پانزدهمین سمینار احتمال
و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان

15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan

انجمن آمار ایران

# نقش فرآیندهای تصادفی در هوش مصنوعی و علم داده: یک مطالعه موردی

**اعظم راستین** [۱]

[۱] دکترای آمار، دانشگاه شهید چمران اهواز

**چکیده:**

## چکیده

در این مقاله، نقش بنیادین فرآیندهای تصادفی در هوش مصنوعی و علم داده مورد بررسی قرار می‌گیرد، با تمرکز ویژه بر یادگیری تقویتی در محیط Gridworld به عنوان یک مطالعه‌ی موردی. فرآیندهای تصادفی، به‌ویژه مدل‌سازی‌های مارکوفی، چارچوبی ریاضی و مفهومی برای تحلیل و پیش‌بینی رفتار عامل‌های هوشمند در محیط‌های نامعین فراهم می‌آورند. در این راستا، الگوریتم Q-Learning در سه سناریوی مختلف (محیط قطعی، محیط با نویز تصادفی، و محیط با پاداش‌های تصادفی) پیاده‌سازی و تحلیل شده است. نتایج عددی، نمودارهای همگرایی، و مسیرهای یادگیری عامل‌ها نشان می‌دهد که چگونه فرآیندهای تصادفی درک دقیق‌تری از تعامل عامل با محیط فراهم کرده و بر کیفیت یادگیری تأثیر می‌گذارند. این مقاله تأکید دارد که ادغام تئوری احتمال و فرآیندهای تصادفی در ساختارهای هوش مصنوعی، مسیر توسعه الگوریتم‌های پایدارتر و تطبیق‌پذیرتر را هموار می‌سازد.

**واژه‌های کلیدی: فرآیند تصادفی، یادگیری تقویتی، MDP، Gridworld، یادگیری سیاست، ارزش‌گذاری، مدل‌سازی محیط**

**کد موضوع‌بندی ریاضی (۲۰۲۰):** 60J20، 68T05، 68T20.

## ۱ مقدمه

**فرآیندهای تصادفی یکی از ابزارهای بنیادی در مدل‌سازی پدیده‌هایی هستند که در آن‌ها عدم قطعیت و تصادف نقش دارند. در دهه‌های اخیر، اهمیت این فرآیندها در حوزه‌هایی مانند هوش مصنوعی و علم داده به طور چشمگیری افزایش یافته است. الگوریتم‌های یادگیری ماشین، تحلیل سری‌های زمانی، یادگیری تقویتی و بسیاری دیگر از شاخه‌های هوش مصنوعی برای مدل‌سازی رفتارهای پیچیده و پیش‌بینی‌های دقیق، به نظریه‌های تصادفی متکی هستند. فرآیندهای تصادفی امکان مدل‌سازی**

[۱] سخنران، rastinstat@gmail.com

پویایی‌هایی را فراهم می‌کنند که در آن‌ها آینده سیستم به صورت احتمالی و بر اساس گذشته یا وضعیت فعلی تعیین می‌شود. به ویژه، زنجیره‌های مارکوف، فرآیندهای تصمیم‌گیری مارکوف، و فرآیندهای گاوسی، ابزارهایی کلیدی در توسعه سیستم‌های هوشمند و تحلیل داده‌های پیچیده هستند. در سال‌های اخیر، هوش مصنوعی و علم داده به‌شدت به مفاهیم احتمالاتی و فرآیندهای تصادفی متکی شده‌اند تا بتوانند رفتار پیچیده و غیرقطعی سیستم‌های واقعی را مدل‌سازی کنند. یکی از حوزه‌هایی که به‌طور مستقیم از این ابزارها بهره‌مند است، یادگیری تقویتی است. یادگیری تقویتی یکی از شاخه‌های مهم یادگیری ماشینی است که به‌طور خاص برای تصمیم‌گیری در محیط‌های پویا و تعاملی طراحی شده‌است. در این روش، عامل از طریق تعامل با محیط و بر اساس بازخورد پاداش، سیاستی بهینه برای بیشینه‌سازی پاداش بلندمدت فرا می‌گیرد (سوتون و باروت ، ۲۰۱۸).

