Pure Mathematics 理论数学, 2024, 14(1), 416-432
Published Online January 2024 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2024.141041

Hans汉斯

随机线性二次控制的资格迹方法

朱亚楠

上海理工大学，理学院，上海

收稿日期：2023年12月20日；录用日期：2023年12月29日；发布日期：2024年1月31日

摘　要

本文研究了强化学习方法在线性二次控制问题(LQR)中的应用。在LQR问题的研究中，常见的方法通过求解代数黎卡提方程得到最优控制，并不直接优化控制增益。本文在策略梯度算法的基础上引入资格迹方法，直接优化控制增益矩阵。考虑已知和未知参数两种情况下，资格迹方法的收敛。在有限时域和高斯噪声的条件下，分别给出了已知和未知参数两种情况下算法的全局收敛保证。参数未知时，利用零阶优化定理近似梯度项，这可以将问题扩展至代价函数非凸的情况。数值模拟结果显示资格迹方法与梯度下降算法相比更快收敛，方差更小。

关键词

线性二次最优控制，梯度下降，资格迹

Eligibility Trace Method for Stochastic Linear Quadratic Control

Yanan Zhu

College of Science, University of Shanghai for Science and Technology, Shanghai

Received: Dec. 20th, 2023; accepted: Dec. 29th, 2023; published: Jan. 31st, 2024

Abstract

This paper studies the application of reinforcement learning method to linear quadratic regulator (LQR) problem. For the study of LQR problem, the usual method is to obtain the optimal control by solving the algebraic Riccati equation, but not to optimize the control gain directly. This paper optimizes the control gain directly, proposes the eligibility trace method on the basis of gradient descent algorithm, and produces global convergence guarantee in the case of known and unknown parameters, in the setting of finite time horizon and Gaussian noise. When the parameters are unknown, the zero-order optimization theorem is used to approximate the gradient term, which

文章引用: 朱亚楠. 随机线性二次控制的资格迹方法[J]. 理论数学, 2024, 14(1): 416-432.
DOI: 10.12677/pm.2024.141041

can extend the problem to cases where the cost function is not convex. Numerical simulation results show that the eligibility trace method has faster convergence and smaller variance than gradient descent algorithm.

Keywords

Linear Quadratic Optimal Control, Gradient Descent, Eligibility Traces

1. 引言

线性二次型控制(LQR)问题是最优控制理论中最基本的问题之一，系统下一时刻的状态由当前状态和控制决定，代价是关于状态和控制的二次函数[1]。问题的目标是找到最优的系统控制，使得代价函数达到最小值。在系统参数完全已知的情况下，LQR 问题已经有广泛的研究，最优控制存在解析表达式[2]，被广泛应用在不同领域中。在实际应用中，获取系统所有参数时不现实的，当系统参数未知时，产生许多不确定性，无法直接求解得到最优控制，这促进了对学习方法的探索。

近年来使用强化学习和数据驱动方法解决不确定动力系统问题取得重要进展，强化学习方法成为解决此类不确定问题的有效手段。强化学习(RL) [3]是交互式的学习方法，通过与环境交互积累经验(采样)，以最大化数值收益信号为导向，不断从经验中学习，最终得到最优策略(控制)。在已有的研究中，利用 RL 方法解决参数未知的 LQR 问题，有基于模型的方法[4] [5] [6]和无模型的方法[7] [8] [9]两种。基于模型的 RL 方法利用经验数据估计系统模型参数，将参数估计值视为真实值进而通过解析表达式求解最优控制[10] [11]; 无模型的 RL 方法直接与系统环境交互，利用经验数据直接学习最优控制而不估计参数[12] [13]。LQR 问题的最优控制是状态的线性函数，在学习方法的框架下，相当于优化矩阵形式的参数。自然的想法是在参数空间中使用梯度下降算法，将控制参数化，即 RL 中的策略梯度算法。这种方法基于代价函数对控制参数的梯度，对参数进行优化。有模型的 RL 方法在模型结构已知时具有更好的表现，无模型方法则更少依赖于模型假设，具有更好的泛化性能[14]。

RL 方法在实践中效果突出，但理论理解较少，相比之下，控制理论已经有较为成熟的研究成果，具有可证明的保证。系统参数已知时，控制优化过程中需要的梯度项有明确的表达式，易于计算；参数未知时，用经验数据对梯度项进行近似，现有的梯度近似有零阶近似、二阶近似以及样条逼近等方法。Dhruv Malik 等人给出了零阶近似方法在线性控制问题中的收敛保证[15]。Maryam Fazel 等人研究了初始状态有噪声干扰的情况下，不同梯度优化方的全局收敛性[16]。Ben Hambly 等人在有限时域和随机状态转移的设置下为零阶近似方法提供了全局收敛保证[17]。Mesbahi A 等人将结果推广至连续时间的 LQR 问题[18] [19]。

LQR 问题描述

有限时域 T 下的随机离散时间 LQR 问题，系统方程为：

$$x_{t+1} = Ax_t + Bu_t + \omega_t \tag{1}$$

其中，$x_t \in \mathbb{R}^d$ 和 $u_t \in \mathbb{R}^k$ 分别是系统在 t 时刻的状态向量和控制向量，x_0 从分布 $\mathcal{D}$ 中随机抽样，$\omega_t \left(t \in [T] = 0,1,\cdots,T-1\right)$ 是独立同分布的零均值系统噪声序列，与 x_0 独立。$A \in \mathbb{R}^{d\times d}$ 和 $B \in \mathbb{R}^{d\times k}$ 是系统转

移矩阵。系统(1)对应的代价函数为：

$$C(\mathbf{K})=E\left[\sum_{t=0}^{T-1}\left(x_t^T Q_t x_t+u_t^T R_t u_t\right)+x_T^T Q_T x_T\right] \tag{2}$$

其中，$Q_t\in\mathbb{R}^{d\times d}$ 是对称正定矩阵，$R_t\in\mathbb{R}^{k\times k}$ 是正定矩阵，$\mathbf{K}=\left(K_0,K_1,\cdots,K_{T-1}\right)$ 是控制序列。

离散时间 LQR 问题的目标是在可行控制空间 $\mathcal{U}$ 中找到使代价函数值最小的最优控制，在系统参数已知时，有限时域的离散时间 LQR 问题的最优控制有显式解，当 (A,B) 可控时，根据最优控制理论[20]，t 时刻的最优控制为当前状态的线性函数：

$$\begin{aligned}&u_t^*=-K_t^* x_t\\&K_t^*=\left(B^T P_{t+1}^* B+R_t\right)^{-1}B^T P_{t+1}^* A\end{aligned}\qquad \forall t\in[T] \tag{3}$$

其中，$K_t^*\in\mathbb{R}^{k\times d}$ 是 t 时刻的最优控制，P_t^* 是以下代数黎卡提方程的解：

$$P_t^*=Q_t+A^T P_{t+1}^* A-A^T P_{t+1}^* B\left(B^T P_{t+1}^* B+R_t\right)^{-1}B^T P_{t+1}^* A \tag{4}$$

黎卡提方程的求解通常令 $P_T=Q_T$，通过递推式(6)进行迭代求解，直到满足精度要求。此时，代价函数达到最小值

$$C\left(\mathbf{K}^*\right)=E\left[\sum_{t=0}^{T-1}x_t^T P_t^* x_t\right].$$

上述通过求解代数黎卡提方程进而得到最优控制的方法有以下缺点：1) 并不直接优化控制增益 K_t，也不直接优化代价函数；2) 在求解黎卡提方程时涉及高维矩阵的求逆运算，消耗的计算量在系统维度高时过大；3) 当系统参数 $\theta=\left(A,B,\{Q_t\}_{t=0}^{T-1},\{R_t\}_{t=0}^{T-1}\right)$ 未知时，这种方法并不适用。

本文研究有限时域下的离散时间随机线性二次控制(DLQR)问题，在策略梯度方法中引入资格迹方法，不求解代数黎卡提方程，直接优化控制增益。资格迹方法是强化学习的基本方法之一，在优化中引入与控制参数向量同维度的资格迹向量，衡量系统控制参数每个分量的重要性，在学习过程中影响参数向量的更新[3]。这种方法不仅适用于分幕式的情况，也适用于持续性问题。在实际效果方面，资格迹方法比策略梯度算法更快收敛，且方差更小，数值模拟结果证明了这一结论。

2. 理论

本文考虑折现的随机 DLQR 问题：

$$\begin{aligned}\min_{\mathbf{K}}\quad & C(\mathbf{K})=E\left[\sum_{t=0}^{T-1}\gamma^t c_t\left(x_t,u_t\right)+\gamma^T c_T\left(x_T\right)\right]\\\text{s.t.}\quad & x_{t+1}=Ax_t+Bu_t+\omega_t\end{aligned} \tag{5}$$

