三、 若某单位反馈控制系统的前向传递函数为 $\frac{2K}{s(2s+1)}$，

(1) 试求系统的传递函数，并将其转换为标准形式。

(2) 若 $K=1/4$，试求系统的阶跃响应。此时系统是否稳定？

(3) 试求 $K=1/4$ 时系统的最大超调量、峰值时间与调节时间。

(4) 若系统输出无振荡，试求 K 的取值范围。

(5) 若 $K=4$，试求系统在单位斜坡输入下的稳态误差。

三、(1) 前向传递函数 $\Phi(s)=\frac{2k}{s(2s+1)}$

∵单位反馈

∴反馈函数 $H(s)=1$

传递函数 $G(s)=\frac{\Phi(s)}{1+\Phi(s)\times H(s)}=\frac{\Phi(s)}{1+\Phi(s)}=\frac{\frac{2k}{s(2s+1)}}{1+\frac{2k}{s(2s+1)}}=\frac{2k}{2s^2+s+2k}$

(2) $K=\frac{1}{4}$，$G(s)=\frac{C(s)}{R(s)}=\frac{\frac{1}{2}}{\frac{1}{2}s^2+s+\frac{1}{2}}=\frac{1}{s^2+2s+1}=\frac{1}{(s+1)^2}$

∵阶跃响应

∴ $R(s)=\frac{1}{s}$

∴ $C(s)=G(s)R(s)=\frac{1}{s(s+1)^2}=\frac{1}{s}-\frac{1}{s+1}-\frac{1}{(s+1)^2}$

通过拉氏反变换，

得阶跃响应 $c(t)=1-e^{-t}-te^{-t}\ (t\geq 0)$

判断稳定性：$s^2+2s+1=0$

作劳斯表：

$$\begin{array}{lll} s^2 & 1 & 1 \\ s^1 & 2 & \\ s^0 & 1 & \end{array}$$

第一列系数全为正，因此系统稳定。

(3) $K=\frac{1}{4}$，$G(s)=\frac{1}{4s^2+2s+1}=\frac{\frac{1}{4}}{s^2+\frac{1}{2}s+\frac{1}{4}}=\frac{\omega_n^2}{s^2+2\zeta\omega_n s+\omega_n^2}$

比较系数得 $\omega_n=\frac{1}{2}$，$\zeta=\frac{1}{2}$

峰值时间 $t_p=\frac{\pi}{\omega_d}=\frac{\pi}{\omega_n\sqrt{1-\zeta^2}}=\frac{\pi}{\frac{1}{2}\times\sqrt{1-\frac{1}{4}}}=7.26s$

最大超调量 $M_p=e^{-\frac{\zeta\pi}{\sqrt{1-\zeta^2}}}=e^{-\frac{\frac{1}{2}\pi}{\sqrt{1-\frac{1}{4}}}}=56.14\%$

调节时间 $t_s=\frac{3}{\zeta\omega_n}=\frac{3}{\frac{1}{2}\times\frac{1}{2}}=12s\ (\pm 5\%)$

$t_s=\frac{4}{\zeta\omega_n}=\frac{4}{\frac{1}{2}\times\frac{1}{2}}=16s\ (\pm 2\%)$

(4) $G(s)=\frac{2k}{2s^2+s+2k}=\frac{k}{s^2+\frac{s}{2}+k}=\frac{\omega_n^2}{s^2+2\zeta\omega_n s+\omega_n^2}$

比较系数得 $\begin{cases} k=\omega_n^2 \\ \frac{1}{2}=2\zeta\omega_n \end{cases} \Rightarrow \zeta=\frac{1}{4\sqrt{k}}$

系统输出无振荡，即 $\zeta\geq 1$，即 $\frac{1}{4\sqrt{k}}\geq 1 \Rightarrow 0\leq k\leq\frac{1}{16}$

(5) $K=4$，$G(s)=\frac{8}{2s^2+s+8}$

∵单位斜坡输入

∴ $K_v=\lim_{s\to 0}sG(s)=\lim_{s\to 0}\frac{8s}{2s^2+s+8}$

稳态误差 $e_{ss}=\frac{1}{K_v}=\lim_{s\to 0}\frac{2s^2+s+8}{8s}=\lim_{s\to 0}\left(\frac{s}{4}+\frac{1}{8}+\frac{1}{s}\right)=\infty$

六、 已知某两输入系统的状态方程如下

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 3 \end{bmatrix} u$$

(1) 试判断该系统的能控性。

(2) 试给出状态反馈向量 $K = [k_1, k_2]$ 使得闭环极点配置在-1 和-3。

六、$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 3 \end{bmatrix} u$

(1) $A = \begin{bmatrix} 3 & 1 \\ -1 & 1 \end{bmatrix}$, $b = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$

$U_c = [b \;\; Ab] = \begin{bmatrix} 1 & 6 \\ 3 & 2 \end{bmatrix}$

秩 $r_c = 2$

∴ 系统完全能控

(2) $K = [k_1, k_2]$，闭环极点配置在 -1 和 -3.

