نشریه علمی‌پژوهشی مهندسی و مدیریت انرژی

سال دوازدهم، شمارهٔ دوم/ تابستان ۱۴۰۱/ صفحه ۲ـ۱۱

نوع مقاله: پژوهشی

تاریخ دریافت: ۱۳۹۹/۱۰/۲۸ تاریخ پذیرش: ۱۴۰۰/۶/۱۴

برنامه‌ریزی بهره‌برداری ریزشبکه‌ها مبتنی بر الگوریتم یادگیری تقویتی عمیق

سعید اسمعیلی[۱]، علیرضا ناطقی[۲]، حسن زارع[۳*]، حسین اصغرپور علمداری[۴]

[۱] دانش‌آموخته دکتری برق دانشگاه علم و صنعت ایران، تهران، ایران
saeidesmaeili.ee86@gmail.com
[۲] استادیار دانشگاه علوم و فنون هوایی شهید ستاری، تهران، ایران
ar_nateghi@yahoo.com
[۳] استادیار گروه مهندسی برق، دانشگاه فنی و حرفه‌ای، تهران، ایران
hassan.zare.tvu@gmail.com
[۴] استادیار گروه مهندسی برق، دانشگاه فنی و حرفه‌ای، تهران، ایران
asgharpour_alamdari@tvu.ac.ir

چکیده: در این مقاله، برنامه‌ریزی بهره‌برداری ریزشبکه‌ها مشتمل بر منابع تولید انرژی و سیستم‌های ذخیره انرژی مبتنی بر یادگیری تقویتی عمیق ارائه شده است. با توجه به خاصیت پویایی مسئله، ابتدا در قالب یک فرایند تصمیم‌گیری مارکوف متشکل از چهارتایی (حالت، اقدام، تابع احتمال انتقال و پاداش) فرمول‌بندی شده است. سپس، الگوریتم گرادیان استراتژی قطعی عمیق به‌منظور یادگیری استراتژی بهینهٔ برنامه‌ریزی بهره‌برداری ریزشبکه با هدف کمینه کردن هزینه‌های بهره‌برداری ارائه شده است. این الگوریتم یک روش بی‌نیاز از مدل، مستقل از استراتژی و بر مبنای معماری عامل-نقاد است که می‌تواند به‌خوبی فضای حالت و اقدام مسئله را به‌صورت پیوسته مدل‌سازی و بر چالش بزرگ بودن ابعاد مسئله غلبه کند. به‌منظور ارزیابی الگوریتم ارائه‌شده، نتایج با الگوریتم یادگیری Q عمیق و روش تحلیلی مقایسه شد. نتایج حاصل از شبیه‌سازی، کارایی الگوریتم گرادیان استراتژی قطعی عمیق ارائه‌شده را از جهت همگرایی، زمان اجرا و هزینهٔ کل نشان دادند.

واژه‌های کلیدی: ریزشبکه، گرادیان استراتژی قطعی عمیق، فرایند تصمیم‌گیری مارکوف، برنامه‌ریزی بهره‌برداری.

* نویسنده مسئول

۱. مقدمه

با گسترش روزافزون منابع انرژی تجدیدپذیر و سیستم‌های ذخیرهٔ انرژی در سطوح ولتاژ پایین و متوسط، اهمیت به‌کارگیری ریزشبکه‌ها[1] در حال افزایش است. در واقع، یک ریزشبکه به مجموعه‌ای از بارهای الکتریکی، منابع تولید انرژی (تجدیدپذیر و فسیلی) و سیستم‌های ذخیرهٔ انرژی[2] گفته می‌شود که توانایی کار در دو حالت متصل به شبکه و منفصل از آن را دارد [۱ــ۳]. استفاده از ریزشبکه‌ها در حوزه‌های مختلف مسکونی [۴]، شبکه‌های توزیع [۵]، صنایع هوافضا [۶] و کشتی‌ها [۷]، منجر به بهبود پارامترهای کیفیت توان، افزایش قابلیت اطمینان سیستم و توزیع پایدار توان شده است. با این حال، برنامه‌ریزی بهره‌برداری آن‌ها با چالش‌های مختلف از جمله عدم قطعیت‌های مربوط به پارامترهای تصادفی (میزان بار الکتریکی مصرفی، میزان توان تولیدی منابع انرژی تجدیدپذیر، قیمت بازار برق و...) و فقدان سیستم‌های پیش‌بینی‌کننده با دقت بالا روبه‌روست.

در همین راستا، تحقیقات زیادی در زمینهٔ برنامه‌ریزی بهره‌برداری ریزشبکه‌ها انجام شده است. در مرجع [۸]، با ارائهٔ یک سیستم مدیریت انرژی میزان هزینهٔ نهایی و آلاینده‌های زیست‌محیطی ریزشبکه توسط برنامه‌ریزی دینامیکی برای بارهای الکتریکی و گرمایی با در نظرگیری عدم قطعیت‌های موجود کمینه شده است. در مرجع [۹]، یک الگوریتم پخش بار بهینهٔ چندهدفه به‌منظور بهبود عملکرد ریزشبکه‌های موجود در شبکهٔ توزیع ارائه شده است که در آن به‌طور همزمان هزینهٔ بهره‌برداری، هزینهٔ کل انرژی تلف‌شده و همچنین انحراف دامنهٔ ولتاژ شین‌های مختلف حداقل شده است. در مرجع [۱۰]، برنامه‌ریزی تصادفی بهره‌برداری کوتاه‌مدت منابع تولیدات پراکنده و سیستم‌های ذخیره‌کنندهٔ انرژی در مجموعهٔ ریزشبکه‌ها ارائه کرده‌اند. مدل چندهدفهٔ ارائه‌شده به‌منظور کاهش هزینه و تلفات به‌صورت برنامه‌نویسی خطی آمیخته‌شده با اعداد صحیح[3] فرمول‌نویسی و توسط الگوریتم ژنتیک حل شده است. به‌طور مشابه، در مرجع [۱۱] به‌منظور تحقق پارامترهای غیرقطعی مربوط به قیمت بازار برق، میزان بار الکتریکی مصرفی و شدت تابش خورشید، یک درخت سناریو دومرحله‌ای متشکل از n سناریو مدل‌سازی شده است.

در یک جمع‌بندی کلی، تاکنون به‌منظور مدل‌سازی عدم قطعیت‌های موجود در برنامه‌ریزی بهره‌برداری ریزشبکه‌ها، انواع روش‌ها از جمله بهینه‌سازی تصادفی[4] [۱۲ و ۱۳]، بهینه‌سازی استوار[5] [۱۴ و ۱۵]، کنترل پیش‌بین مدل[6] [۱۶]، برنامه‌ریزی تقریبی پویا[7] [۱۷] و بهینه‌سازی لیاپانوف[8] [۱۸] ارائه شده است. در عمل، تمامی روش‌های فوق در شرایطی قابلیت پیاده‌سازی دارند که مدل سیستم و متغیرهای تصادفی آن توسط مدل‌های پیش‌بینی یا توسط مقادیر مورد انتظار به‌درستی تخمین زده شده باشند. این موضوع اصلی‌ترین چالش مقالات و روش‌های ارائه‌شدهٔ قبلی است. در مقابل، الگوریتم‌های یادگیری تقویتی[9] روش‌های مبتنی بر داده را بدون نیاز به مدل سیستم و مشخصه‌های تصادفی ارائه می‌دهند [۱۹]. با توجه به خاصیت دینامیکی مسائل برنامه‌ریزی بهره‌برداری، می‌توان آن‌ها را در قالب فرایندهای تصمیم‌گیری مارکوف[10] فرمول‌بندی کرد که در همین راستا الگوریتم‌های یادگیری تقویتی که از روش‌های برنامه‌نویسی پویا استفاده می‌کنند، معمولاً تحت عنوان یک فرایند تصمیم‌گیری مارکف مدل می‌شوند. به‌منظور برطرف کردن چالش مسائل با ابعاد بزرگ و همچنین فضای حالت پیوستهٔ آن‌ها، یادگیری عمیق[11] با به‌کارگیری شبکهٔ عصبی عمیق[12]، تابع ارزش[13] این مسائل را تخمین می‌زند [۲۰].

