Pure Mathematics 理论数学, 2024, 14(5), 629-643
Published Online May 2024 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2024.145216

Hans汉斯

马尔科夫跳变线性系统二次最优控制的资格迹方法

朱亚楠

上海理工大学理学院，上海

收稿日期：2024年4月11日；录用日期：2024年5月12日；发布日期：2024年5月31日

摘　要

本文研究了资格迹方法在马尔科夫跳变线性系统的最优二次控制问题(MJLS-LQR)中的应用。常见的方法通过求解耦合的代数黎卡提方程得到最优控制，并不直接优化策略参数。本文在无模型强化学习方法的基础上引入资格迹，直接优化策略参数。考虑参数已知和参数未知两种情况下，MJLS-LQR问题的资格迹方法。参数未知时，无法利用系统参数信息精确表示资格迹，本文利用零阶优化定理近似资格迹，这可以将问题扩展至代价函数非凸的情况。在有限时域和高斯噪声的条件下，分别给出了两种情况下算法的全局收敛保证。数值模拟结果显示资格迹方法与梯度下降算法相比收敛更快。

关键词

最优控制，马尔科夫跳变系统，资格迹

Eligibility Trace Method for Quadratic Optimal Control of Markovian Jump Linear Quadratic Control

Yanan Zhu

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

Received: Apr. 11th, 2024; accepted: May 12th, 2024; published: May 31st, 2024

Abstract

This paper studies the application of eligibility trace methods in the optimal quadratic control problem of Markov jump linear systems (MJLS-LQR). Common methods obtain optimal control by

文章引用: 朱亚楠. 马尔科夫跳变线性系统二次最优控制的资格迹方法[J]. 理论数学, 2024, 14(5): 629-643.
DOI: 10.12677/pm.2024.145216

solving coupled algebraic Riccati equations, rather than directly optimizing policy parameters. Based on the model-free reinforcement learning method, this paper introduces eligibility traces to directly optimize policy parameters. The eligibility trace method for MJLS-LQR problems is considered under two scenarios: known parameters and unknown parameters. When the parameters are unknown, the system parameter information cannot be used to accurately represent the eligibility trace. This paper utilizes the zero-order optimization theorem to approximate the eligibility trace, which can extend the problem to non-convex cost functions. Global convergence guarantees for the algorithms under both scenarios are provided under the conditions of finite time horizon and Gaussian noise. Numerical simulation results show that the eligibility trace method converges faster compared with the gradient descent algorithm.

Keywords

Optimal Control, Markov Jump Linear Systems, Eligibility Traces

1. 引言

控制理论的一个重要问题是，当系统动力学发生变化时，保证控制系统仍然满足某些预期中的性能要求。这些变化可能由外部环境干扰导致，也可能由系统内部的故障或子系统间连接故障引起。当干扰因素对系统的影响较小时，通过在系统状态方程中引入随机噪声项描述此类不确定性。然而，情况更加复杂时，这种方式无法有效刻画干扰对系统的影响，导致控制反馈的有效性降低，计算成本也会大幅增加，随机跳变系统[1] [2]能够更好地描述这类问题。

马尔科夫跳变线性系统(Markov Jump Linear System, MJLS) [2]是一类重要的随机系统，在通信、控制、金融等领域有广泛的应用。MJLS 具有多个模态，在理想条件下，系统在各个模态之间的跳变转移通过马尔科夫链建模。在复杂的场景中，假设系统在有限的多个模型之间随机转移有确定的概率分布，即从一种模型状态跳转到另一种模型状态的概率是确定的。

在 MJLS 的最优控制问题研究中，常用的方法通过求解一组耦合黎卡提方程组获取最优控制。然而，当系统参数部分已知或未知时，无法取得良好的效果。Tzortzis [3]等研究了模态转移概率不确定情况下，为转移概率矩阵设置模糊集研究 MJLS 的最优控制问题。文献[4]基于矩阵不等式方法研究了系统模态无法有效观测的情况，文献[6] [7] [8]基于黎卡提方程研究了 MJLS 的最优控制问题。随着研究深入，关于转移概率部分未知的离散时间和连续时间马尔科夫跳变线性系统的稳定性问题的理论更加成熟[5] [10]。强化学习(RL)方法[9]和数据驱动方法在解决不确定动力系统的问题中有较大突破，基于采样数据的方法成为解决此类随机系统的最优控制问题的有效手段[11] [12]。RL 方法是交互式的学习方法，系统通过与环境交互积累经验，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)。当系统动力学参数未知时，RL 中的无模型方法直接利用经验数据学习最优控制而不估计系统参数。参数未知时，常用滤波方法估计系统状态参数，代替真实参数求解问题。最著名的是卡尔曼滤波方法[13]。Kim & Smagin [14]，Marcos [15]，Martins [16]将卡尔曼滤波应用在马尔科夫跳变线性系统中，取得了不错的效果。虽然目前理论理解仍然不够完善，但无模型方法在 MJLS 最优控制问题中效果突出[18]。

本文对 MJLS 的策略优化学习方法进行研究，将强化学习和控制理论相结合，提出参数已知和参数

未知两种情况下的策略优化学习方法。在实际应用中，许多动态系统都具有随机性，如通信网络、电力系统、飞行器控制系统[17]等。本文在理论上证明了 RL 方法中的资格迹方法的收敛性，数值分析验证了资格迹方法在 MJLS 最优控制问题中具有较快的收敛性。为解决参数未知的复杂系统的控制问题提供了有效的解决方案和思路。

数值分析部分通过数值模拟验证了不同维度的状态空间下，资格迹方法拥有更快的收敛速度。并研究了资格迹方法中不同衰减参数以及不同模态的系统参数的设置对最终收敛效果的影响。结果显示，衰减参数在合适的范围内，资格迹方法能够获取的最优控制逼近真实最优控制，且收敛速度优于传统方法。

2. 资格迹方法

本文基于 RL 方法中的 actor-critic 框架，基于梯度下降算法提出策略参数优化的资格迹方法[9]，在策略参数优化的过程中用资格迹代替梯度项。时变策略参数 $\mathbf{K}=\left(K_0(\theta_0),K_1(\theta_1),\cdots,K_{T-1}(\theta_{T-1})\right)$，考虑如下有限时域的随机 MJLS-LQR 问题：

$$\begin{aligned}\min_{\mathbf{K}}\quad & V(\mathbf{K})=\mathrm{E}\left[\sum_{t=0}^{T-1}\left(x_t^TQ_t(\theta_t)x_t+u_t^TR_t(\theta_t)u_t\right)+x_T^TQ_T(\theta_T)x_T\right]\\ \text{s.t.}\quad & x_{t+1}=A_t(\theta_t)x_t+B_t(\theta_t)u_t+C_t(\theta_t)\omega_t\end{aligned}\tag{1}$$

其中，$x_t\in\mathbb{R}^d$ 和 $u_t\in\mathbb{R}^k$ 分别表示系统的状态和控制变量，$t\in[T]$，初始状态 x_0 从分布 $\mathcal{D}$ 中随机抽样。$Q_t(\theta_t)$，$R_t(\theta_t)$ 是正定矩阵参数。$A_t(\theta_t),B_t(\theta_t)$ 和 $C_t(\theta_t)$ 是具有合适维数的系统矩阵参数，系统模态参数 $\theta_t\in\Theta$，某一时刻的系统模态 θ_t 由 $\left(A_t(\theta_t),B_t(\theta_t),C_t(\theta_t)\right)$ 给出。假设初始状态协方差矩阵为 $\Sigma_0=E\left[x_0x_0^T\right]$ 正定，独立同分布噪声序列 $\{\omega_t\}_{t=0}^{T-1}$ 满足：

$$\begin{aligned}&E(\omega_t)=0,\\ &E\left(\omega_t\omega_t^T\mathbf{1}_{\{\theta_t=i\}}\right)=W,\forall t\in[T]\\ &E\left(x_t\omega_t^T\mathbf{1}_{\{\theta_t=i\}}\right)=0\end{aligned}\tag{2}$$

