

doi: 10.12052/gdutxb.170081

基于值迭代的自适应动态规划的收敛条件

刘毅, 章云

(广东工业大学 自动化学院, 广东 广州 510006)

摘要: 研究了应用于离散时间非仿射非线性系统的基于值迭代的自适应动态规划的收敛条件, 指出了迭代性能指标函数初始化为半正定函数可保证值迭代收敛到最优, 并给出了证明.

关键词: 自适应动态规划; 值迭代; 收敛

中图分类号: TP273 文献标志码: A 文章编号: 1007-7162(2017)05-0010-05

Convergence Condition of Value-iteration Based Adaptive Dynamic Programming

Liu Yi, Zhang Yun

(School of Automation, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: The convergence condition of value-iteration based adaptive dynamic programming which is applied to discrete time nonlinear non-affine system is studied. Convergence of value-iteration based adaptive dynamic programming is proven. The proof shows that value iteration will converge to the optimal when the initial iterative performance index function is a positive semi-definite function.

Key words: adaptive dynamic programming; value iteration; convergence

动态规划能精确求解哈密顿-雅可比-贝尔曼方程, 是求解动态优化问题的有力工具^[1]. 但动态规划是一种逆向计算方法, 最优控制序列必须在时间上逆向求解, 导致计算量和存储量随着控制序列的长度也即解的时间维度的增加而急剧增长. 这称为“维数灾”问题^[2]. Werbos提出的自适应动态规划是一种采用函数近似结构迭代求取哈密顿-雅可比-贝尔曼方程近似解的算法^[3-6]. 自适应动态规划通过迭代计算, 令迭代控制策略和迭代性能指标函数逐渐逼近最优控制策略和最优性能指标函数, 避免了动态规划在时间上逆向求解所面临的“维数灾”问题. 自适应动态规划又称为自适应评判设计^[7]、近似动态规划^[4]、神经动态规划^[8]、强化学习^[9].

自适应动态规划的迭代方式可分为值迭代和策略迭代^[10-11]. 在策略迭代方面, Murray等给出了一种应用于连续时间仿射非线性系统的策略迭代, 首次指出从稳定控制策略开始进行策略迭代可保证迭代收敛和系统稳定, 并给出了稳定性和收敛性证明. 这

是学术界首次给出的策略迭代的数学证明^[6]. 文献^[12]研究了应用于具有饱和约束的连续非线性系统的策略迭代, 并证明了其收敛性. Liu等提出了一种应用于离散非线性系统的策略迭代, 并给出了迭代控制策略初始化为稳定控制策略可保证迭代收敛到最优的理论证明^[13]. Wang等研究了带误差边界的策略迭代以解决有限时间离散非线性系统的优化问题^[14].

在值迭代方面, Al-Tamimi等^[15]给出了一种应用于离散时间仿射非线性系统的值迭代, 首次指出迭代性能指标函数初始化为零可保证迭代收敛到最优, 并给出了数学证明. Zhang等^[16]定义了新型的性能指标函数解决离散仿射非线性系统的最优跟踪控制问题, 并给出了值迭代的收敛性和最优性证明. 随后Wang等^[17]进行了应用于一般离散时间非仿射非线性系统的值迭代的收敛性和最优性证明. 文献^[15-17]要求迭代性能指标函数初始化为零, 而且迭代控制策略不能保证系统的稳定性. Wei等^[18]给出一种保证迭代控制策略是稳定控制策略的迭代性能指标函数

收稿日期: 2017-04-11

基金项目: 国家自然科学基金资助项目(U1501251, 51307025); 高等学校博士学科点专项科研基金资助项目(20124420130001)

作者简介: 刘毅(1979-), 男, 博士研究生, 主要研究方向为智能控制、优化控制.

通信作者: 章云(1963-), 男, 教授, 主要研究方向为优化控制、非线性控制、多智能体技术等. E-mail: yun@gdut.edu.cn

的初始化方法,并给出了收敛性证明。

文献[19]证明了半正定的初始的迭代性能指标函数的收敛性,极大地放松了值迭代的初始条件。本文证明了半正定的初始的迭代性能指标函数的收敛性,证明方法与文献[19]是不相同的。

1 问题描述

本文所描述的离散时间非仿射非线性系统为

$$\mathbf{x}_{k+1} = \mathbf{F}(\mathbf{x}_k, \mathbf{u}_k), k = 0, 1, 2, \dots, \quad (1)$$

