Gestión segura de la batería de EVs utilizando aprendizaje por refuerzo

Safe EVs battery management using reinforcement learning

Maximiliano Trimboli[1], Nicolás Antonelli[1], Luis Avila[1*] y Mariano de Paula[2]

[1]Laboratorio de Sistemas Inteligentes, CONICET-UNSL
D5730EKQ, San Luis, Argentina
[2]Centro de Investigaciones en Física e Ingeniería del Centro -UNICEN-CICpBA-CONICET, INTELYMEC
Olavarría, B7400JWI, Argentina.
*loavila@unsl.edu.ar

PALABRAS CLAVE:

Safe-RL, SOC, envejecimiento de la batería, variabilidad

RESUMEN

Las baterías de iones de litio son el dispositivo de alimentación estándar de los vehículos eléctricos (EVs) como alternativa de elección para reducir las emisiones de CO2. Pero antes de convertirse en una tecnología fiable, las baterías de iones de litio deben hacer frente a dos grandes retos: las reacciones electroquímicas indeseables debidas a tasas de carga excesivas y el tiempo considerable que tarda un EV en cargarse. Por tanto, es necesario emplear perfiles de corriente equilibrados que eviten tanto los graves efectos de degradación de la batería como las molestias a los usuarios finales. En este trabajo, los autores proponen un enfoque de aprendizaje profundo por refuerzo de exploración segura (SDRL) para determinar los perfiles de carga óptimos en condiciones de funcionamiento variables. Una de las principales ventajas de las técnicas de RL es que pueden aprender de la interacción con el sistema simulado o real incorporando la no linealidad y la incertidumbre derivadas de las condiciones ambientales fluctuantes. Sin embargo, como las técnicas RL tienen que explorar estados indeseables antes de obtener una política óptima, no ofrecen garantías de seguridad. El enfoque propuesto pretende mantener cero violaciones de las restricciones a lo largo de todo el proceso de aprendizaje incorporando una capa de seguridad que corrige la acción si es probable que se viole una restricción. El método propuesto se prueba en el circuito equivalente de una batería de iones de litio considerando condiciones de variabilidad. Los primeros resultados muestran que SDRL es capaz de encontrar políticas de carga optimizadas y seguras teniendo en cuenta un compromiso entre la velocidad de carga y la vida útil de la batería.

KEYWORDS:

Safe-RL, SOC, battery aging, variability

ABSTRACT

Lithium-ion batteries are the standard power source for electric vehicles (EVs) as an alternative of choice to reduce CO2 emissions. But before becoming a reliable technology, lithium-ion batteries must face two major challenges: undesirable electrochemical reactions due to excessive charging rates and the considerable time it takes for an EV to charge. Therefore, it is necessary to use balanced current profiles that avoid both the serious effects of battery degradation and inconvenience to end users. In this work, the authors propose a safe scanning deep reinforcement learning (SDRL) approach to determine optimal load profiles under varying operating conditions. One of the main advantages of RL techniques is that they can learn from the interaction with the simulated or real system by incorporating nonlinearity and uncertainty arising from fluctuating environmental conditions. However, since RL techniques have to explore undesirable states before obtaining an optimal policy, they do not offer security guarantees. The proposed approach aims to maintain zero constraint violations throughout the entire learning process by incorporating a security layer that corrects the action if a constraint is likely to be violated. The proposed method is tested on the equivalent circuit of a lithium-ion battery considering variability conditions. First results show that SDRL is able to find optimized and safe charging policies taking into account a trade-off between charging speed and battery life.

• **Recibido: 26 de junio de 2023** • **Aceptado: 20 de noviembre de 2023** • **Publicado en línea: 1 de febrero de 2024**

ProgMat, 2024, 16, 1; https://progmat.uaem.mx/progmat
DOI: 10.30973/progmat/2024.16.1/4

1. INTRODUCCIÓN

Las baterías de iones de litio son la clave para suministrar energía de forma limpia y segura a los sistemas móviles en general y son fundamentales para hacer frente a los problemas medioambientales actuales [1]. No obstante, mientras los EVs ganan popularidad rápidamente, las baterías de iones de litio deben hacer frente a dos retos principales: las molestias de los usuarios finales y el envejecimiento de la batería [2]. En la actualidad, los conductores están acostumbrados a llenar el depósito de gasolina en cuestión de minutos, pero cargar un vehículo eléctrico (EV) suele llevar más tiempo, dependiendo del tamaño y las especificaciones de la batería. Esto se debe principalmente a la limitación de la fuente de alimentación con una tasa de carga más baja. Por otro lado, una tasa de carga alta puede generar una estructura inestable y problemas cuando las partículas de litio se introducen rápidamente en el ánodo. Cargar las baterías a altas velocidades también puede provocar un sobrecalentamiento, lo que afecta negativamente su estado de salud a lo largo de múltiples ciclos. Es evidente que existe una necesidad de encontrar un equilibrio entre la temperatura del núcleo de la batería y el tiempo de carga [3]. Por lo tanto, los algoritmos que determinan los perfiles de carga desempeñan un papel fundamental en el rendimiento final de la batería.

Existe una amplia literatura sobre el problema de la carga en el menor tiempo posible, desde estrategias simples como la corriente constante-tensión constante (CC/CV), hasta enfoques más creativos como los algoritmos de carga multietapa. Sin embargo, existen menos estudios que aborden el desafío de encontrar perfiles de carga óptimos que maximicen la vida útil de la batería. En el trabajo de [2], se presentó un enfoque de aprendizaje para encontrar una estrategia de carga rápida, mientras que en [4] se utilizó un esquema de control predictivo basado en modelos para diseñar estrategias de carga rápida que también tuvieran en cuenta la salud de la batería. Otros métodos implementaron estrategias multietapa CC-CV para prolongar la vida útil de la batería [5], [6]. Sin embargo, dado que la mayoría de estas estrategias se basan en heurísticas para encontrar la estrategia de carga, no pueden garantizar la optimalidad ni cumplir plenamente con las restricciones de seguridad.

