控制与决策
Control and Decision

基于非策略Q-learning的欺骗攻击下未知线性离散系统最优跟踪控制

宋星星, 储昭碧

引用本文:
宋星星, 储昭碧. 基于非策略Q-learning的欺骗攻击下未知线性离散系统最优跟踪控制[J]. 控制与决策, 2025, 40(5): 1641-1650.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2024.0830

基于非策略 Q-learning 的欺骗攻击下未知线性离散系统最优跟踪控制

宋星星，储昭碧†

(合肥工业大学 电气与自动化工程学院，合肥 230000)

摘　要: 针对多重欺骗攻击下动力学信息未知的线性离散系统, 提出一种非策略 Q-learning 算法解决系统的最优跟踪控制问题. 首先, 考虑加入一个权重矩阵建立控制器通信信道遭受多重欺骗攻击的输入模型, 并结合参考命令生成器构建增广跟踪系统. 在线性二次跟踪框架内将系统的最优跟踪控制表达为欺骗攻击与控制输入同时参与的零和博弈问题. 其次, 设计一种基于状态数据的非策略 Q-learning 算法学习系统最优跟踪控制增益, 解决应用中控制增益不能按照给定要求更新的问题, 并证明在满足持续激励条件的探测噪声下该算法的求解不存在偏差. 同时考虑系统状态不可测的情况, 设计基于输出数据的非策略 Q-learning 算法. 最后, 通过对 F-16 飞机自动驾驶仪的跟踪控制仿真, 验证所设计非策略 Q-learning 算法的有效性以及对探测噪声影响的无偏性.

关键词: 欺骗攻击；最优跟踪；非策略 Q-learning；零和博弈

中图分类号: TP273　　**文献标志码:** A

DOI: 10.13195/j.kzyjc.2024.0830

引用格式: 宋星星, 储昭碧. 基于非策略 Q-learning 的欺骗攻击下未知线性离散系统最优跟踪控制 [J]. 控制与决策, 2025, 40(5): 1641-1650.

Based on off-policy Q-learning: Optimal tracking control for unknown linear discrete-time systems under deception attacks

SONG Xing-xing，*CHU Zhao-bi*†

(College of Electrical and Automation Engineering，Hefei University of Technology，Hefei 230000，China)

Abstract: An off-policy Q-learning algorithm is proposed to solve the optimal tracking control problem for the linear discrete-time system with unknown dynamics information under multiple deception attack. Firstly, a weight matrix is added to establish the input model of multiple deception attacks on the controller communication channel, and an augmented tracking system is constructed with a reference command generator. In the framework of linear quadratic tracking, the optimal tracking control of the system is expressed as a zero-sum game problem between deception attacks and control inputs. Then, an off-policy Q-learning algorithm based on state data is designed to learn the optimal tracking control gain of the system, which solves the problem that the control gain is difficult to update according to the given requirements in applications. It is proved that the solution of the algorithm has no deviation under the probe noise satisfying the persistence of excitation condition. At the same time, considering the situation that the system state cannot be measured, an off-policy Q-learning algorithm based on output data is designed. Finally, through the tracking control simulation of F-16 aircraft autopilot, the effectiveness of the designed off-policy Q-learning algorithm and the unbiasedness effect on detection noise are verified.

Keywords: deception attacks；optimal tracking；off-policy Q-learning；zero-sum game

0　引　言

在自动驾驶汽车、飞行器控制、多智能体等领域中常需要实现最优跟踪控制 (OTC), 使系统在最优条件下强制系统的状态或输出以最小的误差跟踪参考轨迹[1-3]. 线性系统的 OTC 可以由线性二次跟踪器 (LQT) 实现, 通过最小化与跟踪误差及控制输入成本相关的二次性能指标寻找最优跟踪控制器[4-5].

近年来, 网络攻击的存在对工业系统运行的稳定性造成了不可估计的危害, 其中欺骗攻击被认为是最常见和最隐蔽的攻击[6]. 为了解决控制器与执行

收稿日期: 2024-07-11；**录用日期:** 2024-09-25.
基金项目: 安徽省科技重大专项项目 (202103a05020001).
责任编委: 刘向杰.
†通信作者. E-mail: zbchu@hfut.edu.cn.

器间通信网络在欺骗攻击下的系统的 OTC 问题, 考虑将上述情况看作控制器与欺骗攻击同时参与的零和博弈问题. 线性系统的零和博弈需要求解博弈代数 Riccati 方程 (GARE), 当无法建立准确的系统模型时, GARE 求解困难并且无法保证系统的闭环稳定性以及控制器的最优性[7-8]. 因此, 有学者提出了一系列非基于模型的控制方案, Kong 等[9] 利用神经网络逼近n轴刚性机械臂的未知动力学, 结合反步法设计了自适应神经网络控制策略. 但是上述文献中的方法仍然存在建模误差, 得到的控制增益影响系统控制效果.

强化学习 (RL) 利用沿系统轨迹测量的数据对未知系统动力学等进行补偿, 通过判断累积奖励学习最优控制策略[10-12]. 然而实际系统的全状态数据难以收集,导致基于状态反馈的 RL 技术的应用受到限制[13]. 因此, Chen 等[14] 直接利用系统的输入输出数据设计最优控制器, 这种方法无需估计器观察系统状态, 结果不会受到状态估计误差的影响[15]. RL 技术分为策略与非策略, 基于策略的 RL要求待评估的策略应用于系统收集数据, 而非策略的 RL 中用于生成数据的行为策略与待评估的目标策略无关[16-17]. 基于策略的 RL 存在以下缺点: 首先是学习过程中策略需按要求更新, 这可能干扰系统运行并在实际应用中难以实现[18]; 其次, 满足持续性激励 (PE) 条件的探测噪声可能会影响算法的收敛性, 产生优化问题的错误解[19].

基于上述分析, 本文设计一种非策略 Q-learning 算法解决多重欺骗攻击下模型未知的线性离散系统的 OTC 问题. 本文贡献如下: 1) 加入一个权重矩阵建立多重欺骗攻击的输入模型, 在 LQT 框架内将系统的 OTC 描述为欺骗攻击与控制器同时参与的零和博弈问题; 2) 设计状态数据驱动的非策略 Q-learning 算法实现 OTC, 解决了实际系统动力学未知并且欺骗攻击难以按照给定形式更新的问题, 该算法在满足 PE 条件的探测噪声下的求解不存在偏差; 3) 设计输出数据驱动的非策略 Q-learning 算法解决系统状态数据不可测的问题.

符号说明: Φ^{T}为矩阵的转置, Φ^{-1}为矩 阵的逆, I为单位阵, $\mathbb{R}^b$为b维的实向量空间, $\mathbb{R}^{b\times d}$为$b\times d$维矩阵, $\mathrm{diag}\{\cdot\}$为对角阵, $\mathrm{vec}(\cdot)$为矩阵的列构成向量, $\otimes$为克罗内克积运算, $||\cdot||$ 为取欧式范数.

1 问题描述与分析

1.1 欺骗攻击下的跟踪系统描述

考虑如下线性离散系统:

$$x_{k+1}=\mathcal{A}x_k+\mathcal{B}u_k,\ y_k=\mathcal{C}x_k. \tag{1}$$

其中: $x_k\in\mathbb{R}^b$, $u_k\in\mathbb{R}^m$, $y_k\in\mathbb{R}^d$分别为系统状态、控制输入和输出; $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$为适当维数的矩阵.

假设 1 $(\mathcal{A},\mathcal{B})$是可控的, $(\mathcal{A},\mathcal{C})$是可观测的.

期望的系统输出轨迹$r_k\in\mathbb{R}^d$满足

$$r_{k+1}=\mathcal{F}r_k, \tag{2}$$

其中$\mathcal{F}$为已知的参数矩阵.

考虑系统控制器的通信网络遭受多重欺骗攻击的情况, 控制输入被修改为

$$u_k^a=u_k+\sum_{p=1}^{q}\mathcal{D}^p a_k^p. \tag{3}$$

其中: $a_k^p\in\mathbb{R}^m$, $p\in\mathbb{O}=\{1,2,\dots,q\}$为攻击者$p$在时间步长$k$处注入的虚假数据, q为攻击者个数. 矩阵$\mathcal{D}^p$满足$\mathcal{D}^p=\mathrm{diag}\{\xi_1^p,\xi_2^p,\dots,\xi_m^p\}$, $\xi_i^p\in(0,1)$, $i=1,2,\dots,m$, $\xi_i^p=1$表示第i个通道被第p个攻击者攻击, $\xi_i^p=0$表示相应通道不会被攻击. 在本文中假设参数$\mathcal{D}^p$已知, 则根据被攻击者修改后的控制输入 (3) 可知遭受网络攻击的系统 (1) 的状态变为

$$x_{k+1}=\mathcal{A}x_k+\mathcal{B}u_k+\sum_{p=1}^{q}\mathcal{B}\mathcal{D}^p a_k^p. \tag{4}$$

注 1 在假设 1 的基础上, $(\mathcal{A},\mathcal{B}\mathcal{D}^p)$也是可控的.

基于系统 (4) 以及跟踪期望值r_k构造增广系统

$$X_{k+1}=\begin{bmatrix}\mathcal{A}&0\\0&\mathcal{F}\end{bmatrix}\begin{bmatrix}x_k\\r_k\end{bmatrix}+\begin{bmatrix}\mathcal{B}\\0\end{bmatrix}u_k+\sum_{p=1}^{q}\begin{bmatrix}\mathcal{B}\mathcal{D}^p\\0\end{bmatrix}a_k^p\equiv$$

$$AX_k+B_u u_k+\sum_{p=1}^{q}B_a^p a_k^p, \tag{5}$$

其中$X_k=[x_k^{\mathrm{T}},\ r_k^{\mathrm{T}}]^{\mathrm{T}}\in\mathbb{R}^{b+d}$.

假设 2 控制输入和欺骗攻击的虚假数据为

$$\begin{aligned}u_k&=\mathcal{K}_1x_k+\mathcal{K}_2r_k=\mathcal{K}X_k,\\ a_k^p&=\mathcal{L}_{p,1}x_k+\mathcal{L}_{p,2}r_k=\mathcal{L}_pX_k,\ p\in\mathbb{O},\end{aligned} \tag{6}$$

其中$\mathcal{K}$和$\mathcal{L}^p$为待确定的反馈增益.

1.2 欺骗攻击下系统的 OTC 问题分析

欺骗攻击下系统 OTC 设计需要考虑以下问题.

问题 1 对于系统 (5), 控制器输入u在时间步k上的目标是最小化代价函数

$$J(e_k,u_k)=\sum_{k=0}^{\infty}\gamma^k(e_k^{\mathrm{T}}\mathcal{Q}e_k+u_k^{\mathrm{T}}\mathcal{R}u_k). \tag{7}$$

最优控制策略$u_k^*=\underset{u_k}{\mathrm{argmin}}J(e_k,u_k)$.