假设马尔科夫链上的模态具有时不变转移概率，概率矩阵为 $\Pi=\left[\pi_{ij}\right]\in\mathbb{R}^{N\times N}$：

$$\pi_{ij}=P(\theta_{t+1}=j\mid\theta_t=i),\quad i,j=1,2,\cdots,N\tag{3}$$

问题的目标是确定最优策略参数，保证累积代价函数达到最小值。

定义 $P_t^{\mathbf{K}}(\theta_t)$ 为式(4)的解：

$$\begin{aligned}P_t^{\mathbf{K}}(\theta_t)=&Q_t(\theta_t)+K_t^T(\theta_t)R_t(\theta_t)K_t(\theta_t)\\ &+\left(A_t(\theta_t)+B_t(\theta_t)K_t(\theta_t)\right)^T\mathcal{E}_{\theta_t}\left(P_{t+1}^{\mathbf{K}}(\theta_{t+1})\right)\left(A_t(\theta_t)+B_t(\theta_t)K_t(\theta_t)\right)\end{aligned}\tag{4}$$

不至引起歧义时，本文用 P_t 代替 $P_t^{\mathbf{K}}(\theta_t)$，$A_t$ 代替 $A_t(\theta_t)$ 进行论述。

命题 2.1：遵循策略参数的累积代价函数可表示为：

$$\begin{aligned}V(\mathbf{K},x_t)&=E\left[x_t^TP_t(\theta_t)x_t+\sum_{s=t}^{T-1}\omega_s^TC_s^T(\theta_s)\mathcal{E}_{\theta_s}\left(P_{s+1}(\theta_{s+1})\right)C_s(\theta_s)\omega_s\right]\\ &=E\left[x_t^TP_t(\theta_t)x_t+\sum_{s=t}^{T-1}Tr\left[C_s(\theta_s)WC_s^T(\theta_s)\mathcal{E}_{\theta_s}\left(P_{s+1}(\theta_{s+1})\right)\right]\right]\end{aligned}\tag{5}$$

定义 $E_t=\left(R_t+B_t^T\mathcal{E}_{\theta_t}\left(P_{t+1}(\theta_{t+1})\right)B_t\right)K_t-B_t^T\mathcal{E}_{\theta_t}\left(P_{t+1}(\theta_{t+1})\right)A_t$，系统状态协方差矩阵：

$$\Sigma_t=E\left[x_tx_t^T\right]\tag{6}$$

累积损失函数的梯度为：

$$\nabla_t V(\mathbf{K}, x_0) = 2E_t\Sigma_t \tag{7}$$

证明：从 t 时刻到幕结束的累积代价函数为：

$$V(\mathbf{K}, x_t) = \mathrm{E}\left[\sum_{s=t}^{T-1} x_s^T Q_s x_s + u_s^T R_s u_s + x_{T-1}^T Q_{T-1} x_{T-1} + u_{T-1}^T R_{T-1} u_{T-1} + x_T^T Q_T x_T\right] \tag{8}$$

其中，

$$\begin{aligned} E\left[x_T^T Q_T x_T\right] &= E\left[\left(\left(A_{T-1} + B_{T-1}K_{T-1}\right)x_{T-1}\right)^T \mathcal{E}_{\theta_{T-1}}(P_T)\left(A_{T-1} + B_{T-1}K_{T-1}\right)x_{T-1}\right] \\ &\quad + E\left[\left(C_{T-1}w_{T-1}\right)^T \mathcal{E}_{\theta_{T-1}}(P_T)C_{T-1}w_{T-1}\right] \end{aligned}$$

所以，

$$\begin{aligned} V(\mathbf{K}, x_t) &= \mathrm{E}\left[\sum_{s=t}^{T-2} x_s^T Q_s x_s + u_s^T R_s u_s + x_{T-1}^T \mathcal{E}_{\theta_{T-2}}(P_{T-1})x_{T-1} + \left(C_{T-1}w_{T-1}\right)^T \mathcal{E}_{\theta_{T-1}}(P_T)C_{T-1}w_{T-1}\right] \\ &= \mathrm{E}\left[x_t^T Q_t x_t + u_t^T R_t u_t + x_{t+1}^T \mathcal{E}_{\theta_t}(P_{t+1})x_{t+1} + \sum_{s=t+1}^{T-1}\left(C_s w_s\right)^T \mathcal{E}_{\theta_s}(P_{s+1})C_s w_s\right] \\ &= \mathrm{E}\left[x_t^T \mathcal{E}_{\theta_t}(P_{t+1})x_t + \sum_{s=t}^{T-1} Tr\left[C_s W C_s^T \mathcal{E}_{\theta_s}(P_{s+1})\right]\right] \end{aligned}$$

$$V(\mathbf{K}, x_T) = x_T^T \mathcal{E}_{\theta_{T-1}}(P_T)x_T,\ P_T = Q_T\ 。$$

累积代价函数对策略参数 K_t 的偏导为：

$$\begin{aligned} \nabla_t V(\mathbf{K}, x_0) &= \frac{\partial V(\mathbf{K}, x_0)}{\partial K_t} \\ &= \frac{\partial E\left[x_t^T\left(Q_t + K_t^T R_t K_t + \left(A - BK_t\right)^T \mathcal{E}_{\theta_t}(P_{t+1})\left(A - BK_t\right)\right)x_t + \mathbf{K}(-t)\right]}{\partial K_t} \\ &= E\left[2R_t K_t x_t x_t^T - 2B^T \mathcal{E}_{\theta_t}(P_{t+1})\left(A - BK_t\right)x_t x_t^T\right] \\ &= 2E_t\Sigma_t \end{aligned}$$

其中，

$$\mathbf{K}(-t) = \sum_{s=0}^{t-1} x_s^T Q_s x_s + u_s^T R_s u_s + \sum_{s=t+1}^{T-1}\left(C_s w_s\right)^T \mathcal{E}_{\theta_s}(P_{s+1})C_s w_s$$

证毕。

2.1. 参数已知的资格迹方法

本节讨论有限时域情况下，系统模态参数 $\theta_t, t \in [T]$ 和系统参数 Ξ 已知时的资格迹方法[9]。资格迹方法在蒙特卡洛方法和时序差分方法的基础上，定义一个与策略参数相同维度的短时记忆向量 $\boldsymbol{\delta}$，作为衡量策略参数 $\mathbf{K}$ 不同分量的指标。随着迭代次数的增加，参与更新的控制参数的分量对应的资格迹逐渐衰减，直到这一分量再次参与更新。

考虑如下优化策略参数的资格迹方法：

$$\begin{aligned}&\mathbf{K}^{n+1}=\mathbf{K}^{n}-\alpha\boldsymbol{\delta}^{n}\\&\boldsymbol{\delta}^{0}=\nabla C\left(\mathbf{K}^{0}\right)\qquad\qquad\forall t\in[T]\\&\boldsymbol{\delta}^{n}=\lambda\boldsymbol{\delta}^{n-1}+\nabla C\left(\mathbf{K}^{n}\right),n>0\end{aligned}\tag{9}$$

