دانشکده مهندسی کامپیوتر

نظریهٔ ریاضی بازی‌ها

پروژه پایانی

مدرس: وصال حکمی موعد تحویل: تا پایان ۲۷ تیر ۱۴۰۰

مقدمه

بر اساس یک دسته‌بندی متداول در حوزه یادگیری تعادل در بازی‌ها، قوانین یادگیری را به دو دسته «بهم‌وابسته[1]» و «غیر بهم‌وابسته[2]» تقسیم‌بندی می‌نمایند [۱]، [۲]، [۳]. قوانین یادگیری غیر بهم‌وابسته بر اساس فرض «عقلانیت محدود[3]» [۴] برای بازیگران طراحی می‌شوند؛ یعنی، این قوانین دارای پیچیدگی محاسباتی پایین هستند و عمدتاً بر مبنای «پاسخ بهتر[4]» و نه لزوماً «بهترین پاسخ» کار می‌کنند. ضمناً، اطلاعات بسیار محدودی برای بِروزرسانی استراتژی‌ها لازم دارند؛ به طور مشخص، فرض می‌شود که بازیگران فقط توابع سودمندی خود را می‌دانند و حداکثر تنها قادر به مشاهدهٔ اَعمال اتخاذ شده توسط سایرین هستند، ولی از توابع سودمندی آنها اطلاعی ندارند. همچنین، گفته می‌شود که یک قانون یادگیری، «کاملاً غیر بهم‌وابسته[5]» است چنانچه علاوه بر «غیر بهم‌وابسته» بودن، بازیگران برای رسیدن به تعادل حتی نیازی به مشاهدهٔ عمل اتخاذ شدهٔ سایرین هم نداشته باشند. در این نوع یادگیری که اصطلاحاً به «بازیِ ناشناخته[6]» معروف است، بازیگران ممکن است حتی تابع سودمندی خود را هم ندانند و همهٔ آنچه را که از تاریخچهٔ بازی برای بِروزرسانی استراتژی‌های خود استفاده می‌کنند، یک تاریخچهٔ خصوصی شامل اَعمال اتخاذ شده توسط خودشان و مقادیر عددی سودمندی‌های دریافت شده در هر تکرار بازی است. در سناریوهایی که بازیگران دچار محدودیت منابع محاسباتی هستند (مثلاً گره‌های شبکه بی‌سیم) و نیز به دلیل سربار بالای محاسباتی و سیگنالینگِ مورد نیاز قوانینِ یادگیری بهم‌وابسته، استفاده از الگوریتم‌های یادگیری غیر بهم‌وابسته به مراتب واقع-گرایانه‌تر است.

با این حال، نتایج نظری وجود دارد که نشان می‌دهند همگرایی به تعادل NE بر اساس قوانین یادگیری غیربهم‌وابسته در همه انواع بازی‌های چند-نفره امکان‌پذیر نیست [۵]. امّا، قوانین غیربهم‌وابسته‌ای وجود دارند که در "هر" بازی، قادر به شکل-دهیِ رفتار حدّی بازیگران بر اساس مفهوم CE هستند [۶]، [۷]. در بسیاری از این قوانین، از نوعی مفهوم «پشیمانی[7]» برای

[1] Coupled
[2] Uncoupled
[3] Bounded rationality
[4] Better reply
[5] Completely uncoupled
[6] Unknown game
[7] Regret

یادگیری و تطبیق استراتژی هر بازیگر استفاده می‌شود. در ادامه، به معرفی مفهوم پشیمانی و سپس، یادگیری مبتنی بر پشیمانی می‌پردازیم.

مفهوم پشیمانی

به بیان غیر رسمی، «پشیمانی» یک بازیگر در هر تکرار بازی، درجهٔ نارضایتی کلی وی از بابت انتخاب‌هایش، از تکرار نخست بازی تابحال، نسبت به هر عمل جایگزین دیگر را نشان می‌دهد. ایدهٔ اساسی در قوانین مبتنی بر پشیمانی، تغییر گرایش بازیگر به سوی اَعمال جایگزینِ با سودمندیِ بالاتر، بر اساس مَحَکی است که از این شاخصِ پشیمانی استنباط می‌کند. مفید بودن مفهوم پشیمانی بیش از هر چیز، ناشی از سادگی و طبیعی بودن نحوهٔ بازی بر اساس آن است چراکه یک بازیگر برای بازنگری در استراتژی‌اش می‌تواند به طور کامل از توابع سودمندی سایر بازیگران نامطلع باشد. در ادامه، به طور اِجمالی به معرفی دقیق‌تر این قانون یادگیری بر اساس ویراست ارائه شده در [۶] می‌پردازیم.