Las técnicas de aprendizaje por refuerzo (RL) permiten obtener políticas de control óptimas sin depender de un modelo detallado del sistema que se está controlando. RL es un enfoque basado en objetivos que utiliza la experiencia para aprender a realizar tareas complejas. Para una política dada, cada estado concreto tiene un valor asociado que depende de la utilidad futura que se espera alcanzar desde ese estado. Varios trabajos han empleado la RL para abordar el problema de encontrar una política de carga óptima. Por ejemplo, en [7] los autores proponen un enfoque basado en RL para minimizar los costes de carga y mejorar la eficiencia del sistema global. En [8] el autor utiliza una metodología de carga dinámica basada en RL con múltiples modos activos para abordar el problema de extender la vida útil de la batería. En [5], un estudio presenta una estrategia de carga óptima que utiliza el aprendizaje por refuerzo (RL) para prolongar la vida útil de la batería. Este enfoque permite a los usuarios finales personalizar su tiempo de carga en función de sus circunstancias individuales. En [6] se presenta una técnica de carga adaptativa para baterías de iones de litio que utiliza RL y corriente constante multietapa. Por último, [9] presentó una estrategia de carga rápida basada en un marco actor-crítico de gradiente-política para baterías de iones de litio. No obstante, todas estas estrategias requieren imperiosamente de una interacción con el sistema, con lo cual, el costo del aprendizaje podría ser dañino cuando se ensayan acciones que conducen a situaciones dañinas.

El aprendizaje por refuerzo seguro (SRL) es un campo en crecimiento que aborda

problemas de aprendizaje donde es crucial que el agente interactúe con el entorno solo a través de políticas que eviten situaciones no deseadas [10]. Una forma de abordar el problema del SRL es guiar el proceso de aprendizaje restringiendo el conjunto de políticas explorables [11]. Esto se ha logrado mediante el modelado de recompensas [12] [13], agregando restricciones al problema de optimización de políticas [14][15], o mediante el asesoramiento de expertos [16][17]. Sin embargo, estos enfoques no pueden garantizar la seguridad antes de un período de aprendizaje adecuado, ya que la seguridad también se aprende a través de la interacción con el entorno utilizando la política óptima. Un enfoque innovador consiste en agregar directamente una señal de seguridad a la política aprendida para realizar correcciones en las acciones hacia los límites de seguridad [18]. Una ventaja de esta técnica es que proporciona una solución cerrada mediante un modelo linealizado aprendido a partir de trayectorias pasadas generadas con acciones aleatorias.

En este trabajo, proponemos utilizar esta aproximación para abordar el problema de encontrar estrategias de carga rápida basadas en la temperatura del núcleo de la batería. Nuestro objetivo es mantener cero violaciones de restricciones a lo largo del proceso de aprendizaje utilizando una red neuronal preentrenada que predice los cambios en una señal de seguridad en un solo paso de tiempo. El modelo entrenado se incorpora en una capa de seguridad que corrige la acción si existe una alta probabilidad de infringir una restricción. El método propuesto se prueba en un circuito equivalente de una batería de iones de litio, teniendo en cuenta las condiciones de variabilidad.

2. MODELO DE BATERÍA LI-ION

Para adquirir la dinámica eléctrica bajo diferentes condiciones de operación en función de mantener cierto equilibrio entre precisión y complejidad, se propone un modelo formado por dos pares resistencia-capacitor (RC) [19]. Además, dada la alta probabilidad de divergencia entre la temperatura superficial y la temperatura del núcleo bajo altas tasas de corriente, se incorpora un modelo de temperatura de dos estados para obtener predicciones más precisas [20].

2.1. Modelo de circuito equivalente con envejecimiento de la batería

El funcionamiento del subsistema eléctrico, mostrado en la Fig. 1, se basa en el siguiente modelo analítico de espacio de estados:

$$\frac{dSOC(t)}{dt}=\frac{\eta}{C_{bat}}I(t) \tag{1}$$

$$\frac{dV_1(t)}{dt}=\frac{-V_1(t)}{R_1C_1}+\frac{I(t)}{C_1} \tag{2}$$

$$\frac{dV_2(t)}{dt}=\frac{-V_2(t)}{R_2C_2}+\frac{I(t)}{C_2} \tag{3}$$

$$V_T=V_{OC}(SOC)+V_1(t)+V_2(t)+R_0I(t) \tag{4}$$

donde *SOC* representa el estado de carga de la batería, η es la eficiencia de Coulomb, C_{bat} indica la capacidad nominal, V_T expresa la tensión en bornes y V_{OC} es la tensión en circuito abierto, que es una función no lineal del SOC. Los estados analizan la derivada con respecto al tiempo del SOC y las tensiones ($V_1(t)$, $V_2(t)$) a través de los pares RC.

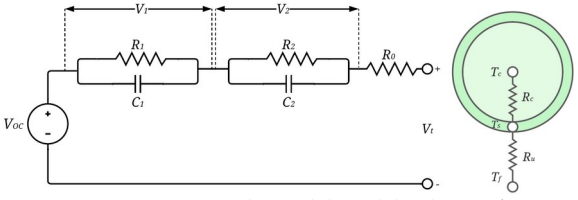


Figura 1. Circuito equivalente del modelo electrotérmico.

Mientras tanto, el funcionamiento térmico viene dado por

$$\frac{dT_c(t)}{dt}=\frac{T_s(t)-T_c(t)}{R_cC_c}+\frac{Q(t)}{C_c} \tag{5}$$

$$\frac{dT_s(t)}{dt}=\frac{T_f(t)-T_s(t)}{R_uC_s}+\frac{T_s(t)-T_c(t)}{R_cC_s} \tag{6}$$

donde $Q(t)=|I(V_{oc}-V_t)|$ representa la generación de calor incluyendo el calentamiento joule y la energía disipada por los sobrepotenciales de los electrodos. R_u y R_c son las resistencias de convección y de conducción del calor. C_s y C_c simbolizan la conducción de calor de la superficie y el núcleo, respectivamente. Los estados ocultos son las temperaturas del núcleo y de la superficie, T_c y T_s respectivamente, mientras que T_f es la temperatura ambiente. Sin embargo, el modelo de circuito equivalente utiliza el valor medio de las temperaturas del núcleo y de la superficie, denominado T_m. Todos los parámetros de los modelos eléctrico y térmico se han ajustado según [21]. El modelo semi-empírico de vida útil [1] adoptó la siguiente ecuación para expresar la correlación entre la pérdida de capacidad (ΔQ_b, en %) y el rendimiento A_h descargado, donde A depende de la tasa C,

$$\Delta Q_b = M(c)\exp\left(\frac{-E_a(c)}{RT_c}\right)A(c)^z \tag{7}$$

donde $M(c)$ representa el factor pre-exponencial dependiente de la tasa C, representado en la ecuación como c. La energía de activación E_a y el factor de ley de potencia z vienen dados por $E_a(c)=31700-370.3c$ y $z=0.55$.