其中，$n\in[N]$是迭代次数，α 是步长参数，λ 是折扣系数，$\mathbf{K}^{n}=\left(K_{0}^{n},K_{1}^{n},\cdots,K_{T-1}^{n}\right)$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。

$$\delta_{t}^{n}=\begin{cases}2E_{t}^{n}\Sigma_{t}^{n} & t=0,n=0\\ \lambda\delta_{t-1}^{n}+2E_{t}^{n}\Sigma_{t}^{n} & t>0,n>0\end{cases},\forall t\in[T]\tag{10}$$

衰减参数λ的取值不同，决定了历史信息对下一步决策的重要程度。$\lambda=0$时，决策时不考虑历史信息，$\lambda=1$时，在决策过程中历史信息与当前信息同样重要。梯度下降算法只考虑梯度更新的平滑度，而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，能够减少参数更新过程中的错误决策次数。

引理 2.2 假设任意可行控制$\mathbf{K}'$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：

$$\begin{aligned}V\left(\mathbf{K}',x\right)-V\left(\mathbf{K},x\right)=&E\left[\sum_{t=0}^{T-1}2Tr\left(x_t'\left(x_t'\right)^{T}\left(K_t'-K_t\right)^{T}E_t\right)\right]\\&+E\left[\sum_{t=0}^{T-1}Tr\left(x_t'\left(x_t'\right)^{T}\left(K_t'-K_t\right)^{T}\left(R_t+B_t^{T}\mathcal{E}_{\theta_t}\left(P_{t+1}\right)B_t\right)\left(K_t'-K_t\right)\right)\right]\end{aligned}\tag{11}$$

证明见附录。

引理 2.3 令$\rho=\max\left\{\max_i\left\|A_i-B_iK_i\right\|,\max_i\left\|A_i-B_iK_i'\right\|\right\}$，$\Delta=K_t-K_t'$，$\mathbf{K}'$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：

$$\left\|\Sigma_{\mathbf{K}}-\Sigma_{\mathbf{K}'}\right\|\le\frac{\rho^{2T}-1}{\rho^{2}-1}\left[(2\rho+1)\left\|B_{\max}\right\|\sum_{t=0}^{T-1}\|\Delta\|+\left\|B_{\max}\right\|^{2}\sum_{t=0}^{T-1}\|\Delta\|^{2}\right]\left(\frac{V\left(\mathbf{K},x_0\right)}{\sigma_{\min}\mathbf{Q}}+T\left\|W_{\max}\right\|\right)\tag{12}$$

其中，$\left\|B_{\max}\right\|=\max_t\left\{\left\|B_t\right\|\right\}$，$W_{\max}=\arg\max_{C_tWC_t^{T}}\left\|C_tWC_t^{T}\right\|$。

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

$$\alpha\le\min\left\{\alpha_1,\frac{\sigma_{\min}\mathbf{Q}}{2T\cdot V\left(\mathbf{K},x_0\right)\cdot\max_t\left\|R_t+B_t^{T}P_{t+1}B_t\right\|}\right\}$$

其中，

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

则

$$V\left(\mathbf{K}',x_0\right)-V\left(\mathbf{K}^{*},x_0\right)\le\left(1-\frac{8\alpha\sigma_{\min}\mathbf{R}\sigma_{\min}^{2}\boldsymbol{\Sigma}}{\left\|\Sigma_{\mathbf{K}^{*}}\right\|}\right)\left(V\left(\mathbf{K},x_0\right)-V\left(\mathbf{K}^{*},x_0\right)\right)\tag{13}$$

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 2.5 假设 $C\left(\mathbf{K}^0\right)$ 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：

$$N \geq \frac{\left\|\Sigma_{\mathbf{K}^*}\right\|}{8 \alpha \sigma_{\min } \boldsymbol{\Sigma} \sigma_{\min } \mathbf{R}} \log \frac{V\left(\mathbf{K}^0, x_0\right)-V\left(\mathbf{K}^*, x_0\right)}{\varepsilon}$$

代价函数值收敛至最优值，即：

$$V\left(\mathbf{K}, x_0\right)-V\left(\mathbf{K}^*, x_0\right) \leq \varepsilon \tag{14}$$

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

$$V\left(\mathbf{K}^1, x_0\right)-V\left(\mathbf{K}^*, x_0\right) \leq\left(1-\frac{8 \alpha \sigma_{\min } \mathbf{R} \sigma_{\min }^2 \boldsymbol{\Sigma}}{\left\|\Sigma_{\mathbf{K}^*}\right\|}\right)\left[V\left(\mathbf{K}, x_0\right)-V\left(\mathbf{K}^*, x_0\right)\right]$$

假设经 $n+1$ 次迭代后，$V\left(\mathbf{K}^{n+1}, x_0\right) \leq V\left(\mathbf{K}^0, x_0\right)$，此时 $K_t^{n+1}=K_t^n-\alpha \delta_t^n$，根据 Cauchy-Schwarz 不等式，

$$\begin{aligned}
\sum_{t=0}^{T-1} \delta_t^n &= \sum_{t=0}^{T-1} \sum_{i=0}^n \lambda^{n-i} \nabla_t V\left(\mathbf{K}^n, x_0\right) \leq \sum_{t=0}^{T-1} \sqrt{n \sum_{i=0}^n\left\|\nabla_t C\left(\mathbf{K}^n\right)\right\|^2} \\
&\leq \sum_{t=0}^{T-1} \sqrt{4 n \sum_{i=0}^n Tr\left(\Sigma_t^i\left(E_t^i\right)^T E_t^i \Sigma_t^i\right)} \\
&\leq \sqrt{T \cdot \sum_{t=0}^{T-1} 4 n \sum_{i=0}^n\left\|\Sigma_t^i\right\|^2 Tr\left(\left(E_t^i\right)^T E_t^i\right)} \\
&\leq \frac{2 V\left(\mathbf{K}, x_0\right)}{\sigma_{\min } \mathbf{Q}} \sqrt{n T \frac{\max _t\left(R_t+B_t^T \mathcal{E}_{\theta_t}\left(P_{t+1}\right) B_t\right)}{\sigma_{\min } \boldsymbol{\Sigma}}\left(V\left(\mathbf{K}, x_0\right)-V\left(\mathbf{K}^*, x_0\right)\right)}
\end{aligned}$$

$$\begin{aligned}
V\left(\mathbf{K}, x_0\right)-V\left(\mathbf{K}^*, x_0\right) &\geq V\left(\mathbf{K}, x_0\right)-V\left(\mathbf{K}', x_0\right) \\
&= E\left[\sum_{t=0}^{T-1} Tr\left(E_t^T\left(R_t+B_t^T \mathcal{E}_{\theta_t}\left(P_{t+1}\right) B_t\right)^{-1} E_t\right)\right] \\
&\geq \frac{\sigma_{\min } \boldsymbol{\Sigma}}{\max _t\left(R_t+B_t^T \mathcal{E}_{\theta_t}\left(P_{t+1}\right) B_t\right)} \sum_{t=0}^{T-1} Tr\left(E_t^T E_t\right)
\end{aligned}$$

结合引理 2.3 的分析，引理 2.3 中的结论仍然成立，即：

$$V\left(\mathbf{K}^{n+1}, x_0\right)-V\left(\mathbf{K}^*, x_0\right) \leq\left(1-\frac{8 \alpha \sigma_{\min } \mathbf{R} \sigma_{\min }^2 \boldsymbol{\Sigma}}{\left\|\Sigma_{\mathbf{K}^*}\right\|}\right)\left[V\left(\mathbf{K}^n, x_0\right)-V\left(\mathbf{K}^*, x_0\right)\right] \tag{15}$$

将 $n+1$ 次的结果进行累积，

$$V\left(\mathbf{K}^{n+1}, x_0\right)-V\left(\mathbf{K}^*, x_0\right) \leq\left(1-\frac{8 \alpha \sigma_{\min } \mathbf{R} \sigma_{\min }^2 \boldsymbol{\Sigma}}{\left\|\Sigma_{\mathbf{K}^*}\right\|}\right)^{n+1}\left[V\left(\mathbf{K}^0, x_0\right)-V\left(\mathbf{K}^*, x_0\right)\right]$$