其中 $\mathbf{x}_k \in \mathbb{R}^n$ 是状态向量, $\mathbf{u}_k \in \mathbb{R}^m$ 是控制向量。 $\forall \mathbf{x}_k, \mathbf{u}_k$, $\mathbf{F}(\mathbf{x}_k, \mathbf{u}_k)$ 连续, $\mathbf{F}(0, 0) = 0$ 。 $\mathbf{x}=0$ 是系统(1)的稳定状态。另外,系统(1)是可控的。

优化目标是寻找控制序列 $\underline{\mathbf{u}}_k = (\mathbf{u}_k, \mathbf{u}_{k+1}, \dots)$,使得性能指标函数(2)最小:

$$J(\mathbf{x}_k, \underline{\mathbf{u}}_k) = \sum_{t=k}^{\infty} U(\mathbf{x}_t, \mathbf{u}_t), \quad (2)$$

其中 $U(\mathbf{x}_t, \mathbf{u}_t)$ 称为效用函数,是连续的正定函数。假设采用状态反馈对系统(1)进行控制,即 $\mathbf{u}_k = \mathbf{u}(\mathbf{x}_k)$ 。另外假设反馈控制满足 $\mathbf{u}(0) = 0$ 。

根据Bellman最优定理,最优性能指标函数满足哈密顿-雅可比-贝尔曼方程:

$$J^*(\mathbf{x}_k) = \min_{\mathbf{u}(\mathbf{x}_k)} \{U(\mathbf{x}_k, \mathbf{u}(\mathbf{x}_k)) + J^*(\mathbf{x}_{k+1})\}. \quad (3)$$

式(3)也可表示为

$$J^*(\mathbf{x}_k) = U(\mathbf{x}_k, \mathbf{u}^*(\mathbf{x}_k)) + J^*(\mathbf{x}_{k+1}), \quad (4)$$

其中 $\mathbf{u}^*(\mathbf{x}_k)$ 是最优控制策略:

$$\mathbf{u}^*(\mathbf{x}_k) = \arg \min_{\mathbf{u}(\mathbf{x}_k)} \{U(\mathbf{x}_k, \mathbf{u}(\mathbf{x}_k)) + J^*(\mathbf{x}_{k+1})\}, \quad (5)$$

即

$$\mathbf{u}^*(\mathbf{x}_k) = \arg \min_{\mathbf{u}(\mathbf{x}_k)} \{U(\mathbf{x}_k, \mathbf{u}(\mathbf{x}_k)) + J^*(\mathbf{x}_{k+1})\}. \quad (6)$$

对于非线性系统,直接求解哈密顿-雅可比-贝尔曼方程是十分困难的。自适应动态规划可通过迭代间接求解哈密顿-雅可比-贝尔曼方程。设 $\mathbf{u}_i(\mathbf{x}_k)$ 是迭代控制策略, $J_i(\mathbf{x}_k)$ 是迭代性能指标函数, $i = 0, 1, 2, \dots$ 是迭代次数。初始化迭代性能指标函数 $J_0(\mathbf{x})$ 为半正定函数,值迭代的迭代过程为

$$\mathbf{u}_i(\mathbf{x}_k) = \arg \min_{\mathbf{u}(\mathbf{x}_k)} \{U(\mathbf{x}_k, \mathbf{u}(\mathbf{x}_k)) + J_i(\mathbf{x}_{k+1})\}, \quad (7)$$

$$J_{i+1}(\mathbf{x}_k) = \min_{\mathbf{u}(\mathbf{x}_k)} \{U(\mathbf{x}_k, \mathbf{u}(\mathbf{x}_k)) + J_i(\mathbf{x}_{k+1})\}. \quad (8)$$

2 收敛性分析

定义一个连续函数的集合:

$$\Gamma = \{\varphi(\mathbf{x}) : \varphi(\mathbf{x}) \geq 0, \varphi(0) = 0\}. \quad (9)$$

Γ 是半正定函数的集合。本文将以一种有别于文献[19]的新方法证明如果 $J_0(\mathbf{x}_k) \in \Gamma$,则 $\lim_{i \rightarrow \infty} \mathbf{u}_i(\mathbf{x}_k) = \mathbf{u}^*(\mathbf{x}_k)$, $\lim_{i \rightarrow \infty} J_i(\mathbf{x}_k) = J^*(\mathbf{x}_k)$ 。

2.1 证明思路

对于优化控制问题,不但要求 $\mathbf{u}(\mathbf{x}_k) \in \mathbb{R}^m$ 是稳定控制策略,还要求性能指标函数(2)是有限的。根据文献[17-18],容许控制策略的定义如定义1。

