形式熵　　正则化随机系统控制　　ms

M. J. T. C. van Zutphen、G. Delimpaltadakis、D. Antunes

arXiv:2603.05021v1 [eess.SY] 5 Mar 2026

摘要：分析和控制系统熵是调控控制系统可预测性的有力工具。受益于此类方法的应用涵盖从强化学习、数据安全到人机协作等多个领域。在连续状态随机系统中，精确的熵分析与控制仍是一项挑战。近年来，连续系统的有限状态抽象使控制综合能够针对阶段成本等目标提供形式性能保证。然而，这些结果并不能扩展到基于熵的性能度量。我们通过如下方式解决这一问题：首先利用传统形式抽象结果获得系统离散化熵的界，然后再获得连续分布的熵与其离散化熵之差的附加界。由此产生的理论能够实现形式化熵感知控制器综合，在可预测性与控制性能之间进行权衡，同时为原始连续系统保留形式保证。更具体地说，我们聚焦于最小化系统轨迹分布相对于均匀分布的 KL 散度——即我们的系统熵度量——与一般累积成本的线性组合。我们指出，针对给定连续分布相对于均匀分布的 KL 散度与其离散化之差所推导的界，在更一般的信息论背景中也可能相关。一组案例研究展示了该方法的有效性。

1 引言

在随机系统中，熵这一概念提供了一种量化可预测性的度量，用以刻画系统可能行为的随机程度或可预测程度。因此，熵优化——即最大化或最小化系统熵——已成为多个领域中的重要概念 [1–13]。提高可预测性有助于提升自动驾驶汽车的乘客舒适度[1]和扰动鲁棒性 [2] ，并促进 [3–5] 更广泛的自主移动智能体之间的协作。相比之下，在数据安全应用中，较高程度的不可预测性对于确保机密性至关重要 [6],，例如监控任务 [7, 8],，防止可预测的自动驾驶汽车行为遭到利用 [9, 10],，以及促进探索并提升强化学习（RL）中的学习鲁棒性 [4, 11–14]。

有限状态随机系统中可预测性的量化通常依赖香农提出的信息熵概念 [15, 16]。尽管近期强化学习文献通常会考虑奖励（或惩罚）输入选择中的随机程度，即策略熵 [11–13], ，我们的工作关注的是对系统行为随机程度的量化与优化，即轨迹熵。对马尔可夫系统中的轨迹熵进行高效分析通常并非易事，因此一直是持续研究的主题，下面我们将对此作简要回顾。

早期关于轨迹熵优化的研究考察了这样一个问题：寻找与给定区间马尔可夫链（IMC）规范一致的最大熵马尔可夫链（MC）；区间马尔可夫链是对马尔可夫链的一种推广，其中转移概率受到约束，必须处于规定的不确定区间内[17]。后续工作 [18] 研究了计算马尔可夫链、区间马尔可夫链和马尔可夫决策过程（MDP）的（最大）熵与熵率的复杂性。之后的研究将这种马尔可夫决策过程方法 [18] 扩展为包含时序逻辑约束 [19, 20], ，开发出能够构造性地获得相应最优策略的算法，并将其应用于实际的监视任务 [21]。文献 [22]提出了一个平衡成本与轨迹熵的贝尔曼方程，以及一个相关的强化学习框架，该框架通过在成本约束下最大化轨迹熵来促进无偏探索。文献 [5]针对建模为区间马尔可夫决策过程（IMDP）的不确定系统，获得了能够最小化熵正则化目标形式上界的策略。文献 [23], 研究了通过最大熵策略实现高效的强化学习探索，再次表明，可以通过求解熵正则化贝尔曼方程找到最大化轨迹熵的策略。文献 [4], 提出了强化学习中的轨迹熵惩罚方法，使得人类和其他智能体能够更容易地预测这些智能体的行为。一种非常相似的方法则最小化动作轨迹熵 [24]。

这些现有方法从根本上都与有限状态设定紧密关联，而许多所关注的系统会自然地在连续域上演化。抽象方法已被发展出来，以便通过精心构造的有限状态表示 [25–35]，对具有复杂规约的连续状态随机系统进行形式化分析与控制。通过对原始系统的状态和动作空间进行离散化，一个有限状态模型——

即该抽象——以一种形式化方式被构造；也就是说，伴随在该抽象上进行的验证与控制设计的形式保证可以传递到原始系统（例如，若该抽象是“安全的”，则原始系统也是“安全的”）。区间马尔可夫决策过程（区间马尔可夫决策过程）通过允许转移概率在区间内变化来推广马尔可夫决策过程，已成为（不确定的）连续马尔可夫系统的标准抽象形式化框架 [5, 25–32, 36]。这些方法借助有限模型，在一般设定下为验证与综合任务提供可靠且计算上可处理的保证。

然而，熵最小化或最大化所面临的独特挑战，无法像期望累积阶段成本 [26,28,30–32,36], 或线性时序逻辑规范 [25,28、30、37–39] 那样自然地融入现有的抽象框架。在本文中，我们证明，将现有的基于抽象的方法应用于系统熵时，无法保留或界定底层系统的熵属性。随后，我们提出一种经过调整的框架，以解决这些不足。

因此，本文的主要目的是建立一套关于基于抽象的熵保证的形式化理论，即通过有限状态抽象，实现对连续状态马尔可夫系统熵的形式化分析与控制。在我们的框架中，我们用轨迹分布相对于均匀分布的 Kullback–Leibler（KL）散度 [16]（下文简称相对于均匀分布的 *KL* 散度）来量化熵；它是系统熵的一种性质良好且无损的替代指标。

我们首先得出若干结果，从而能够通过区间马尔可夫链抽象，形式化地计算连续马尔可夫链轨迹相对于均匀分布的 KL 散度下界。随后，在适当假设下，我们将这些结果加以扩展，以构造同一 KL 散度的形式化上界。在此过程中，我们推导出给定连续分布相对于均匀分布的 KL 散度与其离散化后的相应 KL 散度之间差值的界；这些结果在更广泛的信息论背景下也具有独立意义。

随后，我们将该方法扩展到基于抽象的策略综合，用于连续状态马尔可夫决策过程。在上述构造形式化上界的两种不同方法的基础上，我们提出了一种算法，能够找到使各自上界最小化的策略。最后，我们通过一个数值示例展示离散化分辨率对界的紧致性的影响。我们还考虑一个连续状态控制问题，在该问题中，我们通过熵惩罚项对累积代价优化进行正则化。由我们生成的策略所产生的系统行为，使最终系统轨迹的熵降低。

本文结构安排如下。在第 2 节中，我们介绍必要的预备概念。我们介绍连续状态马尔可夫链，并阐述其两种离散化方法：均值轨迹分布离散化和区间马尔可夫链（区间马尔可夫链）抽象。随后，我们简要讨论紧连续状态马尔可夫系统背景下系统熵的量化。在第 3 节中，我们通过有限状态抽象，针对连续状态马尔可夫系统上的系统熵（以轨迹分布相对于均匀分布的 KL 散度表示）保证，给出主要结果。在第 4 节中，我们将这些结果扩展到马尔可夫决策过程（马尔可夫决策过程），并提出一种界最小化策略综合算法。在第 5 节中，我们给出两个数值实验的结果，以展示上述系统熵抽象方法的有效性。最后，我们在第 6 节中总结全文。

2 马尔可夫链与熵

设 λ 表示 $\mathbb{R}^n$上的勒贝格测度。对于可测集 $A \subseteq \mathbb{R}^n$，我们将 A 上的概率密度函数空间定义为

$$\mathcal{P}(A) := \{q : A \to \mathbb{R}_{\geq 0} \mid q \text{ measurable}, \int_A q(x)\, \mathrm{d}x = 1\},$$

即，所有函数 $q \in \mathcal{P}(A)$均对应概率测度 $\mu(B) = \int_B p(x)\, \mathrm{d}x$ ，对所有可测的 $B \subseteq A$成立，其中 μ 是绝对连续于 λ。

进一步，设 $\mathbb{P}^n$ 表示标准的$(n-1)$维概率单纯形为

$$\mathbb{P}^n := \{p \in \mathbb{R}^n_{\geq 0} \mid \textstyle\sum_{i=1}^n p_i = 1\},$$

即 n 个元素上所有概率分布的集合。

For 为简单起见，我们采用如下约定：$0 \log 0 := \lim_{x \downarrow 0} x \log\ x = 0$。

2.1 马尔可夫链

设离散时间随机动力系统 $\mathcal{M}$ 在一个 n_x 维超矩形状态空间 $\mathcal{X} := [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_{n_x}, b_{n_x}] \subset \mathbb{R}^{n_x}$上演化，时间步为 $k \in \{0, 1, \ldots, K\}$，其中 $K \in \mathbb{N}$——这并不失一般性；

所得结果适用于任意紧致状态空间。设其在时间范围 K 内的马尔可夫动力学由状态转移概率密度（转移核）描述

$$\mathrm{Prob}(x_{k+1} \in A \mid x_k) = \int_A q(x_k, x_{k+1})\ \mathrm{d}x_{k+1},$$

其中， $A \subseteq \mathcal{X}$, $x_k \in \mathcal{X}$表示系统在时间步 $k \in \mathbb{N} \cup \{0\}$的状态，并结合初始状态分布密度 q_0，描述如下

$$\mathrm{Prob}(x_0 \in A) = \int_A q_0(x_0)\ \mathrm{d}x_0.$$

引入动作将推迟到第4节，在讨论控制器设计时进行。请注意，我们已隐含地假设， q_0 和 q 相对于勒贝格测度的概率测度具有绝对连续性。

将所有可能轨迹构成的空间记为 $\mathcal{S} := \mathcal{X}^{K+1}$，它是$n$维实空间的一个超矩形紧致子集，且 $n = (K+1)n_x$，则系统 $\mathcal{M}$ 可通过轨迹 $s \in \mathcal{S}$上的分布T 获得轨迹级描述，如下所示

$$T(s) = q_0(x_0) \prod_{k=1}^{K} q(x_{k-1}, x_k), \tag{1}$$

其中， $s = x_0$、.....、 x_K是 $\mathcal{S}$中的一条轨迹。

2.1.1 轨迹分布离散化

设$(\mathcal{X}_i)_{i \in X}$ 是 $\mathcal{X}$的超矩形网格划分。因此，对于 $i \in X := \{1, 2, \ldots, |X|\}$，状态空间划分 $\mathcal{X}_i$中的每个元素，以及对于$t \in S := \{1, 2, \ldots, |S|\}$，轨迹 $\mathcal{S}_t$的诱导划分中的每个元素都是超矩形，其中 $|S| = |X|^{K+1}$，见图1。令离散概率分布$p = \begin{bmatrix} p_1 & p_2 & \cdots & p_{|S|} \end{bmatrix}^\top$ 表示与划分 T相关联的$(\mathcal{S}_t)_{t \in S}$的离散化，如下所示

$$p_t := \int_{\mathcal{S}_t} T(s)\ \mathrm{d}s, \tag{2}$$

对所有 $t \in S$。

2.1.2 区间马尔可夫链抽象

令状态空间划分 $(\mathcal{X}_i)_{i \in X}$ 进一步将原连续空间马尔可夫过程 $\mathcal{M}$ 抽象为状态空间 X上的区间马尔可夫链 (IMC) [17] 。其初始有限状态分布为

$$\pi_i = \int_{\mathcal{X}_i} q_0(x_0)\ \mathrm{d}x_0, \tag{3a}$$

对于 $i \in X$，其（随时间变化的）马尔可夫核则受限于区间

$$P_k(i) \in \mathbb{P}_i := \{p' \in \mathbb{P}^{|X|} \mid \underline{P}_{ij} \le p'_j \le \overline{P}_{ij}\}, \tag{3b}$$

其中

$$\overline{P}_{ij} = \max_{x \in \mathcal{X}_i} \int_{\mathcal{X}_j} q(x, x')\ \mathrm{d}x', \quad \underline{P}_{ij} = \min_{x \in \mathcal{X}_i} \int_{\mathcal{X}_j} q(x, x')\ \mathrm{d}x', \tag{3c}$$

对所有 i, $j \in X$以及所有 $k \in \{0, 1, \ldots, K\}$。令 $\mathbb{T}$ 表示所有轨迹律的集合 $p' \in \mathbb{P}^{|S|}$，定义于 $S = X^{K+1}$之上且在区间马尔可夫链（3）中可容许，即满足初始分布 π，并允许存在随时间变化的马尔可夫核序列P_k（i）$\in \mathbb{P}_i$，对于 $i \in X$， $k \in \{0, 1, \ldots, K\}$。

总之，我们考虑三个不同的对象。主要对象是连续状态离散时间马尔可夫链 $\mathcal{M}$，它可以由状态空间层面的马尔可夫核或轨迹空间层面的概率分布表示来指定，如下所示

$$\mathcal{M} = (\mathcal{X}, q_0, q, K) \equiv (\mathcal{S}, T). \tag{4}$$

除了这个“底层”系统之外，我们还考虑它在离散轨迹空间上的轨迹层面的离散化 p (2)$S = X^{K+1}$，以及由 $\mathbb{T}$ (3)中的时变马尔可夫链集合定义的 IMC 抽象。

