Aprendizaje reforzado en sistemas de control

Dr.-Ing. Miguel A. Solis

Track Infonor & Tecnologías Emergentes
INFONOR-CHILE

05 Septiembre de 2018

Contenido

Modelo del sistema

Versión lineal de una ecuación de diferencias:

$$y[k+n] + a_{n-1}y[k+n-1] + \ldots + a_0 y[k] = b_m u[k+m] + \ldots + b_0 u[k],$$

por medio del operador de adelanto se obtiene

$$\begin{aligned} x[k+1] &= Ax[k] + Bu[k], \\ y[k] &= Cx[k] + Du[k], \end{aligned}$$

Estabilidad

El análisis de estabilidad depende del criterio que se considere:

- Asintótica
- BIBO (Entrada acotada - Salida acotada)
- MSS (Estabilidad cuadrática media)

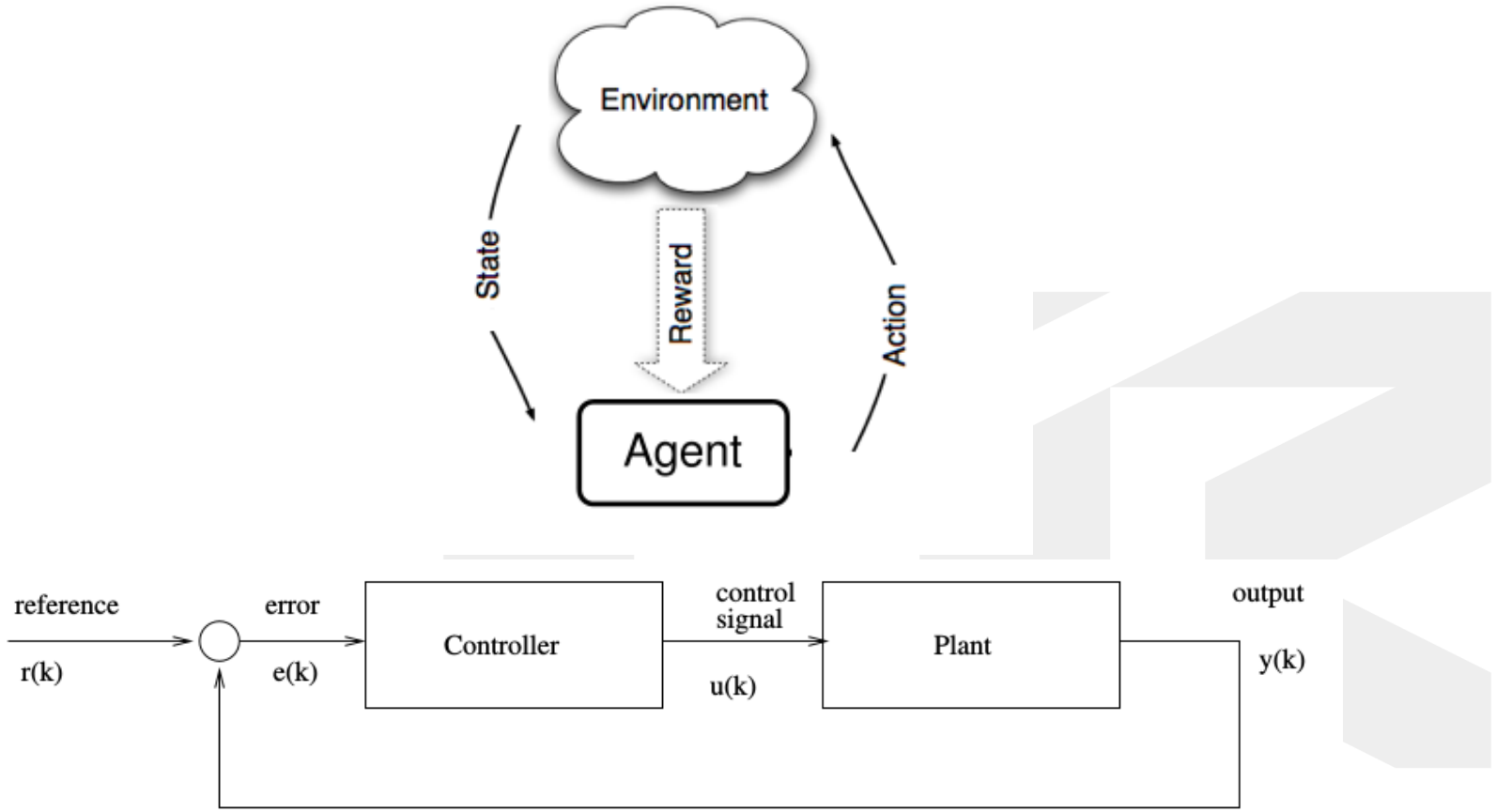
MSS

El sistema es estable en sentido cuadrático medio si y sólo si, para cualquier x_0 y u con v, existe $\mu_x \in \mathbb{R}^{n_x}$ y $M_x \in \mathbb{R}^{n_x \times n_x}$, $M_x \geq 0$, tal que

$$\lim_{k \to \infty} \xi\{x[k]\} = \mu_x,$$

$$\lim_{k \to \infty} \xi\{x[k]x[k]^\intercal\} = M_x.$$

Esquema de realimentación



Problema de aprendizaje reforzado

Se formula como un MDP, compuesto por la tupla $(\mathrm{X}, \mathrm{U}, \mathrm{T}, \mathrm{R})$ donde

- X: corresponde al conjunto de todos los posibles estados.
- U: denota el conjunto de acciones que el agente puede tomar.
- T: $\mathrm{X} \times \mathrm{U} \times \mathrm{X} \to [0, 1]$ es una función de transición de estado, que asigna una probabilidad de ser transferido desde el estado x al estado x' por medio de ejecutar u.
- R: $\mathrm{X} \times \mathrm{U} \to \mathbb{R}$ corresponde a una función (escalar) de recompensas.
- π: $\mathrm{X} \to \mathrm{U}$ es un mapeo de estados a acciones, describiendo la política (acciones a tomar en ciertos estados).