定义1 如果 $\mathbf{u}(\mathbf{x})$ 在 $\mathbb{R}^n$ 上是连续的, $\mathbf{u}(0) = 0$, $\mathbf{u}(\mathbf{x})$ 在 $\mathbb{R}^n$ 上稳定系统(1),并且 $\forall \mathbf{x}_k \in \mathbb{R}^n$, $J(\mathbf{x}_k)$ 是有限的,那么 $\mathbf{u}(\mathbf{x})$ 称为容许控制策略。

定义一个连续函数的集合:

$$\Psi = \left\{ \psi(\mathbf{x}_k) : \psi(\mathbf{x}_k) = \sum_{t=k}^{\infty} U(\mathbf{x}_t, \mathbf{v}(\mathbf{x}_t)) \right\}, \quad (10)$$

其中 $\mathbf{v}(\mathbf{x}) \in \mathbb{R}^m$ 是容许控制策略。若 $\mathbf{x}_k = 0$,则 $\psi(0) = 0$;若 $\mathbf{x}_k \neq 0$,则 $\psi(0) > 0$ 。 $\psi(\mathbf{x}_k)$ 是正定函数。由式(10)可见,如果 $J_i(\mathbf{x}_k) \in \Psi$,则存在容许控制策略 $\mathbf{v}_i(\mathbf{x})$,使得式(11)成立。

$$J_i(\mathbf{x}_k) = \sum_{t=k}^{\infty} U(\mathbf{x}_t, \mathbf{v}_i(\mathbf{x}_t)) = U(\mathbf{x}_k, \mathbf{v}_i(\mathbf{x}_k)) + J_i(\mathbf{x}_{k+1}). \quad (11)$$

下文将首先证明如果 $J_i(\mathbf{x}_k) \in \Psi$,则函数数列 $\{J_i(\mathbf{x}_k)\}$ 单调且收敛到 $J^*(\mathbf{x}_k)$ 。然后证明如果 $J_i(\mathbf{x}_k) \in \Gamma$,则存在 $J_i(\mathbf{x}_k)$ 的上界 $\bar{J}_i(\mathbf{x}_k) \in \Psi$ 和下界 $\bar{J}_i(\mathbf{x}_k) = 0$,使得函数数列 $\{J_i(\mathbf{x}_k)\}$ 随着上下界的迭代收敛而收敛到 $J^*(\mathbf{x}_k)$ 。

2.2 收敛性证明

引理1 若 $J_i(\mathbf{x}_k) \in \Psi$,则 $J_i(\mathbf{x}_k) \geq J^*(\mathbf{x}_k)$ 。

证明 由式(10)可知,当 $J_i(\mathbf{x}_k) \in \Psi$ 时,其相应的容许控制策略 $\mathbf{v}_i(\mathbf{x})$ 是存在的,因此有

$$J_i(\mathbf{x}_k) = \sum_{t=k}^{\infty} U(\mathbf{x}_t, \mathbf{v}_i(\mathbf{x}_t)) \geq \min_{\mathbf{u}(\mathbf{x}_k)} \sum_{t=k}^{\infty} U(\mathbf{x}_t, \mathbf{u}(\mathbf{x}_t)) = J^*(\mathbf{x}_k). \quad (12)$$

引理1证毕。

引理2 若 $J_i(\mathbf{x}_k) \in \Psi$,则 $J_{i+1}(\mathbf{x}_k) \leq J_i(\mathbf{x}_k)$ 。

证明 根据式(8)和(11),得

$$J_i(\mathbf{x}_k) = U(\mathbf{x}_k, \mathbf{v}_i(\mathbf{x}_k)) + J_i(\mathbf{x}_{k+1}) \geq \min_{\mathbf{u}(\mathbf{x}_k)} \{U(\mathbf{x}_k, \mathbf{u}(\mathbf{x}_k)) + J_i(\mathbf{x}_{k+1})\} = J_{i+1}(\mathbf{x}_k). \quad (13)$$

引理2证毕。

引理3 若 $J_i(\mathbf{x}_k) \in \Psi$,则 $J_{i+1}(\mathbf{x}_k) \in \Psi$ 。

证明 由 $J_i(\mathbf{x}_k) \in \Psi$ 可知其相应的容许控制策略 $\mathbf{v}_i(\mathbf{x})$ 是存在的。

1) 根据式(8)和(11), 得

$$J_{i+1}(\mathbf{x}_k) = U(\mathbf{x}_k, \mathbf{u}_i(\mathbf{x}_k)) + J_i(\mathbf{x}_{k+1}) = U(\mathbf{x}_k, \mathbf{u}_i(\mathbf{x}_k)) + \sum_{t=k+1}^{\infty} U(\mathbf{x}_t, \mathbf{v}_i(\mathbf{x}_t)). \quad (14)$$

