como objetivo reunir los algoritmos de referencia en un paquete de software accesible que a su vez incluye una colección diversa de tareas conocidas como entornos. Los entornos deben estar versionados de una manera que asegure que los resultados sean significativos y reproducibles a medida que se actualiza el software. De esta manera, el concepto de entorno permite que el rendimiento de diferentes soluciones basadas en RL se puedan comparar directamente entre sí en un ambiente controlado y bien definido.

II. MPPT PARA SISTEMAS PV

El punto de operación de un arreglo PV se define como la potencia producida debido a la corriente I_{PV} y el voltaje V_{PV}. El MPP es el punto en el cual la potencia generada a partir de la fuente PV se maximiza. Cuando una carga se conecta a una fuente PV, el punto de operación y la potencia producida se definen por la resistencia de la carga eléctrica. Por ejemplo, si el valor resistivo de una carga eléctrica es igual a $R_L = V_{MPP}/I_{MPP}$, entonces el punto de operación coincidirá con el MPP y no hay necesidad de seguir el MPP. Cuando se conecta una carga resistiva diferente, el punto de operación será diferente al MPP. En este caso, la fuente PV no produce la máxima potencia posible y se deben aplicar acciones de control para seguir el MPP.

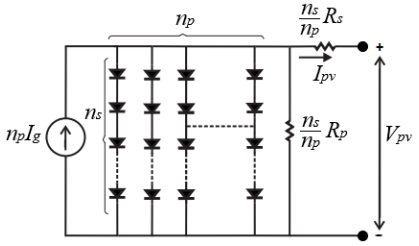

Fig. 1. Circuito equivalente de una celda PV.

A. Modelo de una Celda PV

Un arreglo fotovoltaica PV puede representarse analíticamente por su característica eléctrica de corriente I_{pv} contra voltaje V_{pv}. La Fig. 1 muestra el circuito equivalente de una celda representado como un diodo de unión p-n. Denotando el número de celdas PV en serie y en paralelo como n_s y n_p, respectivamente, la corriente I_{pv} de salida se escribe como la diferencia entre la fotocorriente generada I_g y la de corriente del diodo I_s

$$I_{pv} = n_p I_g - n_p I_s \left(\exp\left[\frac{q}{AkT} \left(\frac{V_{pv}}{n_s} + \frac{I_{pv} R_S}{n_p} \right) \right] - 1 \right), \tag{1}$$

donde A es el factor de idealidad del diodo, k es la constante de Boltzmann, q es la carga de electrones, T es la temperatura en Kelvin, R_S es la resistencia en serie equivalente.

Por su parte, la fotocorriente I_g generada mediante irradiación solar I_{rr} es

$$I_g = \left(I_{sc} + k_i \left(T - T_{ref} \right) \right) \frac{I_{rr}}{1000}, \tag{2}$$

donde I_{sc} es la corriente de cortocircuito a temperatura y radiación de referencia, T_{ref} es la temperatura de referencia de la celda y k_i es el coeficiente de temperatura para la corriente de cortocircuito.

La corriente de saturación I_s de la celda PV varía con la temperatura de acuerdo con la siguiente relación:

$$I_s = I_{RS} \left[\frac{T}{T_{ref}} \right]^3 \exp\left[\frac{qE_g}{Ak} \left(\frac{1}{T_{ref}} - \frac{1}{T} \right) \right], \tag{3}$$

donde I_{RS} es la corriente de saturación inversa y E_g es la energía de la banda prohibida del semiconductor. Finalmente, la potencia P_{pv} entregada por el panel se calcula como:

$$P_{pv} = I_{pv} V_{pv}. \tag{4}$$

B. Problema de Control del MPPT

Las relaciones (1)-(3) muestran claramente la dependencia del modelo de las condiciones de radiación solar y temperatura. A partir de esta representación matemática, la curva de potencia asociada al arreglo PV se obtiene expresando la conexión en serie y en paralelo de todos las celdas fotovoltaicas. Suponiendo que se considera un arreglo que consta de celdas idénticas bajo irradiación solar uniforme, la curva P-V característica tendrá un único pico, como se muestra en

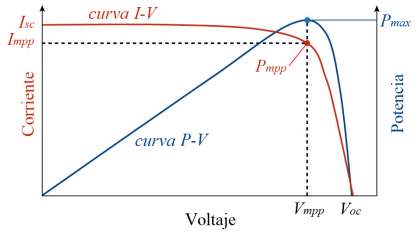

Fig. 2. Curva P-V para diferentes irradiancias solares.

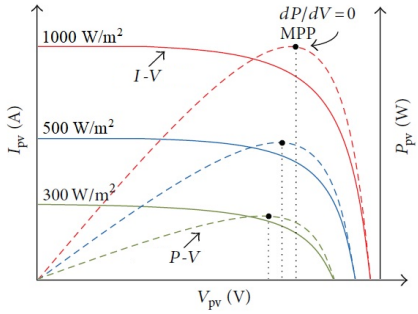

Fig. 3. Arreglo PV con n_s celdas en serie y n_p celdas en paralelo.

la Fig. 2. La figura ilustra la curva característica (*I-V*) para el voltaje de circuito abierto (V_{oc}), la corriente de cortocircuito (I_{sc}) y el punto de operación de máxima potencia de la celda solar. El MPP es entonces el punto único en esta curva donde la potencia generada de la fuente PV se maximiza. Si aumenta la potencia requerida por la carga, el punto de operación se moverá a la izquierda del MPP mientras que si se reduce el punto de operación se moverá a la derecha del MPP. Por lo tanto, el MPP debe ser rastreado de manera continua.