问题 2 对于系统 (5), 欺骗攻击者p在时间步k上的目标是最大化代价函数

$$J(e_k,a_k^p)=\sum_{k=0}^{\infty}\gamma^k(e_k^{\mathrm{T}}\mathcal{Q}e_k-\epsilon^p(a_k^p)^{\mathrm{T}}a_k^p). \tag{8}$$

第p个最优攻击策略为$(a_k^p)^* = \underset{a_k^p}{\operatorname{argmax}} J(e_k, a_k^p)$, 权重参数$\epsilon^p > 0$. 权重矩阵$\mathcal{Q} > 0$, $\mathcal{R} > 0$, 折现因子$0 < \gamma \leqslant 1$, 仅当$\mathcal{F}$为 Hurwitz 时, $\gamma = 1$.

定义欺骗攻击下系统无限视界代价函数并写为依赖于X_k的二次形式

$$\begin{aligned} &J(e_k, u_k, a_k^1, \dots, a_k^p) = \\ &\sum_{k=0}^{\infty} \gamma^k \Big(e_k^{\mathrm{T}} \mathcal{Q} e_k + u_k^{\mathrm{T}} \mathcal{R} u_k - \sum_{p=1}^{q} \epsilon^p (a_k^p)^{\mathrm{T}} a_k^p \Big) = \\ &X_k^{\mathrm{T}} \mathcal{Q}^a X_k + u_k^{\mathrm{T}} \mathcal{R} u_k - \sum_{p=1}^{q} \epsilon^p (a_k^p)^{\mathrm{T}} a_k^p + \\ &\gamma J(X_{k+1}, u_{k+1}, a_{k+1}^1, \dots, a_{k+1}^p) = X_k^{\mathrm{T}} P X_k. \end{aligned} \tag{9}$$

其中: $\mathcal{Q}^a = [\mathcal{C}, -I]^{\mathrm{T}} \mathcal{Q} [\mathcal{C}, -I]$, $P \in \mathbb{R}^{(b+d)\times(b+d)}$.

在假设 2 的基础上, 将上述问题转化为系统多个参与者的零和博弈问题, 最优跟踪问题转化为求解满足下列等式的$u_k, a_k^1, \dots, a_k^p$:

$$\begin{aligned} &J^*(e_k, u_k, a_k^1, \dots, a_k^p) = \\ &\min_{u_k} \max_{a_k} \sum_{k=0}^{\infty} \gamma^k \Big(e_k^{\mathrm{T}} \mathcal{Q} e_k + \\ &u_k^{\mathrm{T}} \mathcal{R} u_k - \sum_{p=1}^{q} \epsilon^p (a_k^p)^{\mathrm{T}} a_k^p \Big). \end{aligned} \tag{10}$$

则受欺骗攻击系统的 LQT Bellman 方程为

$$\begin{aligned} X_k^{\mathrm{T}} P X_k = X_k^{\mathrm{T}} \mathcal{Q}^a X_k + u_k^{\mathrm{T}} \mathcal{R} u_k - \sum_{p=1}^{q} \epsilon^p (a_k^p)^{\mathrm{T}} a_k^p + \\ \gamma X_{k+1}^{\mathrm{T}} P X_{k+1}. \end{aligned} \tag{11}$$

根据式 (11) 定义与欺骗攻击$a_k^1, a_k^2, \dots, a_k^p$相关的 LQT Hamiltonian 方程

$$\begin{aligned} &\breve{H}(X_k, u_k, a_k^1, \dots, a_k^p) = \\ &X_k^{\mathrm{T}} \mathcal{Q}^a X_k + u_k^{\mathrm{T}} \mathcal{R} u_k - \\ &\sum_{p=1}^{q} \epsilon^p (a_k^p)^{\mathrm{T}} a_k^p + \gamma X_{k+1}^{\mathrm{T}} P X_{k+1} - X_k^{\mathrm{T}} P X_k. \end{aligned} \tag{12}$$

令 $\hat{a}_k = [(a_k^1)^{\mathrm{T}}, (a_k^2)^{\mathrm{T}}, \dots, (a_k^q)^{\mathrm{T}}]^{\mathrm{T}}$, $\hat{B}_a = [B_a^1, B_a^2, \dots, B_a^q]$, $\hat{\epsilon} = \operatorname{diag}\{\epsilon^1, \epsilon^2, \dots, \epsilon^q\}$. 分别求解$\partial \breve{H}(X_k, u_k, \hat{a}_k)/\partial u_k = 0$, $\partial \breve{H}(X_k, u_k, \hat{a}_k)/\partial \hat{a}_k = 0$可得最优跟踪控制策略$u_k^*$和最差攻击策略$\hat{a}_k^*$为

$$u_k^* = \mathcal{K}^* X_k, \ \hat{a}_k^* = \mathcal{L}^* X_k. \tag{13}$$

控制增益分别为

$$\begin{aligned} &\mathcal{K}^* = (\mathcal{R} + \gamma B_u^{\mathrm{T}} P B_u - \Upsilon B_u)^{-1} (\Upsilon A - \gamma B_u^{\mathrm{T}} P A), \\ &\mathcal{L}^* = -(\Omega B_a - \hat{\epsilon})^{-1} (\Omega A - \gamma B_a^{\mathrm{T}} P A). \end{aligned} \tag{14}$$

其中: $\Upsilon = \gamma^2 B_u^{\mathrm{T}} P \hat{B}_a (\gamma \hat{B}_a^{\mathrm{T}} P \hat{B}_a - \hat{\epsilon})^{-1} \hat{B}_a^{\mathrm{T}} P$; $\Omega^p = \gamma (B_a^p)^{\mathrm{T}} P (I - \gamma B_u (\mathcal{R} + \gamma B_u^{\mathrm{T}} P B_u)^{-1} B_u^{\mathrm{T}} P)$, 且$\Omega = [(\Omega^1)^{\mathrm{T}}, (\Omega^2)^{\mathrm{T}}, \dots, (\Omega^q)^{\mathrm{T}}]^{\mathrm{T}}$; $\mathcal{L}^* = [(\mathcal{L}_1^*)^{\mathrm{T}}, (\mathcal{L}_2^*)^{\mathrm{T}}, \dots, (\mathcal{L}_q^*)^{\mathrm{T}}]^{\mathrm{T}}$. 得到关于$P$满足的 GARE 为

$$\begin{aligned} &P = \mathcal{Q}^a - \gamma^2 \begin{bmatrix} B_u^{\mathrm{T}} P A \\ \hat{B}_a^{\mathrm{T}} P A \end{bmatrix}^{\mathrm{T}} \chi \begin{bmatrix} B_u^{\mathrm{T}} P A \\ \hat{B}_a^{\mathrm{T}} P A \end{bmatrix} + \gamma A P A, \\ &\chi = \begin{bmatrix} \mathcal{R} + \gamma B_u^{\mathrm{T}} P B_u & \gamma B_u^{\mathrm{T}} P \hat{B}_a \\ \gamma \hat{B}_a^{\mathrm{T}} P B_u & \gamma \hat{B}_a^{\mathrm{T}} P \hat{B}_a - \hat{\epsilon} \otimes I \end{bmatrix}^{-1}. \end{aligned} \tag{15}$$

定理 1　具有折扣因子的 GARE(15) 具有唯一解的条件是$(\mathcal{A}, \mathcal{B})$、$(\mathcal{A}, \mathcal{B}\mathcal{D}^p)$可控, $(\mathcal{A}, \mathcal{C})$可观, 并且折扣因子$\gamma$使$\gamma^{\frac{1}{2}}\mathcal{F}$在单位圆内有特征值.

证明　令$\tilde{A} = \gamma^{\frac{1}{2}} A$, $\tilde{B}_u = \gamma^{\frac{1}{2}} B_u$, $\tilde{B}_a = \gamma^{\frac{1}{2}} \hat{B}_a$, 则式 (15) 可改写为关于系统$(\tilde{A}, \tilde{B}_u, \tilde{B}_a)$的标准 GARE

$$\begin{aligned} &P = \\ &\mathcal{Q}^a + \tilde{A} P \tilde{A} - \begin{bmatrix} \tilde{B}_u^{\mathrm{T}} P \tilde{A} \\ \tilde{B}_a^{\mathrm{T}} P \tilde{A} \end{bmatrix}^{\mathrm{T}} \times \\ &\begin{bmatrix} \mathcal{R} + \tilde{B}_u^{\mathrm{T}} P B_u & \tilde{B}_u^{\mathrm{T}} P \tilde{B}_a \\ \tilde{B}_a^{\mathrm{T}} P \tilde{B}_u & \tilde{B}_a^{\mathrm{T}} P \tilde{B}_a - \hat{\epsilon} \otimes I \end{bmatrix}^{-1} \begin{bmatrix} \tilde{B}_u^{\mathrm{T}} P \tilde{A} \\ \tilde{B}_a^{\mathrm{T}} P \tilde{A} \end{bmatrix}. \end{aligned} \tag{16}$$

标准 GARE 唯一解的存在条件是系统参数矩阵$(\tilde{A}, \tilde{B}_u)$、$(\tilde{A}, \tilde{B}_a)$是稳定的, 并且$(\tilde{A}, \sqrt{\mathcal{Q}^a})$可观. 这就需要$(\gamma^{\frac{1}{2}}\mathcal{A}, \gamma^{\frac{1}{2}}\mathcal{B})$、$(\gamma^{\frac{1}{2}}\mathcal{A}, \gamma^{\frac{1}{2}}\mathcal{B}\mathcal{D}^p)$稳定, 且$\gamma^{\frac{1}{2}}\mathcal{F}$也稳定. 选择使$\gamma^{\frac{1}{2}}\mathcal{F}$在单位圆内有特征值的折扣因子$\gamma$, 在$(\mathcal{A}, \mathcal{B})$、$(\mathcal{A}, \mathcal{B}\mathcal{D}^p)$可控的条件下, $(\tilde{A}, \tilde{B}_u)$、$(\tilde{A}, \tilde{B}_a)$是可稳定的. 并且在$(\mathcal{A}, \mathcal{C})$可观的条件下, 可知$(\tilde{A}, \sqrt{\mathcal{Q}^a})$可观. 最终保证了折现 GARE(15) 存在唯一的正定解P. □

2　模型未知的 OTC RL 算法设计

2.1　基于状态数据模型未知系统的 OTC

根据 LQT Bellman 方程 (11) 定义关于X_k, u_k, $a_k^1, a_k^2, \dots, a_k^p$的无限视界 Q 函数并化为二次形式

$$\begin{aligned} &Q(X_k, u_k, \hat{a}_k) = \\ &X_k^{\mathrm{T}} \mathcal{Q}^a X_k + u_k^{\mathrm{T}} \mathcal{R} u_k - \\ &\sum_{p=1}^{q} \epsilon^p (a_k^p)^{\mathrm{T}} a_k^p + \gamma X_{k+1}^{\mathrm{T}} P X_{k+1} = \\ &\bar{X}_k^{\mathrm{T}} \begin{bmatrix} \mathcal{S}_{XX} & \mathcal{S}_{Xu} & \mathcal{S}_{Xa} \\ \mathcal{S}_{uX} & \mathcal{S}_{uu} & \mathcal{S}_{ua} \\ \mathcal{S}_{aX} & \mathcal{S}_{au} & \mathcal{S}_{aa} \end{bmatrix} \bar{X}_k \equiv \bar{X}_k^{\mathrm{T}} \mathcal{S} \bar{X}_k. \end{aligned} \tag{17}$$

