部分可观测系统中确定性等价策略的次优性界

Berk Bozkurt、Aditya Mahajan、Ashutosh Nayyar 和 Yi Ouyang

arXiv:2602.02814v1 [math.OC] 2 Feb 2026

摘要—**本文提出了随机控制中确定性等价原理的一种推广。对于具有输出反馈和二次代价的线性系统，经典确定性等价原理的一种解释如下：每个时刻的最优动作，都是将随机线性系统的最优状态反馈策略应用于状态的最小均方误差（MMSE）估计值而得到的。受此解释启发，我们研究了一般的（非线性）部分可观测随机系统中的确定性等价策略，允许使用任意状态估计值，而不局限于最小均方误差估计。在此类设置下，确定性等价策略并非最优。对于代价和动力学在适当意义下光滑的模型，我们推导出了确定性等价策略次优性的上界。我们给出了若干示例，以说明这些结果。**

I. 引言

在机器人学、自主系统、金融、医疗保健及其他领域的许多应用中，决策者无法获取系统的完整状态。这类系统通常被建模为部分可观测马尔可夫决策过程（POMDP）。求解 POMDP 的标准方法是将其转换为全可观测马尔可夫决策过程（MDP），并将决策者对当前状态的后验信念视为充分统计量 [2], [3]。关于利用由此得到的信念状态马尔可夫决策过程的结构来求得最优及近似最优策略的算法，已有大量研究。

然而，寻找最优策略是 PSPACE困难的 [4]。大多数用于寻找最优策略的算法，在状态空间与观测空间规模上具有指数级最坏情况复杂度，因此不适用于大规模问题。与此同时，许多启发式方法（如基于点的值迭代方法 [5] ）计算效率较高，但缺乏可证明的性能保证。事实上，寻找近似最优策略同样是 PSPACE困难的 [6], [7]，这表明通用算法可能无法高效处理所有部分可观测马尔可夫决策过程模型。

这些挑战促使人们对识别兼具计算可处理性和良好性能保证的结构化策略类别产生了浓厚兴趣。例如，基于有限窗口内过往观测的策略 [8]（在强化学习中称为帧堆叠），以及更一般地，基于智能体状态的策略；智能体状态是过去观测和动作的递归可更新函数 [9], [10]。已有多篇论文提出了此类策略表现良好的充分条件，包括近似信息状态 [11]、滤波器稳定性 [12]–[14]、弱揭示观测 [15]、较小的覆盖数 [16]、低秩结构 [17]，以及揭示型观测模型 [18]。这些结果表明，对于某个特定的部分可观测马尔可夫决策过程子类，结构化策略可以达到近似最优。

在具有二次代价和高斯噪声的线性系统（即所谓的线性二次高斯问题）中，最优策略可视为具有如下结构的结构化策略：每一时刻的最优动作是最小均方误差状态估计的线性函数，且相应的反馈增益与最优状态反馈 控制（针对将所有随机变量替换为其均值后得到的确定性系统）的反馈增益相同。该结果通常称为随机控制的确定性等价原理 [19]–[21]，并已推广到其他情形，包括具有非线性动力学的系统 [22]–[24] 以及风险敏感控制 [25]。

本文提出将确定性等价原理推广到一般的部分可观测马尔可夫决策过程。我们的推广基于对LQG确定性等价原理略有不同的解释：考虑具有完全可观测状态的随机 系统的最优策略，并在最小均方误差状态估计处评估该策略。强化学习与自适应控制文献中已采用对确定性等价原理的类似观点 [26]，有时称之为部分随机确定性等价 [27]。为清晰起见，我们给出这一解释的形式化描述。

令 $\mathcal{P}$ 表示状态为 $s_t \in \mathcal{S}$、动作为 $a_t \in \mathcal{A}$、输出为 $y_t \in \mathcal{Y}$ 的部分可观测线性系统，其中 $\mathcal{S}$、$\mathcal{A}$ 和 $\mathcal{Y}$ 均为欧几里得空间。令 $\mathcal{M}$ 表示完全可观测随机线性系统，在该系统中，决策者可以获取状态。请注意，完全观测系统 $\mathcal{M}$ 不同于确定性等价中通常假设的系统。与标准确定性等价原理一样，我们假设 $\mathcal{M}$ 是完全可观测的，但不假设 $\mathcal{M}$ 的动力学是确定性的。为简单起见，假设系统运行有限时域 T。令 $\pi^{\mathcal{M}} = (\pi_1^{\mathcal{M}}, \dots, \pi_T^{\mathcal{M}})$ 表示模型 $\mathcal{M}$ 的最优策略，并令 $\mu^{\mathcal{P}} = (\mu_1^{\mathcal{P}}, \dots \mu_T^{\mathcal{P}})$ 表示模型 $\mathcal{P}$ 的最优策略。此外，对于截至时刻 $\mathcal{P}$ 的任意历史 $h_t = (y_1, a_1, y_2, a_2, \dots, y_t)$，

本文的初步版本发表于 [1]。Berk Bozkurt任职于加拿大魁北克省蒙特利尔INLAN。（电子邮箱：berk.bozkurt@mail.mcgill.ca）。Aditya Mahajan任职于加拿大魁北克省蒙特利尔麦吉尔大学电气与计算机工程系。（电子邮箱：aditya.mahajan@mcgill.ca）。Ashutosh Nayyar任职于美国加利福尼亚州洛杉矶南加州大学电气与计算机工程系。（电子邮箱：ashutosn@usc.edu）。Yi Ouyang任职于美国加利福尼亚州圣克拉拉Atmanity。（电子邮箱：ouyangyii@gmail.com）。