Por otro lado, la curva *I-V* de la fuente varía sus características de acuerdo a las condiciones ambientales. Una representación de la operación de una célula solar bajo diferentes irradiaciones se muestra en la Fig. 3. Puede verse que el valor de salida que corresponde al MPP a una irradiación de $1000\,[\mathrm{W/m^2}]$ no coincide con el MPP a $500\,[\mathrm{W/m^2}]$. Del mismo modo, un cambio en la temperatura afectará la potencia entregada por las celdas. El voltaje de salida depende en gran medida de la temperatura y un aumento en la temperatura disminuirá el valor de V_{pv}.

Dado que el control MPPT se puede lograr regulando el voltaje de salida del sistema V_{pv}, este se considera como la variable de optimización. Favorablemente, el MPP coincide con el punto en el cual la derivada de la potencia P_{pv} con respecto al voltaje V_{pv} es cero, esto es:

$$\frac{dP_{pv}}{dV_{pv}} = 0. \tag{5}$$

III. Metodología

Este trabajo propone un enfoque RL para resolver el problema del control del MPPT en un arreglo PV. Un enfoque RL permite resolver el problema sin conocer el comportamiento de la fuente PV o predefinir la dinámica de la misma. El algoritmo RL tiene como objetivo aprender el comportamiento del sistema o su configuración óptima en función de la respuesta de las interacciones con la fuente PV. Para implementar un enfoque de aprendizaje de refuerzo en la operación del arreglo PV, se debe definir un modelo Markoviano de su comportamiento.

A. Aprendizaje por Refuerzo

El aprendizaje por refuerzo supone que hay un agente situado en un entorno y en cada interacción con el mismo el agente toma una acción e influye sobre el entorno recibiendo una recompensa en forma de señal numérica. Un algoritmo de RL buscará entonces 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) [14].

Comúnmente, en las formulaciones RL el problema de control está definido por cuatro elementos, a saber, el espacio de estado $\mathbb{X}$, el espacio de acción $\mathbb{U}$, la probabilidad de transición de estado p y la función de recompensa r. Para el problema de control MPPT, en el tiempo t se toma una acción que corresponde a un valor u_t para la variable manipulada V_{pv}. Durante el proceso de aprendizaje, el agente interactúa con el sistema aplicando una acción $u_t =\in \mathbb{U} \subseteq \mathbb{R}^{n_u}$ y, después de eso, el sistema evoluciona desde el estado $\mathbf{x}_t =\in \mathbb{X} \subseteq \mathbb{R}^{n_x}$ a un estado sucesor $\mathbf{x}_{t+1}$ y el agente recibe una señal numérica r_t llamada recompensa (o castigo) que proporciona una medida de cuán buena (o mala) fue la acción u_t elegida y aplicada en el sistema en el estado $\mathbf{x}_t$ en términos de la transición de estado ocurrida $\mathbf{x}_{t+1}$. 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:

$$J^* = \max_{\pi} J_\pi = \max_{\pi} \mathrm{E}_\pi \{R_t \mid \mathbf{x}_t = \mathbf{x}\} \tag{6}$$

donde J_π es la recompensa total esperada bajo la política de control π.

B. Aprendizaje Profundo por Refuerzos

Uno de los mayores desafíos que han enfrentado desde sus comienzos las técnicas de RL ha sido cómo lidiar con espacios de acciones continuas. Si se discretiza demasiado el espacio de acciones, se termina con un problema de dimensionalidad. Pero una discretización insuficiente del espacio de acción podría despreciar información valiosa sobre la geometría del dominio de acciones. Consecuentemente, los algoritmos RL han estado limitados a entornos de cuadrícula pequeños y discretos, lo que les resta factibilidad en su aplicación para la mayoría de los sistemas dinámicos.

El éxito y la rápida aceptación del enfoque Deep Q-Networks [19], generó una expansión del estudio e implementación de técnicas de RL para resolver problemas de alta dimensionalidad dentro del área de control de sistemas dinámicos [20], [24]. Basándose en el trabajo previo presentado por Silver *et al.* [25], en relación a los gradientes de políticas determinísticas, Lillicrap *et al.* [20] desarrollaron un enfoque actor-crítico denominado gradiente de la política

determinística profundo (DDPG) que tiene la característica de ser *off-policy* y *model-free*.

