Computational Intelligence in Electrical Engineering
Vol. 16, No. 3, 2025
pp. 53-68
Research Paper

Investigating the inequity aversion model in multi-agent reinforcement learning approaches for enhancing power control in wireless cellular networks

Farinaz Alamiyan-Harandi [1], Azam S Hosseinzadeh-Salaty [2], Zeinab Maleki [3], Pouria Ramazi [4]

[1] Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran

[2] Electrical and Computer Engineering Group, Golpayegan College of Engineering, Isfahan University of Technology, Golpayegan, Iran

[3] Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran

[ˈ] Department of Biological Sciences and Dept. of Mathematics and Statistics, University of Calgary, Calgary, Canada

Abstract:

Recently, data-driven deep reinforcement learning algorithms have replaced traditional model-based algorithms in the problem of power allocation in wireless cellular networks. This article focuses on multi-agent reinforcement learning approaches that leverage centralized training and distributed execution. These approaches use an intrinsic reward function to solve the downlink power control problem in a multi-user wireless cellular network, with the aim of maximizing the sum rate. To enhance the quality of the agents' learning and their collaboration and coordination, the inequity aversion model is applied in the structure of multi-agent reinforcement learning methods. This model manipulates the agents' rewards based on the concepts of envy and guilt. The parameters of this model should first be selected properly. Various simulation results indicate that applying the inequity aversion model with appropriately tuned parameters can improve the performance of basic multi-agent reinforcement learning methods in the problem of power control in multi-user wireless cellular networks.

Keywords: Multi-Agent Reinforcement Learning, Wireless Cellular Networks, Resource Management.

https://doi.org/10.22108/isee.2025.143537.1715

مقاله پژوهشی

بررسی مدل تنفر از بی‌عدالتی در راهکارهای یادگیری تقویتی چندعاملی برای ارتقای کنترل توان در شبکه‌های سلولی بی‌سیم

فریناز اعلمی‌یان هرندی[۱]، اعظم سادات حسین‌زاده صلوتی*[۲]، زینب مالکی[۳]، پوریا رمضی[٤]

۱- پژوهشگر پسادکتری، دانشکدهٔ مهندسی برق و کامپیوتر، دانشگاه صنعتی اصفهان، اصفهان، ایران
farinaz.alamiyan@gmail.com

۲- استادیار، گروه مهندسی برق و کامپیوتر، دانشکدهٔ فنی مهندسی گلپایگان، دانشگاه صنعتی اصفهان، گلپایگان، ایران
salaty@iut.ac.ir

۳- دانشیار، دانشکدهٔ مهندسی برق و کامپیوتر، دانشگاه صنعتی اصفهان، اصفهان، ایران
zmaleki@iut.ac.ir

٤- استادیار، گروه بیولوژی و گروه ریاضی و آمار، دانشگاه کلگری، کلگری، کانادا
p.ramazi@gmail.com

چکیده: به‌تازگی، الگوریتم‌های مبتنی بر دادهٔ یادگیری تقویتی عمیق جایگزین الگوریتم‌های مبتنی بر مدل سنتی در مسئلهٔ تخصیص توان در شبکه‌های سلولی بی‌سیم شده‌اند. تمرکز این مقاله بر راهکارهای یادگیری تقویتی چندعاملی است که از آموزش متمرکز و اجرای توزیع‌شده بهره می‌برند. در این راهکارها، از یک تابع پاداش داخلی برای حل مسئلهٔ کنترل توان لینک پایین‌رو در یک شبکهٔ سلولی بی‌سیم چندکاربره، با هدف بیشینه‌سازی نرخ مجموع استفاده می‌شود. در این راستا، برای بهبود کیفیت یادگیری عامل‌ها و همکاری و هماهنگی میان آنها، مدل تنفر از بی‌عدالتی در ساختار یادگیری تقویتی چندعاملی اعمال می‌شود. این مدل بر اساس هر دو مفهوم حسادت و عذاب وجدان، پاداش‌های عامل‌ها را با هدف دست‌یابی به برابری دستکاری می‌کند. نتایج شبیه‌سازی‌های مختلف نشان می‌دهد اعمال مدل تنفر از بی‌عدالتی با پارامترهایی که به طور مناسب تنظیم شده‌اند، می‌تواند عملکرد روش‌های یادگیری تقویتی چندعاملی پایه را در مسئلهٔ تخصیص توان در شبکه‌های سلولی بی‌سیم بهبود بخشد.

واژه‌های کلیدی: یادگیری تقویتی چندعاملی، شبکه‌های سلولی بی‌سیم، مدیریت منابع.

۱- مقدمه

از آنجا که مدیریت منابع رادیویی یک مسئلهٔ غیرمحدب است و افزایش اندازهٔ شبکه در پیچیدگی محاسبات آن نقشی به‌سزا دارد، الگوریتم‌های متمرکز و توزیع‌شدهٔ متنوعی برای مدیریت منابع رادیویی ارائه شده‌اند. این راهکارها از تکنیک‌های مطرح در زمینه‌های مختلف علمی همچون برنامه‌ریزی هندسی [۱]، بهینه‌سازی حداقل وزن‌دار میانگین مربعات خطا[۱] [۲]، نظریهٔ بازی [۳]، نظریهٔ اطلاعات [٤، ۵]، و برنامه‌ریزی کسری[۲] [٦، ۷] بهره‌برداری می‌کنند. با توجه به اینکه شبکه‌های بی‌سیم ماهیتی پویا دارند، هیچ تضمینی وجود ندارد که این الگوریتم‌ها برای مدیریت منابع رادیویی در تمام سناریوها به سطحی قابل قبول از عملکرد دست یابند. اما الگوریتم‌هایی که از طریق تعامل با محیط یک وظیفه را یاد می‌گیرند، قادر هستند به‌خوبی چنین پویایی‌هایی را مدیریت کنند. در این راستا، چارچوب‌هایی که فرایند تصمیم‌گیری خود را بر مبنای حجمی عظیم از

[۱] تاریخ ارسال مقاله : ۱٤۰۳/۰۹/۲۱
تاریخ پذیرش مقاله : ۱٤۰٤/۰۹/۰۹
نام نویسنده مسئول : اعظم سادات حسین‌زاده صلوتی

نشانی نویسنده مسئول: گروه مهندسی برق و کامپیوتر، دانشکده فنی مهندسی گلپایگان، دانشگاه صنعتی اصفهان، گلپایگان، ایران

داده‌های فعلی شبکه‌های ارتباطی بی‌سیم قرار می‌دهند، برای مقابله با این چالش‌ها مناسب هستند. به همین دلیل، تکنیک‌های یادگیری ماشین در حل طیفی وسیع از مسائل مطرح در حوزهٔ شبکه‌های بی‌سیم پیشنهاد شده‌اند.

روش‌های یادگیری تقویتی زیرمجموعه‌ای خاص از الگوریتم‌های یادگیری ماشین هستند که امکان یادگیری را از طریق تعامل با محیط برای یک عامل فراهم می‌کنند. در این روش‌ها، عامل در پاسخ به مشاهدات دریافتی خود، اقداماتی را با هدف بیشینه‌سازی پاداش در محیطِ مسئله اعمال می‌کند. اجرای عمل انتخابی عامل باعث می‌شود محیط به مرحلهٔ بعدی منتقل شود و عامل، علاوه بر دریافت یک مقدار پاداش عددی، مجموعه‌ای جدید از مشاهدات را درک کند. الگوریتم‌های یادگیری در حین این تعاملات، سیاست رفتاری عامل را اصلاح می‌کنند تا عامل پاداش خود را در طول زمان بیشینه کند.

پیچیدگی‌های فضای حالت و عمل در مسائل مختلف به ویژه مسائل دنیای واقعی، پژوهشگران را به سمت بهره‌برداری از راهکارهای یادگیری تقویتی عمیق سوق داده است. در این روش‌ها، از ساختار شبکهٔ عصبی عمیق به عنوان یک تقریب‌زنندهٔ تابع جامع استفاده می‌شود تا تخمینی از احتمال انجام هر عمل از مجموعه عمل‌های مجاز عامل با توجه به هر مشاهده و نیز ارزش هر جفت مشاهده–عمل به دست آید. الگوریتم‌های یادگیری تقویتی عمیق در حل مسائل چالش‌برانگیز تصمیم‌گیری متوالی، به ویژه در انواع بازی‌ها، مانند Atari 2600 و Go به موفقیت‌هایی چشمگیر دست یافته‌اند [۸–۱۰].

به‌تازگی، از تکنیک‌های یادگیری تقویتی عمیق برای حل مسئلهٔ کنترل توان لینک پایین‌رو[3] در شبکه‌های بی‌سیم سلولی به عنوان یک سیستم چندعاملی استفاده شده است و مکانیسم‌هایی برای برنامه‌ریزی ارسال‌ها پیشنهاد شده‌اند تا عملکرد شبکه برای همهٔ کاربران در سراسر شبکه منصفانه باشد. برای نمونه، می‌توان به پژوهش‌های منگ و همکاران [۱۱، ۱۲] اشاره کرد که در آن، یک شبکهٔ عصبی با استفاده از الگوریتم DQN[4] در یک محیط شبیه‌سازی با هدف تخصیص توان در شبکه‌های سلولی چندکاربره آموزش داده شده است. شبکهٔ عصبی حاصل از این آموزش به صورت پویا در سناریوهای واقعی با هدف انتقال یادگیری بهره‌برداری و تنظیم شده است. در کاربرد روش DQN، از نرخ مجموع برای تعریف تابع پاداش استفاده شده است. مقایسهٔ عملکرد روش پیشنهادی منگ و همکاران با الگوریتم‌های مبتنی بر مدل نشان داده است میانگین نرخ مجموع در این روش بیشتر و قابلیت تعمیم‌پذیری بهتر است.