Función de valor

La calidad de cierta política π se cuantifica a través de la función de valor $V^{\pi}(x_k)$, definida como la recompensa acumulada (esperada) desde cierto estado x en el instante k:

$$V^{\pi}(x_k) = \xi\{\sum_{i=0}^{\infty} \gamma^{i} r_{k+i} | \pi\},$$

donde $\gamma \in [0, 1)$ corresponde al factor de descuento.

Política óptima

π^* se dice óptima, si satisface la siguiente expresión:

$$V^{\pi^*}(x) \geq V^{\pi}(x), \quad \forall x, \pi$$

Expresión de optimalidad de Bellman

La política óptima π^* satisface

$$V^*(x) = \max_{u \in \mathrm{U}} \sum_{x' \in \mathrm{X}} Pr\{(x, u, x')\} \left(R(x, u) + \gamma V^*(x')\right),$$

Policy Iteration

1: $\hat{\pi}_0(s) =$ acción aleatoria $\quad \forall s \in \mathcal{S}$
2: $\hat{V}_0(s) = 0 \quad \forall s \in \mathcal{S}$ ▷ arbitrariamente
3: evaluar_politica()
4: mejorar_politica()

Policy Iteration

evaluar_politica():

1: $k = 0$
2: **while** $\Delta > \epsilon$ **do** ▷ (para ϵ pequeño)
3: **for** $s \in \mathcal{S}$ **do**
4: $\hat{V}_{k+1}(s) = \text{máx}_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} Pr\{(s, a, s')\} (R(s, a) + \gamma V(s'))$
5: **end for**
6: $\Delta = \| \hat{V}_{k+1} - \hat{V}_k \|$
7: $k = k + 1$
8: **end while**

Policy Iteration

mejorar_politica() :

1: $politica\_estable = true$
2: $k = 0$
3: **for** $s \in \mathcal{S}$ **do**
4: $\quad \hat{\pi}_{k+1}(s) = argmax_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} Pr\{(s, a, s')\} (R(s, a) + \gamma V(s'))$
5: $\quad$ **if** $\hat{\pi}_k(s) \neq \hat{\pi}_{k+1}(s)$ **then**
6: $\quad\quad politica\_estable = false$
7: $\quad$ **end if**
8: **end for**
9: **if** *politica_estable* **then**
10: $\quad$ **return** $\hat{\pi}_k$
11: **else**
12: $\quad$ evaluar_politica()
13: **end if**

Función de valor óptima

Pueden existir múltiples políticas óptimas, pero la función de valor óptima es única:

$$V^*(x[k]) = \max_{u \in \mathcal{U}} \sum_{x[k+1]} Pr\{(x[k], u[k], x[k+1])\} \left(R(x[k], u[k]) + \gamma V^*(x[k+1])\right),$$

Función de valor estado-acción óptima

La función de valor óptima V^* se relaciona con la función de valor estado-acción óptima Q^* como

$$V^*(x[k]) = \max_{u \in \mathcal{U}} Q^*(x[k], u[k]).$$

TD(0)

1: $\hat{V}(x[0]) = 0 \qquad \forall x \in \mathcal{X}$
2: **for** $k = 1, 2, ..., n$ **do**
3: **Observar** $x[k], R(x[k], u[k]), x[k+1]$
4: $\hat{V}(x[k]) = \hat{V}(x[k]) + \alpha \left(R(x[k], u[k]) + \gamma \hat{V}(x[k+1]) - \hat{V}(x[k]) \right)$
5: **end for**
6: **retornar** $\hat{V}$

Q-learning

1: $\hat{Q}(x[0], u[0]) = 0 \qquad \forall x \in \mathcal{X}, \quad u \in \mathcal{U}$
2: **for** $k = 1, 2, ..., n$ **do**
3: **Observar** $x[k], u[k], R(x[k], u[k]), x[k+1]$
4: $\hat{Q}(x[k], u[k]) = \hat{Q}(x[k], u[k]) + \alpha \left(R(x[k], u[k]) + \gamma \, \text{máx}_{u[k+1]} \hat{Q}(x[k+1], u[k+1]) - \hat{Q}(x[k], u[k]) \right)$
5: **end for**
6: **retornar** $\hat{Q}$

Entonces, suponiendo $\hat{Q} = Q^*$, la acción óptima para cada estado se puede computar fácilmente mediante una operación simple de maximización

$$\pi^*(x[k]) = arg \max_{u[k]} Q^*(x[k], u[k]),$$

Representación en espacio de estados

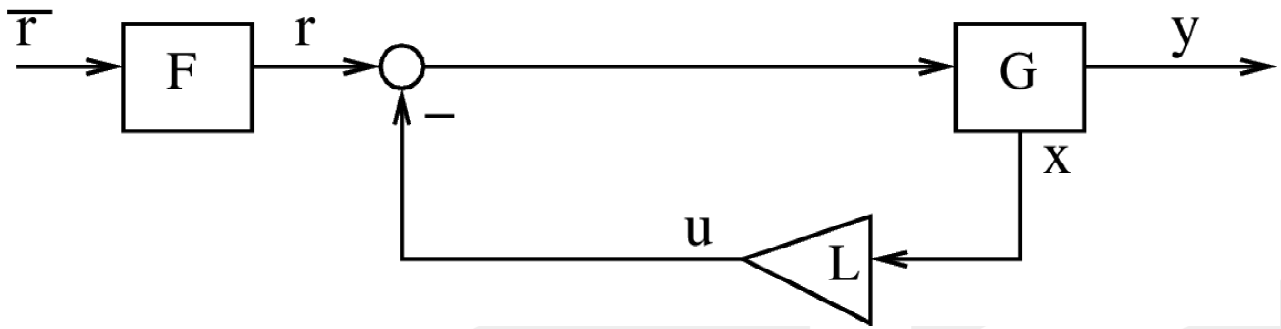
Para modelos de tiempo discreto, la representación en espacio de estados viene dada por

$$x[k+1] = Ax[k] + Bu[k] + v[k],$$
$$y[k] = Cx[k] + w[k],$$

Suposiciones

- El estado inicial $x[0] = x_0$ es una variable aleatoria de segundo orden con media $\mu_{x,0}$ y varianza $P_{x,0} \geq 0$.
- La entrada u se caracteriza en términos del estado x.
- El ruido v es una secuencia de ruido blanco de media cero no correlacionado con x_0, y varianza $P_v \geq 0$.

