دومین کنگره مشترک سیستمهای فازی و هوشمند ایران
2nd Joint Congress on Fuzzy and Intelligent Systems
28-30 October 2008 ۷ الی ۹ آبان ماه ۱۳۸۷

بررسی ارتباط طول عمر کنترل کننده و شیوه ی تصمیم گیری بهینه در حل مساله ی MAB با استفاده از یادگیری تقویتی

سیدمصطفی کلامی هریس[*] محمدباقر نقیبی سیستانی
دانشکده‌ی مهندسی دانشگاه فردوسی مشهد استادیار دانشکده‌ی مهندسی دانشگاه فردوسی مشهد
mb-naghibi@ferdowsi.um.ac.ir [*]m.kalami@gmail.com

چکیده - مسأله‌ی MAB کاربردهای وسیعی در علوم مهندسی، آمار، اقتصاد و روان‌شناسی دارد و در شاخه‌های مختلف علمی و فنی به اشکال متفاوتی ظاهر می‌شود. یکی از چالش‌هایی که در حل این مساله وجود دارد لزوم برقراری تعادل میان بهره‌جویی از اطلاعات فعلی و کسب اطلاعات جدید از محیط می‌باشد. این دو پدیده، به ترتیب به نام‌های بهره‌برداری (Exploitation) و جستجو (Exploration) معروف هستند. روشی که در این مقاله برای حل مسأله‌ی MAB به کار رفته است، روش یادگیری تقویتی است. این روش، که نوعی رویکرد یادگیری غیر نظارت شده را پیاده‌سازی می‌کند، این امکان را فراهم می‌کند که با تغییر پارامترهای تصمیم‌گیری، تعادل مطلوب بین پدیده‌های جستجو و بهره‌برداری به وجود بیایند. در این نوشتار، با انجام آزمایش‌های متعدد، ارتباط میان پارامترهای تصمیم‌گیری و طول بازه‌ی زمانی برای حل مسأله، که به طول عمر کنترل کننده یا عامل یادگیرنده معروف است، مورد بررسی قرار گرفته است.

کلید واژه- Multi-Armed Bandit (MAB)، تخصیص منابع، تصمیم گیری، یادگیری تقویتی، یادگیری ماشینی.

۱- مقدمه

مسأله‌ی Multi-Armed Bandit یا به اختصار MAB، در رشته‌های مختلف مهندسی، آمار، روان‌شناسی و اقتصاد مورد بررسی قرار گرفته و دارای کاربردهای وسیعی است. در علم آمار، این مسأله با عنوان طراحی آزمایش‌ها مطرح است که توسط تامپسون[1] (۱۹۳۳) و سپس به طور کامل‌تری توسط رابینز[2] (۱۹۵۲) معرفی شده است. این مسأله یک مدل آماری است که برای تصمیم‌گیری ارائه شده است و همزمان با این که سعی می‌شود از اطلاعات موجود برای اتخاذ بهترین تصمیمات ممکن استفاده شود، اطلاعات جدیدی نیز از محیط کسب می‌شوند که برای تصمیم‌گیری‌های آتی مورد استفاده قرار می‌گیرند [1, 4, 5]. این مسأله از دیدگاه آماری، توسط بلمن[3] (۱۹۵۶) و سپس به نحوی وسیع‌تر، توسط بِری[4] و فریستِدت[5] (۱۹۸۵) مورد مطالعه قرار گرفته است [3]. نارِندرا[6] و تاتاکار[7] (۱۹۸۹) این مسأله را با چند مهندسی مورد بررسی قرار دادند که همراه با بحث‌های کاملی در خصوص کاربردهای مختلف آن و مسائل مرتبط بود [1, 2]. این مسأله در علم روان‌شناسی، نقش مهمی در بحث یادگیری آماری دارد که توسط استس[8] (۱۹۵۰) و سپس توسط بوش[9] و موستلر[10] (۱۹۸۵) مورد مطالعه قرار گرفته است [1].