假设 1。 我们的工作依赖于以下假设成立：

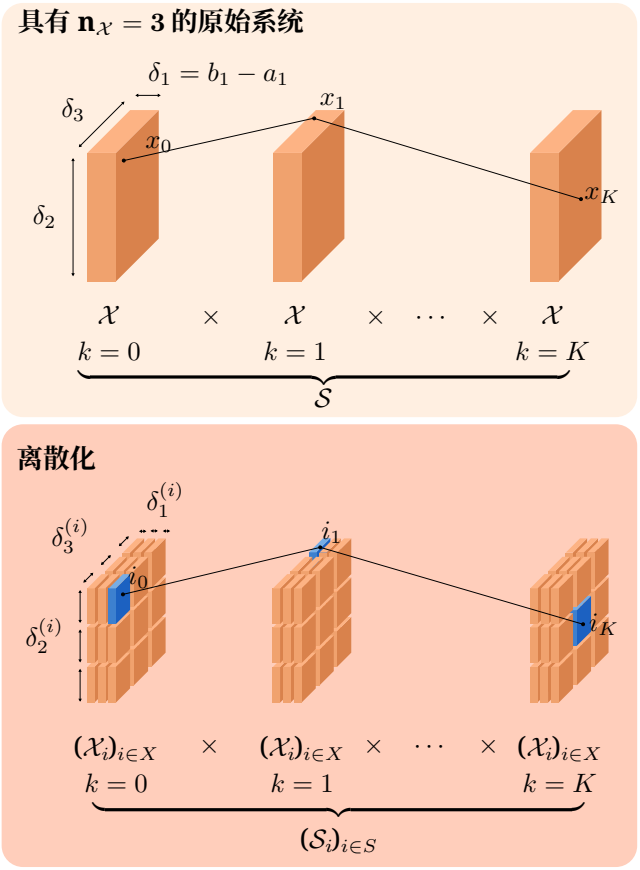


图 1：轨迹空间 $\mathcal{S}$的离散化（$\mathcal{S}_t$）$_{t\in S}$，由状态空间 $\mathcal{X}$的超矩形离散化（$\mathcal{X}_i$）$_{i\in X}$所诱导。

- $\mathcal{X}_i$ 对于每个 $i \in X$ 都是超矩形的，
- q 和 q_0 几乎处处连续且可微，
- ∇q 和 ∇q_0 几乎处处有界，即 $\|\nabla q\|_\infty \le L_{\nabla q}$、$\|\nabla q_0\|_\infty \le L_{\nabla q}$，其中 $L_{\nabla q} \in \mathbb{R}$。

根据这些假设，T 有界、连续且几乎处处可微。进一步地，∇T 几乎处处有界，即 $\|\nabla T\|_\infty \le L$，其中 $L \in \mathbb{R}$。所有连续分布（q、q_0、T）都可以表示为密度。

2.2 熵量化

系统熵通过系统轨迹分布与均匀分布之间的*KL* 散度来量化。对于分布 $T \in \mathcal{P}(\mathcal{S})$，相对于均匀分布的 KL 散度定义为

$$\mathrm{KL}(T\|U) = \int_{\mathcal{S}} T(s) \log \frac{T(s)}{1/\lambda(\mathcal{S})}\, \mathrm{d}s, \tag{5}$$

其中 $U(s) = 1/\lambda(\mathcal{S})$ 是在 $\mathcal{S}$ 上的均匀分布，且 $\mathrm{KL}(\cdot\|U)$ 是 [16] 相对于该均匀分布的分布的 KL散度。

备注1。以下关系有助于理解分布 $T \in \mathcal{P}(\mathcal{S})$ 相对于均匀分布的 *KL* 散度与其微分熵 $h(T)$ 之间的联系

$$\begin{aligned}\mathrm{KL}(T\|U) = \log \lambda(\mathcal{S}) + \int_{\mathcal{S}} T(s) \log T(s)\, \mathrm{d}s \\ = \max_{T' \in \mathcal{P}(\mathcal{S})} \{h(T')\} - h(T),\end{aligned}$$

由于微分熵 $h(T) = -\int_{\mathcal{S}} T(s)\log T(s)\,\mathrm{d}s$ [16]。

在离散化空间中，我们通过相对于均匀分布的离散KL散度来量化可预测性：

$$\mathrm{KL}_D(p\|p^{\mathrm{u}}) = \sum_{t\in S} p_t \log \frac{p_t}{p_t^{\mathrm{u}}} = \sum_{t\in S} p_t \log \frac{p_t\lambda(\mathcal{S})}{\lambda(\mathcal{S}_t)}, \tag{6}$$

其中， p和 p^{u}分别是根据（ T2），对轨迹分布 U和均匀分布 T进行离散化后得到的结果。

尽管相对于均匀分布的 KL 散度与微分熵密切相关，但二者在许多重要方面具有不同的性质。最重要的是，离散分布 [16] 的香农熵 p (2) 在离散化越来越精细时不会收敛到 T 的微分熵；而我们发现，当$\lambda(\mathcal{S}_t)$ 收缩时：$=$ $\max_{t\in S}\lambda(\mathcal{S}_t)$，相对于均匀分布的 KL 散度确实会收敛：

$$\begin{aligned}
\lim_{\overline{\lambda}(\mathcal{S}_t)\to 0} \mathrm{KL}_D(p\|p^{\mathrm{u}}) &= \lim_{\overline{\lambda}(\mathcal{S}_t)\to 0} \sum_{t\in S} p_t \log \frac{p_t\lambda(\mathcal{S})}{\lambda(\mathcal{S}_t)} \\
&= \lim_{\overline{\lambda}(\mathcal{S}_t)\to 0} \sum_{t\in S} \int_{\mathcal{S}_t} T(s)\ \mathrm{d}s \log T(s_t)\lambda(\mathcal{S}) \\
&= \lim_{\overline{\lambda}(\mathcal{S}_t)\to 0} \sum_{t\in S} \int_{\mathcal{S}_t} T(s) \log T(s_t)\lambda(\mathcal{S})\ \mathrm{d}s \\
&= \int_{\mathcal{S}} T(s)\log T(s)\lambda(\mathcal{S})\ \mathrm{d}s \\
&= \mathrm{KL}(T\|U),
\end{aligned}$$

这里我们利用了 $\lim_{\lambda(\mathcal{S}_t)\to 0}\int_{\mathcal{S}_t} T(s)\,\mathrm{d}s \to T(s_t)\lambda(\mathcal{S}_t)$ 对任意 $s_t\in\mathcal{S}_t$ 以及所有 $t\in S$ 均成立这一事实。

3 系统熵的形式化界定

本节将建立所需的理论，以便从原始系统 $\mathcal{M}$ 的 IMC 抽象 $\mathbb{T}$ 中，获得该系统轨迹相对于均匀分布的 KL 散度的形式化上界和下界。

在第 3.1 节中，我们首先提出一种方法，专门用于求得相对于均匀分布的 KL 散度下界，见图 2。要得到有效的上界，还需要额外的结果，因此我们提出了两种独特的上界方法。一种是对标准 IMC 方法进行“全局”修正，可以事后添加到经典 IMC 算法中，以修正其计算出的界；另一种是集成度更高的“局部”方法，它会扩展经典 IMC 算法本身，见图 3。第 3.2 节随后讨论如何应用这些结果，以便在实践中实际计算界。

3.1 界的构造

根据 [40]，的定理 3 和命题 3 中关于区间马尔可夫链抽象完备性的结果可知，对于状态空间 $\mathcal{X}$ 的任意超矩形划分 $(\mathcal{X}_i)_{i\in X}$，由（ p ）所诱导的离散化轨迹分布 2 是可容许的区间马尔可夫链轨迹集 $\mathbb{T}$ 中的一个元素，即

$$p\in\mathbb{T}. \tag{7a}$$

此外，该抽象近似完备，具体而言，随着划分不断细化，

$$\lim_{\overline{\lambda}(\mathcal{X}_i)\to 0} \mathbb{T} = \{p\}, \tag{7b}$$

在离散化分辨率极限 $\lambda(\mathcal{X}_i)\to 0$下，其中收敛是指弱意义上的收敛，即所有有限时刻的联合分布均收敛（等价地，所有有限维边缘分布均逐分量收敛）。这正是现有 IMC 抽象方法能够对离散轨迹行为p作出保证的根本原因 T，尽管它们只能获取区间马尔可夫链 $\mathbb{T}$。

令

$$\overline{\mathrm{KL}}_D(\mathbb{T}\|\rho) := \max_{p'\in\mathbb{T}} \mathrm{KL}_D(p'\|\rho),$$

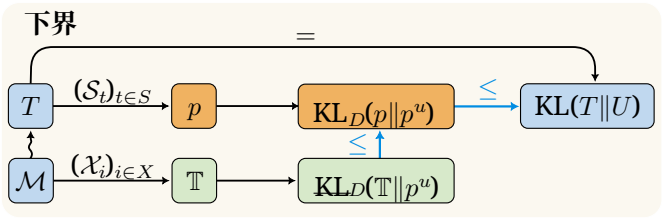


图 2：构成系统 $\mathcal{M}$ 相对于均匀分布的 KL 散度下界的对象，即 $\mathrm{KL}(T\|U)$，也就是我们的度量总系统熵。

$$\underline{\mathrm{KL}}_D(\mathbb{T}\|\rho) := \min_{p'\in\mathbb{T}} \mathrm{KL}_D(p'\|\rho),$$

描述满足区间马尔可夫链的分布相对于任意概率分布的极端 KL 散度值 $\mathbb{T}$，并将其与抽象完备性（7）相结合，我们可以得出结论：对于状态空间 $\mathcal{X}$的任意超矩形划分 $(\mathcal{X}_i)_{i\in X}$ 以及任意概率分布 $\rho \in \mathbb{P}^{|S|}$ ——其中包括$\rho = p^{\mathrm{u}}$这一我们所关注的分布——我们有

$$\overline{\mathrm{KL}}_D(\mathbb{T}\|\rho) \geq \mathrm{KL}_D(p\|\rho) \geq \underline{\mathrm{KL}}_D(\mathbb{T}\|\rho). \tag{8a}$$

此外，当 $\lambda(\mathcal{X}_i) \to 0$趋于极限时，（8a）成立且取等号，因为

$$\begin{aligned}\lim_{\overline{\lambda}(\mathcal{X}_i)\to 0} \overline{\mathrm{KL}}_D(\mathbb{T}\|\rho) &= \lim_{\overline{\lambda}(\mathcal{X}_i)\to 0} \mathrm{KL}_D(p\|\rho) \\ &= \lim_{\overline{\lambda}(\mathcal{X}_i)\to 0} \underline{\mathrm{KL}}_D(\mathbb{T}\|\rho).\end{aligned} \tag{8b}$$

注意，该结果（8）使我们能够利用 IMC 抽象 $\mathbb{T}$ 来获得轨迹分布离散化 p相对于均匀分布的 KL 散度的界。然而，仅凭这一点还不足以刻画其与原始轨迹分布 T相对于均匀分布的 KL 散度之间的关系。为了将这一结果与系统熵水平 $\mathrm{KL}(T\|U)$ 直接联系起来，我们需要以下引理。

引理 1 (连续-离散差异)。 对于定义在连通紧集$\mathcal{S} \subset \mathbb{R}^n$ 上的分布密度 T 及其离散化 $p_t := \int_{\mathcal{S}_t} T(s)\,\mathrm{d}s$ 在划分 $\bigcup_{t\in S} \mathcal{S}_t = \mathcal{S}$, $\lambda(\mathcal{S}_t \cup \mathcal{S}_{t'}) = 0$ 上对所有 $t \neq t'$成立，我们可以将 T相对于均匀分布的 *KL* 散度与其离散化p相对于均匀分布的 *KL* 散度之差写为

$$\mathrm{KL}(T\|U) - \mathrm{KL}_D(p\|p^u) = \mathrm{KL}(T\|T_D), \tag{9}$$

其中 T_D 是 p 的分段连续表示，定义为

$$T_D(s) = \sum_{t\in S} \mathbb{1}_{\mathcal{S}_t}(s)\frac{p_t}{\lambda(\mathcal{S}_t)}, \qquad \mathbb{1}_{\mathcal{S}_t}(s) = \begin{cases}1, & \textit{if } s \in \mathcal{S}_t,\\ 0, & \textit{otherwise},\end{cases}$$

其中 $\mathbb{1}_{\mathcal{S}_t}(s)$ 因此是一个指示函数。

证明：见附录A。 □

虽然引理 1 与（8）相结合后，仍不足以上界约束 $\mathrm{KL}(T\|U)$ 通过抽象 $\mathbb{T}$ 的取值，因为我们既无法获得 T，也无法获得 p，但它足以保证一个下界。