Los enfoques RL basados en el gradiente de la política utilizan un método iterativo en el cual evalúan la política y luego siguen su gradiente para maximizar el rendimiento. DDPG utiliza una política estocástica para lograr una buena exploración, pero estima una política objetivo determinística que es mucho más simple de aprender. Por otro lado, DDPG se basa en un enfoque actor-crítico por lo que utiliza dos modelos de redes neuronales profundas. Estas redes calculan la predicciones de la próxima acción para el estado actual y generan una señal de error de diferencia temporal (TD) en cada paso. La entrada a la red actor es el estado actual, y la salida es un único valor real que representa una acción elegida del espacio de acciones continuo. La salida de la red que modela al crítico es simplemente el valor Q estimado para el estado actual y la acción dada por el actor.

Si bien el algoritmo DDPG es capaz de alcanzar un buen desempeño, presenta un precario comportamiento con respecto al manejo de los hiperparámetros. Además, la función Q modelada tiende a sobreestimar drásticamente los valores de Q, lo que podría conducir a la no convergencia de la política. Por su parte, Twin Delayed DDPG (TD3) [26] es un algoritmo que soluciona este problema mediante tres modificaciones críticas: en primer lugar, TD3 aprende dos funciones Q en lugar de una (de ahí el termino *twin*) y utiliza el menor de los dos valores Q para computar la función de Bellman; en segundo lugar, TD3 actualiza la política con menos frecuencia que la función Q (aproximadamente una actualización de política por cada dos actualizaciones de la función Q); y por último, TD3 agrega algo de ruido a la acción con el fin de dificultar que la política aprenda los errores de la función Q. Estas modificaciones dan como resultado un rendimiento sustancialmente mejor frente a DDPG.

Mas recientemente, Hausknecht *et al.* [27] propusieron el uso de redes neuronales profundas en espacios estructurados (parametrizados) de acción continua para delimitar los gradientes del espacio de acción sugeridos por el crítico. Este enfoque, se centra en el aprendizaje de un pequeño conjunto de acciones discretas, cada una de las cuales está parametrizadas con variables continuas. Esto permite extender el uso de RL a la clase de Procesos de Decisión de Markov (MDP) con espacios de acciones continuas y parametrizadas. Por otro lado, este enfoque reduce los gradientes a medida que los hiperparámetros se acercan a los límites de sus rangos y se invierten si estos exceden el rango de valores (de ahí su denominación *inverted gradients*). De esta manera, este enfoque mantiene activamente los parámetros dentro de los límites, minimizando los problemas de sobreestimación.

C. Entorno MPPT en OpenAI Gym

En el aprendizaje por refuerzo, los agentes aprenden a maximizar las recompensas acumuladas durante la interacción con su entorno observando y tomando acciones en consecuencia. Estos entornos deben satisfacer la propiedad de Markov. OpenAI Gym es un marco estandarizado y abierto que proporciona entornos para entrenar agentes RL. OpenAI Gym se centra en el escenario episódico del aprendizaje por refuerzo, donde la experiencia del agente se divide en una serie de episodios. En cada episodio, el estado inicial del agente se muestrea aleatoriamente de una distribución, y la interacción continúa hasta que el entorno alcanza un estado terminal.

OpenAI Gym es una plataforma de código abierto implementada en Python en la que se pueden entrenar, probar y evaluar algoritmos de aprendizaje ya que proporciona una variedad de entornos de videojuegos y ambientes típicos de los problemas de control clásicos. Aquí describimos cómo la plataforma podría usarse como una herramienta de simulación y prueba para control fotovoltaico. Particularmente, en este trabajo desarrollamos un ambiente de simulación en Gym empleando el modelo expuesto en la Sección II. La ventaja de hacer nuestros desarrollos en la plataforma OpenAI Gym, se debe principalmente a que, además de hacer un aporte concreto a la comunidad de RL, permite comparar el desempeño entre diferentes técnicas de control, en este caso para el problema de seguimiento del punto de máxima potencia de un sistema fotovoltaico.

Un ambiente Gym consiste básicamente en cuatro funciones. La primera es la función de inicialización de la clase *"init"*, que además establece el estado inicial de nuestro problema RL. La segunda función es la de paso *"step"*, que recibe datos de la próxima acción y devuelve una lista de cuatro elementos: el estado siguiente, la recompensa resultante de la última acción, un valor Booleano que informa si el episodio actual ha concluido e información extra sobre el estado del sistema. La función de *"reset"* restablece el estado y otras variables del entorno al estado de inicio y la función de renderizado *"render"* que proporciona información relevante sobre el comportamiento de nuestro entorno hasta el momento. Una vez que el entorno MPPT está completo, podemos crear una instancia del mismo y probarlo con diferentes algoritmos RL [1].

1) Espacio de estados: En un problema de RL, el estado debe ser lo suficientemente descriptivo para incluir toda la información necesaria respecto la condición actual del sistema. Además, un MDP necesita cumplir con la propiedad de Markov, lo que significa que la probabilidad de transición de un estado a otro depende únicamente de la información de este estado y de la acción aplicada.