این مسأله را می‌توان نوع خاصی از مسائل تخصیص منابع به صورت دنباله‌ای در نظر گرفت [7]. در این نوع از مسائل، یک یا چند منبع به چندین دستگاه، ماشین و یا پروژه اختصاص داده می‌شوند و هدف از حل مسأله، بهینه کردن عملکرد سیستم با معیاری مشخص است. مسأله‌ی MAB در موضوعات مختلف علمی و فنی ظاهر می‌شود و کاربردهای فراوانی نظیر سازمان‌دهی حس‌گرها، سیستم‌های تولیدی،

[1] Thompson
[2] Robbins
[3] Bellman
[4] Berry
[5] Fristedt
[6] Narendra
[7] Thathachar
[8] Estes
[9] Bush
[10] Mosteller

اقتصاد، شبکه‌های ارتباطی، آزمایش‌های پزشکی، نظریه‌ی کنترل، و نظریه‌ی جستجو برای آن مطرح شده‌اند [7]. قضیه‌ی مهمی که گیتینس[11] (۱۹۷۴) در رابطه با این مسأله مطرح کرده است، کاربردهای فراوانی در اقتصاد دارد [3, 5].

در این نوشتار، مسأله‌ی MAB با استفاده از روش یادگیری تقویتی حل شده است. با اعمال تغییرات بر روی پارامترهای تصمیم‌گیری، آزمایش‌های متعددی انجام گرفته‌اند، و اثرات تغییر پارامترهای تصمیم‌گیری بر عملکرد عامل یادگیرنده، بررسی شده‌اند. سایر بخش‌های این مقاله به این ترتیب هستند. در بخش ۲، تعریفی جامع از مسأله‌ی MAB ارائه شده است. در بخش ۳، مروری بر یادگیری تقویتی انجام گرفته است. بخش ۴ نیز حاوی نتایج به دست آمده از آزمایش‌های انجام شده است.

۲- تعریف مسأله‌ی MAB [1,2,4,5,7,9,10]

مسأله‌ی MAB در حالت کلاسیک، از k دستگاه و یک کنترل کننده تشکیل شده است. در هر زمان، کنترل کننده می‌تواند فقط یک دستگاه را به کار بگیرد. در این حالت، بقیه‌ی دستگاه‌ها، دست نخورد و ثابت باقی می‌مانند. فرض کنید $t = 0,1,2,\ldots$ نشان دهنده‌ی زمان گسسته باشد. تعداد دفعاتی که دستگاه i تا زمان t توسط کنترل کننده به کار گرفته شده است، با $N_i(t)$ نشان داده می‌شود. حالت دستگاه i در زمان t، تابعی از $N_i(t)$ یه صورت $X_i\left(N_i(t)\right)$ است. خروجی هر دستگاه به صورت تابعی از حالت دستگاه به شکل $R_i\left(X_i\left(N_i(t)\right)\right)$ تعریف می‌شود. هر دستگاه به شکل مجموعه‌ای از زوج‌های مرتب (X_i, R_i) و به صورت زیر قابل تعریف است:

$$M_i = \left\{\left(X_i\left(N_i(t)\right), R_i\left(X_i\left(N_i(t)\right)\right)\right)\right\} \quad N_i(t) = 0,1,2,\ldots,t \quad t = 0,1,2,\ldots \qquad (۱)$$

فرض کنید $U(t) = \left(U_1(t), U_2(t), \ldots, U_k(t)\right)$ عملی است که کنترل کننده در زمان t انجام می‌دهد. مقادیری که $U(t)$ می‌گیرد، همگی بردارهای یکه‌ای هستند که فقط یک مولفه‌ی غیر صفر دارند، که مقدار این مولفه، همواره برابر با یک است و مربوط به دستگاهی است که توسط کنترل کننده انتخاب شده است. فرض کنید $H_i(t) = \left(X_i(0), \ldots, X_i\left(N_i(t)\right)\right)$ باشد. حالت جدید دستگاه i در زمان $t+1$ به صورت زیر تعریف می‌شود:

$$X_i\left(N_i(t+1)\right) = \begin{cases} f_{N_i(t)}\left(H_i(t)\right) & ,U_i(t) = 1 \\ X_i\left(N_i(t)\right) & ,U_i(t) = 0 \end{cases} \qquad (۲)$$