فرآیندهای تصادفی نقشی کلیدی در مدل‌سازی عدم قطعیت در یادگیری تقویتی دارند. مطالعات پیشین نشان داده‌اند که مدل‌سازی دقیق نویزها و تصادفی بودن پاداش یا انتقال حالت می‌تواند به طراحی الگوریتم‌های پایدارتر و واقع‌گرایانه‌تر منجر شود (مراجعه شود به ترون و همکاران (۱۹۹۵) و کرنز و سینگز (۲۰۰۲) ).

در این مقاله، ما محیط $Gridworld$ را به‌عنوان بستری ساده ولی گویا برای بررسی تأثیر فرآیندهای تصادفی انتخاب کرده‌ایم. با استفاده از الگوریتم $Q-learning$ و در سه سناریوی مجزا ـ محیط قطعی، محیط با نویز در انتقال، و محیط با پاداش تصادفی ـ روند یادگیری و عملکرد عامل مورد تحلیل قرار گرفته‌اند.

## ۲ مفاهیم پایه

فرآیند تصادفی به توالی از متغیرهای تصادفی، $\{X_t\}_{t\in T}$، گفته می‌شود که به صورت تابعی از زمان تعریف شده‌اند که در آن $T$ می‌تواند گسسته (مثلاً زمان‌های طبیعی) یا پیوسته (مثلاً بازه‌ای از اعداد حقیقی) باشد. یک مورد خاص از فرآیندهای تصادفی، زنجیره مارکوف است که دارای ویژگی زیر می‌باشد:

$$P(X_{t+1} = x|X_t = x_t, X_{t-1} = x_{t-1}, \ldots) = P(X_{t+1} = x|X_t = x_t)$$

به این خاصیت، ویژگی مارکوفی یا حافظه کوتاه گفته می‌شود. زنجیره‌های مارکوف به طور گسترده‌ای در الگوریتم‌های یادگیری، مدل‌سازی تصمیم‌گیری، و تحلیل سری‌های زمانی به کار می‌روند. فرآیندهای تصادفی مبنای ریاضی مدل‌سازی رفتار سیستم‌هایی هستند که با گذر زمان و تحت تأثیر عدم قطعیت تغییر می‌کنند. در یادگیری تقویتی، به‌ویژه از فرآیندهای تصمیم‌گیری مارکوفی ( $MDP$ ) استفاده می‌شود (پورتمن ، ۲۰۰۵).

**تعریف ۱.۲.** فرآیند تصمیم‌گیری مارکوف ($MDP$): در واقع یک $MDP$، $M = (S, A, P, R, \gamma)$ شامل مجموعه‌ای از حالات ممکن $S$، اعمال ممکن $A$، تابع احتمال انتقال $P(s'|s, a)$ ( احتمال انتقال از حالت $s$ به حالت $s'$ با انتخاب عمل $a$) و تابع پاداش $R(s, a)$ ( پاداش دریافتی برای اجرای عمل $a$ در حالت $s$) و ضریب تنزیل آینده $\gamma \in [0, 1)$

و طبق خاصیت مارکوفی، $P(s_{t+1}|s_t, a_t, s_{t-1}, ..., s.) = P(s_{t+1}|s_t, a_t)$. در واقع، این خاصیت تضمین می‌کند که وضعیت آینده فقط به وضعیت و عمل فعلی بستگی دارد، نه به مسیر طی‌شده قبلی.

**تعریف ۲.۲.** سیاست ($Policy$): تابعی که مشخص می‌کند عامل در هر حالت چه عملی انجام دهد. سیاست بهینه $\pi^*$ باعث بیشینه شدن امید ریاضی پاداش تجمعی می‌شود.

**تعریف ۳.۲.** تابع $Q$: بیانگر ارزش حالت–عمل تحت یک سیاست خاص است:

$$Q(s,a) = \mathbb{E}\left[\sum_{t=\cdot}^{\infty} \gamma^t R(s_t, a_t) \mid s_\cdot = s, a_\cdot = a\right]$$

**هدف اصلی در $MDP$، یافتن سیاست بهینه $\pi^*$ است به‌طوری که تابع ارزش $Q$ بیشینه شود.**

**یادگیری تقویتی نوعی یادگیری ماشینی است که در آن عامل با تعامل با محیط، سیاست بهینه را از طریق آزمون و خطا می آموزد. در این تحقیق از الگوریتم $Q-Learning$ استفاده شده است. $Q-Learning$ یکی از الگوریتم‌های پایه در یادگیری تقویتی است که بدون مدل محیط کار می‌کند:**