对 $\forall \varepsilon>0$，当

$$N \geq \frac{\left\|\Sigma_{\mathbf{K}^*}\right\|}{8\alpha\sigma_{\min}\mathbf{\Sigma}\sigma_{\min}\mathbf{R}}\log\frac{V\left(\mathbf{K}^0,x_0\right)-V\left(\mathbf{K}^*,x_0\right)}{\varepsilon}$$

时，$V\left(\mathbf{K}^N,x_0\right)-V\left(\mathbf{K}^*,x_0\right)\leq\varepsilon$。证毕。

2.2. 系统参数未知的资格迹方法

本节讨论系统模态θ_t和系统参数Ξ未知时的资格迹方法。不同模态下的系统参数间差异间需要满足一定的界限。模态未知，系统使用零阶优化方法近似资格迹，零阶优化方法[19] [20] [21]对目标函数的凸性没有要求，直接以函数值估计函数梯度。在 MJLS 的最优二次控制问题中，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

$$V\left(\mathbf{K},x_0\right)=E_\zeta\left[V\left(\mathbf{K},x_0;\zeta\right)\right] \tag{16}$$

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$\mathbb{U}^r=\left\{\mathbf{U}\in\mathbb{R}^{k\times d}:\left\|\mathbf{U}\right\|_F=r\right\}$，设$\mathcal{P}_{\mathbb{U}}$是$\mathbb{U}^r$上的均匀分布。任意度量$r>0$，以及$\mathbf{U}\sim\mathcal{P}_{\mathbb{U}}$与$\zeta$独立，则$C\left(\mathbf{K}\right)$的梯度估计[22]为：

$$\nabla V\left(\mathbf{K},x_0\right)=\frac{k\times d}{r}V\left(\mathbf{K}+\mathbf{U},x_0\right)\mathbf{U} \tag{1}$$

随着 r 越来越小，近似值越来越精确，但 r 过小容易导致方差过大。

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

$$\hat{\delta}_t^n=\begin{cases}\dfrac{1}{I}\sum_{i=0}^{I-1}\dfrac{D}{r^2}\hat{c}_t^iU_t^i & n=0\\ \lambda\hat{\delta}_t^{n-1}+\dfrac{1}{I}\sum_{i=0}^{I-1}\dfrac{D}{r^2}\hat{c}_t^iU_t^i & n>0\end{cases} \tag{18}$$

其中，

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

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

$$\left\|K_t'-K_t\right\|\leq\min\left\{\left\|K_t\right\|,\frac{\left(\rho^2-1\right)\sigma_{\min}\mathbf{Q}\sigma_{\min}\mathbf{\Sigma}}{2T\left(\rho^{2T}-1\right)\left(2\rho+1\right)\left(V\left(\mathbf{K},x_0\right)+\sigma_{\min}\mathbf{Q}\cdot T\left\|W_{\max}\right\|\right)\left\|B_{\max}\right\|}\right\} \tag{19}$$

则存在

$$h_C\leq\left\{\frac{\rho^2-1}{\left(2\rho+1\right)\left(\rho^{2T}-1\right)}\cdot\frac{1}{\left\|B_{\max}\right\|}\cdot\frac{1}{\left\|W_{\max}\right\|}\cdot\frac{1}{V\left(\mathbf{K}^0,x_0\right)}\right\}$$

$$h_g\leq\left\{\frac{\rho^2-1}{\left(2\rho+1\right)\left(\rho^{2T}-1\right)}\cdot\frac{1}{\left\|B_{\max}\right\|}\cdot\frac{1}{\left\|W_{\max}\right\|}\cdot\frac{\sigma_{\min}\mathbf{Q}\sigma_{\min}\mathbf{\Sigma}}{V\left(\mathbf{K}^0,x_0\right)}\frac{1}{\left\|\mathbf{\Sigma}\right\|}\right\}$$

使得，

$$\left|V\left(\mathbf{K}',x_0\right)-V\left(\mathbf{K},x_0\right)\right|\leq h_C\sum_{t=0}^{T-1}\left\|K_t'-K_t\right\|,\ \left\|\nabla_tV\left(\mathbf{K}',x_0\right)-\nabla_tV\left(\mathbf{K},x_0\right)\right\|\leq h_g\sum_{t=0}^{T-1}\left\|K_t'-K_t\right\|$$

定理 2.7 假设$C\left(\mathbf{K}^0\right)$有界，步长$\alpha$满足引理 2.3 的约束，对$\forall\varepsilon>0$，当迭代次数$N$满足

$$N \geq \frac{\left\|\Sigma_{\mathbf{K}^*}\right\|}{8\alpha\sigma_{\min}\boldsymbol{\Sigma}\sigma_{\min}\mathbf{R}}\log\frac{V\left(\mathbf{K}^0,x_0\right)-V\left(\mathbf{K}^*,x_0\right)}{\varepsilon}$$

代价函数值收敛至最优值，即：

$$V\left(\mathbf{K},x_0\right)-V\left(\mathbf{K}^*,x_0\right)\leq\varepsilon \tag{20}$$

证明与定理 2.4 类似。

表 1 提出了 MJLS-LQ 问题的资格迹算法。

Table 1. Eligibility trace algorithm
表 1. 资格迹算法

算法：资格迹算法

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 $\hat{K}_t^i=K_t+U_t^{ni}\left\|U_t^{ni}\right\|_F=r$。

2) 记录 $\hat{c}_t^i$。

5. 计算资格迹的估计值：

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

6. $\mathbf{K}^{n+1}=\mathbf{K}^n-\alpha\hat{\boldsymbol{\delta}}^n$

3. 数值模拟

当系统状态空间维数 $d=2$ 时，系统参数为

$$A_1=\begin{pmatrix}0.8521 & -1.11\\ 1.035 & 0.7436\end{pmatrix},B_1=\begin{pmatrix}0.831\\ 1.002\end{pmatrix},Q_1=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix},R_1=1$$

$$A_2=\begin{pmatrix}0.6984 & 1.13\\ 1.025 & 0.6521\end{pmatrix},B_2=\begin{pmatrix}0.705\\ 0.849\end{pmatrix},Q_2=\begin{pmatrix}1.105 & 0\\ 0 & 0.92\end{pmatrix},R_2=1$$

系统模态转移概率矩阵为：

$$\Pi=\begin{pmatrix}0.9 & 0.1\\ 0.7 & 0.3\end{pmatrix}$$

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，时域 $T=200$，迭代次数 $N=50$ 和迭代次数 $N=100$，代价函数的收敛情况结果如图 1 和图 2 所示。

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 $T=100$，$N=70$ 时不同的折扣系数对算法性能的影响。

图 4 展示了某次系统模态序列，在本节设定的系统参数下，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，收敛更快，但结果不收敛。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。

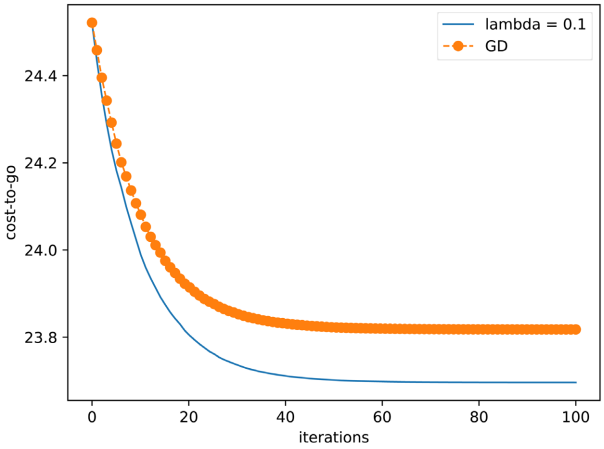

Figure 1. The convergence of $C(K)$ when $d = 2$, $T = 40$
图 1. $d = 2$，$T = 40$ 代价函数的收敛情况