و یا به صورت خلاصه‌تر

$$X_i\left(N_i(t+1)\right) = f_{N_i(t)}\left(H_i(t)\right)U_i(t) + X_i\left(N_i(t)\right)\left(1 - U_i(t)\right) \qquad (۳)$$

که در آن $N_i(t+1) = N_i(t) + U_i(t)$ و تابع $f_{N_i(t)}(.)$ یک تابع معلوم و وابسته به $N_i(t)$ می‌باشد.

خروجی دستگاه i در زمان t عبارت است از:

$$r_i(t) = R_i\left(X_i\left(N_i(t)\right)\right)U_i(t) \qquad (۴)$$

و خروجی کل سیستم در زمان t، به صورت

$$r(t) = \sum_{i=1}^{k} r_i(t) \qquad (۵)$$

تعریف می‌شود. فرض کنید شاخصی به صورت $z(t) = z\left(r(0), r(1), \ldots, r(t)\right)$ برای ارزیابی عملکرد سیستم تعریف شده باشد. هدف از حل مسأله‌ی MAB، یافتن دنباله‌ی کنترلی $\{U(0), \ldots, U(T)\}$ است به نحوی که $z(T)$ بهینه شود. منظور از T بازه‌ای از زمان است که کنترل کننده موظف است سیستم را کنترل کند. از T به عنوان *طول عمر کنترل کننده* نیز یاد می‌شود.

۳- مروری بر یادگیری تقویتی

یادگیری تاثیری است که یادگیرنده از محیط اطرافش می‌گیرد. هدف از یادگیری، کسب نوعی از دانش توسط عامل یادگیرنده است، به نحوی که عامل مذکور بتواند در شرایط مختلف، بهترین تصمیم ممکن را برای به دست آوردن نتایج مناسب، اتخاذ کند [1,2,6,11]. این کار، در واقع یافتن نگاشتی از فضای حالت‌ها به فضای اعمال ممکن در هر حالت است، که این نگاشت بهترین عمل را در هر حالت به دست می‌دهد. یادگیری وقتی اتفاق می‌افتد که عامل با توجه به تجارب جدیدی که به دست می‌آورد، به

[11] Gittins

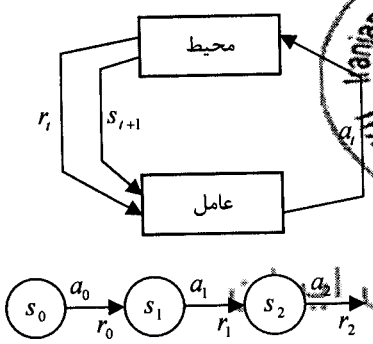
طور متفاوتی عمل کند و عاقبت این عملکرد متفاوت، منجر به عملکرد بهتر نیز بشود. این چنین شیوه‌ی یادگیری، کاملا وابسته به پسخورد[12]ی (فیدبک) است که به عامل یادگیرنده از طرف محیط اطراف یا سایر عوامل برگردانده می‌شود [1,6,11].