در سال‌های اخیر، ادبیات کنترل توان مبتنی بر یادگیری تقویتی چندعاملی[5] در شبکه‌های سلولی جهشی معنادار داشته است. از یک سو، رویکردهای غیرمتمرکز و مقیاس‌پذیر برای کنترل توان ارائه شده‌اند که فقط بر مشاهدات محلی تکیه دارند و همچنان به کارایی نزدیکِ روش‌های متمرکز دست می‌یابند [۱۳، ۱٤]. از سوی دیگر، چارچوب «آموزش متمرکز و اجرای توزیع‌شده»[6] در قالب‌های تعاملی و مشارکتی برای کنترل توان به کار گرفته شده و در برخی موارد با شبکهٔ عصبی حافظهٔ طولانی کوتاه مدت[7] نیز ترکیب شده است تا ناایستایی کانال را بهتر مدل کند [۱۵، ۱٦]. هم‌زمان، جریان روبه‌رشد انصاف‌محور در یادگیری تقویتی از توابع رفاه اجتماعی[8] و شبکه‌های تفکیکِ ارزش[9] برای موازنهٔ کارایی–عدالت در تخصیص منابع بهره گرفته و نتایجی امیدبخش در سناریوهای مخابراتی گزارش کرده است [۱۷، ۱۸]. این خط پژوهشی با ایدهٔ مطرح در این مقاله، یعنی به‌کارگیری پاداشِ داخلیِ «تنفر از بی‌عدالتی»[10] در یادگیری تقویتی چندعاملی هم‌راستاست و جایگاه نظری رویکرد این مقاله را تقویت می‌کند.

اثربخشیِ اعمالِ مدل تنفر از بی‌عدالتی در ساختار یادگیری تقویتی چندعاملی در بهبود کیفیت یادگیری عامل‌ها و همکاری و هماهنگی میان آنها در سناریوهایی همچون بازی‌های پاک‌سازی و برداشت[11] و کنترل ترافیک آزموده شده است [۱۹، ۲۰]. این مدل بر اساس هر دو مفهوم حسادت[12] و عذاب وجدان[13]، پاداش‌های عامل‌ها را با هدف دست‌یابی به برابری دستکاری می‌کند. مدل تنفر از بی‌عدالتی به‌جای تمرکز صرف روی بیشینه‌کردن کارایی کلی شبکه (مانند مجموع توان یا نرخ داده)، انصاف را به عنوان یک معیار اساسی در تابع پاداش عامل یادگیری تقویتی تعبیه می‌کند. این کار از نارضایتی کاربران با منابع ضعیف

جلوگیری می‌کند و پایداری و قابلیت اطمینان شبکه را در بلندمدت افزایش می‌دهد. در این مدل، اگر عامل یادگیری تقویتی دریابد یک کاربر برای مدت طولانی از سرویس محروم مانده است یا نرخ دادهٔ بسیار پایینی دارد، پاداش منفی دریافت می‌کند. این امر عامل را مجبور می‌کند تا برای بهبود وضعیت آن کاربر اقدام کند، حتی اگر این اقدام کمی از کارایی کلی شبکه بکاهد. ترافیک کاربران، تعداد کاربران و شرایط کانال در شبکه‌های مدرن به طور مداوم در حال تغییر است. عامل یادگیری تقویتی می‌تولند این شرایط پیچیده و متغیر را مستقیماً از داده‌ها بیاموزد و یک سیاست تطبیقی ایجاد کند که در شرایط مختلف، بی‌عدالتی را تشخیص و به آن واکنش نشان دهد. بهره‌برداری از مدل تنفر از بی‌عدالتی می‌تولند تعادل هوشمند و پویا بین کارایی و انصاف برقرار کند و با شرایط پیچیده و متغیر شبکه‌های مدرن سازگار شود.

ساختار مقاله به این شرح است: در بخش دوم، مسئلهٔ تخصیص پویای توان لینک پایین‌رو در یک شبکهٔ بی‌سیم مطرح می‌شود. سپس، روش‌های یادگیری تقویتی پایه و مدل تنفر از بی‌عدالتی به‌ترتیب در بخش‌های سوم و چهارم معرفی می‌شوند. در بخش پنجم، عناصر یادگیری تقویتی در مسئلهٔ تحت بررسی شرح داده می‌شوند. آزمایش‌ها و نتایج حاصل از آنها در بخش ششم گزارش می‌شوند و در نهایت، در بخش هفتم، نتیجه‌گیری بیان می‌شود.

۲- بیان مسئله

در اینجا، مسئلهٔ تخصیص پویای توان لینک پایین‌رو در یک شبکهٔ سلولی بی‌سیم با کانال دسترسی چندگانهٔ تداخلی[۱٤] با هدف بیشینه‌سازی نرخ مجموع[۱۵] بررسی می‌شود. به عبارت دیگر، هدف اصلی مسئله انتخاب توان انتقال لینک پایین‌رو در پاسخ به شرایط فیزیکی کانال، تحت محدودیت‌های مقدار بیشینهٔ توان است. محیط مسئله مطابق ساختار ارائه‌شده توسط مِنگ و همکاران [۱۱، ۱۲] یک شبکهٔ سلولی بی‌سیم با یک کانال پخش تداخلی تک‌ورودی تک‌خروجی[۱۶] است که از N سلول تشکیل و در مرکز هر سلول، یک ایستگاه پایه[۱۷] مجهز به یک آنتن فرستنده مستقر شده است. با فرض باندهای فرکانسی مشترک، تعداد K کاربر در هر سلول به طور هم‌زمان توسط فرستنده ایستگاه پایهٔ آن سلول سرویس‌دهی می‌شوند. در شیار زمانی tاُم، بهرهٔ کانال مستقل بین ایستگاه پایهٔ nاُم و کاربر kاُم در سلول jاُم به صورت $g_{n,j,k}^{t}$ در رابطهٔ (۱) تعریف می‌شود:

$$g_{n,j,k}^{t} = \left|h_{n,j,k}^{t}\right|^2 \beta_{n,j,k} \qquad (۱)$$

که در آن، $h_{n,j,k}^{t}$ یک متغیر تصادفی گوسی مختلط است که پوش آن توزیع رایلی دارد و $\beta_{n,j,k}$ مؤلفهٔ مقیاس بزرگ محوشدگی است و هر دو مورد تضعیف هندسی[۱۸] و محوشدگی سایه[۱۹] را شامل می‌شوند و فرض می‌شود مقدار آنها در بازهٔ زمانی مدنظر ثابت است. مطابق مدل Jakes، محوشدگی مسطح[۲۰] مقیاس کوچک می‌تواند به صورت یک فرایند گوس–مارکوف مختلط مرتبهٔ اول به صورت رابطهٔ (۲) مدل شود:

$$h_{n,j,k}^{t} = \rho h_{n,j,k}^{t-1} + n_{n,j,k}^{t} \qquad (۲)$$

که در این رابطه $h_{n,j,k}^{1} \sim \mathcal{CN}(0,1)$ و $n_{n,j,k}^{t} \sim \mathcal{CN}(0,1-\rho^2)$ و ضریب همبستگی ρ به صورت رابطهٔ (۳) در نظر گرفته می‌شود:

$$\rho = J_0(2\pi f_d T_s) \qquad (۳)$$

به طوری که $J_0(.)$ تابع بسل مرتبهٔ صفر نوع اول، f_d فرکانس داپلر بیشینه و T_s دورهٔ زمانی بین شیارهای متوالی است.

با فرض اینکه سیگنال‌های ارسالی از فرستنده‌های متفاوت مستقل از یکدیگر هستند، کانال‌ها در هر شیار زمانی ثابت در نظر گرفته می‌شوند. به این ترتیب، نسبت سیگنال به تداخل و نویز[۲۱] برای لینک پایین‌رو از ایستگاه پایهٔ nاُم به کاربر kاُم ($\mathrm{dl}_{n,k}$) در شیار زمانی t به صورت رابطهٔ (٤) تعریف می‌شود:

$$\gamma_{n,k}^{t} = \frac{g_{n,n,k}^{t}\, p_{n,k}^{t}}{\sum_{k' \neq k} g_{n,n,k}^{t}\, p_{n,k'}^{t} + \sum_{n' \in D_n} g_{n',n,k}^{t} \sum_{i} p_{n',i}^{t} + \sigma^2} \qquad (٤)$$

که در آن، D_n مجموعهٔ تمام سلول‌های تداخلی در اطراف سلول nاُم، $p_{n,k}^{t}$ توان انتشار فرستندهٔ nاُم به گیرندهٔ kاُم در شیار زمانی tاُم و σ^2 توان نویز جمع‌شونده است.