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left[r + \gamma \max_{a'} Q(s',a') - Q(s,a)\right]$$

**که در آن $\alpha, r, s'$ و $\gamma$ بیانگر نرخ یادگیری، پاداش دریافتی در انتقال، حالت بعدی پس از انجام $a$ در $s$ و ضریب تنزیل است. با گذر زمان، مقدارهای $Q$ به سمت مقدارهای بهینه همگرا می‌شوند (واتکینز و دایان ، ۱۹۹۲).**

## ۳ مطالعه موردی: یادگیری تقویتی در محیط $Gridworld$

**محیط $Gridworld$ یک شبکه دوبعدی است که هر خانه از آن یک حالت را نشان می‌دهد. که در آن یک عامل ($Agent$) می‌تواند به چهار جهت اصلی (بالا، پایین، چپ، راست) حرکت کند. بعضی خانه‌ها می‌توانند موانع یا چاله‌هایی باشند که مانع حرکت عامل می‌شوند. هدف عامل، رسیدن به خانه هدف است، در حالی که از موانع اجتناب کند و پاداش حداکثری دریافت کند. در این مطالعه سه نوع محیط تعریف شده است:**

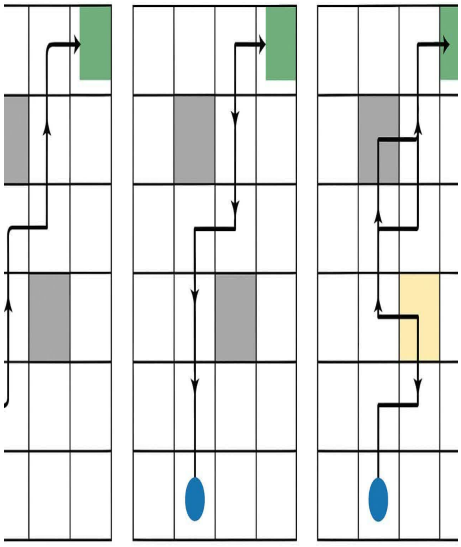
۱. **محیط قطعی: انتقال دقیق و بدون نویز به جهت دلخواه**

۲. **محیط نویزدار: احتمال ٪ ۸۰ برای حرکت مورد نظر و ۲۰ ٪ برای انحراف به جهت دیگر**

۳. **محیط با پاداش‌های تصادفی: مقدار پاداش در برخی خانه‌ها به صورت تصادفی تغییر می‌کند**

**در شکل ۱ این سه محیط متفاوت نمایش داده شده است. این تصویر سه محیط مختلف را در قالب شبکه ۵×۵ نشان می‌دهد که در هر کدام شرایط متفاوتی حاکم است: در محیط قطعی، حرکت عامل با دقت کامل در جهت اعمال‌شده صورت می‌گیرد و پاداش‌ها ثابت هستند. در محیط نویزی، عامل با احتمال ۰.۸ در جهت اعمال‌شده حرکت می‌کند و با احتمال ۰.۲ به جهت‌های دیگر می‌رود. و در محیط با پاداش تصادفی، مقصد حرکت مشخص است، اما مقدار پاداش از یک توزیع تصادفی (مثلاً یکنواخت یا نرمال) گرفته می‌شود.**

**در این مطالعه، محیط $Gridworld$ شامل یک شبکه ۵ × ۵ است که عامل باید از خانه شروع به خانه هدف برسد. عامل در هر مرحله می‌تواند یکی از چهار جهت را انتخاب کند. هدف یافتن سیاست بهینه برای بیشینه کردن پاداش است. پس از اجرای الگوریتم $Q-Learning$، سیاست بهینه استخراج شده نشان‌دهنده بهترین جهت حرکت از هر خانه به سمت هدف است. به عنوان مثال، اگر در خانه‌ای فلش به سمت راست باشد ($\rightarrow$)، به معنی آن است که حرکت به سمت راست بیشترین ارزش را دارد. این سیاست به عامل کمک می‌کند تا با پیمودن مسیر بهینه، در کوتاه‌ترین زمان و بیشترین پاداش به هدف برسد.**



شکل ۱: نمایش سه محیط $Gridworld$ با ویژگی‌های مختلف: محیط قطعی، محیط با نویز، محیط با پاداش تصادفی.