其中 $\gamma\in(0,1)$ 是折现因子，单步的代价 $c_t\left(x_t,u_t\right)=x_t^T Q_t x_t+u_t^T R_t u_t, t\in[T]$，$c_T\left(x_T\right)=x_T^T Q_T x_T$。控制增益遵循以下法则优化：

$$\begin{aligned}&\mathbf{K}^{n+1}=\mathbf{K}^n-\alpha\boldsymbol{\delta}^n\\&\boldsymbol{\delta}^0=\nabla C\left(\mathbf{K}^0\right)\\&\boldsymbol{\delta}^n=\lambda\boldsymbol{\delta}^{n-1}+\nabla C\left(\mathbf{K}^n\right),n>0\end{aligned}\qquad \forall t\in[T] \tag{6}$$

其中，$n\in[N]$ 是迭代次数，α 是步长参数，λ 是衰减参数，$\mathbf{K}^n=\left(K_0^n,K_1^n,\cdots,K_{T-1}^n\right)$ 是第 n 次迭代时的控制序列，$\boldsymbol{\delta}^n=\left(\delta_0^n,\delta_1^n,\cdots,\delta_{T-1}^n\right)$ 是与之对应的资格迹序列。

引理 2.1 在策略参数的第 n 次迭代中，定义状态向量的协方差矩阵为：

$$\Sigma_t = E\left[x_t x_t^T\right],\ \Sigma_{\mathbf{K}} = \sum_{t=0}^{T-1}\Sigma_t \tag{7}$$

则与 K_t 对应的资格迹由以下表达式给出：

$$\delta_t^n = \begin{cases} 2\gamma^t E_t^n \Sigma_t^n & n=0 \\ \lambda\delta_t^{n-1} + 2\gamma^t E_t^n \Sigma_t^n & n>0 \end{cases}, \forall t \in [T] \tag{8}$$

其中，$E_t^n = \left(R_t + \gamma B^T P_{t+1}^n B\right)K_t - \gamma B^T P_{t+1}^n A$。

证明：根据资格迹的定义，

$$\delta_t^n = \lambda\delta_t^{n-1} + \nabla_t C\left(\mathbf{K}^n\right), n>0 \tag{9}$$

随机 DLQR 问题(5)中，代价函数可拆分为初始状态项与噪声项：

$$\begin{aligned} C(\mathbf{K}) &= E\left[\sum_{t=0}^{T-1}\gamma^t c_t\left(x_t,u_t\right) + \gamma^T c_T\left(x_T\right)\right] \\ &= E\left[\sum_{t=0}^{T-2}\gamma^t c_t\left(x_t,u_t\right) + \gamma^{T-1}c_{T-1}\left(x_{T-1},-K_{T-1}x_{T-1}\right) + \gamma^T c_T\left(x_T\right)\right] \\ &= E\left[\sum_{t=0}^{T-2}\gamma^t c_t\left(x_t,u_t\right) + \gamma^{T-1}x_{T-1}^T P_{T-1}x_{T-1} + \gamma^T \omega_{T-1}^T P_T \omega_{T-1}\right] \\ &= E\left[x_0^T P_0 x_0 + \sum_{t=0}^{T-1}\gamma^{t+1}\omega_t^T P_{t+1}\omega_t\right] \end{aligned} \tag{10}$$

上式对 K_t 求偏导，

$$\begin{aligned} \nabla_t C(\mathbf{K}) &= \frac{\partial C(\mathbf{K})}{\partial K_t} \\ &= \frac{\partial E\left[\gamma^t x_t^T\left(Q_t + K_t^T R_t K_t + \gamma\left(A-BK_t\right)^T P_{t+1}\left(A-BK_t\right)\right)x_t + \mathbf{K}(-t)\right]}{\partial K_t} \\ &= E\left[2\gamma^t R_t K_t x_t x_t^T - 2\gamma^{t+1}B^T P_{t+1}\left(A-BK_t\right)x_t x_t^T\right] \\ &= 2\gamma^t E_t \Sigma_t \end{aligned} \tag{11}$$

其中，$\mathbf{K}(-t) = \sum_{j=0}^{t-1}\gamma^j c_j\left(x_j,u_j\right) + \sum_{j=t+1}^{T-1}\gamma^j \omega_j^T P_{j+1}\omega_j$ 证毕。

对衰减参数 λ 的不同设置，是衡量历史信息对下一步决策的重要程度，当 $\lambda=0$ 时，不考虑历史信息，当 $\lambda=1$ 时，则认为历史信息与当前信息同样重要。资格迹方法考虑了当前的损失函数和历史策略梯度的关系，而梯度下降算法只考虑梯度更新的平滑度。资格迹方法能够减少参数更新过程中忽略影响较大的历史动量而造成的错误决策次数。

2.1. 参数已知的资格迹方法

本节讨论系统参数 θ 完全已知时的资格迹方法[3]。资格迹方法结合了蒙特卡洛方法和时序差分方法。算法的核心是定义一个维度与策略参数相同的短时记忆向量 $\boldsymbol{\delta}$，衡量策略参数分量的重要性。随着迭代次数的增加，和参与更新的控制参数分量对应的资格迹分量逐渐衰减，直到这一分量再次参与更新。本文用资格迹代替参数更新中的梯度项，增强了算法的泛化性能。

在分析算法收敛性之前，首先研究不同控制序列下产生的代价函数的差异以及状态协方差矩阵间的差异上界，定义$\|S\|$为矩阵 S 的谱范数。

引理 2.2 [17]假设$\mathbf{K}'$与$\mathbf{K}$产生的代价函数均有界，$\{x_t\}_{t=0}^{T-1}$，$\{u_t\}_{t=0}^{T-1}$，$\{x_t'\}_{t=0}^{T-1}$，$\{u_t'\}_{t=0}^{T-1}$分别是由$\mathbf{K}'$，$\mathbf{K}$生成的序列，令$x_0' = x_0 = x$，则代价差可表示为：

$$\begin{aligned} C(\mathbf{K}') - C(\mathbf{K}) = & E\left[\sum_{t=0}^{T-1} 2Tr\left(x_t'(x_t')^T (K_t' - K_t)^T E_t\right)\right] \\ & + E\left[\sum_{t=0}^{T-1} Tr\left(x_t'(x_t')^T (K_t' - K_t)^T \left(R_t + \gamma B^T P_{t+1} B\right)(K_t' - K_t)\right)\right] \end{aligned} \tag{12}$$

其中，P_t是下述方程的解：

$$P_t = Q_t + K_t^T R_t K_t + \gamma (A - BK_t)^T P_{t+1}(A - BK_t),\ \forall t \in [T] \tag{13}$$

引理 2.3 令$\rho = \max\left\{\max_t \|A - BK_t\|, \max_t \|A - BK_t'\|\right\}$，$\Delta = K_t - K_t'$，$\mathbf{K}'$与$\mathbf{K}$是任意策略，系统状态向量的协方差满足下面的关系：

$$\|\Sigma_{\mathbf{K}} - \Sigma_{\mathbf{K}'}\| \le \frac{\rho^{2T} - 1}{\rho^2 - 1}\left(2\rho\|B\|\sum_{t=0}^{T-1}\|\Delta\| + \|B\|^2\sum_{t=0}^{T-1}\|\Delta\|^2\right)\left(\frac{C(\mathbf{K})}{\sigma_{\min}\mathbf{Q}} + T\|W\|\right) \tag{14}$$

证明详见附录。

上面的分析为收敛保证奠定了基础，证明算法的收敛性之前，引理 2.4 的论证了控制序列经过一次迭代后对代价函数值的影响。

引理 2.4 设$\mathbf{K}^*$是最优至序列，$\mathbf{K}'$由$\mathbf{K}$经一次迭代得到，当

$$\alpha \le \min\left\{\alpha_1, \frac{\sigma_{\min}\mathbf{Q}}{2C(\mathbf{K})T\max_t\left\|R_t + \gamma B^T P_{t+1} B\right\|}\right\}$$

其中，

$$\alpha_1 = \frac{\rho^2 - 1}{2\left(\rho^{2T} - 1\right)(2\rho + 1)\|B\|} \cdot \frac{\sigma_{\min}\mathbf{Q}\sigma_{\min}\Sigma_{\mathbf{K}}}{C(\mathbf{K}) + T\|W\|\sigma_{\min}\mathbf{Q}} \cdot \frac{1}{\max_t\|\delta_t\|}$$

则

$$C(\mathbf{K}') - C(\mathbf{K}^*) \le \left(1 - \frac{4\alpha\sigma_{\min}\mathbf{R}\sigma_{\min^2}\Sigma_{\mathbf{K}}}{\left\|\Sigma_{\mathbf{K}^*}\right\|}\right)C(\mathbf{K}) - C(\mathbf{K}^*) \tag{15}$$