En este modelo, el final de la vida útil (EOL, por sus siglas en inglés) de una batería de automóvil se define por una pérdida de capacidad del 20%. Por lo tanto, el rendimiento A_h correspondiente A_{tol} y el número de ciclos N se describen mediante

$$A_{tol}(c,T_c)=\left[\frac{20}{M(c)\exp\left(\frac{-E_a(c)}{RT_c}\right)}\right]^{\frac{1}{z}} \tag{8}$$

$$N(c,T_c)=\frac{3600\,A_{tol}(c,T_c)}{C_{bat}} \tag{9}$$

Cada ciclo corresponde a un rendimiento de carga de $2C_{bat}$, y dado que A_{tol} es el rendimiento de descarga de A_h, el rendimiento total, incluyendo tanto la carga como la descarga de A_h, debería ser de $2A_{tol}$. Teniendo esto en cuenta, la siguiente ecuación define el estado de salud (SOH) de la batería

$$SOH(t)=SOH(t_0)-\frac{\int_{t_0}^{t} I(\tau)\,d\tau}{2N(c,T_c)C_{bat}} \tag{10}$$

La última ecuación muestra una dependencia explícita entre el proceso de envejecimiento de la batería con la temperatura del núcleo y el perfil de corriente de carga empleado.

2.2 Variabilidad en el estado de la batería

Las condiciones de funcionamiento cambian con la temperatura cuando están sometidas a grandes variaciones de temperatura.

Una alternativa práctica y sencilla para describir el comportamiento fluctuante de las baterías causado por la temperatura es mediante la utilización de un proceso estocástico de difusión. Para ello se incluye un parámetro de escala de ruido σ en el modelo térmico para describir la dinámica de transferencia de calor radial de una batería cilíndrica considerando la dinámica de las temperaturas del núcleo y de la superficie T_c y T_s dada la Fig. 1,

$$\frac{dT_c(t)}{dt}=\frac{T_s(t)-T_c(t)}{R_cC_c}+\frac{Q(t)}{C_c}+\sigma dw \tag{11}$$

Para representar el funcionamiento de la batería bajo diferentes niveles de variabilidad, en la Fig. 2 se obtienen cuatro conjuntos de curvas de temperatura para diferentes escalas del parámetro de ruido σ aumentando de 0 (comportamiento determinista) a 6 mientras que el resto de parámetros del modelo de batería no varían. Obsérvese que, puesto que el SOC no depende directamente de la temperatura (véase Ec. 1), y la corriente de carga $I(t)$ es constante para este ejemplo, presenta un comportamiento lineal.

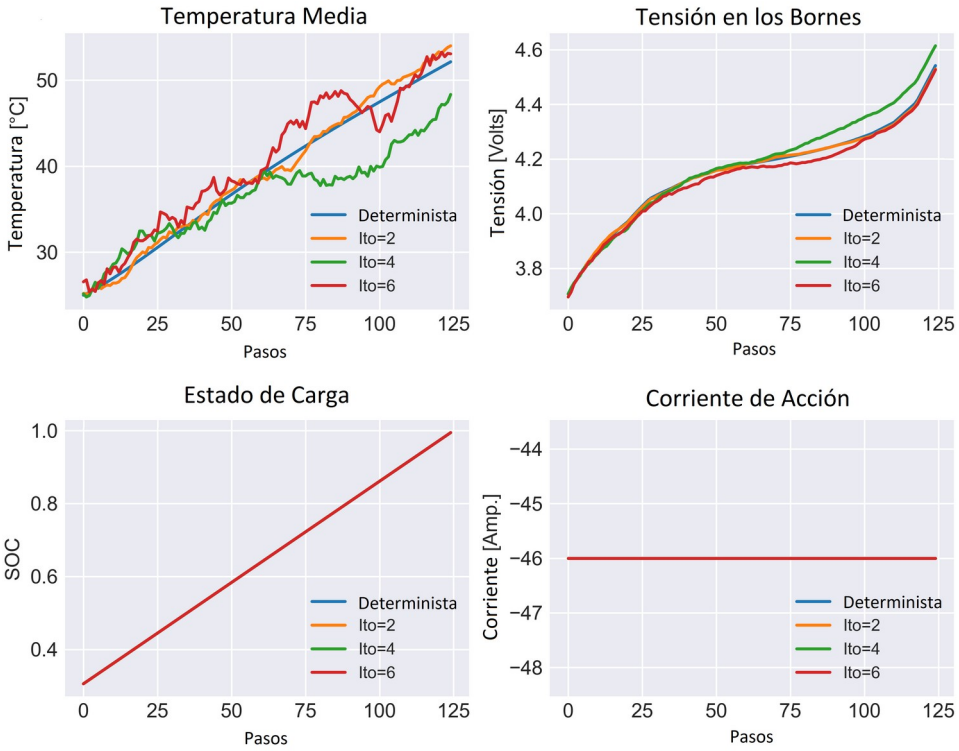


Figura 2. Comportamiento del modelo ante variación de escala Ito.

El objetivo de incluir un proceso estocástico es capturar una representación analítica de la variabilidad potencial entre celdas de batería reales, donde cada escala de ruido dada representa variaciones en todos los parámetros eléctricos debidas al cambio de las condiciones térmicas y de envejecimiento.