پس از آموزش، سیاست بهینه به صورت جهت حرکت از هر خانه به سوی هدف نمایش داده می‌شود. این نتایج نشان می‌دهند که فرآیندهای تصادفی چگونه می‌توانند مسیرهای بهینه را در محیط‌های نامعین بیاموزند.

برای هر یک از سه سناریو، الگوریتم $Q-learning$

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\left[r_t + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a_t)\right]$$

در صد اپیزود و با پارامترهای $\alpha = 0.1$ و $\gamma = 0.95$ اجرا شده و نمودار همگرایی و مسیر یادگیری استخراج شده است.

در شکل ۲ همگرایی مقدار $Q$ در سه سناریو نمایش داده شده است. همان‌طور که مشاهده می‌شود، مقدار خطای $Q$ در محیط قطعی سریع‌تر کاهش می‌یابد. در محیط‌های نویزی و دارای پاداش تصادفی، نرخ همگرایی کندتر و همراه با نوسان است.

در شکل ۳ پاداش تجمعی در طول اپیزودها برای سه محیط نمایش داده شده است. در محیط قطعی، پاداش تجمعی با شیب مثبت مشخص رشد می‌کند، ولی در محیط‌های تصادفی، پاداش‌ها بی‌ثبات‌ترند.

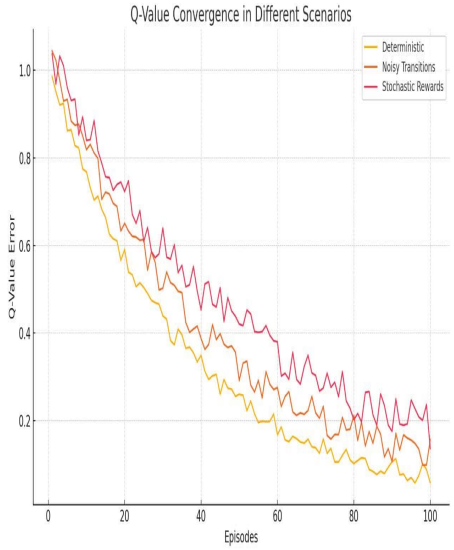
در شکل ۴ مسیر عامل در شبکه نهایی نمایش داده شده است. عامل در طی فرآیند یادگیری، مسیری بهینه برای رسیدن به هدف پیدا کرده است.

تجزیه‌وتحلیل این نمودارها نشان می‌دهد که تصادفی بودن در پاداش یا انتقال، موجب کاهش سرعت یادگیری و همچنین افزایش نوسان در عملکرد عامل می‌شود. بنابراین، استفاده از تکنیک‌های مقاوم در برابر نویز و سیاست‌های اکتشافی مناسب، اهمیت زیادی دارند.

تحلیل نتایج بخش اصلی این مقاله است که نقش فرآیندهای تصادفی را آشکار می‌کند.

- محیط قطعی: در این حالت، عامل به سرعت سیاست بهینه را می‌یابد. همگرایی الگوریتم بسیار سریع است. مسیر عامل کاملاً مستقیم و بهینه است. پاداش تجمعی در اپیزودهای پایانی به حداکثر مقدار ممکن می‌رسد.

- محیط با نویز حرکتی: عامل مجبور است سیاست محافظه‌کارانه‌تری اتخاذ کند، زیرا هر حرکت ممکن است منجر به




شکل ۲: نمودار همگرایی مقدار $Q$ برای سه حالت

حالت ناخواسته شود. این موضوع باعث نوسانات بیشتر در مقادیر $Q$ و تأخیر در همگرایی می‌شود. با این حال، الگوریتم همچنان قادر به یادگیری سیاست نسبتاً خوب است.

- محیط با پاداش‌های تصادفی: در این حالت عامل باید با تغییرات تصادفی پاداش‌ها سازگار شود. مقدار $Q$ دچار نوسانات دائمی می‌شود. مسیر عامل نیز گاه به گاه تغییر می‌کند، بسته به اینکه کدام مسیر اخیراً پاداش بیشتری داشته است. همگرایی کندتر و نوسانی‌تر از سایر سناریوها است.