给定截至时刻 t 的观测和动作，令 $\mathcal{E}_t(h_t)$ 表示给定历史 h_t 时状态的最小均方误差估计。那么，LQG 最优控制的标准结果是

$$\mu_t^{\mathcal{P}}(h_t) = \pi_t^{\mathcal{M}}(\mathcal{E}_t(h_t)).$$

本文考虑上述结果的两种一般化。

1) 我们允许 $\mathcal{E}_t$ 采用状态的任意估计器，而不是仅限于最小均方误差估计器。2) 我们考虑一般的部分可观测马尔可夫决策过程，而不是仅限于线性系统。

在这一一般设定下，我们将确定性等价策略$\mu^{\mathsf{CE}} = (\mu_1^{\mathsf{CE}}, \ldots, \mu_T^{\mathsf{CE}})$定义为

$$\mu_t^{\mathsf{CE}}(h_t) = \pi_t^{\mathcal{M}}(\mathcal{E}_t(h_t)). \tag{1}$$

一般而言， μ^{CE} 并非最优。我们的主要结果是刻画确定性等价策略 μ^{CE} 的次优程度。

我们的结果可视为对部分可观测马尔可夫决策过程中结构化策略次优性差距的刻画。已有大量文献利用预测状态表示 [28], [29],双模拟度量 [30],近似信息状态 (AIS) [11], 和滤波器稳定性 [12]–[14]等工具推导此类次优性差距。我们的分析基于 [11]的近似信息状态近似界。

本文的主要贡献如下:

- 我们在代价与动力学在适当意义下光滑的假设下，推导了确定性等价策略次优性的显式界。我们的界取决于条件期望估计误差的最坏情况值。
- 我们将结果推广到具有状态抽象的情形，其中估计器产生的是抽象状态的估计值而非完整状态，从而使我们的框架能够应用于必须进行状态聚合/量化或特征抽象的大规模系统。
- 我们通过若干示例说明我们的结果，包括：具有有界观测噪声的系统、间歇性退化观测、采用事件触发通信的控制、学习与自适应控制情形，以及对非齐次多粒子系统的控制。这些示例表明，当估计误差较小时，确定性等价策略可以达到接近最优的性能，从而为精确求解部分可观测马尔可夫决策过程提供实用且计算上可行的替代方案。

该结果的初步版本发表于 [1]。那里的分析仅限于估计系统完整状态的确定性等价策略，且结果是在更强的假设下得到的。本文所考虑的状态抽象模型是新的，并且结果是在更弱的假设下推导的。

本文其余部分组织如下。我们在第II节给出系统模型，并定义确定性等价策略。我们在第III节通过若干示例说明我们的结果。我们在第IV 节给出证明，并在第V节做出结论。

记号：我们用大写字母表示随机变量（例如 S、 A等），用相应的小写字母表示其实现值（例如 s、 a等），并用相应的花体字母表示其实现值空间（例如 $\mathcal{S}$、 $\mathcal{A}$等）。下标表示时间，因此 S_t 表示时刻 t的变量。记号 $S_{1:t}$ 是序列（$S_1, \ldots, S_t$）的简写。

我们用 R 表示实数集。对于拓扑空间 $\mathcal{X}$， $\Delta(\mathcal{X})$ 表示 $\mathcal{X}$ 上所有概率测度的集合，而 $\mathcal{B}(\mathcal{X})$ 表示 $\mathcal{X}$ 上所有有界可测实值函数的集合。

我们用 P(·) 表示某个事件的概率，用 E[·] 表示某个随机变量的期望。我们用记号 $\mathrm{P}(S_{t+1} \in M_S | s_t, a_t)$ 简写 $\mathrm{P}(S_{t+1} \in M_S | S_t = s_t, A_t = a_t)$。

给定一个度量空间 $(\mathcal{S}, d_{\mathcal{S}})$，两个概率分布 $\nu_1, \nu_2 \in \Delta(\mathcal{S})$ 之间的 Wasserstein-1 距离定义为

$$d_{\mathrm{Was}}(\nu_1, \nu_2) = \inf_{(S_1, S_2) \sim \Gamma(\nu_1, \nu_2)} \mathbb{E}[d_{\mathcal{S}}(S_1, S_2)],$$

其中， $\Gamma(\nu_1, \nu_2)$ 表示 $\mathcal{S} \times \mathcal{S}$ 上所有边缘分布为 ν_1 和 ν_2 的联合概率分布。对于取值于 $\mathcal{S}$ 的两个随机变量 S_1和 S_2 ，我们有时用 $d_{\mathrm{Was}}(S_1, S_2)$ 表示 S_1 和 S_2 的边缘分布之间的 Wasserstein-1 距离。对于从一个度量空间映射到另一个度量空间的函数 f ， Lip(f) 表示 f 的 Lipschitz 常数。