3. APRENDIZAJE POR REFUERZO SEGURO

3.1. Proceso de decisión de Markov restringido

Estudiamos un caso especial de procesos de decisión de Markov restringidos (CMDP, por sus siglas en inglés) [22] en el que las señales de seguridad observadas deben mantenerse acotadas. Un CMDP se caracteriza por la tupla (S, A, P, R, γ, C), donde S es un espacio de estados, A es un espacio de acciones, $P: S \times A \times S \to [0;1]$ es una función de transición, $R: S \times A \to R$ es una función de recompensa, $\gamma \in (0;1)$ es el factor de descuento, y $C = \{c_i: S \times A \to R \mid i \in [K]\}$ es un conjunto de funciones de restricción inmediata, dado el conjunto K formado por $\{1, \ldots, K\}$. Basándonos en esto, definimos también un conjunto de señales de seguridad $\overline{C} = \{\bar{c}_i: S \to R \mid i \in [K]\}$ como observaciones por estado de los valores de las restricciones inmediatas. Por último, la política $\mu: S \to A$ es un mapeo estacionario de estados a acciones. Investigamos el concepto de exploración segura en el marco de la optimización de políticas. En este contexto, para cada estado, nos aseguramos de que todas las señales de seguridad, denotadas como $\bar{c}_i(\cdot)$, están limitadas por sus correspondientes límites superiores $C_i \in R$:

$$\max_{\theta} E[\sum_{t=0}^{\infty} \gamma^t R(s_t, \mu_\theta(s_t))] \tag{12}$$

$$s.t. \quad \bar{c}_i(s_t) \leqslant C_i \quad \forall i \in [K] \tag{13}$$

donde μ_θ es la política parametrizada.

3.2 Modelo lineal de la señal de seguridad

Encontrar una solución a la Ec. 13 es difícil, incluso para un modelo básico de la dinámica de la batería. El agente RL necesita

exploración para mejorar su política. Sin embargo, al comienzo del entrenamiento, no es posible cumplir las restricciones con una política aleatoria sin conocimiento previo del entorno. Esta afirmación se mantiene incluso cuando la recompensa penaliza los estados inseguros. Para que el agente de RL aprenda a evitar comportamientos no deseados, puede ser necesario violar las restricciones varias veces para que el impacto negativo se propague en la programación dinámica.

No nos centramos en aprender el modelo de transición completo, sino en capturar las funciones de restricción inmediata $c_i(s,a)$. Considerando $[x]^+$ como la operación $max\{x,0\}$; donde $x \in R$, realizamos una linealización para obtener una aproximación de primer orden a $c_i(s,a)$. con respecto a la acción a.

$$\bar{c}_i(s) \triangleq c_i(s,a) \approx \bar{c}_i(s) + g(s,w_i)^\top a \tag{14}$$

donde w_i son los pesos de una red neuronal $g(s,w_i)$, que toma s como entrada y tiene como salida un vector de la misma dimensión que a. Este modelo representa explícitamente cómo los cambios en la señal de seguridad se ven afectados por las acciones utilizando características de estado.

A partir de un conjunto de tuplas $D=(s_j,a_j,s_{j'})$ independientes de la política, entrenamos $g(s;w_i)$ resolviendo

$$\underset{w_i}{arg\,min} \sum_{(s,a,s') \in D} \left(\bar{c}_i(s') - \left(\bar{c}_i(s) + g(s,w_i)^\top a\right)\right)^2 \tag{15}$$

donde D se genera inicializando el agente en una ubicación uniformemente aleatoria para realizar acciones de características similares a lo largo de múltiples episodios que finalizan cuando expira un intervalo de tiempo o cuando se produce una violación de una restricción.

El entrenamiento de $g(s;w_i)$ en el conjunto de datos D se lleva a cabo como una fase de pre-entrenamiento antes del entrenamiento RL, y este proceso se realiza una vez para cada tarea.

3.3. Capa de seguridad mediante optimización analítica

Para resolver el problema (13) utilizamos el Gradiente de Política Determinista Profundo (DDPG, por sus siglas en inglés) [23], un algoritmo de gradiente de política [24] cuya red de política emite directamente una acción escalar en lugar de un vector estado-valor.

Utilizando la acción determinista $\mu_\theta(s)$ seleccionada por la red política profunda, utilizamos una capa adicional situada encima de la red política, cuya función es resolver

$$a^* = \underset{a}{arg\,min}\ \frac{1}{2}\left\|a - \mu_\theta(s)\right\|^2 \tag{16}$$

$$s.t.\ \ \bar{c}_i(s) + g(s,w_i)^\top a \le C_i \ \forall\, i \in [K] \tag{17}$$

Suponiendo que sólo una restricción está activa en un momento dado, podemos aprovechar esta condición para derivar una solución analítica de forma cerrada para la última ecuación. Por lo tanto, suponiendo la existencia de la mencionada solución de forma cerrada denotada por $(a^*, \{\lambda_i^*\}_{i=1}^K)$, donde λ_i^* es el multiplicador de Lagrange óptimo con respecto a la i-ésima restricción, y que $\left|\{i \mid \lambda_i^* > 0\}\right| \le 1$; es decir, a lo sumo una de las restricciones está activa, entonces

$$\lambda_i{}^* = \left[\frac{g(s;w_i)^\top \mu_\theta(s) \mid \bar{c}_i(s) - C_i}{g(s;w_i)^\top g(s;w_i)}\right]^+ \tag{18}$$

y

$$a^* = \mu_\theta(s) - \lambda_{i^*}^* g(s;w_{i^*}) \tag{19}$$

donde $i^* = arg\,max_i \lambda_i^*$

La solución (18) es básicamente una proyección lineal de la acción original $\mu_\theta(s)$ al hiperplano "seguro" cuya inclinación es $g(s;w_i^+)$ e intersección $\bar{c}_{i^*}(s) - C_{i^*}$.

4. CONFIGURACIÓN EXPERIMENTAL

Asumimos los siguientes parámetros para el modelo de batería de iones de litio: una capacidad nominal de celda $C_{bat}=2300\,Ah$, con una corriente de carga máxima $I(t)=46\,A$ y una tensión mínima de terminal $V_{t_{min}}=2\,V$. Inicialmente, las temperaturas del núcleo T_c y de la superficie T_s de la batería coinciden con la temperatura ambiente T_f. La corriente de carga corresponde a la acción elegida por el agente, sobre un espacio de acción continuo, y se aplica al entorno como se muestra en la Fig. 3. El signo negativo corresponde a una corriente de carga. Los límites de alcance para los espacios de estado y acción están definidos por los límites de funcionamiento seguro para una batería de iones de litio según los parámetros: Espacio de Estado S=[5,45]°C, Espacio de Acción A=[-46,0] amp, SOC Inicial SOC_0=0.3, SOH Inicial SOH_0=0.9, Temperatura Ambiente T_f=25°C. El estado y la acción se normalizan al rango [0,1] para aumentar la estabilidad durante el entrenamiento de los algoritmos. Todos los episodios se inicializaron en las mismas condiciones ambientales para facilitar la comparación.