En el problema de control MPPT, la eficiencia de la tarea se define de acuerdo a qué tan lejos del MPP está funcionando un panel PV en condiciones ambientales específicas. En su mayoría, los enfoques RL aplicados al MPPT usan características cualitativas de las curvas *I-V* y *P-V*, como la tendencia del cambio de potencia y/o definen estados discretos [28], [13], [12]. Estos enfoques tienen la ventaja de formar un espacio de estado pequeño, pero son incapaces de describir la operación del arreglo PV bajo condiciones de funcionamiento variables. En este trabajo se define un espacio de estados continuo que corresponde a los valores actuales $[V_{pv}, P_{pv}, \Delta V_{pv}]$, la inclusión del valor ΔV_{pv} permite por un lado determinar en que lado del MPP esta trabajando el arreglo PV y por

[1]En el siguiente enlace se encuentra disponible el ambiente de simulación desarrollado: https://github.com/loavila/mppt-gym

otro vuelve Markoviano al sistema ya que permite definir si el algoritmo viene incrementando la tensión de salida o reduciéndola. La potencia fotovoltaica máxima generada es el resultado de la corriente y el voltaje generados, como se describe en la curva *I-V* presentada en la Fig. 2.

2) Espacio de acciones: El espacio de acciones aplicado al problema de control MPPT es continuo, por lo cual contiene todas las acciones que se pueden aplicar en un arreglo PV para generar un cambio en la operación del sistema. Si bien esto garantiza una alta precisión en la magnitud de la acción, requiere de técnicas de aprendizaje potentes para que el enfoque resulte computacionalmente eficiente. La acción del agente RLMPPT aquí se define como la perturbación deseada ΔV_{pv} aplicada a la variable controlable V_{pv}.

3) Función de recompensa: Para cada acción elegida por el agente y aplicada sobre el sistema, el mismo reacciona y evoluciona hacia un estado sucesor generando una respuesta en forma de recompensa que va desde el entorno hacia el agente.

Intuitivamente, la derivación de recompensa más simple pero efectiva podría ser un tipo de función de acertar o fallar [13]. Una vez que el par de señales observables (V_{pv}, P_{pv}) alcanza el punto más alto, es decir, el verdadero MPP de (V_{mpp}, P_{mpp}) otorga una recompensa positiva al agente RL; de lo contrario el agente RL recibe una recompensa nula o negativa. Una función de recompensa simple, otorga una capacidad de generalización y adaptabilidad para el modelo (*reward engineering*). Por este motivo, para prescindir de información a priori acerca del sistema, y consecuentemente de la región de máxima potenica, se define la siguiente función de recompensa:

$$r_t = \begin{cases} (P/c)^2, & \text{if } P > 0, \\ -10, & \text{if } P \leq 0, \end{cases} \tag{7}$$

dónde c es un factor de normalización. Como puede advertirse, con esta simple función la recompensa obtenida es directamente proporciónal a la potencia obtenida y no hace falta ningún conocimiento previo acerca del sistema para definirla lo que facilita el aprendizaje del agente.

IV. Experimentos

Para evaluar la aplicabilidad y el desempeño de los métodos de aprendizaje por refuerzo profundo, expuestos anteriormente, para resolver el problema de control de seguimiento del punto de máxima potencia en una instalación fotovoltaica, se han planteado un número de escenarios de prueba bajo diferentes condiciones de operación. Particularmente, en cada escenario se han simulado diferentes condiciones de temperatura e irradiancia solar, con el objetivo de comprobar el desempeño de los métodos de aprendizaje por refuerzo profundo en términos de rendimiento (potencia máxima), así como su efectividad para adaptarse a las variaciones en las condiciones climáticas.

Los ensayos se realizaron usando el simulador implementado en la plataforma Open AI Gym, descripto en la sección anterior. En la Fig.4 se presentan las curvas características del modelo implementado en Gym, donde se fija la temperatura en 25°C y se varía la irradiancia solar. Cada una de las curvas

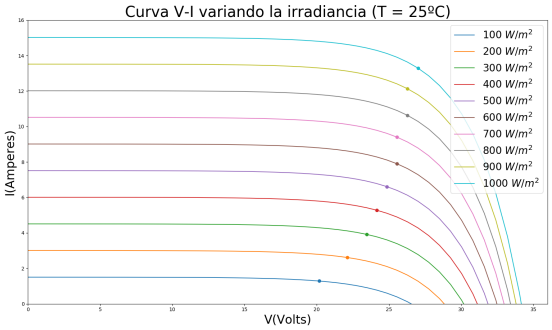

Fig. 4. Curvas de operación $V-I$ variando la irradiancia solar y manteniendo temperatura constante, T = 25ºC.

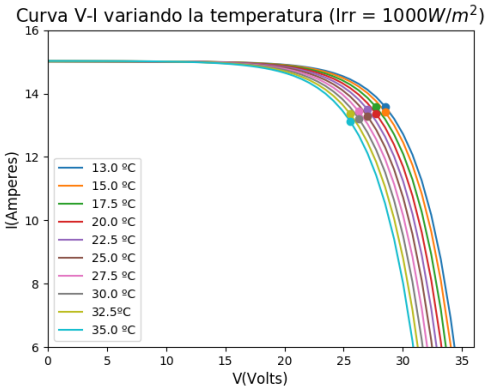

Fig. 5. Curvas de operación $V-I$ con irradiancia solar constante, $Irr = 1000\ W/m^2$, y variando la temperatura.

muestra como varía la relación voltage corriente *V-I* indicando con un círculo lleno, sobre cada una de ellas, el punto de máxima potencia correspondiente para los diferentes niveles de irradiancia solar. Análogamente, en la Fig. 5 se muestran las relaciones *V-I* para diferentes valores de temperatura, para un nivel nominal de irradiancia solar de 1,000 W/m^2. Como se puede ver en ambas figuras, el punto de máxima potencia varía cuando la temperatura y/o la irradiancia solar cambian.