عبارت‌های $\sum_{k'\neq k} g^t_{n,n,k}\, p^t_{n,k'}$ و $\sum_{n'\in D_n} g^t_{n',n,k} \sum_i p^t_{n',i}$ نیز به‌ترتیب معرف تداخلات درون‌سلولی و بین‌سلولی هستند. با فرض پهنای بلند نرمال‌شده، نرخ لینک پایین‌رو $\mathrm{dl}_{n,k}$ به صورت رابطهٔ (۵) تعریف می‌شود:

$$C^t_{n,k} = \log_2(1+\gamma^t_{n,k}). \qquad (۵)$$

با توجه به تعاریف بالا، مسئلهٔ بهینه‌سازی با در نظر گرفتن مجموعهٔ توان غیرمنفی $\boldsymbol{p}$ که در آن، تمام عناصر محدودیت توان بیشینه را ارضا می‌کنند، به صورت بیشینه‌سازی نرخ مجموع در سرتاسر شبکه در رابطهٔ (۶) تعریف می‌شود:

$$\begin{aligned} &\max_{\boldsymbol{p}^t} C(\boldsymbol{g}^t, \boldsymbol{p}^t) \\ &s.t. \quad 0 \le p^t_{n,k} \le P_{max} \quad \forall n,k \end{aligned} \qquad (۶)$$

که در آن، P_{max} بیشترین توان انتشار سیگنال است و مجموعه توان $\boldsymbol{p}^t$، مجموعه بهرهٔ کانال $\boldsymbol{g}^t$ و نرخ مجموع $C(\boldsymbol{g}^t, \boldsymbol{p}^t)$ به صورت رابطه‌های (۷)، (۸) و (۹) تعریف می‌شوند:

$$\boldsymbol{p}^t := \{p^t_{n,k} \,|\forall n,k\} \qquad (۷)$$

$$\boldsymbol{g}^t := \{g^t_{n',n,k} \,|\forall n',n,k\} \qquad (۸)$$

$$C(\boldsymbol{g}^t, \boldsymbol{p}^t) := \sum\nolimits_{n,k} C^t_{n,k}. \qquad (۹)$$

توجه شود که بهینه‌سازی بالا یک مسئلهٔ غیرمحدب و چندجمله‌ای غیرقطعی سخت[۲۲] است. از آنجا که اطلاعات وضعیت کانال[۲۳] در هر شیار زمانی، اطلاعات کافی دربارهٔ راه‌حل بهینه در اختیار قرار می‌دهد، تابعی برای نگاشت این اطلاعات به راه‌حل مسئله موجود و قابل دست‌یابی است. حل این مسئلهٔ بهینه‌سازی به صورت یک سیستم یادگیری تقویتی تک‌عاملی با مشکلاتی همچون افزایش نمایی ابعاد فضای حالت و عمل با افزایش تعداد سلول‌ها و عدم امکان انتقال تمام اطلاعات وضعیت کانال به عامل با تأخیری قابل تحمل مواجه است. به همین دلیل، تخصیص توان به صورت غیرمتمرکز و به عنوان یک مسئلهٔ یادگیری تقویتی چندعاملی مطرح می‌شود و برای کنترل توان انتقال هر لینک پایین‌رو از یک عامل یادگیری تقویتی مجزا بهره‌برداری می‌شود. همهٔ عامل‌ها در شبکهٔ ارتباطی به صورت هم‌زمان و توزیع‌شده عمل می‌کنند. در این بین، عامل مرتبط با لینک پایین‌رو $\mathrm{dl}_{n,k}$ فقط بخشی از اطلاعات وضعیت کانال ($\boldsymbol{g}^t_{n,k}$) را به عنوان ورودی دریافت می‌کند و توان انتقال لینک پایین‌رو مخصوص خود ($p^t_{n,k}$) را به عنوان خروجی تولید می‌کند. اطلاعات جزئی وضعیت کانال $\boldsymbol{g}^t_{n,k}$ به صورت رابطهٔ (۱۰) تعریف می‌شود:

$$\boldsymbol{g}^t_{n,k} = \left\{g^t_{n',n,k} | n' \in \{n, D_n\}\right\}. \qquad (۱۰)$$

به این ترتیب، مسئلهٔ بهینه‌سازی (۶) به صورت مسئلهٔ برنامه‌ریزی چندهدفهٔ[۲۴] رابطهٔ (۱۱) در نظر گرفته می‌شود:

$$\begin{aligned} &\left\{\max_{p^t_{n,k}} C^t_{n,k}(\boldsymbol{g}^t_{n,k}, p^t_{n,k}) |\ \forall n,k\right\} \\ &s.t. \ \ 0 \le p^t_{n,k} \le P_{max} \ \ \forall n,k\,. \end{aligned} \qquad (۱۱)$$

یادگیری چندعاملی برای حل این مسئله حتی به صورتی که در رابطهٔ (۱۱) تعریف شده است، همچنان دشوار است؛ زیرا برای یادگیری تمام پارامترهای ساختار عامل‌ها که به صورت شبکه‌های عصبی عمیق هستند، به زمان یادگیری طولانی و تعداد زیادی دادهٔ آموزشی نیاز است. برای برطرف‌کردن این مشکل، از چارچوب آموزش متمرکز و اجرای توزیع‌شده بهره‌برداری می‌شود. در این چارچوب، تمام عامل‌ها در قالب یک عامل و با ساختاری مشابه در نظر گرفته می‌شوند و سیاستی یکسان برای تمام عامل‌ها به صورت اشتراکی و با استفاده از داده‌های جمع‌آوری‌شده از عملکرد تمام لینک‌های پایین‌رو یاد گرفته می‌شود. این چارچوب، با توجه به اینکه لینک‌ها از نظر مشخصات مکانی ثابت و ابعاد شبکه وسیع هستند، قابل اعمال است؛ زیرا لینک‌ها در فضاهای متمایز تقریباً مشابه یکدیگر هستند؛ بنابراین، سیاست یادگرفته‌شده می‌تواند با بهره‌برداری از مفاهیم انتقال یادگیری[۲۵] بین عامل‌ها به اشتراک گذاشته شود. در واقع، هر عامل این سیاست را به صورت مستقل استفاده می‌کند. به عبارت دیگر، روش یادگیری متمرکز و روش اجرا توزیع‌شده است.

۳- روش‌های یادگیری تقویتی پایه

در این مقاله، سه الگوریتم یادگیری تقویتی عمیق از سه دستهٔ اصلی معماری‌های یادگیری تقویتی شامل معماری

عملگر–تنها، نقاد–تنها و عملگر–نقاد به عنوان روش‌های پایه در نظر گرفته شده‌اند. الگوریتم REINFORCE [۲۱] یک روش یادگیری گرادیان سیاست از دستهٔ معماری‌های عملگر–تنها و مبتنی بر روش مونت–کارلو[26] است. الگوریتم‌های مبتنی بر سیاست به صورت صریح سیاست تصادفی $\boldsymbol{\pi}$ را در قالب یک بردار احتمالاتی در خروجی یک شبکهٔ عصبی با پارامترهای قابل یادگیری $\boldsymbol{\theta}_\pi$ تولید می‌کنند. به عبارت دیگر، یادگیری سیاست $\boldsymbol{\pi}(a|s;\boldsymbol{\theta}_\pi)$ با استفاده از تنظیم پارامترهای یک شبکهٔ عصبی برای هر یک از حالت‌های دریافتی s عمل مناسب a را تولید می‌کند. راهکار کلی در روش گرادیان افزایشی تصادفی[۲۷] به نمونه‌هایی نیاز دارد که امید گرادیان آنها متناسب با گرادیان واقعی معیار عملکرد به عنوان تابعی از پارامترها باشد. هدف الگوریتم REINFORCE مطابق رابطهٔ (۱۲)، بیشینه‌سازی مقدار پاداش مورد انتظار تحت سیاست $\boldsymbol{\pi}$ است:

$$\boldsymbol{\theta}_\pi^* = \arg\max \mathbb{E}_\pi[\textstyle\sum_a \pi(a|s;\boldsymbol{\theta}_\pi) r_s^a]. \qquad (۱۲)$$

گرادیان نمونه‌برداری مونت–کارلو به صورت رابطهٔ (۱۳) محاسبه می‌شود:

$$\nabla\boldsymbol{\theta}_\pi = \mathbb{E}_\pi\left[\nabla_{\boldsymbol{\theta}_\pi} \ln \pi(a|s;\boldsymbol{\theta}_\pi) r_s^a \,|s = s^t, a = a^t\right] \qquad (۱۳)$$

به طوری که ∇ عملگر گرادیان است. از آنجا که شبکهٔ عصبی سیاست به صورت مستقیم سیاست تصادفی را تولید می‌کند، در طول یادگیری با استفاده از این شبکه، عمل‌های تصادفی تولید می‌شوند. اما زمانی که یادگیری انجام شد و سیاست بهینه یاد گرفته شد، این شبکهٔ عصبی عمل بهینه را تولید خواهد کرد. برای کاهش حساسیت این الگوریتم به مقدار تابع پاداش، از مقادیر نرمال‌شدهٔ پاداش به صورت $\tilde{r} = \frac{r-\mu_r}{\sigma_r}$ استفاده می‌شود که در آن، μ_r و σ_r به‌ترتیب میانگین و انحراف معیار پاداش r هستند.

الگوریتم DQN[28] [۸] از معماری نقاد–تنها پیروی می‌کند. این الگوریتم مقدار تابع ارزش $Q(s,a;\theta_q)$ را تخمین می‌زند؛ به طوری که در آن، θ_q بردار پارامترهای قابل تنظیم شبکهٔ عصبی است. انتخاب عمل‌های مناسب در این الگوریتم به دقت تخمین تابع ارزش بستگی دارد و هدف الگوریتم، مطابق رابطهٔ (۱٤)، کمینه‌کردن تابع اتلاف l_2 با یافتن بردار پارامترهای بهینه $\boldsymbol{\theta}_q^*$ است:

$$\boldsymbol{\theta}_q^* = \arg\min_{\theta_q} \tfrac{1}{2}\left(Q(s,a;\theta_q) - r_s^a\right)^2. \qquad (۱٤)$$

گرادیان θ_q به صورت رابطهٔ (۱۵) محاسبه می‌شود:

$$\nabla\theta_q = \left(Q(s,a;\theta_q) - r_s^a\right)\nabla_{\theta_q} Q(s,a;\theta_q). \qquad (۱۵)$$

مطابق رابطهٔ (۱٦)، عمل بهینهٔ a^* به گونه‌ای انتخاب می‌شود که مقدار Q را بیشینه کند:

$$a^* = \arg\max_a Q(s,a;\theta_q). \qquad (۱٦)$$

برای مدیریت درجهٔ کاوشِ الگوریتم در طول فرایند یادگیری، یک سیاست شبه‌حریصانه (ε-greedy) تطبیق داده می‌شود که در آن، پارامتر ε به صورت رابطهٔ (۱۷) تعریف می‌شود:

$$\varepsilon_k := \varepsilon_1 + \tfrac{k-1}{N_e-1}\left(\varepsilon_{N_e} - \varepsilon_1\right), k = 1, \dots, N_e \qquad (۱۷)$$

که در آن، N_e برابر تعداد کل گام‌ها در یک مرحله از فرایند یادگیری است و مقادیر ε_1 و ε_{N_e} به‌ترتیب نشان‌دهندهٔ احتمال کاوش اولیه و نهایی هستند.