证明详见附录。

经过以上分析，下面给出参数已知时，资格迹算法在 DLQR 问题中的全局收敛性保证。

定理 2.4 (收敛性定理)假设$C(\mathbf{K}^0)$有界，步长α满足引理 2.4 的约束，对$\forall\varepsilon > 0$，当迭代次数 N 满足下述条件：

$$N \ge \frac{\left\|\Sigma_{\mathbf{K}^*}\right\|}{4\alpha\sigma_{\min^2}\Sigma_{\mathbf{K}}\sigma_{\min}\mathbf{R}}\log\frac{C(\mathbf{K}^0) - C(\mathbf{K}^*)}{\varepsilon}$$

代价函数值收敛至最优值，即：

$$C(\mathbf{K}^N) - C(\mathbf{K}^*) \le \varepsilon \tag{16}$$

证明：令$\mathbf{K}^1=\mathbf{K}^0-\alpha\boldsymbol{\delta}^0$，根据引理 2.4 的结论，

$$C\left(\mathbf{K}^1\right)-C\left(\mathbf{K}^*\right)\le\left(1-\frac{4\alpha\sigma_{\min}\mathbf{R}\sigma_{\min^2}\Sigma_{\mathbf{K}}}{\left\|\Sigma_{\mathbf{K}^*}\right\|}\right)\left(C\left(\mathbf{K}^0\right)-C\left(\mathbf{K}^*\right)\right)$$

假设经$n+1$次迭代后，$C\left(\mathbf{K}^{n+1}\right)\le C\left(\mathbf{K}^0\right)$，此时$K_t^{n+1}=K_t^n-\alpha\delta_t^n$，根据 Cauchy-Schwarz 不等式，

$$\begin{aligned}\sum_{t=0}^{T-1}\delta_t^n&=\sum_{t=0}^{T-1}\sum_{i=0}^{n}\lambda^{n-i}\nabla_t C\left(\mathbf{K}^n\right)\le\sum_{t=0}^{T-1}\sqrt{n\sum_{i=0}^{n}\left\|\nabla_t C\left(\mathbf{K}^n\right)\right\|^2}\\&\le\sum_{t=0}^{T-1}\sqrt{4n\sum_{i=0}^{n}Tr\left(\Sigma_t^i\left(E_t^i\right)^T E_t^i\Sigma_t^i\right)}\\&\le\sqrt{T\cdot\sum_{t=0}^{T-1}4n\sum_{i=0}^{n}\left\|\Sigma_t^i\right\|^2 Tr\left(\left(E_t^i\right)^T E_t^i\right)}\\&\le\left(\frac{2C\left(\mathbf{K}\right)}{\sigma_{\min}\mathbf{Q}}\right)\sqrt{nT\frac{\max\limits_t\left(R_t+\gamma B^T P_t B\right)}{\sigma_{\min}\Sigma_{\mathbf{K}}}\left(C\left(\mathbf{K}\right)-C\left(\mathbf{K}^*\right)\right)}\end{aligned}$$

最后一个不等式成立基于下面的结果：

$$\begin{aligned}C\left(\mathbf{K}\right)-C\left(\mathbf{K}^*\right)&\ge C\left(\mathbf{K}\right)-C\left(\mathbf{K}'\right)\\&=E\left[\sum_{t=0}^{T-1}Tr\left(E_t^T\left(R_t+\gamma B^T P_{t+1}B\right)^{-1}E_t\right)\right]\\&\ge\frac{\sigma_{\min}\Sigma_{\mathbf{K}}}{\max\limits_t\left(R_t+\gamma B^T P_t B\right)}\sum_{t=0}^{T-1}Tr\left(E_t^T E_t\right)\end{aligned}$$

结合引理 2.3 的分析，引理 2.3 中的结论仍然成立，即：

$$C\left(\mathbf{K}^{n+1}\right)-C\left(\mathbf{K}^*\right)\le\left(1-\frac{4\alpha\sigma_{\min}\mathbf{R}\sigma_{\min}^2\Sigma_{\mathbf{K}}}{\left\|\Sigma_{\mathbf{K}^*}\right\|}\right)\left(C\left(\mathbf{K}^n\right)-C\left(\mathbf{K}^*\right)\right)\tag{17}$$

将$n+1$次的结果进行累积，

$$C\left(\mathbf{K}^{n+1}\right)-C\left(\mathbf{K}^*\right)\le\left(1-\frac{4\alpha\sigma_{\min}\mathbf{R}\sigma_{\min}^2\Sigma_{\mathbf{K}}}{\left\|\Sigma_{\mathbf{K}^*}\right\|}\right)^{n+1}\left(C\left(\mathbf{K}^0\right)-C\left(\mathbf{K}^*\right)\right)$$

对$\forall\varepsilon>0$，当

$$N\ge\frac{\left\|\Sigma_{\mathbf{K}^*}\right\|}{4\alpha\sigma_{\min}^2\Sigma_{\mathbf{K}}\sigma_{\min}\mathbf{R}}\log\frac{C\left(\mathbf{K}^0\right)-C\left(\mathbf{K}^*\right)}{\varepsilon}$$

时，$C\left(\mathbf{K}^N\right)-C\left(\mathbf{K}^*\right)<\varepsilon$。证毕。

2.2. 参数未知的资格迹方法

当系统参数θ未知时，式(13)中对梯度项的表达无效，在控制增益的更新中，性能指标的资格迹无法以解析式精确表达，只能对模型进行仿真得到训练数据。本文使用零阶优化方法[6]近似资格迹，无无需估计系统参数，直接利用目标函数的值优化控制增益。并且零阶优化方法对目标函数的凸性没有要求，直接用函数值估计函数梯度，这大大降低了算法的复杂度。本节首先分析零阶优化方法在随机最优化问题上的拓展[11] [18]，在 DLQR 问题中，参数未知时，在每一步的控制上加入随机噪声进行采样来估计

代价函数值。目标函数可表示为

$$C(\mathbf{K})=E_{\omega}\left[C(\mathbf{K};\omega)\right] \tag{18}$$

这里利用带噪声的代价函数值构造梯度的近似无偏估计。令 $\mathbb{U}^r=\left\{\mathbf{U}\in\mathbb{R}^{k\times d}:\|\mathbf{U}\|_F=r\right\}$，设 $\mathcal{P}_{\mathbb{U}}$ 是 $\mathbb{U}^r$ 上的均匀分布。任意给一个度量 $r>0$，以及 $\mathbf{U}\sim\mathcal{P}_{\mathbb{U}}$ 与 ω 独立，则零阶优化方法对 $C(\mathbf{K})$ 的梯度估计为：

$$\nabla C(\mathbf{K})=\frac{k\times d}{r}C(\mathbf{K}+r\mathbf{U})\mathbf{U} \tag{19}$$

随着 r 越来越小，近似值越来越精确[11]，但 r 过小可能导致估计值有较大的方差。表 1 给出了算法的框架。

Table 1. Eligibility track-gradient algorithm for DLQR problem
表 1. DLQR 的资格迹－梯度算法

算法：资格迹－梯度算法
1. 输入：$\mathbf{K}$，最大迭代次数 M，采样轨迹数 N，幕长 T，随机项参数 r，维度 D
2. for $n=0,1,\cdots,N-1$：
3. for $i=0,1,\cdots,I-1$：
4. for $t=0,1,\cdots,T-1$：
1) 从 $x_0^i\in\mathcal{D}$ 开始，根据 $\left(\mathbf{K}_{-t},\hat{K}_t^i\right)=\left(K_0,\cdots,K_{t-1},\hat{K}_t^i,K_{t+1},\cdots,K_{T-1}\right)$ 采样，其中 $\hat{K}_t^i=K_t+U_t^{ni}\quad\left\|U_t^{ni}\right\|_F=r$．
2) 记录单幕代价 $\hat{c}_t^i$．
5. 计算资格迹的估计值：
$$\hat{\delta}_t^n=\begin{cases}\frac{1}{I}\sum_{i=0}^{I-1}\frac{D}{r^2}\hat{c}_t^iU_t^i & n=0\\ \lambda\hat{\delta}_t^{n-1}+\frac{1}{I}\sum_{i=0}^{I-1}\frac{D}{r^2}\hat{c}_t^iU_t^i & n>0\end{cases}$$
6. $\mathbf{K}^{n+1}=\mathbf{K}^n-\alpha\hat{\boldsymbol{\delta}}^n$

引理 2.5 对给定的 $r>0$ 以及从 $\upsilon^r=\left\{\mathbf{U}\in\mathbb{R}^d:\|\mathbf{U}\|_F=r\right\}$ 中随机抽取的随机向量 $\mathbf{U}$，I 为采样幕数，λ 是折扣系数，资格迹的经验近似为：

$$\hat{\delta}_t^n=\begin{cases}\frac{1}{I}\sum_{i=0}^{I-1}\frac{D}{r^2}\hat{c}_t^iU_t^i & n=0\\ \lambda\hat{\delta}_t^{n-1}+\frac{1}{I}\sum_{i=0}^{I-1}\frac{D}{r^2}\hat{c}_t^iU_t^i & n>0\end{cases} \tag{20}$$