1) Entrenamiento: Cada uno de los algoritmos fue implementado en lenguaje Python y entrenados durante 2,000 episodios. Cada episodio de entrenamiento consta de una duración máxima establecida de 100 segundos. Para todos los casos, se empleó una estrategia de exploración ϵ - $greedy$, con decaimiento lineal a lo largo de los episodios. Se estableció un tamaño del $buffer$ de repetición de experiencia (R) de 50,000 muestras ($|R|$ = 50,000) con un tamaño de lote de selección aleatorio de 64 muestras (M =64). El vector de estado característico del sistema está conformado por la tensión y la potencia en cada instante de muestreo, así como la ultima variación de tensión tal que $\mathbf{x}_t = [V_t, P_t, \Delta V_t]$; y (7) se utilizó como función de recompensa. Se utilizó el modelo fotovoltaico desarrollado en la Sección II, implementado en el ambiente Gym "mppt-v0" disponible en el repositorio GitHub https://github.com/loavila/mppt-gym.

La Fig. 6 muestra una comparación de la evolución en el aprendizaje entre los método de aprendizaje por refuerzos profundo comparados para resolver el problema de control de seguimiento del punto de máxima potencia. En dicha figura

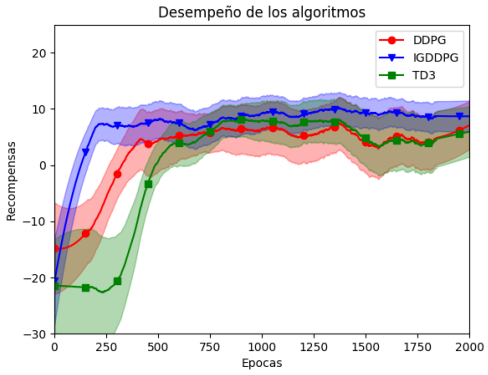

Fig. 6. Comparativa de recompensa media por episodio.

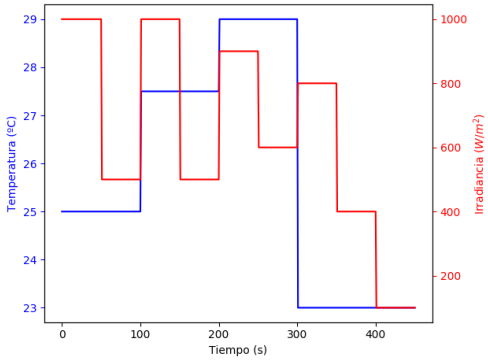

Fig. 7. Condiciones ambientales de temperatura e irradiancia solar.

se indica la evolución de la recompensa promedio lograda por cada algoritmo para 10 fases de entrenamiento independientes entre sí, cada una con una semilla de inicialización diferente (*random seed*). Además, en la Fig. 6 se encuentra sombreada la desviación media de la recompensa obtenida a lo largo de los episodios de entrenamiento. Como puede verse los tres métodos tienen un patrón de evolución similar, aunque el método IGDDPG tiene una leve ventaja sobre los otros dos.

2) Simulaciones y resultados: A continuación presentamos los resultados obtenidos para el seguimiento del punto de máxima potencia para los algoritmos de aprendizaje por refuerzo profundo presentados. Para ello, se expuso a estos sistemas de control a condiciones variables del ambiente, considerando los perfiles de temperatura e irradiancia solar indicados en la Fig. 7. Como puede verse fácilmente en este gráfico los cambios son bruscos y repentinos, lo cual impone una alta exigencia a los algoritmos de control, para adaptarse y encontrar el punto de máxima potencia lo más rápidamente posible. Cabe mencionar que los pares temperatura e irradiancia enfrentados no fueron vistos en la fase de entrenamiento.

En la Fig. 8 se indican los resultados obtenidos para cada uno de los algoritmos de control presentados. Como puede verse, todos ellos tienen una respuesta similar aunque presentan alguna diferencia de comportamiento. De este gráfico, puede advertirse que el algoritmo DDPG y el IGDDPG tienen un comportamiento muy similar entre sí, mientras que el algoritmo TD3 muestra un comportamiento ligeramente distinto, sobre todo en la última fase donde se tienen condiciones de baja temperatura e irradiancia solar.

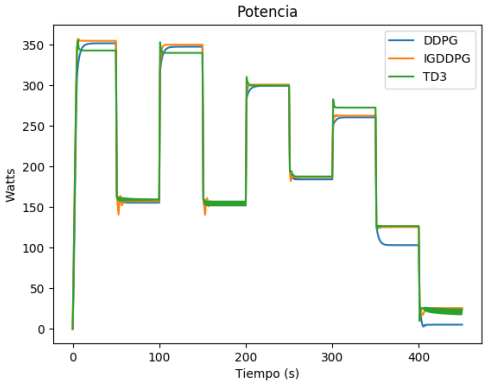

Fig. 8. Perfil de potencia alcanzado por cada algoritmo.