其中: $\bar{X} = [X_k^{\mathrm{T}}, u_k^{\mathrm{T}}, \hat{a}_k^{\mathrm{T}}]^{\mathrm{T}}$; 核矩阵$\mathcal{S} = \mathcal{S}^{\mathrm{T}} \in \mathbb{R}^{l\times l}$, $l = b + d + m(q+1)$. 各矩阵块分别为

$$\mathcal{S}_{XX}=\mathcal{Q}^a+\gamma A^{\mathrm{T}}PA.$$

$$\mathcal{S}_{Xu}=\gamma A^{\mathrm{T}}P\hat{B}_u=\mathcal{S}_{uX}^{\mathrm{T}}.$$

$$\mathcal{S}_{Xa}=\gamma A^{\mathrm{T}}P\hat{B}_a=[\mathcal{S}_{Xa^1},\mathcal{S}_{Xa^2},\dots,\mathcal{S}_{Xa^q}].$$

$$\mathcal{S}_{uu}=\mathcal{R}+\gamma B_u^{\mathrm{T}}PB_u.$$

$$\mathcal{S}_{Xa^p}=\gamma A^{\mathrm{T}}PB_a^p.$$

$$\mathcal{S}_{ua^p}=\gamma B_u^{\mathrm{T}}PB_a^p,\ p\in\mathbb{O}.$$

$$\mathcal{S}_{ua}=\gamma B_u^{\mathrm{T}}P\hat{B}_a=[\mathcal{S}_{ua^1},\mathcal{S}_{ua^2},\dots,\mathcal{S}_{ua^q}].$$

$$\mathcal{S}_{aa}=\gamma\hat{B}_a^{\mathrm{T}}P\hat{B}_a-\hat{\epsilon}\otimes I=$$

$$\begin{bmatrix}\mathcal{S}_{a^1a^1} & \mathcal{S}_{a^1a^2} & \dots & \mathcal{S}_{a^1a^q},\\ \mathcal{S}_{a^2a^1} & \mathcal{S}_{a^2a^2} & \dots & \mathcal{S}_{a^2a^q},\\ \vdots & \vdots & \ddots & \vdots\\ \mathcal{S}_{a^qa^1} & \mathcal{S}_{a^qa^2} & \dots & \mathcal{S}_{a^qa^q}\end{bmatrix}.$$

$$\mathcal{S}_{a^ia^j}=\begin{cases}\gamma(B_a^i)^{\mathrm{T}}PB_a^j,\ i\neq j;\\ \gamma(B_a^i)^{\mathrm{T}}PB_a^i-\epsilon^i\otimes I,\ i=j.\end{cases}\tag{18}$$

根据系统稳定性条件, 控制器与欺骗攻击的最优策略可以分别通过$\partial Q(X_k,u_k,\hat{a}_k)/\partial u_k=0$和$\partial Q(X_k,u_k,\hat{a}_k)/\partial\hat{a}_k=0$求得, 即

$$u_k^*=(\mathcal{S}_{uu}^*-\mathcal{S}_{ua}^*(\mathcal{S}_{aa}^*)^{-1}\mathcal{S}_{au}^*)^{-1}\times(\mathcal{S}_{ua}^*(\mathcal{S}_{aa}^*)^{-1}\mathcal{S}_{aX}^*-\mathcal{S}_{uX}^*)X_k,\tag{19}$$

$$\hat{a}_k^*=(\mathcal{S}_{aa}^*-\mathcal{S}_{au}^*(\mathcal{S}_{uu}^*)^{-1}\mathcal{S}_{ua}^*)^{-1}\times(\mathcal{S}_{au}^*(\mathcal{S}_{uu}^*)^{-1}\mathcal{S}_{uX}^*-\mathcal{S}_{aX}^*)X_k,\tag{20}$$

其中$\hat{a}_k^*=[(a_k^{1,*})^{\mathrm{T}},(a_k^{2,*})^{\mathrm{T}},\dots,(a_k^{q,*})^{\mathrm{T}}]^{\mathrm{T}}$.

2.2 基于状态数据的策略 *Q*-learning 算法

由式 (17) 可知 Q 函数满足 Bellman 方程

$$Q(X_k,u_k,\hat{a}_k)=X_k^{\mathrm{T}}\mathcal{Q}^aX_k-\sum_{p=1}^{q}\epsilon^p(a_k^p)^{\mathrm{T}}a_k^p+u_k^{\mathrm{T}}\mathcal{R}u_k+\gamma Q(X_{k+1},u_{k+1},\hat{a}_{k+1}).\tag{21}$$

将式 (17) 代入 (21) 可得 Q 函数 Bellman 方程为

$$\bar{X}_k^{\mathrm{T}}\mathcal{S}\bar{X}_k=X_k^{\mathrm{T}}\mathcal{Q}^aX_k+u_k^{\mathrm{T}}\mathcal{R}u_k-\sum_{p=1}^{q}\epsilon^p(a_k^p)^{\mathrm{T}}a_k^p+\gamma\bar{X}_{k+1}^{\mathrm{T}}\mathcal{S}\bar{X}_{k+1}.\tag{22}$$

将方程 (22) 的参数线性化与参数矩阵$\mathcal{S}$分离可得

$$(\bar{X}_k^{\mathrm{T}}\otimes\bar{X}_k^{\mathrm{T}}-\gamma\bar{X}_{k+1}^{\mathrm{T}}\otimes\bar{X}_{k+1}^{\mathrm{T}})\mathrm{vec}(\mathcal{S}^{j+1})=(\tilde{\varphi}_k^j)^{\mathrm{T}}\tilde{\mathcal{S}}^{j+1}=\tilde{v}_k^j.\tag{23}$$

其中: $\tilde{v}_k^j=X_k^{\mathrm{T}}\mathcal{Q}^aX_k+u_k^{\mathrm{T}}\mathcal{R}u_k-\sum_{p=1}^{q}\epsilon^p(a_k^p)^{\mathrm{T}}a_k^p$, $\tilde{\varphi}_k^j=\tilde{X}_k-\gamma\tilde{X}_{k+1}$. 未知参数矩阵$\tilde{\mathcal{S}}^{j+1}$有$\frac{1}{2}l(l+1)$个元素, 需要收集$L>\frac{1}{2}l(l+1)$个数据样本. 定义$\tilde{\Psi}^j=[(\tilde{\varphi}_k^j)^{\mathrm{T}},(\tilde{\varphi}_{k+1}^j)^{\mathrm{T}},\dots,(\tilde{\varphi}_{k+L-1}^j)^{\mathrm{T}}]^{\mathrm{T}}$, $\tilde{\mathcal{V}}^j=[(\tilde{v}_k^j)^{\mathrm{T}},(\tilde{v}_{k+1}^j)^{\mathrm{T}},\dots,(\tilde{v}_{k+L-1}^j)^{\mathrm{T}}]^{\mathrm{T}}$, 利用最小二乘法可得

$$\tilde{\mathcal{S}}^{j+1}=(\tilde{\Psi}^j(\tilde{\Psi}^j)^{\mathrm{T}})^{-1}\tilde{\Psi}^j\tilde{\mathcal{V}}^j.\tag{24}$$

在不需要系统动力学信息的情况下结合在线系统状态数据x_k以及实际输入u_k, $a_k^1,a_k^2,\dots,a_k^q$确定参数矩阵$\tilde{\mathcal{S}}$, 设计策略 Q-learning 算法 1.

算法1 基于状态的无模型策略博弈Q-learning算法.

step 1: 初始化. 可容许的矩阵$\tilde{\mathcal{S}}^0$, 以及初始控制策略u_k^0, a_k^1, $a_k^2,\dots,a_k^q$.

step 2: 策略评估. 对于$j=0,1,\dots,$收集数据样本u_k^j, $(a_k^1)^j$, $(a_k^2)^j,\dots,(a_k^q)^j$, x_k^j, 利用最小二乘法求解式(24) 得到$\tilde{\mathcal{S}}^{j+1}$.

step 3: 策略改进. 将$\tilde{\mathcal{S}}^{j+1}$代入式(19)、(20)更新改进后的目标策略$u_k^{j+1}$和$\hat{a}_k^{j+1}$.

step 4: 停止迭代条件.$\|\tilde{\mathcal{S}}^{j+1}-\tilde{\mathcal{S}}^j\|<\varepsilon$($\varepsilon$为设定的小正数).

定理 2 当$j\to\infty$时, 算法 1 学习生成的控制增益$\mathcal{K}^{j+1}$和$\mathcal{L}^{j+1}$分别收敛于式 (14) 中的最优控制策略$\mathcal{K}^*$和$\mathcal{L}^*$, 其中折扣 GARE(15) 的正定解P存在.

证明 将迭代过程中的目标策略u_k^{j+1}、$\hat{a}_k^{j+1}$代入 Q 函数 (21) 中, 并结合核矩阵$\mathcal{S}$的定义 (18) 可得

$$P^{j+1}=[I,(\mathcal{K}^{j+1})^{\mathrm{T}},(\mathcal{L}^{j+1})^{\mathrm{T}}]\,\mathcal{S}^{j+1}\begin{bmatrix}I\\ \mathcal{K}^{j+1}\\ \mathcal{L}^{j+1}\end{bmatrix}=\mathcal{Q}^a-\gamma^2\begin{bmatrix}B_u^{\mathrm{T}}P^jA\\ \hat{B}_a^{\mathrm{T}}P^jA\end{bmatrix}^{\mathrm{T}}\Xi\begin{bmatrix}B_u^{\mathrm{T}}P^jA\\ \hat{B}_a^{\mathrm{T}}P^jA\end{bmatrix}+\gamma AP^{j+1}A,$$

$$\Xi=\begin{bmatrix}\mathcal{R}+\gamma B_u^{\mathrm{T}}P^jB_u & \gamma B_u^{\mathrm{T}}P^j\hat{B}_a\\ \gamma\hat{B}_a^{\mathrm{T}}P^jB_u & \gamma\hat{B}_a^{\mathrm{T}}P^j\hat{B}_a-\hat{\epsilon}\otimes I\end{bmatrix}^{-1}.\tag{25}$$

可知在$\mathcal{S}$上迭代与式 (25) 中P上迭代等价. 已经证明在 (25) 上迭代收敛于 GARE(15) 的解, 即$j\to\infty$时, $\mathcal{S}$收敛于 (18) 各项值, 算法 1 收敛于最优解. □

2.3 基于状态数据的非策略 *Q*-learning 算法设计

将目标策略$u_k^j=\mathcal{K}^jX_k$, $a_k^{p,j}=\mathcal{L}_p^jX_k$代入跟踪系统 (5) 中重写为

$$X_{k+1}=\hat{A}X_k+(B_uu_k-B_u\mathcal{K}^jX_k)+\Big(\sum_{p=1}^{q}B_a^pa_k^p-\sum_{p=1}^{q}B_a^p\mathcal{L}_p^jX_k\Big).\tag{26}$$

其中: $\hat{A}=A+B_u\mathcal{K}^j+\sum_{p=1}^{q}B_a^p\mathcal{L}_p^j$; u_k, a_k^p为行为策略. 定义$\hat{a}_k^j=[(a_k^{1,j})^{\mathrm{T}},(a_k^{2,j})^{\mathrm{T}},\dots,(a_k^{q,j})^{\mathrm{T}}]^{\mathrm{T}}$, $\mathcal{L}^j=[(\mathcal{L}_1^j)^{\mathrm{T}},(\mathcal{L}_2^j)^{\mathrm{T}},\dots,(\mathcal{L}_q^j)^{\mathrm{T}}]^{\mathrm{T}}$.