در نوع خاصی از یادگیری، موسوم به یادگیری تقویتی، هیچ‌گاه عامل یادگیرنده به شکل مستقیم بخوبی یا بدی اعمالی که انجام می‌دهد، آگاهی داده نمی‌شود، بلکه کمیت یا کمیت‌هایی را به عنوان پاداش یا جریمه دریافت می‌کند، که به صورت غیر مستقیم، معیاری از بخوبی یا بدی اعمالش در هر حالت هستند [1,11]. وظیفه‌ی عامل یادگیرنده، یادگیری عملکرد بهینه، صرفا با تکیه بر اطلاعات نهفته در پاداش‌ها یا جرایمی است که دریافت می‌کند. به واقع فرایند یادگیری برای عامل مذکور، مترادف با بیشینه[13] کردن میزان پاداش و یا کمینه[14] کردن میزان جریمه است. در اکثر مواقع می‌توان تصمیم‌گیری در مسائل پیچیده، با تکیه بر اطلاعات مختصر را، به صورت یک مسأله‌ی یادگیری تقویتی بیان نمود. در حیوانات و البته انسان، یادگیری تقویتی سهم قابل توجهی از میزان یادگیری را به خود اختصاص می‌دهد. هنگامی که ما دست خود را در مقابل حرارت قرار می‌دهیم و دستمان می‌سوزد، به سرعت یاد می‌گیریم که این کار را دوباره تکرار نکنیم. لذت‌ها و دردها، نمونه‌ای از پاداش‌ها و جریمه‌هایی هستند که الگوی رفتاری بسیاری از موجودات زنده را تشکیل می‌دهند [6,11].

شکل ۱، عامل یادگیرنده‌ای را نشان می‌دهد که با محیط اطرافش در تعامل است. نحوه‌ی حضور عامل در محیط، به صورت مجموعه‌ی حالات ممکن یا S تعریف می‌شود. در لحظه‌ی t از زمان گسسته، عامل در حالت s_t قرار دارد و اعمالی که عامل یادگیرنده می‌تواند انجام دهد به صورت مجموعه‌ی $A(s_t)$ تعریف می‌شود. با انجام عمل a_t، عامل به حالت بعدی، یعنی s_{t+1} می‌رود و امتیازی به صورت r_t را از محیط اطراف دریافت می‌کند. به این ترتیب دنباله‌ی حالات $\{s_i\}$، دنباله‌ی اعمال $\{a_i\}$ و دنباله‌ی امتیازهای آنی $\{r_i\}$ تعریف می‌شوند. وظیفه‌ی عامل یادگیرنده، پیدا کردن روشی برای انتخاب اعمال در حالات مختلف است، به نحوی که

$$z(T) = z(r_0, r_1, r_2, \ldots, r_T) \quad (1)$$

بیشینه شود. z خروجی مطلوب سیستم است و بسته به نوع مسأله، می‌تواند به اشکال مختلف تعریف شود. T طول عمر عامل یادگیرنده است و نشان دهنده‌ی اندازه‌ی افق زمانی مورد نظر برای مسأله می‌باشد.



شکل ۱: عامل یادگیرنده و نحوه‌ی تعاملش با محیط اطراف

روشی که عامل برای انتخاب اعمال در حالات مختلف به کار می‌برد، در اصطلاح *سیاست* یا *خط مشی* نامیده می‌شود. اگر A اجتماع مجموعه‌ی اعمال در تمام حالات باشد، سیاست را می‌توان به صورت $\pi: S \times A \to [0,1]$ تعریف نمود. هنگامی که عامل در حالت s قرار دارد، $\pi(s,a)$ احتمال انتخاب a به عنوان عمل بعدی، از طرف عامل یادگیرنده است. روش‌های مختلفی برای ایجاد و بهبود نگاشت π وجود دارند که مورد بحث نظریه‌ی یادگیری تقویتی می‌باشند [1,2].

یکی از روش‌های معمول که در این مقاله نیز مورد استفاده قرار گرفته است، استفاده از تعریف تابع ارزش [1,2,8] برای اعمال و حالات است. در این روش برای هر زوج مرتب از اعمال و حالات، یکی عدد حقیقی به عنوان ارزش در نظر گرفته می‌شود. این کار با تعریف نگاشتی به صورت $Q: S \times A \to \square$ امکان‌پذیر است. در هر لحظه از زمان، با توجه به مقادیر نگاشت Q، می‌توان از بین اعمال مختلف، یکی را انتخاب کرد. در واقع Q تخمینی از بهینگی هر عمل در هر حالت است و می‌توان از آن به عنوان معیاری در تصمیم‌گیری استفاده نمود [1,2,8,11].

