Pure Mathematics 理论数学, 2023, 13(3), 659-668
Published Online March 2023 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2023.133071

Hans汉斯

对带熵的随机线性二次最优控制问题的收敛性证明

舒　心

上海理工大学理学院，上海

收稿日期：2023年2月22日；录用日期：2023年3月23日；发布日期：2023年3月30日

摘　要

本文通过矩阵变换将带熵的随机线性二次最优控制问题的解转化为其等价形式后，证明了线性二次方程的二次项系数解的唯一性和迭代式的收敛性，而一次项系数为0，常数项系数只与二次项有关，控制过程的最优概率分布也只与二次项有关。然后用蒙特卡洛随机抽样样本的均值估计期望值，由此设置了算法1，并证明了算法1中的迭代式具有波动性，波动率的大小和随机参数的方差有关，也与蒙特卡洛中的样本数有关，样本数越多，波动对应的方差越小。然后用两个数值案例比较了随机逼近Q-learning算法和蒙特卡洛Q-learning算法，相同迭代次数下，随机逼近Q-learning算法计算时间更少，但误差更大，蒙特卡洛Q-learning算法收敛更快更稳定，并且可以通过增加随机抽取的样本数使误差更小。

关键词

随机线性二次最优控制，收敛性，Q-Learning，蒙特卡洛，随机逼近

The Proof of the Convergence of Stochastic Linear Quadratic Optimal Control Problem with Entropy

Xin Shu

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

Received: Feb. 22nd, 2023; accepted: Mar. 23rd, 2023; published: Mar. 30th, 2023

Abstract

In this paper, after transforming the solution of the stochastic linear quadratic optimal control

文章引用: 舒心. 对带熵的随机线性二次最优控制问题的收敛性证明[J]. 理论数学, 2023, 13(3): 659-668.
DOI: 10.12677/pm.2023.133071

problem with entropy into its equivalent form through matrix transformation, we prove the uniqueness of the solution of the quadratic coefficient of the linear quadratic equation and the convergence of the iterative formula, and the result shows that the coefficient of the first term is 0, the coefficient of the constant term is only related to the quadratic term, and the optimal probability distribution of the control process is only related to the quadratic term. Then, the mean value of random sampling samples in Monte Carlo is used to estimate the expected value, thus algorithm 1 is set up, and it is proved that the iterative formula in algorithm 1 has volatility, the volatility is related to the variance of random parameters and the number of samples in Monte Carlo, the more sample number, the smaller the variance of the fluctuation. Then, two numerical cases are used to compare Q-learning algorithm with stochastic approximation and Q-learning algorithm with Monte Carlo. Under the same number of iterations, Q-learning algorithm with stochastic approximation takes less time to compute, but the error is larger. Q-learning algorithm with Monte Carlo converges faster and more stable. Moreover, the error can be reduced by increasing the number of randomly selected samples.

Keywords

Stochastic Linear Quadratic Optimal Control, The Convergence, Q-Learning, Monte Carlo, Stochastic Approximation

1. 引言

因为广泛的应用背景，线性二次最优控制问题已经有了很多研究[1]-[6]，根据[7]可知，对于给定初始状态 $x_0 = x \in \mathbb{R}^n$，系统

$$x_{t+1} = A_{t+1}x_t + B_{t+1}u_t = \Lambda_{t+1}\begin{bmatrix} x_t \\ u_t \end{bmatrix},\ t = 0,1,2,\cdots$$

对应的折扣问题的价值函数为

$$V(x) = \min_{\pi \in A(x)} \int \pi_t(u) E\left[r_t(x,u) + \gamma V(x_{t+1}) \mid x_t = x \right] \mathrm{d}u$$

的随机线性二次控制问题，可以根据控制过程的概率分布，用熵增加对控制过程的探索。其中 $\mathcal{A}(x)$ 是 $x_t = x$ 下的可行性控制分布集，$r_t(x,u) = \left[x^{\mathrm{T}}, u^{\mathrm{T}} \right] N_{t+1} \begin{bmatrix} x \\ u \end{bmatrix}$ 是 t 时刻时状态 x 控制概率为 $\pi_t(u)$ 下的即时奖励。假设值函数为二次型时 $V(x) = x^{\mathrm{T}}K_2 x + K_1 x + K_0$，可求得随机控制的最优概率分布

$$\begin{aligned} \pi^*(u) &= \frac{\exp\left\{ -\frac{1}{\lambda}\left(u^{\mathrm{T}}M_{uu}u + u^{\mathrm{T}}M_{ux}x + x^{\mathrm{T}}M_{xu}u + x^{\mathrm{T}}M_{xx}x + N_u u + N_x x + K_0 \right) \right\}}{\int \exp\left\{ -\frac{1}{\lambda}\left(u^{\mathrm{T}}M_{uu}u + u^{\mathrm{T}}M_{ux}x + x^{\mathrm{T}}M_{xu}u + x^{\mathrm{T}}M_{xx}x + N_u u + N_x x + K_0 \right) \right\} \mathrm{d}u} \\ &= \mathcal{N}\left(u \mid -\frac{M_{uu}^{+}}{2}\left(2M_{ux}x + N_u^{\mathrm{T}} \right), \frac{\lambda}{2}M_{uu}^{+} \right) \end{aligned}$$