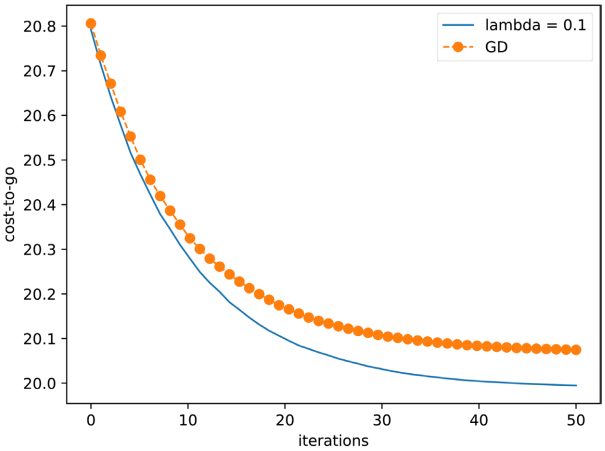

Figure 2. The convergence of $C(K)$ when $d = 2$, $T = 70$
图 2. $d = 2$，$T = 70$ 代价函数的收敛情况

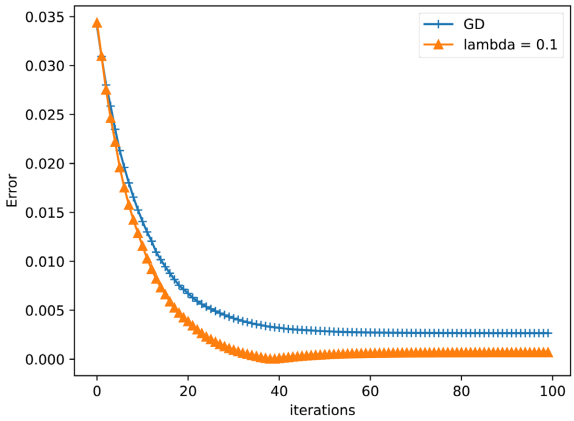

Figure 3. Cost function error variation with $N = 100$
图 3. $N = 100$ 代价函数误差变化

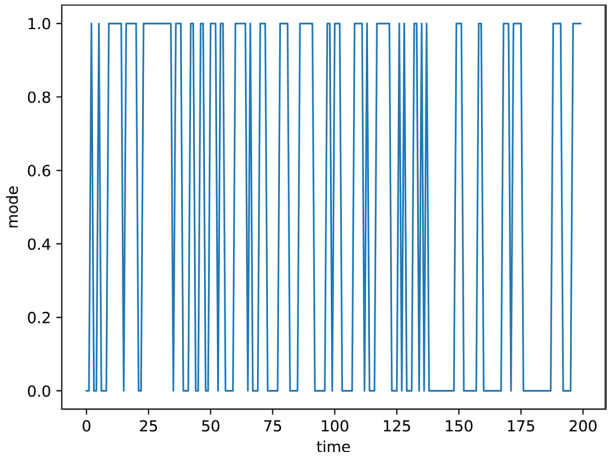

Figure 4. System modal sequence
图 4. 系统模态序列

4. 结论

本文研究了无模型强化学习方法在有限时域 MJLS-LQ 问题中的应用，不同于通过解代数黎卡提方程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] Zhang, Q., Li, L., Yan, X. and Spurgeon, S.K. (2017) Sliding Mode Control for Singular Stochastic Markovian Jump Systems with Uncertainties. *Automatica*, **79**, 27-34. https://doi.org/10.1016/j.automatica.2017.01.002

[2] Costa, O.L., Fragoso, M.D. and Marques, R.P. (2004) Discrete-Time Markov Jump Linear Systems. *IEEE Transactions on Automatic Control*, **51**, 916-917. https://doi.org/10.1109/TAC.2006.874981

[3] Tzortzis, I., Charalambous, C.D. and Hadjicostis, C.N. (2019) Robust LQG for Markov Jump Linear Systems. 2019 *IEEE* 58*th Conference on Decision and Control* (*CDC*), Nice, 11-13 December 2019, 6760-6765. https://doi.org/10.1109/CDC40024.2019.9028886

[4] Todorov, M.G. and Fragoso, M.D. (2014) New Methods for Mode-Independent Robust Control of Markov Jump Linear Systems. 53*rd IEEE Conference on Decision and Control*, Los Angeles, 15-17 December 2014, 4222-4227. https://doi.org/10.1109/CDC.2014.7040047

[5] Wang, Y., Ahn, C.K., Yan, H. and Xie, S. (2020) Fuzzy Control and Filtering for Nonlinear Singularly Perturbed Markov Jump Systems. *IEEE Transactions on Cybernetics*, **51**, 297-308. https://doi.org/10.1109/TCYB.2020.3004226

[6] Guo, Y. and Li, J. (2021) Network-Based Quantized H∞ Control for T-S Fuzzy Singularly Perturbed Systems with Persistent Dwell-Time Switching Mechanism and Packet Dropouts. *Nonlinear Analysis*: *Hybrid Systems*, **42**, Article ID: 101060. https://doi.org/10.1016/j.nahs.2021.101060

[7] Tzortzis, I., Charalambous, C.D. and Hadjicostis, C.N. (2019) Robust LQG for Markov Jump Linear Systems. 2019 *IEEE* 58*th Conference on Decision and Control* (*CDC*), Nice, 11-13 December 2019, 6760-6765. https://doi.org/10.1109/CDC40024.2019.9028886

[8] Lopes, R.O., Mendes, E.M., Tôrres, L.A., Vargas, A.N. and Palhares, R.M. (2020) Finite-Horizon Suboptimal Control of Markov Jump Linear Parameter-Varying Systems. *International Journal of Control*, **94**, 2659-2668. https://doi.org/10.1080/00207179.2020.1728387

[9] Sutton, R.S. and Barto, A.G. (2018) Reinforcement Learning: An Introduction. MIT Press, Cambridge.

[10] Souza, M., Fioravanti, A.R. and Araujo, V.S. (2021) Impulsive Markov Jump Linear Systems: Stability Analysis and H_2 Control. *Nonlinear Analysis*: *Hybrid Systems*, **42**, Article ID: 101089. https://doi.org/10.1016/j.nahs.2021.101089

[11] Chen, Y., Wen, J., Luan, X. and Liu, F. (2020) Robust Control for Markov Jump Linear Systems with Unknown Transition Probabilities—An Online Temporal Differences Approach. *Transactions of the Institute of Measurement and Control*, **42**, 3043-3051. https://doi.org/10.1177/0142331220940208

[12] Park, I.S., Kwon, N.K. and Park, P. (2019) Dynamic Output-Feedback Control for Singular Markovian Jump Systems with Partly Unknown Transition Rates. *Nonlinear Dynamics*, **95**, 3149-3160. https://doi.org/10.1007/s11071-018-04746-0

[13] Zhao, J. and Mili, L. (2019) A Decentralized H-Infinity Unscented Kalman Filter for Dynamic State Estimation Against Uncertainties. *IEEE Transactions on Smart Grid*, **10**, 4870-4880. https://doi.org/10.1109/TSG.2018.2870327

[14] Kim, K.S. and Smagin, V.I. (2020) Robust Filtering for Discrete Systems with Unknown Inputs and Jump Parameters. *Automatic Control and Computer Sciences*, **54**, 1-9. https://doi.org/10.3103/S014641162001006X

[15] Marcos, L.B. and Terra, M.H. (2020) Markovian Filtering for Driveshaft Torsion Estimation in Heavy Vehicles. *Control Engineering Practice*, **102**, Article ID: 104552. https://doi.org/10.1016/j.conengprac.2020.104552

[16] Queiroz de Jesus, G. and Martins Calazans Silva, B. (2022) Robust Estimation for Discrete-Time Markovian Jump Linear Systems in a Data Fusion Scenario. *Intermaths*, **3**, 17-36. https://doi.org/10.22481/intermaths.v3i1.10715

[17] Gray, W.S., González, O.R. and Doğan, M. (2000) Stability Analysis of Digital Linear Flight Controllers Subject to Electromagnetic Disturbances. *IEEE Transactions on Aerospace and Electronic Systems*, **36**, 1204-1218. https://doi.org/10.1109/7.892669

[18] Bertsekas, D.P. (1995) Dynamic Programming and Optimal Control. 3rd Edition, Massachusetts Institute of Technology, Cambridge.

