Suplemento / Supplement

Aprendizaje en Teoría de Juegos

Francisco Sánchez Sánchez[1]

Hace algunos años, prevalecían en la teoría de juegos los jugadores hiperracionales. Estos jugadores no sólo conocían y podían prever cualquier situación a la que pudieran enfrentarse, sino que también podían hacer lo propio con las de sus oponentes, eligiendo en cada caso la mejor opción. Como ejemplo, se pensaba que un jugador con una estrategia ganadora le debería ganar a Karpov.

Actualmente hay una línea de investigación donde esos jugadores con hiperracionalidad se sustituyen por otros que sólo se adaptan al medio. Se inicia con jugadores inexpertos. Los jugadores sólo guardan ponderaciones que reflejan la experiencia acumulada de sus partidas. La idea es comparar esta experiencia, por un lado, con el equilibrio de Nash, y por otro, con el comportamiento observado en experimentos controlados.

A continuación se presentan dos modelos de aprendizaje. Aunque la sencillez de los modelos no resalta la controversia sobre la racionalidad de los jugadores, sí ilustra el uso de esta técnica. En ambos modelos se consideran juegos en forma normal.

Un juego en forma normal es una terna: $(N,\{S_i\}_{i\in N},u)$, donde $N=\{1,...,n\}$ es un conjunto finito de jugadores, S_i es el conjunto de estrategias puras para el jugador i y u: $S \to \mathrm{R}^n$ y es una función de pago, con $S = S_1 \times ...\mathrm{S}_n$. Una partida se lleva a cabo en la siguiente forma: cada jugador de manera independiente elige una estrategia pura del conjunto que le corresponde, y se evalúa la función u en la n-ada de estrategias puras seleccionadas para determinar el pago de cada jugador. En este trabajo supondremos que las funciones de pago son cóncavas.

Usualmente los juegos bipersonales se representan con una tabla de doble entrada, donde el jugador 1 elige un renglón, el segundo una columna y la doble entrada indica el pago para cada jugador. Las estrategias puras del jugador 1 son los renglones y las del jugador 2 las columnas.

Una estrategia mixta para el jugador es una distribución sobre su espacio de estrategias puras. Se supone que el jugador selecciona su estrategia pura de acuerdo a la estrategia mixta que esté usando, y tiene por objeto esconder su elección.

---

1. Centro de Ivestigación de Matemáticas (CIMAT) Aplicadas.

La solución más común en juegos en forma normal es un equilibrio de Nash, el cual es un conjunto de estrategias mixtas, una por jugador, que es estable en el siguiente sentido: si todos los jugadores están usando su correspondiente estrategia y sólo uno la cambia, ese jugador no mejora su pago. En juegos bipersonales de suma cero, esta solución corresponde a la estrategia óptima.

● *Proceso de la jugada ficticia*

En el proceso de la jugada ficticia los jugadores desean deducir las estrategias mixtas de los demás jugadores. Así, los jugadores se comportan como si se enfrentaran a estrategias mixtas dadas, desconocidas de los oponentes. Supóngase que cada jugador actúa bajo este supuesto, de esta forma, cada jugador maneja una colección de ponderaciones, una para cada estrategia pura de sus oponentes. Cada ponderación representa la proporción de veces que su oponente ha usado la estrategia pura asociada.

Así, el proceso se lleva a cabo de la siguiente forma:

a) Cada jugador estima las estrategias mixtas de los demás jugadores:

$$\gamma^i_t(\tilde{\sigma}_j) = \frac{\varpi^i_t(\tilde{\sigma}_j)}{\Sigma_{\sigma j \in Sj}\varpi^i_t(\sigma_j)}$$

donde $\varpi^i_t(\tilde{\sigma}_j)$ es la ponderación que el jugador *i* asigna a la estrategia pura $\tilde{\sigma}_j$, del jugador *j*, a tiempo *t*. Posteriormente, cada jugador elige la estrategia pura que sea la mejor respuesta considerando estas estrategias mixtas dadas (si hubiera más de una, se elige cualquiera de ellas).

b) La partida se lleva a cabo con las elecciones de los jugadores. Cada jugador actualiza sus ponderaciones sumando uno, a las ponderaciones asociadas a estrategias puras usadas en la partida.

