Aprendizaje Reforzado: Regulación cuadrática lineal

Lección 26

Dr. Pablo Alvarado Moya

CE5506 Introducción al reconocimiento de patrones
Área de Ingeniería en Computadores
Tecnológico de Costa Rica

II Semestre, 2019

Contenido

Repaso (1)

- Definimos un MDP como la tupla $\langle \mathcal{S}, \mathcal{A}, \{P_{sa}\}, \gamma, R \rangle$ con estados.
- El factor de degradación $\gamma \in [0,1)$
- Función de recompensa $R : \mathcal{S} \to \mathbb{R}$.
- En la **iteración de valor**, repetimos hasta convergencia

$$V(s) \leftarrow R(s) + \max_{a \in \mathcal{A}} \gamma \sum_{s'} P_{sa}(s') V(s')$$

que al final será equivalente a $V^*(s)$

- Con $V^*(s)$ encontramos π^*

$$\pi(s) = \arg\max_{a \in \mathcal{A}} \sum_{s'} P_{sa}(s') V(s')$$

Repaso (2)

- En el caso de estados continuos, lo que hicimos fue aproximar $V(s)$.
- En esos casos usamos regresión para la aproximación del siguiente estado $\underline{\mathbf{s}}_{(t+1)}$ y un simulador para poder estimar valores esperados.

Plan de acción

- Lo que haremos ahora es proponer algunas modificaciones comunes a MDP que
  - usan otra definición de la función de recompensa R
  - usan otra estrategia para degradar la recompensa
  - nos permiten calcular $V(s)$ exactamente a pesar de continuidad de s

Recompensas por estado acción

- Después de usar una función de recompensa $R : \mathcal{S} \to \mathbb{R}$, regresaremos a la definición que considera también la acción a tomar.
- De nuevo usaremos $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
- Con eso, redefinimos el saldo total como

$$R(s_{(0)},a_{(0)}) + \gamma R(s_{(1)},a_{(1)}) + \gamma^2 R(s_{(2)},a_{(2)}) + \ldots$$

- Queremos encontrar la política que maximiza el valor esperado del saldo total:

$$\mathrm{E}\left[R(s_{(0)},a_{(0)}) + \gamma R(s_{(1)},a_{(1)}) + \gamma^2 R(s_{(2)},a_{(2)}) + \ldots\right]$$

- Esto permite incorporar diferencias entre el costo que conlleva tomar cierta acción.
  Por ejemplo, "deténgase" es distinto a "acelere" en términos energéticos.

Ecuaciones de Bellman

- Las ecuaciones de Bellman serían ahora

$$V^*(s) = \max_{a \in \mathcal{A}} \left(R(s,a) + \gamma \sum_{s'} P_{sa}(s') V^*(s') \right)$$

- En la **iteración de valor**, repetimos hasta convergencia

$$V(s) \leftarrow \max_{a \in \mathcal{A}} \left(R(s,a) + \gamma \sum_{s'} P_{sa}(s') V(s') \right)$$

que al final será equivalente a $V^*(s)$

- La política óptima estará dada por:

$$\pi^*(s) = \arg\max_{a \in \mathcal{A}} \left(R(s,a) + \gamma \sum_{s'} P_{sa}(s') V^*(s') \right)$$

MDP de horizonte finito

- Un MDP de **horizonte finito** es una tupla $\langle \mathcal{S}, \mathcal{A}, \{P_{\underline{\mathbf{sa}}}\}, T, R \rangle$, donde T es el tiempo de horizonte.
- Redefinimos el saldo total como

$$R(s_{(0)}, a_{(0)}) + R(s_{(1)}, a_{(1)}) + \cdots + R(s_{(T)}, a_{(T)})$$

- Queremos encontrar el valor esperado del saldo total, que ahora percibe solo una ventana de tiempo.
- La política óptima puede ser **no-estacionaria**, es decir, que depende del tiempo.
- ¡Las acciones óptimas dependerán de cuándo en el tiempo se deban tomar! pues existirán recompensas no visibles en la ventana de tiempo.

