Decimocuarta Semana Internacional de la Estadística y la Probabilidad
14-18 de junio de 2021

Aproximación numérica vía Q-Learning a un problema de consumo-inversión.

Ruy Alberto López-Ríos[a], Hugo Adán Cruz-Suárez[b], Fernando Velasco Luna[c]
[a,b,c] *Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias Físico Matemáticas, Puebla, Puebla, México.*
[a] ruy.lopez@alumno.buap.mx, [b] hcs@fcfm.buap.mx

Resumen

El trabajo se centra en un problema de consumo-inversión en tiempo discreto con un horizonte infinito. Este problema se formula como un proceso de decisión de Markov con una utilidad descontada total esperada como criterio de rendimiento, y función de utilidad dependiente del consumo de tipo logarítmico y exponencial.

Se tiene como objetivo el presentar procedimientos de aproximación para la solución a través de algoritmos *machine learning*, específicamente, la técnica de *Q-Learning*. Estos métodos toman mayor ventaja al trabajar con espacios de estados y de acciones muy grandes. Se proporcionan resultados numéricos del problema.

Palabras clave: Problema de consumo-inversión, Programación dinámica, Q-Learning

Introducción

El control óptimo estocástico es un área de las matemáticas dedicada a resolver problemas de optimización cuya evolución en el tiempo es susceptible a ser influenciado por variables aleatorias. Los procesos de control de Markov (PCM) son problemas de control estocástico, también conocidos como procesos de decisión de Markov o procesos de Markov controlados. La técnica básica para resolver problemas de control de Markov es la programación dinámica.

Los métodos numéricos para resolver problemas de control óptimo presentan un limitante al momento de discretizar el tiempo o los espacios de estados o acciones. Con la finalidad de aumentar la eficiencia y precisión de tales métodos se han creado versiones que combinan técnicas de aproximación iterativas y recurrir a otras herramientas que eviten o que minimicen el efecto de tales complicaciones.

La técnica Q-Learning pertenece a una clase de algoritmos machine learning, y es un algoritmo de refuerzo por aprendizaje el cual fue introducido por Watkins en 1989, ver (Watkins). Q-learning es un método asíncrono de programación dinámica que permite al controlador aprender a actuar óptimamente en dominios Markovianos, en (Watkins) se demuestra que la solución en los algoritmos de Q-learning convergen a los valores de acciones óptimos con probabilidad uno. Por lo tanto, Q-learning es una técnica adecuada para implementar en la solución del problema de consumo- inversión.

El trabajo está organizado como sigue:

Sección 1: Preliminares. Sección 2: presenta el método Q-Learning. Sección 3: se muestra un algoritmo de aproximación para un problema de consumo inversión.

Preliminares

Procesos de control de Markov

Definición. Un modelo de control de Markov es una quíntupla

$$(X, A, \{A(x)|x \in X\}, Q, r)$$

la cual consiste de los siguientes elementos:
a) un espacio de Borel X, llamado espacio de estados;
b) un espacio de Borel A, llamado espacio de control (o acción);

c) una familia $\{A(x)|x \in X\}$ de subconjuntos medibles A(x) de A, donde A(x) denota el conjunto de controles (o acciones) admisibles cuando el sistema está en el estado $x \in X$, y con la propiedad de que el conjunto

$K = \{(x,a)|x \in X, a \in A(x)\}$ de parejas estado-acción admisibles es un subconjunto medible de $X \times A$;

d) un kernel estocástico Q sobre X dado K llamada ley de transición.

Denotamos x_t y a_t el estado del sistema y el control (o acción) aplicado en el tiempo t, respectivamente, la evolución del sistema puede ser descrita como sigue. Si el sistema está en el estado $x_t = x \in X$, que representa el capital en el tiempo t y el control $a_t = a \in A(x)$ es aplicado, que representa la inversión, entonces dos cosas ocurren:

i) una recompensa r(x,a) es incurrida. Dado el contexto general de aplicaciones en este escrito, consideraremos de ahora en adelante a la función de recompensa r como una función de utilidad u del consumo x − a, es decir $r(x, a) := u(x, a) = u(x - a)$, y

ii) el sistema se mueve al siguiente estado x_{t+1}, que es una variable aleatoria con valores en X con distribución $Q(\cdot|x, a)$.

Una vez que ocurre la transición al nuevo estado, un nuevo control es elegido y el proceso es repetido.