其中，

$$\hat{c}_t^i=\sum_{t=0}^{T-1}\gamma^t\left(\left(x_t^i\right)^TQ_tx_t^i+\left(u_t^i\right)^TR_tu_t^i\right)+\left(x_T^i\right)^TQ_Tx_T^i$$

引理 2.6 假设任意不同控制 $\mathbf{K}'$ 与 $\mathbf{K}$ 的分量满足：

$$\|K_t'-K_t\|\le\min\left\{\|K_t\|,\frac{\left(\rho^2-1\right)\sigma_{\min}\mathbf{Q}\sigma_{\min}\Sigma_{\mathbf{K}}}{2T\left(\rho^{2T}-1\right)(2\rho+1)\left(C(\mathbf{K})+\sigma_{\min}\mathbf{Q}T\|W\|\right)\|B\|}\right\} \tag{21}$$

则存在

$$h_C \le \left\{ \frac{\rho^2-1}{(2\rho+1)\left(\rho^{2T}-1\right)} \cdot \frac{1}{\|B\|} \cdot \frac{1}{\|W\|} \cdot \frac{1}{C\left(\mathbf{K}^0\right)} \right\},\ h_g \le \left\{ \frac{\rho^2-1}{(2\rho+1)\left(\rho^{2T}-1\right)} \cdot \frac{1}{\|B\|} \cdot \frac{1}{\|W\|} \cdot \frac{\sigma_{\min}\mathbf{Q}}{C\left(\mathbf{K}^0\right)} \frac{1}{\left\|\Sigma_{\mathbf{K}^0}\right\|} \right\}$$

使得，

$$\left|C(\mathbf{K}')-C(\mathbf{K})\right| \le h_C \sum_{t=0}^{T-1} \left\|K_t'-K_t\right\|,\ \left\|\nabla_t C(\mathbf{K}')-\nabla_t C(\mathbf{K})\right\| \le h_g \sum_{t=0}^{T-1} \left\|K_t'-K_t\right\|$$

证明详见附录。

引理 2.7 对 $\forall \varepsilon > 0$，当 $r \le \frac{\varepsilon}{4h_g}$ 时，资格迹的经验近似满足：

$$\left\|\hat{\delta}_t^n - \delta_t^n\right\|_F \le \varepsilon,\ \forall t \in [T]$$

证明详见附录。

定理 2.8 (收敛性定理)假设 $C\left(\mathbf{K}^0\right)$ 有界，步长 α 满足引理 2.3 的约束，对 $\forall \varepsilon > 0$，当迭代次数 N 满足

$$N \ge \frac{\left\|\Sigma_{\mathbf{K}^*}\right\|}{2\alpha\sigma_{\min^2}\Sigma_{\mathbf{K}}\sigma_{\min}\mathbf{R}} \log \frac{C\left(\mathbf{K}^0\right)-C\left(\mathbf{K}^*\right)}{\varepsilon}$$

代价函数值收敛至最优值，即：

$$C\left(\mathbf{K}^N\right)-C\left(\mathbf{K}^*\right) \le \varepsilon \tag{22}$$

证明：令 $K_t' = K_t - \alpha\delta_t$，根据引理 2.4，有

$$C(\mathbf{K}')-C\left(\mathbf{K}^*\right) \le \left(1-\frac{4\alpha\sigma_{\min}\mathbf{R}\sigma_{\min}^2\Sigma_{\mathbf{K}}}{\left\|\Sigma_{\mathbf{K}^*}\right\|}\right)\left(C(\mathbf{K})-C\left(\mathbf{K}^*\right)\right) \tag{23}$$

令 $\hat{K}_t' = K_t - \alpha\hat{\delta}_t$，当样本数足够多时，只要 $C(\mathbf{K})-C\left(\mathbf{K}^*\right) \ge \varepsilon$ 时，

$$C\left(\hat{\mathbf{K}}'\right)-C\left(\mathbf{K}^*\right) \le \left(1-\frac{4\alpha\sigma_{\min}\mathbf{R}\sigma_{\min}^2\Sigma_{\mathbf{K}}}{\left\|\Sigma_{\mathbf{K}^*}\right\|}\right)\left(C(\mathbf{K})-C\left(\mathbf{K}^*\right)\right) \tag{24}$$

因为

$$C\left(\hat{\mathbf{K}}'\right)-C\left(\mathbf{K}^*\right) = C\left(\hat{\mathbf{K}}'\right)-C(\mathbf{K}')+C(\mathbf{K}')-C\left(\mathbf{K}^*\right)$$

首先讨论右式第一项，根据引理 2.6，当 $\forall t \in T$，若

$$\left\|\hat{K}_t'-K_t'\right\| \le \frac{\alpha\sigma_{\min}\mathbf{R}\sigma_{\min}^2\Sigma_{\mathbf{K}}}{T\left\|\Sigma_{\mathbf{K}^*}\right\|h_C}\varepsilon$$

成立，则有

$$\left|C\left(\hat{\mathbf{K}}'\right)-C(\mathbf{K}')\right| \le \frac{\alpha\sigma_{\min}\mathbf{R}\sigma_{\min}^2\Sigma_{\mathbf{K}}}{\left\|\Sigma_{\mathbf{K}^*}\right\|}\varepsilon$$

因为 $\hat{K}_t'-K_t' = \alpha\left(\delta_t-\hat{\delta}_t\right)$，所以只需证明

$$\left\|\delta_t-\hat{\delta}_t\right\| \le \frac{\sigma_{\min}\mathbf{R}\sigma_{\min}^2\Sigma_{\mathbf{K}}}{T\left\|\Sigma_{\mathbf{K}^*}\right\|h_C}\varepsilon \tag{25}$$

根据引理 2.7，上式成立。综上所述

$$C\left(\hat{\mathbf{K}}'\right)-C\left(\mathbf{K}^*\right)\leq\left(1-\frac{3\alpha\sigma_{\min}\mathbf{R}\sigma_{\min}^2\Sigma_{\mathbf{K}}}{\left\|\Sigma_{\mathbf{K}^*}\right\|}\right)\left(C(\mathbf{K})-C\left(\mathbf{K}^*\right)\right) \tag{26}$$

基于以上结论，与定理 2.4 相同的证明保证了收敛性。

3. 数值模拟

当状态维度 $d=2$ 时，设定系统参数为

$$A=\begin{pmatrix}0.9 & -1.1\\ 1.0 & 0.8\end{pmatrix}, B=\begin{pmatrix}0.8\\ 1\end{pmatrix}, Q=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}, R=1$$

折现因子 $\gamma=0.99$，比较资格迹方法与梯度下降算法的收敛情况。在折扣系数 $\lambda=0.1$ 的条件下，设定指数衰减的步长参数，分别令时域 $T=20$，迭代次数 $N=40$；时域 $T=50$，迭代次数 $N=70$，代价函数的收敛情况见见图 1 和图 2：

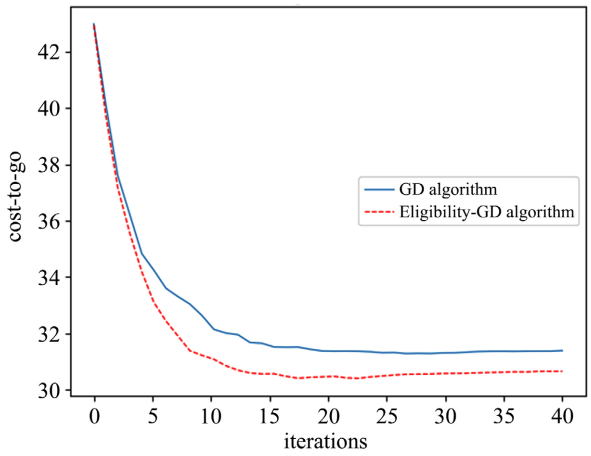

Figure 1. The convergence of $C(K)$ when $d=2$, $T=20$
图 1. $d=2$，$T=20$，代价函数的收敛情况

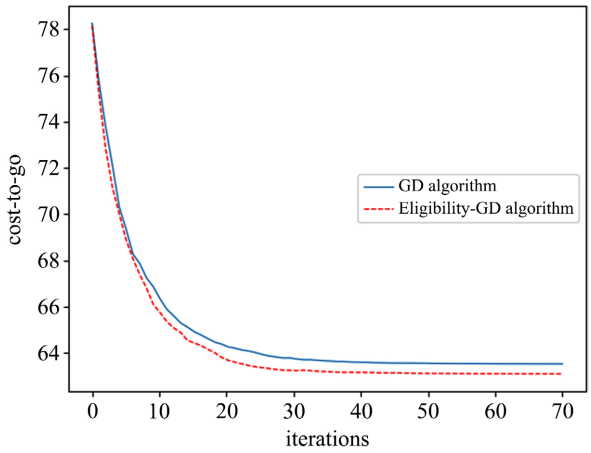

Figure 2. The convergence of $C(K)$ when $d=2$, $T=50$
图 2. $d=2$，$T=50$，代价函数的收敛情况