الگوریتم DDPG[29] [۲۲] منطبق با معماری عملگر–نقاد و مبتنی بر گرادیان سیاست قطعی است که روی فضای عمل پیوسته عمل می‌کند. در این الگوریتم، یک عملگر حالت محیط s را مشاهده و از طریق نگاشت شبکهٔ عصبی $A(s;\theta_a)$، عمل قطعی a را تولید می‌کند؛ به طوری که θ_a معرف پارامترهای عملگر است. نقاد مقدار Q متناظر با هر زوج حالت–عمل را از طریق شبکهٔ عصبی $C(s_c,a;\theta_c)$ پیش‌بینی می‌کند که در آن، θ_c پارامترهای قابل تنظیم نقاد و s_c حالت محیط است که توسط نقاد مشاهده شده است. نقاد و عملگر با یکدیگر همکاری دارند و سیاست قطعی بهینه را با حل مسئلهٔ بهینه‌سازی توأم رابطه‌های (۱۸) و (۱۹) به دست می‌آورند:

$$\boldsymbol{\theta}_a^* = \arg\max_{\boldsymbol{\theta}_a} C(s_c,a;\boldsymbol{\theta}_c)\ |_{a=A(s;\boldsymbol{\theta}_a)} \qquad (۱۸)$$

$$\boldsymbol{\theta}_c^* = \arg\max_{\boldsymbol{\theta}_c} \tfrac{1}{2}\left(C(s_c,a;\boldsymbol{\theta}_c)|_{a=A(s;\boldsymbol{\theta}_a)} - r_s^a\right)^2. \qquad (۱۹)$$

عملگر تلاش می‌کند تا مقادیر ارزش ارزیابی‌شده توسط نقاد را به بیشینه کند و هدف نقاد دقیق‌ترکردن این ارزیابی است. عملگر و نقاد هر دو مشتق‌پذیر هستند و طبق قاعدهٔ زنجیره، گرادیان آنها به صورت رابطه‌های (۲۰) و (۲۱) محاسبه می‌شود:

$$\nabla\boldsymbol{\theta}_a = \nabla_a C(s_c, a; \boldsymbol{\theta}_c)\ |_{a=A(s;\boldsymbol{\theta}_a)}\ \nabla_{\boldsymbol{\theta}_a} A(s; \boldsymbol{\theta}_a) \quad (۲۰)$$

$$\nabla\boldsymbol{\theta}_c = (C(s_c, a; \boldsymbol{\theta}_c) - r_s^a)\nabla_{\boldsymbol{\theta}_c} C(s_c, a; \boldsymbol{\theta}_c)\ |_{a=A(s;\boldsymbol{\theta}_a)}\ . \quad (۲۱)$$

مطابق رابطهٔ (۲۲)، عمل قطعی به طور مستقیم از طریق عملگر به دست می‌آید:

$$a^* = A(s; \boldsymbol{\theta}_a). \quad (۲۲)$$

مشابه سیاست شبه‌حریصانهٔ پویا، عمل حاصل از کاوش برای مرحلهٔ k در بازهٔ $[0, P_{max}]$ و به صورت رابطهٔ (۲۳) تعریف می‌شود:

$$a := [A(s; \boldsymbol{\theta}_a) + n^k]_0^{P_{max}} \quad (۲۳)$$

که در آن، n^k نویز اضافی است و از توزیع یکنواخت رابطهٔ (۲٤) پیروی می‌کند:

$$n^k \sim u(-\tfrac{P_{max}}{k}, \tfrac{P_{max}}{k})\ . \quad (۲٤)$$

نقاد $C(s_c, a; \boldsymbol{\theta}_c)$ را می‌توان به عنوان یک شبکهٔ کمکی برای انتقال گرادیان در یادگیری در نظر گرفت که پس از اتمام فرایند یادگیری در تولید عمل دخالتی ندارد. به همین دلیل، نقاد با وجود آنکه باید مشتق‌پذیر باشد، لزوماً آموزش‌پذیر نیست. در اینجا، یک نقاد تا حدودی مستقل از مدل پیشنهاد می‌شود که از هر دو امکان بهره‌برداری از دانش پیشین و انعطاف‌پذیری ساختار شبکهٔ عصبی برخوردار است.

الگوریتم گرادیان سیاست REINFORCE بر اساس یک سیاست تصادفی توسعه یافته است، اما نمونه‌برداری در فضای عمل پیوسته یا با ابعاد بالا ناکارآمد است. در الگوریتم DDPG، از گرادیان سیاست قطعی برای غلبه بر این مشکل استفاده می‌شود. از سوی دیگر، با آنکه عملکرد نقاد $C(s_c, a; \boldsymbol{\theta}_c)$ در للگوریتم DDPG با تخمین‌گر ارزش $Q(s, a; \boldsymbol{\theta}_q)$ در الگوریتم DQN مشابه است، این دو ساختار از نظر ورودی‌ها با یکدیگر تفاوت دارند. در الگوریتم DQN، ارزش جفت حالت–عمل با دریافت حالت محیط تخمین زده می‌شود؛ اما در الگوریتم DDPG، این ارزش با دریافت حالت محیط و عمل عامل محاسبه می‌شود.

۴- مدل تنفر از بی‌عدالتی

برای بهبود الگوریتم‌های یادگیری تقویتی چندعاملی، پژوهشگران پاداش‌های دریافتی از محیط را با پاداش‌های داخلی[۳۰] غنی‌سازی می‌کنند تا جنبه‌هایی از رفتار عامل‌ها را در یادگیری لحاظ کنند که لزوماً توسط پاداش محیطی کدگذاری نمی‌شوند. به این ترتیب، برای هر عامل یادگیر در گام زمانی، پاداشی به صورت ترکیب خطی از پاداش محیطی r_{env} و پاداش داخلی $r_{intrinsic}$ به صورت رابطهٔ (۲۵) تعریف می‌شود:

$$r = a\, r_{env} + b\, r_{intrinsic}\ . \quad (۲۵)$$

که در آن، a و b ضرایب ثابت هستند.

مدل تنفر از بی‌عدالتی [۲۳] برای بهبود عملکرد الگوریتم‌های یادگیری تقویتی چندعاملی بر اساس مفاهیم حسادت و عذاب وجدان، پاداش‌های عامل‌ها را دستکاری می‌کند. در این مدل، مطابق رابطهٔ (۲٦)، هر عامل k پاداش خود (r_k) را با پاداش هر یک از دیگر عامل‌ها (r_j) مقایسه می‌کند. نتیجهٔ هر مقایسه یکی از دو وضعیت حسادت یا عذاب وجدان را برای عامل در مقابل عامل همکارش مشخص می‌کند: وضعیت عذاب وجدان در نتیجهٔ یک نابرابری مطلوب[۳۱] رخ می‌دهد، یعنی زمانی که عامل در مقایسهٔ پاداش خود با دیگری به این نتیجه برسد که درآمد بیشتری کسب کرده است $(r_k - r_j > 0)$. حسادت بر اثر یک نابرابری نامطلوب[۳۲] زمانی رخ می‌دهد که عامل نسبت به همکارش پاداش کمتری دریافت کرده باشد $(r_j - r_k > 0)$. در این دو وضعیت، عامل مطابق مدل تنفر از بی‌عدالتی با دستکاری پاداش دریافتی خود، اقدام به تنبیه یا تشویق خود می‌کند. در رابطهٔ (۲٦)، مقدار پاداش نهایی عامل kاُم، U_k تعداد کل عامل‌ها، N و α_k پارامترهای قابل تنظیم β_k مدل هستند.

$$U_k(r_k, \dots, r_N) = r_k - \frac{\alpha_k}{N-1}\sum_{j\neq k}\max(r_j - r_k, 0) - \frac{\beta_k}{N-1}\sum_{j\neq k}\max(r_k - r_j, 0). \quad (۲٦)$$

۵- عناصر یادگیری تقویتی در مسئلهٔ تحت بررسی

انتخاب حالت محیط به عنوان آنچه از اطلاعات محیط توسط هر عامل درک می‌شود، یکی از موضوع‌های بسیار مهم و تأثیرگذار در روند یادگیری عامل‌هاست. بدیهی است، اطلاعات وضعیت فعلی کانال $(\boldsymbol{g}_{n,k}^t)$ که به صورت جزئی

توسط هر عامل درک می‌شود، یکی از مهم‌ترین ویژگی‌های تشکیل‌دهندهٔ حالت محیط است. در اینجا، صورت لگاریتمی این اطلاعات پس از نرمال‌سازی آنها به شکل رابطهٔ (۲۷) استفاده می‌شود:

$$\mathbf{\Gamma}_{n,k}^{t} = \log_2(1 + \frac{\boldsymbol{g}_{n,k}^{t}}{g_{n,n,k}^{t}}) \otimes \mathbf{1}_K \qquad (۲۷)$$

که در آن، $\otimes$ ضرب کرونیکر[۳۳] و $\mathbf{1}_K$ یک بردار است که با تعداد K عدد یک پر شده است. برای نرمال‌سازی مقادیر $\boldsymbol{g}_{n,k}^{t}$، از بهرهٔ کانال لینک پایین‌رو $\mathrm{dl}_{n,k}$ استفاده می‌شود و ترجیح این است که از بازنمایی لگاریتمی این مقادیر استفاده شود؛ زیرا دامنه‌ها معمولاً از نظر مقدار بزرگی تفاوت زیادی دارند. تعداد عناصر $\mathbf{\Gamma}_{n,k}^{t}$ برابر $K(|D_n|+1)$ است و با تغییر در تعداد کاربران هر سلول تغییر می‌کند.