Para evitar una manipulación excesiva durante el aprendizaje se usó la siguiente función de recompensa que depende del valor real del SOC:

$$R=\begin{cases} SOC-1 & si\ \ 0<SOC<1 \\ -1 & de\ \ otra\ \ forma \end{cases} \tag{20}$$

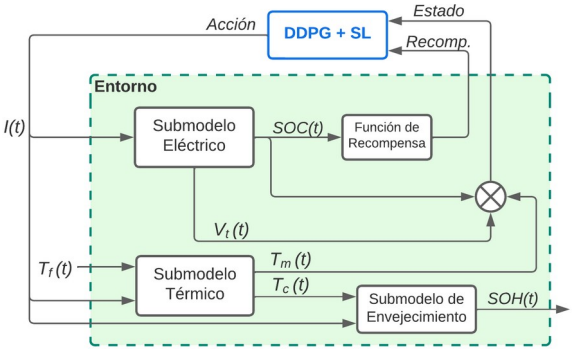


Figura 3. Esquema simplificado de interacción agente-entorno.

Como se muestra en la Fig. 4, el algoritmo DDPG consta de una red actor con su correspondiente red objetivo, ambas formadas por redes neuronales totalmente conectadas (FCNN, por sus siglas en inglés) de 128-64. A su vez, la red crítica y su red objetivo están formadas por 64-128-32 FCNN. Utilizamos el optimizador Adam con una tasa de aprendizaje de actor α_σ=0.00001 y una tasa de aprendizaje crítica α_ρ=0.0001. El aprendizaje del agente consiste en muchas épocas de entrenamiento, cada una de ellas seguida de una fase de evaluación después de un número fijo de pasos. Cada episodio finaliza cuando se alcanza una longitud de paso máxima o el agente alcanza el SOC máximo. Para obtener las señales de seguridad c_i, se realiza una época de pre-entrenamiento para calcular los multiplicadores de Lagrange y el término de corrección de la acción que son utilizados cada vez que la red actor le envía información sobre el estado y acción a la capa de seguridad. Todos los hiper-parámetros utilizados para el entrenamiento se muestran en la Tabla 1.

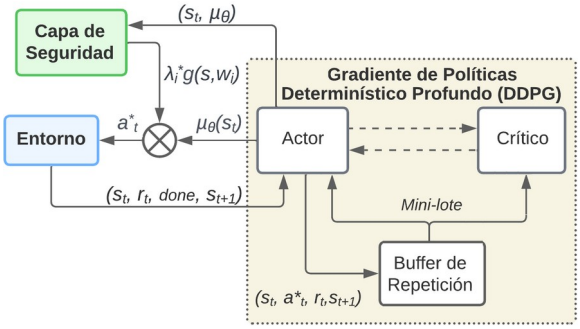


Figura 4. Esquema simplificado de la fase de aprendizaje.

Tabla 1. Principales hiperparámetros del algoritmo DDPG.

Símbolo	Denominación	Rango
σ	Capas de red actor	[128,64]
ϕ	Capas de red crítico	[64, 128, 32]
e_{DDPG}	Épocas	100
m_t	Pasos por época de entrenamiento	6000
m_e	Pasos por época de evaluación	1500
l_{max}	Duración máxima del episodio	300
B	Tamaño de lote	256
D	Tamaño de búfer de repetición	1000000
γ	Factor de descuento	0.99
α_σ	Tasa de aprendizaje del actor	0.00001
α_ρ	Tasa de aprendizaje del crítico	0.0001
ϵ	Rango de ruido de acción	0.01
	Optimizador	Adam
C	Capas del modelo de restricción	[2, 2]
e_{SL}	Épocas	5
α_{SL}	Tasa de aprendizaje del modelo de restricción	0.001

4.1. Resultados experimentales

Todas las pruebas experimentales fueron realizadas mediante Python utilizando diferentes librerías tanto para la programación de los algoritmos (Pytorch, por ejemplo) como para el modelado de la batería (OpenAI Gym) y la generación de curvas (Matplotlib), entre otras.

Para comparar la eficacia del método propuesto, aplicamos una técnica DDPG sin restricciones y una estrategia DDPG con conformación de recompensas. El modelado de recompensas manipula la señal de recompensa para proporcionar al agente un conocimiento experto que le permita evitar zonas no deseadas. En concreto, el agente es penalizado con una recompensa negativa cuando la temperatura T_m supera el margen de temperatura M. El margen M es el límite superior a partir del cual la política empieza a corregir sus acciones para que la temperatura no supere las restricciones. Para determinar el mejor valor para M, fijamos la penalización en r=-1 cuando el agente está por encima de M y realizamos series de 10 simulaciones con diferentes semillas para $M \in \{0.05, 0.09, 0.12, 0.15\}$. Se determinaron las violaciones de restricciones acumuladas, es decir, las veces que el agente está fuera de los límites de seguridad, para cada ejecución.

Se obtuvieron métricas relevantes durante la fase de aprendizaje, que se muestran en la Tabla 2. Comparamos para cada técnica el número de episodios y el número medio de pasos por episodio para converger. El número de violaciones representa el porcentaje del total de episodios en los que la temperatura supera el límite.

Observando el número medio de pasos necesarios para completar una carga de batería, podemos ver que DDPG carga más rápido la batería a costa de un mayor número de violaciones de las restricciones. Por su parte, la estrategia SDRL permite cargar ligeramente más rápida la batería que los métodos RS sin violar ninguna restricción. Como ya se ha dicho, las baterías de iones de litio son propensas a sobrecalentarse cuando la corriente de carga es elevada, lo que provoca su degradación a lo largo de varios ciclos. Podemos ver en la última fila el *SOH* resultante después del proceso de entrenamiento. SDRL supera con creces al resto de algoritmos, DDPG+RS con *M*=0.12 alcanza un 40% de degradación (teniendo en cuenta un SOH_0=0.9) pero con un tiempo de carga más largo mientras que DDPG sin restricciones alcanza un 50% de degradación del SOH.

Tabla 2. Información de carga de la batería después del entrenamiento.