图 1 和图 2 的结果说明，资格迹算法比梯度下降算法具有更快的收敛速度。资格迹方法中折扣系数的取值对最终结果有显著影响，图 3 展示了 $T=50$，$N=70$ 时不同的折扣系数对算法性能的影响：

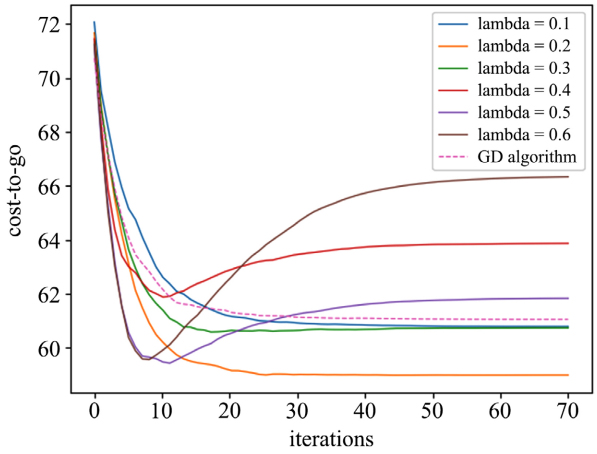

Figure 3. The convergence of $C(K)$ with different λ
图 3. 不同 λ，代价函数的收敛情况

在本节设定的系统参数下，图 3 的结果显示，折扣系数 $\lambda < 0.3$ 时，资格迹算法表现优于策略梯度算法，当 $\lambda > 0.3$ 后，结果出现不收敛的情况，随 λ 的增大，结果偏离越大。λ 是过去梯度信息的权重，说明在这一数值范例中，过去梯度信息对问题求解只能提供少量信息。

当状态维度 $d = 4$ 时，设定系统参数为

$$A = \begin{pmatrix} 0.3 & 0.3 & 0.1 & 0.2 \\ 0.2 & 0.2 & 0.3 & 0.1 \\ 0.3 & 0.1 & 0.3 & 0.4 \\ 0.3 & 0.2 & 0.15 & 0.1 \end{pmatrix}, B = \begin{pmatrix} 0.4 & 0.3 \\ -0.5 & 0.1 \\ 1.0 & 0.2 \\ 0.2 & 0.9 \end{pmatrix}$$

$$Q = \begin{pmatrix} 0.9 & 0.2 & -0.5 & 0.15 \\ 0.2 & 1.1 & 0.15 & 0.1 \\ -0.5 & 0.15 & 0.9 & -0.8 \\ 0.15 & 0.1 & -0.8 & 0.88 \end{pmatrix}, R = \begin{pmatrix} 0.9 & -0.5 \\ -0.5 & 0.8 \end{pmatrix}$$

设定 $T = 50$，$N = 100$，其余参数与二维系统相同，代价函数收敛情况如图 4 所示：

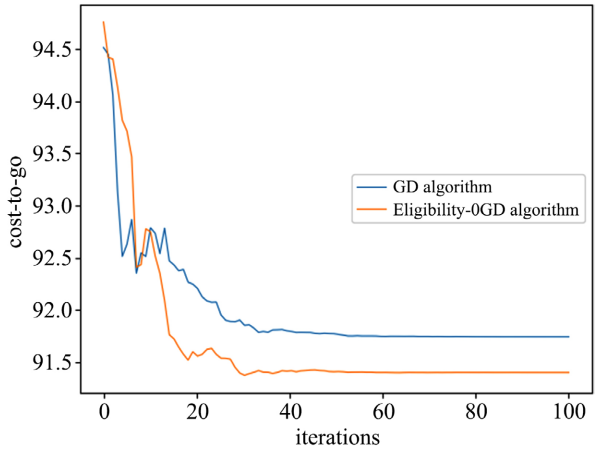

Figure 4. The convergence of $C(K)$ when $d = 4$, $T = 50$
图 4. $d = 4$，$T = 50$，代价函数的收敛情况

上述数值结果说明资格迹方法的收敛效果优于传统的资格梯度方法，并且在算法初期方差更小。

4. 结论

本文研究了无模型强化学习方法在有限时域离散 LQR 问题中的应用，不同于通过解代数黎卡提方程得到最优控制的方法，本文直接优化控制增益，在梯度下降算法的基础上引入资格迹方法，并给出在参数已知和参数未知两种情况下算法的收敛保证。在初始代价函数有界的条件下，算法可以扩展至无限时域。数值模拟验证了算法的收敛性，展示了不同参数设置对结果的影响。另一个方向是基于有模型的强化学习方法，在更少样本量的基础上，进一步达到更好的收敛结果。

致　谢

感谢张老师在论文写作过程中给出的指导和建议。

参考文献

[1] Birge, J. and Louveaux, F. (2011) Introduction to Stochastic Programming. Springer Science & Business Media, Heidelberg. https://doi.org/10.1007/978-1-4614-0237-4

[2] Kučera, V. (1992) Optimal Control: Linear Quadratic Methods: Brian D. O. Anderson and John B. Moore. *Automatica*, **28**, 1068-1069. https://doi.org/10.1016/0005-1098(92)90166-D

[3] Sutton, R.S. and Barto, A.G. (2018) Reinforcement Learning: An Introduction. 2nd ed., the MIT Press, Cambridge.

[4] Basei, M., Guo, X., Hu, A. and Zhang, Y. (2020) Logarithmic Regret for Episodic Continuous-Time Linear-Quadratic Reinforcement Learning over a Finite-Time Horizon. *Computation Theory eJournal*. https://doi.org/10.2139/ssrn.3848428

[5] Dean, S., Mania, H., Matni, N., Recht, B. and Tu, S. (2017) On the Sample Complexity of the Linear Quadratic Regulator. *Foundations of Computational Mathematics*, **20**, 633-679. https://doi.org/10.1007/s10208-019-09426-y

[6] Ren, Z., Zhong, A. and Li, N. (2021) LQR with Tracking: A Zeroth-Order Approach and Its Global Convergence. 2021 *American Control Conference* (*ACC*), New Orleans, LA, 25-28 May 2021, 2562-2568. https://doi.org/10.23919/ACC50511.2021.9483417

[7] Bertsekas, D.P. (2011) Approximate Policy Iteration: A Survey and Some New Methods. *Journal of Control Theory and Applications*, **9**, 310-335. https://doi.org/10.1007/s11768-011-1005-3

[8] Mania, H., Guy, A. and Recht, B. (2018) Simple Random Search Provides a Competitive Approach to Reinforcement Learning. arXiv preprint arXiv:1803.07055

[9] Abbasi-Yadkori, Y., Lazic, N. and Szepesvari, C. (2019) Model-Free Linear Quadratic Control via Reduction to Expert Prediction. *The 22nd International Conference on Artificial Intelligence and Statistics*, Naha, 16-18 April 2019, 3108-3117.

[10] Mahdi, I. and Braga-Neto, U.M. (2018) Finite-Horizon lqr Controller for Partially-Observed Boolean Dynamical Systems. *Automatica*, **95**, 172-179. https://doi.org/10.1016/j.automatica.2018.05.028

[11] Zhang, H. and Li, N. (2022) Data-Driven Policy Iteration Algorithm for Continuous-Time Stochastic Linear-Quadratic Optimal Control Problems. *Asian Journal of Control*, **26**, 481-489. https://doi.org/10.1002/asjc.3223

[12] Farjadnasab, M. and Babazadeh, M. (2022) Model-Free LQR Design by Q-Function Learning. *Automatica*, **137**, Article ID: 110060. https://doi.org/10.1016/j.automatica.2021.110060

[13] Yaghmaie, F.A., Gustafsson, F.K. and Ljung, L. (2023) Linear Quadratic Control Using Model-Free Reinforcement Learning. *IEEE Transactions on Automatic Control*, **68**, 737-752. https://doi.org/10.1109/TAC.2022.3145632

[14] Tu, S. and Recht, B. (2019) The Gap between Model-Based and Model-Free Methods on the Linear Quadratic Regulator: An Asymptotic Viewpoint. *Conference on Learning Theory*, USA, 9 December 2019, 3036-3083.

[15] Malik, D., Pananjady, A., Bhatia, K., Khamaru, K., Bartlett, P.L. and Wainwright, M.J. (2018) Derivative-Free Methods for Policy Optimization: Guarantees for Linear Quadratic Systems. *Journal of Machine Learning Research*, **21**, 1-21.