برای آنکه ابعاد ورودی ساختارهای شبکهٔ عصبی در عامل یادگیرنده کاهش داده شوند و متغیربودن تعداد عناصر $\mathbf{\Gamma}_{n,k}^{t}$ مدیریت شود، مجموعهٔ جدید $\tilde{\mathbf{\Gamma}}_{n,k}^{t}$ و اندیس‌های $I_{n,k}^{t}$ از طریق مرتب‌سازی نزولی عناصر $\mathbf{\Gamma}_{n,k}^{t}$ در یک لیست و انتخاب I_c عنصر ابتدای لیست به همراه اندیس آنها در $\mathbf{\Gamma}_{n,k}^{t}$ ایجاد می‌شوند. در واقع، از آنجا که تعداد تداخلات واقعی که در رابطهٔ نسبت سیگنال دریافتی به تداخل و نویز اثرگذار هستند از تعداد عناصر $\mathbf{\Gamma}_{n,k}^{t}$ بسیار کمتر است و سیگنال تداخلی بسیاری از این عناصر به صفر نزدیک است، مرتب‌سازی عناصر $\mathbf{\Gamma}_{n,k}^{t}$ و انتخاب تعدادی از آنها نه فقط فضای حالت را کوچک‌تر می‌کند، بلکه تقریبی از عبارت مربوط به تداخل را در مخرج رابطهٔ نسبت سیگنال دریافتی به تداخل و نویز (رابطهٔ ٤) حاصل می‌کند.

با توجه به اینکه کانال به صورت یک فرایند مارکوف مدل شده است و این فرایند به زمان وابسته است، آخرین حل، یعنی $p_{n,k}^{t-1}$، می‌تواند نقطهٔ شروع بهبودیافته‌ای را برای ادامهٔ یادگیری فراهم کند؛ بنابراین، مقادیر توان انتقال و نرخ لینک پایین‌رو که در زمان قبلی محاسبه شده‌اند، برای تمام عناصر مجموعهٔ $\tilde{\mathbf{\Gamma}}_{n,k}^{t}$ به ورودی ساختار شبکهٔ عصبی عامل یادگیرنده اضافه می‌شوند؛ به طوری که:

$$\tilde{\boldsymbol{p}}_{n,k}^{t-1} = \{p_{n,k}^{t-1} | (n,k) \in I_{n,k}^{t}\}, \qquad (۲۸)$$

$$\tilde{\boldsymbol{C}}_{n,k}^{t-1} = \{C_{n,k}^{t-1} | (n,k) \in I_{n,k}^{t}\}. \qquad (۲۹)$$

در اینجا، فضای حالت با در نظر گرفتن گروه ویژگی رابطهٔ (۳۰) اجرا شده است:

$$f = \{\tilde{\mathbf{\Gamma}}_{n,k}^{t}, \tilde{\boldsymbol{p}}_{n,k}^{t-1}, \tilde{\boldsymbol{C}}_{n,k}^{t-1}\}. \qquad (۳۰)$$

مشابه مراحلی که برای پیش‌پردازش حالت محیط s بیان شد، ورودی ساختار نقاد s_c در الگوریتم DDPG به صورت $\tilde{\boldsymbol{C}}_{n,k}^{t}$ مطابق رابطهٔ (۲۹) تعریف می‌شود که برای ایجاد آن از تعداد I_c عنصر ابتدای لیست نزولی مرتبط‌شدهٔ $\boldsymbol{C}_{n,k}^{t} = \{C_{n,k}^{t} | \forall n,k\}$ استفاده می‌شود.

توان لینک پایین‌رو که مقداری پیوسته و غیرمنفی است و با مقدار توان بیشینهٔ P_{max} محدود شده است، به عنوان عمل هر عامل در نظر گرفته می‌شود. در الگوریتم‌های REINFORCE و DQN، بازهٔ مجاز برای توان انتشار سیگنال به تعدادی سطح توان به صورت پلکانی گسسته‌سازی می‌شود و این سطوح توان انتشار به عنوان عمل‌های گسستهٔ هر عامل یادگیری تقویتی به کار گرفته می‌شوند. از آنجا که توان‌های انتشار در بازهٔ مجاز خود از نظر بزرگی مقدار تفاوت زیادی دارند، برای گسسته‌سازی بهتر توان انتقال از نرمال‌سازی لگاریتمی بهره‌برداری می‌شود؛ بنابراین، مجموعهٔ عمل‌های مجاز یک عامل (A) به صورت رابطهٔ (۳۱) تعریف می‌شود:

$$A \coloneqq \left\{0, \left\{P_{min}\left(\frac{P_{max}}{P_{min}}\right)^{\frac{b}{|A|-2}} | b = 0, \dots, |A|-2\right\}\right\} \qquad (۳۱)$$

که در آن، P_{min} کمینهٔ غیرصفر توان انتشار است و $|A|$ تعداد عمل‌های مجاز عامل یادگیری تقویتی است.

یکی دیگر از روش‌های پایه که از ساختار عملگر–نقاد بهره‌برداری می‌کند الگوریتم DDPG است که در آن، عملگر می‌تواند مقادیر توان انتشار را از یک بازهٔ پیوستهٔ محدودشده تولید کند. این عملگر برای تولید خروجی از رابطهٔ (۳۲) استفاده می‌کند و در آن، همان عمل انتخابی عامل و معادل با توان انتشار تولید شده است:

$$a \coloneqq P_{max} \cdot \frac{1}{1+\exp(-x)}, \qquad (۳۲)$$

به طوری که $\boldsymbol{x}$ خروجی لایهٔ آخر شبکهٔ عصبی عملگر در عامل یادگیرنده است.

تابع پاداش به گونه‌ای طراحی می‌شود که نرخ انتقال

اطلاعات را بهبود ببخشـــد و میزان تداخلات در لینک‌های همســـایه را کاهش دهد. یکی از گزینه‌ها برای طراحی تابع پاداش استفاده از میانگین نرخ مجموع است و بر این اساس اســـتفاده می‌شـــود که مجموع پاداش تمام عامل‌ها با نرخ مجموع شبکه برابر باشد. در تابع پاداش، همسایگی هر عامل لحاظ و تابع پاداش به صـورت محلی برای هر عامل تعریف می‌شود:

$$r_{n,k}^t = C_{n,k}^t + \alpha\left(\sum_{n,k'\neq k} C_{n,k'}^t + \sum_{n'\in D_n,i} C_{n',i}^t\right), \quad (۳۳)$$

که در آن، $\alpha \in \mathbb{R}^+$ ضـــریب وزنی برای اثر تداخل و $\mathbb{R}^+$ معرف مجموعهٔ مقادیر حقیقی مثبت اســـت. اگر تعداد ســـلول‌های شـــبکه (N) به قدر کافی بزرگ باشـــد، جمع پاداش‌های محلی با نرخ مجموع متناسب می‌شود:

$$\sum_{n,k} r_{n,k}^t \propto C(\boldsymbol{g}^t, \boldsymbol{p}^t). \quad (۳٤)$$

در ترکیب مدل تنفر از بی‌عدالتی با روش‌های یادگیری تقویتی، مقادیر حاصـــل از رابطهٔ (۳۳) برای هر کدام از عامل‌های kاُم و jاُم جایگزین مقادیر r_k و r_j در رابطهٔ (۲٦) می‌شوند. با وجود آنکه در مدل تنفر از بی‌عدالتی، اثرگذاری نابرابری‌های مطلوب و نامطلوب در تغییرات مقدار پاداش عامل ممکن است به صورت تنبیه یا تشویق باشد، تا آنجا که بررسی‌ها نشان می‌دهند، تمام پژوهش‌های قبلی فقط از بخشی از مدل و آن هم با اثر تنبیهی بهره‌برداری می‌کنند. در این پژوهش، مجموعه‌ای از مقادیر به عنوان ضرایب مدل در نظر گرفته می‌شـــوند و اثر آنها در یادگیری عامل‌ها و بهبود نتایج بررســـی می‌شـــود. مطابق نتایجی که در ادامه ارائه می‌شود، استفاده از ضرایب منفی و تشویق عامل به کمک افزودن ضـریبی مثبت از تفاوت‌های محاسـبه‌شـده به پاداش عامل در برخی از روش‌های یادگیری تقویتی در مســـئلهٔ تخصـیص منابع شـبکه‌های بی‌سـیم می‌تواند به بهبود نتایج کمک کند.

برای وضـــوح بیشـــتر کاربرد مدل تنفر از بی‌عدالتی در روش‌های یادگیری تقویتی، گام‌های الگوریتم DQN برای عامل k به عنوان نمونه در زیر آورده شده‌اند:

1. *Input*: Episode times N_e, exploration times T, learning rate η_q, initial and final exploration probability ε_1, ε_{N_e}.
2. *Initialization*: Initialize DQN $Q(s,a;\theta_q)$ with random parameters θ_q.
3. **for** $k = 1$ *to* N_e **do**
4. Update ε_k by (17).
5. Receive initial state s^1.
6. **for** $t = 1$ *to*T **do**
7. **if** rand() < ε_k then
8. Randomly select action $a^t \in A$ with uniform probability.
9. **else**
10. Select action a^t by (16).
11. **end if**
12. Execute action a^t, achieve all agents' rewards $r_1^t,\ldots, r_N^t$, and observe new state s^{t+1}.
13. Calculate $U_k^t(r_k^t,\ldots,r_N^t)$ by (26) as r^t.
14. Calculate gradient $\nabla\theta_q$ by (15),
15. Update parameter along negative gradient direction: $\theta_q \leftarrow \theta_q - \eta_q\nabla\theta_q$.
16. $s^t \leftarrow s^{t+1}$
17. **end for**
18. **end for**
19. Output: learned DQN $Q(s,a;\theta_q)$.