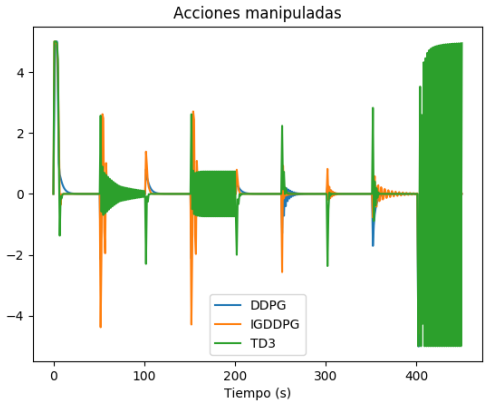

Fig. 9. Acciones de control de cada algoritmo.

La Fig. 9 muestra las acciones tomadas por cada uno de los algoritmos para seguir el punto de máxima potencia en las condiciones establecidas de temperatura e irradiancia solar (Fig. 7). Como puede verse en la Fig. 9 el algoritmo DDPG presenta un comportamiento más suave, en tanto que el IGDDPG es algo más reactivo frente a las variaciones ambientales. Por su parte, a pesar de que el algoritmo TD3 tiene un desempeño adecuado presenta un comportamiento más agresivo que los otros dos.

Por su característica intrínseca, el algoritmo TD3 tiene mayor número de hiperparámetros que el DDPG y el IGDDPG. Esto hace que este algoritmo pueda requerir mayor cantidad de episodios de entrenamiento cuando la inicialización de tales hiperparámentros diste demasiado de la óptima. En otras palabras, esto puede pensarse como que a igual inicialización y cantidad de episodios de entrenamiento, el algoritmo TD3 puede tomar más tiempo en alcanzar su desempeño óptimo. Sin embargo, los resultados alcanzados por el algoritmo TD3 son satisfactorios como se muestra en la Fig. 8.

La Tabla I presenta las condiciones de ensayo así como los valores teóricos de máxima potencia (P^*) para las diferentes condiciones de temperatura e irradiancia. En las últimas dos columnas se indican los valores teóricos óptimos de tensión (V^*) y corriente (I^*) para el punto de máxima potencia (P^*).

La Tabla II muestra el resumen de los resultados obtenidos para cada uno de los algoritmos. Como puede verse, todos los

TABLA I
PUNTOS MÁXIMOS TEÓRICOS PARA DIFERENTES PARES DE TEMPERATURA E IRRADIANCIA SOLAR

Temperatura	Irradiancia (W/m^2)	Pmax*	V*	I*
25,0	1000,0	359,00	26,79	13,39
25,0	500,0	164,22	24,75	6,63
27,5	1000,0	353,11	26,56	13,29
27,5	500,0	161,19	24,31	6,63
29,0	900,0	310,40	25,87	11,99
29,0	600,0	196,11	24,75	7,92
23,0	800,0	283,11	26,56	10,65
23,0	400,0	129,42	24,53	5,27
23,0	100,0	26,39	20,50	1,28

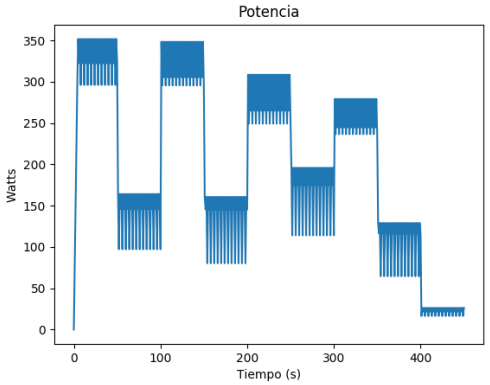

Fig. 10. Perfil de potencia alcanzado por el método P&O.

algoritmos de aprendizaje por refuerzo profundo presentados tienen un desempeño más que favorable para ser empleados en sistemas de seguimiento del punto de máxima potencia en fuentes fotovoltaicas.

Finalmente, para comparar el desempeño de los algoritmos propuestos contra uno tradicional, en la Fig. 10 y la Fig. 11 se muestran los resultados obtenidos de aplicar el bien conocido método de perturbación y observación (P&O), considerando las mismas condiciones. Como puede verse en la Fig. 10, la potencia entregada usando P&O tiene un comportamiento más variable que los anteriores (Fig.8), lo cual es una situación no deseable y que siempre se busca mitigar. Por otra parte, como se ve en la Fig. 11 las acciones de control son mucho más agresivas que en el caso de los algoritmos presentados anteriormente, lo cual también implica más esfuerzo de control en detrimento de la performance del sistema de control.

V. CONCLUSIONES

Este trabajo presenta el modelado de un arreglo PV en términos de un MDP desarrollado en la plataforma libre OpenAI Gym. Esto es una ventaja distintiva para resolver el problema de control MPPT por medio de técnicas de RL con modelos profundos (*DRL*). Estas técnicas monitorean el estado actual del arreglo y ajustan la perturbación a la tensión de operación del mismo para alcanzar la máxima transferencia de potencia a la carga.