Se busca optimizar un criterio de rendimiento, uno de ellos es la utilidad total descontada esperada:

$$V(\pi, x) := \mathbb{E}_x^{\pi}\left[\sum_{t=0}^{\infty} \alpha^t\, u(x_t - a_t)\right],$$

donde α (0 < α < 1) es llamado factor de descuento. El problema de control óptimo correspondiente a maximizar es una utilidad de consumo α-descontada.

Políticas

Definición. Se define el espacio H_t de historias admisibles hasta el tiempo t como

$$H_0 := X$$

$$H_t := K^t \times X = K \times H_{t-1} \quad \text{para } t=1,2,...$$

Una política de control aleatorizada o política de control es una sucesión $\pi = \{\pi_t\}^{\infty}_{t=0}$ de kérneles estocásticos π_t sobre A dada la t-historia admisible como: $h_t = (x, a_0, x_1, a_1, ..., x_{n-1}, a_{t-1}, x_t)$,

para cada t = 0,1,..., con $(x_t, a_t) \in K$. Esto es, para $t \geq 0$, π_t es un kernel estocástico si satisface las siguientes propiedades:

a) $\pi_t(\cdot \mid h_t)$ es una medida de probabilidad sobre X, para cada $h_t \in K^t \times X$.
b) $\pi_t(B \mid \cdot)$ es una variable aleatoria para cada $B \in B(X)$, donde B(X) denota la σ-álgebra de Borel de X.
Es decir, se satisface la restricción

$$\pi_t\big(A(x_t)|h_t\big) = 1 \quad \forall h_t \in H_t, \;\; t = 0, 1, \dots .$$

El conjunto de todas las políticas es denotado por Π.

Definición. Una política π^* satisfaciendo

$$V(\pi^*, x) = \sup_{\Pi} V(\pi, x) =: V^*(x),$$

se dice ser política óptima α−descontada, y V^* es llamada función de valor (función objetivo o función de utilidad óptima) α−descontada.

Definición. Una función medible $\vartheta : X \to R$ se dice ser una solución de la ecuación de optimalidad con utilidad α−descontada si satisface

$$\vartheta(x) = \max_{a \in \mathbb{A}(x)} \left\{ u(x,a) + \alpha \int_{\mathbb{X}} \vartheta(y)\, Q(dy|x,a) \right\}$$

Las funciones de iteración de valor están definidas como:

$$\vartheta_n(x) = \max_{a \in \mathbb{A}(x)} \left\{ u(x,a) + \alpha \int_{\mathbb{X}} \vartheta_{n-1}(y) Q(dy|x,a) \right\}$$

Se puede demostrar que V^* es una solución de la ecuación de optimalidad con utilidad α−descontada, es decir,

$$V^*(x) = \max_{a \in \mathbb{A}(x)} \left\{ u(x,a) + \alpha \int_{\mathbb{X}} V^*(y)\, Q(dy|x,a) \right\},$$

y que V^* satisface:

$$V^*(x) = \lim_{n \to \infty} \vartheta_n(x),$$

Requerimos de algunas suposiciones para garantizar la existencia de solución para el problema de optimización, por ejemplo, ver (Hernández-Lerma):

Suposición

a) La función de utilidad u(·) es semicontinua inferior, no-positiva e inf-compacta sobre K.
b) Q es fuertemente continua.

Una de las grandes complicaciones de la programación dinámica es combatir la maldición de la dimensionalidad originada al manejar espacios de estados o acciones demasiado grandes, por lo cual se suele optar por métodos de refuerzo de aprendizaje. La siguiente Sección trata específicamente del método *Q-Learning*.

Q-Learning

Los algoritmos de Refuerzo de Aprendizaje (RL) utilizan la función objetivo (o función de valor) de la programación dinámica. En RL la función objetivo es almacenada en funciones llamadas Q-Factores. Ver (Gosavi).

Definición Para $\tilde{x_i} \in \tilde{X}$, $\tilde{a} \in \tilde{A}(\tilde{x_i})$ se define la función de Q-factores como sigue:

$$Q(\tilde{x}_i, \tilde{a}) := u(\tilde{a}) + \alpha \sum_{j=1}^{|\tilde{\mathbb{X}}|} J^*(\tilde{x}_j)\, \hat{\mathcal{Q}}(\tilde{x}_j \,|\, \tilde{x}_i, \tilde{a}).$$

Donde $\tilde{X}$ y $\tilde{A}$, son los espacios discretizados de los correspondientes espacios de estados y acciones X y A.