	DDPG	DDPG+RS (M=0.05)	DDPG+RS (M=0.12)	DDPG+SL
Promedio de pasos para una carga completa	154.25	205.83	251	174.43
% de episodios con violación de restricciones	30.96%	9.09%	1.52%	0%
Valor SOH en la convergencia	0.3978	0.4872	0.5304	0.8602

Las Figs. 5-8 muestran las curvas obtenidas para diferentes políticas de carga. Se observan diferentes comportamientos en función de la limitación de temperatura. El algoritmo DDPG (Fig. 5) carga la batería lo más rápido posible utilizando la máxima corriente permitida, violando las restricciones. Debido a que la política no tiene en cuenta la temperatura, la corriente es casi constante, y *SOC*=1 se alcanza alrededor del paso 150. DDPG+RS con *M*=0.05, también utiliza la corriente máxima permitida para cargar la batería hasta que la temperatura alcanza el margen, como en la Fig. 6. A continuación, la política limita la corriente y aumenta a 200 el número de pasos necesarios para alcanzar *SOC*=1. Dado que DDPG+RS con *M*=0.12 utiliza un margen mayor, como se muestra en la Fig. 7, empieza a corregir el perfil de corriente antes. A expensas de un gran número de pasos para cargar la batería, podemos observar un perfil de corriente constante sin violaciones de las restricciones. La estrategia SDRL propuesta opera cerca de la restricción de temperatura pero sin violaciones, como se observa en la Fig. 8, mientras que el tiempo de carga se encuentra entre la política DDPG sin restricciones y los valores alcanzados con RS.

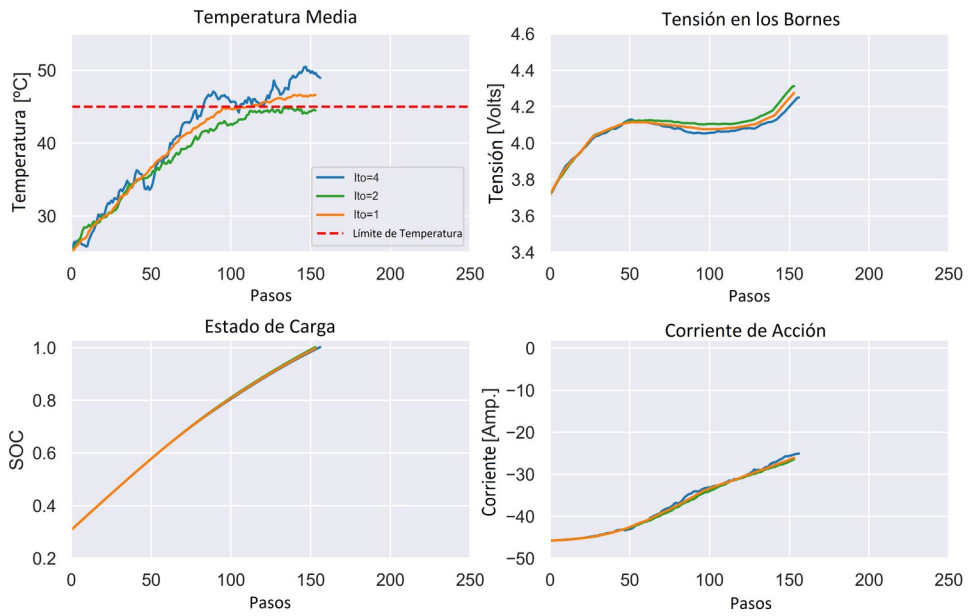


Figura 5. Proceso de carga de batería utilizando únicamente DDPG.

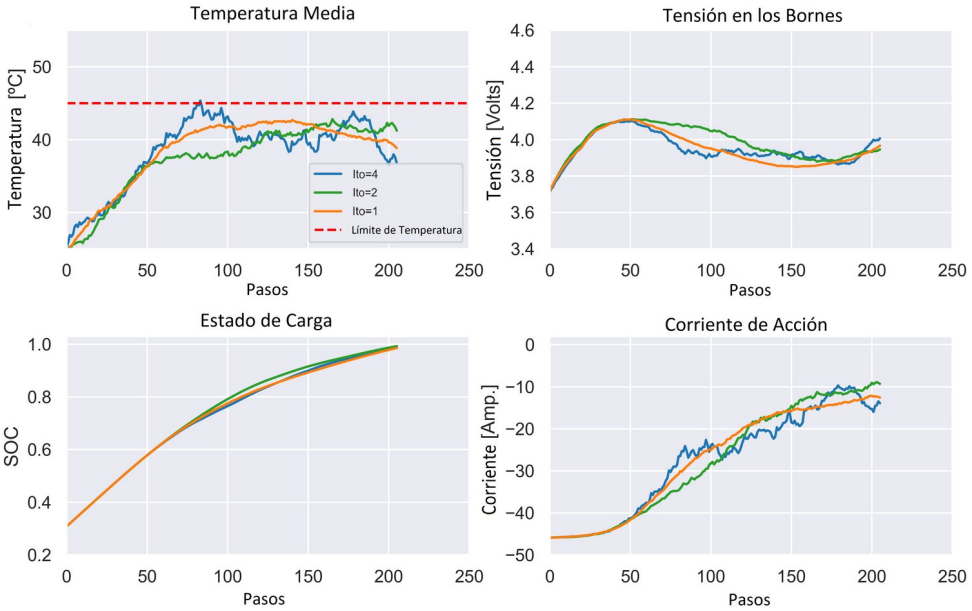


Figura 6. Proceso de carga de batería mediante DDPG+RS con margen M=0.05

5. CONCLUSIONES

El tiempo de carga de las baterías es una preocupación para los propietarios de vehículos de gasolina que consideran cambiar a vehículos eléctricos (EVs). Aunque cargar la batería de forma más rápida puede reducir el tiempo de espera, puede tener efectos electroquímicos no deseados. Por lo tanto, es importante encontrar perfiles de carga que mantengan la batería en un rango óptimo para maximizar su vida útil. En este trabajo, utilizamos un enfoque de SRL para obtener perfiles de carga de alta calidad para baterías Li-ion. Nuestro enfoque garantiza que nunca se violen las restricciones durante el proceso

de aprendizaje. Para ello, empleamos el algoritmo DDPG para calcular la política de carga y agregamos una capa de seguridad que resuelve de manera analítica una formulación de corrección de acciones para cada estado. Nuestra técnica proporciona una solución elegante y cerrada mediante el uso de un modelo linealizado aprendido a partir de trayectorias pasadas que incluyen acciones aleatorias.