در مرجع [۲۱] یک رویکرد تطبیقی یادگیری تقویتی برای برنامه‌ریزی بهره‌برداری مجموعه ریزشبکه‌ها که به‌صورت عامل محور مدل شده‌اند، ارائه شده است. در این مقاله تلفیق دو رویکرد اکتشاف و بهره‌برداری[14] در یادگیری تقویتی توسط استنتاج فازی انجام شده است. بدین منظور برای تصمیم‌گیری و تنظیم پارامتر یادگیری بر اساس توان تولیدی هر عامل به‌صورت تطبیقی استفاده می‌شود.

به‌طور مشابه، در مرجع [۲۲] سیستم مدیریت انرژی ریزشبکهٔ هوشمند کشتی به‌منظور برنامه‌ریزی بهره‌برداری سیستم‌های ذخیرهٔ انرژی و پنل‌های خورشیدی با استفاده از الگوریتم یادگیری Q عمیق[15] ارائه شده است. همچنین در مراجع [۲۳] و [۲۴] به‌ترتیب مسئلهٔ برنامه‌ریزی بهره‌برداری از ریزشبکهٔ خانگی و یک سیستم ترکیبی پنل‌خورشیدی-باتری ارائه شده است. به‌منظور مدل‌سازی و

1. Microgrid (MG)
2. Energy Storage System (ESS)
3. Mixed-Integer Linear Programming (MILP)

4. Stochastic Optimization (SO)
5. Robust Optimization (RO)
6. Model Predictive Control (MPC)
7. Approximate Dynamic Programming (ADP)
8. Lyapunov Optimization
9. Reinforcement Learning (RL)
10. Markov Decision Process (MDP)
11. Deep Learning (DL)
12. Deep Neural Network (DNN)
13. Value function
14. Exploration and exploitation
15. Deep Q-Learning (DQL)

بهینه‌سازی مسئلهٔ بیان‌شده، الگوریتم یادگیری Q عمیق در دو حالت تک‌عاملی و چندعاملی ارائه شده است. گفتنی است که این الگوریتم یک روش بی‌نیاز از مدل می‌باشد که به‌خوبی فضای حالت پیوسته مسئله را مدل‌سازی و بر چالش بزرگ بودن ابعاد مسئله غلبه کرده است. در مقابل، ضعف اصلی روش‌های مبتنی بر یادگیری Q عمیق، ناتوانی آن‌ها در مدل‌سازی فضای اقدام به‌صورت پیوسته است که در عالم واقعیت در مسائل برنامه‌ریزی بهره‌برداری وجود دارد.

در همین راستا در این مقاله به‌منظور رفع چالش پیوستگی فضای اقدام، برنامه‌ریزی بهره‌برداری ریزشبکه توسط الگوریتم گرادیان استراتژی قطعی عمیق[1] که یک الگوریتم استراتژی‌محور[2] می‌باشد، ارائه شده است.

نوآوری‌های اصلی مقاله به شرح زیر است:

- فرمول‌بندی مسئلهٔ برنامه‌ریزی بهره‌برداری ریزشبکه در قالب فرایند تصمیم‌گیری مارکوف؛
- ارائهٔ الگوریتم *DDPG* (نسخه‌ای از یادگیری تقویتی عمیق) به‌منظور مدل‌سازی فضای اقدام مسئله به‌صورت پیوسته با درنظرگیری عدم قطعیت‌های موجود.

در بخش دوم، مدل‌سازی سیستم و فرمول‌نویسی مسئلهٔ برنامه‌ریزی بهره‌برداری ریزشبکه ارائه شده است. بخش سوم این مقاله، مدل‌سازی مسئله بر اساس فرایند تصمیم‌گیری مارکوف را بیان می‌کند. در بخش چهارم، الگوریتم پیشنهادی این مقاله ارائه شده است. در بخش پنجم، شبیه‌سازی و تحلیل نتایج بررسی شده است. در پایان، نتیجه‌گیری مقاله ارائه شده است.

۲. مدل‌سازی سیستم و فرمول‌نویسی مسئله

یک ریزشبکه با قابلیت بهره‌برداری در دو حالت متصل به شبکه و منفصل از آن، که شامل مجموعه‌ای از بارهای الکتریکی، پنل‌های خورشیدی[3]، میکروتوربین‌ها[4] و سیستم‌های ذخیرهٔ انرژی می‌باشد، به‌صورت شکل (۱) در نظر گرفته شده است. برنامه‌ریزی بهره‌برداری این ریزشبکه به تعداد T بازهٔ زمانی ($t \in \{1,...,T\}$) تقسیم شده است.

۱.۲. مدل میکروتوربین

فرض شده است که تعداد M میکروتوربین $(MT_1, MT_2, ..., MT_M)$ در ریزشبکهٔ مورد نظر قرار داده شده است. توان خروجی mامین میکروتوربین $P_t^{MT_m}$ باید از رابطهٔ (۱) در بازهٔ زمانی t پیروی کند.

$$P_{\min}^{MT_m} \le P_t^{MT_m} \le P_{\max}^{MT_m}, \ \forall m \in \{1,...,M\}, \quad (۱)$$

که در آن، $P_{\min}^{MT_m}$ و $P_{\max}^{MT_m}$ به‌ترتیب حداقل و حداکثر توان مجاز تولیدی توسط میکروتوربین mام است.

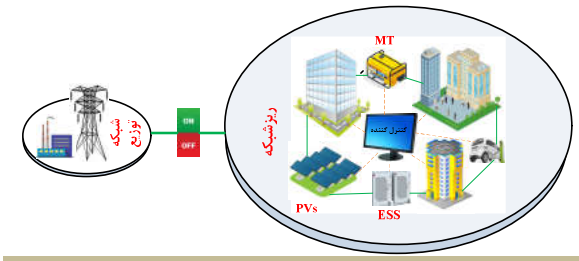
هزینهٔ بهره‌برداری میکروتوربین mام با استفاده از تابع درجهٔ دوم در رابطهٔ (۲) محاسبه می‌شود.

$$C_t^{MT_m} = a_m (P_t^{MT_m})^2 + b_m P_t^{MT_m} + c_m, \quad \forall m \in \{1,...,M\}, \quad (۲)$$

که در آن، a_m، b_m، c_m ضرایب هزینهٔ میکروتوربین هستند.

۲.۲. مدل سیستم ذخیرهٔ انرژی

روابط (۳) تا (۸) سیستم ذخیرهٔ انرژی در نظر گرفته‌شده در این مقاله را که به‌صورت مجموعه باتری‌های لیتیوم-یون است، مدل‌سازی می‌کنند.