[19] 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

[20] 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.

[21] Hambly, B.M., Xu, R., and Yang, H. (2020) Policy Gradient Methods for the Noisy Linear Quadratic Regulator over a Finite Horizon. DecisionSciRN: Other Decision-Making in Economics (Topic).

[22] 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**, 1-51.

附　录

引理 2.2 证明：

令 $\Upsilon_s = \omega_s^T C_s^T \mathcal{E}_{\theta_s}(P_{s+1}) C_s \omega_s$，

$$\begin{aligned} V(\mathbf{K}', x) - V(\mathbf{K}, x) &= E\left[x^T P_0 x + \sum_{s=0}^{T-1} \Upsilon_s \right] - V(\mathbf{K}, x) \\ &= E\left[x^T P_0 x + \sum_{s=0}^{T-1} \left(\Upsilon_s + V(\mathbf{K}, x_s') - V(\mathbf{K}, x_s') \right) \right] - V(\mathbf{K}, x) \\ &= E\left[x^T P_0 x + \sum_{s=0}^{T-1} \left(\Upsilon_s + V(\mathbf{K}, x_{s+1}') - V(\mathbf{K}, x_s') \right) \right] \end{aligned}$$

令 $J(\mathbf{K}, x_s, u_s) = x_s^T Q_s x_s + u_s^T R_s u_s + E\left[V(\mathbf{K}, x_{s+1}) \right]$

$$\begin{aligned} V(\mathbf{K}', x) - V(\mathbf{K}, x) &= E\left[x^T P_0 x + \sum_{s=0}^{T-1} \left(\Upsilon_s + V(\mathbf{K}, x_{s+1}') - V(\mathbf{K}, x_s') \right) \right] \\ &= E\left[\sum_{s=0}^{T-1} J(\mathbf{K}, x_s', u_s') - V(\mathbf{K}, x_s') \right] \end{aligned}$$

$$\begin{aligned} & J(\mathbf{K}, x_s', u_s') - V(\mathbf{K}, x_s') \\ &= (x_s')^T Q_s x_s' + (u_s')^T R_s u_s' + E\left[V(\mathbf{K}, x_{t+1}) \right] - V(\mathbf{K}, x_s') \\ &= (x_s')^T \left(Q_s + (K_s')^T R_s K_s' + (A_s + B_s K_s')^T \mathcal{E}_{\theta_s}(P_{s+1}) (A_s + B_s K_s') \right) x_s' - (x_s')^T P_s x_s' \\ &= (x_s')^T \left(Q_s + (K_s' - K_s + K_s)^T R_s (K_s' - K_s + K_s) \right) x_s' \\ &\quad + (x_s')^T \left(A_s + B_s (K_s' - K_s + K_s) \right)^T \mathcal{E}_{\theta_s}(P_{s+1}) \left(A_s + B_s (K_s' - K_s + K_s) \right) x_s' \\ &\quad - (x_s')^T \left(Q_s + K_s^T R_s K_s + (A_s + B_s K_s)^T \mathcal{E}_{\theta_s}(P_{s+1}) (A_s + B_s K_s) \right) x_s' \\ &= 2 (x_s')^T (K_s' - K_s)^T \left(\left(R_s + B_s^T \mathcal{E}_{\theta_s}(P_{s+1}) B_s \right) K_s - B_s^T \mathcal{E}_{\theta_s}(P_{s+1}) A_s \right) x_s' \\ &\quad + (x_s')^T (K_s' - K_s)^T \left(R_s + B_s^T \mathcal{E}_{\theta_s}(P_{s+1}) B_s \right) (K_s' - K_s) x_s' \\ &= 2 Tr\left(x_s' (x_s')^T (K_s' - K_s)^T E_s \right) \\ &\quad + Tr\left(x_s' (x_s')^T (K_s' - K_s)^T \left(R_s + B_s^T \mathcal{E}_{\theta_s}(P_{s+1}) B_s \right) (K_s' - K_s) \right) \end{aligned}$$

引理 2.3 证明：

定义线性算子：$\mathcal{F}_{K_t}(X) = (A_t + B_t K_t) X (A_t + B_t K_t)^T$，$\mathcal{G}_t(\Sigma) = \mathcal{F}_{K_t} \circ \mathcal{F}_{K_{t-1}} \circ \cdots \circ \mathcal{F}_{K_0}$

系统状态协方差矩阵为：

$$\begin{aligned} \Sigma_{t+1} &= E\left(x_{t+1} x_{t+1}^T \right) \\ &= E\left(\left((A_t + B_t K_t) x_t + C_t \omega_t \right) \left((A_t + B_t K_t) x_t + C_t \omega_t \right)^T \right) \\ &= (A_t + B_t K_t) \Sigma_t (A_t + B_t K_t)^T + C_t W C_t^T \\ &= \mathcal{F}_{K_t}(\Sigma_t) + C_t W C_t^T \end{aligned}$$

$$
\begin{aligned}
&= \left(A_t + B_t K_t\right)\left(\mathcal{F}_{K_{t-1}}\left(\Sigma_{t-1}\right) + C_{t-1} W C_{t-1}^T\right)\left(A_t + B_t K_t\right)^T + C_t W C_t \\
&= \mathcal{F}_{K_t} \circ \mathcal{F}_{K_{t-1}}\left(\Sigma_{t-1}\right) + \mathcal{F}_{K_t}\left(C_{t-1} W C_{t-1}^T\right) + C_t W C_t^T \\
&= \mathcal{F}_{K_t} \circ \mathcal{F}_{K_{t-1}} \circ \mathcal{F}_{K_{t-2}}\left(\Sigma_{t-2}\right) + \mathcal{F}_{K_t} \circ \mathcal{F}_{K_{t-1}}\left(C_{t-2} W C_{t-2}^T\right) + \mathcal{F}_{K_t}\left(C_{t-1} W C_{t-1}^T\right) + C_t W C_t^T \\
&= \mathcal{G}_t\left(\Sigma_0\right) + \sum_{s=0}^{t-1} \mathcal{F}_{K_t} \circ \cdots \circ \mathcal{F}_{K_{t-s}}\left(C_{t-s-1} W C_{t-s-1}^T\right) + C_t W C_t^T
\end{aligned}
$$

$$
\begin{aligned}
\sum_{t=0}^{T-1}\left\|\left(\mathcal{F}_{K_t} - \mathcal{F}_{K_t'}\right)(X)\right\| &= \sum_{t=0}^{T-1}\left\|\left(A_t + B_t K_t\right) X \left(A_t + B_t K_t\right)^T - \left(A_t + B_t K_t'\right) X \left(A_t + B_t K_t'\right)^T\right\| \\
&= \sum_{t=0}^{T-1}\left\|\left(A_t + B_t K_t\right) X \left(B_t \Delta_t\right)^T + \left(B_t \Delta_t\right) X \left(A_t + B_t K_t\right)^T - \left(B_t \Delta_t\right) X \left(B_t \Delta_t\right)^T\right\| \\
&\le \sum_{t=0}^{T-1}\|X\|\left(2\left\|A_t + B_t K_t\right\|\left\|B_t\right\|\left\|K_t - K_t'\right\| + \left\|B_t\right\|^2\left\|K_t - K_t'\right\|^2\right) \\
&\le \left(2\rho\left\|B_{\max}\right\|\sum_{t=0}^{T-1}\left\|K_t - K_t'\right\| + \left\|B_{\max}\right\|^2 \sum_{t=0}^{T-1}\left\|K_t - K_t'\right\|^2\right)\|X\|
\end{aligned}
$$