将目标策略代入到 Q 函数 (17) 中以及 Bellman 方程 (11) 中, 可以得到依赖于u_k^j, $\hat{a}_k^j$的 Q 函数

$$
\begin{aligned}
&Q(X_k,u_k^j,a_k^{p,j})=\\
&X_k^{\mathrm{T}}\mathcal{Q}^aX_k+(u_k^j)^{\mathrm{T}}\mathcal{R}u_k^j-\\
&(\hat{a}_k^j)^{\mathrm{T}}(\hat{\epsilon}\otimes I)\hat{a}_k^j+\gamma(X_{k+1}^{j+1})^{\mathrm{T}}PX_{k+1}^{j+1}=\\
&X_k^{\mathrm{T}}\mathcal{Q}^aX_k+(u_k^j)^{\mathrm{T}}\mathcal{R}u_k^j-(\hat{a}_k^j)^{\mathrm{T}}(\hat{\epsilon}\otimes I)\hat{a}_k^j+\\
&\gamma\Big(AX_k+B_uu_k^j+\sum_{p=1}^{q}B_a^pa_k^{p,j}\Big)^{\mathrm{T}}P^{j+1}\times\\
&\Big(AX_k+B_uu_k^j+\sum_{p=1}^{q}B_a^pa_k^{p,j}\Big).
\end{aligned}\tag{27}
$$

得到依赖于目标策略的 Lyapunov 方程

$$
\begin{aligned}
&P^{j+1}=\\
&\mathcal{Q}^a+(\mathcal{K}^j)^{\mathrm{T}}\mathcal{R}\mathcal{K}^j-(\mathcal{L}^j)^{\mathrm{T}}\hat{\epsilon}\mathcal{L}^j+\gamma\hat{A}^{\mathrm{T}}P^{j+1}\hat{A}.
\end{aligned}\tag{28}
$$

沿着系统 (26) 的轨迹结合 Q 函数 (27), 推导出非策略 Bellman 方程

$$
\begin{aligned}
&X_k^{\mathrm{T}}\mathcal{Q}^aX_k+(u_k^j)^{\mathrm{T}}\mathcal{R}u_k^j-(\hat{a}_k^j)^{\mathrm{T}}(\hat{\epsilon}\otimes I)\hat{a}_k^j=\\
&X_k^{\mathrm{T}}P^{j+1}X_k-\gamma X_{k+1}^{\mathrm{T}}P^{j+1}X_{k+1}-\\
&\gamma(B_u(u_k-u_k^j)+\hat{B}_a(\hat{a}_k-\hat{a}_k^j))^{\mathrm{T}}P^{j+1}\times\\
&(B_u(u_k-u_k^j)+\hat{B}_a(\hat{a}_k-\hat{a}_k^j))+\\
&2\gamma X_{k+1}^{\mathrm{T}}P^{j+1}(B_u(u_k-u_k^j)+\hat{B}_a(\hat{a}_k-\hat{a}_k^j)).
\end{aligned}\tag{29}
$$

将系统 (26) 和依赖于目标策略的 Lyapunov 方程 (28) 代入到方程 (29) 中, 并根据核矩阵$\mathcal{S}$的定义重写非策略 Bellman 方程

$$
\begin{aligned}
&X_k^{\mathrm{T}}\mathcal{Q}^aX_k+u_k^{\mathrm{T}}\mathcal{R}u_k-(\hat{a}_k)^{\mathrm{T}}(\hat{\epsilon}\otimes I)\hat{a}_k=\\
&X_k^{\mathrm{T}}\mathcal{S}_{XX}^{j+1}X_k-\gamma X_{k+1}^{\mathrm{T}}\mathcal{S}_{XX}^{j+1}X_{k+1}+2X_k^{\mathrm{T}}\mathcal{S}_{Xu}^{j+1}u_k-\\
&2\gamma X_{k+1}^{\mathrm{T}}\mathcal{S}_{Xu}^{j+1}u_{k+1}^j+2X_k^{\mathrm{T}}\mathcal{S}_{Xa}^{j+1}\hat{a}_k-\\
&2\gamma X_{k+1}^{\mathrm{T}}\mathcal{S}_{Xa}^{j+1}\hat{a}_{k+1}^j-\gamma(u_{k+1}^j)^{\mathrm{T}}\mathcal{S}_{uu}^{j+1}u_{k+1}^j+\\
&u_k^{\mathrm{T}}\mathcal{S}_{uu}^{j+1}u_k+\hat{a}_k^{\mathrm{T}}\mathcal{S}_{aa}^{j+1}\hat{a}_k-\gamma(\hat{a}_{k+1}^j)^{\mathrm{T}}\mathcal{S}_{aa}^{j+1}\hat{a}_{k+1}^j+\\
&2u_k^{\mathrm{T}}\mathcal{S}_{ua}^{j+1}\hat{a}_k-2\gamma(u_{k+1}^j)^{\mathrm{T}}\mathcal{S}_{ua}^{j+1}\hat{a}_{k+1}^j.
\end{aligned}\tag{30}
$$

定义

$$
\begin{aligned}
&\bar{\mathcal{S}}^{j+1}=[\mathrm{vec}(\mathcal{S}_{XX}^{j+1});\mathrm{vec}(\mathcal{S}_{Xu}^{j+1});\mathrm{vec}(\mathcal{S}_{Xa}^{j+1});\\
&\qquad\mathrm{vec}(\mathcal{S}_{uu}^{j+1});\mathrm{vec}(\mathcal{S}_{ua}^{j+1});\mathrm{vec}(\mathcal{S}_{aa}^{j+1})];\\
&\psi_k^j=\left[\varphi_1^j,\varphi_2^j,\varphi_3^j,\varphi_4^j,\varphi_5^j,\varphi_6^j\right];\\
&\varphi_1^j=X_k^{\mathrm{T}}\otimes X_k^{\mathrm{T}}-\gamma X_{k+1}^{\mathrm{T}}\otimes X_{k+1}^{\mathrm{T}};\\
&\varphi_2^j=2(X_k^{\mathrm{T}}\otimes u_k^{\mathrm{T}}-\gamma X_{k+1}^{\mathrm{T}}\otimes(u_{k+1}^j)^{\mathrm{T}});\\
&\varphi_3^j=2(X_k^{\mathrm{T}}\otimes \hat{a}_k^{\mathrm{T}}-\gamma X_{k+1}^{\mathrm{T}}\otimes(\hat{a}_{k+1}^j)^{\mathrm{T}});\\
&\varphi_4^j=u_k^{\mathrm{T}}\otimes u_k^{\mathrm{T}}-\gamma(u_{k+1}^j)^{\mathrm{T}}\otimes(u_{k+1}^j)^{\mathrm{T}};\\
&\varphi_5^j=2(u_k^{\mathrm{T}}\otimes \hat{a}_k^{\mathrm{T}}-\gamma(u_{k+1}^j)^{\mathrm{T}}\otimes(\hat{a}_{k+1}^j)^{\mathrm{T}});\\
&\varphi_6^j=\hat{a}_k^{\mathrm{T}}\otimes \hat{a}_k^{\mathrm{T}}-\gamma(\hat{a}_{k+1}^j)^{\mathrm{T}}\otimes(\hat{a}_{k+1}^j)^{\mathrm{T}};\\
&\upsilon_k^j=X_k^{\mathrm{T}}\mathcal{Q}^aX_k+u_k^{\mathrm{T}}\mathcal{R}u_k-(\hat{a}_k)^{\mathrm{T}}(\hat{\epsilon}\otimes I)\hat{a}_k.
\end{aligned}\tag{31}
$$

结合定义 (31) 可将非策略 Bellman 方程 (30) 写为

$$
\psi_k^j\bar{\mathcal{S}}^{j+1}=\upsilon_k^j,\tag{32}
$$

其中$\bar{\mathcal{S}}\in\mathbb{R}^{\frac{1}{2}l(l+1)}$. 未知参数矩阵$\bar{\mathcal{S}}^{j+1}$有$\dfrac{1}{2}l(l+1)$个元素, 需要收集$L>\dfrac{1}{2}l(l+1)$个数据样本. 定义$\Psi^j=[(\psi_k^j)^{\mathrm{T}},(\psi_k^j)^{\mathrm{T}},\dots,(\psi_{k+L-1}^j)^{\mathrm{T}}]^{\mathrm{T}}$, $\mathcal{V}^j=[(\upsilon_k^j)^{\mathrm{T}},(\upsilon_{k+1}^j)^{\mathrm{T}},\dots,(\upsilon_{k+L-1}^j)^{\mathrm{T}}]^{\mathrm{T}}$, 利用最小二乘法可得

$$
\bar{\mathcal{S}}^{j+1}=((\Psi^j)^{\mathrm{T}}\Psi^j)^{-1}(\Psi^j)^{\mathrm{T}}\mathcal{V}^j.\tag{33}
$$

在不需要系统动力学信息的基础上结合在线系统状态数据x_k以及实际输入u_k, $a_k^1,a_k^2,\dots,a_k^q$确定参数矩阵$\bar{\mathcal{S}}$, 设计非策略 Q-learning 算法 2.

算法2 基于状态的无模型非策略博弈Q-learning算法.

step 1: 初始化. 可容许的矩阵$\bar{\mathcal{S}}^0$, 以及初始控制策略u_k^0, a_k^1, $a_k^2,\dots,a_k^q$.

step 2: 策略评估. 对于$j=0,1,\dots,$收集数据样本u_k^j, $(a_k^1)^j$, $(a_k^2)^j,\dots,(a_k^q)^j$, x_k^j, 利用最小二乘法求解式(33)可得$\bar{\mathcal{S}}^{j+1}$.

step 3: 策略改进. 将$\bar{\mathcal{S}}^{j+1}$代入式(19)、(20)更新改进后的目标策略u_k^{j+1}和$\hat{a}_k^{j+1}$.

step 4: 停止迭代条件.$\|\bar{\mathcal{S}}^{j+1}-\bar{\mathcal{S}}^j\|<\varepsilon$($\varepsilon$为设定的小正数).

定理 3　基于状态数据的无模型非策略博弈 Bellman 方程 (30), 对引入到系统动力学 (5) 中保证 PE 条件的任意噪声都不敏感.