并根据概率分布可以求得线性二次最优控制各项系数的迭代式：

$$K_2=\Pi\left(E\left(N_{t+1}+\gamma\Lambda_{t+1}^{\mathrm{T}}K_2\Lambda_{t+1}\right)\right)$$

$$K_1=\Gamma_1\left(E\left(N_{t+1}+\gamma\Lambda_{t+1}^{\mathrm{T}}K_2\Lambda_{t+1}\right),E\left(\gamma K_1\Lambda_{t+1}\right)\right)$$

$$K_0=\frac{1}{1-\gamma}\Gamma_0\left(M\left(K_2\right),N\left(K_1\right)\right)$$

其中的函数映射定义为

$$\Pi(P):=P_{xx}-P_{xu}P_{uu}^{+}P_{ux}\text{ with }P_{xx}\in\mathbb{R}^{n\times n},P_{xu}\in\mathbb{R}^{n\times m}$$

$$\Gamma_1(P,Q):=Q_x-Q_uP_{uu}^{+}P_{ux}\text{ with }Q_x\in\mathbb{R}^{1\times n},Q_u\in\mathbb{R}^{1\times m}$$

$$\Gamma_0(P,Q):=-\frac{1}{4}Q_uP_{uu}^{+}Q_u^{\mathrm{T}}+\frac{\lambda}{2}\ln\left((\lambda\pi e)^m\det\left(P_{uu}^{+}\right)\right)+\frac{\lambda m}{2}$$

$$M\left(K_2\right)=E\left(N_{t+1}+\gamma\Lambda_{t+1}^{\mathrm{T}}K_2\Lambda_{t+1}\right)$$

$$N\left(K_1\right)=E\left(\gamma K_1\Lambda_{t+1}\right)$$

根据推导出来的系数的迭代公式定义函数

$$Q_2^*=E\left[N+\gamma\Lambda^{\mathrm{T}}K_2\Lambda\right]$$

$$Q_1^*=E\left[\gamma K_1\Lambda\right]$$

由此可得到

$$Q_2^*=E\left[N+\gamma\Lambda^{\mathrm{T}}\Pi\left(Q_2^*\right)\Lambda\right]$$

$$Q_1^*=E\left[\gamma\Gamma_1\left(Q_2^*,Q_1^*\right)\Lambda\right]$$

所以最终 $K_2=\Pi\left(Q_2^*\right)$，$K_1=\Gamma_1\left(Q_2^*,Q_1^*\right)$。

参考文献[7]只给出了问题解的迭代形式，但是没有证明解的唯一性以及根据迭代式算得的结果的收敛性。因此本文在参考文献[7]的基础上继续研究，证明了 Q_2^* 和 Q_1^* 解的唯一性，同时确定 $Q_1^*=0$，所以在后续计算过程中不需要再继续计算 Q_1^*，减少了计算过程。然后证明了当期望可准确求出时，迭代得到的 $Q_2(t)$ 是收敛的，但使用蒙特卡洛近似期望后，$Q_2(t)$ 具有波动性，波动方差与随机参数有关，但是也可以通过调节蒙特卡洛方法中的随机抽样样本数来减少误差。

2. 解的唯一性

为了求得 Q_2^* 和 Q_1^*，使用迭代式

$$Q_2(t+1)=Q_2(t)+\alpha\left[E\left(N_{t+1}+\gamma\Lambda_{t+1}^{\mathrm{T}}\Pi\left(Q_2(t)\right)\Lambda_{t+1}\right)-Q_2(t)\right]$$

$$Q_1(t+1)=Q_1(t)+\alpha\left[E\left(\gamma\Gamma_1\left(Q_2(t),Q_1(t)\right)\Lambda_{t+1}\right)-Q_1(t)\right]$$

但是参考文献[7]没有证明 Q^* 解的唯一性以及 $Q(t)$ 的收敛性，因此本文继续证明 Q 值的收敛问题。假设线性二次最优控制问题有解，可以将问题转化为等价形式再求解。假设 $E\left[\|N\|^2+\gamma\left\|\Lambda^{\mathrm{T}}\Lambda\right\|^2\right]$ 是有限的，$E[N]$ 是正定矩阵。首先定义一个关于 $Q_2(t)$ 的函数

$$F_2\left(Q_2(t)\right)\triangleq E\left(N_{t+1}+\gamma\Lambda_{t+1}^{\mathrm{T}}\Pi\left(Q_2(t)\right)\Lambda_{t+1}\right)$$