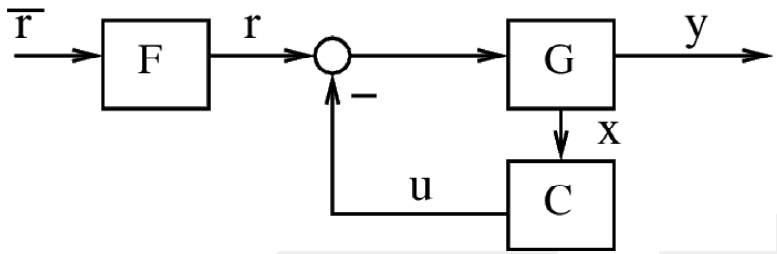
Realimentación de estado



- para seguimiento de trayectoria:

$$F = -\left(C\left(A - BL - I\right)^{-1} B\right)^{-1}.$$

Modelo



$$x_c[k+1] = A_c x_c[k] + B_c x[k],$$
$$u[k] = C_c x_c[k] + D_c x[k],$$

Prefiltro

Entonces, el prefiltro viene dado por

$$F = -\left(C \left(A - I - BD_c + BC_c \left(A_c - I \right)^{-1} B_c \right)^{-1} B \right)^{-1}$$

MSS

Un controlador MSS estabilizará (en sentido cuadrático medio) la planta si y solo si los autovalores de $\bar{A}$ están dentro del circulo unitario, donde $\bar{A}$ es una matriz por bloques dada por

$$\bar{A} = \begin{bmatrix} A - BD_c & -BC_c \\ B_c & A_c \end{bmatrix}.$$

MSS con observador

Introduciendo un observador K:

$$\bar{\bar{A}} = \begin{bmatrix} A - KC & 0_{nx \times nx_c} & 0_{nx \times nx} \\ -B_c & A_c & B_c \\ BD_c & -BC_c & A - BD_c \end{bmatrix}.$$

Simulación

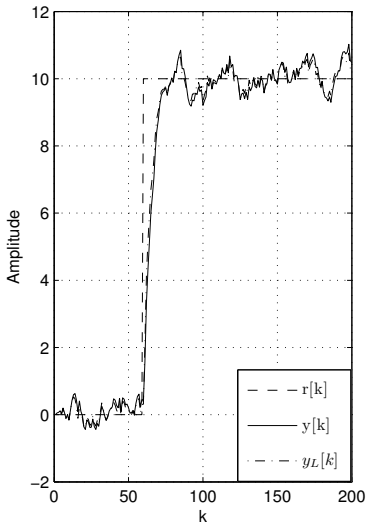
- Considere el sistema:

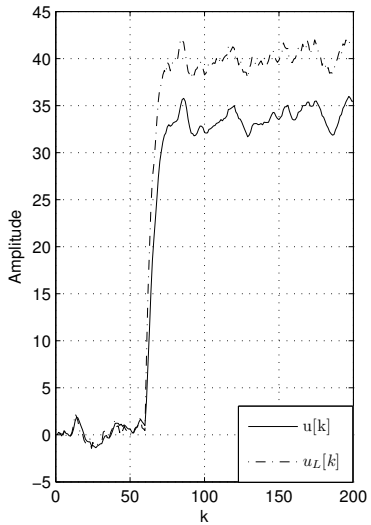
$$A = \begin{bmatrix} 0{,}5 & 0 \\ 0{,}7 & 1{,}2 \end{bmatrix},$$
$$B = \begin{bmatrix} 0 \\ 0{,}1 \end{bmatrix},$$
$$C = \begin{bmatrix} 1 & 1 \end{bmatrix}.$$

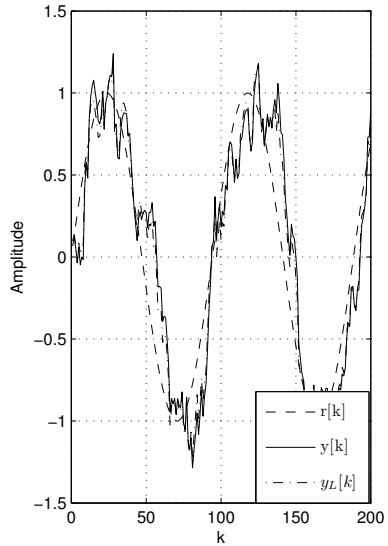
- y el controlador:

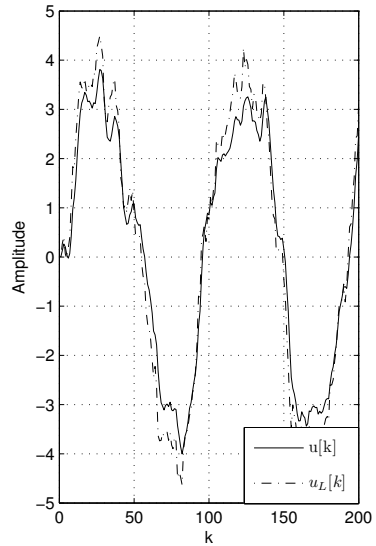
$$A_c = 0{,}4 \qquad B_c = \begin{bmatrix} 1 & -1{,}52 \end{bmatrix},$$
$$C_c = -0{,}5 \qquad D_c = \begin{bmatrix} 0{,}3 & 2{,}1 \end{bmatrix}.$$

Simulación (mediciones completas)