II. 系统模型与主要结果

考虑一个离散时间部分可观测马尔可夫决策过程（ POMDP），记为 $\mathcal{P}$，其状态空间为 $\mathcal{S}$，观测空间为 $\mathcal{Y}$，动作空间为 $\mathcal{A}$ ，并在有限时域 T上运行。令 $S_t \in \mathcal{S}$ 表示系统的状态， $Y_t \in \mathcal{Y}$表示控制器的观测， $A_t \in \mathcal{A}$ 表示控制器在时刻 t采取的控制动作。我们假设 $\mathcal{S}$， $\mathcal{Y}$ 和 $\mathcal{A}$ 是度量空间，并用 $d_{\mathcal{S}}$ 表示 $\mathcal{S}$上的度量。

初始状态和观测（S_1， Y_1）按照概率分布 $\xi \in \Delta(\mathcal{S} \times \mathcal{Y})$ 分布。假设动力学和观测具有马尔可夫性。具体而言，我们假设存在随机核P_t： $\mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S} \times \mathcal{Y})$ ， $t \in \{1, \ldots, T-1\}$，使得对任意 $t \in \{1, \ldots, T-1\}$，分别属于 $\mathcal{S}$和$\mathcal{Y}$ 的任意博雷尔子集 M_S， M_Y，以及分别为 $S_{1:t}$、 $Y_{1:t}$、 $A_{1:t}$的任意实现 $s_{1:t}$、 $y_{1:t}$ 和 $a_{1:t}$ ，我们有

$$\begin{aligned} &\mathbb{P}(S_{t+1} \in M_S, Y_{t+1} \in M_Y | s_{1:t}, y_{1:t}, a_{1:t}) \\ &\quad = \mathbb{P}(S_{t+1} \in M_S, Y_{t+1} \in M_Y | s_t, a_t) \\ &\quad =: P_t(M_S, M_Y | s_t, a_t). \end{aligned} \tag{2}$$

我们用记号 $P_{S,t}(\cdot|s_t, a_t)$ 和 $P_{Y,t}(\cdot|s_t, a_t)$ 表示 $P_t(\cdot, \cdot|s_t, a_t)$ 的状态和观测边缘分布。

在每个时刻 t，系统都会产生单步成本 $c_t(S_t, A_t)$，该成本一致有界，即存在一个 $c_{\max} \in \mathbb{R}$，使得 $\sup_{s \in \mathcal{S}, a \in \mathcal{A}} |c_t(s, a)| \le c_{\max}$。

控制器可以获取截至时间 t 的观测和动作历史 $h_t = \{y_{1:t}, a_{1:t-1}\}$ 。令 $\mathcal{H}_t$ 表示 h_t 的实现值空间。令 $\mu = (\mu_1, \ldots, \mu_T)$ 表示任意历史

依赖型确定性策略。策略μ 的价值函数定义为

$$W_t^{\mathcal{P},\mu}(h_t) = \mathbb{E}^{\mu}\left[\sum_{\tau=t}^{T} c_\tau(s_\tau, a_\tau) \,\middle|\, h_t\right]$$

其中E^{μ} 表示由策略 μ诱导的系统变量上联合概率测度下的期望。最优价值函数定义为

$$W_t^{\mathcal{P}}(h_t) = \inf_{\mu} W_t^{\mathcal{P},\mu}(h_t),$$

其中下确界取遍所有依赖历史的策略。

在部分可观测马尔可夫决策过程中寻找最优策略的标准方法是使用基于信念状态的动态规划 [2], [3], , 这些方法在计算上具有挑战性。如引言中所讨论的，确定性等价策略提供了一种有吸引力的替代方法。在本节其余部分，我们刻画此类策略的次优性。

A. 确定性等价策略

考虑上述随机系统的状态反馈控制器，其中控制器在时刻 t可以获取状态 S_t 。该系统是一个有限时域马尔可夫决策过程（马尔可夫决策过程） $\mathcal{M}$，其状态空间为 $\mathcal{S}$，动作空间为$\mathcal{A}$，动力学为$P_{S,\ t}$，以及每步代价 c_t。

我们需要一个技术性假设，以确保马尔可夫决策过程 $\mathcal{M}$存在最优策略。

定义1（可测选择）若一个MDP $\langle \mathcal{S}$、 $\mathcal{A}$、$\{P_{S,t}\}_{t=1}^{T-1}$、$\{c_t\}_{t=1}^{T}$，$T\rangle$ 满足可测选择，则对于每个可测函数 $V\colon \mathcal{S} \to \mathrm{R}$以及每个时刻 $t \in \{1,\dots,T-1\}$，都存在一个可测选择器 $\pi\colon \mathcal{S} \to \mathcal{A}$ 使得

$$\inf_{a\in\mathcal{A}}\left\{c_t(s,a) + \int_{\mathcal{S}} V(s')P_{S,t}(ds'|s,a)\right\}$$
$$= c_t(s,\pi(s)) + \int_{\mathcal{S}} V(s')P_{S,t}(ds'|s,\pi(s)) =: V_+(s),$$

且上文定义的 $V_+\colon \mathcal{S} \to \mathrm{R}$ 是一个可测函数。

假设1 模型 $\mathcal{M}$ 满足可测选择。■

满足可测选择的 $\mathcal{M}$ 的一个推论是：对于该MDP [31]，存在一个最优策略 $\pi^{\mathcal{M}} = (\pi_1^{\mathcal{M}}, \dots, \pi_T^{\mathcal{M}})$，其中 $\pi_t^{\mathcal{M}}\colon \mathcal{S} \to \mathcal{A}$，并且存在相应的最优价值函数（$V_1^{\mathcal{M}}, \dots, V_T^{\mathcal{M}}$） ， $V_t^{\mathcal{M}}\colon \mathcal{S} \to \mathrm{R}$。