经K引入状态反馈后的系统矩阵为

$A - bK = \begin{bmatrix} 3-k_1 & 1-k_2 \\ -1-3k_1 & 1-3k_2 \end{bmatrix}$

其特征多项式为

$|sI - (A-bK)| = s^2 + (k_1+3k_2-4)s + 2k_1 - 10k_2 + 4$

由期望的闭环极点给出的特征多项式为

$(s+1)(s+3) = s^2+4s+3$

比较两式得 $\begin{cases} k_1+3k_2-4=4 \\ 2k_1-10k_2+4=3 \end{cases}$

解得 $k_1 = \frac{77}{16}$, $k_2 = \frac{17}{16}$

∴ 状态反馈向量 $K = [\frac{77}{16}, \frac{17}{16}]$.

八、 Q 学习

(1) 表格型 Q 学习算法的流程是什么样的?

(2) $Q(s,a)$ 的更新公式是什么?

(3) 寻找最佳策略的策略迭代方法与值迭代方法分别是什么?

八、(1) 表格型 Q 学习算法流程:

对所有的 $s\in S$ 和 $a\in A(s)$,初始化 $Q(s,a)$ 和 $Model(s,a)$

无限循环:

(a) $S \leftarrow$ 当前(非终止)状态

(b) $A \leftarrow \varepsilon$-贪心(S,Q)

(c) 采取动作 A;观察产生的收益 R 以及状态 S'

(d) $Q(S,A) \leftarrow Q(S,A) + \alpha[R+\gamma \max_a Q(S',a) - Q(S,A)]$

(e) $Model(S,A) \leftarrow R, S'$ (假设环境是确定的)

(f) 重复 n 次循环:

$S \leftarrow$ 随机选择之前观察到的状态

$A \leftarrow$ 随机选择之前在状态 S 下采取过的动作 A

$R, S' \leftarrow Model(S,A)$

$Q(S,A) \leftarrow Q(S,A) + \alpha[R+\gamma \max_a Q(S',a) - Q(S,A)]$

(2) $Q(s,a)$ 的更新公式为:

$Q(s,a) \leftarrow Q(s,a) + \alpha[R+\gamma \max_a Q(s',a) - Q(s,a)]$

(3) 策略迭代方法:从任意一个状态价值函数开始,依据给定的策略,结合贝尔曼期望方程、状态转移概率和奖励同步迭代更新状态价值函数。

$$V_{k+1}(s) = \sum_{a\in A} \pi(a|s) \sum_{s',r} p(s',r|s,a)[r+\gamma V_k(s')]$$

值迭代方法:只计算第一次价值函数 $V_*(s)$ 的 ~~状~~ 策略迭代方法。

$$V_{k+1}(s) = \max_a \sum_{s',r} p(s',r|s,a)[r+\gamma V_k(s')]$$

十、 蒙特卡洛策略梯度方法

(1) 试给出蒙特卡洛策略梯度方法 (REINFORCE) 的算法流程。

(2) 试给出策略梯度优化的目标函数，并推导策略梯度公式。

(3) 什么是策略？什么是回报？

十.(1) 蒙特卡洛策略梯度方法的算法流程

输入：一个可导的参数化策略 $\pi(a|s,\theta)$

算法参数：步长 $\alpha>0$

初始化策略参数 $\theta\in R^{d'}$

无限循环 (对于每一幕)：

根据 $\pi(\cdot|\cdot,\theta)$，生成一幕序列 $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$

对于幕的每一步循环，$t=0,1,\ldots,T-1$：

$G \leftarrow \sum_{k=t+1}^{T}\gamma^{k-t-1}R_k$

$\theta \leftarrow \theta + \alpha\gamma^t G\nabla\ln\pi(A_t|S_t,\theta)$

(2) 目标函数：$G_t = \sum_{k=t+1}^{T}\gamma^{k-t-1}R_k$

策略梯度公式推导：

由策略梯度定理可知：$\nabla J(\theta)\propto\sum_s\mu(s)\sum_a q_\pi(s,a)\nabla\pi(a|s,\theta)$

$$
\begin{aligned}
\nabla J(\theta) &\propto \sum_s\mu(s)\sum_a q_\pi(s,a)\nabla\pi(a|s,\theta)\\
&= E_\pi\left[\sum_a q_\pi(S_t,a)\nabla\pi(a|S_t,\theta)\right]\\
&= E_\pi\left[\sum_a \pi(a|S_t,\theta)q_\pi(S_t,a)\frac{\nabla\pi(a|S_t,\theta)}{\pi(a|S_t,\theta)}\right]\\
&= E_\pi\left[q_\pi(S_t,A_t)\frac{\nabla\pi(A_t|S_t,\theta)}{\pi(A_t|S_t,\theta)}\right]\\
&= E_\pi\left[G_t\frac{\nabla\pi(A_t|S_t,\theta)}{\pi(A_t|S_t,\theta)}\right]
\end{aligned}
$$

$$\theta_{t+1} = \theta_t + \alpha G_t\frac{\nabla\pi(A_t|S_t,\theta_t)}{\pi(A_t|S_t,\theta_t)}$$

(3) 策略：从状态到每个动作的选择概率之间的映射，用 π 来表示。

回报：通常指期望回报，表示时刻 t 后接收的收益总和，用 G_t 来表示。

考虑到折扣率 γ，G_t 可由下式表示：

$$G_t = R_{t+1}+\gamma R_{t+2}+\gamma^2R_{t+3}+\cdots=\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}$$