Probabilidades de transición no estacionarias

- Hasta ahora hemos supuesto que

$$\underline{\mathbf{s}}_{(t+1)} \sim P_{\underline{\mathbf{s}}_t, \underline{\mathbf{a}}_t}$$

  con $P_{\underline{\mathbf{s}}_t, \underline{\mathbf{a}}_t}$ estacionario (no depende del tiempo).
- Estas probabilidades pueden depender del tiempo (no-estacionarias):

$$\underline{\mathbf{s}}_{(t+1)} \sim P^{(t)}_{\underline{\mathbf{s}}_t, \underline{\mathbf{a}}_t}$$

- Ejemplos:
  1. probabilidades de transición de un avión dependen de cuánto combustible tiene (pues la masa se reduce con el tiempo) o las condiciones climáticas.
  2. modelo de tránsito depende de la hora del día

Recompensas no-estacionarias

- Para generalizar el modelo anterior, también podemos suponer que la función de recompensa es no-estacionaria.
- Redefinimos el saldo total como

$$R_{(0)}(s_{(0)},a_{(0)}) + R_{(1)}(s_{(1)},a_{(1)}) + \ldots + R_{(T)}(s_{(T)},a_{(T)})$$

Estimación de función de valor (1)

- Queremos un algoritmo que estime la política óptima
- Modifiquemos la definición de función de valor óptima:

$$V^*_{(t)}(\underline{\mathbf{s}}) = \mathrm{E}\left[R_{(t)}(\underline{\mathbf{s}}_t,\underline{\mathbf{a}}_t) + \ldots + R_{(T)}(\underline{\mathbf{s}}_{(T)},\underline{\mathbf{a}}_{(T)})\middle|\pi^*,\underline{\mathbf{s}}_{(t)} = \underline{\mathbf{s}}\right]$$

- La **iteración de valor** para este caso es

$$V^*_{(t)}(\underline{\mathbf{s}}) = \max_{\underline{\mathbf{a}}\in\mathcal{A}}\left(R_{(t)}(\underline{\mathbf{s}},\underline{\mathbf{a}}) + \sum_{\underline{\mathbf{s}}'\in\mathcal{S}} P^{(t)}_{\underline{\mathbf{sa}}}(\underline{\mathbf{s}}')V^*_{(t+1)}(\underline{\mathbf{s}}')\right)$$

que es un problema de programación dinámica solucionable con recursión.

Estimación de función de valor (2)

- Para iniciar la recursión usamos

$$V^*_{(T)}(\underline{\mathbf{s}}) = \max_{\underline{\mathbf{a}} \in \mathcal{A}} R_{(T)}(\underline{\mathbf{s}}, \underline{\mathbf{a}})$$

pues no consideramos nada luego de T, y retrocedemos en el tiempo.

- Luego encontramos la política óptica con

$$\pi^*_{(t)}(\underline{\mathbf{s}}) = \arg\max_{\underline{\mathbf{a}} \in \mathcal{A}} \left(R_{(t)}(\underline{\mathbf{s}}, \underline{\mathbf{a}}) + \sum_{\underline{\mathbf{s}}' \in \mathcal{S}} P_{\underline{\mathbf{sa}}}(\underline{\mathbf{s}}') V^*_{(t+1)}(\underline{\mathbf{s}}') \right)$$

- Una diferencia de esta versión con tiempo de horizonte es que el resultado es directo, y no requiere iteraciones para convergencia asintótica.

Estimación de función de valor (3)

- En la implementación, calculamos

$$V^*_{(T)}, V^*_{(T-1)}, V^*_{(T-2)}, \cdots V^*_{(t)}$$

y con ellos calculamos

$$\pi^*_{(T)}, \pi^*_{(T-1)}, \pi^*_{(T-2)}, \cdots \pi^*_{(t)}$$

Sistemas dinámicos lineales

- Vamos a usar los dos cambios anteriores: recompensas por estado-acción, y MDP con horizonte temporal finito, para resolver un tipo muy particular de sistemas, que sin embargo se encuentran a menudo.
- **Regulación cuadrática lineal** nos permitirá aplicar los conceptos anteriores a casos de espacios de estado y de acción contínuos.