۶- آزمایش‌ها و نتایج

برای انجام آزمایش‌ها و مقایســـهٔ الگوریتم‌های یادگیری تقویتی، از تنظیمات بیان‌شـده در جدول (۱) برگرفته از [۱۱، ۱۲] اسـتفاده شـده اسـت. تنظیمات مربوط به سـاختارهای شـــبکهٔ عصـــبی عمیق از جمله توابع فعال‌ســـازی و تعداد نرون‌ها در هر لایه در جدول (۲) آورده شده است.

در مقایســـهٔ الگوریتم‌های یادگیری و ارزیابی عملکرد آنها، از معیار نرخ مجموع اســـتفاده می‌شـــود. از آنجا که مقداردهی اولیهٔ پارامترها و تولید نمونه‌های آموزشـــی در آزمایش‌ها به صـورت تصـادفی انجام می‌شـود، در سـنجش عملکرد الگوریتم از میانگین نتایج چندین آزمایش اســـتفاده می‌شـود. نمادهای $\bar{C}$ ، σ_c^2 و $\bar{C}^*$ به‌ترتیب به عنوان واریانس نرخ مجموع، میانگین نرخ جمع و میانگین نرخ مجموع، ۲۰ درصـد برتر نسـبت به تکرارهای مسـتقل آزمایش‌ها تعریف می‌شـوند. معیار $\bar{C}^*$ یک شـاخص برای اندازه‌گیری عملکرد الگوریتم‌هایی است که به‌خوبی آموزش دیده‌اند.

در اینجا، نتایج حاصـل از آموزش سـه الگوریتم یادگیری تقویتی عمیق DDPG، DQN و REINFORCE کـه بـه عنوان روش‌های پایه معرفی شـدند، با اسـتفاده از تابع پاداش حاصـل از مدل تنفر از بی‌عدالتی گزارش می‌شـود. برای هر

یک از روش‌های پایه، ترکیب‌هایی متفاوت از مقادیر $\alpha, \beta \in \{-1, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 1\}$ در مدل تنفر از بی‌عدالتی لحاظ می‌شوند و عامل‌ها به کمک آن روش پایه در محیط آموزش تحت یادگیری قرار می‌گیرند. در فرایند یادگیری، ۵۰۰۰ مرحلهٔ ۱۰گامی در نظر گرفته می‌شوند، معیار متوسط نرخ مجموع پس از هر ۱۰۰ مرحله محاسبه و با بهترین نتیجهٔ قبلی مقایسه می‌شود و در صورت بهبود نتیجه، ساختار یادگرفته‌شده به عنوان بهترین ساختار به همراه نتیجهٔ حاصل‌شده برای مقایسه‌های بعدی ذخیره می‌شوند.

جدول (۱): پارامترهای شبیه‌سازی محیط در آزمایش‌ها

عنوان	متغیر	مقدار
تعداد سلول‌ها در شبکه	N	25
بیشترین تعداد کاربران در هر سلول	K	4
فضای درونی	R_{min}	km 0.01
نیم‌فاصلهٔ سلول به سلول	R_{max}	km 1
فرکانس داپلر	f_d	10 Hz
دورهٔ زمانی*	T_s	20 ms
محوشدگی مقیاس بزرگ**	β	$-120.9 - 37.6 \log_{10} d + 10 \log_{10} z$
توان نویز گوسی سفید جمع‌شونده	σ^2	−114 dBm
کران پایین توان انتشار	P_{min}	5 dBm
کران بالای توان انتشار	P_{max}	38 dBm
بیشینهٔ نسبت سیگنال به تداخل و نویز	$SINR$	30 dB
تعداد سلول‌های همسایه برای هر سلول	$\|Dn\|, \forall n$	18
تعداد تداخلات انتخابی از ابتدای لیست‌های مرتب‌شده	I_c	16
تعداد سطوح توان (تعداد عمل ممکن برای هر عامل)	$\|A\|$	10
ابعاد فضای حالت با در نظر گرفتن ویژگی‌های f_1 و f_2	$\|S\|$	48, 32
ضریب وزن	α	1
تعداد مراحل[۳٤] یادگیری	N_e	5000
تعداد شیارهای زمانی در هر مرحله	T	10

* برای شبیه‌سازی اثرات محوشدگی به کار گرفته شده است.

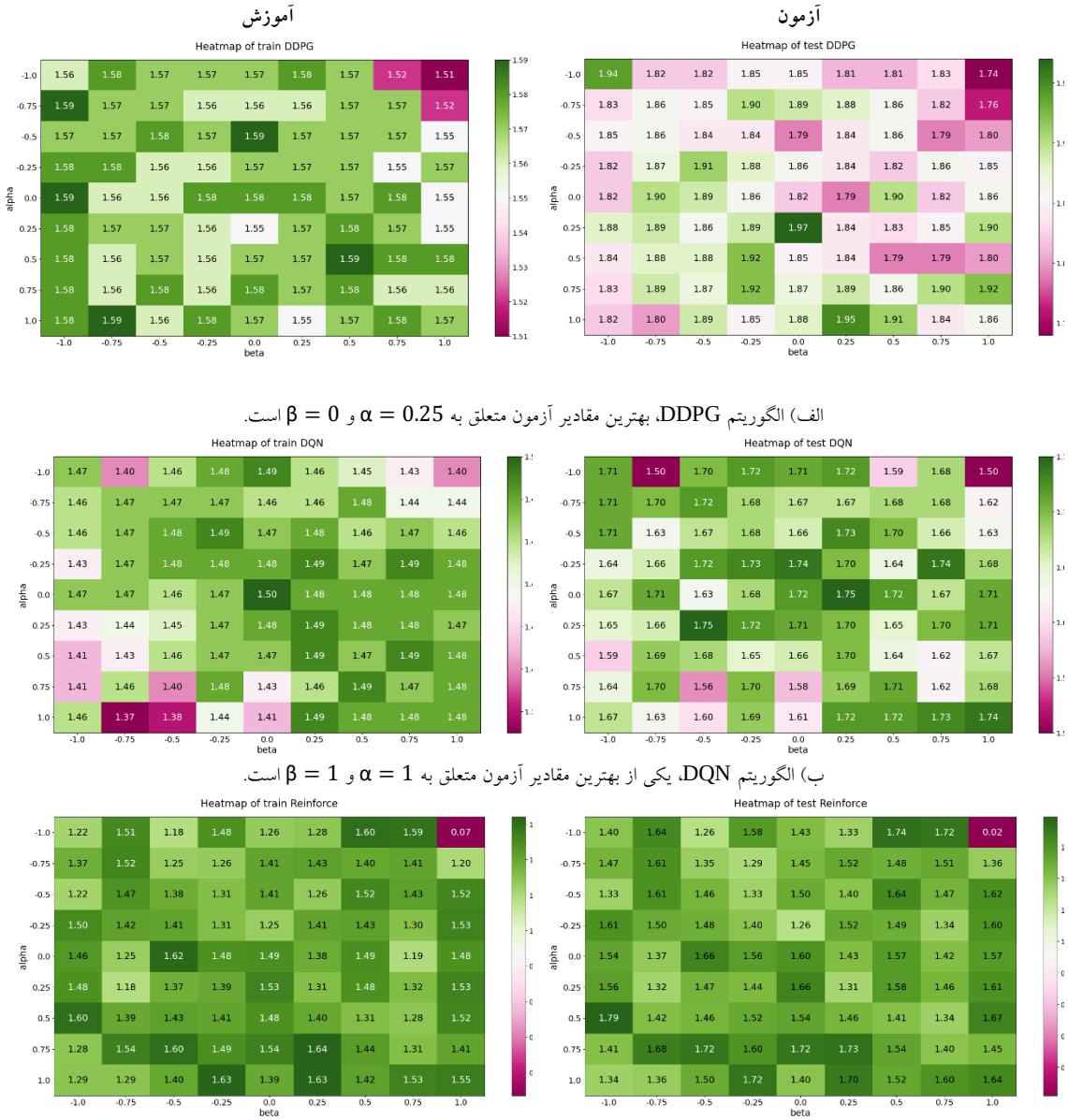
** در این رابطه، d طول لینک است و متغیر تصادفی z با توزیع لگاریتمی نرمال از $\ln z \sim N(0, \sigma_z^2)$ پیروی می‌کند که در آن، σ_z^2 برابر $8\ dB$ است.

جدول (۲): تنظیمات یادگیری و پارامترهای ساختارهای شبکهٔ عصبی عمیق در الگوریتم‌ها

تنظیمات	الگوریتم			
	REINFORCE	DQN	DDPG	
			Actor	Critic
نرخ یادگیری	$\eta_\pi = 1 \times 10^{-4}$	$\eta_q = 1 \times 10^{-3}$	$\eta_a = 1 \times 10^{-4}$	$\eta_c = 1 \times 10^{-3}$
روش کاوش	رابطهٔ (۱۷)	$\varepsilon_1 = 0.2$ $\varepsilon_{N_e} = 1 \times 10^{-4}$	رابطهٔ (۲۳)	-
لایهٔ خروجی شبکهٔ عصبی	$softmax, \|A\|$	$linear, \|A\|$	رابطهٔ (۳۲) , 1	$linear, 1$
لایهٔ پنهان شبکهٔ عصبی	$ReLU, 64$ $softmax, 128$	$ReLU, 64$ $softmax, 128$	$ReLU, 64$ $softmax, 128$	$ReLU, 64$
لایهٔ ورودی شبکهٔ عصبی	$linear, \|S\|$	$linear, \|S\|$	$linear, \|S\|$	$linear, I_c$

$ReLU$: $f(x) = \max(0, x)$, $linear$: $f(x) = x$, $softmax$: $f(x) = \frac{e^x}{\sum e^x}$

پس از پایان فرایند یادگیری، ساختار یادگرفته‌شدهٔ عامل‌ها در محیط آزمون ارزیابی می‌شود. برای فرایند آزمون، ۵۰۰ مرحلهٔ ۳۰۰گامی در نظر گرفته شده‌اند که ارزیابی نتایج آنها با در نظر گرفتن معیار $\bar{C}$ انجام می‌شود. شکل (۱) این نتایج را برای فرایندهای آموزش و آزمون هر یک از روش‌های یادگیری تقویتی به صورت نمودارهای حرارتی و با توجه به مقادیر مختلف پارامترهای مدل تنفر از بی‌عدالتی نشان می‌دهد. بهبود نتایج $\bar{C}$ با استفاده از برخی از ترکیب‌های پارامترهای α و β برای هر یک از روش‌های پایه در این نمودارها مشهود است. با توجه به اینکه در برخی از پژوهش‌ها، علاوه بر معیار $\bar{C}$ از معیار $\bar{C}^*$ هم استفاده شده است، شکل (۲) مقایسهٔ نتایج را با توجه به این دو معیار به صورت نمودارهای خطی برای ۸۱ ترکیب مختلف پارامترهای α و β نشان می‌دهد.