[16] Fazel, M., Ge, R. Kakade, S.M. and Mesbahi, M. (2018) Global Convergence of Policy Gradient Methods for the Linear Quadratic Regulator. *International Conference on Machine Learning*, Stockholm, 10-15 July 2018, 1467-1476.

[17] Hambly, B.M., Xu, R. and Yang, H. (2021) Policy Gradient Methods for the Noisy Linear Quadratic Regulator over a Finite Horizon. *SIAM Journal on Control and Optimization*, **59**, 3359-3391. https://doi.org/10.1137/20M1382386

[18] Shamir, O. (2017) An Optimal Algorithm for Bandit and Zero-Order Convex Optimization with Two-Point Feedback.

The Journal of Machine Learning Research, **18**, 1703-1713.

[19] Bu, J., Mesbahi, A. and Mesbahi, M. (2020) Policy Gradient-Based Algorithms for Continuous-Time Linear Quadratic Control. arXiv: 2006.09178.

[20] Bertsekas, D.P. (1995) Dynamic Programming and Optimal Control. 3rd Edition, Athena Scientific, Nashua, NH.

附录

引理 2.3 证明：

在对 $\Sigma_{\mathbf{K}}$ 进行分析前，定义以下线性算子以方便证明。令

$$\mathcal{F}_{K_t}(\Sigma)=(A-BK_t)\Sigma(A-BK_t)^T,\ \ \mathcal{G}_t(\Sigma)=\mathcal{F}_{K_t}\circ\mathcal{F}_{K_{t-1}}\circ\cdots\circ\mathcal{F}_{K_0}$$

$$\begin{aligned}\Sigma_{t+1}&=E\left(((A-BK_t)x_t+\omega_t)((A-BK_t)x_t+\omega_t)^T\right)\\&=(A-BK_t)\Sigma_t(A-BK_t)^T+W=\mathcal{F}_{K_t}(\Sigma_t)+W\\&=(A-BK_t)\left(\mathcal{F}_{K_{t-1}}(\Sigma_{t-1})+W\right)(A-BK_{t-1})^T+W\\&=\mathcal{F}_{K_t}\circ\mathcal{F}_{K_{t-1}}(\Sigma_{t-1})+\mathcal{F}_{K_t}(W)+W\\&=\mathcal{F}_{K_t}\circ\mathcal{F}_{K_{t-1}}\circ\mathcal{F}_{K_{t-2}}(\Sigma_{t-2})+\mathcal{F}_{K_t}\circ\mathcal{F}_{K_{t-1}}(W)+\mathcal{F}_{K_t}(W)+W\\&=\mathcal{G}_t(\Sigma_0)+\sum_{s=0}^{t}\prod_{u=0}^{s}\mathcal{F}_{K_t}\circ\cdots\circ\mathcal{F}_{K_{t-u}}(W)+W\end{aligned}$$

$$\Sigma_{\mathbf{K}}=\Sigma_0+\sum_{t=0}^{T-1}\left[\mathcal{G}_t(\Sigma_0)+\sum_{s=0}^{t}\prod_{u=0}^{s}\mathcal{F}_{K_t}\circ\cdots\circ\mathcal{F}_{K_{t-u}}(W)\right]+TW$$

$$\begin{aligned}\sum_{t=0}^{T-1}\left\|(\mathcal{G}_t'-\mathcal{G}_t)(\Sigma_0)\right\|&\le\sum_{t=0}^{T-1}\left\|\left(\mathcal{F}_{K_t'}\circ\mathcal{G}_{t-1}'-\mathcal{F}_{K_t'}\circ\mathcal{G}_{t-1}+\mathcal{F}_{K_t'}\circ\mathcal{G}_{t-1}-\mathcal{F}_{K_t}\circ\mathcal{G}_{t-1}\right)(\Sigma_0)\right\|\\&\le\sum_{t=0}^{T-1}\left\|\mathcal{F}_{K_t'}\right\|\left\|(\mathcal{G}_{t-1}'-\mathcal{G}_{t-1})(\Sigma_0)\right\|+\left\|\mathcal{G}_{t-1}\right\|\left\|\left(\mathcal{F}_{K_t'}-\mathcal{F}_{K_t}\right)(\Sigma_0)\right\|\\&\le\sum_{t=0}^{T-1}\rho^2\left\|(\mathcal{G}_{t-1}'-\mathcal{G}_{t-1})(\Sigma_0)\right\|+\rho^{2(t+1)}\left\|\left(\mathcal{F}_{K_t'}-\mathcal{F}_{K_t}\right)(\Sigma_0)\right\|\\&\le\frac{\rho^{2T}-1}{\rho^2-1}\sum_{t=0}^{T-1}\left\|\left(\mathcal{F}_{K_t'}-\mathcal{F}_{K_t}\right)(\Sigma_0)\right\|\end{aligned}$$

$$\begin{aligned}\sum_{t=0}^{T-1}\left\|\left(\mathcal{F}_{K_t}-\mathcal{F}_{K_t'}\right)(\Sigma_0)\right\|&=\sum_{t=0}^{T-1}\left\|(A-BK_t)\Sigma_0(A-BK_t)^T-(A-BK_t')\Sigma_0(A-BK_t')^T\right\|\\&=\sum_{t=0}^{T-1}\left\|(A-BK_t)\Sigma_0(B\Delta)^T+(B\Delta)\Sigma_0(A-BK_t)^T-(B\Delta)\Sigma_0(B\Delta)^T\right\|\\&\le\sum_{t=0}^{T-1}\left\|\Sigma_0\right\|\left(2\left\|A-BK_t\right\|\left\|B\right\|\left\|K_t-K_t'\right\|+\left\|B\right\|^2\left\|K_t-K_t'\right\|^2\right)\\&\le\left(2\rho\left\|B\right\|\sum_{t=0}^{T-1}\left\|K_t-K_t'\right\|+\left\|B\right\|^2\sum_{t=0}^{T-1}\left\|K_t-K_t'\right\|^2\right)\left\|\Sigma_0\right\|\end{aligned}$$

同理可得，

$$\begin{aligned}&\textstyle\sum_{t=0}^{T-1}\left\|\sum_{s=0}^{t}\prod_{u=0}^{s}\left(\mathcal{F}_{K_t'}\circ\cdots\circ\mathcal{F}_{K_{t-u}'}-\mathcal{F}_{K_t}\circ\cdots\circ\mathcal{F}_{K_{t-u}}\right)(W)\right\|\\&\le\frac{\rho^{2T}-1}{\rho^2-1}\sum_{t=0}^{T-1}\left\|\sum_{s=0}^{t}\left(\mathcal{F}_{K_t'}-\mathcal{F}_{K_t}\right)(W)\right\|\end{aligned}$$

综上所述，

$$\begin{aligned}\left\|\Sigma_{\mathbf{K}}-\Sigma_{\mathbf{K}'}\right\| &\le \sum_{t=0}^{T-1}\left\|\left(\mathcal{G}_t'-\mathcal{G}_t\right)\left(\Sigma_0\right)\right\|+\left\|\sum_{s=0}^{t}\prod_{u=0}^{s}\left(\mathcal{F}_{K_t'}\circ\cdots\circ\mathcal{F}_{K_{t-u}'}-\mathcal{F}_{K_t}\circ\cdots\circ\mathcal{F}_{K_{t-u}}\right)(W)\right\| \\ &\le \frac{\rho^{2T}-1}{\rho^2-1}\left(\sum_{t=0}^{T-1}\left\|\left(\mathcal{F}_{K_t'}-\mathcal{F}_{K_t}\right)\left(\Sigma_0\right)\right\|+\sum_{t=0}^{T-1}\left\|\sum_{s=0}^{t}\left(\mathcal{F}_{K_t'}-\mathcal{F}_{K_t}\right)(W)\right\|\right) \\ &\le \frac{\rho^{2T}-1}{\rho^2-1}\left(\sum_{t=0}^{T-1}\left\|\mathcal{F}_{K_t'}-\mathcal{F}_{K_t}\right\|\right)\left(\left\|\Sigma_0\right\|+T\left\|W\right\|\right) \\ &\le \frac{\rho^{2T}-1}{\rho^2-1}\left(\sum_{t=0}^{T-1}\left\|\mathcal{F}_{K_t'}-\mathcal{F}_{K_t}\right\|\right)\left(\left\|\Sigma_0\right\|+T\left\|W\right\|\right) \\ &\le \frac{\rho^{2T}-1}{\rho^2-1}\left(2\rho\|B\|\sum_{t=0}^{T-1}\|\Delta\|+\|B\|^2\sum_{t=0}^{T-1}\|\Delta\|^2\right)\left(\frac{C(\mathbf{K})}{\sigma_{\min}\mathbf{Q}}+T\left\|W\right\|\right)\end{aligned}$$