所以

$$Q_2(t+1)=Q_2(t)+\alpha\left[F_2(Q_2(t))-Q_2(t)\right]=\alpha F_2(Q_2(t))+(1-\alpha)Q_2(t)$$

根据文献[6]的 Theorem1.1 可知 Q_2^* 有唯一解，$Q_2(t)$ 是收敛的。

然后证明 $Q_1(t)$ 的收敛性。因为 $E[N]$ 是正定的，所以存在 $\varepsilon_0\in(0,1)$ 使得 $E[N]\geq\varepsilon_0 I_d$ (矩阵 $A\geq B$ 意味着 $A-B$ 是半正定矩阵)，所以

$$K_2=\Pi\left(E\left(N+\gamma\Lambda^{\mathrm{T}}K_2\Lambda\right)\right)\geq\Pi\left(E[N]\right)>0$$

同理 $E\left[Q_2^*\right]>0$，令 L 和 M 为可逆矩阵且

$$L^{\mathrm{T}}KL=I_n,M^{\mathrm{T}}Q_{2uu}^*M=I_m$$

根据文献[6]的证明过程，定义一个 $n+m$ 的矩阵

$$C:=\begin{bmatrix}I_n & O\\ \Gamma & I_m\end{bmatrix}\begin{bmatrix}L & O\\ O & M\end{bmatrix}=\begin{bmatrix}L & O\\ \Gamma L & M\end{bmatrix}$$

其中 $\Gamma:=\Gamma\left(Q_2^*\right)=-\left(Q_{2uu}^*\right)^{-1}Q_{2ux}^*$ 是矩阵 Q_2 的子矩阵运算的简化形式。利用矩阵 C 首先对 Q_2 及相关矩阵做矩阵变换

$$\tilde{\Lambda}_t:=L^{-1}\Lambda_t C$$

$$\tilde{N}_t:=C^{\mathrm{T}}N_t C$$

$$\tilde{Q}_2(t):=C^{\mathrm{T}}Q_2(t)C$$

变换后的矩阵具有以下关系

$$\tilde{Q}_2=C^{\mathrm{T}}Q_2^*C=\begin{bmatrix}I_n & O\\ O & I_m\end{bmatrix}$$

然后对矩阵 $Q_1(t)$ 做变换

$$\tilde{Q}_1(t):=Q_1(t)C$$

变换后可得到以下关系式

$$\tilde{Q}_1=Q_1^*C=E\left[\gamma\Gamma_1\left(\tilde{Q}_2,\tilde{Q}_1\right)\tilde{\Lambda}\right]=\left[\gamma\tilde{Q}_{1x}\tilde{\Lambda}\right]$$

然后定义一个关于 $\tilde{Q}_1$ 的函数

$$T\left(\tilde{Q}_1(t)\right)=E\left[\gamma\tilde{Q}_{1x}(t)\tilde{\Lambda}_t\right]$$

根据 T 函数定义可得 $0=T(0)$，因为 $E[N]>0$ 可以推得 $E\left[\tilde{N}\right]>0$，又因为 $\tilde{Q}_2$ 的性质 $I_d=C^{\mathrm{T}}Q_2^*C=E\left[\tilde{N}+\gamma\tilde{\Lambda}^{\mathrm{T}}\tilde{\Lambda}\right]$，所以存在一个正数 $\beta<1$ 使得

$$E\left(\tilde{\Lambda}^{\mathrm{T}}\tilde{\Lambda}\right)=I_d-E\left[\tilde{N}\right]\leq\beta I_d$$

根据矩阵 2 范数的定义知 $\left\|E\left[\tilde{\Lambda}\right]\right\|_2=\sqrt{\lambda_{\max}}$，$\lambda_{\max}$ 表示为 $\tilde{\Lambda}^{\mathrm{T}}\tilde{\Lambda}$ 的最大特征值，因为 $E\left[\tilde{\Lambda}^{\mathrm{T}}\tilde{\Lambda}\right]\leq\beta I_d$，所以 $\lambda_{\max}\leq\beta$，所以

$$\frac{\left\|T\left(\tilde{Q}_1\right)\right\|_2}{\left\|\tilde{Q}_1\right\|_2}=\frac{\left\|E\left[\gamma\tilde{Q}_{1x}\tilde{\Lambda}\right]\right\|_2}{\left\|\tilde{Q}_1\right\|_2}\leq\frac{\left\|\gamma\tilde{Q}_{1x}\right\|\left\|E\left[\tilde{\Lambda}\right]\right\|_2}{\left\|\tilde{Q}_1\right\|_2}\leq\gamma\sqrt{\beta}<1$$