یکی از روش‌های انتخاب، روشی است که، به روش انتخاب حریصانه یا Greedy مشهور است. این روش توصیه می‌کند، در هر حالت عملی انتخاب شود که مقدار تابع ارزش بیشینه

[12] Feedback
[13] Maximum
[14] Minimum

شود. در روش دیگری که به نام ε-Greedy مشهور است، عددی ثابت و کوچک مانند ε در بازه‌ی $[0,1]$ در نظر گرفته می‌شود و با احتمال $1-\varepsilon$، عامل دقیقا مثل روش Greedy عمل می‌کند. اما با احتمال ε، عمل بعدی از بین همه‌ی اعمال ممکن انتخاب می‌شود. به این ترتیب این امکان به عامل یادگیرنده داده می‌شود که در بین همه‌ی اعمال جستجو کند و اطلاعات بیشتری در مورد محیط کسب نماید. انتخاب عملی که بیشترین پاداش را دارد معمولا با نام بهره‌برداری[15] و امتحان سایر اعمال نیز با نام جستجو[16] معرفی می‌شوند. با تغییر پارامتر ε می‌توان موازنه‌ی مناسبی بین پدیده‌های جستجو و بهره‌برداری به دست آورد [1,2,8,11].

اگر از روش انتخاب ε-Greedy برای انتخاب اعمال استفاده شود، احتمال انتخاب عمل a، هنگامی که عامل در حالت s قرار دارد، به صورت زیر خواهد بود [1,2,8]:

$$\Pr(a|s)=\begin{cases}\dfrac{Q(s,a)}{\sum_{a'\in A}Q(s,a')} & ,\varepsilon_0\le\varepsilon\\ 1 & ,a=a^*(s),\varepsilon_0>\varepsilon\\ 0 & ,a\neq a^*(s),\varepsilon_0>\varepsilon\end{cases} \quad (۷)$$

که در آن $a^*(s)$، نشان دهنده‌ی عملی است که دارای بیشترین Q در حالت s است. به عبارت دیگر داریم:

$$a^*(s)=\arg\max_{a\in A}Q(s,a) \quad (۸)$$

و ε_0 عددی تصادفی، با توزیع یکنواخت و در بازه‌ی $[0,1]$ است و در ابتدای هر مرحله‌ی تصمیم‌گیری ایجاد می‌شود [1,2,8,11].

۴- نتایج آزمایش‌ها

در این بخش نتایج مربوط به آزمایش‌های انجام شده بر روی مسأله‌ی MAB که با استفاده از یادگیری تقویتی حل شده است، ارائه می‌شوند. تعداد دستگاه‌های مسأله‌ی مورد بررسی، برابر با ۱۰ دستگاه در نظر گرفته شده است. خروجی یا امتیاز هر دستگاه، به صورت کمیتی تصادفی و با توزیع نرمال در نظر گرفته شده است. میانگین این توزیع برای دستگاه i، به صورت μ_i و انحراف معیار (واریانس) آن یک در نظر گرفته شده است. توزیع‌های مربوط به دستگاه‌ها، کاملا مستقل از هم می‌باشند. همچنین میانگین امتیاز دستگاه‌ها، همگی نمونه‌هایی از یک کمیت تصادفی هستند که دارای توزیع نرمال با میانگین صفر و انحراف معیار یک است.