令

$$\mathbf{v}_{i+1}(\mathbf{x}_t) = \begin{cases} \mathbf{u}_i(\mathbf{x}_t), t = k; \\ \mathbf{v}_i(\mathbf{x}_t), t > k. \end{cases} \quad (15)$$

则式(14)重写为

$$J_{i+1}(\mathbf{x}_k) = \sum_{t=k}^{\infty} U(\mathbf{x}_t, \mathbf{v}_{i+1}(\mathbf{x}_t)). \quad (16)$$

2) 下面证明 $\mathbf{v}_{i+1}(\mathbf{x})$ 是容许控制策略. 根据引理2以及式(8), 得

$$J_i(\mathbf{x}_{k+1}) - J_i(\mathbf{x}_k) \leq J_i(\mathbf{x}_{k+1}) - J_{i+1}(\mathbf{x}_k) = -U(\mathbf{x}_k, \mathbf{u}_i(\mathbf{x}_k)). \quad (17)$$

根据式(10), $J_i(\mathbf{x}_k)$ 是正定函数. 式(17)指出 $J_i(\mathbf{x}_{k+1}) - J_i(\mathbf{x}_k)$ 是负定函数. 因此 $J_i(\mathbf{x}_k)$ 是Lyapunov函数^[20], $\mathbf{u}_i(\mathbf{x})$ 是稳定控制策略. $\mathbf{v}_i(\mathbf{x})$ 是容许控制策略, 因此 $\mathbf{v}_{i+1}(\mathbf{x})$ 是稳定控制策略. 另外, $\mathbf{v}_i(\mathbf{x})$ 是容许控制策略, $\forall \mathbf{x}_k \in \mathbb{R}^n$, $J_i(\mathbf{x}_k)$ 是有限的. 根据引理2, $J_{i+1}(\mathbf{x}_k)$ 也是有限的. 综上所述, $\mathbf{v}_{i+1}(\mathbf{x})$ 满足了定义1的要求, $\mathbf{v}_{i+1}(\mathbf{x})$ 是容许控制策略, $J_{i+1}(\mathbf{x}_k) \in \Psi$.

引理3证毕.

引理1、2和3表明, 若 $J_i(\mathbf{x}_k) \in \Psi$, 则函数数列 $\{J_i(\mathbf{x}_k), J_{i+1}(\mathbf{x}_k), \dots\}$ 单调不增且以 $J^*(\mathbf{x}_k)$ 为下界. 此数列的极限是存在的.

定理1 若 $J_i(\mathbf{x}_k) \in \Psi$, 则 $\lim_{i \rightarrow \infty} J_i(\mathbf{x}_k) = J^*(\mathbf{x}_k)$, $\lim_{i \rightarrow \infty} \mathbf{u}_i(\mathbf{x}_k) = \mathbf{u}^*(\mathbf{x}_k)$.

证明 采用文献[17]的方法进行证明.

设 $\lim_{i \rightarrow \infty} J_i(\mathbf{x}_k) = J_{\infty}(\mathbf{x}_k)$, $\lim_{i \rightarrow \infty} \mathbf{u}_i(\mathbf{x}_k) = \mathbf{u}_{\infty}(\mathbf{x}_k)$. 根据式(8)和引理2, 有

$$J_{\infty}(\mathbf{x}_k) \leq J_{i+1}(\mathbf{x}_k) = U(\mathbf{x}_k, \mathbf{u}_i(\mathbf{x}_k)) + J_i(\mathbf{x}_{k+1}). \quad (18)$$

若 $i \rightarrow \infty$, 则

$$J_{\infty}(\mathbf{x}_k) \leq U(\mathbf{x}_k, \mathbf{u}_{\infty}(\mathbf{x}_k)) + J_{\infty}(\mathbf{x}_{k+1}). \quad (19)$$

另一方面, 根据引理2, 有

$$J_{\infty}(\mathbf{x}_k) \leq J_i(\mathbf{x}_k), \quad (20)$$

于是得

$$J_{\infty}(\mathbf{x}_{k+1}) \leq J_i(\mathbf{x}_{k+1}) \quad (21)$$

以及

$$J_{i+1}(\mathbf{x}_k) = U(\mathbf{x}_k, \mathbf{u}_i(\mathbf{x}_k)) + J_i(\mathbf{x}_{k+1}) \geq U(\mathbf{x}_k, \mathbf{u}_i(\mathbf{x}_k)) + J_{\infty}(\mathbf{x}_{k+1}). \quad (22)$$