Planteando el nuevo problema LQR

- Un MDP se especifica con quíntupla: $\langle \mathcal{S}, \mathcal{A}, \{P_{\underline{\mathbf{sa}}}\}, T, R \rangle$
- Asumiremos el espacio de estados $\mathcal{S} = \mathbb{R}^n$
- El espacio de acciones será $\mathcal{A} = \mathbb{R}^d$.
- Para las probabilidades de transición usaremos:

$$P_{\underline{\mathbf{sa}}}^{(t)} : \underline{\mathbf{s}}_{(t+1)} = \mathbf{A}_{(t)}\underline{\mathbf{s}}_{(t)} + \mathbf{B}_{(t)}\underline{\mathbf{a}}_{(t)} + \underline{\mathbf{w}}_{(t)}$$

  con el ruido gaussiano $\underline{\mathbf{w}}_t \sim \mathcal{N}(\underline{\mathbf{0}}, \mathbf{\Sigma}_{\underline{\mathbf{w}}})$, y las matrices conocidas *a-priori* $\mathbf{A}_{(t)} \in \mathbb{R}^{n \times n}$, $\mathbf{B}_{(t)} \in \mathbb{R}^{n \times d}$
- Posteriormente veremos que $\underline{\mathbf{w}}_t$ no es muy importante y lo podremos ignorar.

Recompensa cuadrática en LQR

- Supondremos que la recompensa está dada por

$$R_{(t)}\left(\underline{\mathbf{s}}_{(t)},\underline{\mathbf{a}}_{(t)}\right) = -\left(\underline{\mathbf{s}}_{(t)}^{T}\mathbf{U}_{(t)}\underline{\mathbf{s}}_{(t)} + \underline{\mathbf{a}}_{(t)}^{T}\mathbf{V}_{(t)}\underline{\mathbf{a}}_{(t)}\right)$$

  con $\mathbf{U}_{(t)} \in \mathbb{R}^{n\times n}$ y $\mathbf{V}_{(t)} \in \mathbb{R}^{d\times d}$, ambas positivas semidefinidas, por lo que $R_{(t)} \leq 0$.
- Por ejemplo:
  - Supongamos un sistema diseñado tal que queremos $\underline{\mathbf{s}}_{(t)} \approx 0$.
  - Podríamos elegir $\mathbf{U}_{(t)} = \mathbf{I}$, $\mathbf{V}_{(t)} = \mathbf{I}$ de modo que

$$R_{(t)}\left(\underline{\mathbf{s}}_{(t)},\underline{\mathbf{a}}_{(t)}\right) = -\left(\|\underline{\mathbf{s}}_{(t)}\|^2 + \|\underline{\mathbf{a}}_{(t)}\|^2\right)$$

  donde la intención del término $\|\underline{\mathbf{a}}_{(t)}\|^2$ es desalentar un uso arbitrario y exagerado de acciones.

Dinámica no estacionaria

- Derivaremos ahora el caso general de dinámica no estacionaria.
- En este caso tendremos variación temporal en la recompensa y en la dinámica del sistema.
- Para simplificar la comprensión, las derivaciones pueden suponerse estacionarias, es decir, con $\mathbf{M} = \mathbf{M}_{(1)} = \ldots = \mathbf{M}_{(T)}$ con todas las matrices involucradas $\mathbf{A}, \mathbf{B}, \mathbf{U}$ y $\mathbf{V}$.

Modelo lineal

- La suposición principal en lo que sigue es que la **dinámica** del sistema es **lineal**:

$$\underline{\mathbf{s}}_{(t+1)} = \mathbf{A}\underline{\mathbf{s}}_{(t)} + \mathbf{B}\underline{\mathbf{a}}_{(t)}$$