Realizamos experimentos utilizando un circuito equivalente que simula la dinámica de la batería en diversas condiciones de funcionamiento y comparamos los resultados con métodos de referencia. En general, nuestro modelo propuesto demostró ser eficiente en términos de tiempo de carga, suavidad del perfil de corriente y mantenimiento de la vida útil de la batería.

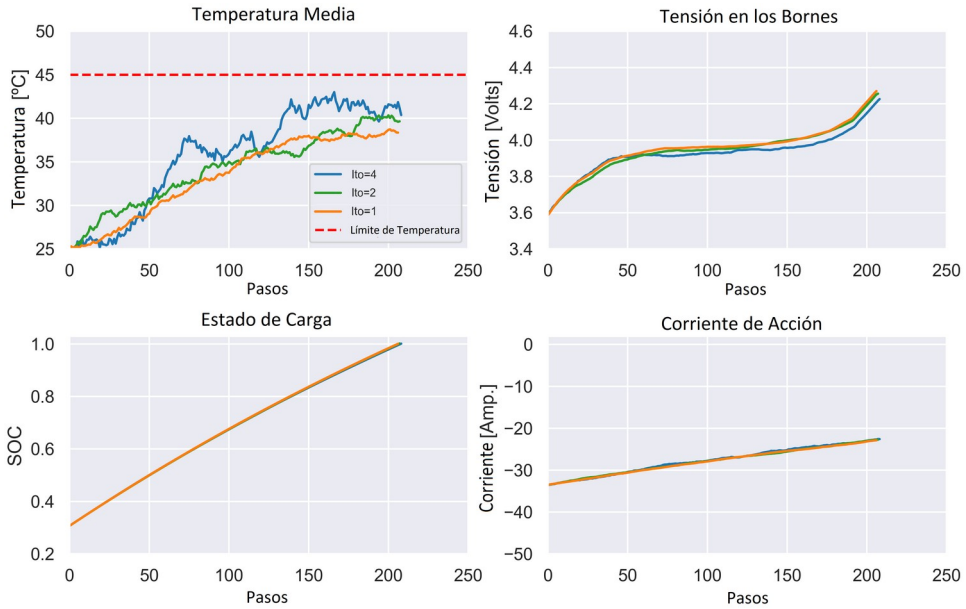


Figura 7. Proceso de carga de batería mediante DDPG+RS con margen M=0.12

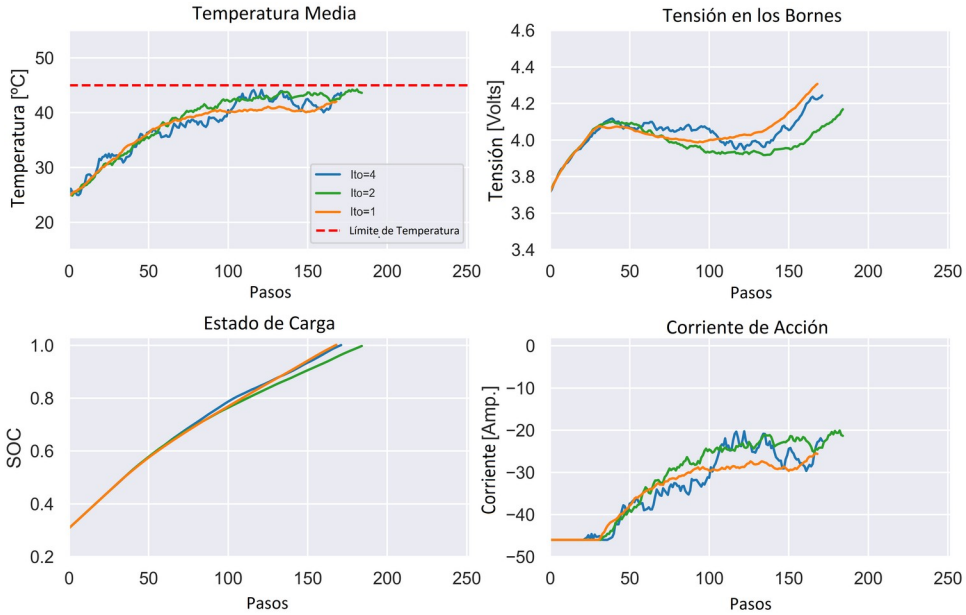


Figura 8. Proceso de carga de batería mediante DDPG con capa de seguridad (Método propuesto)

REFERENCIAS

[1] Campbell, I. D., Gopalakrishnan, K., Marinescu, M., Torchio, M., Offer, G. J., Raimondo, D. Optimising lithium-ion cell design for plug-in hybrid and battery electric vehicles. *Journal of Energy Storage*. 2019, 22, 228-238. doi: 10.1016/j.est.2019.01.006.

[2] Danilov, D., Notten, P. H. L. Adaptive battery management systems for the new generation of electrical vehicles. In 2009 *IEEE Vehicle Power and Propulsion Conference*, 2009, 317-320. doi: 10.1109/VPPC.2009.5289835.

[3] Xing, Y., Ma, E. W., Tsui, K. L., Pecht, M. Battery management systems in electric and hybrid vehicles. *Energies*. 2011, 4(11), 1840-1857. doi: 10.3390/en4111840.

[4] Yan, W., Zhang, B., Zhao, G., Weddington, J., Niu, G. Uncertainty management in Lebesgue-sampling-based diagnosis and prognosis for lithium-ion battery. *IEEE Transactions on Industrial Electronics*. 2017, 64(10), 8158-8166. doi: 10.1109/TIE.2017.2701790.

[5] Kim, M., Lim, J., Ham, K. S., Kim, T. Optimal charging method for effective Li-ion battery life extension based on reinforcement learning. In *Proc. of the 38th ACM/SIGAPP Symposium on Applied Computing*. 2023, 1659-1661. doi: 10.1145/3555776.3577800.

[6] Tunuguntla, S. T. Adaptive charging techniques for Li-ion battery using Reinforcement Learning (Doctoral dissertation), 2021.

[7] Chang, F., Chen, T., Su, W., Alsafasfeh, Q. Control of battery charging based on reinforcement learning and long short-term memory networks. *Computers & Electrical Engineering*. 2020, 85, 106670. doi: j.compeleceng.2020.106670.