证明　通过引入探测噪声信号α_k和$\beta_k^1,\beta_k^2,\dots,\beta_k^q$, 行为策略变为$\grave{u}_k=u_k+\alpha_k$, $\grave{a}_k^1=a_k^1+\beta_k^1,\dots,\grave{a}_k^q=a_k^q+\beta_k^q$, 定义$\grave{a}_k=[(\grave{a}_k^1)^{\mathrm{T}},(\grave{a}_k^2)^{\mathrm{T}},\dots,(\grave{a}_k^q)^{\mathrm{T}}]^{\mathrm{T}}$. 核矩阵$\mathcal{S}$是行为策略为$u_k$、$\hat{a}_k$时式 (22) 的解. 核矩阵$\grave{\mathcal{S}}$是加入探测噪声后行为策略为$\grave{u}_k$、$\grave{a}_k$时式 (22) 的解. 参考文献 [19], 经过简单推导可知$\mathcal{S}=\grave{\mathcal{S}}$. 因此, 在加入噪声激励后, 迭代求解的 Q 函数不存在偏差. □

定理 4　当探测噪声为 0 时, 策略博弈 Q-learning 算法 1 与非策略博弈 Q-learning 算法 2 等价, 并且当$j\to\infty$时, 基于状态数据的无模型非策略博弈 Q-learning 算法学习生成的控制增益$\mathcal{K}_k^{j+1}$和$\mathcal{L}_k^{j+1}$分别收敛于式 (14) 中的最优控制策略$\mathcal{K}_k^*$和$\mathcal{L}_k^*$, 其中折扣 GARE(15) 的正定解P存在.

证明　将式 (26) 代入基于状态数据的非策略 Bellman 方程 (30) 中并结合控制输入u_k以及攻击输入$a_k^p,p\in\mathbb{O}$的定义 (13), 化简后与策略 Bellman 方程 (22) 相等, 因此策略博弈 Q-learning 算法 1 与非策略博弈 Q-learning 算法 2 等价, 算法 2 也收敛于最优解, 并且折扣 GARE(15) 的正定解P存在. □

2.4　基于输出数据的模型未知的 OTC

利用系统输出数据进行 OTC 设计. 定义

$$\begin{aligned}
&\bar{u}_{k-1,k-N}=[u_{k-1}^{\mathrm{T}},u_{k-2}^{\mathrm{T}},\dots,u_{k-N}^{\mathrm{T}}]^{\mathrm{T}},\\
&\bar{y}_{k-1,k-N}=[\,y_{k-1}^{\mathrm{T}},y_{k-2}^{\mathrm{T}},\dots,y_{k-N}^{\mathrm{T}}]^{\mathrm{T}},\\
&\bar{a}_{k-1,k-N}^{1}=[(a_{k-1}^{1})^{\mathrm{T}},(a_{k-2}^{1})^{\mathrm{T}},\dots,(a_{k-N}^{1})^{\mathrm{T}}]^{\mathrm{T}},\\
&\qquad\vdots\\
&\bar{a}_{k-1,k-N}^{q}=[(a_{k-1}^{q})^{\mathrm{T}},(a_{k-2}^{q})^{\mathrm{T}},\dots,(a_{k-N}^{q})^{\mathrm{T}}]^{\mathrm{T}},
\end{aligned}\tag{34}$$

其中 $\bar{u}_{k-1,k-N}$, $\bar{a}_{k-1,k-N}^{p}$, $\bar{y}_{k-1,k-N}$分别为在时间$[k-N,k]$内输入、欺骗攻击、输出信号组成的向量.

引理 1 记$\bar{W}_a=[W_a^1,W_a^2,\dots,W_a^q]$, $\tilde{a}_{k-1,k-N}=[(\bar{a}_{k-1,k-N}^{1})^{\mathrm{T}},(\bar{a}_{k-1,k-N}^{2})^{\mathrm{T}},\dots,(\bar{a}_{k-1,k-N}^{q})^{\mathrm{T}}\,]^{\mathrm{T}}$, $Z_k=[\bar{u}_{k-1,k-N}^{\mathrm{T}},\tilde{a}_{k-1,k-N}^{\mathrm{T}},\bar{y}_{k-1,k-N}^{\mathrm{T}},r_{k-N}^{\mathrm{T}}]^{\mathrm{T}}$, 在$(\mathcal{A},\mathcal{C})$可观的条件下, 式 (5) 在$[k-N,k]$上可写成扩展状态方程

$$\begin{aligned}
X_k=&W_u\bar{u}_{k-1,k-N}+\sum_{p=1}^{q}W_a^p\bar{a}_{k-1,k-N}^{p}+\\
&W_r r_{k-N}+W_y\bar{y}_{k-1,k-N}=\\
&[W_u,\bar{W}_a,W_y,W_r]Z_k=WZ_k.
\end{aligned}\tag{35}$$

其中

$$\begin{aligned}
&W_u=\begin{bmatrix}\Lambda_N-\mathcal{A}^N(M_N^{\mathrm{T}}M_N)^{-1}M_N^{\mathrm{T}}V_N\\0\end{bmatrix},\\
&W_a^p=\begin{bmatrix}\Lambda_N^p-\mathcal{A}^N(M_N^{\mathrm{T}}M_N)^{-1}M_N^{\mathrm{T}}V_N^p\\0\end{bmatrix},\\
&W_y=\begin{bmatrix}\mathcal{A}^N(M_N^{\mathrm{T}}M_N)^{-1}M_N^{\mathrm{T}}\\0\end{bmatrix},\ W_r=\begin{bmatrix}0\\\mathcal{F}^N\end{bmatrix},\\
&\Lambda_N=[\mathcal{B},\mathcal{A}\mathcal{B},\dots,\mathcal{A}^{N-1}\mathcal{B}],\\
&\Lambda_N^p=[\mathcal{B}\mathcal{D}^p,\mathcal{A}\mathcal{B}\mathcal{D}^p,\dots,\mathcal{A}^{N-1}\mathcal{B}\mathcal{D}^p],\\
&M_N=[\,(\mathcal{C}\mathcal{A}^{N-1})^{\mathrm{T}},(\mathcal{C}\mathcal{A}^{N-2})^{\mathrm{T}},\dots,\mathcal{C}^{\mathrm{T}}]^{\mathrm{T}},\\
&V_N=\begin{bmatrix}0&\mathcal{C}\mathcal{B}&\mathcal{C}\mathcal{A}\mathcal{B}&\dots&\mathcal{C}\mathcal{A}^{N-2}\mathcal{B}\\0&0&\mathcal{C}\mathcal{B}&\dots&\mathcal{C}\mathcal{A}^{N-3}\mathcal{B}\\\vdots&\vdots&\ddots&\ddots&\vdots\\0&0&0&\dots&\mathcal{C}\mathcal{B}\\0&0&0&\dots&0\end{bmatrix},\\
&V_N^p=\begin{bmatrix}0&\mathcal{C}\mathcal{B}\mathcal{D}^p&\mathcal{C}\mathcal{A}\mathcal{B}\mathcal{D}^p&\dots&\mathcal{C}\mathcal{A}^{N-2}\mathcal{B}\mathcal{D}^p\\0&0&\mathcal{C}\mathcal{B}\mathcal{D}^p&\dots&\mathcal{C}\mathcal{A}^{N-3}\mathcal{B}\mathcal{D}^p\\\vdots&\vdots&\ddots&\ddots&\vdots\\0&0&0&\dots&\mathcal{C}\mathcal{B}\mathcal{D}^p\\0&0&0&\dots&0\end{bmatrix}.
\end{aligned}\tag{36}$$

证明过程参考文献 [10].

通过式 (9) 和 (35) 中X_k与Z_k的关系, 可以得到利用系统输入输出数据序列构造的代价函数为

$$J(X_k,u_k,a_k^1,\dots,a_k^p)=X_k^{\mathrm{T}}PX_k=Z_k^{\mathrm{T}}\bar{P}Z_k,\tag{37}$$

其中$\bar{P}=W^{\mathrm{T}}PW$.

定义 $\hat{Z}_k=[\bar{u}_{k-1,k-N}^{\mathrm{T}},\tilde{a}_{k-1,k-N}^{\mathrm{T}},\bar{y}_{k-1,k-N}^{\mathrm{T}},r_{k-N}^{\mathrm{T}},u_k^{\mathrm{T}},\hat{a}_k^{\mathrm{T}}]^{\mathrm{T}}$, $\bar{\mathcal{Q}}^a=W^{\mathrm{T}}\mathcal{Q}^aW$. 根据 Q 函数 (17) 可得利用系统输入输出数据表示的博弈 Q 函数

$$\begin{aligned}
&Q(X_k,u_k,a_k^1,\dots,a_k^p)=\\
&Z_k^{\mathrm{T}}\bar{\mathcal{Q}}^aZ_k+u_k^{\mathrm{T}}\mathcal{R}u_k-\\
&\hat{a}_k^{\mathrm{T}}(\hat{\epsilon}\otimes I)\hat{a}_k+\gamma Z_{k+1}^{\mathrm{T}}\bar{P}Z_{k+1}=\hat{Z}_k^{\mathrm{T}}\hat{\mathcal{H}}\hat{Z}_k.
\end{aligned}\tag{38}$$

核矩阵$\hat{\mathcal{H}}=\hat{\mathcal{H}}^{\mathrm{T}}\in\mathbb{R}^{l_2\times l_2}$, $l_2=(N+1)(\tilde{m}+d)$, $\tilde{m}=m(q+1)$, 定义为

$$\hat{\mathcal{H}}=\begin{bmatrix}\mathcal{H}_{\bar{u}\bar{u}}&\mathcal{H}_{\bar{u}a}^{a}&\mathcal{H}_{\bar{u}\bar{y}}&\mathcal{H}_{\bar{u}r}&\mathcal{H}_{\bar{u}u}&\mathcal{H}_{\bar{u}\hat{a}}\\\mathcal{H}_{\tilde{a}\bar{u}}&\mathcal{H}_{\tilde{a}\tilde{a}}&\mathcal{H}_{\tilde{a}\bar{y}}&\mathcal{H}_{\tilde{a}r}&\mathcal{H}_{\tilde{a}u}&\mathcal{H}_{\tilde{a}\hat{a}}\\\mathcal{H}_{\bar{y}\bar{u}}&\mathcal{H}_{\bar{y}a}^{a}&\mathcal{H}_{\bar{y}\bar{y}}&\mathcal{H}_{\bar{y}r}&\mathcal{H}_{\bar{y}u}&\mathcal{H}_{\bar{y}\hat{a}}\\\mathcal{H}_{r\bar{u}}&\mathcal{H}_{ra}&\mathcal{H}_{r\bar{y}}&\mathcal{H}_{rr}&\mathcal{H}_{ru}&\mathcal{H}_{r\hat{a}}\\\mathcal{H}_{u\bar{u}}&\mathcal{H}_{u\tilde{a}}&\mathcal{H}_{u\bar{y}}&\mathcal{H}_{ur}&\mathcal{H}_{uu}&\mathcal{H}_{u\hat{a}}\\\mathcal{H}_{\hat{u}}&\mathcal{H}_{\hat{a}\tilde{a}}&\mathcal{H}_{\hat{a}\bar{y}}&\mathcal{H}_{\hat{a}r}&\mathcal{H}_{\hat{a}u}&\mathcal{H}_{\hat{a}\hat{a}}\end{bmatrix}.\tag{39}$$

根据稳定性条件, 通过$\partial Q(X_k,u_k,\hat{a}_k)/\partial u_k=0$和$\partial Q(X_k,u_k,\hat{a}_k)/\partial\hat{a}_k=0$求得最优控制策略

$$\begin{aligned}
&u_k^*=(\mathcal{H}_{uu}-\mathcal{H}_{u\hat{a}}\mathcal{H}_{\hat{a}\hat{a}}^{-1}\mathcal{H}_{\hat{a}u})^{-1}[\mathcal{H}_{u\hat{a}}\mathcal{H}_{\hat{a}\hat{a}}^{-1}\mathcal{H}_1-\mathcal{H}_2]Z_k,\\
&\hat{a}_k^*=(\mathcal{H}_{\hat{a}\hat{a}}-\mathcal{H}_{\hat{a}u}\mathcal{H}_{uu}^{-1}\mathcal{H}_{u\hat{a}})^{-1}[\mathcal{H}_{\hat{a}u}\mathcal{H}_{uu}^{-1}\mathcal{H}_2-\mathcal{H}_1]Z_k.
\end{aligned}\tag{40}$$

其中: $\mathcal{H}_1=[\mathcal{H}_{\hat{a}\bar{u}},\mathcal{H}_{\hat{a}\tilde{a}},\mathcal{H}_{\hat{a}\bar{y}},\mathcal{H}_{\hat{a}r}]$, $\mathcal{H}_2=[\mathcal{H}_{u\bar{u}},\mathcal{H}_{u\tilde{a}},\mathcal{H}_{u\bar{y}},\mathcal{H}_{ur}]$.