- Realizamos m experimentos con el sistema:

$$\underline{\mathbf{s}}_{(0)}^{(1)} \xrightarrow{\underline{\mathbf{a}}_{(0)}^{(1)}} \underline{\mathbf{s}}_{(1)}^{(1)} \xrightarrow{\underline{\mathbf{a}}_{(1)}^{(1)}} \underline{\mathbf{s}}_{(2)}^{(1)} \xrightarrow{\underline{\mathbf{a}}_{(2)}^{(1)}} \cdots \underline{\mathbf{s}}_{(T)}^{(1)} \cdots$$

$$\underline{\mathbf{s}}_{(0)}^{(2)} \xrightarrow{\underline{\mathbf{a}}_{(0)}^{(2)}} \underline{\mathbf{s}}_{(1)}^{(2)} \xrightarrow{\underline{\mathbf{a}}_{(1)}^{(2)}} \underline{\mathbf{s}}_{(2)}^{(2)} \xrightarrow{\underline{\mathbf{a}}_{(2)}^{(2)}} \cdots \underline{\mathbf{s}}_{(T)}^{(2)}$$

$$\vdots$$

$$\underline{\mathbf{s}}_{(0)}^{(m)} \xrightarrow{\underline{\mathbf{a}}_{(0)}^{(m)}} \underline{\mathbf{s}}_{(1)}^{(m)} \xrightarrow{\underline{\mathbf{a}}_{(1)}^{(m)}} \underline{\mathbf{s}}_{(2)}^{(m)} \xrightarrow{\underline{\mathbf{a}}_{(2)}^{(m)}} \cdots \underline{\mathbf{s}}_{(T)}^{(m)}$$

Estimando modelo dinámico

- Para encontrar el modelo dinámico entonces optimizamos

$$\arg\min_{\mathbf{A},\mathbf{B}} \frac{1}{2} \sum_{i=1}^{m} \sum_{t=0}^{T-1} \left\| \underline{\mathbf{s}}_{(t+1)} - \left(\mathbf{A}\underline{\mathbf{s}}_{(t)} + \mathbf{B}\underline{\mathbf{a}}_{(t)} \right) \right\|^2$$

- Otra forma de llegar a un sistema lineal es **linealizando** un sistema no-lineal.

Linealización

- Supongamos que tenemos una función f no-lineal tal que

$$\underline{\mathbf{s}}_{(t+1)} = f\left(\underline{\mathbf{s}}_{(t)}, \underline{\mathbf{a}}_{(t)}\right)$$

- Usamos una serie de Taylor de primer orden alrededor de un punto $(\bar{\underline{\mathbf{s}}}_{(t)}, \bar{\underline{\mathbf{a}}}_{(t)})$ como aproximación lineal.
- Obviemos por un instante la dependencia de la acción:

$$\begin{aligned}\underline{\mathbf{s}}_{(t+1)} &= f(\underline{\mathbf{s}}_{(t)}) \\ &\approx f(\bar{\underline{\mathbf{s}}}_{(t)}) + \mathbf{J}_f(\bar{\underline{\mathbf{s}}}_{(t)})(\underline{\mathbf{s}}_{(t)} - \bar{\underline{\mathbf{s}}}_{(t)})\end{aligned}$$

  con $\mathbf{J}_f(\bar{\underline{\mathbf{s}}}_{(t)})$ el jacobiano de f evaluado en $\bar{\underline{\mathbf{s}}}_{(t)}$.
- La linealización obviamente solo es válida en una pequeña vecindad alrededor de $(\bar{\underline{\mathbf{s}}}_{(t)}, \bar{\underline{\mathbf{a}}}_{(t)})$
- Por ello, empíricamente se debe linealizar alrededor de estado observado con frecuencia.

Ejemplo

- Por ejemplo, con el péndulo invertido

$$\underline{\mathbf{s}}_{(t+1)} = f\left(\begin{bmatrix} x \\ \dot{x} \\ \theta \\ \dot{\theta} \end{bmatrix}\right)$$

si elegimos que el centro corresponde a $x = 0$, y queremos que la base esté estática ($\dot{x} = 0$), que el poste esté vertical ($\theta = 0$) y que no se esté cayendo ($\dot{\theta} = 0$), entonces querremos linealizar alrededor de $\bar{\underline{\mathbf{s}}}_{(t)} = \underline{\mathbf{0}}$.