所以 $T(\cdot)$ 是关于矩阵 2 范数的压缩映射，且不动点为 0。所以对于迭代式

$$\tilde{Q}_1(t+1)=\tilde{Q}_1(t)+\alpha_t\left(T\left(\tilde{Q}_1(t+1)\right)-\tilde{Q}_1(t)\right)$$

此时 $\tilde{Q}_1(t)$ 是收敛的。

又因为 $\tilde{Q}_1=T\left(\tilde{Q}_1\right)$，所以 $\tilde{Q}_1$ 是 $T(\cdot)$ 的不动点，由不动点唯一性可得 $\tilde{Q}_1=0$，所以最终线性二次问题只需要求二次项系数 K_2 和常数项 K_0 的值。

$$K_2=\Pi\left(M\left(K_2\right)\right)=\Pi\left(E\left(N_{t+1}+\gamma\Lambda_{t+1}^{\mathrm{T}}K_2\Lambda_{t+1}\right)\right)$$

$$K_0=\frac{1}{1-\gamma}\left[\frac{\lambda}{2}\ln\left((\lambda\pi e)^m\det\left(Q_{2uu}^{+}\right)\right)+\frac{\lambda m}{2}\right]$$

3. 算法及其收敛性

根据 Q-learning 算法，迭代公式为

$$Q_2(t+1)=Q_2(t)+\alpha\left[E\left(N_{t+1}+\gamma\Lambda_{t+1}^{\top}\Pi\left(Q_2(t)\right)\Lambda_{t+1}\right)-Q_2(t)\right]$$

因为参数是随机的，所以无法求得准确期望值，因此使用蒙特卡洛法求期望的近似值。蒙特卡洛由 Metropolis [8]提出，是根据随机抽样来估计数学函数或者复杂系统[9]。蒙特卡洛应用广泛于不同领域，目前已有大量理论研究和实践经验[10] [11] [12]。蒙特卡洛通过随机抽样生成大量样本，然后用样本统计量估计问题的解。本文中使用抽样样本均值来估计期望值，计算方法为

$$E\left(N_{t+1}+\gamma\Lambda_{t+1}^{\mathrm{T}}\Pi\left(Q_2(t)\right)\Lambda_{t+1}\right)\approx\frac{1}{s}\sum_{k=1}^{s}\left[N_k+\gamma\Lambda_k^{\mathrm{T}}\Pi\left(Q_2(t)\right)\Lambda_k\right]$$

其中 s 为随机生成的样本数，本文中设置样本数量为 $s=200$。由此设置算法 1，见表 1。其中学习率满足 $\sum_{t=0}^{\infty}\alpha_t=+\infty$，$\sum_{t=0}^{\infty}\alpha_t^2<+\infty$。

Table 1. Q-learning algorithm with Monte-Carlo
表 1. 蒙特卡洛 Q-learning 算法

算法 1：蒙特卡洛 Q-learning 算法
1) 初始化矩阵使 $Q_2(0)=I_d$
2) for $t=0,T$:
3) $Q_2(t+1)=Q_2(t)+\alpha\left[\frac{1}{s}\sum_{k=1}^{s}\left[N_k+\gamma\Lambda_k^{\mathrm{T}}\Pi\left(Q_2(t)\right)\Lambda_k\right]-Q_2(t)\right]$
4) $K_0(t+1)=\frac{1}{1-\gamma}\Gamma_0\left(Q_2(t+1),0\right)$
5) $t=t+1$
6) end for

然后证明算法 1 (见表 1)中 $Q_2(t)$ 的收敛性。根据算法 1 中的迭代式定义关于 $Q_2(t)$ 的一个函数

$$\tilde{F}_2\left(Q_2(t)\right)\triangleq\frac{1}{s}\sum_{k=1}^{s}\left[N_k+\gamma\Lambda_k^{\mathrm{T}}\Pi\left(Q_2(t)\right)\Lambda_k\right]$$

所以 $Q_2(t+1)=Q_2(t)+\alpha\left[\tilde{F}_2\left(Q_2(t)\right)-Q_2(t)\right]=\alpha\tilde{F}_2\left(Q_2(t)\right)+(1-\alpha)Q_2(t)$，可以转化为以下等式

$$\begin{aligned}Q_2(t+1)&=\alpha\tilde{F}_2\left(Q_2(t)\right)+(1-\alpha)Q_2(t)\\&=\alpha F_2\left(Q_2(t)\right)+(1-\alpha)Q_2(t)+\alpha\left[\tilde{F}_2\left(Q_2(t)\right)-F_2\left(Q_2(t)\right)\right]\end{aligned}$$