定理 1 （熵抽象下界）。

$$\mathrm{KL}(T\|U) \geq \mathrm{KL}_D(p\|p^{\mathrm{u}}) \geq \underline{\mathrm{KL}}_D(\mathbb{T}\|p^{\mathrm{u}}),$$

and

$$\mathrm{KL}(T\|U) = \lim_{\overline{\lambda}(\mathcal{X}_i)\to 0} \mathrm{KL}_D(p\|p^{\mathrm{u}}) = \lim_{\overline{\lambda}(\mathcal{X}_i)\to 0} \underline{\mathrm{KL}}_D(\mathbb{T}\|p^{\mathrm{u}}).$$

证明： 第一个（不）等式是将引理 1 中的结果与其右侧 KL 散度的非负性结合得出的，而第二个则由（8）得出，见图 2。 □

我们注意到，$\mathrm{KL}_D(\mathbb{T}\|p^{\mathrm{u}})$ 和 $\mathrm{KL}_D(\mathbb{T}\|p^{\mathrm{u}})$ 的值可以通过标准的 IMDP 方法 [36],获得，并结合相对于均匀分布的 KL 散度的递归性质；这一性质可通过扩展 [5], 中的方法直接得到，见附录中的引理 7。

要获得基于抽象的上界，以约束相对系统熵，还需要进一步的洞见。我们首先给出一个一般性引理，用于从上界约束离散化 p相对于其原始连续分布 T的相对于均匀分布的 KL 散度之间的差异。由于该引理适用于一般的连续空间分布及其离散化（因此不依赖马尔可夫性质或抽象），我们认为它在更广泛的信息论背景下也会有用。除了要求假设1 成立外，我们还推导出这一差异的解析上界，无需了解实际分布密度 T。

由于每个轨迹空间离散化区间 $\mathcal{S}_t$ 都是超矩形，我们可以用其分量区间来描述，如下所示

$$\mathcal{S}_t = [\alpha_1^{(t)}, \beta_1^{(t)}] \times [\alpha_2^{(t)}, \beta_2^{(t)}] \times \cdots \times [\alpha_n^{(t)}, \beta_n^{(t)}], \tag{10}$$

对于所有 $t \in S$。令长度 维度 $j \in \{1, 2, \ldots, n\}$ 中元素 $\mathcal{S}_t$ 在 $t \in S$ 下的长度记为 $\delta_j^{(t)} := \beta_j^{(t)} - \alpha_j^{(t)}$.

引理 2（离散化差异上界）。对于定义在连通紧集 T 上的概率分布 $\mathcal{S} \subset \mathbb{R}^n$，在满足假设 1的条件下，及其在超矩形划分 p (2) $(\mathcal{S}_t)_{t\in S}$上得到的离散化分布， T 的相对于均匀分布的 *KL* 散度与其离散化分布 p 的相对于均匀分布的 *KL* 散度之差，上界为

$$\mathrm{KL}(T\|U) - \mathrm{KL}_D(p\|p^{\mathrm{u}}) \le \varepsilon(p, n, L, \overline{\delta}),$$

其中

$$\varepsilon(p, n, L, \overline{\delta}) := \sum_{t\in\{1,2,\ldots,|p|\}} p_t \log\left(1 + \frac{n}{2p_t} L\overline{\delta}^{n+1}\right), \tag{11}$$

与 $\overline{\delta} := \max_{t\in S} \max_{j\in\{1,2,\ldots,n\}} \delta_j^{(t)}$ 。

证明： 见附录B。 □

接下来，我们将引理 2 的结果具体化，用于构造两种不同的上界方法。首先是“全局”方法，它对 $\mathcal{M}$所需的信息极少，并且可以事后应用，以修正更传统的抽象方法 [5,36], 并恢复形式化上界保证。其次是更集成的方法，以“局部”方式利用引理 2 的结果，在每个时间步修正经典的抽象方法熵递归 [5] ，从而更充分地利用来自 $\mathcal{M}$的信息，并通常得到更不保守的上界。我们将这些结果作为两个完整的、基于抽象的 IMC 熵上界定理提出，即定理 2 和3 （见图 3）。

定理 2（全局离散化差异界）. 在假设 1下，我们有 e

$$\begin{aligned}\mathrm{KL}(T\|U) &\le \mathrm{KL}_D(p\|p^{\mathrm{u}}) + \varepsilon(\mathbb{1}|S|^{-1}, n, L, \overline{\delta}) \\ &\le \overline{\mathrm{KL}}_D(\mathbb{T}\|p^{\mathrm{u}}) + \varepsilon(\mathbb{1}|S|^{-1}, n, L, \overline{\delta}),\end{aligned}$$

其中 $\mathbb{1}|S|^{-1} := \begin{bmatrix}|S|^{-1} & \cdots & |S|^{-1}\end{bmatrix}^\top$，因此

$$\varepsilon(\mathbb{1}|S|^{-1}, n, L, \overline{\delta}) = \log\left(1 + \frac{n}{2}|S|L\overline{\delta}^{n+1}\right).$$

discr 上的这一上界 *etization* 差异对于任意满足条件的离散化方案进一步收敛到零

$$|S|\overline{\delta}^{n+1} \to 0, \tag{12}$$

从而得到

$$\lim_{\overline{\lambda}(\mathcal{X}_i)\to 0} \overline{\mathrm{KL}}_D(\mathbb{T}\|p^{\mathrm{u}}) + \varepsilon(\mathbb{1}|S|^{-1}, n, L, \overline{\delta}) = \mathrm{KL}(T\|U).$$

证明：见附录 C，我们在其中同时推导了界 $\varepsilon(\mathbb{1}|S|^{-1}$， n， L， $\delta)$，以及其收敛条件。 □

注 2。收敛要求(12) 并不十分严格，因为可以证明，例如，当 $N \to \infty$ 时，将 $\mathcal{X}$ 的每个维度均匀分布地划分为 N 个区域的简单基于网格的离散化方案能够直接满足该要求。让我们注意到，在上述方案下， $|S| = |X|^{K+1} = N^{n_x(K+1)}$、$-\delta = -\delta_x/N$，其中 $-\delta_x = \max_{i\in\{1,2,\ldots,n_x\}} b_i - a_i$，最后， $n = (K+1)n_x$，由此即可确认这一点。这些量可以代入要求(12) ，得到 $|S|\delta^{n+1} = \delta^{n_x(K+1)+1}/N$)；对于 $N \to \infty$，该式总会趋于零，因为 $\delta^{n_x(K+1)+1}$ 是常数。

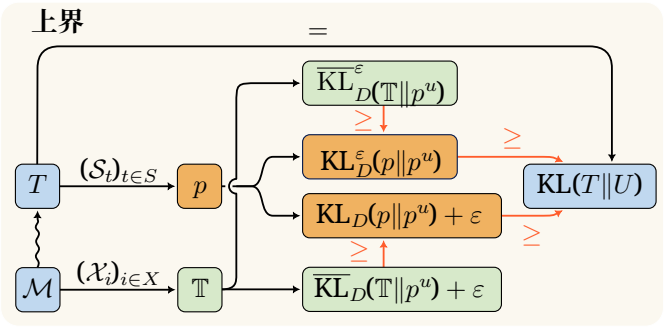


图 3：共同构成系统 $\mathcal{M}$ 相对于均匀分布的 KL 散度（即 $\mathrm{KL}(T\|U)$）的形式化上界的对象，即我们的度量总系统熵。

令 $\Phi:\mathbb{P}^{|X|}\times\mathbb{R}^{|X|}\to\mathbb{R}$，定义为

$$\Phi(p',V):=\sum_{j\in X}p'_j\log\frac{p'_j\lambda(\mathcal{X})}{\lambda(\mathcal{X}_j)}+\sum_{j\in X}p'_jV_j, \tag{13}$$

并令 $\Phi^\varepsilon:\mathbb{P}^{|X|}\times\mathbb{R}^{|X|}\to\mathbb{R}$，进一步定义为

$$\Phi^\varepsilon(p',V):=\Phi(p',V)+\varepsilon(p',n_x,L_{\nabla q},\bar{\delta}).$$

定理 3 (局部离散化差异界)。 在假设 *1*下，连续轨迹 *KL* 散度相对于均匀分布的值由以下式子给出上界

$$\mathrm{KL}(T\|U)\le\mathrm{KL}^{\varepsilon}_{D}(\mathbb{T}\|p^{\mathrm{u}}),$$

其中

$$\mathrm{KL}^{\varepsilon}_{D}(\mathbb{T}\|p^{\mathrm{u}}):=\Phi^\varepsilon(\pi,V^\varepsilon_0), \tag{14a}$$

并且，对于 $k=K-1,K-2,\ldots,0$,

$$V^\varepsilon_k(i)=\max_{P_k(i)\in\mathbb{P}_i}\Phi^\varepsilon(P_k(i),V^\varepsilon_{k+1}), \tag{14b}$$

对所有 $i\in X$，均由 $V^\varepsilon_K(i)=0$ 初始化。此外，对于满足 $|S|\delta^{n+1}\to0$ 的基于网格的离散化，我们得到收敛，如

$$\lim_{\bar{\lambda}(\mathcal{X}_i)\to0}\mathrm{KL}^{\varepsilon}_{D}(\mathbb{T}\|p^{\mathrm{u}})=\mathrm{KL}(T\|U).$$

证明： 参见附录 D。 □

3.2 注意事项和算法

定理 2 和定理 3 的结果，结合用于计算相对于均匀分布的 KL 散度的递归算法（参见附录中的引理 7），提供了两种实用方法，可用于获得 T 相对于均匀分布的 KL 散度界。算法 1 同时计算这两种方法，得到以下界值：(i) $\mathrm{KL}_D(\mathbb{T}\|U)$ ($\le\mathrm{KL}(T\|U)$)，(ii) $\mathrm{KL}_D(\mathbb{T}\|U)+\varepsilon(\mathbb{1}|S|^{-1},n,L,\delta)$ ($\ge\mathrm{KL}(T\|U)$)，以及 (iii) $\mathrm{KL}^{\varepsilon}_{D}(\mathbb{T}\|U)$ ($\ge\mathrm{KL}(T\|U)$)。

最后，引理 3 有助于求得 L 转移分布密度的梯度上界 T，前提是能够获得初始状态分布 q_0 和马尔可夫核 q。

引理 3 （梯度幅度的界 T）。 给定假设 *1* 和常数 $L_q\in\mathbb{R}_{\ge0}$，使得

$$\|q\|_\infty\le L_q,\quad \|q_0\|_\infty\le L_q, \tag{15}$$

then

$$\|\nabla T\|_\infty\le2(L_q)^KL_{\nabla q}=:L.$$

证明： 参见附录G。 □

算法 1 求取上界和下界

输入： $L, L_{\nabla q}, X, (\mathcal{X}_i)_{i\in X}, K, \pi, \underline{P}, \overline{P}. n_x \Leftarrow$维度$(\mathcal{X}_0), \delta \Leftarrow$最大$_{i\in X}$最大$_{j\in\{1,2,\ldots,n_x\}} b_j^{(i)} - a_j^{(i)}$, $|S| \Leftarrow |X|^{K+1}, \varepsilon \Leftarrow$对数$(1 + \frac{1}{2}n|S|L\delta^{n+1})$, $\mathbb{P}_i \Leftarrow \{p' \in \mathbb{P}^{|X|} : P_{ij} \le p'_j \le P_{ij}\}, \forall i \in X.$ 迭代取值以求得上界和下界： $V_K(i) = V_K(i) = V_K^\varepsilon(i) \Leftarrow 0, \forall i \in X,$ 对于 $k \in \{K-1, K-2, \ldots, 0\}$ 执行对于 $i \in X$ 执行

$$\underline{V}_k(i) \Leftarrow \min_{p'\in\mathbb{P}_i} \Phi(p', \underline{V}_{k+1}),$$
$$\overline{V}_k(i) \Leftarrow \max_{p'\in\mathbb{P}_i} \Phi(p', \overline{V}_{k+1}),$$
$$V_k^\varepsilon(i) \Leftarrow \max_{p'\in\mathbb{P}_i} \Phi^\varepsilon(p', V_{k+1}^\varepsilon),$$

结束 结束合并并返回： $\mathrm{KL}_D(\mathbb{T}\|p^u) \Leftarrow \pi^\top V_0$、 $\mathrm{KL}_D(\mathbb{T}\|p^u) + \varepsilon(\mathbb{1}|S|^{-1}, n, L, \delta) \Leftarrow \pi^\top V_0 + \varepsilon$、 $\mathrm{KL}_D^\varepsilon(\mathbb{T}\|p^u) \Leftarrow \pi^\top V_0^\varepsilon$。

4 形式化熵正则化控制

利用上述针对形式化熵抽象所建立的理论，我们现在能够构建一种形式化抽象控制器综合 方法。这里，我们首先介绍马尔可夫决策过程（MDP）的符号表示，然后提出一种基于抽象的通用控制器综合方法，该方法能够在系统熵（以其轨迹分布相对于均匀分布的 KL 散度量化）与累积成本之间进行权衡，并能将形式保证延续至原始连续状态系统。