Linealización general

- Considerando las acciones, la extensión con el jacobiano es directa:

$$\begin{aligned}\underline{\mathbf{s}}_{(t+1)} &= f(\underline{\mathbf{s}}_{(t)}, \underline{\mathbf{a}}_{(t)}) \\ &\approx f(\bar{\underline{\mathbf{s}}}_{(t)}, \bar{\underline{\mathbf{a}}}_{(t)}) + \mathbf{J}_f(\bar{\underline{\mathbf{s}}}_{(t)}, \bar{\underline{\mathbf{a}}}_{(t)}) \begin{bmatrix} \underline{\mathbf{s}}_{(t)} - \bar{\underline{\mathbf{s}}}_{(t)} \\ \underline{\mathbf{a}}_{(t)} - \bar{\underline{\mathbf{a}}}_{(t)} \end{bmatrix}\end{aligned}$$

  donde de nuevo $(\bar{\underline{\mathbf{s}}}_{(t)}, \bar{\underline{\mathbf{a}}}_{(t)})$ corresponden al centro de la serie, que es constante en la aproximación, seleccionado adecuadamente para la aplicación concreta.
- Con la linealización, ¿podemos reexpresar lo anterior como

$$\underline{\mathbf{s}}_{(t+1)} = \mathbf{A}\underline{\mathbf{s}}_{(t)} + \mathbf{B}\underline{\mathbf{a}}_{(t)}?$$

Lidiando con el intercepto

- Si expresamos la formula anterior como productos de matrices

$$\underline{\mathbf{s}}_{(t+1)} \approx \underbrace{f\left(\bar{\underline{\mathbf{s}}}_{(t)}, \bar{\underline{\mathbf{a}}}_{(t)}\right)}_{\underline{\mathbf{c}}} + \mathbf{J}_f\left(\bar{\underline{\mathbf{s}}}_{(t)}, \bar{\underline{\mathbf{a}}}_{(t)}\right) \begin{bmatrix} \underline{\mathbf{s}}_{(t)} - \bar{\underline{\mathbf{s}}}_{(t)} \\ \underline{\mathbf{a}}_{(t)} - \bar{\underline{\mathbf{a}}}_{(t)} \end{bmatrix}$$

$$= \mathbf{A}\underline{\mathbf{s}}_{(t)} + \mathbf{B}\underline{\mathbf{a}}_{(t)} + \underbrace{\underline{\mathbf{c}} - \mathbf{A}\bar{\underline{\mathbf{s}}}_{(t)} - \mathbf{B}\bar{\underline{\mathbf{a}}}_{(t)}}_{\bar{\underline{\mathbf{c}}}}$$

- De la teoría de señales y sistemas sabemos que un sistema con intercepto **no** es lineal (sino afín)
- Si $\bar{\underline{\mathbf{c}}} \neq \underline{\mathbf{0}}$ podemos extender el estado $\underline{\mathbf{s}}_{(t)}$ en una dimensión con un 1 y así considerar el intercepto implícitamente en la matriz $\mathbf{A}$.
- Por ejemplo: $\underline{\mathbf{s}} = \begin{bmatrix} 1 & x & \dot{x} & \theta & \dot{\theta} \end{bmatrix}^T$ y el intercepto estará en la primera columna de $\mathbf{A}$

Resumiendo el problema LQR (1)

- Un MDP de horizonte finito para LQR se especifica con la quíntupla: $\langle \mathcal{S}, \mathcal{A}, \{P_{sa}\}, T, R \rangle$ con $\mathcal{S} = \mathbb{R}^n$, $\mathcal{A} = \mathbb{R}^d$.
- Suponemos que para las probabilidades de transición se cumple

$$P_{\underline{\mathbf{sa}}}^{(t)} : \underline{\mathbf{s}}_{(t+1)} = \mathbf{A}_{(t)}\underline{\mathbf{s}}_{(t)} + \mathbf{B}_{(t)}\underline{\mathbf{a}}_{(t)} + \underline{\mathbf{w}}_{(t)}$$

- Para las recompensas usamos

$$R_{(t)}\left(\underline{\mathbf{s}}_{(t)}, \underline{\mathbf{a}}_{(t)}\right) = -\left(\underline{\mathbf{s}}_{(t)}^T \mathbf{U}_{(t)} \underline{\mathbf{s}}_{(t)} + \underline{\mathbf{a}}_{(t)}^T \mathbf{V}_{(t)} \underline{\mathbf{a}}_{(t)}\right)$$

con $\mathbf{U}_{(t)} \in \mathbb{R}^{n \times n}$ y $\mathbf{V}_{(t)} \in \mathbb{R}^{d \times d}$, ambas positivas semidefinidas.