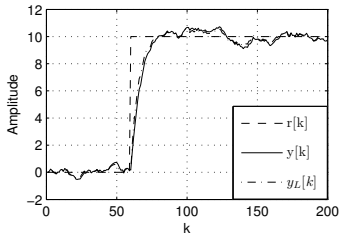


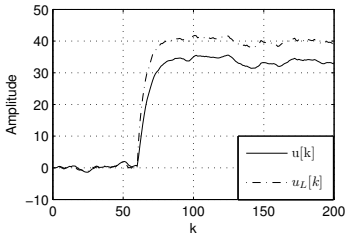


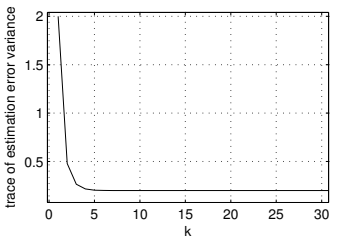


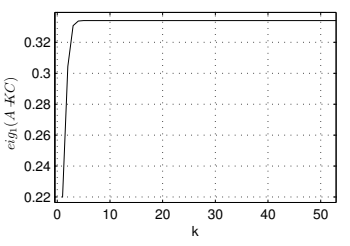
centro de innovación y robótica

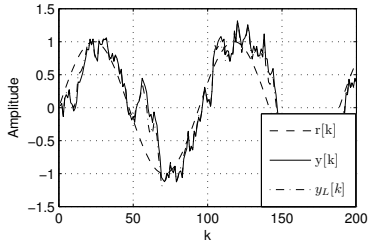
Simulación (con observador)

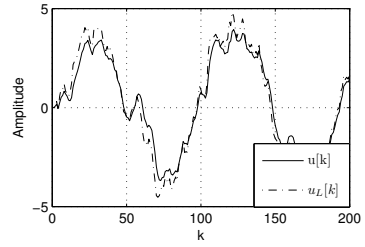


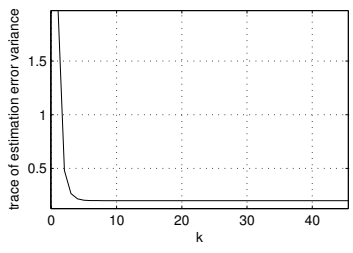


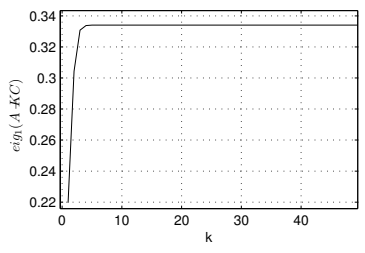












LQT

El desempeño es medido por

$$J[k] = \xi\{\sum_{i=k}^{\infty} \gamma^{i-k} \left(z^{\intercal}[i] Q_a z[i] + u^{\intercal}[i] R u[i]\right)\},$$

con

$$z[k] = \begin{bmatrix} (r[k] - y[k]) \\ x_c[k] \end{bmatrix}, \quad Q_a = \begin{bmatrix} Q_1 & 0 \\ 0 & Q_2 \end{bmatrix},$$

donde ambos Q_1 y Q_2 son matrices definidas positivas, penalizando el error de control y evitando que la dinámica del mismo controlador crezca sin cota respectivamente.

Formulación

Para que el problema de LQT parezca más un problema de aprendizaje reforzado, defina la función de valor $V(X[k])$ como

$$V(X[k]) = -J[k],$$

donde $X[k]$ corresponde al vector de estado aumentado que contiene al estado interno del proceso a controlar.

Función de valor

Entonces, la función de valor puede ser escrita como

$$V\left(X[k]\right) = \xi\{\sum_{i=k}^{\infty} \gamma^{i-k}\left(tr\left(P_w Q_1\right) + X^{\intercal}[i]\bar{Q}X[i] + u^{\intercal}[i]Ru[i]\right)\},$$

con Q_1 como se definió anteriormente, y $\bar{Q}$ dado por

$$\bar{Q} = \begin{bmatrix} C^{\intercal}Q_1 C & 0 & -C^{\intercal}Q_1 \\ 0 & Q_2 & 0 \\ -Q_1 C & 0 & Q_1 \end{bmatrix}.$$

Función de valor óptima

Entonces, suponiendo que la función de valor óptima es cuadrática en términos del vector de estado aumentado, i.e.,

$$V^*\left(X[k]\right) = X^{\mathsf{T}}[k]PX[k] + g[k],$$

para alguna matriz simétrica y estacionaria $P > 0$, y $g[k]$ tal que

$$g[k+1] = \left(\frac{1}{\gamma}\right) g[k] + tr\left(\frac{1}{\gamma} P_w Q_1 + P_v P_{11}\right),$$

Los parámetros (T_r, C_c, D_c) vendrán dados por

$$
\begin{aligned}
T_r &= \gamma Z^{-1} B^\intercal P_{13} F, \\
C_c &= -\gamma Z^{-1} B^\intercal P_{12} A_c, \\
D_c &= -\gamma Z^{-1} M,
\end{aligned}
$$

con

$$
\begin{aligned}
Z &= \left(R + \gamma B^\intercal P_{11} B\right), \\
M &= \left(A^\intercal P_{11} B + B_c^\intercal P_{21} B\right),
\end{aligned}
$$

Ecuaciones algebraicas de Riccatti

$$
\begin{aligned}
P_{11} &= C^\intercal Q_1 C + \gamma \left(A^\intercal P_{11} A + B_c^\intercal P_{21} A \right. \\
&\qquad \left. + A^\intercal P_{12} B_c + B_c^\intercal P_{12} B_c \right) - \gamma^2 M Z^{-1} M^\intercal, \\
P_{12} &= \gamma \left(A^\intercal P_{12} A_c + B_c^\intercal P_{22} A_c \right) - \gamma^2 M Z^{-1} B^\intercal P_{12} A_c, \\
P_{22} &= Q_2 + \gamma A_c^\intercal P_{22} A_c - \gamma^2 A_c^\intercal P_{21} B Z^{-1} B^\intercal P_{12} A_c, \\
P_{13} &= -C^\intercal Q_1 + \gamma \left(A^\intercal P_{13} F + B_c^\intercal P_{23} F \right) \\
&\qquad - \gamma^2 M Z^{-1} B^\intercal P_{12} F, \\
P_{23} &= \gamma A_c^\intercal P_{23} F - \gamma^2 A_c^\intercal P_{21} B Z^{-1} B^\intercal P_{13} F,
\end{aligned}
$$

donde $P_{ij} = P_{ji}^\intercal$, como consecuencia de la simetría de P.