4.1 基本预备知识

设一个连续状态马尔可夫决策过程的参数化形式为

$$\mathcal{D} = (\mathcal{X}, \tilde{U}, q_0, [q^u]_{u\in\tilde{U}}, g, K),$$

其中 $\mathcal{X}$、 q_0 和 K 如上所述， U 为有限动作集， $[q^u]_{u\in\tilde{U}}$ 为由 $u \in U$ 索引的有限马尔可夫核集合，且为阶段成本函数 $g : \mathcal{X} \times U \to \mathbb{R}$ （不失一般性，省略 g 的时间依赖性）。定义离散化状态空间 X 上的策略为函数 $\mu : \{0, 1, \ldots, K-1\} \times X \to U$。由 $\mathcal{D}$ 在策略 μ 下诱导的轨迹分布密度 T^μ 定义如下：

$$T^\mu(s) = q_0(x_0) \prod_{k=0}^{K-1} q^{\mu_k(\mathrm{i}(x_k))}(x_k, x_{k+1}),$$

其中 $s = x_0$、.....、 x_K 是 $\mathcal{S} := \mathcal{X}^{K+1}$ 中的一条轨迹， $\mathrm{i}(x_k)$ 返回满足 $x_k \in \mathcal{X}_i$ 的离散状态 $i \in X$。

我们的目标是最小化标准的期望累积阶段成本，并以 T^μ 相对于均匀分布的 KL 散度所衡量的轨迹熵进行正则化。此处我们考虑一个正的正则化项，因此在最小化代价函数时会惩罚可预测的行为，但该方法也可直接推广到鼓励可预测性。具体而言，我们考虑有限时域熵正则化目标

$$\underbrace{\mathbb{E}\left[\sum_{k=0}^{K-1} g(x_k, \mu_k(\mathrm{i}(x_k)))\right]}_{J^\mu} + \mathrm{KL}(T^\mu\|U). \tag{16}$$

为简化符号表示，省略了显式的正则化权重因子。

˜我们将该抽象建模为一个有限状态马尔可夫决策过程，其状态空间为 X，动作空间为 U，转移˜概率具有不确定性；这通常称为区间马尔可夫决策过程（IMDP）。对于每个状态-动作对（i， u） $\in X \times U$，转移核被约束在

$$P(i,u) \in \mathbb{P}_i(u) := \{p' \in \mathbb{P}^{|X|} \mid \underline{P}_{ij}(u) \le p'_j \le \overline{P}_{ij}(u)\}, \tag{17}$$

其中

$$\begin{aligned} \underline{P}_{ij}(u) &= \min_{x \in \mathcal{X}_i} \int_{\mathcal{X}_j} q^u(x,x')\ \mathrm{d}x', \\ \overline{P}_{ij}(u) &= \max_{x \in \mathcal{X}_i} \int_{\mathcal{X}_j} q^u(x,x')\ \mathrm{d}x', \end{aligned} \tag{18}$$

˜对所有 i， $j \in X$， $u \in U$。我们进一步将保守的抽象阶段成本 [36], 定义为

$$\overline{g}(i,u) = \max_{x \in \mathcal{X}_i} g(x,u), \quad \underline{g}(i,u) = \min_{x \in \mathcal{X}_i} g(x,u), \tag{19}$$

˜对所有 $i \in X$，以及所有 $u \in U$。

我们通过优化（最小化）与定理 2 和 3中的技术分别相关的两个上界，来综合策略；也就是说，我们的方法得到两个优化上界的策略及其对应的上界值。此外，根据定理 1的结果，我们还计算每个策略对应的下界；该下界刻画了在采用相同策略时，实际连续系统的性能。

4.2 形式熵正则化的上界优化

我们将上界熵递推框架（算法 1）与用于阶段成本的标准 IMDP 鲁棒动态规划相结合。我们计算：(i) 使经过 μ 全局修正的上界最小化的策略（参见定理 2）；以及 (ii) 使经过 μ^ε 局部修正的上界最小化的策略（参见定理 3）。如算法 2所述，两者均通过在抽象模型上进行反向动态规划计算。.

修改控制器综合算法，使其最小化上界（而非像当前那样最大化下界），只需将相应的最大化与最小化互换即可轻松实现。

注3. 请注意，算法 $p' \in \mathbb{P}_i(u)$ 中的最大化问题 *2* 是在多面体上对凸目标函数求最大值，即一个凹问题。因此，这也可以等价地表述为在极点上进行有限次优化，与仅考虑成本的 *IMDP* 设置类似 *[5, 36]*, ，从而保证得到全局最优解。相反，在我们算法中鼓励可预测性的变体里，相对于均匀分布的 *KL* 散度带有负号，因此形成一个凸问题，可以高效地求得数值解。

因此，将这些策略应用于原始的连续状态系统时，它们会使相对于均匀分布的 KL 散度保证值与阶段成本上界之间的差异最小化，其中

$$\begin{aligned} \mu &= \arg\min_{\mu'} \left\{ \overline{J}^{\mu'} + \overline{\mathrm{KL}}_D(\mathbb{T}^{\mu'} \| p^{\mathrm{u}}) + \varepsilon(\mathbb{1}|S|^{-1}, n, L, \overline{\delta}) \right\}, \\ \mu^\varepsilon &= \arg\min_{\mu'} \left\{ \overline{J}^{\mu'} + \mathrm{KL}^\varepsilon_D(\mathbb{T}^{\mu'} \| p^{\mathrm{u}}) \right\}, \end{aligned}$$

其中

$$\overline{J}^{\mu} := \mathbb{E}\left[\sum_{k=0}^{K-1} \overline{g}(\mathrm{i}(x_k), \mu_k(\mathrm{i}(x_k))) \right]. \tag{20}$$

此外，所得界限适用于原始的连续状态系统，因为

$$\begin{aligned} \underline{J}^{\mu} + \underline{\mathrm{KL}}_D(\mathbb{T}^{\mu} \| U) &\le J^{\mu} + \mathrm{KL}(T^{\mu} \| U) \\ &\le \overline{J}^{\mu} + \overline{\mathrm{KL}}_D(\mathbb{T}^{\mu} \| U) + \varepsilon(\mathbb{1}|S|^{-1}, n, L, \overline{\delta}), \end{aligned}$$

算法 2 计算形式上界最小化控制策略

输入： $L, L_{\nabla q}, X, (\mathcal{X}_i)_{i\in X}, K, \pi, \underline{P}, \overline{P}.n_x \Leftarrow$维数$(\mathcal{X}_0), \delta \Leftarrow$ 最大$_{i\in X}$ 最大$_{j\in\{1,2,\ldots,n_x\}} b_j^{(i)} - a_j^{(i)}$,$|S| \Leftarrow |X|^{K+1}, \varepsilon \Leftarrow$ 对数$(1 + \frac{1}{2}n|S|L\delta^{n+1}), \mathbb{P}_i(u) \Leftarrow \{p' \in \mathbb{P}^{|X|} :\ P_{ij}(u) \le p'_j \le P_{ij}(u)\}, \forall i \in X, u \in \tilde{U}$. $\underline{g}(i,u) \Leftarrow$最大$_{x\in\mathcal{X}_i}\, g(x,u)$, 对所有 $u \in U$, 所有 $i \in X$, $\overline{g}(i,u) \Leftarrow$最小$_{x\in\mathcal{X}_i}\, g(x,u)$, 对所有 $u \in U$, 所有 $i \in X$, 通过值迭代求出上界和下界：$V_K(i) = V_K(i) = V_K^{\varepsilon}(i) = V_K^{\varepsilon}(i) \Leftarrow 0,\ \forall i \in X$, 对于 $k \in \{K-1, K-2, \ldots, 0\}$ 执行

对于 $i \in X$，**执行**

$$
\begin{aligned}
\mu_k(i) &\Leftarrow \arg\min_{u\in\tilde{U}} \max_{p'\in\mathbb{P}_i(u)} \overline{g}(i,u) + \Phi(p', \overline{V}_{k+1}),\\
\mu_k^{\varepsilon}(i) &\Leftarrow \arg\min_{u\in\tilde{U}} \max_{p'\in\mathbb{P}_i(u)} \overline{g}(i,u) + \Phi^{\varepsilon}(p', V_{k+1}^{\varepsilon}),\\
\underline{V}_k(i) &\Leftarrow \min_{p'\in\mathbb{P}_i(\mu_k(i))} \underline{g}(i,\mu_k(i)) + \Phi(p', \underline{V}_{k+1}),\\
\underline{V}_k^{\varepsilon}(i) &\Leftarrow \min_{p'\in\mathbb{P}_i(\mu_k^{\varepsilon}(i))} \underline{g}(i,\mu_k^{\varepsilon}(i)) + \Phi(p', \underline{V}_{k+1}^{\varepsilon}),\\
\overline{V}_k(i) &\Leftarrow \max_{p'\in\mathbb{P}_i(\mu_k(i))} \overline{g}(i,\mu_k(i)) + \Phi(p', \overline{V}_{k+1}),\\
\overline{V}_k^{\varepsilon}(i) &\Leftarrow \max_{p'\in\mathbb{P}_i(\mu_k^{\varepsilon}(i))} \overline{g}(i,\mu_k^{\varepsilon}(i)) + \Phi^{\varepsilon}(p', \overline{V}_{k+1}^{\varepsilon}),
\end{aligned}
$$

end

end

合并并返回：$J^{\mu} + \mathrm{KL}_D(\mathbb{T}^{\mu}\|U) \Leftarrow \pi^{\top}V_0, J^{\mu^{\varepsilon}} + \mathrm{KL}_D(\mathbb{T}^{\mu^{\varepsilon}}\|U) \Leftarrow \pi^{\top}V_0^{\varepsilon}$, $J^{\mu} + \mathrm{KL}_D(\mathbb{T}^{\mu}\|U) + \varepsilon\, \mathbb{1}|S|^{-1}\, \mathrm{KL}(\,)(,\ n,\ L,\ \delta) \Leftarrow \pi^{\top}V_0 + \varepsilon, J^{\mu^{\varepsilon}} + \mathrm{KL}_D^{\varepsilon}(\mathbb{T}^{\mu^{\varepsilon}}\|U) \Leftarrow \pi^{\top}V_0^{\varepsilon}$。

对于最优策略 μ，以及

$$
\begin{aligned}
\underline{J}^{\mu^{\varepsilon}} + \underline{\mathrm{KL}}_D(\mathbb{T}^{\mu^{\varepsilon}}\|U) &\le J^{\mu^{\varepsilon}} + \mathrm{KL}(T^{\mu^{\varepsilon}}\|U)\\
&\le \overline{J}^{\mu^{\varepsilon}} + \mathrm{KL}_D^{\varepsilon}(\mathbb{T}^{\mu^{\varepsilon}}\|U),
\end{aligned}
$$

对于最优策略 μ^{ε}，其中 $\underline{J}^{\mu}$ 表示将阶段成本上界 20替换为下界 g 后的 g。

5 示例

下文通过抽象后的现实世界系统简化版本，展示熵量化和熵调节控制器综合的行为。

5.1 马尔可夫链熵界

本节通过数值研究，考察我们基于抽象的熵界方法的收敛行为。算法 1 应用于一个高度简化但极具代表性的多维（截断）高斯转移模型示例。

我们旨在研究的精确系统由下式描述：

$$
x_{k+1} = x_k + w_k, \qquad w_k \sim \mathcal{N}(0, \Sigma), \tag{21}
$$

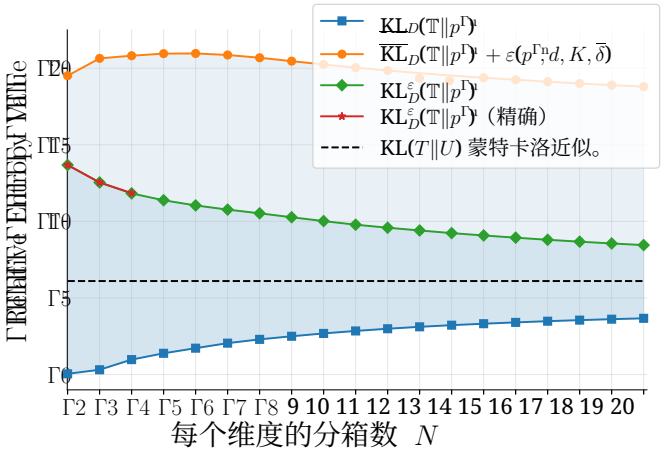


图 4：相对于均匀分布的 KL 散度的均匀抽象界，随离散化分辨率变化；离散化分辨率表示为每个维度上等分的数量 N。

其中 $x_k \in \mathcal{X} \subset \mathbb{R}^{n_x}$、$\Sigma \in \mathbb{R}^{n_x \times n_x}$ 位于紧致超矩形状态空间上

$$\mathcal{X} = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_{n_x}, b_{n_x}],$$

并结合均匀分布的放置，在结果为 $x_k + w_k \notin \mathcal{X}$ 时将 x_{k+1} 放回 $\mathcal{X}$。

记随机系统（21）处于状态$x \in \mathcal{X}$时，在状态空间 $\mathcal{X}$ 之外放置的概率质量为 $o(x, \Sigma) = 1 - \int_{\mathcal{X}} \mathcal{N}(x'; x, \Sigma)\, \mathrm{d}x'$，我们可以解析地描述其马尔可夫核 $q(x, x')$ 如下：

$$q(x, x') = \begin{cases} \mathcal{N}(x'; x, \Sigma) + o(x, \Sigma)/\lambda(\mathcal{X}), & \text{for } x' \in \mathcal{X}, \\ 0 & \text{otherwise}, \end{cases}$$