现在，我们使用MDP $\mathcal{M}$ 的最优策略 $\pi^{\mathcal{M}}$ ，为POMDP $\mathcal{P}$定义一个可行策略。假设给定一系列状态估计函数 $\{\mathcal{E}_t\}_{t=1}^{T}$，其中$\mathcal{E}_t\colon \mathcal{H}_t \to \mathcal{S}$。例如， $\mathcal{E}_t$ 可以是条件均值或最大后验概率（MAP）估计器，它们依赖于给定观测与动作历史的状态条件分布。或者，该估计器也可以是最近若干次观测和动作的简单函数（例如线性函数）。

我们称依赖历史的策略 $\mu^{\mathcal{E}} = (\mu_1^{\mathcal{E}}, \dots\dots, \mu_T^{\mathcal{E}})$ 是关于马尔可夫决策过程 $\mathcal{M}$ =确定性等价的 $\langle \mathcal{S}$、 $\mathcal{A}$、 $\{P_{S,t}\}_{t=1}^{T-1}$、 $\{c_t\}_{t=1}^{T}$、 $T\rangle$ 以及估计器 $\{\mathcal{E}_t\colon \mathcal{H}_t \to \mathcal{S}\}_{t\ge 1}$ 如果

$$\mu_t^{\mathcal{E}}(h_t) = \pi_t^{\mathcal{M}}(\mathcal{E}_t(h_t)). \tag{3}$$

换言之，确定性等价策略将 $\mathcal{E}_t(h_t)$ 视为马尔可夫决策过程 $\mathcal{M}$ 的状态的无误差估计，然后按照 $\mathcal{M}$ 的最优策略采取行动。如前所述，在使用条件均值作为估计值时，这类策略在线性二次高斯设定下是最优的，但一般而言并非最优。我们希望为确定性等价策略的次优性提供一个界。具体而言，我们关注策略 $\mu^{\mathcal{E}}$的价值函数与部分可观测马尔可夫决策过程的最优价值函数之间差距的上界，即 $W^{\mathcal{P},\ \mu^{\mathcal{E}}}_{\ t}(h_t) - W_t^{\mathcal{P}}(h_t)$ 的一个上界。我们的结果在以下关于单步成本和系统动力学“光滑性”的技术假设下给出了这样的界。

假设2 存在一列凹且非递减的函数 F_t^P， $F_t^c\colon \mathrm{R}_{\ge 0} \to \mathrm{R}_{\ge 0}$，$t \in \{1,\dots,T\}$，使得对于任意 s， $s' \in \mathcal{S}$ 和 $a \in \mathcal{A}$，均有

$$d_{\mathrm{Was}}(P_{S,t}(\cdot|s,a), P_{S,t}(\cdot|s',a)) \le F_t^P(d_{\mathcal{S}}(s,s')) \tag{4}$$

and

$$\left|c_t(s,a) - c_t(s',a)\right| \le F_t^c(d_{\mathcal{S}}(s,s')). \tag{5}$$

■

注1 当 F_t^P、 F_t^c为线性函数，即 $F_t^P(x) = L_t^P x$和$F_t^c(x) = L_t^c x$满足某些正常数 L_t^P和 L_t^c时，假设2便简化为假设动力学和单步成本是利普希茨连续的。这是关于动力学和单步成本光滑性的标准假设，并意味着马尔可夫决策过程 $\mathcal{M}$ [32]的价值函数也具有光滑性（利普希茨连续性）。具体而言，按照 [32],中的论证，我们有

$$\mathrm{Lip}(V_t^{\mathcal{M}}) \le L_t^c + L_{t+1}^P \,\mathrm{Lip}(V_{t+1}^{\mathcal{M}}),$$

对其展开后，可得到 $V_t^{\mathcal{M}}$ 的利普希茨常数关于 $\{L_\tau^P\}_{\tau=t+1}^{T-1}$ 和 $\{L_\tau^c\}_{\tau=t}^{T}$ 的上界。

我们对确定性等价策略 $\mu_t^{\mathcal{E}}(h_t)$ （定义见（3））的次优性差距所给出的界，取决于状态估计函数 $\mathcal{E}_t$产生的估计值的质量，我们使用状态空间上的度量 $d_{\mathcal{S}}$ 对其进行评估。对于每个时刻 t，我们定义

$$\eta_t := \sup_{h_t\in\mathcal{H}_t} \mathbb{E}[d_{\mathcal{S}}(S_t, \mathcal{E}_t(h_t))|h_t]. \tag{6}$$

备注2 注意，式（6）的右端不依赖于策略，因为给定历史 S_t h_t 时当前状态的条件概率分布与策略无关 [2], [3]。

假设 3 对于 $t = 1,\ 2,\ \dots\dots,\ T$有 $\eta_t < \infty$ 其中η_t 由式（6）给出。■

我们现在可以陈述第一个结果。

定理1 定义

$$\varepsilon_t = F_t^c(\eta_t) \text{ and } \delta_t = F_t^P(\eta_t) + \eta_{t+1}$$