Resumiendo el problema LQR (2)

- **A**, **B**, **U** y **V** son dadas en la especificación del problema LQR
- La meta es encontrar una política que maximice la recompensa en el tiempo de horizonte finito:

$$\text{máx} : R_{(0)}(s_{(0)},a_{(0)}) + R_{(1)}(s_{(1)},a_{(1)}) + \ldots + R_{(T)}(s_{(T)},a_{(T)})$$

Estrategia de solución

- La estrategia será primero encontrar $V^*_{(T)}(\underline{\mathbf{s}}_{(T)})$ y luego con programación dinámica (recursivamente) encontrar los valores en tiempos anteriores a T.
- Luego encontraremos la política usando $V^*_{(t)}(\underline{\mathbf{s}}_{(t)})$.

Encontrando la función de valor y la política

- El inicio de la recursión se encuentra con:

$$
\begin{aligned}
V^*_{(T)}(\underline{\mathbf{s}}_{(T)}) &= \max_{\underline{\mathbf{a}}_{(T)} \in \mathcal{A}} R_{(T)}(\underline{\mathbf{s}}_{(T)}, \underline{\mathbf{a}}_{(T)}) \\
&= \max_{\underline{\mathbf{a}}_{(T)} \in \mathcal{A}} \left(- \left(\underline{\mathbf{s}}^T_{(T)} \mathbf{U}_{(T)} \underline{\mathbf{s}}_{(T)} + \underline{\mathbf{a}}^T_{(T)} \mathbf{V}_{(T)} \underline{\mathbf{a}}_{(T)} \right) \right) \\
&= -\underline{\mathbf{s}}^T_{(T)} \mathbf{U}_{(T)} \underline{\mathbf{s}}_{(T)}
\end{aligned}
$$

  donde se ha considerado que con $\underline{\mathbf{a}}_{(T)} = \underline{\mathbf{0}}$ se obtiene el máximo.

- Entonces, la acción óptima de la política en T es

$$
\pi^*_{(T)} \left(\underline{\mathbf{s}}_{(T)} \right) = \arg \max_{\underline{\mathbf{a}}_{(T)}} R_{(T)} \left(\underline{\mathbf{s}}_{(T)}, \underline{\mathbf{a}}_{(T)} \right) = \underline{\mathbf{0}}
$$

Paso de programación dinámica

- El paso de programación dinámica que buscamos produce $V^*_{(t)}$ a partir de $V^*_{(t+1)}$, como lo hicimos antes.
- Para el caso de estados finitos teníamos

$$V^*_{(t)}(\underline{\mathbf{s}}_{(t)}) = \max_{\underline{\mathbf{a}}_{(t)}} \left(R_{(t)}(\underline{\mathbf{s}}_{(t)}, \underline{\mathbf{a}}_{(t)}) + \sum_{\underline{\mathbf{s}}'_{(t+1)}} P^{(t)}_{\underline{\mathbf{s}}_{(t)}\underline{\mathbf{a}}_{(t)}}(\underline{\mathbf{s}}'_{(t+1)}) V^*_{(t+1)}(\underline{\mathbf{s}}'_{(t+1)}) \right)$$

- Para el caso continuo, la suma se convierte en integral, pero podemos expresar directamente el resultado como el valor esperado:

$$V^*_{(t)}(\underline{\mathbf{s}}_{(t)}) = \max_{\underline{\mathbf{a}}_{(t)}} \left(R_{(t)}(\underline{\mathbf{s}}_{(t)}, \underline{\mathbf{a}}_{(t)}) + \mathrm{E}_{\underline{\mathbf{s}}_{(t+1)} \sim P_{\underline{\mathbf{s}}_{(t+1)}\underline{\mathbf{a}}_{(t)}}} \left[V^*_{(t+1)}(\underline{\mathbf{s}}_{(t+1)}) \right] \right)$$

- Por usar LQR, resulta que la expresión anterior se puede expresar de forma cuadrática.