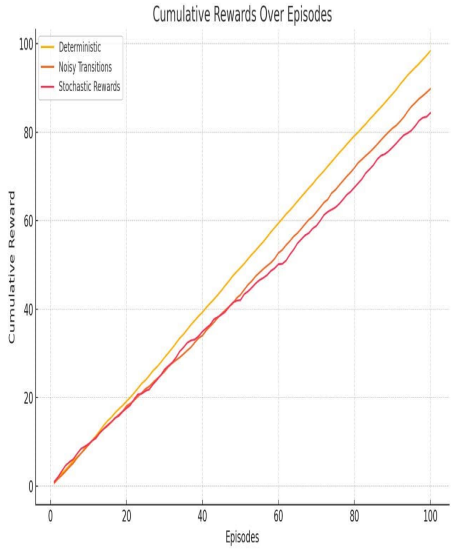
## بحث و نتیجه‌گیری

مطالعه حاضر نشان داد که فرآیندهای تصادفی نقش کلیدی در تصمیم‌گیری هوشمند دارند. حتی در محیط‌های ساده‌ای مانند $Gridworld$، وجود نویز و تصادفی‌بودن پاداش می‌تواند یادگیری عامل را با چالش روبرو کند. $Q-learning$با وجود این چالش‌ها، قادر است در بلندمدت سیاست‌های کارآمدی یاد بگیرد. مطالعه موردی ارائه‌شده به‌خوبی نشان می‌دهد که چگونه مفاهیم ریاضی می‌توانند به راه‌حل‌های عملی در مسائل هوش مصنوعی منجر شوند.

برای کارهای آینده، استفاده از الگوریتم‌های پیشرفته‌تر مانند $DQN$ و بررسی اثر حافظه و شبکه‌های عصبی پیشنهاد می‌شود.

## مراجع

**Bertsekas, D. P. (2017).** ***Dynamic Programming and Optimal Control*****. Athena Scientific.**




شکل ۳: نمودار پاداش تجمعی در طول اپیزودها برای سه حالت

**Durrett, R. (2019). Probability: Theory and Examples. Cambridge University Press.**

**Kearns, M., and Singh, S. (2002). Near-optimal reinforcement learning in polynomial time. Machine Learning, 49(2), 209-232.**

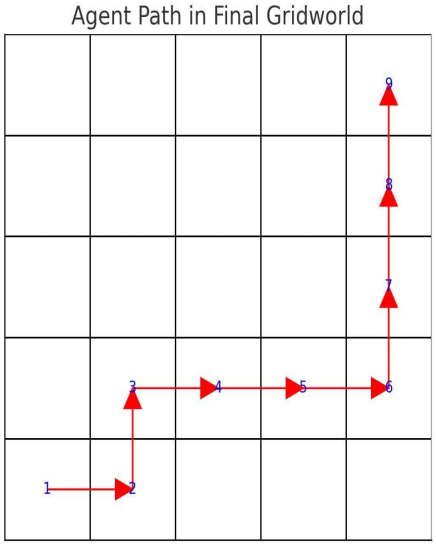
**Mnih, V., Kavukcuoglu, K. and Silver, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529–533.**

**Puterman, M. L. (2005). Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley.**

**Sutton, R. S., Barto, A. G. (2018). Reinforcement Learning: An Introduction (2nd ed.). MIT Press.**

**Thrun, S, Burgard, W., and Fox, D. (1995). Probabilistic robotics. MIT press.**

**Watkins, C. J. C. H., and Dayan, P. (1992). ”Q-learning”. Machine Learning, 8(3-4), 279–292.**




شکل ۴: مسیر طی‌شده عامل در شبکه $Gridworld$ نهایی

# The Role of Stochastic Processes in Artificial Intelligence and Data Science: A Case Study

**Azam Rastin**[1]

[1]**Shahid Chamran university of Ahvaz**

---

**Abstract:**

**Stochastic processes are a foundational component in modeling uncertainty and dynamics within Artificial Intelligence (AI) and Data Science. This paper presents a theoretical and empirical investigation into the role of stochastic processes in Reinforcement Learning (RL), focusing on a case study using the classical Gridworld environment. We formalize the decision-making process as a Markov Decision Process (MDP), and examine how stochastic transition dynamics influence learning policies and value estimation. Experimental results demonstrate how varying the stochasticity of the environment impacts policy convergence, optimality, and stability. These findings provide critical insights**

**into the importance of accurate modeling of environment dynamics for robust AI systems.**

**Keywords: Most important keywords of article (3-5 items) should appear here.**

**Mathematics Subject Classification (2020):** 60J20، 68T05، 68T20**.**