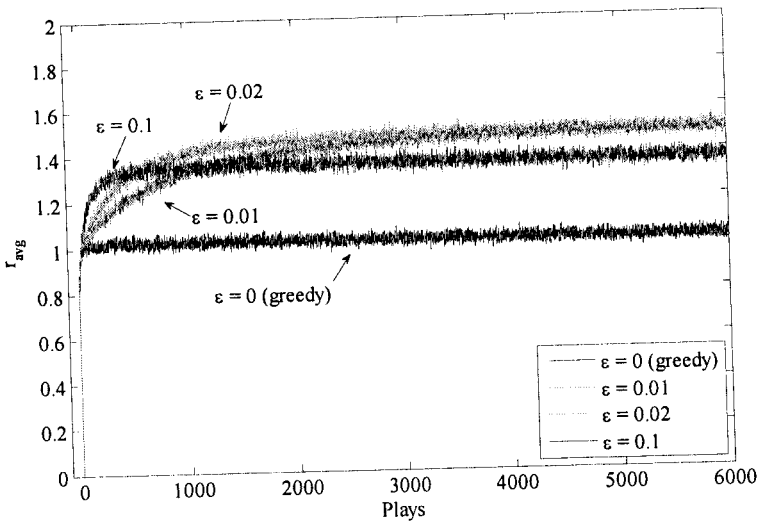
انتخاب هر دستگاه، نشان دهنده‌ی عملی است که عامل می‌تواند انجام دهد. لذا مجموعه‌ی اعمال قابل انجام برای عامل دارای ۱۰ عضو است. در اولین آزمایش، عامل یادگیرنده ۶۰۰۰ بار از بین ۱۰ عمل موجود، عملی را انتخاب می‌کند. کل این فرایند ۲۰۰۰ بار تکرار می‌شود و میانگین نتایج حاصل از آن‌ها محاسبه می‌شود. این آزمایش برای ۴ مقدار مختلف از ε که عبارتند از صفر، ۰/۰۱ ، ۰/۰۲ و ۰/۱ ، تکرار شده است. برای حل این مسأله، ازتعریف زیر برای Q استفاده شده است:

$$Q(a)=\frac{r_a(0)+r_a(1)+\cdots+r_a(t)}{N_a(t)} \quad (۹)$$

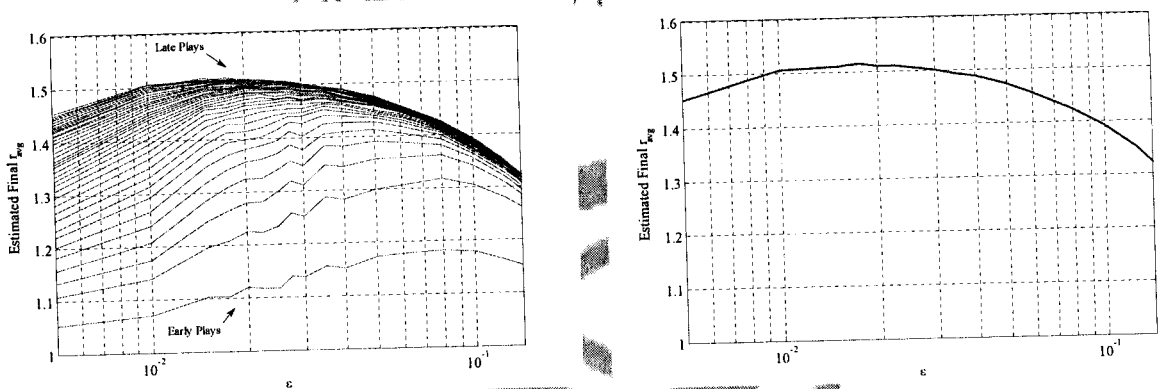
که در آن، t نشان‌دهنده‌ی زمان حال و N_a نشان دهنده‌ی تعداد دفعاتی است که عمل a انتخاب شده است. توجه کنید که، همان طور که در بخش ۲ گفته شد، اگر عمل a در لحظه‌ی τ انتخاب نشود، مقدار $r_a(\tau)$ صفر خواهد بود.

در شکل ۲، نمودار امتیاز متوسطی که در هر مرحله به دست آمده است بر حسب مرحله و به تفکیک مقدار ε نشان داده شده است. با توجه به شکل ۲، به ازای مقادیر مختلف ε ، مقادیر متفاوتی برای متوسط امتیازهای نهایی که توسط عامل کسب می‌شود، به دست می‌آیند. اگر به ازای مقادیر مختلف ε ، آزمایش تکرار شود و متوسط امتیاز نهایی کسب شده برای هر حالت محاسبه گردد، نمودار شکل ۳ به دست می‌آید.