引理 2 核矩阵$\hat{\mathcal{H}}$与核矩阵$\mathcal{S}$的关系为

$$\hat{\mathcal{H}}=\begin{bmatrix}H_{11}&H_{12}&H_{13}\\H_{21}&H_{22}&H_{23}\\H_{31}&H_{32}&H_{33}\end{bmatrix}.\tag{41}$$

其中

$$\begin{aligned}
&H_{11}=W^{\mathrm{T}}\mathcal{S}_{XX}W=\begin{bmatrix}\mathcal{H}_{\bar{u}\bar{u}}&\mathcal{H}_{\bar{u}\tilde{a}}&\mathcal{H}_{\bar{u}\bar{y}}&\mathcal{H}_{\bar{u}r},\\\mathcal{H}_{\tilde{a}\bar{u}}&\mathcal{H}_{\tilde{a}\tilde{a}}&\mathcal{H}_{\tilde{a}\bar{y}}&\mathcal{H}_{\tilde{a}r},\\\mathcal{H}_{\bar{y}\bar{u}}&\mathcal{H}_{\bar{y}\tilde{a}}&\mathcal{H}_{\bar{y}\bar{y}}&\mathcal{H}_{\bar{y}r}\\\mathcal{H}_{r\bar{u}}&\mathcal{H}_{r\tilde{a}}&\mathcal{H}_{r\bar{y}}&\mathcal{H}_{rr}\end{bmatrix},\\
&H_{12}=W^{\mathrm{T}}\mathcal{S}_{Xu}=[\mathcal{H}_{\bar{u}u}^{\mathrm{T}},\mathcal{H}_{\tilde{a}u}^{\mathrm{T}},\mathcal{H}_{\bar{y}u}^{\mathrm{T}},\mathcal{H}_{ru}^{\mathrm{T}}]^{\mathrm{T}},\\
&H_{13}=W^{\mathrm{T}}\mathcal{S}_{Xa}=[\mathcal{H}_{\bar{u}\hat{a}}^{\mathrm{T}},\mathcal{H}_{\tilde{a}\hat{a}}^{\mathrm{T}},\mathcal{H}_{\bar{y}\hat{a}}^{\mathrm{T}},\mathcal{H}_{r\hat{a}}^{\mathrm{T}}]^{\mathrm{T}},\\
&H_{22}=\mathcal{S}_{uu}=\mathcal{H}_{uu},\ H_{23}=\mathcal{S}_{ua}=\mathcal{H}_{u\hat{a}},\\
&H_{33}=\mathcal{S}_{aa}=\mathcal{H}_{\hat{a}\hat{a}}.
\end{aligned}\tag{42}$$

结合 Q 函数 (17) 与式 (35) 可证该引理.

2.5 基于输出数据的非策略 Q-learning 算法设计

结合式 (35) 与 (30) 得到基于输出数据的非策略 Bellman 方程为

$$\begin{aligned}
&Z_k^{\mathrm{T}}\bar{\mathcal{Q}}^aZ_k+u_k^{\mathrm{T}}\mathcal{R}u_k-(\hat{a}_k)^{\mathrm{T}}(\hat{\epsilon}\otimes I)\hat{a}_k=\\
&Z_k^{\mathrm{T}}W^{\mathrm{T}}\mathcal{S}_{XX}^{j+1}WZ_k-\gamma Z_{k+1}^{\mathrm{T}}W^{\mathrm{T}}\mathcal{S}_{XX}^{j+1}WZ_{k+1}+\\
&2Z_k^{\mathrm{T}}W^{\mathrm{T}}\mathcal{S}_{Xu}^{j+1}u_k-2\gamma Z_{k+1}^{\mathrm{T}}W^{\mathrm{T}}\mathcal{S}_{Xu}^{j+1}u_{k+1}^{j}+\\
&2Z_k^{\mathrm{T}}W^{\mathrm{T}}\mathcal{S}_{Xa}^{j+1}\hat{a}_k-2\gamma Z_{k+1}^{\mathrm{T}}W^{\mathrm{T}}\mathcal{S}_{Xa}^{j+1}\hat{a}_{k+1}^{j}+\\
&u_k^{\mathrm{T}}\mathcal{S}_{uu}^{j+1}u_k-\gamma(u_{k+1}^{j})^{\mathrm{T}}\mathcal{S}_{uu}^{j+1}u_{k+1}^{j}+\\
&\hat{a}_k^{\mathrm{T}}\mathcal{S}_{aa}^{j+1}\hat{a}_k-\gamma(\hat{a}_{k+1}^{j})^{\mathrm{T}}\mathcal{S}_{aa}^{j+1}\hat{a}_{k+1}^{j}+\\
&2u_k^{\mathrm{T}}\mathcal{S}_{ua}^{j+1}\hat{a}_k-2\gamma(u_{k+1}^{j})^{\mathrm{T}}\mathcal{S}_{ua}^{j+1}\hat{a}_{k+1}^{j}.
\end{aligned}\tag{43}$$

定义

$$
\begin{aligned}
&\bar{\psi}_k^j = [\bar{\varphi}_1^j, \bar{\varphi}_2^j, \bar{\varphi}_3^j, \bar{\varphi}_4^j, \bar{\varphi}_5^j, \bar{\varphi}_6^j];\\
&\bar{\varphi}_1^j = Z_k^{\mathrm{T}} \otimes Z_k^{\mathrm{T}} - \gamma Z_{k+1}^{\mathrm{T}} \otimes Z_{k+1}^{\mathrm{T}};\\
&\bar{\varphi}_2^j = 2(Z_k^{\mathrm{T}} \otimes u_k^{\mathrm{T}} - \gamma Z_{k+1}^{\mathrm{T}} \otimes (u_{k+1}^j)^{\mathrm{T}});\\
&\bar{\varphi}_3^j = 2(Z_k^{\mathrm{T}} \otimes \hat{a}_k^{\mathrm{T}} - \gamma Z_{k+1}^{\mathrm{T}} \otimes (\hat{a}_{k+1}^j)^{\mathrm{T}});\\
&\bar{\varphi}_4^j = u_k^{\mathrm{T}} \otimes u_k^{\mathrm{T}} - \gamma (u_{k+1}^j)^{\mathrm{T}} \otimes (u_{k+1}^j)^{\mathrm{T}};\\
&\bar{\varphi}_5^j = 2(u_k^{\mathrm{T}} \otimes \hat{a}_k^{\mathrm{T}} - \gamma (u_{k+1}^j)^{\mathrm{T}} \otimes (\hat{a}_{k+1}^j)^{\mathrm{T}});\\
&\bar{\varphi}_6^j = \hat{a}_k^{\mathrm{T}} \otimes \hat{a}_k^{\mathrm{T}} - \gamma (\hat{a}_{k+1}^j)^{\mathrm{T}} \otimes (\hat{a}_{k+1}^j)^{\mathrm{T}};\\
&\bar{\mathcal{H}}^{j+1} = [\mathrm{vec}(H_{11}^{j+1}); \mathrm{vec}(H_{12}^{j+1}); \mathrm{vec}(H_{13}^{j+1});\\
&\qquad \mathrm{vec}(H_{22}^{j+1}); \mathrm{vec}(H_{23}^{j+1}); \mathrm{vec}(H_{33}^{j+1})];\\
&\bar{v}_k^j = Z_k^{\mathrm{T}} \bar{\mathcal{Q}}^a Z_k + u_k^{\mathrm{T}} \mathcal{R} u_k - (\hat{a}_k)^{\mathrm{T}} (\epsilon \otimes I) \hat{a}_k. \qquad (44)
\end{aligned}
$$

与基于状态数据的非策略 Q-learning 算法推理类似, 根据引理 2 和式 (44) 中的定义重写方程 (43) 可得

$$\bar{\psi}_k^j \bar{\mathcal{H}}^{j+1} = \bar{v}_k^j, \qquad (45)$$

其中 $\bar{\mathcal{H}} \in \mathbb{R}^{\frac{1}{2}l_2(l_2+1)}$. 由于未知矩阵 $\bar{\mathcal{H}}$ 有 $\dfrac{1}{2}l_2(l_2+1)$ 个元素, 需要收集 $L_1 > \dfrac{1}{2}l_2(l_2+1)$ 个数据样本. 定义 $\bar{\Psi}^j = [(\bar{\psi}_k^j)^{\mathrm{T}}, (\bar{\psi}_{k+1}^j)^{\mathrm{T}}, \dots, (\bar{\psi}_{k+L-1}^j)^{\mathrm{T}}]^{\mathrm{T}}$, $\bar{\Upsilon}^j = [(\bar{v}_k^j)^{\mathrm{T}}, (\bar{v}_{k+1}^j)^{\mathrm{T}}, \dots, (\bar{v}_{k+L-1}^j)^{\mathrm{T}}]^{\mathrm{T}}$, 利用最小二乘法可得

$$\bar{\mathcal{H}}^{j+1} = ((\bar{\Psi}^j)^{\mathrm{T}} \bar{\Psi}^j)^{-1} (\bar{\Psi}^j)^{\mathrm{T}} \bar{\Upsilon}^j. \qquad (46)$$

在不需要系统动力学信息的基础上结合在线系统输出数据 y_k 及输入 u_k, $a_k^1, a_k^2, \dots, a_k^q$ 确定参数矩阵 $\bar{\mathcal{H}}$, 设计基于输出数据非策略博弈 Q-learning 算法 3.

算法3 基于输出数据的无模型非策略博弈Q-learning算法.

step 1: 初始化. 可容许的矩阵 $\hat{\mathcal{H}}^0$, 以及初始控制策略 u_k^0, a_k^1, $a_k^2, \dots, a_k^q$.

step 2: 策略评估. 对于 $j = 0, 1, \dots$, 收集数据样本 u_k^j, $(a_k^1)^j$, $(a_k^2)^j$, $\dots$, $(a_k^q)^j$, y_k^j, 利用最小二乘法求解式(47)可得 $\bar{\mathcal{H}}^{j+1}$.

step 3: 策略改进. 利用式(41)更新改进后的目标策略.

step 4: 停止迭代条件. $\|\hat{\mathcal{H}}^{j+1} - \hat{\mathcal{H}}^j\| < \varepsilon$ (ε为设定的小正数).