其中，$\lambda(\mathcal{X}) = \prod_{j=1}^{n_x}(b_j - a_j)$ 表示 $\mathcal{X}$ 的勒贝格测度。进一步假设由 $\hat{x}_0, \Sigma_0$ 参数化的高斯初始状态分布，我们有

$$q_0(x_0) = \mathcal{N}(x_0; \hat{x}_0, \Sigma_0) + o(\hat{x}_0, \Sigma_0)/\lambda(\mathcal{X}), \quad \text{for } x_0 \in \mathcal{X}.$$

尽管对于这一特定场景，可以解析地得到 L_q 和 $L_{\nabla q}$ 的表达式，但我们转而在实现中以数值方式获取这些值，这凸显了即使对于更复杂的分布，这一步骤仍然是可处理的。

将熵抽象方法应用于该连续状态马尔可夫链，以针对 $K = 4$、$n_x = 2$、$a_j = 0$、$b_j = 1$，求解 $j \in \{1, 2\}$，并按照前述方法将 $\mathcal{X}$ 的每个维度均匀离散化为 N 个区域，可得到收敛的界，如图4所示。请注意，对于 $N \in \{2, 3, 4\}$，我们运行了算法1，分别采用精确最大化（在 $\mathbb{P}_i(u)$ 的所有极点上进行，见注3）和一种效率更高的数值替代方案。由于多面体 $\mathbb{P}_i(u)$ 中极点的数量随 N 呈指数级增长，因此对于更大的 N，这种方法会变得难以处理。鉴于数值解在 $N \in \{2, 3, 4\}$ 处与精确解一致，我们相信其在 $N \geq 5$ 时的结果是准确的。

5.2 马尔可夫决策过程控制器综合

本节通过一个自动驾驶示例，在阶段成本最小化与系统熵最小化之间进行权衡；在该示例中，我们既希望车速较快，也希望行为具有可预测性。通过翻转正则化项的符号，可以利用算法2来鼓励可预测性。由于最小化相对于均匀分布的KL正则化项负值的上界（即最大化可预测性），等价于最大化原始项的下界，因此定理2 和3 中引入的额外机制并不涉及策略优化。只有在我们希望获得所得策略的性能下界时，才需要应用这些机制。

我们考虑一个简化的最短时间与最小熵粗糙地形下降问题，其中较高的速度与较不可预测的行为相关。令状态 $v_k \in \mathcal{X} := [0, 1]$ 描述车辆的速度，输入$u_k \in U := \{0, 5, 10, 15, 20\}$ 描述加性加速度，并将速度相关随机性建模为另一项加性（随机）加速度 $w_k \in [-1, 1]$，时间为 $k \in \{0, 1, \ldots, K\}$。阶段成本 $g(v_k, u_k) = -v_k$

会随着速度增加而降低，而不可预测性则会随着速度增加而增强，因为系统动力学遵循

$$v_{k+1} = 0.8v_k + 0.01u_k + w_k, \quad w_k \sim d(\,\cdot\,|v_k),$$

其中我们使用左、中、右三个点

$$\begin{bmatrix} l(v) \\ m(v) \\ r(v) \end{bmatrix} = (1-v)\begin{bmatrix} 0 \\ 0.05 \\ 0.1 \end{bmatrix} + v\begin{bmatrix} -0.8 \\ -0.2 \\ 0 \end{bmatrix},$$

来定义 w_k 关于 v_k 的三角概率密度，如下

$$d(\omega|\nu) = \frac{2}{r(\nu)-l(\nu)} \begin{cases} \frac{\omega - l(\nu)}{m(\nu)-l(\nu)}, & \text{if } \omega \in [l(\nu), m(\nu)], \\ \frac{r(\nu)-\omega}{r(\nu)-m(\nu)}, & \text{if } \omega \in [m(\nu), r(\nu)], \\ 0, & \text{otherwise}, \end{cases}$$

使得

$$\mathrm{Prob}(w_k \in A | v_k = \nu) = \int_A d(\omega|\nu)\ \mathrm{d}\omega.$$

函数（l， m， r）会扩大 w_k 的支持集，并使其随着速度增加而向下移动，从而模拟高速下更难预测的扰动，同时确保对于所有 $v_k \in [0,1]$ 和所有 $u_k \in U$，闭环动力学几乎必然满足 $v_{k+1} \in [0,1]$。通过最小化熵正则化控制性能的上界，我们得到能够保证达到或超过相应性能水平的策略，例如，从而确保受控系统满足某些安全规范。

我们将状态空间 $\mathcal{X}$ 均匀划分为 $N = 80$ 个区域，并根据（18）得到一个区间马尔可夫决策过程抽象。求解由此产生的鲁棒动态规划（算法 2）会生成一种在速度与不可预测性之间进行权衡的策略。为了说明正则化对系统轨迹的影响，我们将代价函数 g 缩放 $\varphi \in \{2.3, 2.56\}$，并计算一个纯粹的最短时间策略 μ_{DP} （使用标准的区间马尔可夫决策过程阶段代价上界最小化 [36]，不进行熵正则化）。

仿真结果如图 5所示，其中在 $\mu(\varphi = 2.56)$ 下的轨迹方差看似大于其他情况，但这是由于轨迹出现了粗略的分岔，要么一路向上，要么一路向下。三种策略相对于均匀分布的 KL 散度通过蒙特卡洛仿真得到，为

$$\begin{aligned} &\mathrm{KL}(T^{\mu_{\mathrm{DP}}} \| U) = 19.2, \\ &\mathrm{KL}(T^{\mu(\varphi=2.56)} \| U) = 25.3, \\ &\mathrm{KL}(T^{\mu(\varphi=2.3)} \| U) = 32.1, \end{aligned}$$

这表明提高熵正则化的相对水平具有明显效果。更加强调熵最小化会使系统避开高速区域，以防止不可预测性增加，转而倾向于扰动分布更窄的中等速度。相比之下，最短时间控制器纯粹将系统驱动至高速度，从而导致显著更高的轨迹熵。

我们所得到的界还会进一步

$$-36.3 \le J^{\mu(\varphi=2.56)} - \mathrm{KL}(T^{\mu(\varphi=2.56)} \| U) \le -34.8,$$

当我们以 $\varphi = 2.56$缩放 g 时，以及

$$-36.2 \le J^{\mu(\varphi=2.3)} - \mathrm{KL}(T^{\mu(\varphi=2.3)} \| U) \le -34.3,$$

当 g 被 $\varphi = 2.3$缩放时，其中 J^{μ}对任意策略 μ定义为（16）。因此，两种方法在本例中所达到的界间隙均仅为总目标的5%。

6 结论与讨论

本文提出了一种形式化抽象框架，用于分析和控制连续状态随机系统中的轨迹熵。我们利用区间马尔可夫决策过程抽象，推导出轨迹分布相对于均匀分布的 KL 散度的可计算上界和下界；随着离散化分辨率提高，这些界限均收敛于真实的相对于均匀分布的 KL 散度

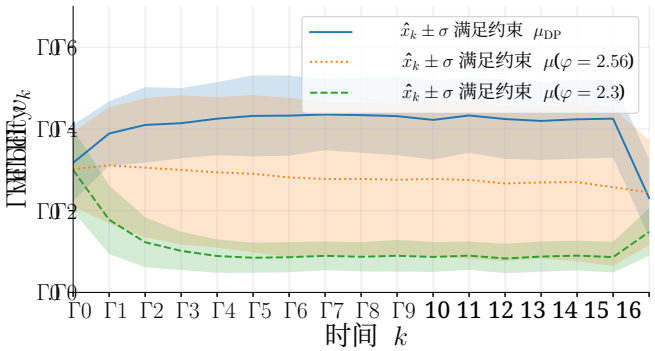


图 5：颠簸山丘自动驾驶车辆示例系统在未正则化策略 μ_{DP}、采用全局熵校正 ε μ 的熵正则化策略，以及局部校正熵正则化策略 μ^{ε} 下的行为。

。由此得到的理论支持形式化的熵感知控制器综合，可在可预测性与控制性能之间进行权衡，同时为原始连续系统保留形式保证。数值研究证实，所提出的界限会随着离散化细化而收敛，且综合得到的策略能够在实践中有效调节熵。未来工作将着重于进一步降低基于抽象的界限的保守性，并将该方法扩展至无限时域、更丰富的规约以及基于学习的系统模型。

参考文献

[1] O. X. Kuiper, J. E. Bos, E. A. Schmidt, C. Diels, and S. Wolter，“了解即将发生的事情：不可预测的运动会加重晕动症”，人因，第 62 卷，第 8 期，页 1339–1348，2020年。[2] R. Wijlens, B. J. V. Englebert, A. Takamatsu, M. Makita, H. Sato, T. Wada, J. C. F. de Winter, M. M. van Paassen, and M. Mulder，“迈向舒适：评估运动可预测性对自动驾驶车辆晕动症的影响”，工效学，第 68 卷，页 697–715，2024年。[3] W. Schwarting, A. Pierson, J. Alonso-Mora, S. Karaman, and D. Rus，“自主驾驶车辆的社会行为”，美国国家科学院院刊，第 116 卷，第 50 期，页 24 972–24 978，2019年。[4] D. J. Ornia, G. Delimpaltadakis, J. Kober, and J. Alonso-Mora，“通过熵率最小化实现可预测的强化学习动态”，机器学习研究汇刊，2025年6月，已接收并发表于 OpenReview。[5] M. J. T. C. van Zutphen, G. Delimpaltadakis, M. Heemels, and D. J. Antunes，“通过熵正则化实现可预测的区间马尔可夫决策过程”，载于第 *63* 届 *IEEE* 决策与控制会议（*CDC*）论文集。意大利米兰：IEEE，2024年，页 8754–8759。[6] F. Biondi, A. Legay, P. Malacaria, and A. Wasowski，“量化随机化协议的信息泄露”，理论计算机科学，第 597 卷，页 62–87，2015年。[7] L. Guo, H. Pan, X. Duan, and J. He，“平衡多机器人巡逻中的效率与不可预测性：一种基于多智能体强化学习的方法”，载于*2023*年 *IEEE* 机器人与自动化国际会议（*ICRA*）论文集。IEEE，2023年，页 3504–3509。[8] H. Guo, Q. Kang, W.-Y. Yau, M. H. Ang, and D. Rus，“Em-patroller：通过稳态分布近似实现熵最大化的多机器人巡逻”，*IEEE*机器人与自动化快报，第 8 卷，第 9 期，页 5712–5719，2023年。[9] P. Liu 等，“准备好在公共道路上欺负自动驾驶车辆了吗？”事故分析与预防，第 145 卷，页 105714，2020年，研究了混合交通中人类驾驶员对自动驾驶车辆采取激进驾驶行为的意图。