PI con dinámica conocida

1: **(Evaluación de política)** Resuelva para cada variable del lado izquierdo de la ARE, en base a los datos de la iteración anterior, es decir:

2: **(Mejora de política)** Actualizar los parámetros A_c y B_c con estas nuevas soluciones, tal que

$$\bar{\lambda}\left(\begin{bmatrix} \bar{A}_{11}^{(l)} & \bar{A}_{12}^{(l)} \\ B_c^{(l)} & A_c^{(l)} \end{bmatrix}\right) < 1,$$

donde

$$\begin{aligned}
\bar{A}_{11}^{(l)} &= A - \gamma B Z^{-1,(l)} \left(A^\intercal P_{11}^{(l)} B + B_c^{(l)} P_{21}^{(l)} B\right), \\
\bar{A}_{12}^{(l)} &= -\gamma B Z^{-1,(l)} B^\intercal P_{12}^{(l)} A_c^{(l)}, \\
Z^{(l)} &= \left(R + \gamma B^\intercal P_{11}^{(l)} B\right).
\end{aligned}$$

PI con dinámica conocida

Entonces, el resto de los parámetros se actualizan como

$$
\begin{aligned}
T_r^{(l)} &= \gamma Z^{-1,(l)} B^\intercal P_{13}^{(l)} F, \\
C_c^{(l)} &= -\gamma Z^{-1,(l)} B^\intercal P_{12}^{(l)} A_c^{(l)}, \\
D_c^{(l)} &= -\gamma Z^{-1,(l)} \left(A^\intercal P_{11}^{(l)} B + B_c^{\intercal,(l)} P_{21}^{(l)} B \right).
\end{aligned}
$$

Función Q para LQT

Considere la función Q para LQT dada por

$$Q\left(X[k], u[k]\right) = X^{\intercal}\bar{Q}X[k] + u^{\intercal}[k]Ru[k] + tr\left(P_w Q_1\right) + \gamma\left(X^{\intercal}[k+1]PX[k+1] + g[k+1]\right),$$

Función Q para LQT

que es equivalente a

$$Q\left(X[k], u[k]\right) = \begin{bmatrix} X[k] \\ u[k] \end{bmatrix}^{\mathsf{T}} H \begin{bmatrix} X[k] \\ u[k] \end{bmatrix} + tr\left(P_w Q_1 + \gamma P_v P_{11}\right) + \gamma g[k+1],$$

con H simétrica dada por

$$H = \begin{bmatrix} H_{xx} & H_{xu} \\ H_{ux} & H_{uu} \end{bmatrix},$$

Función Q para LQT

donde

$$H_{uu} = R + \gamma B^\intercal P_{11} B,$$

$$H_{xu} = \gamma \begin{bmatrix} B^\intercal P_{11} A + B^\intercal P_{12} B_c \\ B^\intercal P_{12} A_c \\ B^\intercal P_{13} F \end{bmatrix},$$

Ley de control

Entonces, encontrando las raíces de la primera derivada, se puede obtener la ley de control en términos de H dada por

$$u[k] = -H_{uu}^{-1} H_{ux} X[k].$$

Función Q para LQT

Sea $S[k]$ dado por

$$S[k] = \begin{bmatrix} 1 \\ X[k] \\ u[k] \end{bmatrix},$$

con la función Q para LQT equivalente a

$$Q\left(X[k], u[k]\right) = S^{\mathsf{T}}[k]\bar{H}S[k],$$

Ecuación de optimalidad de Bellman

con $\bar{H}$ dado por

$$\bar{H} = \begin{bmatrix} tr\left(P_w Q_1 + \gamma P_v P_{11}\right) + \gamma g[k+1] & 0 & 0 \\ 0 & H_{xx} & H_{xu} \\ 0 & H_{ux} & H_{uu} \end{bmatrix},$$

llevando a

$$S^{\intercal}[k]\bar{H}S[k] = X^{\intercal}[k]\bar{Q}X[k] + u^{\intercal}[k]Ru[k] + \gamma S^{\intercal}[k+1]\bar{H}S[k+1].$$

PI con dinámica desconocida

1: **(Evaluación de política)** En base a las observaciones de $S[k]$, $S[k+1]$ y $\left(X^{\intercal}[k]\bar{Q}X[k] + u^{\intercal,(l)}[k]Ru^{(l)}[k]\right)$ para la l-ésima iteración, usar regresión para encontrar $\bar{H}$,

$$S^{\intercal}[k]\bar{H}^{(l+1)}S[k] = X^{\intercal}[k]\bar{Q}X[k] + u^{\intercal,(l)}[k]Ru^{(l)}[k] + \gamma S^{\intercal}[k+1]\bar{H}^{(l+1)}S[k+1].$$

2: **(Mejora de política)** Actualizar la ley de control

$$u^{(l+1)}[k] = -\left(H_{uu}^{-1}\right)^{(l+1)} H_{ux}^{(l+1)} X[k].$$

Simulación

Considere el sistema lineal dado por

$$x[k+1] = \begin{bmatrix} 0{,}5 & 0 \\ 0{,}7 & 1{,}2 \end{bmatrix} x[k] + \begin{bmatrix} 0 \\ 0{,}1 \end{bmatrix} u[k] + v[k],$$