[15] Exploitation
[16] Exploration



شکل ۲: نمودار امتیاز متوسط کسب شده در هر مرحله



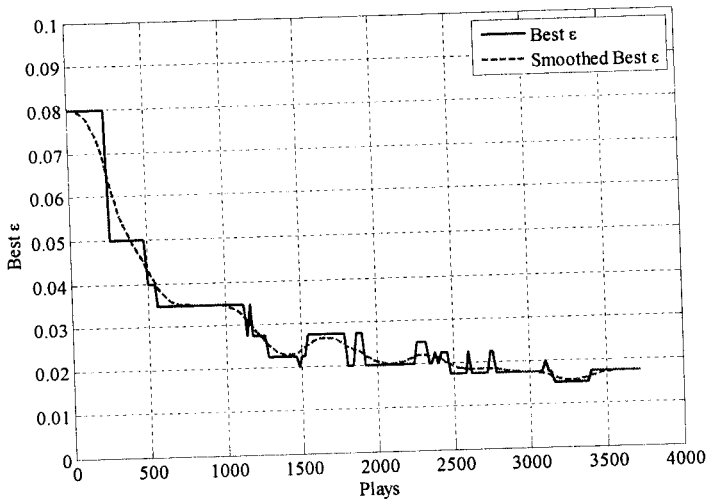
شکل ۳: متوسط امتیاز نهایی بر حسب ε به صورت نیم-لگاریتمی

شکل ۴: منحنی‌های مربوط به متوسط امتیازهای کسب شده در مراحل مختلف آزمایش، بر حسب ε به صورت نیم-لگاریتمی

همان‌طور که از شکل ۳ برمی‌آید، به ازای مقدار تقریبی ۰/۰۲۲ برای ε، بهترین نتیجه به دست می‌آید. مشاهده می‌شود که در صورتی که مقداری مناسب برای ε انتخاب شود، می‌توان متوسط امتیاز نهایی را تا حد ۱/۵۲ بالا برد. در شکل ۳ دیده می‌شود که با فاصله گرفتن از مقدار بهینه‌ی ε، مقدار متوسط امتیاز نهایی کاهش می‌یابد و نمودار تقریبا حول مقدار بهینه‌ی ε، متقارن است.

اگر به ازای مراحل مختلف، متوسط امتیازهای کسب شده، محاسبه و بر حسب ε ترسیم شوند، دسته‌ای از منحنی‌ها به دست می‌آیند که نمونه‌ای از این منحنی‌ها در شکل ۴ قابل مشاهده هستند. منحنی‌های پایین‌تر مربوط به مراحل اولیه‌ی آزمایش و منحنی‌های بالاتر مربوط به مراحل نهایی آزمایش می‌باشند. مشاهده می‌شود که تراکم منحنی‌ها در مراحل نهایی بیشتر است. در این شکل دیده می‌شود که برای هر مرحله، مقداری برای ε می‌توان پیدا کرد، به نحوی که امتیاز کسب شده برای آن مرحله را بیشینه کند. نمودار این مقادیر بهینه ε به ازای مراحل مختلف آزمایش به صورت شکل ۵ است. در این شکل، نموداری به صورت خط‌چین ترسیم شده است که منحنی به دست آمده از آزمایش‌ها را تقریب می‌زند و تغییرات نرمی دارد.

مقادیر بیشتر ε، نشان دهنده اهمیت جستجو برای عامل یادگیرنده هستند. در مقابل، هر چه مقدار ε کمتر شود، عامل یادگیرنده به تجربه کردن اعمال جدید، کمتر رغبت می‌کند. مقدار بهینه‌ای که برای ε به دست می‌آید، موازنه‌ای بین دو پدیده‌ی جستجو و استخراج را به وجود می‌آورد که منجر به نتیجه‌گیری بهتر از آزمایش‌ها می‌شود.