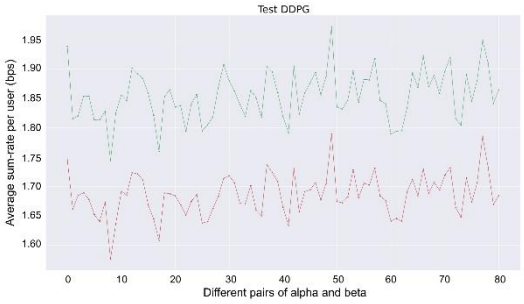


الف) الگوریتم DDPG، بهترین مقادیر آزمون متعلق به $\alpha = 0.25$ و $\beta = 0$ است.

ب) الگوریتم DQN، یکی از بهترین مقادیر آزمون متعلق به $\alpha = 1$ و $\beta = 1$ است.

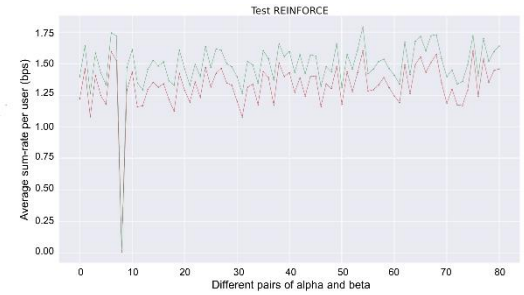
پ) الگوریتم REINFORCE، یکی از بهترین مقادیر آزمون متعلق به $\alpha = 0.75$ و $\beta = 0.25$ است.

شکل (۱): مقایسهٔ معیار $\bar{C}$ در فرایند آموزش و آزمون عملکرد ساختار یادگرفته‌شده توسط هر یک از الگوریتم‌ها با در نظر گرفتن ترکیب‌هایی متفاوت از مقادیر پارامترهای α و β در مدل تنفر از بی‌عدالتی. در این مقایسه، بهترین مقادیر با رنگ سبز تیره قابل مشاهده هستند. ترکیب $\alpha = 0$ و $\beta = 0$ معرف نتیجهٔ حاصل از اعمال روش پایهٔ هر الگوریتم است.



الف) الگوریتم DDPG

ب) الگوریتم DQN



پ) الگوریتم REINFORCE

شکل (۲): مقایسهٔ معیار $\bar{C}$ (قرمز رنگ) و $\bar{C}^*$ (سبز رنگ) در سنجش عملکرد ساختار یادگرفته‌شده به‌ازای مقادیر مختلف پارامترهای α و β در مدل تنفر از بی‌عدالتی در محیط آزمون.

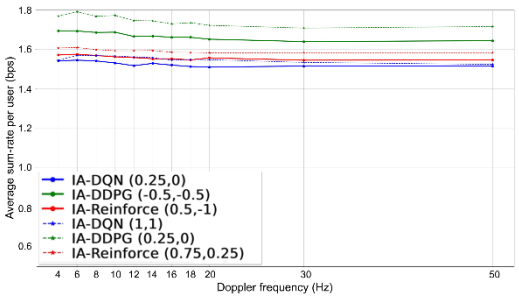
برای انجام آزمایش‌های تعمیم‌پذیری[۳۵]، از میان ترکیب‌هایی از پارامترهای مدل تنفر از بی‌عدالتی که نتایج آزمون بهتری داشته‌اند، دو ترکیب انتخاب شده‌اند و عملکرد آنها در محیط‌های شبیه‌سازی با مشخصات متفاوت بررسی شده است. جدول (۳) مقادیر معیار $\bar{C}$ را برای هر یک از این ترکیب‌ها با روش پایه مقایسه می‌کند. در این آزمایش‌ها، از مشخصات بیان‌شده در جدول (۱) استفاده شده است.

ارزیابی تعمیم‌پذیری الگوریتم‌های یادگیری تقویتی با مدل تنفر از بی‌عدالتی در شکل (۳) که با توجه به مقادیر مختلف پارامترهای شبکه انجام شده است، نشان می‌دهد از میان ترکیب‌های بیان‌شده در جدول (۳)، بهترین نتایج مربوط به الگوریتم DDPG با پارامترهای $\alpha = 0.25$ و $\beta = 0$ است.

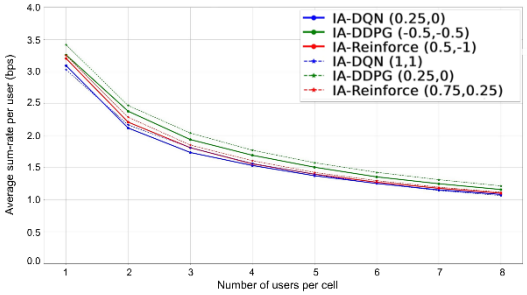
نتایج تعمیم‌پذیری بهترین ترکیبات پارامترهای مدل تنفر از بی‌عدالتی در هر روش یادگیری تقویتی با روش‌های پایه در شکل (٤) برتری بهره‌برداری از این مدل را در الگوریتم‌های DDPG و REINFORCE به‌وضوح نشان می‌دهد. در این شکل، نتایج استفاده از بیشترین توان و کاربرد توان تصادفی به عنوان دو روش اولیه و نتایج دو الگوریتم مبتنی بر مدل به نام‌های برنامه‌ریزی کسری [۷] و حداقل وزن‌دار میانگین مربعات خطا [۲] به عنوان روش‌های محک با نتایج الگوریتم‌های یادگیری تقویتی مورد بحث مقایسه شده است.

جدول (۳): مقایسهٔ نتایج برخی از بهترین ترکیب‌های پارامترهای مدل تنفر از بی‌عدالتی در روش‌های پایه

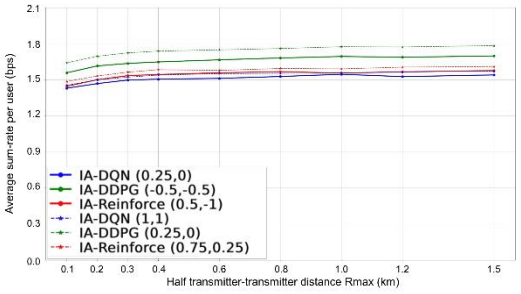
متغیر ارزیابی	REINFORCE	IA-REINFORCE		DQN	IA-DQN		DDPG	IA-DDPG	
		$\alpha = 0.5$ $\beta = -1$	$\alpha = 0.75$ $\beta = 0.25$		$\alpha = 1$ $\beta = 1$	$\alpha = 0.25$ $\beta = 0$		$\alpha = 0.25$ $\beta = 0$	$\alpha = -0.5$ $\beta = -0.5$
$\bar{C}$ در آموزش	1.49	1.60	1.64	1.50	1.48	1.48	1.58	1.55	1.58
$\bar{C}$ در آزمون	1.60	1.79	1.73	1.72	1.74	1.71	1.82	1.97	1.84



(الف) فرکانس داپلر

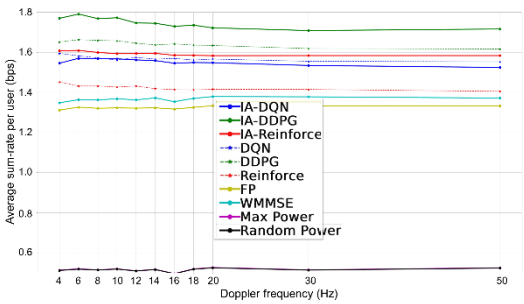


(ب) بیشترین تعداد کاربران در هر سلول شبکه

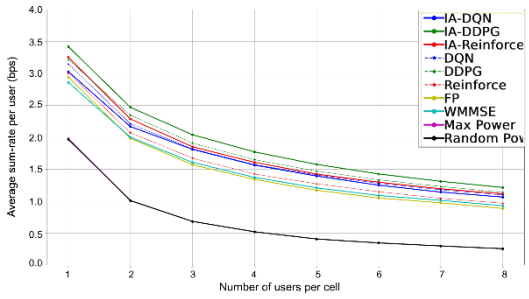


(پ) نیم فاصله سلول به سلول

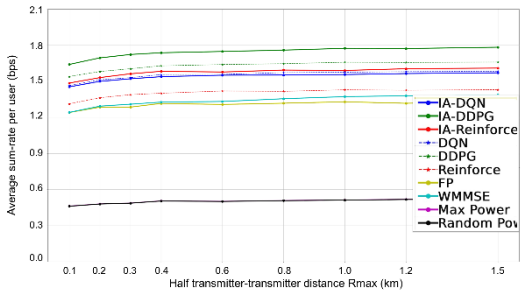
شکل (۳): بررسی تعمیم‌پذیری الگوریتم‌های یادگیری تقویتی با توجه به پارامترهای انتخابی مدل تنفر از بی‌عدالتی. این بررسی با توجه به مقادیر مختلف پارامترهای شبکه انجام شده است.