其中 η_t 由 (6) 给出。然后，在假设1、2 和3 下，确定性等价策略 $\mu^{\mathcal{E}}$ （定义于(3)）满足

$$W_t^{\mathcal{P},\mu^{\mathcal{E}}}(h_t) - W_t^{\mathcal{P}}(h_t) \le 2\alpha_t \tag{7}$$

其中

$$\alpha_t = \varepsilon_t + \sum_{\tau=t}^{T-1} \left[\delta_\tau \operatorname{Lip}(V_{\tau+1}^{\mathcal{M}}) + \varepsilon_{\tau+1}\right] \tag{8}$$

且 $\{V_t^{\mathcal{M}}\}_{t=1}^T$ 是马尔可夫决策过程 $\mathcal{M}$的最优价值函数。我们将在下一小节陈述并证明这一结果的一个更一般的版本。

备注 3 确定性等价策略并不适用于所有模型。例如，在某些部分可观测马尔可夫决策过程中，智能体在采取控制动作之前，可以选择支付一定代价来采取感知动作，从而获知马尔可夫决策过程的真实状态。在此类模型中，确定性等价策略永远不会选择感知动作。因此，如果感知动作的代价不太高，那么偶尔支付感知代价以了解真实状态的策略，可能优于确定性等价策略。

B. 使用状态抽象的确定性等价策略

对于状态空间较大或为连续状态空间的问题，由于维数灾难，可能难以找到马尔可夫决策过程 $\pi^{\mathcal{M}}$ 的最优策略 $\mathcal{M}$。对于此类大规模马尔可夫决策过程，通常通过考虑由状态聚合或状态量化得到的抽象马尔可夫决策过程，为马尔可夫决策过程 $\mathcal{M}$求得一个近似最优策略。在这种情况下，自然会考虑一种确定性等价策略，该策略基于抽象马尔可夫决策过程的最优策略以及对抽象状态的估计值。本节将正式定义此类策略，并给出定理 1 在此类情形下的推广。

假设存在配备抽象状态空间 $\tilde{\mathcal{S}}$、一个（可测）状态抽象函数 $\phi\colon \mathcal{S} \to \tilde{\mathcal{S}}$，以及两个随机核 λ^P，$\lambda^c\colon \tilde{\mathcal{S}} \to \Delta(\mathcal{S})$，使得对于每个 $\tilde{s}_t \in \tilde{\mathcal{S}}$，$\lambda^P(\phi^{-1}(\tilde{s}_t)|\tilde{s}_t) = 1$和$\lambda^c(\phi^{-1}(\tilde{s}_t)|\tilde{s}_t) = 1$。

我们构造一个抽象马尔可夫决策过程 $\tilde{\mathcal{M}} := \langle \tilde{\mathcal{S}}, \mathcal{A}, \{\tilde{P}_t\}_{t=1}^{T-1}, \{\tilde{c}_t\}_{t=1}^{T}, T\rangle$，其中动力学 $\tilde{P}_t\colon \tilde{\mathcal{S}} \times \mathcal{A} \to \tilde{\mathcal{S}}$，单步成本 $\tilde{c}_t\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ 定义如下：对于任意可测的 $M_{\tilde{\mathcal{S}}} \subset \mathcal{S}$，

$$\tilde{P}_t(\tilde{S}_{t+1} \in M_{\tilde{\mathcal{S}}} | \tilde{s}_t, a_t) = \int_{\phi^{-1}(\tilde{s}_t)} P_{\mathcal{S},t}\big(\phi(S_{t+1}) \in M_{\tilde{\mathcal{S}}} | s_t, a_t\big) \lambda^P(ds_t|\tilde{s}_t) \tag{9}$$

and

$$\tilde{c}_t(\tilde{s}_t, a_t) = \int_{\phi^{-1}(\tilde{s}_t)} c_t(s_t, a_t) \lambda^c(ds_t|\tilde{s}_t). \tag{10}$$

抽象马尔可夫决策过程的成本函数可以看作是在 $\phi^{-1}(\tilde{s}_t)$ 中对原马尔可夫决策过程代价在所有状态上的加权平均；类似的解释同样适用于抽象模型中的动力学。

注意，当 $\mathcal{S} = \mathcal{S}$ 且 $\phi(s) = s$ 时，抽象马尔可夫决策过程 $\mathcal{M}$等于马尔可夫决策过程 $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \{P_{S,t}\}_{t=1}^{T-1}, \{c_t\}_{t=1}^{T}, T\rangle$。

我们对马尔可夫决策过程 $\mathcal{M}$施加可测选择假设。

假设4 模型 $\mathcal{M}$ 满足可测选择条件。■

如第 II-A 节所述，可测选择的一个推论是存在一个最优策略 $\pi^{\mathcal{M}} = (\pi_1^{\mathcal{M}}, \ldots\ldots, \pi_T^{\mathcal{M}})$，其中 $\pi_t^{\tilde{\mathcal{M}}}\colon \tilde{\mathcal{S}} \to \mathcal{A}$，并具有相应的最优价值函数$(V_1^{\tilde{\mathcal{M}}}, \ldots\ldots, V_T^{\tilde{\mathcal{M}}})$，$V_t^{\tilde{\mathcal{M}}}\colon \tilde{\mathcal{S}} \to \mathbb{R}$，对应马尔可夫决策过程 $\tilde{\mathcal{M}}$ [31]。