若 $i \rightarrow \infty$, 则

$$J_{\infty}(\mathbf{x}_k) \geq U(\mathbf{x}_k, \mathbf{u}_{\infty}(\mathbf{x}_k)) + J_{\infty}(\mathbf{x}_{k+1}). \quad (23)$$

由不等式(19)和(23)可得

$$J_{\infty}(\mathbf{x}_k) = U(\mathbf{x}_k, \mathbf{u}_{\infty}(\mathbf{x}_k)) + J_{\infty}(\mathbf{x}_{k+1}). \quad (24)$$

另外根据式(7), 当 $i \rightarrow \infty$ 时,

$$\mathbf{u}_{\infty}(\mathbf{x}_k) = \arg \min_{\mathbf{u}(\mathbf{x}_k)} \{U(\mathbf{x}_k, \mathbf{u}(\mathbf{x}_k)) + J_{\infty}(\mathbf{x}_{k+1})\}. \quad (25)$$

可见式(24)、(25)与式(3)、(5)一样.

定理1证毕.

下面证明如果 $J_i(\mathbf{x}_k) \in \Gamma$, 则存在 $J_i(\mathbf{x}_k)$ 的上界 $\bar{J}_i(\mathbf{x}_k) \in \Psi$ 和下界 $\bar{J}_i(\mathbf{x}_k) = 0$, 使得函数数列 $\{J_i(\mathbf{x}_k)\}$ 随着上下界的迭代收敛而收敛到 $J^*(\mathbf{x}_k)$. 在给出该证明前, 需要首先证明 $J_i(\mathbf{x}_k)$ 的上界 $\bar{J}_i(\mathbf{x}_k) \in \Psi$ 是存在的.

引理4 $\forall J_i(\mathbf{x}_k) \in \Gamma, \exists S(\mathbf{x}_k) \in \Psi$, 使得 $J_i(\mathbf{x}_k) \leq S(\mathbf{x}_k)$.

证明 1) $\mathbf{x}_k \neq 0$

由于系统(1)是可控的, 因此 $\forall \mathbf{x}_k \neq 0$, 存在有限正整数 η 和控制策略 $\mu(\mathbf{x})$, 使得 $\mathbf{x}_{k+\eta} = 0$. 另一方面, 任取一控制 $\Lambda(0) \neq 0$, 存在 $\mathbf{X} \neq 0$, 使得 $\mathbf{F}(0, \Lambda(0)) = \mathbf{X}$ 成立. 当然, 当 $\mathbf{x}_k = \mathbf{X}$ 时, 存在有限正整数 $\hat{\eta}$ 和控制策略 $\mu(\mathbf{x})$, 使得 $\mathbf{x}_{k+\hat{\eta}} = 0$. 构造一个函数:

$$\sum_{t=0}^p \left\{ U(\mathbf{x}_{k+\eta+t(\hat{\eta})}, \Lambda(\mathbf{x}_{k+\eta+t(\hat{\eta}+1)})) + \sum_{\tau=k+\eta+t(\hat{\eta}+1)+1}^{k+\eta+t(\hat{\eta}+1)\hat{\eta}} U(\mathbf{x}_{\tau}, \mu(\mathbf{x}_{\tau})) \right\}. \quad (26)$$

在式(26)中, $\mathbf{x}_{k+\eta} = 0$, $\mathbf{x}_{k+\eta+1} = \mathbf{X}$, $\mathbf{x}_{k+\eta+\hat{\eta}+1} = 0$, $\mathbf{x}_{k+\eta+\hat{\eta}+2} = \mathbf{X}$, $\mathbf{x}_{k+\eta+2\hat{\eta}+2} = 0, \dots$. 也即是

$$\begin{cases} \mathbf{x}_{k+\eta} = \mathbf{x}_{k+\eta+(t+1)(\hat{\eta}+1)} = 0 \\ \mathbf{x}_{k+\eta+t(\hat{\eta}+1)+1} = \mathbf{X} \end{cases}, t = 0, 1, \dots, p. \quad (27)$$

于是式(26)可重写为

$$H(\mathbf{x}_k, p) = \sum_{t=k}^{k+\eta-1} U(\mathbf{x}_t, \mu(\mathbf{x}_t)) + p \left\{ U(\mathbf{x}_{k+\eta}, \Lambda(\mathbf{x}_{k+\eta})) + \sum_{\tau=k+\eta+1}^{k+\eta+\hat{\eta}} U(\mathbf{x}_{\tau}, \mu(\mathbf{x}_{\tau})) \right\}. \quad (28)$$