因为前半部分 $Q_2(t+1)=\alpha F_2\left(Q_2(t)\right)+(1-\alpha)Q_2(t)$ 是收敛的，所以主要考虑 $\tilde{F}\left(Q_2(t)\right)-F_2\left(Q_2(t)\right)$ 的收敛性。因为随机变量独立同分布，所以 $E\left[\tilde{F}_2\left(Q_2(t)\right)\right]=F_2\left(Q_2(t)\right)$，所以 $E\left[\left(\tilde{F}\left(Q_2(t)\right)-F_2\left(Q_2(t)\right)\right)\right]=0$，令 $N_k+\gamma\Lambda_k^{\mathrm{T}}\Pi\left(Q_2(t)\right)\Lambda_k=x_k$，

$$\begin{aligned}var\left[\tilde{F}_2\left(Q_2(t)\right)-F_2\left(Q_2(t)\right)\right]&=E\left[\left(\tilde{F}_2\left(Q_2(t)\right)-F_2\left(Q_2(t)\right)\right)^2\right]\\&=E\left[\left(\frac{1}{s}\sum_{k=1}^{s}x_k-E\left(\frac{1}{s}\sum_{k=1}^{s}x_k\right)\right)^2\right]\\&=var\left(\frac{1}{s}\sum_{k=1}^{s}x_k\right)\end{aligned}$$

这里的方差代表着矩阵的协方差矩阵，因为方差不为 0，所以算法 1 中的 $Q_2(t)$ 不收敛。根据方差计算结果可知，$Q_2(t)$ 在一定范围内波动，波动率与随机参数的方差有关，也与随机生成抽样的样本数有关，样本数越多，对应的方差越小。

然后根据随机逼近理论考虑另一种算法。Robbins-Monro 首创了随机逼近理论[13]，然后被快速应用到随机优化问题中[14] [15]。随机逼近算法是为了解决寻根问题，其中的函数是一个期望值。[16]结合随机逼近思想证明了 Q-learning 的收敛性，[6]将随机逼近算法应用到线性二次问题中。根据文献[6]，本文也考虑使用随机逼近算法。根据随机逼近思想，迭代式为

$$Q_2(t+1)=Q_2(t)+\alpha\left(N_{t+1}+\gamma\Lambda_{t+1}^{\mathrm{T}}\Pi\left(Q_2(t)\right)\Lambda_{t+1}-Q_2(t)\right)$$

因此设置算法 2，见表 2。Dai [6]已经证明线性二次问题中使用随机逼近的 Q-learning 算法是收敛的。

Table 2. Q-learning algorithm with stochastic approximation
表 2. 随机逼近 Q-learning 算法

算法 2：随机逼近 Q-learning 算法
1) 初始化矩阵使 $Q_2(0)=I_d$
2) for $t=0,T$:
3) $Q_2(t+1)=Q_2(t)+\alpha\left(N_{t+1}+\gamma\Lambda_{t+1}^{\mathrm{T}}\Pi\left(Q_2(t)\right)\Lambda_{t+1}-Q_2(t)\right)$
4) $K_0(t+1)=\frac{1}{1-\gamma}\Gamma_0\left(Q_2(t+1),0\right)$
5) $t=t+1$
6) end for

4. 数值分析

比较两种算法的可行性和有效性，选择两个不同的例子，其中折扣因子 $\gamma=0.99$，算法中的学习率 $\alpha=\frac{t}{5+t}$。首先考虑状态空间 $n=2$，控制空间 $m=1$ 时的情况，此时使 $\Lambda_t=\Lambda^{(0)}+\omega_t^{(1)}\Lambda^{(1)}+\omega_t^{(2)}\Lambda^{(2)}$，$N_t=N^{(0)}$，其中 $\omega_t^{(1)}$，$\omega_t^{(2)}$ 独立同分布且服从标准正态分布。

$$\Lambda^{(0)}=\begin{bmatrix}-1 & -0.1 & -0.2\\ 2.6 & 0.5 & 0.5\end{bmatrix},\quad \Lambda^{(1)}=\begin{bmatrix}0.6 & 0.075 & 0.125\\ -0.8 & 0.1 & -0.375\end{bmatrix}$$

$$\Lambda^{(2)}=\begin{bmatrix}-0.06 & -0.06 & 0.02\\ 0.2 & 0.23 & -0.09\end{bmatrix},\quad N^{(0)}=\begin{bmatrix}3.11 & 1.5626 & -0.2798\\ 1.5626 & 1.816175 & -1.021425\\ -0.2798 & -1.021425 & 0.91585\end{bmatrix}$$