引理 2.6 证明：

令 $\phi_t=Q_t+K_t^TR_tK_t$

$$\begin{aligned}C(\mathbf{K}')-C(\mathbf{K}) &= \sum_{t=0}^{T-1}\gamma^t\left[Tr\left(\Sigma_t'\left(Q_t+\left(K_t'\right)^TR_tK_t'\right)\right)-Tr\left(\Sigma_t\left(Q_t+K_t^TR_tK_t\right)\right)\right] \\ &= \sum_{t=0}^{T-1}\gamma^t\left[Tr\left(\mathcal{G}_t'\left(\Sigma_0\right)\phi_t'-\mathcal{G}_t\left(\Sigma_0\right)\phi_t\right)+Tr\left(\left(\phi_t'-\phi_t\right)W\right)\right] \\ &\quad+\sum_{t=0}^{T-1}\gamma^t\left[Tr\left(\sum_{s=0}^{t}\prod_{u=0}^{s}\mathcal{F}_{K_t'}'\circ\cdots\circ\mathcal{F}_{K_{t-u}'}'(W)\phi_t'-\sum_{s=0}^{t}\prod_{u=0}^{s}\mathcal{F}_{K_t}\circ\cdots\circ\mathcal{F}_{K_{t-u}}(W)\phi_t\right)\right]\end{aligned}$$

首先看右边第一项，

$$\left\|\sum_{t=0}^{T-1}\gamma^tTr\left(\mathcal{G}_t'\left(\Sigma_0\right)\phi_t'-\mathcal{G}_t\left(\Sigma_0\right)\phi_t\right)\right\|\le Tr\left(\Sigma_0\right)\sum_{t=0}^{T-1}\gamma^t\left\|Tr\left(\mathcal{G}_t'\left(\phi_t'\right)-\mathcal{G}_t\left(\phi_t\right)\right)\right\|$$

根据引理 2.3 证明中的定义，

$$\begin{aligned}\sum_{t=0}^{T-1}\gamma^t\left\|Tr\left(\mathcal{G}_t'\left(\phi_t'\right)-\mathcal{G}_t\left(\phi_t\right)\right)\right\| &\le \sum_{t=0}^{T-1}\gamma^t\left\|Tr\left(\mathcal{G}_t'\left(\phi_t'\right)-\mathcal{G}_t\left(\phi_t'\right)+\mathcal{G}_t\left(\phi_t'\right)-\mathcal{G}_t\left(\phi_t\right)\right)\right\| \\ &\le \frac{\rho^{2T}-1}{\rho^2-1}\left[\left(2\rho+1\right)\|B\|\sum_{t=0}^{T-1}\gamma^t\left\|K_t-K_t'\right\|\left\|\phi_t'\right\|\right]+\sum_{t=0}^{T-1}\gamma^t\left\|\mathcal{G}_t\right\|\left\|\left(K_t'\right)^TR_tK_t'-K_t^TR_tK_t\right\| \\ &\le \frac{\rho^{2T}-1}{\rho^2-1}\left[\left(2\rho+1\right)\|B\|\sum_{t=0}^{T-1}\gamma^t\left\|K_t-K_t'\right\|\cdot\sum_{t=0}^{T-1}\left\|\phi_t'\right\|+\sum_{t=0}^{T-1}\gamma^t\left\|\left(K_t'\right)^TR_tK_t'-K_t^TR_tK_t\right\|\right] \\ &\le \frac{\rho^{2T}-1}{\rho^2-1}\left(2\rho+1\right)\|B\|\left(\sum_{t=0}^{T-1}\left\|Q_t+K_t^TR_tK_t\right\|\right)\sum_{t=0}^{T-1}\gamma^t\left\|K_t-K_t'\right\| \\ &\quad+\frac{\rho^{2T}-1}{\rho^2-1}\left[\left(2\rho+1\right)\|B\|\sum_{t=0}^{T-1}\left\|K_t-K_t'\right\|+1\right]\sum_{t=0}^{T-1}\gamma^t\left\|\left(K_t'\right)^TR_tK_t'-K_t^TR_tK_t\right\|\end{aligned}$$

第二项，

$$\sum_{t=0}^{T-1}\gamma^tTr\left(\left(\phi_t'-\phi_t\right)W\right)\le Tr\left(W\right)\sum_{t=0}^{T-1}\gamma^t\left\|\left(K_t'\right)^TR_tK_t'-K_t^TR_tK_t\right\|$$

第三项，

$$\sum_{t=0}^{T-1}\gamma^t Tr\left[\sum_{s=0}^{t}\prod_{u=0}^{s}\mathcal{F}'_{K_t}\circ\cdots\circ\mathcal{F}'_{K_{t-u}}(W)\phi'_t-\sum_{s=0}^{t}\prod_{u=0}^{s}\mathcal{F}_{K_t}\circ\cdots\circ\mathcal{F}_{K_{t-u}}(W)\phi_t\right]$$

$$=\sum_{t=0}^{T-1}\gamma^t Tr\left[\sum_{s=0}^{t}\left[\prod_{u=0}^{s}\mathcal{F}'_{K_t}\circ\cdots\circ\mathcal{F}'_{K_{t-u}}(W)\left(\phi'_t-\phi_t+\phi_t\right)-\prod_{u=0}^{s}\mathcal{F}_{K_t}\circ\cdots\circ\mathcal{F}_{K_{t-u}}(W)\phi_t\right]\right]$$

$$\le\sum_{t=0}^{T-1}\sum_{s=0}^{t}Tr(W)\left\|\prod_{u=0}^{s}\mathcal{F}'_{K_t}\circ\cdots\circ\mathcal{F}'_{K_{t-u}}\right\|^2\sum_{t=0}^{T-1}\gamma^t\left\|\left(K'_t\right)^T R_tK'_t-K_t^T R_tK_t\right\|$$

$$+\sum_{t=0}^{T-1}\gamma^t\left\|\sum_{s=0}^{t}\prod_{u=0}^{s}\mathcal{F}'_{K_t}\circ\cdots\circ\mathcal{F}'_{K_{t-u}}(W)-\prod_{u=0}^{s}\mathcal{F}_{K_t}\circ\cdots\circ\mathcal{F}_{K_{t-u}}(W)\right\|\left[\sum_{t=0}^{T-1}\left\|Q_t+K_t^T R_tK\right\|\right]$$

$$\le\frac{(T-1)\left(\rho^{4(T-1)}-1\right)}{\rho^4-1}Tr(W)\sum_{t=0}^{T-1}\gamma^t\left\|\left(K'_t\right)^T R_tK'_t-K_t^T R_tK_t\right\|$$

$$+\frac{T\left(\rho^{2T}-1\right)}{\rho^2-1}\left[(2\rho+1)\|W\|\|B\|\sum_{t=0}^{T-1}\left\|K_t-K'_t\right\|\right]\left[\sum_{t=0}^{T-1}\left\|Q_t+K_t^T R_tK\right\|\right]$$

其中，

$$\sum_{t=0}^{T-1}\gamma^t\left\|\left(K'_t\right)^T R_tK'_t-K_t^T R_tK_t\right\|=\sum_{t=0}^{T-1}\gamma^t\left\|\left(K'_t-K_t+K_t\right)^T R_t\left(K'_t-K_t+K_t\right)-K_t^T R_tK_t\right\|$$

$$\le\sum_{t=0}^{T-1}\gamma^t\left\|K_t-K'_t\right\|^2\|R_t\|+2\gamma^t\|K_t\|\|R_t\|\left\|K_t-K'_t\right\|$$

$$\le\sum_{t=0}^{T-1}3\gamma^t\|K_t\|\|R_t\|\left\|K_t-K'_t\right\|$$

根据假设$\left\|K_t-K'_t\right\|\le\|K_t\|$，倒数第二个不等式成立，综上所述，

$$C(\mathbf{K}')-C(\mathbf{K})\le\left\{\frac{\rho^{2T}-1}{\rho^2-1}\left[Tr(\Sigma_0)(2\rho+1)\|B\|\sum_{t=0}^{T-1}\left\|K_t-K'_t\right\|+1+Tr(W)\right]+Tr(W)\right\}\sum_{t=0}^{T-1}3\gamma^t\|K_t\|\|R_t\|\left\|K_t-K'_t\right\|$$

$$+\frac{(T-1)\left(\rho^{4(T-1)}-1\right)}{\rho^4-1}Tr(W)\sum_{t=0}^{T-1}3\gamma^t\|K_t\|\|R_t\|\left\|K_t-K'_t\right\|$$

$$+\frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\left(T\|W\|+\sum_{t=0}^{T-1}\left\|Q_t+K_t^T R_tK_t\right\|\right)\sum_{t=0}^{T-1}\gamma^t\left\|K_t-K'_t\right\|$$

$$\le h_1\sum_{t=0}^{T-1}\left\|K_t-K'_t\right\|+h_2\left[\sum_{t=0}^{T-1}\left\|K_t-K'_t\right\|\right]^2$$