$\forall \mathbf{x}_k \neq 0$, 式(28)的每一项均正定且有限. $J_i(\mathbf{x}_k)$ 是半正定且有限的. 存在一个有限正整数 P , 使得式(29)成立:

$$J_i(\mathbf{x}_k) \leq H(\mathbf{x}_k, P). \quad (29)$$

设 $\theta(\mathbf{x}) \in \mathbb{R}^m$ 是零控制策略. 构造函数:

$$S(\mathbf{x}_k) = H(\mathbf{x}_k, P) + \sum_{t=k+\eta+(P+1)(\hat{\eta}+1)}^{\infty} U(\mathbf{x}_t, \theta(\mathbf{x}_t)). \quad (30)$$

根据 $\mathbf{F}(0, 0) = 0$, 由于 $\mathbf{x}_t = 0, t = k + \eta + (P+1)(\hat{\eta}+1)$, 那么 $\theta(\mathbf{x})$ 令 $\mathbf{x}_t = 0, t > k + \eta + (P+1)(\hat{\eta}+1)$. 式(30)第二项等于0. 所以有

$$J_i(\mathbf{x}_k) \leq H(\mathbf{x}_k, P) = S(\mathbf{x}_k). \quad (31)$$

令 $T = k + \eta + p(\bar{\eta} + 1)$, $p = 0, 1, \dots, P$, 并设:

$$\omega(\mathbf{x}_t) = \begin{cases} \mu(\mathbf{x}_t), t < k + \eta + (P + 1)(\bar{\eta} + 1) \text{ and } t \neq T; \\ \Lambda(\mathbf{x}_t), t = T; \\ \theta(\mathbf{x}_t), t \geq k + \eta + (P + 1)(\bar{\eta} + 1). \end{cases} \quad (32)$$

式(30)可重写为

$$S(\mathbf{x}_k) = \sum_{t=k}^{\infty} U(\mathbf{x}_t, \omega(\mathbf{x}_t)). \quad (33)$$

由 $\mathbf{x}_t = 0, t \geq k + \eta + (P + 1)(\bar{\eta} + 1)$ 可知 $\omega(\mathbf{x})$ 是稳定控制策略. 另一方面, 因为 $H(\mathbf{x}_k, P)$ 是有限的, 那么根据式(31)可知 $S(\mathbf{x}_k)$ 也是有限的. $\omega(\mathbf{x})$ 是容许控制策略.

2) $\mathbf{x}_k = 0$

给定一个零控制策略, 并令

$$S(\mathbf{x}_k) = \sum_{t=k}^{\infty} U(\mathbf{x}_t, 0), \quad (34)$$

由 $F(0, 0) = 0$ 可知, 在零控制策略下, 状态保持为零. 于是有

$$J_i(\mathbf{x}_k) = S(\mathbf{x}_k) = 0, \quad (35)$$

因此当 $\mathbf{x}_k = 0$ 时, 零控制策略是容许控制策略.

根据式(31)~(35), 可得到以下结论. $\forall \mathbf{x}_k \in \mathbb{R}^n$, 存在

$$S(\mathbf{x}_k) = \sum_{t=k}^{\infty} U(\mathbf{x}_t, \bar{\omega}(\mathbf{x}_t)), \quad \bar{\omega}(\mathbf{x}_t) = \begin{cases} \omega(\mathbf{x}_t), \mathbf{x}_k \neq 0; \\ 0, \mathbf{x}_k = 0. \end{cases} \quad (36)$$

使得 $J_i(\mathbf{x}_k) \leq S(\mathbf{x}_k)$ 成立. 并且由于 $\bar{\omega}(\mathbf{x})$ 是容许控制策略, 因此 $S(\mathbf{x}_k) \in \Psi$.

引理4证毕.

定理 2 若 $J_i(\mathbf{x}_k) \in \Gamma$, 则 $\lim_{i \rightarrow \infty} u_i(\mathbf{x}_k) = \mathbf{u}^*(\mathbf{x}_k)$,

$$\lim_{i \rightarrow \infty} J_i(\mathbf{x}_k) = J^*(\mathbf{x}_k).$$

证明 设 $\tilde{J}_i(\mathbf{x}_k) = 0$. 根据式(9), $J_i(\mathbf{x}_k)$ 是半正定的, 因此有

$$\tilde{J}_i(\mathbf{x}_k) \leq J_i(\mathbf{x}_k). \quad (37)$$

另外根据引理4, 存在 $\tilde{J}_i(\mathbf{x}_k) \in \Psi$, 使得

$$J_i(\mathbf{x}_k) \leq \tilde{J}_i(\mathbf{x}_k). \quad (38)$$

根据值迭代(7)和(8), 定义两个不同初始的迭代性能指标函数的迭代过程:

$$\begin{cases} \tilde{\mathbf{u}}_i(\mathbf{x}_k) = \arg \min_{\mathbf{u}(\mathbf{x}_k)} \{U(\mathbf{x}_k, \mathbf{u}(\mathbf{x}_k)) + \tilde{J}_i(\mathbf{x}_{k+1})\}; \\ \tilde{J}_{i+1}(\mathbf{x}_k) = U(\mathbf{x}_k, \tilde{\mathbf{u}}_i(\mathbf{x}_k)) + \tilde{J}_i(\mathbf{x}_{k+1}). \end{cases} \quad (39)$$

$$\begin{cases} \bar{\mathbf{u}}_i(\mathbf{x}_k) = \arg \min_{\mathbf{u}(\mathbf{x}_k)} \{U(\mathbf{x}_k, \mathbf{u}(\mathbf{x}_k)) + \tilde{J}_i(\mathbf{x}_{k+1})\}; \\ \bar{J}_{i+1}(\mathbf{x}_k) = U(\mathbf{x}_k, \bar{\mathbf{u}}_i(\mathbf{x}_k)) + \tilde{J}_i(\mathbf{x}_{k+1}). \end{cases} \quad (40)$$

设 $q=0$, 把不等式(37)和(38)合并得

$$\tilde{J}_{i+q}(\mathbf{x}_k) \leq J_{i+q}(\mathbf{x}_k) \leq \bar{J}_{i+q}(\mathbf{x}_k), \quad (41)$$

也可以重写为

$$\tilde{J}_{i+q}(\mathbf{x}_{k+1}) \leq J_{i+q}(\mathbf{x}_{k+1}) \leq \bar{J}_{i+q}(\mathbf{x}_{k+1}). \quad (42)$$

于是可推出

$$\begin{aligned} \tilde{J}_{i+q+1}(\mathbf{x}_k) &= U(\mathbf{x}_k, \bar{\mathbf{u}}_{i+q}(\mathbf{x}_k)) + \tilde{J}_{i+q}(\mathbf{x}_{k+1}) \geq \\ &U(\mathbf{x}_k, \bar{\mathbf{u}}_{i+q}(\mathbf{x}_k)) + J_{i+q}(\mathbf{x}_{k+1}) \geq \\ \min_{\mathbf{u}(\mathbf{x}_k)} \{U(\mathbf{x}_k, \mathbf{u}(\mathbf{x}_k)) + J_{i+q}(\mathbf{x}_{k+1})\} &= J_{i+q+1}(\mathbf{x}_k), \end{aligned} \quad (43)$$

以及

$$\begin{aligned} J_{i+q+1}(\mathbf{x}_k) &= U(\mathbf{x}_k, \mathbf{u}_{i+q}(\mathbf{x}_k)) + J_{i+q}(\mathbf{x}_{k+1}) \geq \\ &U(\mathbf{x}_k, \mathbf{u}_{i+q}(\mathbf{x}_k)) + \tilde{J}_{i+q}(\mathbf{x}_{k+1}) \geq \\ \min_{\mathbf{u}(\mathbf{x}_k)} \{U(\mathbf{x}_k, \mathbf{u}(\mathbf{x}_k)) + \tilde{J}_{i+q}(\mathbf{x}_{k+1})\} &= \tilde{J}_{i+q+1}(\mathbf{x}_k). \end{aligned} \quad (44)$$

将不等式(43)和(44)合并为

$$\tilde{J}_{i+q+1}(\mathbf{x}_k) \leq J_{i+q+1}(\mathbf{x}_k) \leq \bar{J}_{i+q+1}(\mathbf{x}_k). \quad (45)$$

当 $q = 1, 2, \dots$ 时, 不等式(41)~(45)仍然成立. 由数学归纳法得

$$\tilde{J}_{i+q}(\mathbf{x}_k) \leq J_{i+q}(\mathbf{x}_k) \leq \bar{J}_{i+q}(\mathbf{x}_k), \quad q = 0, 1, 2, \dots \quad (46)$$

根据定理1, $\lim_{i \rightarrow \infty} \tilde{J}_i(\mathbf{x}_k) = J^*(\mathbf{x}_k)$. 文献[17]已证明 $\lim_{i \rightarrow \infty} \tilde{J}_i(\mathbf{x}_k) = J^*(\mathbf{x}_k)$. 因此 $\lim_{i \rightarrow \infty} J_i(\mathbf{x}_k) = J^*(\mathbf{x}_k)$, 同时有 $\lim_{i \rightarrow \infty} \mathbf{u}_i(\mathbf{x}_k) = \mathbf{u}^*(\mathbf{x}_k)$.