定理 5　当 $j \to \infty$ 时, 基于输出数据的无模型非策略博弈 Q-learning 算法学习生成的控制增益 $\mathcal{K}_k^{j+1}$ 和 $\mathcal{L}_k^{j+1}$ 分别收敛于式 (14) 中的最优控制策略 $\mathcal{K}_k^*$ 和 $\mathcal{L}_k^*$, 其中折扣 GARE(15) 的正定解 P 存在.

证明　将迭代过程中的目标策略 u_k^{j+1}、$\hat{a}_k^{j+1}$ 代入非策略 Q 函数 Bellman 方程 (44) 中, 并结合核矩阵 $\hat{\mathcal{H}}$ 的定义以及式 (42), 与定理 2 的证明类似, 可知算法 3 收敛于最优解. □

3　仿真结果

在仿真实验中以 F-16 飞机自动驾驶仪为例, 分别展示受到网络攻击后, 使用基于策略的状态反馈算法 1 和非策略的状态反馈算法 2 以及非策略的输出反馈算法 3 下系统的最优跟踪控制性能.

F-16 飞机短周期动力学有 3 种状态为 $x = [\alpha, q, \delta_e]^{\mathrm{T}}$. 其中: α 为迎角, q 为俯仰速率, δ_e 为升降舵偏转角[13]. 使用零阶保持离散化技术将飞机动力学的连续时间模型离散化并考虑两个通道存在欺骗攻击的情况, 将系统建模为

$$x_{k+1} = \mathcal{A}x_k + \mathcal{B}u_k + \sum_{p=1}^{2} \mathcal{B}\mathcal{D}^p a_k^p, \ y_k = \mathcal{C}x_k. \qquad (47)$$

其中

$$
\begin{aligned}
&\mathcal{A} = \begin{bmatrix} 0.906\,488 & 0.081\,601\,2 & -0.000\,5 \\ 0.074\,134\,9 & 0.901\,21 & -0.000\,708\,383 \\ 0 & 0 & 0.132\,655 \end{bmatrix},\\
&\mathcal{B} = \begin{bmatrix} -0.001\,508\,08 & 0.009\,518\,92 \\ -0.009\,6 & 0.000\,383\,73 \\ 0.867\,345 & 0 \end{bmatrix},\\
&\mathcal{D}^1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \ \mathcal{D}^2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \ \mathcal{C} = [1, 1, 1]. \qquad (48)
\end{aligned}
$$

权重矩阵 $Q = 0.9$, $R = \begin{bmatrix} 0.25 & 0 \\ 0 & 0.001 \end{bmatrix}$, $\epsilon^1 = 0.83$, $\epsilon^2 = 0.81$, 状态初值为 $x = [10, 5, -2]^{\mathrm{T}}$, 折现因子 $\gamma = 0.88$, 参考轨迹参数为 $\mathcal{F} = 1, r_1 = 5$. 利用式 (14)、(15) 可以计算得到受欺骗攻击的系统的最优跟踪控制策略 $\mathcal{K}^*$、攻击增益 $\mathcal{L}^*$、矩阵 P 分别为

$$
\begin{aligned}
&\mathcal{K}^* = \\
&\begin{bmatrix} -0.931\,7 & -0.933\,3 & -0.132\,4 & 0.942\,0 \\ -10.111\,1 & -10.174\,1 & -0.364\,8 & 11.000\,2 \end{bmatrix},\\
&\mathcal{L}^* = \begin{bmatrix} 0.280\,6 & 0.281\,1 & 0.039\,9 & -0.283\,7 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0.012\,5 & 0.012\,6 & 0.000\,5 & -0.013\,6 \end{bmatrix},\\
&P = \\
&\begin{bmatrix} 1.901\,4 & 1.907\,7 & 0.936\,1 & -1.989\,5 \\ 1.907\,7 & 1.914\,0 & 0.936\,2 & -1.996\,6 \\ 0.936\,1 & 0.936\,2 & 0.905\,1 & -0.936\,6 \\ -1.989\,5 & -1.996\,6 & -0.936\,6 & 2.098\,0 \end{bmatrix}. \qquad (49)
\end{aligned}
$$

3.1　算法 1 结果

欺骗攻击下系统 (47) 执行状态数据驱动的无模型策略博弈 Q-learning 算法 1 的结果如图 1 和图 2 所示. 为了保证学习过程的 PE 条件, 在控制输入前 10 000 步中加入探测噪声. 图 1 显示了控制策略 $\mathcal{K}$、攻击增益 $\mathcal{L}$、核矩阵 $\mathcal{S}$ 逐步学习收敛到最优增益 $\mathcal{K}^*$, $\mathcal{L}^*$ 与最优核矩阵 $\mathcal{S}^*$ 的过程. 图 2 显示系统 (47) 的输出 y 在学习过程结束后跟踪上期望轨迹 r.

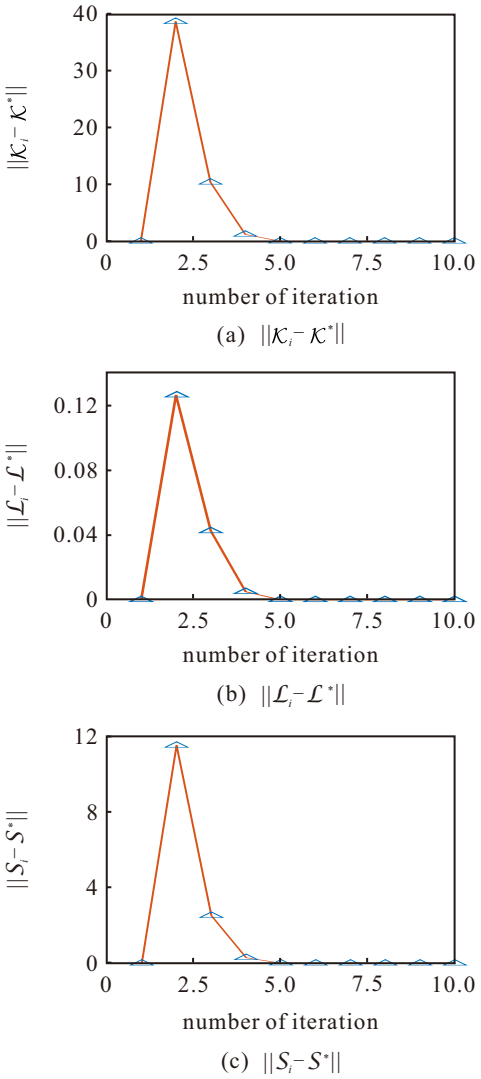

图1 执行算法 1 时$\mathcal{K}_i$、$\mathcal{L}_i$、$\mathcal{S}_i$收敛到最优值的过程

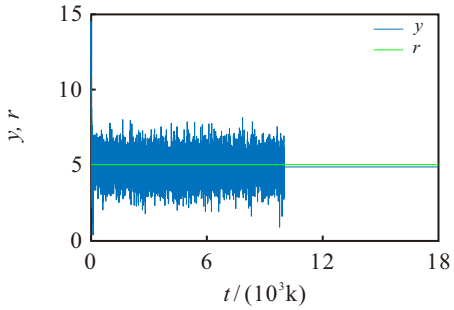

图2 算法 1 学习过程中系统输出值y和参考轨迹r

3.2 算法 2 结果

在与算法 1 相同的条件下, 系统 (47) 执行状态数据驱动的无模型非策略博弈 Q-learning 算法 2 的结果如图 3 和图 4 所示. 图 3 显示了控制策略$\mathcal{K}$、攻击增益$\mathcal{L}$、核矩阵$\mathcal{S}$逐步学习收敛到最优增益$\mathcal{K}^*$, $\mathcal{L}^*$与最优核矩阵$\mathcal{S}^*$的过程. 由图 4 可以看出, 系统 (47) 的输出y在学习过程结束后跟踪上期望轨迹r, 并给出了算法 2 执行过程中系统的所有输入u^a的变化过程.

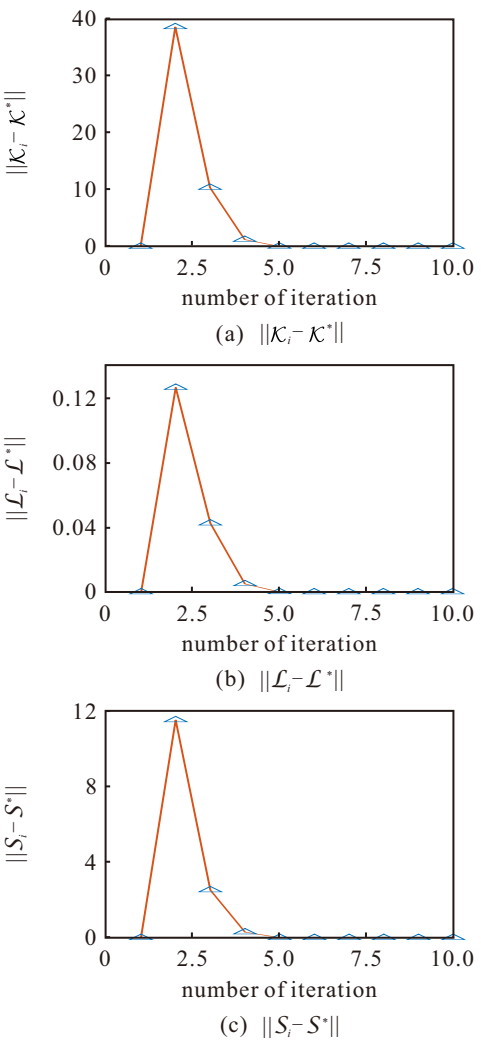

图3 执行算法 2 时$\mathcal{K}_i$、$\mathcal{L}_i$、$\mathcal{S}_i$收敛到最优值的过程

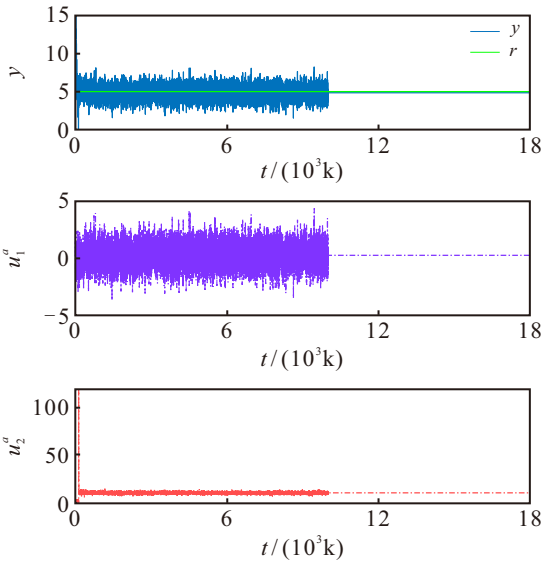

图4 算法 2 学习过程中系统输出y和输入u^a的变化

3.3 算法 3 结果

当系统 (47) 状态不可测时, 考虑$N=2$的情况执行输入输出数据驱动的无模型非策略博弈 Q-learning 算法 3, 结果如图 5 和图 6 所示. 为了保证学习过程的 PE 条件, 在控制输入前 28 000 步中加入探

测噪声. 图 5 显示了控制策略$\mathcal{K}$、攻击增益$\mathcal{L}$、核矩阵$\hat{\mathcal{H}}$逐步学习收敛到最优增益$\mathcal{K}^*$、$\mathcal{L}^*$与核矩阵$\hat{\mathcal{H}}^*$的过程. 由图 6 中可以看出, 系统输出y在学习过程结束后跟踪上期望轨迹r, 并给出了算法 3 执行过程中系统的所有输入u^a的变化过程.