A la estrategia mixta γ^i_t que se deriva de las ponderaciones que el jugador *i* va actualizando, también se le conoce como las creencias del jugador *i* en el tiempo *t*. Denotaremos por σ^{-i} a la estrategia de los oponentes del jugador *i* contenida en σ y por la mejor respuesta del jugador *i* para σ^{-i} entenderemos el argumento que maximiza $u^i(\sigma^i,\sigma^{-i})$ al variar σ^i. Nótese que como las funciones de utilidad se suponen cóncavas, la mejor respuesta es única. Por último, diremos que un equilibrio de Nash es estricto, si para cada jugador *i*, σ^i es la mejor respuesta a σ^{-i}, esto es, el jugador *i* prefiere estrictamente σ^i a cualquier otra respuesta.

Fudenberg y Levin (1998) demuestran la siguiente proposición, que de alguna manera garantiza el buen comportamiento del proceso:

Proposición

a) Si σ es un equilibrio de Nash estricto y σ es jugado en t en el proceso de la jugada ficticia, entonces σ se juega en lo sucesivo.
b) Cualquier estrategia pura que sea estado "estacionario" del proceso de la jugada ficticia es un equilibrio de Nash.
c) Si las distribuciones marginales empíricas convergen, las estrategias correspondientes al producto es un equilibrio de Nash.

Demostración. Primero demostremos la parte a) supongamos que las evaluaciones γ^i_t de los jugadores corresponden a un equilibrio de Nash estricto σ. Cuando este equilibrio es jugado, cada creencia del jugador i en el periodo $t + 1$ es una combinación convexa de γ^i_t y la actualización de σ^{-i}:$\gamma^i_{t+1}=(1- \alpha_i)\gamma^i_t + \alpha_i\delta(\sigma^{-i})$. Como las utilidades esperadas son lineales en probabilidad:

$$u^i(\sigma^i,\gamma^i_{t+1}) = \alpha_t u^i(\sigma^i,\delta(\sigma^{-i})) + (1- \alpha_i)\, u^i(\sigma^i,\gamma^i_t)$$

y así, si σ^{-i} es la mejor respuesta del jugador i para γ^i_t esta es la mejor respuesta estricta para γ^i_{t+1}.

Para demostrar el inciso b) nótese que cualquier n-ada de estrategias puras σ que se juegue indefinidamente provoca que las evaluaciones empíricas de los jugadores converjan a ella. Si esta σ no es un equilibrio de Nash, alguno de los jugadores se desviará eventualmente.

Por último, según el inciso c) si el producto de las distribuciones empíricas converge a σ, entonces las creencias convergen a σ y de aquí, si σ no fuera un equilibrio de Nash, algún jugador eventualmente desearía desviarse.

En general el modelo es muy certero, sin embargo hay que tener cuidado al usarlo ya que el proceso puede ciclar, como se muestra en el siguiente ejemplo:

		J.2	
		A	B
J.1	A	(0,0)	(1,1)
	B	(1,1)	(0,0)

Supóngase que este juego es jugado con el proceso de la jugada ficticia, con ponderaciones iniciales (1,1.5) para cada jugador. En el primer periodo, ambos piensan que el otro va a jugar *B* con mayor probabili-

dad, para esta elección la mejor respuesta es jugar *A* y así lo hacen. En el siguiente periodo los pesos actualizados son (2,1.5) y ambos juegan *B*. El resultado es la sucesión alterna (B,B), (A,A), (B,B), (A,A)... Las distribuciones convergen a la estrategia mixta (1/2,1/2) las cuales forman un equilibrio de Nash, sin embargo el pago que reciben en cada periodo es (0,0). Nótese que la distribución empírica sobre las casillas no es una distribución independiente.

- *Modelo de Adaptación de Roth y Erev*

Se pretende construir modelos donde el aprendizaje de los jugadores satisfaga las leyes de aprendizaje que se manejan en la literatura de psicología. De entre estas leyes, sobresalen por su importancia las siguientes dos:

1. Las alternativas que han dado buen resultado en el pasado son más susceptibles de ser usadas en el futuro.
2. La curva de aprendizaje tiende a ser más pronunciada al principio y más plana después.