令 $\mathcal{F}_{K_t}(X) = \mathcal{F}_t, \mathcal{F}_{K_t'}(X) = \mathcal{F}_t'$

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

同理可得：

$$
\sum_{s=0}^{t-1}\left(\mathcal{F}_t' \circ \cdots \circ \mathcal{F}_{t-s}' - \mathcal{F}_t \circ \cdots \circ \mathcal{F}_{t-s}\right)\left(C_{t-s-1} W C_{t-s-1}^T\right) \le \frac{\rho^{2T} - 1}{\rho^2 - 1}\left(\sum_{s=0}^{t-1}\left\|\mathcal{F}_t' - \mathcal{F}_t\right\|\right)\left\|W_{\max}\right\|
$$

综上所述：

$$
\begin{aligned}
\left\|\Sigma_{\mathbf{K}} - \Sigma_{\mathbf{K}'}\right\| &\le \sum_{t=0}^{T-1}\left[\left\|\left(\mathcal{G}_t' - \mathcal{G}_t\right)\left(\Sigma_0\right)\right\| + \sum_{s=0}^{t-1}\left\|\left(\mathcal{F}_t' \circ \cdots \circ \mathcal{F}_{t-u}' - \mathcal{F}_t \circ \cdots \circ \mathcal{F}_{t-u}\right)\left(C_{t-u-1} W C_{t-u-1}^T\right)\right\|\right] \\
&\le \frac{\rho^{2T} - 1}{\rho^2 - 1}\left[\sum_{t=0}^{T-1}\left\|\left(\mathcal{F}_t' - \mathcal{F}_t\right)\left(\Sigma_0\right)\right\| + \sum_{t=0}^{T-1}\left\|\sum_{s=0}^{t}\left(\mathcal{F}_t' - \mathcal{F}_t\right)\left(W_{\max}\right)\right\|\right] \\
&\le \frac{\rho^{2T} - 1}{\rho^2 - 1}\left[\sum_{t=0}^{T-1}\left\|\mathcal{F}_t' - \mathcal{F}_t\right\|\right]\left(\left\|\Sigma_0\right\| + T\left\|W_{\max}\right\|\right) \\
&\le \frac{\rho^{2T} - 1}{\rho^2 - 1}\left[2\rho\left\|B_{\max}\right\|\sum_{t=0}^{T-1}\|\Delta\| + \left\|B_{\max}\right\|^2 \sum_{t=0}^{T-1}\|\Delta\|^2\right]\left(\frac{V\left(\mathbf{K}, x_0\right)}{\sigma_{\min}\mathbf{Q}} + T\left\|W_{\max}\right\|\right)
\end{aligned}
$$

引理 2.4 证明：

$$
V\left(\mathbf{K}', x_0\right) - V\left(\mathbf{K}^*, x_0\right) = V\left(\mathbf{K}', x_0\right) - V\left(\mathbf{K}, x_0\right) + V\left(\mathbf{K}, x_0\right) - V\left(\mathbf{K}^*, x_0\right)
$$

$$
\begin{aligned}
&V\left(\mathbf{K}', x_0\right)-V\left(\mathbf{K}, x_0\right)\\
&=\sum_{t=0}^{T-1}\left[-2\alpha Tr\left(\Sigma_t'\delta_t^T E_t\right)+\alpha^2 Tr\left(\Sigma_t'\delta_t^T\left(R_t+B_t^T\mathcal{E}_{\theta_t}\left(P_{t+1}\right)B_t\right)\delta_t\right)\right]\\
&=\sum_{t=0}^{T-1}\left[-2\alpha Tr\left(\left(\Sigma_t'-\Sigma_t+\Sigma_t\right)\delta_t^T E_t\right)+\alpha^2 Tr\left(\Sigma_t'\delta_t^T\left(R_t+B_t^T\mathcal{E}_{\theta_t}\left(P_{t+1}\right)B_t\right)\delta_t\right)\right]\\
&\le\sum_{t=0}^{T-1}\left[-2\alpha Tr\left(\delta_t^T\delta_t-\left(\Sigma_t'-\Sigma_t\right)\delta_t^T E_t\Sigma_t\Sigma_t^{-1}\right)+\alpha^2 Tr\left(\Sigma_t'\delta_t^T\left(R_t+B_t^T\mathcal{E}_{\theta_t}\left(P_{t+1}\right)B_t\right)\delta_t\right)\right]\\
&\le\sum_{t=0}^{T-1}\left[-2\alpha Tr\left(\delta_t^T\delta_t\right)+2\alpha\frac{\left\|\Sigma_t'-\Sigma_t\right\|}{\sigma_{\min}\boldsymbol{\Sigma}}Tr\left(\delta_t^T\delta_t\right)\right]+\sum_{t=0}^{T-1}\left[\alpha^2\left\|R_t+B_t^T\mathcal{E}_{\theta_t}\left(P_{t+1}\right)B_t\right\|\left\|\Sigma_t'\right\|Tr\left(\delta_t^T\delta_t\right)\right]\\
&\le\alpha'_{para}\sum_{t=0}^{T-1}Tr\left(\delta_t^T\delta_t\right)
\end{aligned}
$$

$$
\alpha'_{para}:=-2\alpha\sum_{t=0}^{T-1}1-\frac{\left\|\Sigma_t'-\Sigma_t\right\|}{\sigma_{\min}\Sigma_{\mathbf{K}}}-\frac{\alpha}{2}\left\|\Sigma_{\mathbf{K}'}\right\|\left\|R_t+B_t^T\mathcal{E}_{\theta_t}\left(P_{t+1}\right)B_t\right\|
$$

$$
\begin{aligned}
\frac{\sum_{t=0}^{T-1}\left\|\Sigma_t'-\Sigma_t\right\|}{\sigma_{\min}\boldsymbol{\Sigma}}&\le\frac{\rho^{2T}-1}{\rho^2-1}\left[\sum_{t=0}^{T-1}\left\|\mathcal{F}_{K_t'}-\mathcal{F}_{K_t}\right\|\right]\frac{\left(\left\|\Sigma_0\right\|+T\left\|W_{\min}\right\|\right)}{\sigma_{\min}\boldsymbol{\Sigma}}\\
&\le\frac{\rho^{2T}-1}{\rho^2-1}\left[2\rho\left\|B_{\max}\right\|\sum_{t=0}^{T-1}\left\|\Delta_t\right\|+\left\|B_{\max}\right\|^2\sum_{t=0}^{T-1}\left\|\Delta_t\right\|^2\right]\frac{\left(\left\|\Sigma_0\right\|+T\left\|W_{\min}\right\|\right)}{\sigma_{\min}\boldsymbol{\Sigma}}
\end{aligned}
$$