Otras de las ventajas de nuestra propuesta radica en que las técnicas empleada permiten trabajar con espacios de observación de alta dimensión y con una lista de acciones continua.

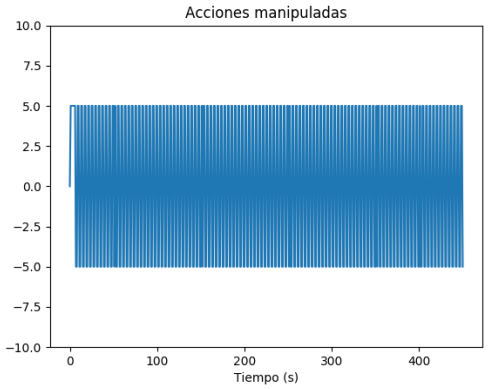

Fig. 11. Acciones manipuladas usando el método P&O.

Esto permite obtener una alta precisión en el valor de potencia final, a la vez que mantienen el espacio de acción manejable.

Los algoritmos se implementaron sin ninguna configuración adicional y la función de recompensa se definió de la manera mas simple posible para garantizar la capacidad de generalización y adaptabilidad de la política obtenida. Los algoritmos propuestos para el control MPPT fueron corridos bajo distintas condiciones de operación (variando temperatura e irradiancia) en el entorno OpenAI Gym con el fin de evaluar su eficiencia y validar su rendimiento.

Este trabajo reúne los avances recientes en el aprendizaje profundo y el aprendizaje por refuerzo, obteniendo resultados sólidos en un dominio con espacio de estado y acciones continuos. Si bien, el uso de aproximadores de función no lineales anula cualquier garantía de convergencia, los resultados experimentales demuestran que el aprendizaje en todos los casos es estable.

Como trabajo a futuro se propone incorporar más modelos al entorno OpenAI Gym creado, por ejemplo aquellos que incluyen condiciones de sombreado parcial, con el fin de añadir mayor complejidad en la tarea de aprendizaje. Además, se considera incorporar datos de condiciones ambientales reales con el fin de contar con perfiles de temperatura e irradiancia con una amplia variabilidad.

REFERENCIAS

[1] D. Baimel, S. Tapuchi, Y. Levron, and J. Belikov, "Improved fractional open circuit voltage mppt methods for pv systems," *Electronics*, vol. 8, no. 3, p. 321, 2019.

[2] S. Hadji, J.-P. Gaubert, and F. Krim, "Maximum power point tracking (mppt) for photovoltaic systems using open circuit voltage and short circuit current," in *3rd international conference on systems and control*. IEEE, 2013, pp. 87–92.

[3] K. Tsang and W. Chan, "Model based rapid maximum power point tracking for photovoltaic systems," *Energy conversion and management*, vol. 70, pp. 83–89, 2013.

[4] P. Bhatnagar and R. Nema, "Maximum power point tracking control techniques: State-of-the-art in photovoltaic applications," *Renewable and Sustainable Energy Reviews*, vol. 23, pp. 224–241, 2013.

[5] P. Jain, S. N. Joshi, N. Gupta, and K. G. Sharma, "Analysis of mppt techniques in grid connected pv system," in *2018 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE)*. IEEE, 2018, pp. 1–4.

[6] R. I. Putri, S. Wibowo, and M. Rifa'i, "Maximum power point tracking for photovoltaic using incremental conductance method," *Energy Procedia*, vol. 68, pp. 22–30, 2015.

TABLA II
RESUMEN DE RESULTADOS

Temperatura	Irradiancia (W/m^2)	P*			MPP						Error		
					DDPG		IGDDPG		TD3				
		DDPG	IGDDPG	TD3	Vmpp	Impp	Vmpp	Impp	Vmpp	Impp	DDPG	IGDDPG	TD3
25,0	1000,0	351,09	357,23	355,19	24.98	14.08	25.97	13.75	25.53	13.91	2,20%	0,49%	1,06%
25,0	500,0	164,18	163,88	163,34	24.98	6.57	24.25	6.75	23.91	6.83	0,03%	0,21%	0,54%
27,5	1000,0	347,61	349,97	353,09	24.85	13.98	25.27	13.84	26.52	13.32	1,56%	0,89%	0,01%
27,5	500,0	160,93	161,19	160,85	24.85	6.47	24.34	6.62	23.85	6.74	0,16%	0,00%	0,21%
29,0	900,0	299,47	300,90	310,39	23.42	12.79	23.61	12.74	26.03	11.92	3,52%	3,06%	0,01%
29,0	600,0	193,93	194,49	194,52	23.42	8.28	23.61	8.23	25.68	7.57	1,11%	0,83%	0,81%
23,0	800,0	260,58	263,29	282,92	22.40	11.63	22.71	11.59	26.80	10.55	7,96%	7,00%	0,07%
23,0	400,0	126,01	126,69	129,13	22.40	5.62	22.64	5.59	23.88	5.41	2,64%	2,12%	0,23%
23,0	100,0	24,71	25,90	26,33	17.43	1.42	18.92	1.37	19.91	1.32	6,37%	1,86%	0,23%

[7] K. Ishaque, Z. Salam, and G. Lauss, "The performance of perturb and observe and incremental conductance maximum power point tracking method under dynamic weather conditions," *Applied Energy*, vol. 119, pp. 228–236, 2014.