$$\le h_C\sum_{t=0}^{T-1}\left\|K_t-K'_t\right\|$$

下面证明梯度项，

$$\left\|\nabla_tC(\mathbf{K}')-\nabla_tC(\mathbf{K})\right\|=2\left\|E'_t\Sigma'_t-E_t\Sigma_t\right\|\le2\left\|E'_t-E_t\right\|\|\Sigma'_t\|+2\|E_t\|\left\|\Sigma'_t-\Sigma_t\right\|$$

$$\|\Sigma'_t\|\le\|\Sigma_{\mathbf{K}'}\|\le\left\|\Sigma_{\mathbf{K}'}-\Sigma_{\mathbf{K}}\right\|+\|\Sigma_{\mathbf{K}}\|\le\frac{C(\mathbf{K})}{\sigma_{\min}\mathbf{Q}}+\|\Sigma_{\mathbf{K}}\|$$

$$\left\|E'_t-E_t\right\|=\left\|R_t\left(K'_t-K_t\right)-B^T\left(P'_{t+1}-P_{t+1}\right)A+B^T\left(P'_{t+1}-P_{t+1}\right)BK'_t+B^TP_{t+1}B\left(K'_t-K_t\right)\right\|$$

$$\le\left(\|R_t\|+\|B\|^2\|P_0\|\right)\sum_{t=0}^{T-1}\left\|K_t-K'_t\right\|+\|B\|\left\|P'_0-P_0\right\|\|A\|+2\|B\|^2\left\|P'_0-P_0\right\|\sum_{t=0}^{T-1}\|K_t\|$$

$$\left\|P_0'-P_0\right\|\le 3\left\|K_0\right\|\left\|R_0\right\|\left\|K_0'-K_0\right\|+\left\|\sum_{t=0}^{T-1}\gamma^t Tr\left(\mathcal{G}_t'\left(\Sigma_0\right)\phi_t'-\mathcal{G}_t\left(\Sigma_0\right)\phi_t\right)\right\|$$

$$+\frac{\rho^{2T}-1}{\rho^2-1}\left[\left(2\rho+1\right)\left\|B\right\|\left\|Q_T\right\|\sum_{t=0}^{T-1}\left\|K_t-K_t'\right\|\right]$$

$$\left\|E_t\right\|\le\sum_{t=0}^{T-1}\left\|E_t\right\|\le\sum_{t=0}^{T-1}\sqrt{Tr\left(E_t^TE_t\right)}\le\sqrt{T\cdot\frac{\max\limits_t\left\|R_t+\gamma B^TP_{t+1}B\right\|}{\sigma_{\min}\Sigma_{\mathbf{K}}}\left(C\left(\mathbf{K}\right)-C\left(\mathbf{K}^*\right)\right)}$$

$$\left\|\Sigma_t'-\Sigma_t\right\|\le\left\|\left(\mathcal{G}_t'-\mathcal{G}_t\right)\left(\Sigma_0\right)\right\|+\left\|\sum_{s=0}^{t}\prod_{u=0}^{s}\left(\mathcal{F}_{K_t'}\circ\cdots\circ\mathcal{F}_{K_{t-u}'}-\mathcal{F}_{K_t}\circ\cdots\circ\mathcal{F}_{K_{t-u}}\right)\left(W\right)\right\|$$

$$\le\left[\rho^{2t}\left(2\rho+1\right)\left\|B\right\|\left\|\Sigma_0\right\|+\frac{\rho^{2T}-1}{\rho^2-1}\left(2\rho+1\right)\left\|B\right\|\left\|W\right\|\right]\sum_{t=0}^{T-1}\left\|K_t-K_t'\right\|$$

$$\left\|\nabla_tC\left(\mathbf{K}'\right)-\nabla_tC\left(\mathbf{K}\right)\right\|\le h_g\sum_{t=0}^{T-1}\left\|K_t'-K_t\right\|$$

引理 2.7 证明过程

定义

$$\tilde{\nabla}_tC\left(\mathbf{K}\right)\triangleq\frac{1}{I}\sum_{i=0}^{I-1}\frac{D}{r^2}\left(C\left(\mathbf{K}+\mathbf{U}^i\right)U_t^i\right)$$

则

$$\hat{\nabla}_tC\left(\mathbf{K}\right)-\nabla_tC\left(\mathbf{K}\right)=\hat{\nabla}_tC\left(\mathbf{K}\right)-\tilde{\nabla}_tC\left(\mathbf{K}\right)+\tilde{\nabla}_tC\left(\mathbf{K}\right)-\nabla_tC\left(\mathbf{K}\right)$$

因为 $E_{x_0,\boldsymbol{\omega}}\left[\hat{\nabla}_tC\left(\mathbf{K}\right)\right]=\tilde{\nabla}_tC\left(\mathbf{K}\right)$，根据霍夫丁不等式，对 $\forall\varsigma>0$，

$$P\left\{\left\|\hat{\nabla}_tC\left(\mathbf{K}\right)-\tilde{\nabla}_tC\left(\mathbf{K}\right)\right\|\le\varsigma\right\}\ge1-2\exp\left(-\frac{2I\varsigma^2}{L^2}\right)$$

因为

$$\begin{aligned}\left\|\hat{c}_t^i\right\|&\le\sum_{t=0}^{T-1}\gamma^t\left\|\left(\left(x_t^i\right)^TQ_tx_t^i+\left(u_t^i\right)^TR_tu_t^i\right)+\left(x_T^i\right)^TQ_Tx_T^i\right\|\\&\le\sum_{t=0}^{T-1}\gamma^t\left\|x_t^i\right\|^2\left\|Q_t+\left(K_t^i\right)^TR_tK_t^i\right\|\\&\le\max_t\left\|Q_t+\left(K_t^i\right)^TR_tK_t^i\right\|\sum_{t=0}^{T-1}\gamma^t\left\|\Sigma_t^i\right\|\\&\le\max_t\left\|Q_t+\left(K_t^i\right)^TR_tK_t^i\right\|\left(\frac{C\left(\mathbf{K}\right)}{\sigma_{\min}\mathbf{Q}}+\left\|\Sigma_{\mathbf{K}^0}\right\|\right)\triangleq L\end{aligned}$$

所以，$\left\|\hat{\nabla}_tC\left(\mathbf{K}\right)-\tilde{\nabla}_tC\left(\mathbf{K}\right)\right\|_F\le\frac{\varepsilon}{2}$。

又 $\nabla_tC^r\left(\mathbf{K}\right)=\frac{D}{r^2}E_{\mathbf{U}\sim\mathbb{B}_r}\left[C\left(\mathbf{K}+\mathbf{U}_t\right)U_t\right]$，$\tilde{\nabla}_tC\left(\mathbf{K}\right)\triangleq\frac{1}{I}\sum_{i=0}^{I-1}\frac{D}{r^2}\left(C\left(\mathbf{K}+\mathbf{U}^i\right)U_t^i\right)$

所以，$E\left[\tilde{\nabla}_tC\left(\mathbf{K}\right)\right]=\nabla_tC^r\left(\mathbf{K}\right)$，$\left\|\tilde{\nabla}_tC\left(\mathbf{K}\right)-\nabla_tC^r\left(\mathbf{K}\right)\right\|_F\le\frac{\varepsilon}{4}$

$$\left\|\nabla_tC^r\left(\mathbf{K}\right)-\nabla_tC\left(\mathbf{K}\right)\right\|_F\le\left\|\nabla_tC^r\left(\mathbf{K}\right)-\nabla_tC\left(\mathbf{K}+\mathbf{U}_t\right)\right\|_F+\left\|\nabla_tC\left(\mathbf{K}+\mathbf{U}_t\right)-\nabla_tC\left(\mathbf{K}\right)\right\|_F$$

根据引理 2.9，当 $r \le \frac{\varepsilon}{4h_g}$ 时，$\left\|\nabla_t C\left(\mathbf{K}+\mathbf{U}_t\right)-\nabla_t C\left(\mathbf{K}\right)\right\|_F \le h_g \left\|\mathbf{U}_t\right\|_F \le \frac{\varepsilon}{8}$

所以，$\left\|\nabla_t C^r\left(\mathbf{K}\right)-\nabla_t C\left(\mathbf{K}\right)\right\|_F \le \frac{\varepsilon}{4}$

综上所述，$\left\|\hat{\nabla}_t C\left(\mathbf{K}\right)-\nabla_t C\left(\mathbf{K}\right)\right\| \le \varepsilon$。根据资格迹与梯度的关系，

$$\left\|\hat{\delta}_t^n-\delta_t^n\right\|_F=\sum_{i=0}^{n}\lambda^{n-i}\left\|\hat{\nabla}_t C\left(\mathbf{K}^n\right)-\nabla_t C\left(\mathbf{K}^n\right)\right\|_F \le \sum_{i=0}^{n}\lambda^{n-i}\varepsilon$$

因为 $\lambda \in (0,1)$，由 ε 的任意性可知结论成立。