شکل (۱): ساختار سیستم ریزشبکه

$$P_{\min}^{ESS_e} U_t^{ch_e} \le P_t^{ch_e} \le P_{\max}^{ESS_e} U_t^{ch_e}, \quad (۳)$$

$$P_{\min}^{ESS_e} U_t^{dch_e} \le P_t^{dch_e} \le P_{\max}^{ESS_e} U_t^{dch_e}, \quad (۴)$$

$$E_t^e = E_{t-1}^e + \eta^{ch_e} P_t^{ch_e} - \frac{P_t^{dch_e}}{\eta^{dch_e}}, \quad \forall t \ge 1, \quad (۵)$$

$$E_t^e = E^{init_e}, \ t = 0, \quad (۶)$$

$$E_{\min}^e \le E_t^e \le E_{\max}^e, \quad (۷)$$

$$U_t^{ch_e} + U_t^{dch_e} \le 1, \quad (۸)$$

که در آن‌ها، $P_t^{ch_e} / P_t^{dch_e}$ و E_t^e به‌ترتیب توان شارژ/دشارژ باتری‌ها و میزان انرژی ذخیره‌شده در باتری eام در زمان tام هستند. $E_{\min}^e / E_{\max}^e$، $\eta^{ch_e} / \eta^{dch_e}$ بازده شارژ/دشارژ باتری eام، حداقل/حداکثر انرژی قابل ذخیره در باتری eام و E^{init_e} انرژی اولیهٔ ذخیره‌شده است. رابطهٔ (۳) و (۴) به‌ترتیب محدودیت میزان توان شارژ و دشارژ باتری‌ها در زمان tام را تعیین می‌کند. رابطهٔ (۵) میزان انرژی ذخیره‌شده در باتری eام و رابطهٔ (۷) محدودیت آن را بیان می‌کنند. رابطهٔ (۸) نیز از شارژ و دشارژ شدن همزمان باتری جلوگیری می‌کند.

1. Deep Deterministic Policy Gradient (DDPG)
2. Policy-based
3. Photovoltaic (PV)
4. Micro Turbine (MT)

۳.۲. مدل شبکهٔ توزیع بالادست

همان طور که بیان شد، ریزشبکهٔ مورد نظر می‌تواند در دو حالت متصل به شبکهٔ توزیع و منفصل از آن عمل کند؛ یعنی می‌تواند توان اضافی تولیدی خود را به بازار عمده‌فروشی بفروشد یا کمبود توان مورد نیاز خود را از بازار عمده‌فروشی خریداری نماید. میزان توان قابل مبادله بین ریزشبکه و شبکهٔ توزیع توسط روابط (۹) تا (۱۰) محدود می‌شود.

$$0 \le P_t^{G+}, P_t^{G-} \le P_{\max}^{G}, \quad (۹)$$

$$P_t^{G+}.P_t^{G-} = 0, \quad (۱۰)$$

که در آن، P_t^{G+} و P_t^{G-} به‌ترتیب میزان توان خریداری‌شده از و فروخته‌شده به بازار عمده‌فروشی توسط ریزشبکه است. رابطهٔ (۱۰) مانع از آن می‌شود که ریزشبکه در یک زمان مشخص هم فروشندهٔ توان و هم خریدار آن باشد.

۴.۲. قید تعادل توان

به‌منظور تأمین پایدار انرژی، باید قید تعادل توان (۱۱) در زمان t برقرار باشد.

$$\sum_m P_t^{MT_m} + \sum_e (P_t^{dch_e} - P_t^{ch_e}) + \sum_p P_t^{PV_p} + P_t^{G+} - P_t^{G-} = P_t^{L} \quad (۱۱)$$

که در آن، $P_t^{PV_p}$ میزان توان تولیدی پنل‌خورشیدی pام و P_t^{L} مجموع بار الکتریکی مصرفی ریزشبکه در زمان tام است.

۵.۲. تابع هدف

در این مقاله، تابع هدف به‌صورت مجموع هزینه‌های ریزشبکه، نشان‌داده‌شده در رابطهٔ (۱۲)، در نظر گرفته شده است.

$$F_{\cos t} = Min \sum_{t=1}^{T} \left(C_t^{MT} + C_t^{ESS} + C_t^{G}\right), \quad (۱۲)$$

این تابع هدف شامل سه بخش هزینه است. جمله اول مربوط به هزینه بهره‌برداری میکروتوربین‌هاست که توسط (۲) قابل محاسبه است. جمله دوم رابطهٔ (۱۲)، مربوط به هزینهٔ بهره‌برداری، تعمیر و نگهداری باتری‌هاست که در (۱۳) ارائه شده است.

$$C_t^{ESS} = \sum_e \rho^{ESS_e} \left(P_t^{dch_e} + P_t^{ch_e}\right), \quad (۱۳)$$

که در آن ρ^{ESS_e} هزینهٔ تعمیر و نگهداری و بهره‌برداری از باتری eام است.

جملهٔ سوم در رابطهٔ (۱۲) مربوط به هزینهٔ تبادل (خریدوفروش) توان بین ریزشبکه و بازار برق عمده‌فروشی می‌باشد که توسط (۱۴) قابل محاسبه است.

$$C_t^{G} = \sum_e (\rho_t^{G+} P_t^{G+} - \rho_t^{G-} P_t^{G-}), \quad (۱۴)$$

که در آن، ρ_t^{G+} و ρ_t^{G-} به‌ترتیب قیمت خرید توان از بازار برق عمده‌فروشی و قیمت فروش توان به آن در زمان tام است.

۳. مدل‌سازی بر اساس فرایند تصمیم‌گیری مارکوف

با توجه به خاصیت دینامیکی مسائل برنامه‌ریزی بهره‌برداری، می‌توان آن‌ها را در قالب یک فرایند تصمیم‌گیری مارکوف مدل‌سازی کرد که منطبق بر نظریهٔ یادگیری تقویتی است. در همین راستا، در زمان tام، ابتدا عامل[1]، حالت محیط[2] $s_t \in S$ را دریافت می‌کند و اقدام[3] $a_t \in A$ را انجام می‌دهد. سپس، حالت محیط با توجه به تابع احتمال انتقال حالت $P(s_{t+1} | s_t, a_t)$ که منجر به پاداش[4] $R_{t+1} = r(s_t, a_t)$ می‌شود، به حالت جدید s_{t+1} منتقل می‌شود. در حالت کلی می‌توان این فرایند مارکوفی را در قالب چهارتایی (S, A, P, r) تعریف کرد. در ادامهٔ این بخش، این چهار جزء با توجه به سیستم ریزشبکهٔ ارائه‌شده بررسی می‌شوند.

۱.۳. حالت

برای مسئلهٔ برنامه‌ریزی بهره‌برداری ارائه‌شده در (۱) تا (۱۴)، فضای حالت در زمان t به‌صورت (۱۵) است.

$$s_t = (P_{t-23}^{L}, ..., P_t^{L}, \rho_{t-23}^{G+}, ..., \rho_t^{G+}, \rho_{t-23}^{G-}, ..., \rho_t^{G-}, P_t^{PV_1}, ..., P_t^{PV_p}, E_{t-1}^{1}, ..., E_{t-1}^{e}), \ s_t \in S, \quad (۱۵)$$

که شامل میزان بار الکتریکی مصرفی در ۲۴ ساعت برنامه‌ریزی، قیمت برق خریداری‌شده از/فروخته‌شده به بازار عمده‌فروشی در طول روز، میزان توان خروجی پنل‌های خورشیدی در زمان t و میزان انرژی ذخیره‌شده در سیستم‌های انرژی در بازهٔ زمانی قبلی است.

۲.۳. اقدام

تعریف مجموعه کنش‌ها از اهمیت بسزایی برخوردار است و طوری باید تعریف شود که باعث کاهش هزینهٔ کل در تمامی حالت‌های سیستم شود. حالت a_t به‌صورت زیر تعریف می‌شود:

$$a_t = (P_t^{MT_1}, ..., P_t^{MT_m}, P_t^{G+}, \rho_t^{G-}, P_t^{ch_1}, ..., P_t^{ch_e}, P_t^{dch_1}, ..., P_t^{dch_e}), \ a_t \in A(s_t). \quad (۱۶)$$

همان طور که ملاحظه می‌شود فضای اقدام این مسئله پیوسته است که در این مقاله در ادامه با ارائهٔ یک الگوریتم مناسب این چالش رفع می‌شود.