[8] A. I. Dounis, P. Kofinas, C. Alafodimos, and D. Tseles, "Adaptive fuzzy gain scheduling pid controller for maximum power point tracking of photovoltaic system," *Renewable energy*, vol. 60, pp. 202–214, 2013.

[9] K. Loukil, H. Abbes, H. Abid, M. Abid, and A. Toumi, "Design and implementation of reconfigurable mppt fuzzy controller for photovoltaic systems," *Ain Shams Engineering Journal*, 2019.

[10] A. I. Dounis, P. Kofinas, G. Papadakis, and C. Alafodimos, "A direct adaptive neural control for maximum power point tracking of photovoltaic system," *Solar Energy*, vol. 115, pp. 145–165, 2015.

[11] M. Arjun and J. Zubin, "Artificial neural network based hybrid mppt for photovoltaic modules," in *2018 International CET Conference on Control, Communication, and Computing*. IEEE, 2018, pp. 140–145.

[12] P. Kofinas, S. Doltsinis, A. Dounis, and G. Vouros, "A reinforcement learning approach for mppt control method of photovoltaic sources," *Renewable Energy*, vol. 108, pp. 461–473, 2017.

[13] R. C. Hsu, C.-T. Liu, W.-Y. Chen, H.-I. Hsieh, and H.-L. Wang, "A reinforcement learning-based maximum power point tracking method for photovoltaic array," *International Journal of Photoenergy*, vol. 2015, 2015.

[14] R. S. Sutton, A. G. Barto *et al.*, *Introduction to reinforcement learning*. MIT press Cambridge, 1998, vol. 2, no. 4.

[15] M. Riedmiller, "Neural fitted q iteration–first experiences with a data efficient neural reinforcement learning method," in *European Conference on Machine Learning*. Springer, 2005, pp. 317–328.

[16] T. Degris, P. M. Pilarski, and R. S. Sutton, "Model-free reinforcement learning with continuous action in practice," in *2012 American Control Conference (ACC)*. IEEE, 2012, pp. 2177–2182.

[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[19] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.

[20] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[22] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International conference on machine learning*, 2015, pp. 1889–1897.

[23] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[24] I. Carlucho, M. De Paula, S. Wang, Y. Petillot, and G. G. Acosta, "Adaptive low-level control of autonomous underwater vehicles using deep reinforcement learning," *Robotics and Autonomous Systems*, vol. 107, pp. 71–86, 2018.

[25] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic Policy Gradient Algorithms," *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 387–395, 2014.

[26] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," *arXiv preprint arXiv:1802.09477*, 2018.

[27] M. Hausknecht and P. Stone, "Deep reinforcement learning in parameterized action space," *arXiv preprint arXiv:1511.04143*, 2015.

[28] A. Youssef, M. E. Telbany, and A. Zekry, "Reinforcement learning for online maximum power point tracking control," *Journal of Clean Energy Technologies*, vol. 4, no. 4, 2016.

Luis Avila es Ingeniero Electrónico graduado en la Universidad Nacional de San Luis (UNSL), Argentina. Doctor en Ingeniería en la Universidad Tecnológica Nacional (UTN-FRSF), Argentina. Investigador del Concejo Nacional de Investigaciones Científicas y Técnicas de Argentina (CONICET) en el LIDIC en Monitoreo inteligente del comportamiento de agentes autónomos. Profesor en la Facultad de Ingeniería de la UNSL.

Mariano De Paula es Ingeniero Industrial graduado en la Universidad Nacional del Centro de la Provincia de Buenos Aires (UNCPBA), Argentina. Doctor en Ingeniería en la Universidad Tecnológica Nacional (UTN-FRSF), Argentina. Investigador del Concejo Nacional de Investigaciones Científicas y Técnicas de Argentina (CONICET) en el grupo INTELYMEC en Control cognitivo inteligente aplicado a la robótica. Profesor en la Facultad de Ingeniería de la UNCPBA.

Ignacio Carlucho es ingeniero Electromecánico graduado de la Universidad Nacional del Centro, Argetina. Doctorando en Ingenieria, como becario CONICET, dentro del grupo INTELYMEC. Sus áreas de investigación son el control inteligente, más específicamente el aprendizaje por refuerzos aplicado a robótica subacuática.

Carlos Sanchez Reinoso es Ingeniero Electrónico y Doctor en Ciencias de la Ingeniería. Investigador del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina. Profesor en la Facultad de Tecnología y Ciencias Aplicadas de la Universidad Nacional de Catamarca, Argentina, e Investigador del Grupo en Modelado, Optimización y Control (MOC). Sus investigaciones abordan métodos de inteligencia computacional para el modelado y optimización de sistemas de energía.