根据算法收敛性证明过程知道，蒙特卡洛算法的波动大小和随机抽样的样本数有关，样本量越大方差越小。

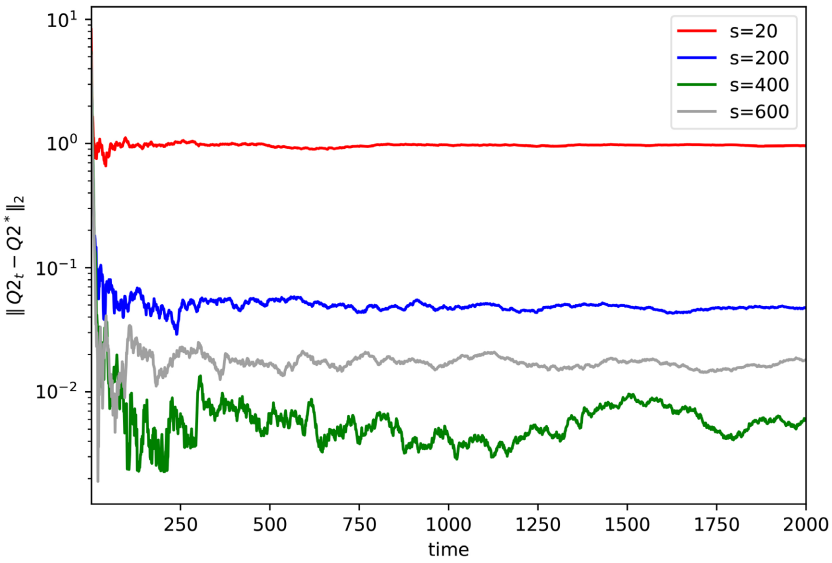

Figure 1. The convergence for different values of s
图 1. s 取不同值时的收敛情况

由图 1 可以看出，随着抽样数 s 的增加，误差越来越小(见图 1)。蒙特卡洛 Q-learning 可以通过改变 s 值来控制误差，但是抽样数 s 越大计算所需时间也越长，因此从计算时长和精确度考虑，选 $s=200$。

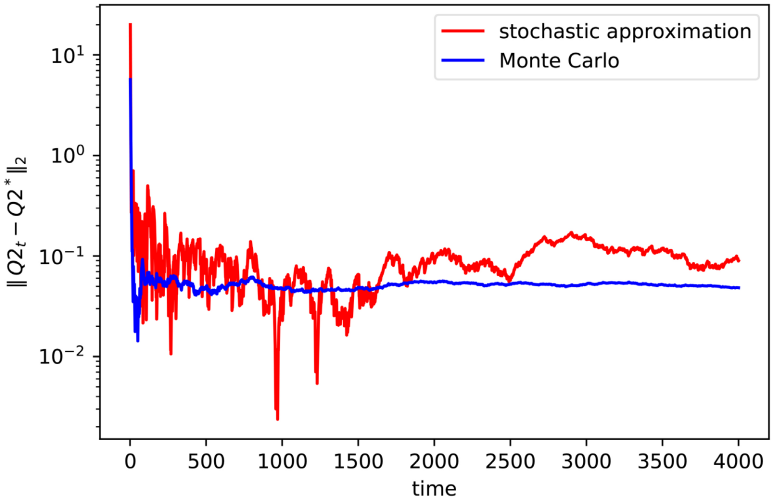

Figure 2. The convergence of Q_2 when $n=2$
图 2. $n=2$ 时 Q_2 的收敛情况

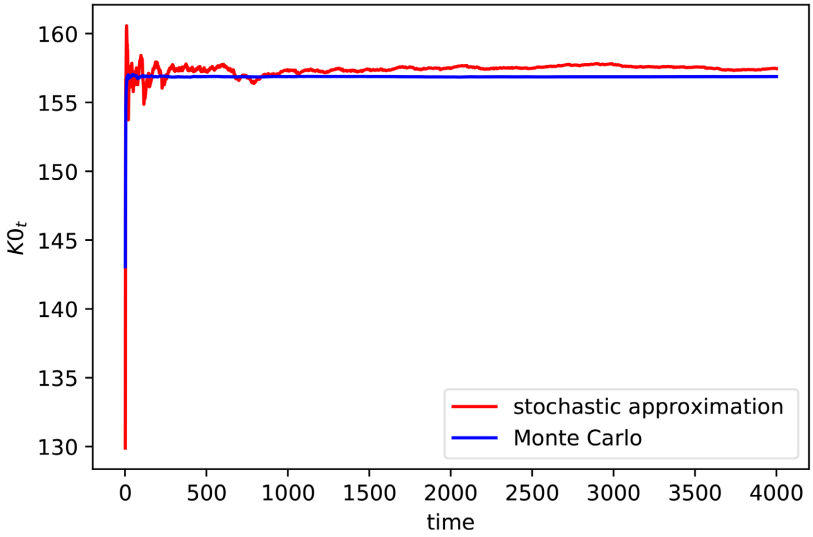

Figure 3. The convergence of K_0 when $n = 2$
图 3. $n = 2$ 时 K_0 的收敛情况

图 2 是状态空间 $n = 2$ 时，不同算法下真实值 Q^* 与算法计算出来的值 $Q_2(t)$ 之间的差值取对数后的结果(见图 2)，图 3 是根据 $Q_2(t)$ 计算得到 K_0 (见图 3)。