ایده الگوریتم «تطبیق پشیمانی[۸]»

در الگوریتم تطبیق پشیمانی [۶]، بازیگران، اَعمالی را که بابت بازی نکردنِ بقدر کافیِ آنها در گذشته "پشیمان" هستند، تقویت می‌کنند. در واقع، هر بازیگری مثل n یک ماتریس پشیمانی $R_k^n\langle i,j\rangle$ دارد که در آن برای هر زوج عمل $i,j \in \mathcal{A}^n$، مابه‌التفاوت متوسط سودمندی‌ای نگهداری می‌شود که می‌توانست حاصل گردد مشروط بر آنکه هر بار که بازیگر n عمل i را در گذشته واقعاً اتخاذ کرده است بجایش عمل j را انتخاب می‌کرد؛ به عبارت دیگر، در تکرار k-اُم بازی،

$$R_k^n\langle i,j\rangle = \frac{1}{k}\sum_{\tau=1}^{k}[u^n(j,\boldsymbol{a}_\tau^{-n}) - u^n(i,\boldsymbol{a}_\tau^{-n})].\,\mathbb{I}_{\{a_\tau^n=i\}}.$$

برای تکرار $(k+1)$-اُم، اگر عمل برگزیدهٔ جاری توسط بازیگر n، عملِ $a_k^n = i$ باشد، بازیگرِ n با احتمال $T_{k+1}^n(a_{k+1}^n = j|a_k^n = i)$ که متناسب با متوسط پشیمانی $R_k^n\langle i,j\rangle$ تعیین می‌شود، از عمل جاری i به عمل دیگری مثل j گذار می‌کند و با احتمال $1 - \sum_{j\in\mathcal{A}^n, j\neq i} T_{k+1}^n(a_{k+1}^n = j|a_k^n = i)$ هم مجدداً همان عمل جاری i را اتخاذ می‌نماید. طبیعی است که هدف هر بازیگر، انتخاب دنباله‌ای از اَعمال است که فارغ از اینکه استراتژی‌های سایرین در انتخاب اَعمالشان چه باشد، منجر به پشیمانی صفر (یا منفی) برایش گردد.

همگرایی رفتار بازیگران به تعادل CE

کمّیت $\boldsymbol{z}_k(\boldsymbol{a})$ را به این صورت تعریف می‌کنیم: "تعداد دفعاتی که پروفایل عمل $\boldsymbol{a}$ در طی k تکرار اول بازی واقعاً پیاده-سازی شده باشد تقسیم بر k." در واقع، $\boldsymbol{z}_k(.)$ نمایانگر «توزیع تجربی» بازی است و از جنس یک توزیع احتمال روی فضای پروفایل اَعمال $\boldsymbol{\mathcal{A}}$ می‌باشد. به سادگی می‌توان نشان داد که اگر کلیهٔ درایه‌های ماتریس پشیمانیِ بازیگران به مقادیر منفی یا صفر همگرا شوند، توزیع $\boldsymbol{z}_k(.)$ هم به مجموعهٔ تعادل‌های همبستهٔ بازی همگرا خواهد شد. در [۶] نشان داده شده است که در

[8] Regret matching

واقع، اجرای قانون «تطبیق پیشمانی» توسط بازیگران، موجب همگرایی ماتریس پشیمانی آنها به همسایگی صفر شده و در نتیجه، رفتار حدّی آنها نیز به *polytope* تعادل‌های همبسته بازی همگرا می‌شود.

به این ترتیب، بر اساس قانون تطبیق پشیمانی، رفتار محلی و نه کاملاً عقلانی بازیگران می‌تواند به ظهور یک رفتار سراسری عقلانی مثل CE منجر شود. در الگوریتم تطبیق پشیمانی، به هر عملی که صرفاً "بهتر" محسوب شود، احتمال مثبت نسبت داده می‌شود. بر این اساس، رفتار یک بازیگر، فاصله زیادی دارد با رفتار یک تصمیم‌ساز عقلانی- که به صورت بهینه بر اساس باورهایی که کمابیش دقیق از محیطش شکل داده- عمل می‌کند. در عوض، رفتار یک بازیگر بیشتر به یک فرد با گرایش رفتاریِ "واکنشی" شباهت دارد که صرفاً به تقویت تصمیمات با پیامدهای خوشایند می‌پردازد. خلاصه اینکه، با استفاده از یادگیری مبتنی بر پشیمانی، بازیگران می‌توانند استراتژی‌های خود را به صورت توزیع‌شده با هم هماهنگ‌سازی نمایند به نحوی که توزیع رفتار جمعی آنها با مجموعهٔ تعادل‌های همبسته بازی همخوانی پیدا کند.

شِبه کُد الگوریتم «تطبیق پشیمانی»

Let $\mathbf{1}_N = [1, \cdots, 1]^T$ denote an $N \times 1$ column vector of ones, $U^n_{\max}$ and $U^n_{\min}$ represent the upper and lower bounds on the payoff function $U^n(\cdot)$ for agent n, respectively, and $I_{\{.\}}$ denote the indicator function.

Initialization: Set $\mu^n > A^n \left| U^n_{\max} - U^n_{\min} \right|$, $\mathrm{p}^n_0 = \frac{1}{A^n} \cdot \mathbf{1}_{A^n}$, and $R^n_0 = \mathbf{0}$, Set $k = \mathbf{0}$.