En (López-Ríos) puede encontrarse el algoritmo modificado Q-Learning para procesos de decisión de Markov con recompensa descontada, basado en el algoritmo clásico en la literatura de iteración de valor conjuntamente con el algoritmo de Robbins-Monro (Robbins & Monro).

Paso 1: Inicializar los Q-Factores. Para todo $(\tilde{x}, \tilde{a})$ donde $\tilde{x} \in \tilde{X}$ y $\tilde{a} \in \tilde{A}(x)$, $Q(\tilde{x}, \tilde{a}) = 0$.

Hacer k = 0, el número de transiciones de estado.
El algoritmo se ejecutará un número k_{max} de iteraciones. Iniciar la simulación del sistema en un estado arbitrario $\tilde{x}_i \in \tilde{X}$.

Paso 2: En el estado actual $\tilde{x}_i$, seleccionar una acción $\tilde{a}$ con probabilidad $1/ ||\tilde{A}(\tilde{x}_i)||$

Paso 3: Simular la acción $\tilde{a}$. Producir el siguiente estado $\tilde{x}_j$. Sea $u(\tilde{a})$ la utilidad inmediata ganada en la transición al estado $\tilde{x}_j$ desde al estado actual $\tilde{x}_i$ bajo la influencia de la acción $\tilde{a}$. Incrementar el valor de k en una unidad. Actualizar el valor de la función λ, conocida como tasa de refuerzo (o aprendizaje).

Paso 4: Actualizar $Q(\tilde{x}_i, \tilde{a})$ usando la ecuación:

$$Q(\bar{x}_i, \bar{a}) = (1 - \lambda_k) Q(\bar{x}_i, \bar{a}) + \lambda_k \left\{ u(\bar{a}) + \alpha \max_{\bar{b} \in \bar{A}(\bar{x}_j)} Q(\bar{x}_j, \bar{b}) \right\}$$

Paso 5: Si $k<k_{max}$, hacer $\tilde{x}_i \leftarrow \tilde{x}_j$ e ir al Paso2. De otra manera, ir al Paso6.

Paso 6: Para cada $\tilde{x}_i \in \tilde{X}$, seleccionar

$$d(\bar{x}_i) \in \operatorname*{argmax}_{\bar{b} \in \bar{A}(\bar{x}_i)} Q(\bar{x}_i, \bar{b}).$$

Aproximación numérica a un problema de consumo-inversión

Presentamos una versión controlada del problema presentado en (Cruz-Suárez, Montes-De-Oca & Zacarías). Primero, supongamos que la riqueza de un inversor es gobernada por una ley definida por la ecuación en diferencias, donde x_t es la riqueza actual al tiempo t, para t = 0,1,2, ..., $\delta < 1$, y $\{\xi_t\}$ es una sucesión de variables aleatorias independientes e idénticamente distribuidas con función de densidad Δ. Suponga que el inversor quiere administrar de manera óptima su capital actual x_t, dedicando parte de este capital a su consumo a_t y el resto, $h(x_t)-a_t$, a inversión. En particular, considere la función de producción h(x) con $x \in X := [0,\infty)$ llamado el espacio de estados. La ley de transición es entonces dada por:

$$x_{t+1} = (h(x_t) - a_t)\,\xi_t,$$

y X_0 = x conocida. Préstamos no son permitidos. Por lo tanto, $a_t \in [0, h(x_t)]$ denota el consumo al tiempo t, $A(x_t) := [0, h(x_t)]$ es el espacio de acciones admisibles al tiempo t, y $A:=[0, +\infty)$ es el conjunto de acciones admisibles.

La dinámica descrita en este sistema de consumo-inversión es como sigue: si el sistema es observado en el tiempo t, el estado considerado es $x_t = x \in X = [0,+\infty)$ y la acción $a_t = a \in A(x)$ es aplicada. Una utilidad u(a) es obtenida y el sistema se mueve al siguiente estado, $x_{t+1} \in X$, por medio de la ley de transición descrita. Este proceso se repite mientras se acumulan las recompensas descontadas en cada tiempo t hasta un

horizonte infinito de acuerdo a un criterio de rendimiento.