然后考虑状态空间 $n = 3$，控制空间 $m = 1$ 时的情况，令 $\Lambda_t = \Lambda^{(0)} + \omega_t^{(1)}\Lambda^{(1)}$，其中 $\omega_t^{(1)}$ 标准正态分布。

$$\Lambda^{(0)} = \begin{bmatrix} -0.7718 & 0.3632 & 0.1619 & 0.7298 \\ 0.0335 & 0.1955 & -0.0709 & 0.3275 \\ -0.0738 & 0.2609 & 0.5275 & -0.5730 \end{bmatrix},$$

$$\Lambda^{(1)} = \begin{bmatrix} -0.4505 & 0.0671 & 0.1783 & 0.1651 \\ -0.900 & -0.0628 & -0.1045 & -0.4122 \\ -0.6539 & -0.4185 & -0.2444 & 0.9814 \end{bmatrix}, \quad N^{(0)} = I_d$$

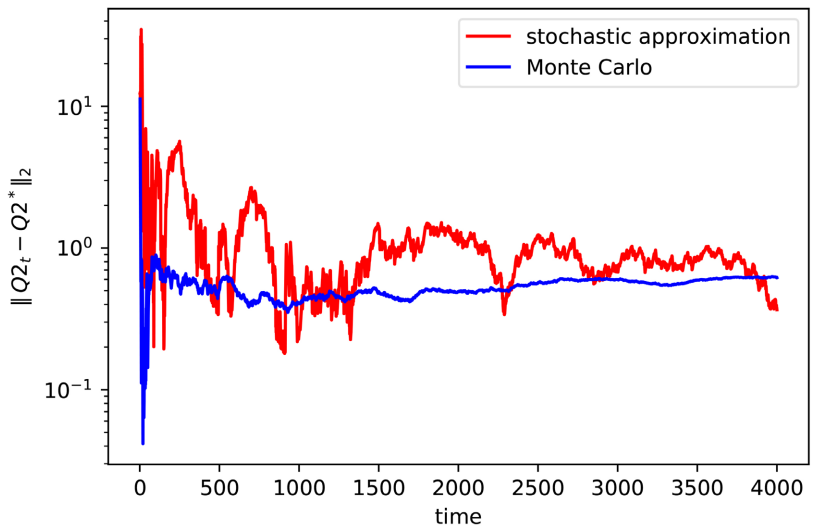

Figure 4. The convergence of Q_2 when $n = 3$
图 4. $n = 3$ 时 Q_2 的收敛情况

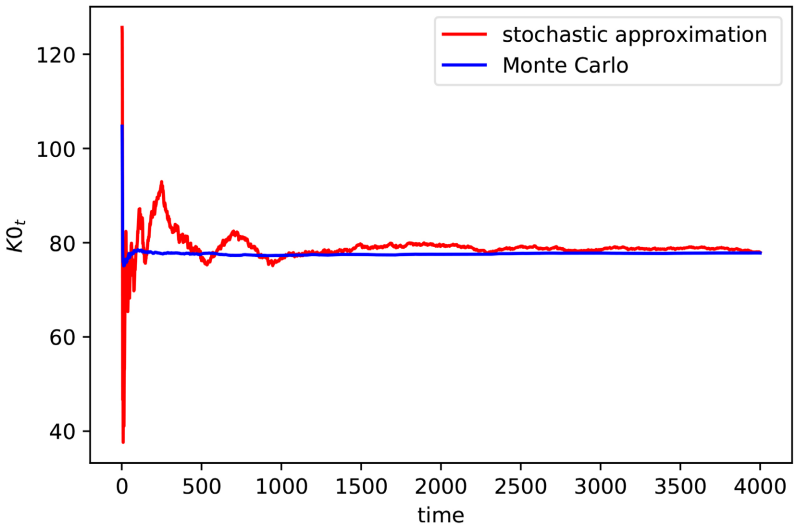

Figure 5. The convergence of K_0 when $n = 3$
图 5. $n = 3$ 时 K_0 的收敛情况

同样的，图 4 是状态空间 $n = 3$ 时，不同算法的计算误差取对数后的值(见图 4)，图 5 是对应的 K_0 结果(见图 5)。

根据数值实例的运算结果可以看出，通过改变抽取的随机样本的个数，我们可以很容易地改变蒙特卡洛 Q-learning 算法的精确度。对于不同的实例两种算法都能使结果趋近于真实值。蒙特卡洛 Q-learning 算法收敛更快更稳定误差值更小，迭代次数大于 1000 次以后，误差值就趋于稳定，而随机逼近 Q-learning 算法计算的结果误差一直存在波动，误差总体大于蒙特卡洛 Q-learning 算法。在相同迭代次数下，随机逼近 Q-learning 算法计算时间所用更少，蒙特卡洛 Q-learning 算法计算结果更准确。