1. Agent
2. State
3. Action
4. Reward

۳.۳. تابع احتمال انتقال

با توجه به زوج (s_t, a_t) در زمان tام، به‌کمک تابع احتمال انتقال حالت در رابطهٔ (۱۷)، حالت بعدی ریزشبکه به s_{t+1} منتقل می‌شود.

$$P(s_{t+1}|s_t,a_t) = P(P_{t+1}^{L}|P_t^{L}).P(P_{t+1}^{PV}|P_t^{PV}).\\ P(\rho_{t+1}^{G+}|\rho_t^{G+}).P(\rho_{t+1}^{G-}|\rho_t^{G-}).P(E_{t+1}|E_t,a_t), \quad (۱۷)$$

که در آن، تابع احتمال انتقال انرژی ذخیره‌شده در باتری $P(E_{t+1}|E_t,a_t)$ توسط (۵) قابل محاسبه است. درحالی‌که تابع احتمال انتقال مربوط به توان خروجی پنل خورشیدی $P(P_{t+1}^{PV}|P_t^{PV})$، توان مبادله‌شده با بازار برق $P(\rho_{t+1}^{G+}|\rho_t^{G+})$ و $P(\rho_{t+1}^{G-}|\rho_t^{G-})$ و میزان بار الکتریکی مصرفی $P(P_{t+1}^{L}|P_t^{L})$ با توجه به وجود پارامترهای نایقینی در آن‌ها، در دسترس نیستند. در این مقاله با ارائهٔ الگوریتم بی‌نیاز از مدل این چالش برطرف شده است.

۴.۳.تابع پاداش

در این مقاله، تابع پاداش به‌صورت تابع هدف ارائه‌شده در (۱۲) با علامت منفی در نظر گرفته شده است.

$$r(s_t,\mathrm{a}_t) = -(C_t^{MT} + C_t^{ESS} + C_t^{G}). \quad (۱۸)$$

در هر مرحله، مقدار پاداش به‌دست‌آمده در نتیجهٔ انتقال از حالت s_t به حالت جدید s_{t+1} از طریق اتخاذ اقدام a_t به دست می‌آید.

۵.۳. ضریب تنزیل[1]

ضریب تنزیل $\gamma \in [0,1]$ میزان تأثیر پاداش لحظه‌ای $(\gamma = 0)$ در پاداش تجمعی $(\gamma = 1)$ را مشخص کرده و سبب کراندار شدن آن شده است. هرچقدر مقدار ضریب تنزیل به صفر نزدیک‌تر باشد، پاداش لحظه‌ای مدنظر بوده، و در مقابل مقادیر نزدیک به یک، بیان‌کنندهٔ پاداش در آینده است.

۴. الگوریتم برنامه‌ریزی بهره‌برداری بی‌نیاز از مدل

به‌منظور حل مسئلهٔ برنامه‌ریزی بهره‌برداری مدل‌سازی‌شده بر اساس فرایند تصمیم‌گیری مارکوف در بخش ۳، الگوریتم DDPG مبتنی بر یادگیری تقویتی عمیق[2] طراحی شده است. در این بخش، ابتدا اصول اساسی یادگیری تقویتی عمیق ارائه و سپس در ادامه، الگوریتم DDPG طراحی شده است.

۱.۴. یادگیری تقویتی عمیق

یادگیری تقویتی به‌عنوان یکی از روش‌های یادگیری ماشین از روان‌شناسی رفتارگرایی الهام گرفته است و به همین دلیل بر رفتارهایی تمرکز دارد که عامل باید برای حداکثر نمودن پاداش انجام دهد. یادگیری تقویتی یک روش یادگیری به‌منظور انتخاب رفتار بهتر با توجه به پاداش و تنبیه است بدون اینکه نیاز باشد تا عامل نحوهٔ انجام عمل را بداند [۲۵].

در الگوریتم‌های مبتنی بر یادگیری تقویتی، نوع اقدام عامل از قبل تعیین نمی‌شود، بلکه عامل با جست‌وجو بر اساس سعی و خطا رفتاری را دنبال می‌کند که بیشترین پاداش را به دست آورد؛ به‌گونه‌ای که اولویت کسب سود بیشتر در کوتاه‌مدت در مقایسه با بلندمدت است. در همین راستا برای رسیدن به پاداش بیشتر همواره دو رویکرد اصلی دنبال می‌شود: رویکرد بهره‌مندانه (حریصانه)[3] و رویکرد اکتشافی (تصادفی)[4]. یکی از چالش‌های اصلی الگوریتم‌های ارائه‌شده، تشکیل یک تعادل با ترکیب دو رویکرد فوق است [۲۶]. همان طور که در بخش قبل بیان شد، یادگیری تقویتی بر اساس تعامل با محیط شکل می‌گیرد که منطبق بر فرایندهای تصمیم‌گیری مارکوف است.

در زمان t انتخاب اقدام مناسب a_t بر اساس استراتژی[5] $\pi = (\mu_1,...,\mu_T)$ به‌صورت $a_t = \mu_t(s_t)$ مشخص می‌شود. به‌عبارت دیگر، روشی برای نگاشت حالت عامل به اقدام مناسب است. در همین راستا، انتخاب مناسب اقدام با هدف بیشینه کردن مجموع پاداش‌ها در بازهٔ نامتناهی طبق رابطهٔ (۱۹) محاسبه می‌شود.

$$R_t = \sum_{i=t}^{\infty} \gamma_{i-t} r_i(s_i,a_i). \quad (۱۹)$$

با در اختیار داشتن زوج حالت–اقدام در زمان t، تابع ارزش–اقدام[6] یا به‌اصطلاح تابع Q طبق رابطهٔ (۲۰) محاسبه می‌شود.

$$Q(s_t,a_t) = \mathrm{E}[R_t|s_t,a_t]. \quad (۲۰)$$

به‌تازگی با به‌کارگیری الگوریتم‌های مبتنی بر یادگیری Q عمیق، چالش بزرگ بودن ابعاد مسئله رفع و به‌خوبی فضای حالت پیوسته مسئله مدل‌سازی شده است. به همین منظور، طبق رابطهٔ (۲۱) با کمینه کردن تابع تلفات، پارامترهای شبکهٔ عصبی عمیق بهینه می‌شوند [۲۷].

$$L(\theta) = (r_t + \gamma \max_a Q(s_{t+1},a') - Q(s_t,a_t))^2. \quad (۲۱)$$

گفتنی است فضای حالت و اقدام در مسائل برنامه‌ریزی بهره‌برداری در عالم واقعی، پیوسته هستند در صورتی که الگوریتم‌های مبتنی بر یادگیری Q عمیق، توانایی مدل‌سازی فضای اقدام به‌صورت پیوسته را ندارند. بدین منظور در ادامه، الگوریتم DDPG قابل پیاده‌سازی روی برنامه‌ریزی بهره‌برداری ریزشبکه‌ها به‌منظور حل این چالش ارائه می‌شود.