Dado un capital inicial $X_0 = x \in X$, un plan $\pi \in \Pi$, y una utilidad de consumo u=Log(a): K→R. El criterio de rendimiento usado para evaluar la calidad del plan $\pi \in \Pi$ es dado por:

$$V(\pi, x) := \mathbb{E}_x^{\pi}\left[\sum_{t=1}^{\infty} \alpha^t \, Log(\tilde{a}_t)\right], \quad \pi \in \Pi, x \in \mathbb{X},$$

La función de producción y la función de utilidad de consumo satisfacen ciertas condiciones usuales. Ver (De La Fuente).

Suposición. En el contexto del problema previo, las funciones de producción $h(x) := x^{\delta}$ y utilidad u(a) := Log(a) satisfacen:

a) $h \in C^2((0, \infty), (0, \infty))$,
b) h es una función cóncava sobre X,
c) $h' > 0$ y $h(0)=0$,
d) $u \in C^2((0, \infty), R)$ es estrictamente creciente y estrictamente cóncava, y,
e) u' es una función invertible, $u'(0) = \infty$, y $\lim_{a\to\infty} u'(a) = 0$.

El algoritmo para aplicar Q-learning al proceso de decisión de Markov con recompensas descontadas por medio de iteración de valores y el algoritmo de Robbins-Monro, ver (Gosavi) y (Robbins & Monro) se presenta a continuación:

Algoritmo. Q-Learning modificado

Initialización Especificar k_{max} y parámetros de $\{\xi\}$;
for $k = 0$ **to** k_{max} **do**
  Simular ξ_k;
  $a_k \longleftarrow x^{\delta} - \frac{x}{\xi_k}$ tal que $a_k \geq 0$ y $a_k < x^{\delta}$;
  $a_k^r \longleftarrow \text{Round}[a_k]$;
end
Poner $r_1 := Round\{Min\{a_k^r\}\}$, $r_2 := Round\{Max\{a_k^r\}\}$;
for $j = 0$ **to** $(r_2 - r_1)100$ **do**
  $\tilde{a}_j \longleftarrow r_1 + j/100 \quad (\tilde{a}_j \in \tilde{\mathbb{A}}(\tilde{x}))$;
  $Q_0(\tilde{x}, \tilde{a}_j) \longleftarrow 0$;
end
repeat
  for $x = \tilde{x}_i \in \tilde{\mathbb{X}}$ *y* $\tilde{a} \in \tilde{\mathbb{A}}(\tilde{x}_i)$ **do**
    $Q_{k+1}(\tilde{x}_i, \tilde{a}_j) \longleftarrow Q_k(\tilde{x}_i, \tilde{a}_j) + \lambda_k\left[Log(\tilde{a}_j) + \alpha \max_{\tilde{b}\in\tilde{\mathbb{A}}(\tilde{x}_j)} Q_k(\tilde{x}_j, \tilde{b}) - Q_k(\tilde{x}_i, \tilde{a}_j)\right]$
  end
  Simular el siguiente estado $\tilde{x}_j$, hasta retornar a $x = \tilde{x}_i$;
  $k \longleftarrow k + 1$;
until $k \geq k_{max}$;
return $\tilde{\pi}^*(x) \in \underset{\tilde{b}\in\tilde{\mathbb{A}}(x)}{argmax}\, Q_{k_{max}}(x, \tilde{b})$.

Considerando el capital inicial es x y $\alpha = 0.5$ es el factor de descuento. La transición del sistema es regida por:

$$x_{t+1} = (x_t^{\delta} - a_t)\,\xi_t,$$

donde $\delta=0.75$ y $\{\xi_t\}$ es tal que $Log(\xi_t) \sim N(0,0.4)$,

Para el método Q-Learning utilizamos el intervalo de paso $\lambda_k(x_i, a) := Log(k)/k$ como tasa de refuerzo o aprendizaje y $k_{max} = 1000$.

En (Cruz-Suárez, Montes-De-Oca & Zacarías) se encuentra que la política óptima es dada por

$$a^*(x) = x^{\delta}(1 - \alpha\delta), \quad x \in \mathbb{X}.$$

Resultado que se requiere para fines comparativos entre la solución exacta y la aproximada.