(الف) فرکانس داپلر



(ب) بیشترین تعداد کاربران در هر سلول شبکه



(پ) نیم‌فاصلهٔ سلول به سلول

شکل (٤): بررسی تعمیم‌پذیری بهترین ترکیب پارامترهای مدل تنفر از بی‌عدالتی در هر یک از الگوریتم‌های یادگیری تقویتی و مقایسهٔ نتایج با روش‌های پایه. این بررسی با توجه به مقادیر مختلف پارامترهای شبکه انجام شده است. در این شکل، منحنی‌های مربوط به الگوریتم‌های یادگیری تقویتی پایه به صورت خط‌چین نشان داده شده‌اند.

۷- نتیجه‌گیری

در این مقاله، عملکرد الگوریتم‌های یادگیری تقویتی عمیق برای حل مسئلهٔ تخصیص پویای توان توزیع‌شده در یک شبکهٔ سلولی بی‌سیم با کانال دسترسی چندگانهٔ تداخلی با هدف بیشینه‌سازی نرخ مجموع بررسی شد. در تنظیمات شبکهٔ تحت بررسی، همکاری بین‌سلولی در نظر گرفته شد و برای مدیریت فرایند یادگیری، از روش آموزش متمرکز و اجرای توزیع‌شده بهره‌برداری شد. مدل تنفر از بی‌عدالتی به عنوان یک ساختار پاداش داخلی برای بهبود عملکرد الگوریتم‌های یادگیری تقویتی چندعاملی پیشنهاد و اثربخشی آن در بیشینه‌سازی نرخ مجموع مطالعه شد. نتایج آزمایش‌ها نشان می‌دهد تنظیم پارامترهای مدل تنفر از بی‌عدالتی تأثیری به‌سزا در بهبود عملکرد این الگوریتم‌ها به ویژه الگوریتم‌های DDPG و REINFORCE دارد. ارزیابی تعمیم‌پذیری نیز نشان داد بهره‌برداری از این پاداش داخلی با ضرایب منفی و

تشـویق عامل به کمک افزودن ضـریبی مثبت از تفاوت‌های محاسـبه‌شـده به پاداش عامل در فرایند یادگیری الگوریتم DDPG بهترین عملکرد را دارد. علاوه بر این، همهٔ رویکردهای مبتنی بر داده از روش‌های مبتنی بر مدل پیشرفته بهتر عمل می‌کنند و همچنین، عملکرد تعمیم‌پذیری بهتری دارند.

الگوریتم‌های یادگیری تقویتی عمیق به عنوان دسته‌ای از الگوریتم‌های مبتنی بر داده، راهکاری امیدوارکننده برای شـبکه‌های هوشـمند آینده اسـت و الگوریتم DDPG ترکیب‌شده با مدل تنفر از بی‌عدالتی را می‌توان برای مسائلی متنوع با فضای حالت/عمل گسسته یا پیوسته و مسائل بهینه‌سازی توأم متغیرهای متعدد اعمال کرد. این الگوریتم را می‌توان برای بسیاری از مسائل مانند زمان‌بندی کاربر، مدیریت کانال و تخصیص توان در شبکه‌های ارتباطی مختلف اعمال کرد.

مراجع

[1] A. Gjendemsjo, D. Gesbert, G. E. Oien, S. G. Kiani, "Binary power control for sum rate maximization over multiple interfering links", IEEE Trans. on Wireless Communications, Vol. 7, No. 8, August 2008. https://doi.org/10.1109/TWC.2008.070227

[2] Q. Shi, M. Razaviyayn, Z. -Q. Luo, C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel", IEEE Trans. on Signal Processing, Vol. 59, No. 9, September 2011. https://doi.org/10.1109/TSP.2011.2147784

[3] L. Song, D. Niyato, Z. Han, E. Hossain, "Game-theoretic resource allocation methods for device-to-device communication", IEEE Wireless Communications, Vol. 21, No. 3, June 2014. https://doi.org/10.1109/MWC.2014.6845058

[4] N. Naderializadeh, A. S. Avestimehr, "ITLinQ: A new approach for spectrum sharing in device-to-device communication systems", IEEE Journal on Selected Areas in Communications, Vol. 32, No. 6, June 2014. https://doi.org/10.1109/jsac.2014.2328102

[5] X. Yi, G. Caire, "ITLinQ+: An improved spectrum sharing mechanism for device-to-device communications", 49th Asilomar Conference on Signals, Systems and Computers, pp. 1310-1314, November 2015. https://doi.org/10.1109/ACSSC.2015.742135

[6] K. Shen, W. Yu, "FPLinQ: A cooperative spectrum sharing strategy for device-to-device communications", IEEE International Symposium on Information Theory (ISIT), pp. 2323-2327, June 2017. https://doi.org/10.1109/ISIT.2017.8006944

[7] K. Shen, W. Yu, "Fractional programming for communication systems—Part I: Power control and beamforming", IEEE Trans. on Signal Processing, Vol. 66, February 2018. https://doi.org/10.1109/TSP.2018.2812733

[8] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, "Human-level control through deep reinforcement learning", Nature, Vol. 518, February 2015. https://doi.org/10.1038/nature14236

[9] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, "Mastering the game of Go with deep neural networks and tree search", Nature, Vol. 529, January 2016. https://doi.org/10.1038/nature16961

[10] C. Berner, G. Brockman, B. Chan, V. Cheung, P. Dębiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, "Dota 2 with large scale deep reinforcement learning", arXiv preprint, December 2019. https://doi.org/10.48550/arXiv.1912.06680

[11] F. Meng, P. Chen, L. Wu, J. Cheng, "Power allocation in multi-user cellular networks: Deep reinforcement learning approaches", IEEE Trans. on Wireless Communications, Vol. 19, June 2020. https://doi.org/10.1109/TWC.2020.3001736

[12] F. Meng, P. Chen, L. Wu, "Power allocation in multi-user cellular networks with deep Q learning approach", IEEE International Conference On Communications (ICC), pp. 1-6, December 2018. https://doi.org/10.1109/ICC.2019.8761431

[13] Z. Wang, J. Zong, Y. Zhou, Y. Shi, V. W. Wong, "Decentralized multi-agent power control in wireless networks with frequency reuse", IEEE Trans. on Communications, Vol. 70, December 2021. https://doi.org/10.1109/TCOMM.2021.3135540

[14] Y. Zhang, D. Guo, "Multi-Agent Reinforcement Learning for Multi-Cell Spectrum and Power Allocation", IEEE Trans. on Communications, Vol. 99, January 2025. https://doi.org/10.1109/TCOMM.2025.3534565

[15] A. Kopic, E. Perenda, H. Gacanin, "A

collaborative multi-agent deep reinforcement learning-based wireless power allocation with centralized training and decentralized execution", IEEE Trans. on Communications, Vol. 72, November 2024. https://doi.org/10.1109/TCOMM.2024.3409530

[16] H. Kim, J. So, "Distributed multi-agent deep reinforcement learning-based transmit power control in cellular networks", Sensors, Vol. 25, No. 13, June 2025. https://doi.org/10.3390/s25134017

[17] Z. Huang, T. Li, C. Song, Z. Li, J. Wang, X. Liu, H. Chen, X. Zhao, Y. Cao, "Joint spectrum and power allocation scheme based on value decomposition networks in D2D communication networks", EURASIP Journal on Wireless Communications and Networking, Vol. 79, September 2024. https://doi.org/10.1186/s13638-024-02393-1.

[18] A. Reuel, D. Ma, "Fairness in reinforcement learning: A survey", Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 1218–1230, 7 February 2025. https://doi.org/10.5555/3716662.3716769

[19] F. Alamiyan-Harandi, P. Ramazi, "Environmental-impact-based multi-agent reinforcement learning", Applied Sciences, Vol. 14, No. 15, July 2024. https://doi.org/10.3390/app14156432

[20] M. Hassanjani, F. Alamiyan-Harandi, P. Ramazi, "Inequity aversion reduces travel time in the traffic light control problem", arXiv preprint, 2023. https://doi.org/10.48550/ arXiv:2302.12053

[21] R. S. Sutton, D. McAllester, S. Singh, Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation", Advances in Neural Information Processing Systems, Vol. 12, 1999. https://proceedings.neurips.cc/paper_files/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf

[22] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, D. Wierstra, "Continuous control with deep reinforcement learning", arXiv preprint, 2015. https://doi.org/1048550/ arXiv:1509.02971

[23] E. Hughes, J. Z. Leibo, M. Phillips, K. Tuyls, E. Dueñez-Guzman, A. García Castañeda, I. Dunning, T. Zhu, K. McKee, R. Koster, "Inequity aversion improves cooperation in intertemporal social dilemmas", Advances in Neural Information Processing Systems, Vol. 31, 2018. https://proceedings.neurips.cc/paper_files/paper/2018/file/7fea637fd6d02b8f0adf6f7dc36aed93-Paper.pdf

---

[1] Weighted Minimum Mean Squared Error (WMMSE)
[2] Fractional Programming (FP)
[3] Downlink
[4] Deep Q Network
[5] Multi Agent Reinforcement Learning (MARL)
[6] Centralized Training Distributed Execution (CTDE)
[7] Long Short Term Memory (LSTM)
[8] Social welfare functions
[9] Value Decomposition Networks
[10] Inequity Aversion (IA)
[11] Cleanup and Harvest Games
[12] Envy
[13] Guilt
[14] Interfering Multiple-Access Channel (IMAC)
[15] Sum-rate
[16] Single-Input Single-Output (SISO) Interfering Broadcast Channel (IBC)
[17] Base Station (BS)
[18] Geometric attenuation
[19] Shadow fading
[20] Flat fading
[21] Signal-to-Interference-plus-Noise Ratio (SINR)

[22] NP hard
[23] Channel State Information (CSI)
[24] Multi-objective programming
[25] Transfer Learning
[26] Monte-Carlo
[27] Stochastic gradient ascent
[28] Deep Q Network
[29] Deep Deterministic Policy Gradient
[30] Intrinsic Rewards
[31] Advantageous Inequity
[32] Disadvantageous Inequity
[33] Kronecker product
[34] Episode
[35] Generalization