1. Discount factor
2. Deep Reinforcement Learning (DRL)
3. Exploiting (Greedy)
4. Exploring (Random)
5. Policy
6. Action-Value function

۲.۴. راه‌حل ارائه‌شده: الگوریتم DDPG

الگوریتم DDPG یک روش بی‌نیاز از مدل، مستقل از استراتژی و بر مبنای معماری عامل–نقاد[1] می‌باشد که روش مواجهه آن با مسائل، یادگیری استراتژی‌هایی در فضای اقدام با ابعاد بالا و پیوسته است. در واقع، این روش دو معماری عامل–نقاد و گرادیان استراتژی قطعی[2] را ترکیب کرده، از تئوری شبکه Q عمیق استفاده می‌کند. در این راستا، بافر بازپخش[3] و دو مجموعهٔ شبکهٔ عصبی عمیق با ساختار یکسان و پارامترهای متفاوت را که در عامل و نقاد به‌روزرسانی می‌شوند، به کار می‌گیرد. به همین منظور، ضرایب وزنی θ^{μ} و θ^{Q} برای دو شبکهٔ عصبی عمیق در نظر گرفته شده است.

ابتدا، برای تعیین استراتژی متغیر با زمان $\pi=(\mu_1,...,\mu_T)$، T-1 شبکه عامل $\{\mu_t(s\mid\theta^{\mu_t})\}_{t=1}^{T-1}$ آموزش داده می‌شود تا به‌جای تنها یک شبکه عامل، تابع استراتژی $\{\mu_t(s)\}_{t=1}^{T-1}$ تخمین زده شود.

سپس به‌منظور حل چالش ناپایداری مسئله و دستیابی به همگرایی سریع‌تر، از روش استنتاج معکوس[4] استفاده شده است که در آن T-1 مرحلهٔ فرایند تصمیم‌گیری مارکوف (شروع از مرحلهٔ T-1 و ادامه آن به‌صورت معکوس تا مرحلهٔ ۱ در نظر گرفته می‌شود. در مرحلهٔ آموزش، برای هر پلهٔ زمانی $t\in\{T-1,...,1\}$، الگوریتم DDPG اعمال شده است تا استراتژی قطعی $a_t=\mu_t(s_t\mid\theta^{\mu_t})$ مربوط به عامل و تابع ارزش–اقدام $Q(s_t,a_t\mid\theta^{Q})$ مربوط به نقاد به دست آیند. سپس، ضرایب وزنی عامل ذخیره می‌شوند، و عامل و نقاد آموزش دیده‌شده در زمان t به‌عنوان عامل و نقاد هدف (Q',μ') برای پلهٔ زمانی قبلی، یعنی $t\text{-}1$ استفاده می‌شوند.

شبکهٔ عامل با استفاده از رابطهٔ (۲۲) محاسبه می‌شود:

$$\nabla_{\theta^{\mu}}J\approx\frac{1}{N}\sum\nabla_{a}Q(s,a\mid\theta^{Q})|_{s=s,a=\mu(s)}\cdot\nabla_{\theta^{\mu}}\mu(s\mid\theta^{\mu})|_{s}\ . \qquad (۲۲)$$

که در آن، N یک مقدار تصادفی برای رویکرد اکتشافی است. رابطهٔ فوق از دو بخش تشکیل شده است: بخش اول آن مربوط به نقاد می‌باشد که با محاسبه گرادیان تابع ارزش–اقدام، میزان انطباق اقدام عامل به مقادیر بزرگ‌تر ارزش–اقدام را نشان می‌دهد؛ بخش دوم مربوط به عامل است که در آن گرادیان تابع استراتژی محاسبه می‌شود. مفهوم این بخش این است که عامل چگونه با تغییر پارامتر خود θ^{μ} می‌تواند اقدام مورد نظر را انجام دهد.

شبکهٔ نقاد با استفاده از رابطهٔ (۲۳) و (۲۴) به‌روزرسانی می‌شود:

$$L=\frac{1}{N}\sum(r+\gamma Q'-Q)^2 \qquad (۲۳)$$

$$y=r+\gamma Q'(s',\mu'(s'\mid\theta^{\mu'})|_{\theta^{Q'}}) \qquad (۲۴)$$

رابطهٔ (۲۳) بیان می‌کند که با کمینه‌سازی خطای تفاوت زمانی[5]، می‌توان پارامترهای شبکهٔ نقاد را به‌روزرسانی کرد [۲۸]. در ضمن مقادیر هدف y نیز توسط (۲۴) محاسبه می‌شوند.

دقت شود که الگوریتم DDPG همواره به‌منظور حل یک دوره از فرایند مارکوف استفاده می‌شود که هر مرتبه آموزش مشتمل بر دو پلهٔ زمانی می‌باشد و در آن فقط برای پلهٔ زمانی اول نیاز به آموزش شبکهٔ عامل و نقاد است.

شبه‌کد مربوط به فرایند آموزش DDPG در الگوریتم ۱ به‌طور خلاصه آورده شده است.

الگوریتم ۱: شبه‌کد DDPG ارائه‌شده

۱ مقداردهی اولیهٔ شبکهٔ عامل $\mu_t(s\mid\theta^{\mu_t})$ و $\theta^{\mu_t}=\theta^{\mu_0}$ و شبکهٔ نقاد $Q(s,a\mid\theta^{Q})$ و $\theta^{Q}=\theta^{Q_0}$
۲ مقداردهی اولیهٔ شبکهٔ هدف Q' و μ'
۳ **برای** $t=T-1,...,1$ **انجام بده:**
۴ شروع بافر بازپخش
۵ درنظرگیری یک مقدار تصادفی برای N
۶ **برای** مرتبهٔ آموزش $e=1,...,E$ **انجام بده:**
۷ حالت s_t^e را دریافت کن
۸ با توجه به استراتژی فعلی، اقدام مناسب a_t^e را انتخاب کن
۹ پاداش r_t^e را دریافت کن و سپس حالت جدید s_{t+1}^e را مشاهده کن
۱۰ چهارتایی $(s_t^e,a_t^e,r_t^e,s_{t+1}^e)$ را در بافر ذخیره کن
۱۱ **اگر** $t=T-1$ هست، **آنگاه**
۱۲ $y_t=r_t+\gamma r_T(s_{t+1},\mu(ss_{t+1}))$
۱۳ **در غیر این صورت**
۱۴ $y_t=r_t+\gamma Q'(s_{t+1},\mu'(s_{t+1}\mid\theta^{\mu'})|_{\theta^{Q'}})$
۱۵ **پایان**
۱۶ با کمینه کردن تابع تلفات، شبکهٔ نقاد را به‌روزرسانی کن:
$L=\frac{1}{N}\sum(y_t-Q(s_t,a_t\mid\theta^{Q}))$
۱۷ شبکهٔ عامل را به‌روزرسانی کن:
$\nabla_{\theta^{\mu}}J\approx\frac{1}{N}\sum\nabla_{a}Q(s,a\mid\theta^{Q})|_{s=s,a=\mu(s)}\nabla_{\theta^{\mu}}\mu(s\mid\theta^{\mu})|_{s}$
۱۸ **پایان**
۱۹ شبکهٔ هدف را به‌روزرسانی کن
۲۰ **پایان**

1. Actor-Critic
2. Deterministic policy gradient
3. Replay buffer
4. Backward induction
5. Temporal difference error

۵. شبیه‌سازی و تحلیل نتایج

در این بخش ابتدا مورد مطالعه معرفی شده است، سپس نتایج حاصل از شبیه‌سازی ارائه و تحلیل‌شده، با یکدیگر مقایسه می‌شوند.