Expresión cuadrática para función de valor

- Supongamos que

$$V^*_{(t+1)}(\underline{\mathbf{s}}_{(t+1)}) = \underline{\mathbf{s}}^T_{(t+1)} \mathbf{\Phi}_{(t+1)} \underline{\mathbf{s}}_{(t+1)} + \psi_{(t+1)}$$

  con $\mathbf{\Phi}_{(t+1)} \in \mathbb{R}^{n \times n}$, y $\psi_{(t+1)} \in \mathbb{R}$
- Se puede demostrar que en un paso de la programación dinámica anterior, entonces

$$V^*_{(t)}(\underline{\mathbf{s}}_{(t)}) = \underline{\mathbf{s}}^T_{(t)} \mathbf{\Phi}_{(t)} \underline{\mathbf{s}}_{(t)} + \psi_{(t)}$$

  para nueva matriz $\mathbf{\Phi}_{(t)}$ y valor $\psi_{(t)}$.

- Con el primer paso habíamos encontrado que

$$V^*_{(T)}(\underline{\mathbf{s}}_{(T)}) = -\underline{\mathbf{s}}^T_{(T)}\mathbf{U}_{(T)}\underline{\mathbf{s}}_{(T)}$$

de modo que $\mathbf{\Phi}_{(T)} = -\mathbf{U}_{(T)}$, $\psi_{(T)} = 0$ para entonces

$$V^*_{(T)}(\underline{\mathbf{s}}_{(T)}) = \underline{\mathbf{s}}^T_{(T)}\mathbf{\Phi}_{(T)}\underline{\mathbf{s}}_{(T)} + \psi_{(T)}$$

- Luego de varias derivaciones algebraicas se obtiene que

$$V^*_{(t)}(\underline{\mathbf{s}}_{(t)}) = \max_{\underline{\mathbf{a}}_{(t)}} \Bigg(-\underline{\mathbf{s}}^T_{(t)}\mathbf{U}_{(t)}\underline{\mathbf{s}}_{(t)} - \underline{\mathbf{a}}^T_{(t)}\mathbf{V}_{(t)}\underline{\mathbf{a}}_{(t)} + \mathrm{E}_{\underline{\mathbf{s}}_{(t+1)}\sim P_{\underline{\mathbf{s}}_{(t)}\underline{\mathbf{a}}_{(t)}}} \underbrace{\left[\underline{\mathbf{s}}^T_{(t+1)}\mathbf{\Phi}_{(t+1)}\underline{\mathbf{s}}_{(t+1)} + \psi_{(t+1)}\right]}_{V^*_{(t+1)}(\underline{\mathbf{s}}_{(t+1)})} \Bigg)$$

con $P_{\underline{\mathbf{s}}_{(t)}\underline{\mathbf{a}}_{(t)}} = \mathcal{N}(\mathbf{A}_{(t)}\underline{\mathbf{s}}_{(t)} + \mathbf{B}_{(t)}\underline{\mathbf{a}}_{(t)}, \mathbf{\Sigma}_{\underline{\mathbf{w}}})$.

- La expresión anterior es a su vez cuadrática.

Encontrando las acciones óptimas

- Queremos maximizar lo anterior, que es una función cuadrática.
- Derivamos e igualamos a cero, con lo que se obtiene

$$\underline{\mathbf{a}}_{(t)} = \underbrace{\left(\mathbf{B}_{(t)}^T \mathbf{\Phi}_{(t+1)} \mathbf{B}_{(t)} - \mathbf{V}_{(t)}\right)^{-1} \mathbf{B}_{(t)}^T \mathbf{\Phi}_{(t+1)} \mathbf{A}_{(t)}}_{\mathbf{L}_{(t)}} \underline{\mathbf{s}}_{(t)}$$

$$= \mathbf{L}_{(t)} \underline{\mathbf{s}}_{(t)}$$

- Lo anterior implica que la acción óptima es una función lineal del estado $\underline{\mathbf{s}}_{(t)}$
- Para calcular la política óptima tenemos entonces:

$$\pi^*_{(t)}(\underline{\mathbf{s}}_{(t)}) = \arg\max_{\underline{\mathbf{a}}_{(t)}} R_{(t)}(\underline{\mathbf{s}}_{(t)}, \underline{\mathbf{a}}_{(t)}) + \mathrm{E}_{\underline{\mathbf{s}}_{(t+1)}}\left[V^*_{(t+1)}(\underline{\mathbf{s}}_{(t+1)})\right]$$

$$= \mathbf{L}_{(t)} \underline{\mathbf{s}}_{(t)}$$

Ecuación de Riccati de tiempo discreto

- Si tomamos lo anterior y lo introducimos en la recursión de la programación dinámica, encontramos:

$$V^*_{(t)}(\underline{\mathbf{s}}_{(t)}) = \underline{\mathbf{s}}^T_{(t)} \mathbf{\Phi}_{(t)} \underline{\mathbf{s}}_{(t)} + \psi_{(t)}$$

con la ecuación de Riccati de tiempo discreto:

$$\begin{aligned}
\mathbf{\Phi}_{(t)} =& \mathbf{A}^T_{(t)} \left(\mathbf{\Phi}_{(t+1)} - \mathbf{\Phi}_{(t+1)} \mathbf{B}_{(t)} \left(\mathbf{B}^T_t \mathbf{\Phi}_{(t+1)} \mathbf{B}_t - \mathbf{V}_{(t)} \right)^{-1} \mathbf{B}^T_{(t)} \mathbf{\Phi}_{(t+1)} \right) \mathbf{A}_{(t)} \\
& - \mathbf{U}_{(t)} \\
\psi_{(t)} =& - \text{tr} \left[\mathbf{\Sigma}_{\underline{\mathbf{w}}} \mathbf{\Phi}_{(t+1)} \right] + \psi_{(t+1)}
\end{aligned}$$

- Observe la dependencia recursiva de los cálculos, que nos permite calcular la función de valor óptima a través de expresar $\mathbf{\Phi}_{(t)}$ en términos de $\mathbf{\Phi}_{(t+1)}$

Resumiendo...

- El algoritmo para encontrar la solución exacta de un LQR de horizonte finito se resume como sigue:
  1. Inicialice $\mathbf{\Phi}_T = -\mathbf{U}_{(T)}$, $\psi_{(T)} = 0$.
  2. Calcule recursivamente $\mathbf{\Phi}_{(t)}$, $\psi_{(t)}$ usando $\mathbf{\Phi}_{(t+1)}$, $\psi_{(t+1)}$ con las ecuaciones de Riccati para $t = T-1, T-2, \ldots, 0$.
  3. Calcule la matrix $\mathbf{L}_{(t)}$ usando $\mathbf{\Phi}_{(t+1)}$, $\psi_{(t+1)}$.
  4. Calcule la política $\pi^*(\underline{\mathbf{s}}_{(t)}) = \mathbf{L}_{(t)}\underline{\mathbf{s}}_{(t)}$.
- Observe que considerando la linealidad del sistema, y las suposición de que la recompensa es cuadrática, podemos obtener de forma exacta la función de valor óptima, a pesar de que los espacios de acción y de estado son contínuos.
- Este algoritmo escala con $\mathcal{O}(n^3)$ con n la dimensión del espacio de estados, que es mucho mejor que el crecimiento exponencial del método de discretización.

Efectos del ruido

- Nótese que la política óptima **no** depende de $\psi_{(t)}$ ni $\psi_{(t+1)}$.
- Como no depende de $\psi_{(t)}$, ¡el ruido no tiene efecto en la política óptima! pues $\mathbf{\Sigma}_{\underline{\mathbf{w}}}$ solo aparece en los $\psi_{(t)}$.
- Esta es una propiedad única para los sistemas LQR.
- Si cambiamos la suposición de linealidad del sistema o la forma cuadrática de la recompensa, ya no se cumple esto.
- Note que la función de valor **sí** depende de $\psi_{(t)}$ y por tanto sí es afectada por el ruido.

Resumen

Este documento ha sido elaborado con software libre incluyendo LaTeX, Beamer, GNUPlot, GNU/Octave, XFig, Inkscape, GNU-Make y Subversion en GNU/Linux