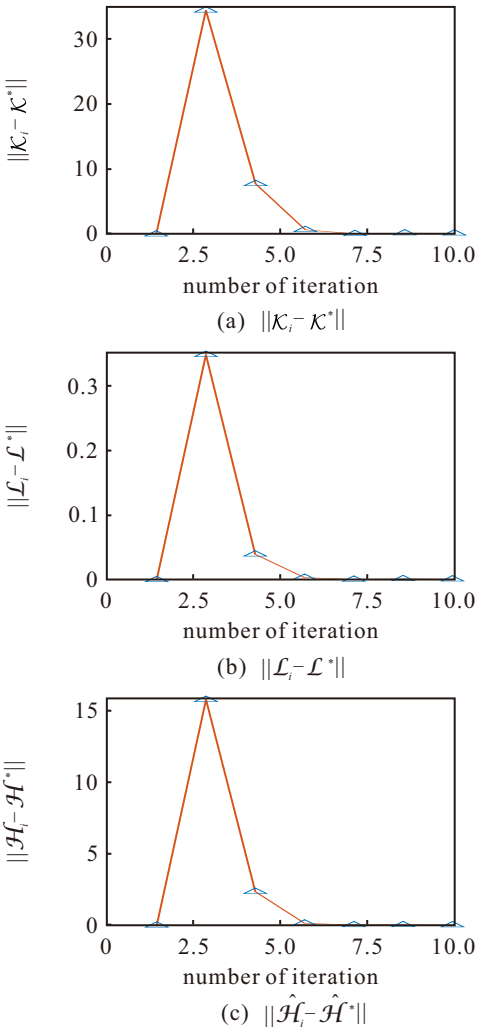

图5　执行算法 3 时$\mathcal{K}_i$、$\mathcal{L}_i$、$\hat{\mathcal{H}}_i$收敛到最优值的过程

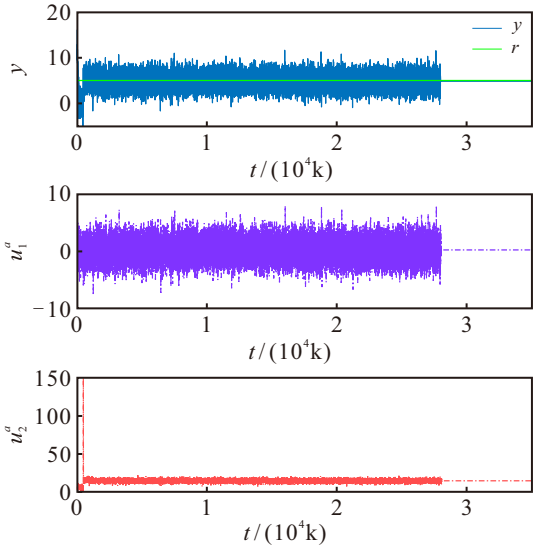

图6　算法 3 学习过程中系统输出y与输入u^a的变化

3.4　不同的探测噪声对算法 1 及算法 2 的影响对比

在相同的条件下向系统 (48) 中分别加入不同的探测噪声, 观察对基于策略的算法 1 及非策略算法 2 的影响. 定义探测噪声形式为$w(a,b,d,e)=a\sin(bk)+\sin(dk)\cos(ek)$, 加入的探测噪声为$n_1=[w_1;w_2]$, 其中 4 种不同的探测噪声参数由表 1 给出. 在 4 种不同的探测噪声下学习迭代相同的次数, 执行算法 1 和算法 2 后得到的控制策略、攻击增益与最优控制增益之间误差的范数$||\mathcal{K}_i-\mathcal{K}^*||$、$||\mathcal{L}_i-\mathcal{L}^*||$分别由表 2 和表 3 给出. 可以看出, 在不同探测噪声条件下, 与策略算法 1 相比, 非策略算法 2 具有无偏性的优点, 这也验证了定理 3 的准确性.

表1　探测噪声

Situations	Parameters	
	w_1	w_2
case 1	$w(1,0.4,0.3,1.5)$	$w(1,0.6,0.3,1.5)$
case 2	$w(10,0.4,0.3,1.5)$	$w(10,0.6,0.3,1.5)$
case 3	$w(1,10.4,10.3,14.5)$	$w(1,10.6,10.3,14.5)$
case 4	$w(10,10.4,10.3,14.5)$	$w(10,10.6,10.3,14.5)$

表2　算法 1 的执行结果

Situations	Norm	
	$\|\|\mathcal{K}_i-\mathcal{K}^*\|\|$	$\|\|\mathcal{L}_i-\mathcal{L}^*\|\|$
case 1	2.564e-05	1.917e-07
case 2	6.643e-06	9.876e-08
case 3	2.032e-04	7.734e-07
case 4	9.135e-04	1.038e-05

表3　算法 2 的执行结果

Situations	Norm	
	$\|\|\mathcal{K}_i-\mathcal{K}^*\|\|$	$\|\|\mathcal{L}_i-\mathcal{L}^*\|\|$
case 1	4.704e-05	3.934e-07
case 2	1.22e-06	3.662e-06
case 3	7.596e-06	2.527e-06
case 4	3.634e-06	3.923e-06

4　结　论

本文提出了一种非策略 Q-learning 算法解决欺骗攻击下动力学未知的线性离散系统的 OTC 问题. 首先, 通过在攻击输入中加入一个权重矩阵描述控制器通信信道受到多重欺骗攻击的情况, 并在 LQT 框架内建立了跟踪系统模型. 其次, 将 OTC 描述为系统控制器与欺骗攻击同时参与系统运行的零和博弈问题. 设计基于状态数据的非策略 Q-learning 算法学习系统的 OTC 增益, 解决了系统动力学未知以及欺骗攻击难以按照给定要求更新的问题, 并证明了在满足 PE 条件的探测噪声下该算法的

求解不存在偏差. 同时, 设计了基于输出数据的非策略 Q-learning 算法处理系统状态不可用的问题. 最后, 通过对 F-16 飞机自动驾驶仪的跟踪控制仿真, 验证了所设计非策略 Q-learning 算法的有效性以及对探测噪声影响的无偏性.

参考文献 (References)

[1] Mu C X, Ni Z, Sun C Y, et al. Air-breathing hypersonic vehicle tracking control based on adaptive dynamic programming[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(3): 584-598.

[2] Rosaline A D, Somarajan U. Structured H_∞ controller for an uncertain deregulated power system[J]. IEEE Transactions on Industry Applications, 2019, 55(1): 892-906.

[3] 赵诗影, 闫泽, 孟庆鑫, 等. 基于宽度学习系统的气动波纹管驱动器无模型跟踪控制[J]. 控制与决策, 2024, 39(1): 121-128.
(Zhao S Y, Yan Z, Meng Q X, et al. Model-free tracking control of pneumatic bellow actuator based on broad learning system[J]. Control and Decision, 2024, 39(1): 121-128.)

[4] Kiumarsi B, Lewis F L, Modares H, et al. Reinforcement Q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics[J]. Automatica, 2014, 50(4): 1167-1175.

[5] Ali Asad Rizvi S, Pertzborn A J, Lin Z L. Reinforcement learning based optimal tracking control under unmeasurable disturbances with application to HVAC systems[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(12): 7523-7533.

[6] Xu Y, Yang C Y, Zhou L N, et al. Adaptive event-triggered synchronization of neural networks under stochastic cyber-attacks with application to Chua's circuit[J]. Neural Networks, 2023, 166: 11-21.

[7] 王龙, 黄锋. 多智能体博弈、学习与控制[J]. 自动化学报, 2023, 49(3): 580-613.
(Wang L, Huang F. An interdisciplinary survey of multi-agent games, learning, and control[J]. Acta Automatica Sinica, 2023, 49(3): 580-613.)

[8] Zhang L, Fan J L, Xue W Q, et al. Data-driven H_∞ optimal output feedback control for linear discrete-time systems based on off-policy Q-learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(7): 3553-3567.

[9] Kong L H, He W, Dong Y T, et al. Asymmetric bounded neural control for an uncertain robot by state feedback and output feedback[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2021, 51(3): 1735-1746.

[10] Lewis F L, Vamvoudakis K G. Reinforcement learning for partially observable dynamic processes: Adaptive dynamic programming using measured output data[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2011, 41(1): 14-25.

[11] Yang X, Xu M M, Wei Q L. Dynamic event-sampled control of interconnected nonlinear systems using reinforcement learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(1): 923-937.

[12] 温广辉, 杨涛, 周佳玲, 等. 强化学习与自适应动态规划:从基础理论到多智能体系统中的应用进展综述[J]. 控制与决策, 2023, 38(5): 1200-1230.
(Wen G H, Yang T, Zhou J L, et al. Reinforcement learning and adaptive/approximate dynamic programming: A survey from theory to applications in multi-agent systems[J]. Control and Decision, 2023, 38(5): 1200-1230.)

[13] 赵振根, 程磊. 基于增量式Q学习的固定翼无人机跟踪控制性能优化[J]. 控制与决策, 2024, 39(2): 391-400.
(Zhao Z G, Cheng L. Performance optimization for tracking control of fixed-wing UAV with incremental Q-learning[J]. Control and Decision, 2024, 39(2): 391-400.)

[14] Chen C, Xie L H, Jiang Y, et al. Robust output regulation and reinforcement learning-based output tracking design for unknown linear discrete-time systems[J]. IEEE Transactions on Automatic Control, 2023, 68(4): 2391-2398.

[15] Gao W N, Jiang Z P. Adaptive optimal output regulation of time-delay systems via measurement feedback[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(3): 938-945.

[16] Yang Y L, Guo Z S, Xiong H Y, et al. Data-driven robust control of discrete-time uncertain linear systems via off-policy reinforcement learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(12): 3735-3747.

[17] 李金娜, 尹子轩. 基于非策略Q-学习的网络控制系统最优跟踪控制[J]. 控制与决策, 2019, 34(11): 2343-2349.
(Li J N, Yin Z X. Off-policy Q-learning: Optimal tracking control for networked control systems[J]. Control and Decision, 2019, 34(11): 2343-2349.)

[18] Wang J, Mi X R, Shen H, et al. Optimal control for interconnected multi-area power systems with unknown dynamics: An off-policy Q-learning method[J]. IEEE Transactions on Circuits and Systems, 2024, 71(5): 2849-2853.

[19] Kiumarsi B, Lewis F L, Jiang Z P. H_∞ control of linear discrete-time systems: Off-policy reinforcement learning[J]. Automatica, 2017, 78: 144-152.

作者简介

宋星星 (1996–), 女, 博士生, 主要研究方向为优化控制、强化学习, E-mail: xxsong0125@163.com;

储昭碧 (1970–), 男, 教授, 博士生导师, 主要研究方向为机器视觉、智能系统与装备, E-mail: zbchu@hfut.edu.cn.