۱.۵. مورد مطالعه

در این مقاله، ریزشبکهٔ واقعی Ergon Energy [۲۹] واقع در شمال غربی کوئینزلند، مشتمل بر باتری‌های یکسان از نوع ردکس[1] [۳۰]، پنل‌های خورشیدی با ظرفیت ۵۰۰ کیلووات و میکروتوربین‌های ۲ مگاواتی به‌عنوان مورد مطالعه در نظر گرفته شده است که قابلیت اتصال به شبکهٔ توزیع بالادست را دارد. اطلاعات یک سال مربوط به میزان شدت تابش خورشید و بار الکتریکی مصرفی ریزشبکه به‌ترتیب از [۳۱] و [۳۲] استخراج شده است. اطلاعات مربوط به ۲۱ روز اول هرماه به‌عنوان مجموعه آموزش و اطلاعات مربوط به روزهای باقی‌مانده به‌عنوان مجموعه تست در نظر گرفته شده است. درمجموع، اطلاعات یک‌ساعته ۲۵۲ روز برای مجموعهٔ آموزش و ۱۱۴ روز سال به‌عنوان مجموعهٔ تست استفاده می‌شود. جزئیات مربوط به پارامترهای منابع تولید پراکنده و باتری‌ها در جدول (۱) ارائه شده است [۲۲]. در ضمن، پارامترهای مورد نیاز برای پیاده‌سازی الگوریتم DDPG در جدول (۲) آورده شده است.

الگوریتم ارائه‌شده در Tensorflow نسخهٔ ۱/۱۳ پیاده‌سازی شده و در محیط نرم‌افزار پایتون نسخهٔ ۳/۶/۸ شبیه‌سازی شده است.

جدول (۱): مشخصات مربوط به منابع تولید پراکنده و باتری‌ها [۲۲]

	$P_{\min}^{MT_m}$	$P_{\max}^{MT_m}$	ضرایب هزینهٔ MTها		
MT			a_m	b_m	c_m
	۰	۲	۰/۰۵	۶	۱۰۰
ESS	$P_{\min}^{ESS_e}$	$P_{\max}^{ESS_e}$	$E_{\min}^{e}$	$E_{\max}^{e}$	$\eta^{ch_e} / \eta^{dch_e}$
	۰	۱	۰/۵	۲	۰/۹۲

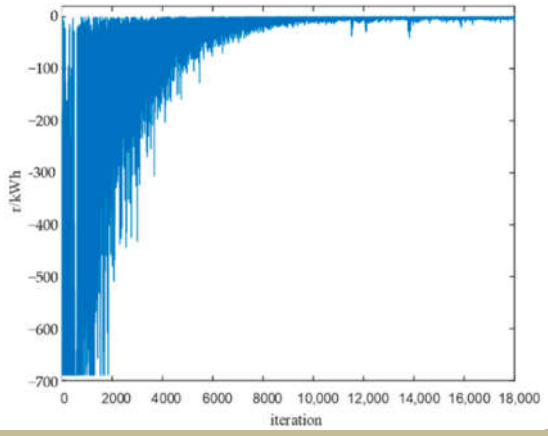
جدول (۲): پارامترهای مربوط به پیاده‌سازی الگوریتم DDPG [۳۳]

پارامتر	مقدار
نرخ یادگیری شبکهٔ عامل	10^{-4}
نرخ یادگیری شبکهٔ نقاد	10^{-5}
سایز بافر بازپخش	۲۰۰۰۰
نرخ رویکرد اکتشافی	۰/۹۹
نرخ تنزیل	۰/۹۶

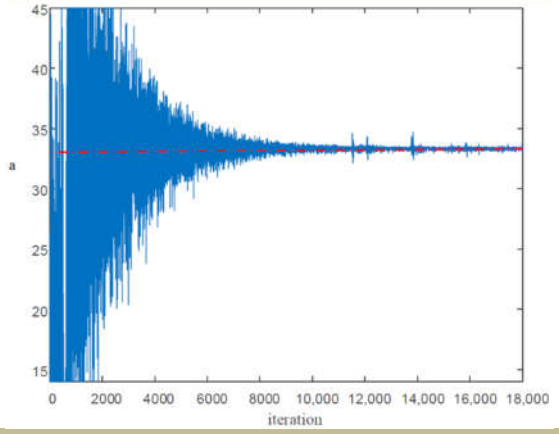
۲.۵. نتایج شبیه‌سازی

ابتدا عامل با توجه به اطلاعات یک‌ساعته ۲۵۲ روز مورد نظر آموزش دیده می‌شود تا بتوان برنامه‌ریزی بهینه را انجام داد. ارزیابی پاداش تجمعی الگوریتم DDPG ارائه‌شده در شکل (۲) نشان داده شده است. همان طور که ملاحظه می‌شود، تا ۲۰۰۰ مرتبه آموزش، اقدام از فضای اقدام به‌صورت تصادفی انتخاب شده است. سپس عامل با توجه به اطلاعات ذخیره‌شده در بافر بازپخش آموزش دیده و در نتیجه پاداش تجمعی با نوسان اندک به صفر نزدیک شده است. این اقدام بهینه منجر به یادگیری موفقیت‌آمیز استراتژی قطعی الگوریتم DDPG شده است که در شکل (۳) نشان داده می‌شود.

در این مقاله به‌منظور ارزیابی الگوریتم DDPG ارائه شده، نتایج حاصل از شبیه‌سازی با الگوریتم DQN و روش تحلیلی مقایسه شده است. همان طور که پیش‌تر بیان شد، فضای اقدام در DDPG پیوسته است، درحالی‌که این فضا برای الگوریتم DQN در بازه‌های ۰/۱ گسسته‌سازی شده است.



شکل (۲): پاداش تجمعی در حین فرایند آموزش

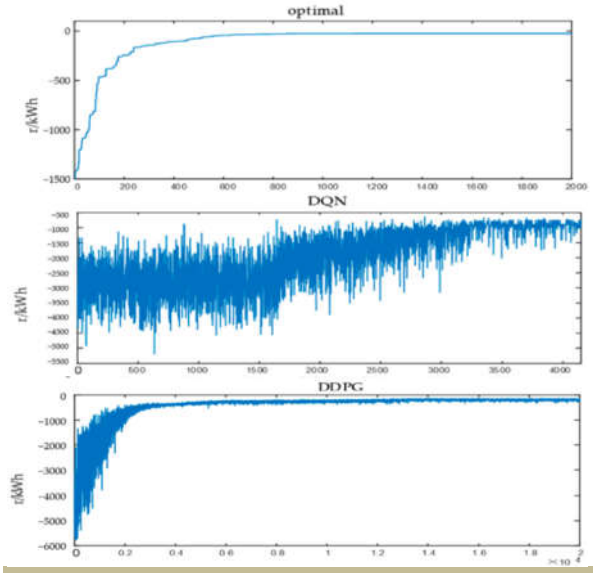


شکل (۳): منحنی تعداد دفعات تکرار اقدام

در شکل (۴)، فرایند همگرایی سه الگوریتم تحلیلی، DQN و DDPG از منظر پاداش تجمعی با یکدیگر مقایسه شده است. همان طور که ملاحظه می‌شود، روش تحلیلی و DDPG به مقادیر بهینه همگرا می‌شوند، درحالی‌که الگوریتم DQN حتی در تکرارهای زیاد نیز نوسانی بوده و نمی‌تواند همگرا شود. همان طور که در جدول (۳) نشان داده شده است، روش‌های تحلیلی، DQN و DDPG به‌ترتیب در ۶۰۰ تکرار و زمان اجرا ۱/۷ دقیقه، ۳۹۰۰ تکرار و زمان اجرا ۲۶

1. REDOX

دقیقه، و ۳۳۰۰ تکرار و زمان اجرا ۸ دقیقه همگرا می‌شوند. ملاحظه می‌شود که زمان اجرا برای الگوریتم DQN بسیار زیاد است.