Step 1: Action Selection. Choose $\mathrm{a}^n_k \sim \mathrm{p}^n_k$

Step 2: Information Exchange. Share decisions a^n_k with others.

Step 3: Regret Update.

$$R^n_{k+1} = R^n_k + \frac{1}{k+1}\left[B^n\left(\mathbf{a}_k\right) - R^n_k\right]$$

where $B^n\left(\mathbf{a}_k\right) = \left[b^n_{ij}\left(\mathbf{a}_k\right)\right]$ is an $A^n \times A^n$ matrix with elements

$$b^n_{ij}\left(\mathbf{a}_k\right) = \left[U^n\left(j, \mathbf{a}^{-n}_k\right) - U^n\left(i, \mathbf{a}^{-n}_k\right)\right] \cdot I\left\{\mathrm{a}^n_k = i\right\}.$$

Step 4: Update Action Selection Probability

$$p^n_{k+1}(i) = \begin{cases} \frac{1}{\mu^n}\left|R^n_{k+1}\left\langle \mathrm{a}^n_k, i \right\rangle\right|^+, & i \neq \mathrm{a}^n_k \\ 1 - \sum_{j \neq i} p^n_{k+1}(j), & i = \mathrm{a}^n_k \end{cases}$$

where $|x|^+ = \max\{0, x\}$.

Step 5: Recursion. Set $k \leftarrow k+1$, and go Step 1.

خروجی‌های مورد انتظار از پروژه

۱- الگوریتم تطبیق پشیمانی را در *MATLAB* پیاده‌سازی نمایید (برای بازی n نفره).

۲- برای بازی دو-نفرهٔ زیر، *polytope* حاوی کلیه تعادل‌های همبسته را ترسیم کنید (از طریق برنامه‌ریزی خطی یا *LP* رؤوس *polytope* را تعیین کرده و خود *polytope* را ترسیم کنید).

	D	C
D	(0,0)	(7,2)
C	(2,7)	(6,6)

۳- برای بازی فوق، در قالب نمودار ***U1-U2***، فضای بردارهای ***utility*** کل تعادل‌های همبسته و ***convex hull*** تعادل‌های نَش را ترسیم نمایید.

۴- نمودار همگرایی متوسط پشیمانی بازیگر ۱ به همسایگی صفر را در طول زمان ترسیم نمایید (نمودار، باید همگرایی تک تک درایه‌های ماتریس پشیمانی بازیگر ۱ به صفر را نشان دهد).

۵- بررسی کنید آیا پس از همگرایی، مقدار حدّیِ «توزیع تجربی» بازی یا همان $\boldsymbol{z}_k(.)$ که از اجرای الگوریتم تطبیق پشیمانی به دست می‌آید، داخل *polytope* تعادل‌های همبسته بازی فوق است یا با آن فاصله دارد؟ * توزیع تجربی بازی را می‌توان با استفاده از رابطه بازگشتی زیر (هر بار که کلیه بازیگران عمل انفرادی خود را مشخص کردند و در نتیجه، تکلیف عمل جمعی تکرار $k+1$-اُم روشن شد) بِروزرسانی نمود:

$$\boldsymbol{z}_{k+1} = \boldsymbol{z}_k + \frac{1}{k+1}\left[\boldsymbol{e}_{\mathbf{a}_{k+1}} - \boldsymbol{z}_k\right].$$

در رابطه فوق، $z_0(\boldsymbol{a})$ برای کلیه اَعمال جمعی $\boldsymbol{a} \in \boldsymbol{\mathcal{A}} = \times_{i=1}^{n} \mathcal{A}^n$ صفر است و $\boldsymbol{e}_i$ عبارتست از بُردار پایه‌ای که اندیس i-اُم آن ۱ است و مابقی مؤلفه‌های آن صفر می‌باشد.

مراجع

[1] H. P. Young, “Strategic learning and its limits,” in *Arne Ryde Memorial Lectures Series*, New York, NY, USA: Oxford Univ. Press, 2004.

[2] M.S. Talebi, "Uncoupled Learning Rules for Seeking Equilibria in Repeated Plays: An Overview," [Online]. Available: http://arxiv.org/abs/1310.5660

[3] S. Hart, "Adaptive heuristics," *Econometrica*, Vol. 73, No. 5, pp. 1401-1430, 2005 .

[4] H. A. Simon, The Sciences of the Artificial. MIT Press, 1969.

[5] S. Hart and A. Mas-Colell, “Uncoupled dynamics do not lead to Nash equilibrium,” *Amer. Econ. Rev.*, vol. 93, pp. 1830-1836, 2003.

[6] S. Hart and A. Mas-Colell, “A simple adaptive procedure leading to correlated equilibrium,” *Econometrica*, vol. 68, no. 5, pp. 1127–1150, Sep. 2000.

[7] N. Cesa-Bianchi, G. Lugosi, Prediction, Learning, and Games. Cambridge University Press, Cambridge, MA, 2006.