Ahora, para construir el modelo, supóngase que cada jugador tiene una ponderación por cada estrategia pura, la cual le indica la propensión a usar dicha estrategia pura. Para jugar una partida, el jugador construye la estrategia mixta que corresponde a sus ponderaciones y con base en ella selecciona su tirada. El resultado de la partida le sirve para corregir sus ponderaciones y reinicia el ciclo. En resumen, el modelo básico queda descrito por las dos siguientes ecuaciones:

$$\omega^i_{t+1}(\sigma_i) = \begin{cases} \omega^i_t(\sigma_i) + x & \text{si el jugacor } i \text{ jugó } \sigma, \text{ y recibió} \\ & \text{un pago } x \text{ de otra forma} \\ \omega^i_t(\sigma_i) & \end{cases}$$

$$p^i_{t+1}(\sigma_i) = \frac{\omega^i_t(\sigma_i)}{\Sigma_{\sigma_i \in S_i} \varpi^i_t(\sigma_i)}$$

Aunque en este modelo, la mayoría de las veces, el aprendizaje de los jugadores converge a algún equilibrio de Nash y satisface las leyes de psicología que se mencionaron, hay algunos juegos donde esto no sucede. Así, se han realizado algunas modificaciones al modelo básico para mejorar su desempeño. Algunas de ellas son:

1. Permitir que la probabilidad de una estrategia pura llegue a cero en tiempo finito. Es decir, si $p^i_{t+1}(\sigma_{ik}) < \mu \Rightarrow \omega^i_{tk}(\sigma_i) = 0$, donde μ es un valor pequeño preestablecido.

2. Afectar también a las estrategias conceptualmente cercanas. Sólo asignar $(1 - \varepsilon)\,x$ a la ponderación de σ_i y repartir εx entre las estrategias puras conceptualmente cercanas.

3. Se reduce gradualmente la importancia de la experiencia pasada, con solo multiplicar las ponderaciones por una constante menor que uno después de cada iteración.

4. Considerar puntos de referencia en los pagos para determinar si un resultado refuerza positiva o negativamente, por ejemplo, un resultado se puede considerar bueno si esta por encima de la media y malo en caso contrario.

Así, una forma alternativa de actualizar las ponderaciones sería:

$$\omega^{i}_{t+1}(\sigma_{ij}) = (1 - \varphi)\,\omega^{i}_{t}(\sigma_{ij}) + (x - \rho(t))$$

donde:

a) φ es un parámetro de olvido. Reduce gradualmente la importancia de la experiencia pasada.

b) $\rho(t)$ es un punto de referencia. Se refuerza positiva o negativamente de acuerdo a si el pago esta por arriba o por debajo de este. Este parámetro se puede actualizar usando la siguiente expresión:

$$\rho(t+1) = \begin{cases} (1 - \omega^{+})\ \rho(t) + \omega^{+}x & si\ x \geq \rho(t) \\ (1 - \omega^{-})\ \rho(t) + \omega^{-}x & si\ x < \rho(t) \end{cases}$$

donde ω^{+} y ω^{-} castigan en forma distinta las experiencias positivas y negativas.

El modelo planteado satisface otras leyes de aprendizaje que también son importantes, como son:

1. También las alternativas que se parecen a las que han tenido éxito son más susceptibles de ser usadas en el futuro.

2. Las experiencias recientes son más importantes que las anteriores.

3. Se generan puntos de referencia que determinan si el resultado refuerza positiva o negativamente.

4. Se supone que los reforzamientos negativos tienen un efecto mayor que los positivos

Alvin Roth e Ido Erev (1996) han estudiado un gran número de juegos con este tipo de modelos. En general, cuando la solución no es controvertida, el aprendizaje converge a un equilibrio de Nash, pero lo más interesante es que en juegos donde el comportamiento observado no coincide con algún equilibrio de Nash, el aprendizaje si lo hace.

Como ejemplo, considere el juego de ultimátum. Dos jugadores se tienen que repartir $100. El jugador 1 propone cómo repartirlo y el jugador 2 sólo puede aceptar o rechazar la oferta. Si el segundo jugador acepta, el monto se reparte de acuerdo a la propuesta y si la rechaza, ambos obtienen cero. Considerando $S_1 = \{1,...,99\}$, este juego tiene un solo equilibrio de Nash: el jugador 1 se queda con $99 y el segundo con $1.

Esta solución es muy mala. No es creíble que el segundo jugador acepte $1. En experimentos controlados, ofertas donde el jugador 1 se queda con más de $70, ya casi siempre son rechazadas y este comportamiento es el que se observa en el modelo de aprendizaje anterior.

● *Referencias*

Alvin E. Roth e Ido Erev (1995) "Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term," *Games and Economic Behavior* 8, 164-212.

—— (1996) "On the Need for Low Racionality, Cognitive Game Theory: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria," Working Paper, MIT.

Drew Fudenberg y David K. Levin (1998) *Theory of Learning in Games*, MIT Press.