شکل (۴): منحنی همگرایی پاداش تجمعی برای روش تحلیلی، DQN و DDPG

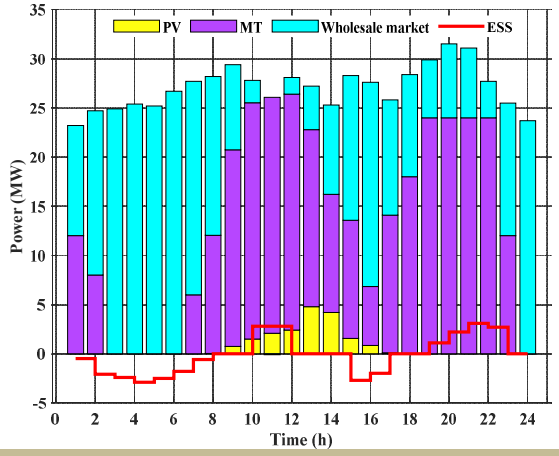
جدول (۳): مقایسهٔ الگوریتم‌های تحلیلی، DQN و DDPG

الگوریتم	تعداد کل تکرار	تکرار همگرایی	هزینهٔ کل (دلار)	زمان اجرا (دقیقه)
تحلیلی	۲۰۰۰	۶۰۰	۳۸۵۵۲	۱/۷
DQN	۴۰۰۰	۳۹۰۰	۴۹۷۲۰	۲۶
DDPG	۲۰۰۰۰	۳۳۰۰	۳۷۹۰۴	۸

همان طور که در جدول (۳) ارائه شده است، هزینهٔ کل محاسبه‌شده توسط الگوریتم DDPG و روش تحلیلی به یکدیگر نزدیک بوده، اما مقدار محاسبه‌شده توسط الگوریتم DQN نسبت به آن‌ها حدوداً ۳۰٪ بیشتر می‌باشد که علت اصلی آن گسسته‌سازی فضای اقدام و بروز خطای بیشتر در فرایند یادگیری است.

همان طور که در بخش‌های قبلی بیان شد، نقطهٔ قوت روش‌های مبتنی بر یادگیری تقویتی عمیق همچون DDPG نسبت به روش‌های تحلیلی، بی‌نیاز از مدل بودن آن‌هاست، درحالی‌که در الگوریتم‌های مبتنی بر یادگیری تقویتی نیاز به صرف زمان بیشتر برای فرایند یادگیری است، روند بهینه‌سازی برنامه‌ریزی بهره‌برداری در مدت‌زمان خیلی کم (چند ثانیه) قابل پردازش است. به همین علت، این الگوریتم‌ها قابلیت استفاده در کاربردهای زمان واقعی را دارند. در ضمن، زمان اجرا برنامه در الگوریتم DDPG نسبت به الگوریتم DQN کاهش پیدا کرده که علت اصلی آن استفاده از معماری عامل–نقاد در الگوریتم DDPG است.

شکل (۵) برنامه‌ریزی بهره‌برداری میزان توان مبادله‌شده با بازار عمده‌فروشی و منابع انرژی گسترده (میکروتوربین‌ها، پنل‌های خورشیدی و منابع ذخیرهٔ انرژی) را نشان می‌دهد. با توجه به مقایسهٔ هزینهٔ تولید توان توسط میکروتوربین‌ها با قیمت برق بازار عمده‌فروشی در ساعت‌های مختلف، ملاحظه می‌شود که در ساعت‌های ۱ تا ۲ و ساعت‌های ۷ تا ۲۳ میکروتوربین‌ها تولید توان خواهند داشت. با توجه به اوج قیمت برق بازار عمده‌فروشی در ساعت‌های ۱۰ تا ۱۲ و ساعت‌های ۱۹ تا ۲۲، مقدار توان تولیدی میکروتوربین‌ها حداکثر است (هر میکروتوربین ۲ مگاوات تولید می‌کند). از طرفی دیگر، با توجه به عدم تولید توان توسط میکروتوربین‌ها در ساعت‌های کم‌باری (یعنی ساعت‌های ۲ تا ۷ و ساعت ۲۴)، بخش عمدهٔ تأمین توان مصرفی ریزشبکه توسط بازار عمده‌فروشی انجام شده است. در ضمن، پنل‌های خورشیدی در ساعت‌های آفتابی (یعنی ساعت ۹ تا ۱۶) با توجه به شرایط رديابی نقطهٔ حداکثر توان، توان تولید کرده‌اند. حداکثر مقدار توان تولیدی مجموع ده پنل خورشیدی نصب‌شده در ساعت ۱۳ می‌باشد که به مقدار ۴/۷ مگاوات رسیده است. همچنین ملاحظه می‌شود که باتری‌های نصب‌شده به‌عنوان سیستم‌های ذخیرهٔ انرژی در ساعت‌های کم‌باری با توجه به قیمت پایین‌تر برق بازار عمده‌فروشی، شارژ شده و در ساعت‌های پرباری با توجه به قیمت بالاتر برق بازار عمده‌فروشی، تخلیه می‌شوند.



شکل (۵): برنامه‌ریزی بهره‌برداری ریزشبکهٔ مورد مطالعه

۶. نتیجه‌گیری

در این مقاله، به‌منظور برنامه‌ریزی بهره‌برداری بهینهٔ منابع تولید انرژی (میکروتوربین‌ها و پنل‌های خورشیدی) و باتری‌های الکتریکی در ریزشبکه، الگوریتم DDPG ارائه شد. این الگوریتم یک روش بی‌نیاز از مدل می‌باشد که به‌خوبی فضای حالت و اقدام پیوستهٔ مسئله را مدل‌سازی نموده است و می‌تواند بر چالش بزرگ بودن ابعاد مسئله غلبه کند.

در این مدل نشان داده شد که در ساعت‌های کم‌باری با توجه به قیمت پایین‌تر برق بازار عمده‌فروشی، میکروتوربین‌ها توان کمتری تولید کرده، باتری‌ها شارژ می‌شوند، در مقابل در ساعت‌های پرباری با توجه به قیمت بالاتر برق بازار عمده‌فروشی، میکروتوربین‌ها توان بیشتری تولید کرده، باتری‌ها تخلیه می‌شوند.

به‌منظور ارزیابی الگوریتم DDPG ارائه‌شده، نتایج حاصل از شبیه‌سازی با الگوریتم DQN و روش تحلیلی مقایسه شد. ملاحظه شد که روش تحلیلی و DDPG در تعداد تکرار کم به مقادیر بهینه همگرا می‌شوند، درحالی‌که الگوریتم DQN حتی در تکرارهای زیاد نیز نوسانی بوده، نمی‌تواند همگرا شود. در ضمن، هزینهٔ کل محاسبه‌شده توسط الگوریتم DDPG و روش تحلیلی با یکدیگر اختلاف اندکی دارند. ملاحظه شد که روش‌های تحلیلی در شرایطی قابلیت پیاده‌سازی دارند که مدل سیستم و متغیرهای تصادفی آن توسط مدل‌های پیش‌بینی یا توسط مقادیر مورد انتظار به‌درستی تخمین زده شده باشند. در مقابل، نقطهٔ قوت اصلی استفاده از الگوریتم‌های مبتنی بر یادگیری تقویتی عمیق همچون DDPG نسبت به روش‌های تحلیلی، بی‌نیاز از مدل بودن آن‌هاست.