现在我们使用最优策略 $\pi^{\mathcal{M}}$ 对应马尔可夫决策过程 $\mathcal{M}$ 为部分可观测马尔可夫决策过程 $\mathcal{P}$定义一个可行策略。该策略类似于第 II-A 节中的确定性等价策略，但有以下两点不同：(i) 我们估计抽象状态 S_t，并且 (ii) 使用的最优策略 $\mathcal{M}$。为方便起见，我们沿用第 II-A 节中的部分记号。

假设给定一列抽象状态估计函数 $\{\mathcal{E}_t\}_{t=1}^T$，其中 $\mathcal{E}_t\colon \mathcal{H}_t \to \tilde{\mathcal{S}}$。如果依赖历史的策略 $\mu^{\mathcal{E}} = (\mu_1^{\mathcal{E}}, \ldots, \mu_T^{\mathcal{E}})$ 满足以下条件，我们就称其为关于抽象马尔可夫决策过程 $\tilde{\mathcal{M}} = \langle \tilde{\mathcal{S}}, \mathcal{A}, \{\tilde{P}_t\}_{t=1}^{T-1}, \{\tilde{c}_t\}_{t=1}^{T}, T\rangle$ 和估计器 $\{\mathcal{E}_t\colon \mathcal{H}_t \to \tilde{\mathcal{S}}\}_{t\ge 1}$ 的确定性等价 策略：

$$\mu_t^{\mathcal{E}}(h_t) = \pi_t^{\tilde{\mathcal{M}}}(\mathcal{E}_t(h_t)). \tag{11}$$

换言之，确定性等价策略将 $\mathcal{E}_t(h_t)$视为抽象马尔可夫决策过程状态的无误差估计，并使用 $\mathcal{M}$ 的最优策略来选择动作。与之前一样，我们关注策略 $\mu^{\mathcal{E}}$的价值函数与部分可观测马尔可夫决策过程的最优价值函数之间差距的上界，即 $W_t^{\mathcal{P},\mu^{\mathcal{E}}}(h_t) - W_t^{\mathcal{P}}(h_t)$的上界。

我们对模型和状态抽象作出以下技术性假设。

假设5 存在一列凹且非递减的函数 $F_t^P, F_t^c\colon \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$，$t \in \{1, \ldots, T\}$，使得对于任意 $s, s' \in \mathcal{S}$ 和 $a \in \mathcal{A}$，均有

$$d_{\mathrm{Was}}(P_{S,t}^{\phi}(\cdot|s,a), P_{S,t}^{\phi}(\cdot|s',a)) \le F_t^P(d_{\tilde{\mathcal{S}}}(\phi(s), \phi(s'))), \tag{12}$$

and

$$\big|c_t(s,a) - c_t(s',a)\big| \le F_t^c(d_{\tilde{\mathcal{S}}}(\phi(s), \phi(s'))) \tag{13}$$

其中 $P_{S,t}^{\phi}$ 是从 $\mathcal{S} \times \mathcal{A}$ 到 Δ 的随机核（$\tilde{\mathcal{S}}$），定义如下

$$P_{S,t}^{\phi}(M_{\tilde{\mathcal{S}}}|s_t, a_t) = P_{\mathcal{S},t}\big(\phi(S_{t+1}) \in M_{\tilde{\mathcal{S}}}|s_t, a_t\big)$$

对于 $\mathcal{S}$ 的所有博雷尔子集 $M_{\tilde{\mathcal{S}}}$。 ■

假设5意味着 $\phi\colon \mathcal{S} \to \mathcal{S}$ 是一个良好的状态抽象，具体而言：如果 $\phi(s)$ 接近于 $\phi(s')$，那么对于任意动作 a，s 和 s' 处的单步成本接近，且给定 s 和 s' 时，下一个抽象状态的概率分布也接近。

我们的结果取决于 $\mathcal{E}_t$ 所生成估计值的质量。为此，对于每个时刻 t，我们定义条件期望估计误差的最坏情况值

$$\eta_t := \sup_{h_t \in \mathcal{H}_t} \mathbb{E}[d_{\tilde{\mathcal{S}}}(\phi(S_t), \mathcal{E}_t(h_t))|h_t]. \tag{14}$$

正如我们在备注 2 中所述，公式 (14) 右侧不依赖于策略，因为条件概率

状态 S_t 在历史 h_t 下的分布与策略无关 [2], [3]。

假设 6 对于 $t = 1, 2, \ldots\ldots, T$，有 $\eta_t < \infty$ 其中η_t 由式（14）给出。■

现在可以陈述主要结果。

定理2 定义

$$\varepsilon_t = F_t^c(\eta_t) \text{ and } \delta_t = F_t^P(\eta_t) + \eta_{t+1}$$

其中 η_t 由 （14）给出。在假设 *4*、*5* 和*6*成立的条件下，确定性等价策略 $\mu^{\mathcal{E}}$ （定义见 （11））满足

$$W_t^{\mathcal{P},\mu^{\mathcal{E}}}(h_t) - W_t^{\mathcal{P}}(h_t) \le 2\alpha_t \tag{15}$$

其中

$$\alpha_t = \varepsilon_t + \sum_{\tau=t}^{T-1} \left[\delta_\tau \operatorname{Lip}(V_{\tau+1}^{\tilde{\mathcal{M}}}) + \varepsilon_{\tau+1}\right] \tag{16}$$