$$y[k] = \begin{bmatrix} 1 & 1 \end{bmatrix} x[k] + w[k],$$

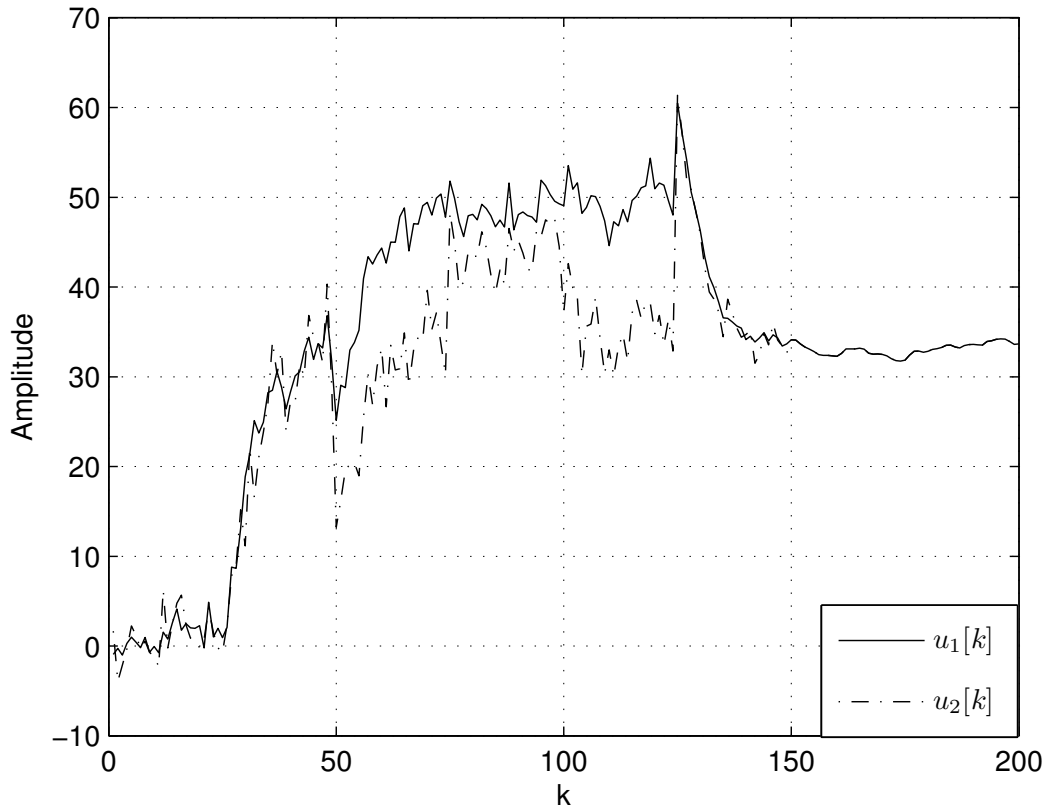
Simulación

Se fija arbitrariamente un modelo para el generador F de la señal de referencia

$$F[k] = \begin{cases} 0 & k < 30 \\ 10 & k \geq 30 \end{cases}$$

y $r[0] = 1$. Además, se establecen los pesos de penalización para el índice de desempeño en $Q_1 = 5$, $Q_2 = 5$ and $R = 1$, y el factor de descuento $\gamma = 0{,}8$.