شکل ۵: نمودار مقادیر بهینه ε به ازای مراحل مختلف آزمایش

به وضوح مشاهده می‌شود که با افزایش تعداد مراحل، مقدار بهینه‌ای که برای ε به دست می‌آید، کمتر می‌شود. به عبارت دیگر، هر چه عامل یادگیرنده و یا کنترل کننده عمر بیشتری داشته باشد، می‌بایست عملکرد حریصانه‌تری از خود نشان دهد. مقادیر بیشتر ε نشان دهنده‌ی ریسک‌پذیری بیشتری هستند و نتایج به دست آمده ریسک بیشتری را به عامل‌های کوتاه مدت توصیه می‌کنند. اگر قرار است عاملی به مدت طولانی، مسأله‌ی MAB را تحت کنترل داشته باشد، می‌بایست ریسک کم‌تری انجام دهد و عملکرد محتاطانه‌تری داشته باشد. به این ترتیب کنترل کننده می‌تواند امتیاز بیشتری را کسب کند.

۵- نتیجه‌گیری

در این مقاله، با آزمایش‌های متعددی که با پارامترهای مختلف برای حل مسأله‌ی MAB ترتیب داده شد، شرایط لازم برای ایجاد موازنه‌ی بهینه بین پدیده‌های جستجو و استخراج در یادگیری تقویتی مورد بررسی قرار گرفتند. مشاهده شد که می‌توان پارامتر ε را به نحوی تعیین کرد که بیشترین امتیاز ممکن را با هر تعداد معین از اعمال به دست آورد. برای طول عمر کمترِ عامل، مقدار بهینه‌ی ε بیشتر است و با افزایش طول عمر عامل، مقدار بهینه‌ی ε رفته رفته کم‌تر می‌شود. به عنوان مطالعات پیشنهادی، می‌توان سایر روش‌های انتخاب و سایر مسائل و روش‌های مطرح در یادگیری تقویتی را مورد بررسی قرار داد.

مراجع

[1] Richard S. Sutton and Andrew G. Barto, *Reinforcement Learing: An Introduction*. MIT Press, Cambridge, MA, 2002.
[2] Leslie Pack Kaelbling, Michael L. Littman and Andrew W. Moore, "Reinforcement Learning: A Survey," in *Journal of Artificial Intelligence Research*, Vol. 4, pp. 237-285, 1996.
[3] Berry, D. and B. Fristedt, *Bandit Problems: Sequential Allocation of Experiments*, Chapman and Hall, London, 1985.
[4] Dirk Bergemann and Juuso Välimäki, "Bandit Problems," in *Discussion Papers of Helsinki Center of Economic Research*, Vol. 93, 2006.
[5] Braz Camargo, "Good news and bad news in two-armed bandits," in *Journal of Economic Theory*, Vol. 135, pp. 558-566, 2007.
[6] Tom M. Mitchell, *Machine Learning*. McGraw Hill, 1997.
[7] A. Mahajan and D. Teneketzis, "Chapter 6: Multi-Armed Bandit Problems," in *Foundations and Applications of Sensor Management*, Springer, 2008.
[8] D.E. Koulouriotis and A. Xanthopoulosa, "Reinforcement learning and evolutionary algorithms for non-stationary multi-armed bandit problems," in *Applied Mathematics and Computation*, Vol. 196, Issue 2, pp. 913-922, 2008.
[9] Peter Auer et al, "The non-stochastic multi-armed bandit problem," in *SIAM Journal on Computing*, Vol. 32, No. 1, pp. 48-77, 2002.
[10] P. Auer, N. Cesa-Bianchi and Paul Fischer, "Finite-time Analysis of the Multiarmed Bandit Problem," in *Machine Learning*, Vol. 47, pp. 235–256, 2002.
[11] S. I. Reynolds, "Reinforcement Learning with Exploration," *Ph.D. Thesis*, School of Computer Science, The University of Birmingham, UK, 2002.