Tabla 1. Política óptima aproximada y exacta

x	$\hat{a}^*(x)$	$a^*(x)$
0.25	0.26	0.220
0.50	0.40	0.371
0.75	0.52	0.503
1.00	0.64	0.625
1.25	0.70	0.738
1.50	0.84	0.847
1.75	1.07	0.950
2.00	1.12	1.051
2.25	1.12	1.148
2.50	1.24	1.242
2.75	1.33	1.334
3.00	1.44	1.424
3.25	1.49	1.512
3.50	1.69	1.599
3.75	1.65	1.684
4.00	1.79	1.767
4.25	1.79	1.850
4.50	1.84	1.931
4.75	1.94	2.010
5.0	2.10	2.089

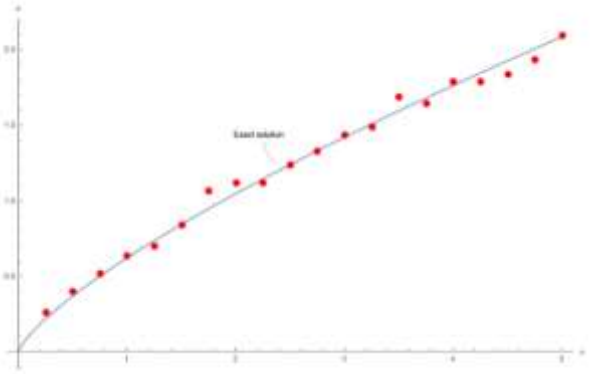

Figura 1. Política óptima aproximada

La Tabla 1 muestra las acciones óptimas aproximadas para algunos valores de x, obtenidas usando el método descrito en el algoritmo Q-Learning, contrastadas con la política óptima exacta. La Figura 1 muestra otras acciones óptimas aproximadas, denotadas por puntos rojos, que exhiben un comportamiento similar que el de la política óptima exacta, graficada por una línea sólida azul.

El algoritmo modificado propuesto de Q-Learning es una buena aproximación para las acciones óptimas con los parámetros dados, sin necesidad de recurrir a la solución cerrada, si es que existe.

Conclusiones

Q-learning encuentra una política óptima en el sentido de maximizar el valor esperado de la recompensa total en cada iteración en todos los estados por donde transite el sistema, comenzando desde el estado actual.

El entorno se establece típicamente en la forma de un proceso de decisión de Markov, porque muchos algoritmos de aprendizaje por refuerzo para este contexto utilizan técnicas de programación dinámica.

El método propuesto en este trabajo ofrece una alternativa de solución a la política óptima de un problema de control. Este método también es útil cuando no se dispone de una solución cerrada o cuando las herramientas de resolución clásicas no encuentran la política óptima directamente.

En las estrategias de inversión óptimas que se derivan en forma cerrada, suelen proponerse funciones candidatas con la forma general de la solución. Estas funciones pueden ser muy difíciles de encontrar cuando se utilizan funciones de producción o de utilidad más complejas o desconocidas.

Los métodos de refuerzo de aprendizaje tienen una ventaja en términos de capacidad de implementación, asignando valores de bondad a los pares estado-acción, poniéndolos a prueba y descartando selecciones de acciones que produzcan resultados adversos.

Referencias

Cruz-Suárez, H. A., Montes-de-Oca, R., & Zacarías, G., (2011). A consumption-investment problem modelled as a discounted Markov decision process, **Kybernetika**, 47(6), 909-929.

De La Fuente, A. (2000). **Mathematical Methods and Models.** Cambridge: Cambridge University Press.

Gosavi, A., (2015). **Simulation-Based Optimization. Parametric Optimization Techniques and Reinforcement Learning**, Springer. 2a. ed.

Hernández-Lerma, O., Laserre, J. B., **Discrete-Time Markov Control Processes. Basic Optimality Criteria**, Applications of Mathematics, Stochastic Modelling and Applied Probability, Vol. 30, Springer, USA, (1996).

Robbins, H., & Monro, S. A stochastic approximation method, **Annals of Mathematical Statistics**, University of North Carolina, 22(3), pp. 400-407. (1951).

López-Ríos, R. A., & Cruz-Suárez H. A., (2021). Q-learning approach to a Consumption-Investment Problem, **International Journal of Statistics and Probability**, 10(2),

Watkins, C., & Dayan, P., (1992). **Q-Learning. Technical note**. Machine Learning, (8), pp. 279-292. Boston, MA: Kluwer Academic Publishers.