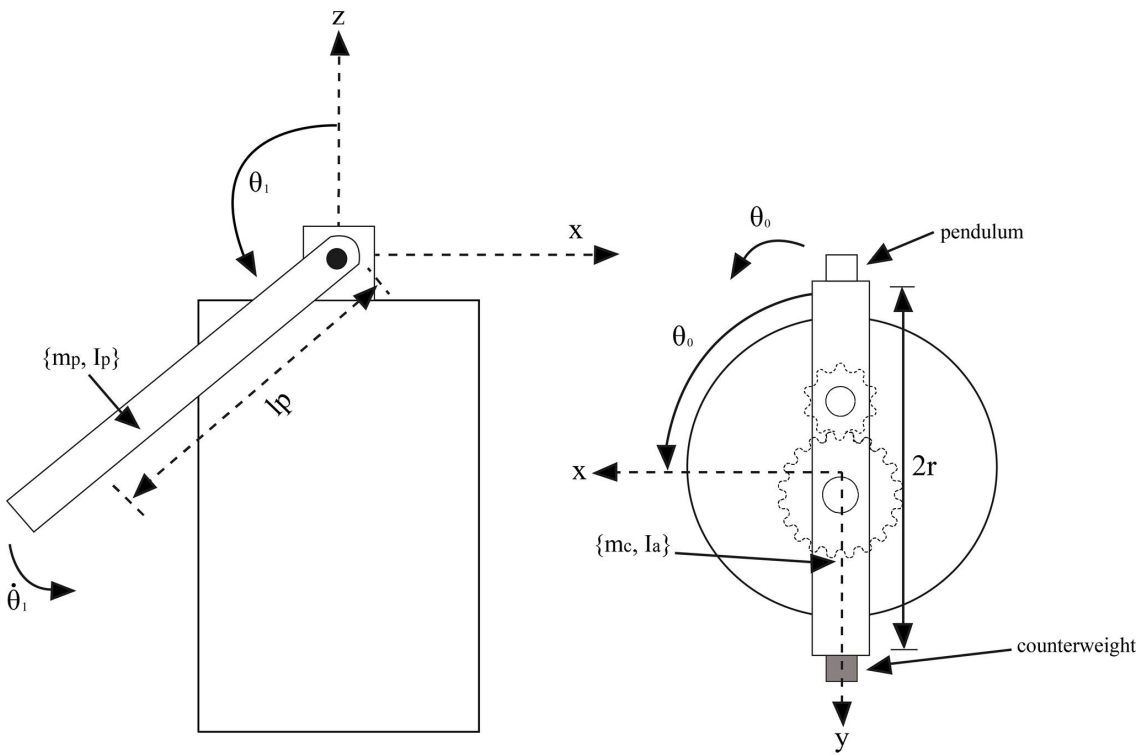
Simulación



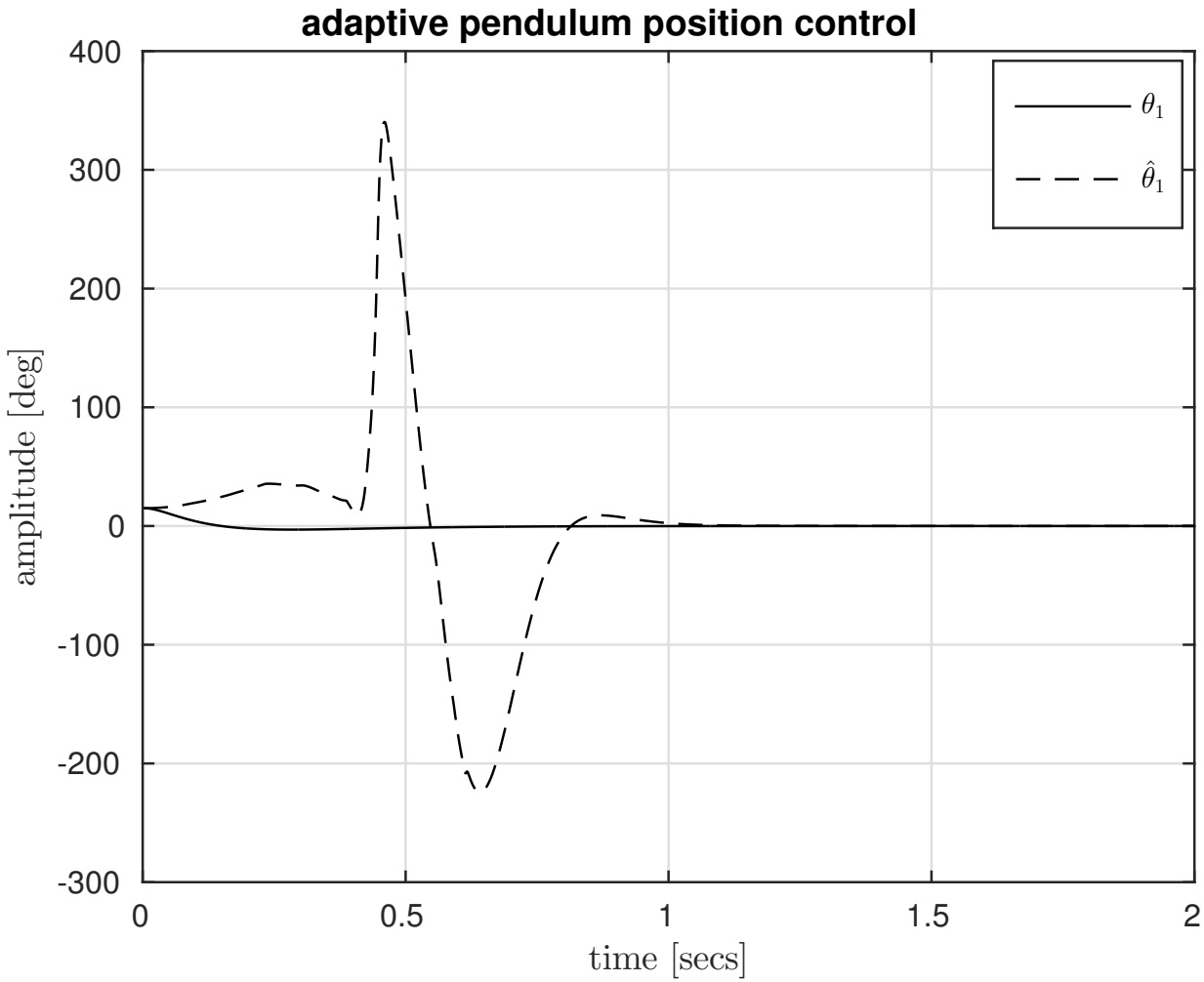
u_1 y u_2: PI con dinámica conocida y desconocida respectivamente.

centro de innovación y robótica

Péndulo de Furuta



Enfoque adaptivo



Por simulación:

Solución de ARE

Solución analítica de la ARE respectiva:

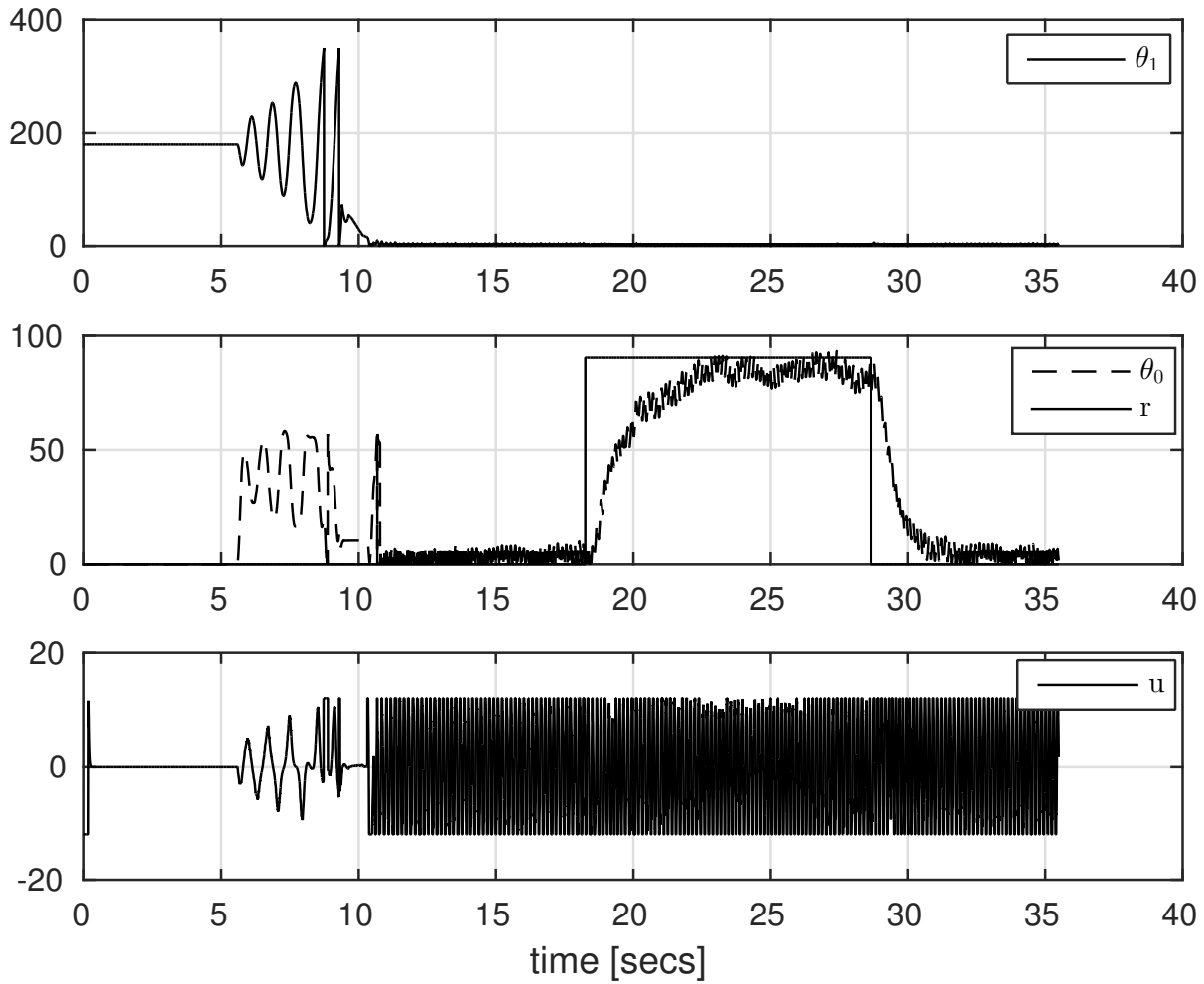
$$\begin{bmatrix} ,0046 & ,0013 & -,0152 & -,0019 & 0 \\ \cdot & ,0103 & -,1321 & -,0165 & -,0003 \\ \cdot & \cdot & 1,7318 & ,2139 & ,0036 \\ \cdot & \cdot & \cdot & ,0268 & ,0005 \\ \cdot & \cdot & \cdot & \cdot & 0 \end{bmatrix} \cdot 10^6,$$