5. 结论

本文在原有加熵的随机线性二次最有控制问题的基础上，证明了解的唯一性，然后证明了算法中迭代式的收敛性。结果表示对于一般形式的线性二次最优控制问题，加熵后根据贝尔曼原理求得的一次项系数为零，常数项系数只与二次项系数有关，由此在后续计算中只需要计算二次项系数与常数项系数的值，减少了运算步骤。然后证明了蒙特卡洛 Q-learning 算法中，由于用样本均值代替原有期望，所以根据迭代式求得的值具有波动性，波动率大小和随机参数的方差有关，也与蒙特卡洛算法中选取的样本数量有关，样本数量越多对应的方差越小，反之波动越大。最后比较了蒙特卡洛 Q-learning 算法和随机逼近 Q-learning 算法。迭代相同次数下随机逼近 Q-learning 算法计算更快，蒙特卡洛 Q-learning 算法收敛更快更稳定，而且可以通过改变 s 值的大小改变算法的准确度。

致　谢

感谢导师的辛苦指导，给了我很多意见和帮助。感谢师兄师妹的鼓励和帮助。

参考文献

[1] Pronzato, L., Kulcsár, C. and Walter, E. (1996) An Actively Adaptive Control Policy for Linear Models. *IEEE Transactions on Automatic Control*, **41**, 855-858. https://doi.org/10.1109/9.506238

[2] Chen, S., Li, X. and Zhou, X.Y. (1998) Stochastic Linear Quadratic Regulators with Indefinite Control Weight Costs. *SIAM Journal on Control and Optimization*, **36**, 1685-1702. https://doi.org/10.1137/S0363012996310478

[3] Chen, S. and Zhou, X.Y. (2000) Stochastic Linear Quadratic Regulators with Indefinite Control Weight Costs. II. *SIAM Journal on Control and Optimization*, **39**, 1065-1081. https://doi.org/10.1137/S0363012998346578

[4] Rami, M.A., Moore, J.B. and Zhou, X.Y. (2002) Indefinite Stochastic Linear Quadratic Control and Generalized Differential Riccati Equation. *SIAM Journal on Control and Optimization*, **40**, 1296-1311. https://doi.org/10.1137/S0363012900371083

[5] Wang, T., Zhang, H. and Luo, Y. (2016) Infinite-Time Stochastic Linear Quadratic Optimal Control for Unknown Discrete-Time Systems Using Adaptive Dynamic Programming Approach. *Neurocomputing*, **171**, 379-386. https://doi.org/10.1016/j.neucom.2015.06.053

[6] Du, K., Meng, Q. and Zhang, F. (2022) A Q-Learning Algorithm for Discrete-Time Linear-Quadratic Control with Random Parameters of Unknown Distribution: Convergence and Stabilization. *SIAM Journal on Control and Optimization*, **60**, 1991-2015. https://doi.org/10.1137/20M1379605

[7] 舒心. 带熵的随机线性二次最优控制问题[J]. 应用数学进展, 2022, 11(12): 8836-8845. https://doi.org/10.12677/AAM.2022.1112931

[8] Metropolis, N. and Ulam, S. (1949) The Monte Carlo Method. *Journal of the American Statistical Association*, **44**, 335-341. https://doi.org/10.1080/01621459.1949.10483310

[9] Harrison, R.L. (2010) Introduction to Monte Carlo Simulation. *AIP Conference Proceedings*, **1204**, 17-21. https://doi.org/10.1063/1.3295638

[10] James, F. (1980) Monte Carlo Theory and Practice. *Reports on Progress in Physics*, **43**, Article No. 1145. https://doi.org/10.1088/0034-4885/43/9/002

[11] Glasserman, P. (2004) Monte Carlo Methods in Financial Engineering. Springer, New York. https://doi.org/10.1007/978-0-387-21617-1

[12] Ferrenberg, A.M. and Swendsen, R.H. (1988) New Monte Carlo Technique for Studying Phase Transitions. *Physical Review Letters*, **61**, 2635-2638. https://doi.org/10.1103/PhysRevLett.61.2635

[13] Robbins, H. and Monro, S. (1951) A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, **22**, 400-407. https://doi.org/10.1214/aoms/1177729586

[14] Lai, T.L. (2003) Stochastic Approximation. *The Annals of Statistics*, **31**, 391-406. https://doi.org/10.1214/aos/1051027873

[15] Nemirovski, A., Juditsky, A., Lan, G. and Shapiro, A. (2009) Robust Stochastic Approximation Approach to Stochastic Programming. *SIAM Journal on Optimization*, **19**, 1574-1609. https://doi.org/10.1137/070704277

[16] Tsitsiklis, J.N. (1994) Asynchronous Stochastic Approximation and Q-Learning. *Machine Learning*, **16**, 185-202. https://doi.org/10.1007/BF00993306