定理2证毕.

定理2表明, $J_0(\mathbf{x}_k)$ 允许初始化为任意的半正定函数.

3 总结

本文给出了一种证明方法, 通过构造迭代性能指标函数的上下界, 证明了初始化为任意半正定函数的迭代性能指标函数可保证值迭代收敛到最优.

参考文献:

- [1] 张海舰, 成思源, 骆少明, 等. 基于动态规划法的B样条主动轮廓模型[J]. 广东工业大学学报, 2005, 22(4): 26-30.
ZHANG H J, CHENG S Y, LUO S M, et al. B-Spline active contour based on dynamic programming [J]. Journal of Guangdong University of Technology, 2005, 22(4): 26-30.
- [2] BELLMAN R E. Dynamic Programming [M]. Princeton: Princeton University Press, 1957.
- [3] WERBOS P J. Advanced forecasting methods for global crisis warning and models of intelligence [J]. General Systems Yearbook, 1977, 22(6): 25-38.
- [4] MILLER W T, SUTTON R S, WERBOS P J. A menu of designs for reinforcement learning over time, in neural net-

- works for control [M]. Cambridge: MIT Press, 1991.
- [5] WERBOS P J. Approximate dynamic programming for real-time control and neural modeling, in handbook of intelligent control [M]. New York: Van Nostrand Reinhold, 1992.
 - [6] MURRAY J J, COX C J, LENDARIS G G, *et al.* Adaptive dynamic programming [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C, 2002, 32(2): 140-153.
 - [7] PROKHOROV D V, WUNSCH D C. Adaptive critic designs [J]. IEEE Transactions on Neural Networks, 1997, 8(5): 997-1007.
 - [8] BERTSEKAS D P, TSITSIKLIS J N. Neuro-dynamic programming [M]. Belmont: Athena Scientific, 1996.
 - [9] SUTTON R S, BARTO A G. Reinforcement learning: an introduction [M]. Cambridge: The MIT Press, 1998.
 - [10] LEWIS F L, LIU D. Reinforcement learning and adaptive dynamic programming for feedback control [J]. IEEE Circuits & Systems Magazine, 2009, 9(3): 32-50.
 - [11] LEWIS F L, VRABIE D, VAMVOUDAKIS K G. Reinforcement learning and feedback control: using natural decision methods to design optimal adaptive controllers [J]. IEEE Control Systems, 2012, 32(6): 76-105.
 - [12] ABU-KHALAF M, LEWIS F L. Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach [J]. Automatica, 2005, 41(5): 779-791.
 - [13] LIU D, WEI Q. Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems [J]. IEEE Transactions on Neural Networks and Learning Systems, 2014, 25(3): 621-634.
 - [14] WANG F, JIN N, LIU D, WEI Q. Adaptive dynamic programming for finite-horizon optimal control of discrete-time nonlinear systems with ε -error bound [J]. IEEE Transactions on Neural Networks, 2011, 22(1): 24-36.
 - [15] AL-TAMIMI A, LEWIS F L, ABU-KHALAF M. Discrete-time nonlinear HJB solution using approximate dynamic programming: convergence proof [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2008, 38(4): 943-949.
 - [16] ZHANG H, WEI Q, LUO Y. A novel infinite-time optimal tracking control scheme for a class of discrete-time nonlinear systems via the greedy HDP iteration algorithm [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2008, 38(4): 937-942.
 - [17] WANG D, LIU D, WEI Q, *et al.* Optimal control of unknown nonaffine nonlinear discrete-time systems based on adaptive dynamic programming [J]. Automatica, 2012, 48(8): 1825-1832.
 - [18] WEI Q, LIU D. A novel iterative θ -adaptive dynamic programming for discrete-time nonlinear systems [J]. IEEE Transactions on Automation Science and Engineering, 2014, 11(4): 1176-1190.
 - [19] WEI Q, LIU D, LIN H. Value iteration adaptive dynamic programming for optimal control of discrete-time nonlinear systems [J]. IEEE Transactions on Cybernetics, 2016, 46(3): 840-853.
 - [20] LIAO X, WANG L, YU P. Stability of dynamical systems [M]. Amsterdam: Elsevier Press, 2007.