且 $\{V_t^{\tilde{\mathcal{M}}}\}_{t=1}^T$ 是马尔可夫决策过程 $\tilde{\mathcal{M}}$ 的最优价值函数。

˜我们将在第 IV 节证明这一结果。注意，当 $\mathcal{S} = \mathcal{S}$且 $\phi(s) = s$ 时，定理 2 可简化为定理 1。如下方推论所示，定理 2 还为在原始马尔可夫决策过程中使用抽象马尔可夫决策过程的策略提供次优性界；这可以看作是 [33] 的有限时域版本。

推论 *1* 对于马尔可夫决策过程 $\mathcal{M}$ 的任意马尔可夫策略 π，令 $V^{\mathcal{M},\pi}: \mathcal{S} \to R$ 表示策略 π 的价值函数。给定抽象马尔可夫决策过程 $\mathcal{M}$ 的最优策略 $\pi^{\mathcal{M}} = (\pi_1^{\mathcal{M}}, \ldots, \pi_T^{\mathcal{M}})$˜，定义马尔可夫决策过程 $\mathcal{M}$ 的一个可行策略 $\bar{\pi}$ 如下：˜ $\bar{\pi} = (\pi_1^{\mathcal{M}} \circ \phi, \ldots, \pi_T^{\mathcal{M}} \circ \phi)$。那么，在假设 *4* 和 *5* 下，对于任意时刻 t 以及 S_t 的任意实现 s_t，我们有

$$V_t^{\mathcal{M},\bar{\pi}}(s_t) - V^{\mathcal{M}}(s_t) \le 2\alpha_t$$

其中 α_t 由 (16) 给出，且 $\varepsilon_t = F_t^c(0)$ 和 $\delta_t = F_t^P(0)$。

证明 这是定理2 的直接推论，只需考虑如下平凡情形：$Y_t = S_t$ （因此，$h_t = (s_{1:t}, a_{1:t-1})$），并令 $\mathcal{E}(h_t) = \phi(s_t)$，这意味着对于所有 t，$\eta_t = 0$。在这种情况下，$W_t^{\mathcal{P}}(h_t) = V_t^{\mathcal{M}}(s_t)$ $\mu^{\mathcal{E}} = \bar{\pi}$且 $W^{\mathcal{P},\mu^{\mathcal{E}}}(h_t) = V^{\mathcal{M},\bar{\pi}}(s_t)$。■

三、说明性示例

本节将给出若干示例，以说明确定性等价策略可能适用的观测模型（及相应的估计器）。我们将运用所得结果，针对特定观测模型推导确定性等价策略次优性的显式界。

A. 有界观测噪声

1） 系统模型： 考虑一个部分可观测马尔可夫决策过程，其中 $\mathcal{Y} = \mathcal{S}$，且系统动力学 P_t 满足

$$d_{\mathcal{S}}(Y_t, S_t) \le r,$$

其中 $r \in [0, \infty)$此外，我们假设马尔可夫决策过程模型$\mathcal{M}$ 满足可测选择（假设1），并且动力学和每步代价均满足利普希茨条件，即存在非负有限常数 L_t^P和 L_t^c，使得假设2在 $F_t^P(x) = L_t^P x$ 以及 $F_t^c(x) = L_t^c x$ 下成立（见备注1）。

2） 确定性等价策略：对于此例，我们考虑相对于原始马尔可夫决策过程 $\mathcal{M}$ 的确定性等价策略，即取状态抽象函数为 $\phi(s) = s$。此外，我们将状态估计取为最近一次观测，即 $\mathcal{E}_t(h_t) = y_t$。于是，确定性等价策略表示为

$$\mu_t^{\mathcal{E}}(h_t) = \pi_t^{\mathcal{M}}(y_t).$$

3) 次优性界：我们假设条件1 和2 均成立。此外，

$$\mathbb{E}[d_{\mathcal{S}}(S_t, \mathcal{E}_t(H_t))|h_t] = \mathbb{E}\big[d_{\mathcal{S}}(S_t, Y_t)|h_t\big] \le r.$$

因此，$\eta_t \le r$ 且假设3 也成立。此外，定理 ε_t 和 δ_t 中的1均可由以下内容上界：

$$\varepsilon_t \le rL_t^c \quad \text{and} \quad \delta_t \le r(1 + L_t^P).$$

因此，定理1 中的界可明确写为

$$W_t^{\mathcal{P},\mu^{\mathcal{E}}}(h_t) - W_t^{\mathcal{P}}(h_t) \le 2rL_t^{\mathcal{M}} \tag{17}$$

其中

$$L_t^{\mathcal{M}} = \left[L_t^c + \sum_{\tau=t}^{T-1} \left[(1 + L_\tau^P) \operatorname{Lip}(V_{\tau+1}^{\mathcal{M}}) + L_{\tau+1}^c\right]\right] \tag{18}$$

是一个常数，取决于动力学、单步成本和最优马尔可夫决策过程价值函数的利普希茨常数。

该界随 r 线性变化，这意味着随着观测逐渐接近底层状态（即“观测噪声”变小），确定性等价策略的性能会趋近于最优部分可观测马尔可夫决策过程策略的性能。

B. 间歇性退化观测

1) 系统模型： 考虑一个部分可观测马尔可夫决策过程，其中 $\mathcal{Y} = \mathcal{S}$ 且系统动力学 P_t 使得控制器要么获得良好观测（由事件 E_t表示），要么获得不良观测（由 E_t^c表示）。在 E_t下，$d_{\mathcal{S}}(Y_t, S_t) \le r$；而在 E_t^c下，$d_{\mathcal{S}}(Y_t, S_t) \le R$，其中 $0 \le r \le R < \infty$。我们假设，对于任意历史 h_t，$\mathrm{P}(E_t^c|h_t) \le p$。