[8] Triki, M., Ammari, A. C., Wang, Y., Pedram, M. Reinforcement learning-based dynamic power management of a battery-powered system supplying multiple active modes. In 2013 *European Modelling Symposium*, 2013, 437-442. doi: 10.1109/EMS.2013.74.

[9] Park, S., Pozzi, A., Whitmeyer, M., Perez, H., Joe, W. T., Raimondo, D. M., Moura, S. Reinforcement learning-based fast charging control strategy for li-ion batteries. In 2020 *IEEE Conference on Control Technology and Applications* (CCTA), 2020, 100-107. doi: 10.1109/CCTA41146.2020.9206314.

[10] Chow, Y., Nachum, O., Faust, A., Duenez-Guzman, E., Ghavamzadeh, M. Lyapunov-based safe policy optimization for continuous control. 2019, *arXiv preprint* 1901.10031. doi: 10.48550/arXiv.1901.10031.

[11] Cheng, R., Orosz, G., Murray, R. M., Burdick, J. W. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *Proc. of the AAAI Conference on Artificial Intelligence*. 2019, 33, 3387-3395. doi: 10.1609/aaai.v33i01.33013387.

[12] Grzes, M. Reward shaping in episodic reinforcement learning, *Proc. of the Int. Joint Conf. on Autonomous Agents and Multiagent Systems*, AAMAS, 2017, 1, 565–573.

[13] Dong, Y., Tang, X., Yuan, Y. Principled reward shaping for reinforcement learning via Lyapunov stability theory. *Neurocomputing*, 2020, 393, 83-90. doi: 10.1016/j.neucom.2020.02.008.

[14] Achiam, J., Held, D., Tamar, A., Abbeel, P. Constrained policy optimization. In *International conference on machine learning*, 2017, 22-31.

[15] Junges, S., Jansen, N., Dehnert, C., Topcu, U., Katoen, J. P. Safety-constrained reinforcement learning for MDPs. In *International Conference on tools and algorithms for the construction and analysis of systems*, 2016, 130-146.

[16] Abbeel, P., Ng, A. Y. Apprenticeship learning via inverse reinforcement learning, In *Proceedings, Twenty-First International Conference on Machine Learning*, ICML 2004, New York, USA, ACM Press, 2004, 1-8.

[17] Zhang, X., Ma, H. Pretraining deep actor-critic reinforcement learning algorithms with expert demonstrations. 2018, *arXiv preprint* 1801.10459. doi: 10.48550/arXiv.1801.10459.

[18] Dalal, G., Dvijotham, K., Vecerik, M., Hester, T., Paduraru, C., Tassa, Y. Safe exploration in continuous action spaces. 2018, *arXiv preprint* 1801.08757. doi: 10.48550/arXiv.1801.08757.

[19] Perez, H. E., Hu, X., Dey, S., Moura, S. J. Optimal charging of Li-ion batteries with coupled electro-thermal-aging dynamics. *IEEE Transactions on Vehicular Technology*. 2017 66(9), 7761-7770. doi: 10.1109/TVT.2017.2676044.

[20] Lin, X., Perez, H. E., Mohan, S., Siegel, J. B., Stefanopoulou, A. G., Ding, Y., Castanier, M. P. A lumped-parameter electro-thermal model for cylindrical batteries. *Journal of Power Sources*. 2014, 257, 1-11. doi: 10.1016/j.jpowsour.2014.01.097.

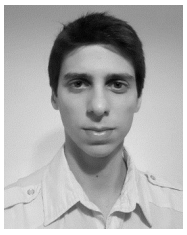
[21] Perez, H. E., Siegel, J. B., Lin, X., Stefanopoulou, A. G., Ding, Y., Castanier, M. P. Parameterization and validation of an integrated electro-thermal cylindrical lfp battery model. In *Dynamic Systems and Control Conference*. 2012, 45318, 41-50. doi: 10.1115/DSCC2012-MOVIC2012-8782.

[22] Altman, E. *Constrained Markov decision processes*. Routledge, 2021.

[23] Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Wierstra, D. Continuous control with deep reinforcement learning. 2015, *arXiv preprint* 1509.02971. doi: 10.48550/arXiv.1509.02971.

[24] Baxter, J., Bartlett, P. L. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*. 2001, 15, 319-350. doi: 10.1613/jair.806.

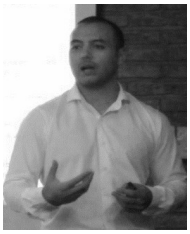
ACERCA DE LOS AUTORES



Maximiliano Trimboli es ingeniero mecatrónico graduado en la Facultad de Ingeniería y Ciencias Agropecuarias de la Universidad Nacional de San Luis (FICA-UNSL), Argentina. Allí además ejerce un cargo como Auxiliar

Docente en el área de automatización. Becado por el Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), se encuentra realizando el Doctorado en Ciencias de la Computación. Desarrolla tareas de investigación en el Laboratorio de Sistemas Inteligentes (LSI) relacionadas a métodos de aprendizaje automático aplicados en el campo de las energías renovables.

Nicolás Antonelli es un ingeniero electromecánico graduado en la Universidad Nacional General Sarmiento (UNGS), Argentina. Está realizando, mediante una beca del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), un doctorado en ciencias de la computación en la Universidad Nacional de San Luis (UNSL). Desarrolla tareas de investigación en el Laboratorio de Sistemas Inteligentes (LSI) relacionadas a métodos de aprendizaje automático aplicados en el campo de las energías renovables.



Luis Avila es ingeniero electrónico graduado en la Universidad Nacional de San Luis (UNSL), Argentina. Obtuvo su Doctorado en Ingeniería en la Universidad Tecnológica Nacional (UTN-FRSF). Es investigador del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) trabajando en el Laboratorio de Sistemas Inteligentes (LSI) de la UNSL, donde además ejerce un cargo como Profesor.

Mariano de Paula es ingeniero industrial graduado de la Universidad Nacional del Centro de la provincia de Buenos Aires, Argentina. Obtuvo un doctorado en Ingeniería de la Universidad Tecnológica Nacional (UTN-FRSF), Argentina. Es investigador en el Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) y lleva a cabo sus actividades en el INTELYMEC-UNCPBA. Además, es Profesor Adjunto en la Facultad de Ingeniería de la UNCPBA.