[10] J. Liu, X. Qi, P. Hang, and J. Sun，“增强自主驾驶车辆的社会决策：一种结合交互取向识别的混合策略博弈方法”，*IEEE*车辆技术汇刊，第 73 卷，编号 12385–12398，2024年，预印本见 arXiv:2312.11843。[11] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine，“基于深度能量策略的强化学习”，载于第 *34* 届国际机器学习会议论文集，第 70 卷。PMLR，2017年，页 1352–1361。[12] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine，“Soft actor-critic：基于随机演员的离策略最大熵深度强化学习”，载于第 *35* 届国际机器学习会议论文集，机器学习研究论文集系列，第 80 卷。PMLR，2018年，页 1861–1870。[13] Z. Ahmed, N. Le Roux, M. Norouzi, and D. Schuurmans，“理解熵对策略优化的影响”，载于第 *36* 届国际机器学习会议论文集，机器学习研究论文集系列，第 97 卷。PMLR，2019 年，页 151–160。[14] Y. Ashlag, U. Koren, M. Mutti, E. Derman, P.-L. Bacon, and S. Mannor，“面向鲁棒强化学习的状态熵正则化”，载于神经信息处理系统进展（*NeurIPS 2025*）——口头报告环节，*F*、*G*、*H*展厅，2025年12月，口头报告，2025年12月4日上午 10:00。[15] C. E. Shannon，“通信的数学理论”，贝尔系统技术期刊，第 27 卷，第 3 期，页 379–423，1948年7月。[16] T. M. Cover and J. A. Thomas，信息论基础。美国纽约州纽约：John Wiley & Sons，1991年。[17] F. Biondi, A. Legay, B. F. Nielsen, and A. Wasowski，“马尔可夫过程上的熵最大化”，逻辑与代数编程方法期刊，第 83 卷，第 5–6 期，页 384–399，2014年。[18] T. Chen and T. Han，“计算马尔可夫模型最大熵的复杂度”，载于第 *34* 届软件技术与理论计算机科学基础会议（*FSTTCS 2014*），莱布尼茨国际信息学论文集系列，V. Raman 和 S. P. Suresh 编，第 29 卷。德国达格施图尔：达格施图尔城堡—莱布尼茨信息学中心，2014年，页 571–583。[19] Y. Savas, M. Ornik, M. Cubuktepe, M. O. Karabag, and U. Topcu，“时序逻辑约束下马尔可夫决策过程的熵最大化”，*IEEE*自动控制汇刊，第 65 卷，第 4 期，页 1552–1567，2020年4月，提前在线发表：2019年6月12日。[20] Y. Chen, S. Li, and X. Yin，“线性时序逻辑任务下马尔可夫决策过程的熵率最大化”，*arXiv* 预印本 *arXiv:2211.12805*，2022年。[21] ——，“面向监视任务的马尔可夫决策过程熵率最大化”，*IFAC-PapersOnLine*，第 56 卷，第 2 期，页 4601–4607，2023年。[22] A. Srivastava and S. M. Salapaka，“参数化马尔可夫决策过程与强化学习问题——基于最大熵原理的框架”，*IEEE*控制论汇刊，第 52 卷，第 9 期，页 9339–9351，2021年。[23] D. Tiapkin, D. Belomestny, D. Calandriello, E. Moulines, R. Munos, A. Naumov, P. Perrault, Y. Tang, M. Valko, and P. Menard，“最大熵探索的快速收敛率”，载于第 *40* 届国际机器学习会议论文集，机器学习研究论文集系列，A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett 编，第 202 卷。PMLR，2023年7月23–29日，页 34 161–34 221。[24] B. You, C. Wang, and H. Liu，“面向可预测鲁棒控制的轨迹熵强化学习”，*arXiv* 预印本 *arXiv:2505.04193*，2025年。[25] S. E. Z. Soudjani and A. Abate，“通过形式化抽象对恒温控制负载群体进行聚合与控制”，*IEEE*控制系统技术汇刊，第 23 卷，第 3 期，页 975–990，2015年。[26] S. Haesaert, S. Esmaeil Zadeh Soudjani, and A. Abate，“通过近似相似关系与策略细化验证一般马尔可夫决策过程”，*SIAM*控制与优化期刊，第 55 卷，第 4 期，页 2333–2367，2017年。[27] E. M. Hahn, V. Hashemi, H. Hermanns, M. Lahijanian, and A. Turrini，“多目标区间马尔可夫决策过程：从鲁棒策略到帕累托曲线”，*ACM*建模与计算机仿真汇刊（*TOMACS*），第 29 卷，第 4 期，页 27:1–27:31，2019年。

[28] A. Lavaei、S. Soudjani、A. Abate和M. Zamani，“随机混合系统的自动验证与综合：综述”，*Automatica*，第146卷，第110617页，2022年。[29] M. Dutreix、J. Huh和S. Coogan，“面向 ω-正则目标的随机系统基于抽象的综合”，非线性分析：混合系统，第45卷，第101204页，2022年。[30] T. Badings、L. Romão、A. Abate、D. Parker、H. A. Poonawala、M. Stoelinga和N. Jansen，“基于形式化抽象的非高斯噪声动力系统鲁棒控制”，人工智能研究期刊，第76卷，第341–391页，2023年。[31] G. Delimpaltadakis、M. Lahijanian、M. Mazo和L. Laurenti，“具有连续动作空间的区间马尔可夫决策过程”，载于第*26*届*ACM*混合系统：计算与控制国际会议（*HSCC*）论文集。美国纽约州纽约市：美国计算机协会，2023年，第1–10页。[32] F. B. Mathiesen、M. Lahijanian和L. Laurenti，“IntervalMDP.jl：区间马尔可夫决策过程的加速值迭代”，*IFAC-PapersOnLine*，第58卷，第11期，第1–6页，2024年，第8届国际自动控制联合会混合系统分析与设计会议（ADHS 2024）。[33] M. Suilen、T. Badings、E. M. Bovy、D. Parker和N. Jansen，“鲁棒马尔可夫决策过程：人工智能与形式化方法的交汇之处”，载于验证原理：漫游概率图景——献给*Joost-Pieter Katoen*六十华诞的论文集，第三部分，计算机科学讲义丛书，第15262卷。施普林格，2024年，第126–154页。[34] A. Banse、G. Delimpaltadakis、L. Laurenti、M. Mazo和R. M. Jungers，“从转移算子视角看随机系统的记忆依赖抽象”，载于第*28*届*ACM*混合系统：计算与控制国际会议（*HSCC '25*）论文集。美国纽约州纽约市：美国计算机协会，2025年5月，HSCC 2025，美国加利福尼亚州尔湾。[35] G. Delimpaltadakis、L. Laurenti和M. J. Mazo，“随机事件触发控制采样行为的形式化分析”，*IEEE*自动控制汇刊，第69卷，第7期，第4491–4505页，2024年。[36] R. Givan、S. Leach和T. Dean，“有界参数马尔可夫决策过程”，人工智能，第122卷，第1-2期，第71–109页，2000年。[37] M. Lahijanian、S. B. Andersson和C. Belta，“离散时间随机系统的形式验证与综合”，*IEEE*自动控制汇刊，第60卷，第8期，第2031–2045页，2015年。[38] L. Laurenti、M. Lahijanian、A. Abate、L. Cardelli和M. Kwiatkowska，“连续时间线性随机混合过程的形式化高效综合”，*IEEE*自动控制汇刊，第66卷，第1期，第17–32页，2021年。[39] T. S. Badings、N. Jansen、L. Romão和A. Abate，“基于形式化抽象的随机与不确定动力学模型构造即正确控制”，载于第五届自主系统形式化方法国际研讨会论文集（*FMAS@iFM 2023*），EPTCS丛书，第395卷。开放出版协会，2023年，第144–152页。[40] Y. Meng和J. Liu，“用于随机系统验证的鲁棒完备有限状态抽象”，载于时间系统的形式建模与分析：第*20*届国际会议，*FORMATS 2022*，波兰华沙，*2022*年*9*月*13–15*日，论文集，计算机科学讲义丛书，S. Bogomolov和D. Parker编，第13465卷。施普林格，2022年，第80–97页。

附录

熵抽象下界（引理1）

将 T (5) 的相对于均匀分布的 KL 散度写成关于区间 $\mathcal{S}_t$ 的求和形式，如下：

$$\mathrm{KL}(T\|U) = \sum_{t\in S} \int_{\mathcal{S}_t} T(s)\log(T(s)\lambda(\mathcal{S}))\,\mathrm{d}s,$$

结合 p (6) 的相对于均匀分布的 KL 散度的显式表达式以及恒等式（2），可将离散化差（9）表示为关于 $t \in S$ 的单一求和形式，如下：

$$\begin{aligned}\mathrm{KL}(T\|U) - \mathrm{KL}_D(p\|p^{\mathrm{u}}) = \\ \sum_{t\in S}\int_{\mathcal{S}_t} T(s)\left(\log(T(s)\lambda(\mathcal{S})) - \log\left(\frac{p_t\lambda(\mathcal{S})}{\lambda(\mathcal{S}_t)}\right)\right)\,\mathrm{d}s,\end{aligned} \tag{22}$$

其中我们利用了这一事实：对数 $\left(\frac{p_t\lambda}{\lambda(\mathcal{S})\,(\mathcal{S}_t)}\right)$ 是关于 $s \in \mathcal{S}_t$ 的常数。将（22）求和式中的各项合并为一个对数项，可得

$$\begin{aligned}\mathrm{KL}(T\|U) - \mathrm{KL}_D(p\|p^{\mathrm{u}}) =& \sum_{t\in S}\int_{\mathcal{S}_t} T(s)\log\left(\frac{T(s)}{p_t/\lambda(\mathcal{S}_t)}\right)\mathrm{d}s \\ =& \,\mathrm{KL}(T\|T_D),\end{aligned}$$

其中我们注意到

$$T_D(s) = \left\{\frac{p_t}{\lambda(\mathcal{S}_t)}, \quad \text{when} \quad s \in \mathcal{S}_t.\right.$$

□

B 离散化差异上界（引理 2）

记离散化差为 Δ，即利用引理 1，我们定义

$$\begin{aligned}\Delta :=& \,\mathrm{KL}(T\|U) - \mathrm{KL}_D(p\|p^{\mathrm{u}}) \\ =& \sum_{t\in S}\int_{\mathcal{S}_t} T(s)\log\left(\frac{T(s)}{p_t/\lambda(\mathcal{S}_t)}\right)\,\mathrm{d}s.\end{aligned} \tag{23}$$

我们的目标是在不知道底层分布密度 T 的具体信息的情况下，找到一个可计算的上界 ε，使得 $\varepsilon \geq \Delta$。我们仅假设 T是一个连续分布密度，且满足 $\|\nabla T\|_\infty \leq L$，并且根据（2）诱导出离散分布 p。换言之， ε 必须满足

$$\varepsilon \geq \max_{\tau}\sum_{t\in S}\int_{\mathcal{S}_t}\tau(s)\log\left(\frac{\tau(s)}{p_t/\lambda(\mathcal{S}_t)}\right)\,\mathrm{d}s, \tag{24a}$$

$$\text{s.t. } \tau : \mathcal{S} \to \mathbb{R}_{\geq 0}, \tag{24b}$$

$$\int_{\mathcal{S}_t}\tau(s)\,\mathrm{d}s = p_t, \forall t \in S, \tag{24c}$$

$$\|\nabla\tau(s)\|_\infty \leq L, \ \forall t \in S, \tag{24d}$$

$$\tau \text{ continuous on } \mathcal{S}. \tag{24e}$$

显然，由于对数和求和的单调性，表达式

$$\begin{aligned}\sum_{t\in S}\int_{\mathcal{S}_t}\tau(s)\log\left(\frac{\max_{c\in\mathcal{S}_t}\{\tau(c)\}}{p_t/\lambda(\mathcal{S}_t)}\right)\,\mathrm{d}s = \\ \sum_{t\in S} p_t\log\left(\frac{\max_{c\in\mathcal{S}_t}\{\tau(c)\}}{p_t/\lambda(\mathcal{S}_t)}\right),\end{aligned} \tag{25}$$

对于任意在 $\mathcal{S}_t$ 上关于 $t \in S$ 连续的 τ，目标函数 (24a) 的值都有一个上界，其中包括使 (24) 最大化的最优 τ^*。我们不再针对每个 $t \in S$ 求出 $\max_{c\in\mathcal{S}_t}\{\tau(c)\}$ 的精确值，并在 τ 上优化由此得到的目标函数 (25)，而是针对每个单独的 $t \in S$，在给定（放宽后的）约束 (24b-24e) 的条件下，为 $\max_{c\in\mathcal{S}_t}\{\tau(c)\}$ 构造解析上界。将这些上界代入 (25)，即可保证所得值是 (24) 的上界，因此可作为 ε 的表达式。

B.1 最大 tau 值的上界

关于 $\max_{c\in\mathcal{S}_t}\{\tau(c)\}$ 的上界，对于每个 $t\in S$ 以及每个 τ满足 (24b-24e) 的情形，我们定义 T_t^* 为以下优化问题的解

$$\overline{T}_t^* := \max_{\tau}\max_{c\in\mathcal{S}_t}\{\tau(c)\}, \tag{26a}$$

$$\text{s.t. } \tau:\mathcal{S}\to\mathbb{R}, \tag{26b}$$

$$\int_{\mathcal{S}_t}\tau(s)\,\mathrm{d}s = p_t, \tag{26c}$$

$$\|\nabla\tau(s)\|_\infty \le L, \tag{26d}$$

$$\tau \text{ continuous on } \mathcal{S}_t, \tag{26e}$$

相对于原问题（24b-24e）进行的约束松弛包括：（i）省略τ上的非负约束；（ii）将全局连续性约束替换为局部约束。

注意，（26）可以改写为

$$\overline{T}_t^* := \max_{c\in\mathcal{S}_t}\overline{T}_t(c), \tag{27}$$

当 $T_t(c)$ 是针对所选轨迹 $c\in\mathcal{S}_t$ 的内层问题的解时，表示为

$$\overline{T}_t(c) := \max_{\tau}\ \tau(c), \quad \text{s.t. } (26b-26e). \tag{28}$$

因此，我们不再搜索所有点 $s\in\mathcal{S}_t$ 上可达到的最高值，而是先考察在选定点 $c\in\mathcal{S}_t$ 上能够达到的最高值。

如正文所述，由于每个划分 $\mathcal{S}_t$ 都是超矩形，因此我们可以用其各个分量区间 $\mathcal{S}_t=[\alpha_1^{(t)},\beta_1^{(t)}]\times[\alpha_2^{(t)},\beta_2^{(t)}]\times\cdots\times[\alpha_n^{(t)},\beta_n^{(t)}]$ 来描述它，其中所有 $t\in S$ 均成立。

引理 4 (方程 (28) 的解的导数)。对于函数 τ_c^*，若要使 (28)达到最大值，对于选定的 $c\in\mathcal{S}_t$，必须满足

$$\nabla_j\tau_c^*(s) = \begin{cases} L & for\ s_j\in[\alpha_j^{(t)},c_j),\\ -L & for\ s_j\in[c_j,\beta_j^{(t)}],\end{cases} \tag{29}$$