con la correspondiente obtenida a través de aprendizaje

$$\begin{bmatrix} 0 & ,0001 & -,0014 & -,0002 & 0 \\ 0 & -,0999 & 1,2757 & ,1634 & ,0028 \\ -,0001 & ,3407 & -4,3501 & -,5572 & -,0097 \\ 0 & ,1281 & -1,6356 & -,2095 & -,0037 \\ 0 & 0 & -,0003 & 0 & 0 \end{bmatrix} \cdot 10^{10}.$$

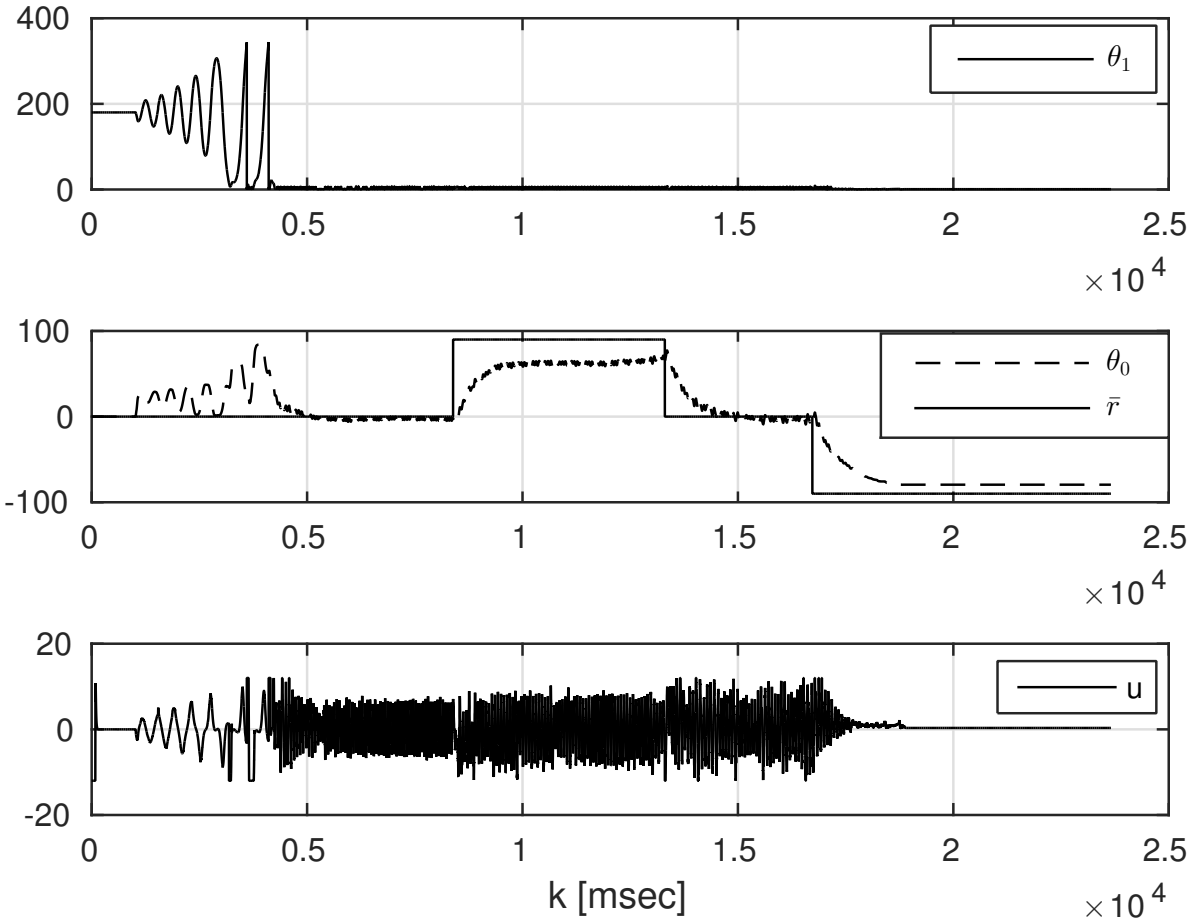
Control conmutado

Levantamiento y realimentación (clásica) de estado:



Control conmutado

Levantamiento y realimentación dinámica de estado:



Conclusiones

- Problema de control formulado como MDP
- Controlador con realimentación de estado agrega más parámetros, pero no afectan en un enfoque libre de modelos
- Enfoque adaptivo es útil cuando se conoce el modelo, pero más aún con dinámica desconocida

Algunas referencias

- R. Sutton and A. Barto **Introduction to reinforcement learning**, 1998.
- D. Bertsekas **Dynamic programming and optimal control**, 1995.
- B. Kiumarsi, F.L.Lewis, H. Modares, A. Karimpour and M-B. Naghibi-Sistani **Reinforcement Q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics**, Automatica, 2014.
- M.A. Solis, M. Olivares and H. Allende. **Stabilizing Dynamic State Feedback Controller Synthesis: A Reinforcement Learning Approach**, Studies in Informatics and Control, 2016.

Preguntas