$$
\left\|B_{\max}\right\|\left\|K_t'-K_t\right\|=\alpha\left\|\delta_t\right\|\le\frac{\sigma_{\min}\mathbf{Q}\sigma_{\min}\boldsymbol{\Sigma}}{2V\left(\mathbf{K}, x_0\right)}\le\frac{1}{2}
$$

$$
2\rho\left\|B_{\max}\right\|\sum_{t=0}^{T-1}\left\|\Delta\right\|+\left\|B_{\max}\right\|^2\sum_{t=0}^{T-1}\left\|\Delta\right\|^2\le\left(2\rho+1\right)\left\|B_{\max}\right\|\sum_{t=0}^{T-1}\alpha\left\|\delta_t\right\|
$$

$$
\begin{aligned}
\frac{\sum_{t=0}^{T-1}\left\|\Sigma_t'-\Sigma_t\right\|}{\sigma_{\min}\boldsymbol{\Sigma}}&\le\frac{\rho^{2T}-1}{\rho^2-1}\left[\sum_{t=0}^{T-1}\left\|\mathcal{F}_{K_t'}-\mathcal{F}_{K_t}\right\|\right]\frac{\left(\left\|\Sigma_0\right\|+T\left\|W_{\min}\right\|\right)}{\sigma_{\min}\boldsymbol{\Sigma}}\\
&\le\frac{\rho^{2T}-1}{\rho^2-1}\left[2\rho\left\|B_{\max}\right\|\sum_{t=0}^{T-1}\left\|\Delta_t\right\|+\left\|B_{\max}\right\|^2\sum_{t=0}^{T-1}\left\|\Delta_t\right\|^2\right]\frac{\left(\left\|\Sigma_0\right\|+T\left\|W_{\min}\right\|\right)}{\sigma_{\min}\boldsymbol{\Sigma}}\\
&\le\frac{\rho^{2T}-1}{\rho^2-1}\left[\left(2\rho+1\right)\left\|B_{\max}\right\|\sum_{t=0}^{T-1}\alpha\left\|\delta_t\right\|\right]\frac{\left(V\left(\mathbf{K}, x_0\right)+T\sigma_{\min}\mathbf{Q}\left\|W_{\min}\right\|\right)}{\sigma_{\min}\mathbf{Q}\sigma_{\min}\boldsymbol{\Sigma}}\\
&\le\frac{1}{2}
\end{aligned}
$$

$$
\left\|\Sigma_{\mathbf{K}'}\right\|\le\left\|\Sigma_{\mathbf{K}'}-\Sigma_{\mathbf{K}}\right\|+\left\|\Sigma_{\mathbf{K}}\right\|\le\frac{1}{2}\sigma_{\min}\boldsymbol{\Sigma}+\frac{V\left(\mathbf{K}, x_0\right)}{\sigma_{\min}\mathbf{Q}}\le\frac{1}{2}\left\|\Sigma_{\mathbf{K}'}\right\|+\frac{V\left(\mathbf{K}, x_0\right)}{\sigma_{\min}\mathbf{Q}}
$$

$$
\begin{aligned}
\frac{\alpha}{2}\left\|\Sigma_{\mathbf{K}'}\right\|\sum_{t=0}^{T-1}\left\|R_t+B_t^T\mathcal{E}_{\theta_t}\left(P_{t+1}\right)B_t\right\|&\le\frac{\alpha}{2}\cdot\frac{2V\left(\mathbf{K}, x_0\right)}{\sigma_{\min}\mathbf{Q}}\cdot\sum_{t=0}^{T-1}\left\|R_t+B_t^T\mathcal{E}_{\theta_t}\left(P_{t+1}\right)B_t\right\|\\
&\le\frac{\alpha V\left(\mathbf{K}, x_0\right)}{\sigma_{\min}\mathbf{Q}}T\cdot\max_t\left\|R_t+B_t^T\mathcal{E}_{\theta_t}\left(P_{t+1}\right)B_t\right\|\\
&\le\frac{1}{2}
\end{aligned}
$$

$$
\begin{aligned}
&V\left(\mathbf{K},x_0\right)-V\left(\mathbf{K}^*,x_0\right)\\
&=\sum_{t=0}^{T-1}\left[-2Tr\left(x_t^*\left(x_t^*\right)^T\left(K_t-K_t^*\right)^T E_t\right)-Tr\left(x_t^*\left(x_t^*\right)^T\left(K_t-K_t^*\right)^T\left(R_t+B_t^T\mathcal{E}_{\theta_t}\left(P_{t+1}^{\mathbf{K}^*}\right)B_t\right)\left(K_t-K_t^*\right)\right)\right]\\
&=\sum_{t=0}^{T-1}\left[-Tr\left(x_t^*\left(x_t^*\right)^T\left(K_t-K_t^*+\left(R_t+B_t^T\mathcal{E}_{\theta_t}\left(P_{t+1}^{\mathbf{K}^*}\right)B_t\right)^{-1}E_t\right)^T\left(R_t+B_t^T\mathcal{E}_{\theta_t}\left(P_{t+1}^{\mathbf{K}^*}\right)B_t\right)\right.\right.\\
&\quad\left.\left.\left(K_t-K_t^*+\left(R_t+B_t^T\mathcal{E}_{\theta_t}\left(P_{t+1}^{\mathbf{K}^*}\right)B_t\right)^{-1}E_t\right)\right)+Tr\left(x_t^*\left(x_t^*\right)^T E_t^T\left(R_t+B_t^T\mathcal{E}_{\theta_t}\left(P_{t+1}^{\mathbf{K}^*}\right)B_t\right)^{-1}E_t\right)\right]\\
&\le\sum_{t=0}^{T-1}\left[Tr\left(x_t^*\left(x_t^*\right)^T E_t^T\left(R_t+B_t^T\mathcal{E}_{\theta_t}\left(P_{t+1}^{\mathbf{K}^*}\right)B_t\right)^{-1}E_t\right)\right.\\
&\le\frac{\left\|\Sigma_{\mathbf{K}^*}\right\|}{\sigma_{\min}\mathbf{R}}\sum_{t=0}^{T-1}Tr\left(E_t^T E_t\right)\le\frac{\left\|\Sigma_{\mathbf{K}^*}\right\|}{4\sigma_{\min}\mathbf{R}}\sum_{t=0}^{T-1}Tr\left(\Sigma_t^{-1}\delta_t^T\delta_t\Sigma_t^{-1}\right)\\
&\le\frac{\left\|\Sigma_{\mathbf{K}^*}\right\|}{4\sigma_{\min}\mathbf{R}\sigma_{\min}^2\mathbf{\Sigma}}\sum_{t=0}^{T-1}Tr\left(\delta_t^T\delta_t\right)
\end{aligned}
$$

$$
\begin{aligned}
&V\left(\mathbf{K}',x_0\right)-V\left(\mathbf{K},x_0\right)\\
&\le-2\alpha\frac{4\sigma_{\min}\mathbf{R}\sigma_{\min}^2\mathbf{\Sigma}}{\left\|\Sigma_{\mathbf{K}^*}\right\|}\left[V\left(\mathbf{K},x_0\right)-V\left(\mathbf{K}^*,x_0\right)\right]+\left[V\left(\mathbf{K},x_0\right)-V\left(\mathbf{K}^*,x_0\right)\right]\\
&\le\left(1-\frac{8\alpha\sigma_{\min}\mathbf{R}\sigma_{\min}^2\mathbf{\Sigma}}{\left\|\Sigma_{\mathbf{K}^*}\right\|}\right)\left[V\left(\mathbf{K},x_0\right)-V\left(\mathbf{K}^*,x_0\right)\right]
\end{aligned}
$$