对于 $j\in\{1,2,\dots,n\}$，对所有 $s\in\mathcal{S}_t$ 均成立，其中，选择让 c_j 属于第二个区间而非第一个区间是任意的，且不影响下文。

证明：设一个满足 (29) 及 (28) 中所有约束的候选最优解为 $\tau_c^*(s)$。那么，我们可以将所有其他可行解写成该候选解的扰动形式，即 $\tau(s):=\tau_c^*(s)+\gamma(s)$。为使这些 $\tau(s)$ 仍然可行（满足 (28) 中的积分约束和导数约束），扰动 $\gamma(s)$ 必须满足 $\int_{s\in\mathcal{S}_t}\gamma(s)\,\mathrm{d}s=0$，并且

$$\nabla_j\gamma(s)\in\begin{cases}[-2K,0] & \text{for } s_j\in[\alpha_j^{(t)},c_j),\\ [0,2K] & \text{for } s_j\in[c_j,\beta_j^{(t)}],\end{cases}$$

对于 $j\in\{1,2,\dots,n\}$。

注意，对于任意 $s\in\mathcal{S}_t$，$\gamma(s)$ 在 $s_j\in[\alpha_j^{(t)},c_j)$ 时沿 j 方向必须是非递增的，而在 $s_j\in[c_j,\beta_j^{(t)}]$时必须是非递减的。因此，若取点 $s\in\mathcal{S}_t$ 处的 $\gamma(s)$ 并固定所有满足 $r\in\{1,2,\dots,n\}\setminus\{j\}$ 的 s_r，则由 s_j 参数化所得的直线在 $s_j=c_j$ 处取得最小值，即对于任意 $s\in\mathcal{S}_t$，$\gamma(\begin{bmatrix}s_1 & s_2 & \cdots & s_j & \cdots & s_n\end{bmatrix}^\top)\ge\gamma($ $\begin{bmatrix}s_1 & s_2 & \cdots & c_j & \cdots & s_n\end{bmatrix}^\top)$。显然，当我们用 c 中对应的元素替换 s 中不止一个元素时，这一点仍然成立。这意味着，对于所有 $s\in\mathcal{S}_t$，$\gamma(c)\le\gamma(s)$；换言之，所有可容许扰动中，γ 的最小值位于 c。

由于对于所有 $s\in\mathcal{S}_t$，$\gamma(c)\le\gamma(s)$，且 $\int_{\mathcal{S}_t}\gamma(s)\,\mathrm{d}s=0$，因此必有 $\gamma(c)\le 0$，且仅当 $\gamma(s)\equiv 0$ 时取等号。由所有相对于 $\tau_c^*(s)$ 的可容许扰动 $\gamma(s)$，对于 $s\in\mathcal{S}_t$ 均必须满足 $\gamma(c)\le 0$ 这一事实，我们得出结论：不存在能够优于 $\tau_c^*(c)$ 的函数 $\tau(s)$；也就是说，对于所有满足 (28) 的 $\tau(s)$，均有 $\tau(c)\le\tau_c^*(c)$。 □

如正文所述，对于任意 $j\in\{1,2,\dots,n\}$、$t\in S$，定义 $\mathcal{S}_t$ 的第 j 维的区间长度记为 $\delta_j^{(t)}:=\beta_j^{(t)}-\alpha_j^{(t)}$。

引理 5 （（28）的解）。 对于任意 $c \in \mathcal{S}_t$， （28）的解具有如下闭式表示

$$\overline{T}_t(c) = \frac{p_t}{\lambda(\mathcal{S}_t)} + \frac{1}{2\lambda(\mathcal{S}_t)} L \cdot \sum_{j=1}^{n} \left(\prod_{\ell \in \{1,\dots,n\} \setminus \{j\}} \delta_\ell^{(t)} \right) \left(\left(c_j - \alpha_j^{(t)} \right)^2 + \left(\beta_j^{(t)} - c_j \right)^2 \right).$$

证明 对（29）进行积分，同时满足连续性，可得 τ_c^* 的形式，如下

$$\tau_c^*(s) = o_c(s)^\top (s - c) + C_c, \tag{30}$$

其中

$$o_{c,j}(s_j) := \begin{cases} L & \text{for } s_j \in [\alpha_j^{(t)}, c_j), \\ -L & \text{for } s_j \in [c_j, \beta_j^{(t)}], \end{cases}$$

对于所有维度 $j \in \{1, 2, \dots, n\}$。

由于 $\tau_c^*(s)$ 对于 $s \in \mathcal{S}_t$ 是（28）的一个解，因此有 $T_t(c) = \tau_c^*(c)$ 对于所有 $c \in \mathcal{S}_t$。从方程（30）可以明显看出， $\tau_c^*(c) = C_c$，即 $T_t(c) = C_c$。

The C_c 的值由积分约束 $p_t = \int_{s \in \mathcal{S}_t} T(s)\, \mathrm{d}s$ 确定，其结果为 s

$$\begin{aligned} p_t &= \int_{s \in \mathcal{S}_t} o_c(s)^\top (s - c) + C_c \, \mathrm{d}s \\ &= \int_{s \in \mathcal{S}_t} o_c(s)^\top (s - c) \, \mathrm{d}s + \lambda(\mathcal{S}_t) C_c, \end{aligned} \tag{31}$$

其中 $\int_{s \in \mathcal{S}_t} o_c(s)^\top (s - c)\, \mathrm{d}s = \sum_{j=1}^{n} \int_{s \in \mathcal{S}_t} o_{c,\ j}(s_j)(s_j - c_j)\, \mathrm{d}s$，其中，对于每个 $j \in \{1, 2, \dots, n\}$，我们得到

$$\begin{aligned} &\int_{s \in \mathcal{S}_t} o_{c,j}(s_j)(s_j - c_j) \, \mathrm{d}s \\ &\qquad = \left(\prod_{\ell \in \{1,\dots,n\} \setminus \{j\}} \delta_\ell^{(t)} \right) \int_{s_j \in [\alpha_j^{(t)}, \beta_j^{(t)}]} o_{c,j}(s_j)(s_j - c_j) \, \mathrm{d}s_j, \end{aligned}$$

由于积分中的函数只取决于 s_j，而不取决于 $\ell \in \{1, 2, \dots, n\} \setminus \{j\}$ 的任何其他 s_ℓ。我们计算上述积分，得到

$$\begin{aligned} &\int_{s_j \in [\alpha_j^{(t)}, \beta_j^{(t)}]} o_{c,j}(s_j)(s_j - c_j) \, \mathrm{d}s_j \\ &\qquad = -L \frac{1}{2} \left(\left(c_j - \alpha_j^{(t)} \right)^2 + \left(\beta_j^{(t)} - c_j \right)^2 \right), \end{aligned}$$

对于每个 $j \in \{1, 2, \dots, n\}$。

将这些结果代回（31）可得

$$p_t = \lambda(\mathcal{S}_t) C_c - \frac{1}{2} L \cdot \sum_{j=1}^{n} \left(\prod_{\ell \in \{1,\dots,n\} \setminus \{j\}} \delta_\ell^{(t)} \right) \left(\left(c_j - \alpha_j^{(t)} \right)^2 + \left(\beta_j^{(t)} - c_j \right)^2 \right),$$

重新排列后可得

$$C_c = \frac{p_t}{\lambda(\mathcal{S}_t)} + \frac{1}{2\lambda(\mathcal{S}_t)} L \cdot \sum_{j=1}^{n} \left(\prod_{\ell \in \{1,\dots,n\} \setminus \{j\}} \delta_\ell^{(t)} \right) \left(\left(c_j - \alpha_j^{(t)} \right)^2 + \left(\beta_j^{(t)} - c_j \right)^2 \right),$$

其中我们回顾 $\overline{T_t}(c) = \tau_c^*(c) = C_c$。 □

定理 4（关于（27）的解）。 方程（27）的解 具有形式 $T_t^* = \frac{p_t}{\lambda(\mathcal{S}_t)} + \frac{1}{2}L\sum_{j=1}^{n}\delta_j^{(t)}$。

证明 由于 $T_t^* = \max_{c\in\mathcal{S}_t} T_t(c)$，并且根据引理 5 我们知道

$$\begin{aligned}\overline{T}_t(c) &= \frac{p_t}{\lambda(\mathcal{S}_t)} + \frac{1}{2\lambda(\mathcal{S}_t)}L \\ &\cdot\sum_{j=1}^{n}\left(\prod_{\ell\in\{1,\dots,n\}\setminus\{j\}}\delta_\ell^{(t)}\right)\left(\left(c_j-\alpha_j^{(t)}\right)^2+\left(\beta_j^{(t)}-c_j\right)^2\right),\end{aligned}$$

由于 $(c-\alpha)^2+(\beta-c)^2$ 是一个在 $(\alpha+\beta)/2$ 处取得最小值的抛物线，因此显然当 $c_j \in \{\alpha_j^{(t)}, \beta_j^{(t)}\}$ 时，对于所有 $j \in \{1,2,\dots,n\}$，该式取得最大值。注意，因此例如当 $c_j = \alpha_j^{(t)}$ 时，对于所有 $j \in \{1,2,\dots,n\}$，我们得到

$$\begin{aligned}\overline{T}_t^* &= \max_{c\in\mathcal{S}_t} C_c \\ &= \frac{p_t}{\lambda(\mathcal{S}_t)} + \frac{1}{2\lambda(\mathcal{S}_t)}L\sum_{j=1}^{n}\left(\prod_{\ell\in\{1,\dots,n\}\setminus\{j\}}\delta_\ell^{(t)}\right)\left(\beta_j^{(t)}-\alpha_j^{(t)}\right)^2 \\ &= \frac{p_t}{\lambda(\mathcal{S}_t)} + \frac{1}{2\lambda(\mathcal{S}_t)}L\sum_{j=1}^{n}\delta_j^{(t)}\lambda(\mathcal{S}_t) = \frac{p_t}{\lambda(\mathcal{S}_t)} + \frac{1}{2}L\sum_{j=1}^{n}\delta_j^{(t)}.\end{aligned}$$

□

引理 2的证明： 将上界 $T_t^* \geq \max_{c\in\mathcal{S}_t}\{\tau(c)\}$ 解（见定理 4）代入（25），得到

$$\begin{aligned}\Delta &\leq \sum_{t\in S} p_t \log\left(\frac{\overline{T}_t^*}{p_t/\lambda(\mathcal{S}_t)}\right) \\ &= \sum_{t\in S} p_t \log\left(1+\frac{1}{2p_t}L\lambda(\mathcal{S}_t)\sum_{j=1}^{n}\delta_j^{(t)}\right),\end{aligned}$$

我们进一步利用 $\delta := \max_{t\in S}\max_{j\in\{1,2,\dots,n\}}\delta_j^{(t)}$ 对其取上界；对此注意到 $\max_{t\in S}\lambda(\mathcal{S}_t) \leq (\delta)^n$，从而得到一个更简单的上界（对于均匀超立方体划分，这不会引入额外的保守性）：

$$\Delta \leq \sum_{t\in S} p_t \log\left(1+\frac{n}{2p_t}L\delta^{n+1}\right). \tag{32}$$

□

C 全局离散化差界（定理 2）

引理 6（凹性）。 形如

$$f(x) = \begin{cases} x\log(1+Cx^{-1}) & for\ x>0, \\ 0 & for\ x=0, \end{cases}$$

的函数在 $x \geq 0$ 上是凹函数，其中 $C \geq 0$。

证明 二阶导数 $f''(x) = \frac{C}{\ln(2)}\frac{-C}{x(x+C)^2}$ 显然对所有 $x>0, C\geq 0$ 都为负。 □

定理 2 的证明 詹森不等式指出，对于任意凹函数 $f(x)$，相对于任意有限点集 p_1，p_2，……，p_n，以及非负权重 α_1，α_2，……，α_n，且满足 $\sum_i \alpha_i = 1$，以下不等式成立 $f\left(\sum_{t\in S}\alpha_t p_t\right) \geq \sum_{t\in S}\alpha_t f(p_t)$。

利用我们的 T 离散化满足 $\sum_{t\in S} p_t = 1$ 且 $p_t \geq 0$ 对所有 $t \in S$ 均成立这一事实，并且（32）关于 p_t 是凹函数（引理6），由詹森不等式可得

$$f(1/|S|) = f\left(\sum_{t\in S}\frac{1}{|S|}p_t\right) \geq \sum_{t\in S}\frac{1}{|S|}f(p_t),$$

将其乘以 $|S|$，得到 $|S|f\ (1/|S|)\ \geq \sum_{t\in S} f\ (p_t)$，对于任意满足和为一的 $(p_t)_{t\in S}$。因此，无论真实的 p为何，对（32）的求和都可以由上界限定为

$$\sum_{t\in S} p_t \log(1+\frac{nL}{2p_t}\overline{\delta}^{n+1}) \leq |S|\left(1/|S|\log(1+\frac{n}{2}|S|L\overline{\delta}^{n+1})\right).$$