مراجع

[1] Amini Badri A, Taghizadegan Kalantari N. Reliability Evaluation of Active Radial Distribution Systems Based on an Improved Classification Algorithm. JEM., Vol.10, No. 3, pp. 2-11, 2020

[2] Shahgholian G, Fani B, Moazzami M, Keyvani B, Karimi H. An Improvement inthe Reactive Power Sharing by the Use of Modified Droop Characteristics in Autonomous Microgrids. JEM.; Vol.9, No. 3, pp. 64-71, 2019

[3] Esmaeili, S., Anvari-Moghaddam, A. and Jadid, S., "*Retail market equilibrium and interactions among reconfigurable networked microgrids*", Sustainable Cities and Society, Vol.49, pp.101628, 2019

[4] Ahmad, S., Alhaisoni, M.M., Naeem, M., Ahmad, A. and Altaf, M., "*Joint energy management and energy trading in residential microgrid system*", IEEE Access, Vol.8, pp.123334-123346, 2020.

[5] Zhu, J., Yuan, Y. and Wang, W., "*An exact microgrid formation model for load restoration in resilient distribution system*", International Journal of Electrical Power & Energy Systems, Vol.116, p.105568, 2020.

[6] Huang, Z. and Dinavahi, V., "*An efficient hierarchical zonal method for large-scale circuit simulation and its real-time application on more electric aircraft microgrid*", IEEE Transactions on Industrial Electronics, Vol.66, pp.5778-5786, 2020.

[7] Zhaoxia, X., Tianli, Z., Huaimin, L., Guerrero, J.M., Su, C.L. and Vásquez, J.C., "*Coordinated control of a hybrid-electric-ferry shipboard microgrid*", IEEE Transactions on Transportation Electrification, Vol.5, pp.828-839, 2020.

[8] Liu, D., Xu, Y., Wei, Q. and Liu, X., "*Residential energy scheduling for variable weather solar energy based on adaptive dynamic programming*", IEEE/CAA Journal of Automatica Sinica, Vol.5, pp.36-46, 2017.

[9] Karimi, H. and Jadid, S., "*Optimal microgrid operation scheduling by a novel hybrid multi-objective and multi-attribute decision-making framework*", Energy, Vol.186, p.115912, 2019.

[10] Farzin, H., Fotuhi-Firuzabad, M. and Moeini-Aghtaie, M., "*A stochastic multi-objective framework for optimal scheduling of energy storage systems in microgrids*", IEEE Transactions on Smart Grid, Vol.8, pp.117-127, 2017.

[11] Kim, H.J., Kim, M.K. and Lee, J.W., "*A two-stage stochastic p-robust optimal energy trading management in microgrid operation considering uncertainty with hybrid demand response*", International Journal of Electrical Power & Energy Systems, Vol.124, p.106422, 2021.

[12] Shuai, H., Fang, J., Ai, X., Tang, Y., Wen, J. and He, H., "*Stochastic optimization of economic dispatch for microgrid based on approximate dynamic programming*", IEEE Transactions on Smart Grid, Vol.10, pp.2440-2452, 2018.

[13] Zakaria, A., Ismail, F.B., Lipu, M.H. and Hannan, M.A., "*Uncertainty models for stochastic optimization in renewable energy applications*", Renewable Energy, Vol.145, pp.1543-1571, 2020.

[14] Craparo, E., Karatas, M. and Singham, D.I., "*A robust optimization approach to hybrid microgrid operation using ensemble weather forecasts*", Applied energy, Vol.201, pp.135-147, 2017.

[15] Qiu, H., Long, H., Gu, W. and Pan, G., "*Recourse-Cost Constrained Robust Optimization for Microgrid Dispatch With Correlated Uncertainties*", IEEE Transactions on Industrial Electronics, 2020.

[16] Zhang, Y., Fu, L., Zhu, W., Bao, X. and Liu, C., "*Robust model predictive control for optimal energy management of island microgrids with uncertainties*", Energy, Vol.164, pp.1229-1241, 2018.

[17] Zhu, J., Mo, X., Zhu, T., Guo, Y., Luo, T. and Liu, M., "*Real-time stochastic operation strategy of a microgrid using approximate dynamic programming-based spatiotemporal decomposition approach*", IET Renewable Power Generation, Vol.13, pp.3061-3070, 2019.

[18] Li, B., Chen, T., Wang, X. and Giannakis, G.B., "*Real-time energy management in microgrids with reduced battery capacity requirements*", IEEE Transactions on Smart Grid, Vol.10, pp.1928-1938, 2017.

[19] Liu, W., Zhuang, P., Liang, H., Peng, J. and Huang, Z., "*Distributed economic dispatch in microgrids based on cooperative reinforcement learning*", IEEE transactions on neural networks and learning systems, Vol.29, pp.2192-2203, 2018.

[20] Khooban, M.H. and Gheisarnejad, M., "*A Novel Deep Reinforcement Learning Controller Based Type-II Fuzzy System: Frequency Regulation in Microgrids*", IEEE Transactions on Emerging Topics in Computational Intelligence, 2020.

[21] Kofinas, P., Vouros, G. and Dounis, A.I., "*Energy*

management in solar microgrid via reinforcement learning using fuzzy reward", Advances in Building Energy Research, Vol.12, pp.97-115, 2018.

[22] Zeng, P., Li, H., He, H. and Li, S., "*Dynamic energy management of a microgrid using approximate dynamic programming and deep recurrent neural network learning*", IEEE Transactions on Smart Grid, Vol.10, pp.4435-4445, 2018.

[23] Xu, Xu, Youwei Jia, Yan Xu, Zhao Xu, Songjian Chai, and Chun Sing Lai., "*A multi-agent reinforcement learning-based data-driven method for home energy management*", IEEE Transactions on Smart Grid, Vol. 11, pp. 3201-3211, 2020.

[24] Huang, Bin, and Jianhui Wang, "*Deep-Reinforcement-Learning-Based Capacity Scheduling for PV-Battery Storage System*", *IEEE Transactions on Smart Grid*, Vol. 12, pp. 2272-2283, 2020.

[25] Han, M., Zhao, J., Zhang, X., Shen, J. and Li, Y., "*The reinforcement learning method for occupant behavior in building control: A review*", Energy and Built Environment, 2020.

[26] Dos Santos Mignon, A. da Rocha, R.L.D.A., "*An Adaptive Implementation of ε-Greedy in Reinforcement Learning*", Procedia Computer Science, Vol.109, pp.1146-1151, 2017.

[27] Afrasiabi, M., Mohammadi, M., Rastegar, M. and Kargarian, A., "*Multi-agent microgrid energy management based on deep learning forecaster*", Energy, Vol.186, p.115873, 2019.

[28] Devraj, Adithya M., Ioannis Kontoyiannis, Sean P. Meyn, "*Differential temporal difference learning*", IEEE Transactions on Automatic Control, 2020.

[29] Beere, N., McPhail, D. and Sharma, R., "*A general methodology for utility microgrid planning: A Cairns case study*", In 2015 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC), pp. 1-5, 2015.

[30] Carpinelli, G., Celli, G., Mocci, S., Mottola, F., Pilo, F. and Proto, D., "*Optimal integration of distributed energy storage devices in smart grids*", IEEE Transactions on smart grid, Vol.4, pp.985-995, 2013.

[31] Rabiee, A., Sadeghi, M., Aghaeic, J. and Heidari, A., "*Optimal operation of microgrids through simultaneous scheduling of electrical vehicles and responsive loads considering wind and PV units uncertainties*", Renewable and Sustainable Energy Reviews, Vol.57, pp.721-739, 2016.

[32] Ustun, T.S., Ozansoy, C. and Zayegh, A., "*Recent developments in microgrids and example cases around the world—A review*", Renewable and Sustainable Energy Reviews, Vol.15, pp.4030-4041, 2011.

[33] Lei, L., Tan, Y., Dahlenburg, G., Xiang, W. and Zheng, K., "*Dynamic Energy Dispatch Based on Deep Reinforcement Learning in IoT-Driven Smart Isolated Microgrids*", IEEE Internet of Things Journal, 2020.