此外，我们可以断定，任何满足 $|S|\delta^{n+1}\to 0$ 的离散化方案都能确保 $\varepsilon \rightrightarrows 0$，因为 $\varepsilon \geq \Delta \to 0$ 可由$\log(1)=0$这一事实推出，从而意味着对日益精细的离散化具有收敛性。 □

D 局部离散化差界（定理 3）

相对于均匀分布的 KL 散度具有马尔可夫系统上熵度量众所周知的递归性质 [5,19,20,22]。令 Φ如（13），并令 $\mathbb{R}^{\mathcal{X}}$ 表示将 $\mathcal{X}$中的值映射到实数轴的函数空间，我们定义泛函 $\phi:\ \mathcal{P}\ (\mathcal{S})\ \times\mathbb{R}^{\mathcal{X}}\to\mathbb{R}$，为

$$\phi(q,\mathcal{V}) := \int_{\mathcal{X}} q(x')\log\frac{q(x')}{1/\lambda(\mathcal{X})}\ \mathrm{d}x' + \int_{\mathcal{X}} q(x')\mathcal{V}(x')\ \mathrm{d}x'.$$

引理 7（离散KL散度对均匀分布的递归性）。 设 $p\in\mathbb{P}^{|S|}$ 是$S=X^{K+1}$上的离散轨迹分布，由初始分布 π 和马尔可夫核 $\{P_k\}_{k=0}^{K-1}$ 诱导。则 p 相对于均匀分布的离散*KL*散度具有如下形式

$$\mathrm{KL}_D(p\|p^{\mathrm{u}}) = \Phi(\pi, V_0), \tag{33a}$$

where, for $k=K-1,K-2,\ldots,0,$

$$V_k(i) = \Phi(P_k(i,\cdot), V_{k+1}), \tag{33b}$$

对所有 $i\in X$，由 $V_K(i)=0$ 初始化。

证明：参见附录E。

引理 *8*（连续*KL*散度对均匀分布）。给定一个马尔可夫系统 $\mathcal{M}=(\mathcal{X},q_0,q,K)$，其诱导的轨迹分布为 $T(1)$，相对于均匀分布的 *KL* 散度可表示为 $KL(T\|U)=\phi(q_0,\mathcal{V}_0)$，其中，对于 $k=K-1,K-2,\ldots,0$, $\mathcal{V}_k(x)=\phi(q(x,\cdot)\ \mathcal{V}_{k+1})$，由 $\mathcal{V}_K(x)=0$ 初始化，对所有 $x\in\mathcal{X}$ 均成立。

证明：参见附录F。

D.1 局部离散化差界

定理*3*的证明：本节将引理2、7和8 结合起来，通过修正后的离散空间动态规划递推，对连续系统对均匀分布的KL散度进行界定。

对于 $x\in\mathcal{X}_i$，定义其诱导的离散转移概率 $P_x(j) := \int_{\mathcal{X}_j} q(x,\ x')\ \mathrm{d}x',\qquad j\in X$。根据构造，对于所有 $x\in\mathcal{X}_i$, $P_x(\cdot)\ \in\mathbb{P}_i$。

由引理 2，可得对所有 $x\in\mathcal{X}$，都有

$$\begin{aligned}\phi(q(x,\cdot),0)-\Phi(P_x,0) &= \mathrm{KL}(q(x,\cdot)\|U_{\mathcal{X}}) - \mathrm{KL}_D(P_x\|p^u_{\mathcal{X}})\\ &\leq \varepsilon(P_x,n_x,L_{\nabla q},\overline{\delta}).\end{aligned}$$

整理后得到

$$\Phi(P_x,0)+\varepsilon(P_x,n_x,L_{\nabla q},\overline{\delta})\geq\phi(q(x,\cdot),0),\qquad\forall x\in\mathcal{X}_i.$$

引理 9。 若对所有 $i\in X$ 都满足以下逐点条件

$$\text{(i)}\qquad \Phi^{\varepsilon}(P_x,0)\ \geq\ \phi(q(x,\cdot),0),\qquad\forall x\in\mathcal{X}_i, \tag{34}$$

则 (14)，以 $V_K^{\varepsilon}(i)\ =0$ 初始化后，对所有 $i\in X$，保证条件

$$\text{(ii)}\quad V_k^{\varepsilon}(i)\geq\max_{x\in\mathcal{X}_i}\mathcal{V}_k(x),$$

对所有 i、$k\in X\times\{0,1,\ldots,K\}$ 均成立，并且（由此）可得

$$\Phi^{\varepsilon}(\pi,V_0^{\varepsilon})\geq\phi(q_0,\mathcal{V}_0).$$

我们证明这蕴含条件（ii）。当 $k = K$ 时，该断言显然成立。假设条件（ii）在时间$k+1$成立，并考虑时间 k。那么，对于所有 $i \in X$,

$$
\begin{aligned}
V_k^\varepsilon(i) &= \max_{p' \in \mathbb{P}_i} \Phi^\varepsilon(p', V_{k+1}^\varepsilon) \\
&= \max_{p' \in \mathbb{P}_i} \Big(\Phi^\varepsilon(p', 0) + \sum_{j \in X} p'_j V_{k+1}^\varepsilon(j) \Big) \\
&\geq \max_{p' \in \mathbb{P}_i} \Big(\Phi^\varepsilon(p', 0) + \sum_{j \in X} p'_j \max_{x' \in \mathcal{X}_j} \mathcal{V}_{k+1}(x') \Big) \\
&\geq \max_{x \in \mathcal{X}_i} \Big(\Phi^\varepsilon(P_x, 0) + \sum_{j \in X} \Big(\int_{\mathcal{X}_j} q(x, x') \, \mathrm{d}x' \Big) \max_{x'' \in \mathcal{X}_j} \mathcal{V}_{k+1}(x'') \Big) \\
&\geq \max_{x \in \mathcal{X}_i} \Big(\phi(q(x, \cdot), 0) + \sum_{j \in X} \int_{\mathcal{X}_j} q(x, x') \max_{x'' \in \mathcal{X}_j} \mathcal{V}_{k+1}(x'') \, \mathrm{d}x' \Big) \\
&\geq \max_{x \in \mathcal{X}_i} \Big(\phi(q(x, \cdot), 0) + \sum_{j \in X} \int_{\mathcal{X}_j} q(x, x') \, \mathcal{V}_{k+1}(x') \, \mathrm{d}x' \Big) \\
&= \max_{x \in \mathcal{X}_i} \Big(\phi(q(x, \cdot), 0) + \int_{\mathcal{X}} q(x, x') \, \mathcal{V}_{k+1}(x') \, \mathrm{d}x' \Big) = \max_{x \in \mathcal{X}_i} \mathcal{V}_k(x).
\end{aligned}
$$

T这就证明了条件（ii）。最后的结论 $\Phi^\varepsilon(\pi,\ V_0^\varepsilon) \geq \phi(q_0,\ \mathcal{V}_0)$ 可类似地推出 . □

E Disc相对于均匀分布的 KL 散度的均匀递归性（引理 7）

我们定义一种替代性的离散 KL 散度记法，用来表示定义在相同字母表 X 上的两个随机变量 X 和 Y 之间的散度:

$$
\mathrm{KL}_D(\mathsf{X} \| \mathsf{Y}) := \sum_{i \in X} \mathrm{Prob}(\mathsf{X} = i) \log \frac{\mathrm{Prob}(\mathsf{X} = i)}{\mathrm{Prob}(\mathsf{Y} = i)}. \tag{35}
$$

引理 7 的证明：我们先通过归纳法证明

$$
V_0(i) = \mathrm{KL}_D(\mathsf{X}_{1:K} | \mathsf{X}_0 = i \| \mathsf{U}_{1:K} | \mathsf{U}_0 = i), \tag{36}
$$

对于所有 $i \in X$，按照（33b）进行迭代时。

若在第 $k+1$ 次迭代中，我们有

$$
V_{k+1}(i) = \mathrm{KL}_D(\mathsf{X}_{k+2:K} | \mathsf{X}_{k+1} = i \| \mathsf{U}_{k+2:K} | \mathsf{U}_{k+1} = i), \tag{37}
$$

对所有 $i \in X$，那么对于迭代 k，利用 (13)、(33，我们有

$$
\begin{aligned}
V_k(i) &= \underbrace{\sum_{j \in X} P_k(i,j) \log \frac{P_k(i,j)\lambda(\mathcal{X})}{\lambda(\mathcal{X}_j)}}_{\mathrm{KL}_D(\mathsf{X}_{k+1} | \mathsf{X}_k = i \| \mathsf{U}_{k+1} | \mathsf{U}_k = i)} \\
&\qquad + \underbrace{\sum_{j \in X} P_k(i,j) V_{k+1}(j)}_{\mathrm{KL}_D(\mathsf{X}_{k+2:K} | \mathsf{X}_{k+1} \| \mathsf{U}_{k+2:K} | \mathsf{U}_{k+1})} \\
&= \mathrm{KL}_D(\mathsf{X}_{k+1:K} | \mathsf{X}_k = i \| \mathsf{U}_{k+1:K} | \mathsf{U}_k = i).
\end{aligned} \tag{38}
$$

其次，由于我们以 $V_K(i) = 0,\ i \in X$ 初始化，因此有

$$
\begin{aligned}
V_{K-1}(i) &= \sum_{j \in X} P_{K-1}(i,j) \log \frac{P_{K-1}(i,j)\lambda(\mathcal{X})}{\lambda(\mathcal{X}_j)} \\
&= \mathrm{KL}_D(\mathsf{X}_K | \mathsf{X}_{K-1} = i \| \mathsf{U}_K | \mathsf{U}_{K-1} = i),
\end{aligned}
$$

满足 (37) 的 $k = K-2$。利用恒等式 (36)，我们现在可以通过注意到以下内容来验证恒等式 (33a)

$$
\begin{aligned}
\Phi(\pi, V_0) &= \mathrm{KL}_D(\mathsf{X}_0 \| \mathsf{U}_0) + \mathrm{KL}_D(\mathsf{X}_{1:K} | \mathsf{X}_0 \| \mathsf{U}_{1:K} | \mathsf{U}_0) \\
&= \mathrm{KL}_D(\mathsf{X}_0 \cdots \mathsf{X}_K \| \mathsf{U}_0 \cdots \mathsf{U}_K),
\end{aligned}
$$

最后一步是根据条件 KL 散度的链式法则得出的。 □

F 连续条件下 KL散度至均匀递归性（引理 8）

我们来定义一种替代的 KL 散度记法，用于表示同一紧字母表 $\mathcal{X}$ 上两个随机变量 X 和 Y 之间的散度：

$$\mathrm{KL}(\mathsf{X}\|\mathsf{Y}) := \int_{\mathcal{X}} q(\mathsf{X}=x)\log\frac{q(\mathsf{X}=x)}{q(\mathsf{Y}=x)}\,\mathrm{d}x, \tag{39}$$

其中 $\mathrm{Prob}(\mathsf{X}\in A) = \int_A q_{\mathsf{X}}(x)\,\mathrm{d}x$，即 $q(\cdot)$ 是相应的概率密度。记号 $q(\mathsf{X}=x)$ 和 $q_{\mathsf{X}}(x)$ 可互换使用，表示同一对象。

引理 *8* 的证明：该证明可直接通过将上述连续条件（39）应用于引理 7 的证明中的相同步骤得出。□

G 梯度界定（引理 3）

引理 *3* 的证明：注意到 $\mathcal{S} = \mathcal{X}^{K+1}$，因此我们可以将轨迹 $s\in\mathcal{S}$ 解释为 $s = x_0x_1\cdots x_K$。对于 $k\in\{1,2,\ldots,K-1\}$，$T(x_0x_1\cdots x_K) = q_0(x_0)\prod_{k=1}^{K} q(x_{k-1},\ x_k)$ 关于 $x_{k,\ j}$ 中元素 $j\in\{1,2,\ldots,n_x\}$ 的导数为

$$\begin{aligned}\nabla_{x_{k,j}}T = q_0(x_0)\Big(\prod_{t\in\{1,\ldots,K\}\setminus\{k\}} q(x_{t-1},x_t)\Big)\frac{\partial}{\partial x_{k,j}}q(x_{k-1},x_k)\\ + q_0(x_0)\Big(\prod_{t\in\{1,\ldots,K\}\setminus\{k+1\}} q(x_{t-1},x_t)\Big)\frac{\partial}{\partial x_{k,j}}q(x_k,x_{k+1}).\end{aligned}$$

根据上界 $L_{\nabla q}\in\mathbb{R}_{\geq 0}$ 和 $L_q\in\mathbb{R}_{\geq 0}$ (15)，对于所有 $j\in\{1,2,\ldots,n_x\}$，我们将上述方程上界为

$$\|\nabla_{x_{k,j}}T(x_{0:K})\|_\infty \leq 2(L_q)^K L_{\nabla q},$$

对于任意 $k\in\{1,2,\ldots,K-1\}$。对 $k\in\{0,K\}$ 重复上述步骤可得 $\nabla_{x_{k,\ j}}T(x_{0:K}) \leq (L_q)^K L_{\nabla q}$，其中$k\in\{0,K\}$。□