此外，我们假设马尔可夫决策过程模型 $\mathcal{M}$满足可测选择条件（假设 1），且动力学和单步成本均为利普希茨连续的，即存在非负有限常数 L_t^P和 L_t^c，使得假设 2在 $F_t^P(x) = L_t^P x$ 以及 $F_t^c(x) = L_t^c x$ 下成立（见注释 1）。

2) 确定性等价策略： 与第 III-A 节中的示例一样，我们取 $\phi(s) = s$ 和 $\mathcal{E}_t(h_t) = y_t$。于是，确定性等价策略可表示为

$$\mu_t^{\mathcal{E}}(h_t) = \pi_t^{\mathcal{M}}(y_t).$$

3) 次优性界：我们假设条件 1 和2 均得到满足。此外，

$$\begin{aligned}\mathbb{E}[d_{\mathcal{S}}(S_t, \mathcal{E}_t(H_t))|h_t] &= \mathbb{E}\big[d_{\mathcal{S}}(S_t, Y_t)|h_t\big] \\ &\le (1-p)r + pR\end{aligned} \tag{19}$$

因此，我们有 $\eta_t \le (1-p)r + pR$，且假设3 也成立。此外，定理 ε_t 和 δ_t 中的1可由以下表达式上界：

$$\varepsilon_t \le [(1-p)r + pR]L_t^c \quad \text{and} \quad \delta_t \le [(1-p)r + pR](1 + L_t^P).$$

因此，定理1 中的界可明确写为

$$W_t^{\mathcal{P},\mu^{\mathcal{E}}}(h_t) - W_t^{\mathcal{P}}(h_t) \le 2[(1-p)r + pR]L_t^{\mathcal{M}} \tag{20}$$

其中 $L_t^{\mathcal{M}}$ 定义见式（18）。

这些结果表明，当状态估计误差 $[(1-p)r + pR]$ 较小时，无论这是由于观测噪声较小，还是由于观测频繁地准确，确定性等价策略都能取得接近最优的性能。这些界根据估计误差和模型的利普希茨常数，对次优性进行了定量刻画。

C. 有界观测噪声下的状态量化

1) 系统模型：考虑第 III-A 节中的有界观测噪声模型，其中 $\mathcal{Y} = \mathcal{S}$ 且观测模型满足 $d_{\mathcal{S}}(Y_t, S_t) \le r$。此外，给定一个抽象模型 $\hat{\mathcal{M}}$，它通过将状态空间$\mathcal{S}$ 划分为有限个 $\{\Psi_k\}_{k=1}^K$ 量化单元构造而成，每个量化单元都有一个代表元素 $\tilde{s}_k \in \Psi_k$。抽象状态空间为 $\tilde{\mathcal{S}} = \{\tilde{s}_1, \dots, \tilde{s}_K\}$，状态抽象函数为 $\phi : \mathcal{S} \to \tilde{\mathcal{S}}$ 由 $\phi(s) = \tilde{s}_k$ 给出，对所有 $s \in \Psi_k$, $k \in \{1, \dots, K\}$均成立。此外，随机核 $\lambda^P(\cdot \mid \tilde{s}_k)$ 和$\lambda^c(\cdot \mid \tilde{s}_k)$ 是定义在 $\tilde{s}_k$, $k \in \{1, \dots, K\}$上的狄拉克 δ 测度。抽象动力学 $\tilde{P}$ 和抽象代价 $\tilde{c}$按照式（9）和式（10）构造。我们还在 $\tilde{\mathcal{S}}$ 上定义度量 $d_{\tilde{\mathcal{S}}}$ 如下：

$$d_{\tilde{\mathcal{S}}}(\tilde{s}_1, \tilde{s}_2) := d_{\mathcal{S}}(\tilde{s}_1, \tilde{s}_2), \quad \forall \tilde{s}_1, \tilde{s}_2 \in \tilde{\mathcal{S}} \subset \mathcal{S},$$

我们假设 4 和 5 均得到满足 d.

2）确定性等价策略：在本例中，我们考虑相对于抽象马尔可夫决策过程 $\mathcal{M}$的确定性等价策略。此外，我们将抽象状态估计取为最近一次观测的量化值，即 $\mathcal{E}_t(h_t) = \phi(y_t)$。于是，确定性等价策略为

$$\mu_t^{\mathcal{E}}(h_t) = \pi_t^{\tilde{M}}\big(\phi(y_t)\big).$$

3）次优性界：我们假设假设 4 和5均得到满足。令 R 表示量化单元的最大半径，即

$$R := \max_{k \in \{1,\dots,K\}} \sup_{s \in \Psi_k} d_{\mathcal{S}}(s, \phi(s)).$$

于是，根据三角不等式，我们有

$$\begin{aligned}
&\mathbb{E}[d_{\tilde{\mathcal{S}}}(\phi(S_t), \mathcal{E}_t(h_t)) | h_t] = \mathbb{E}[d_{\mathcal{S}}(\phi(S_t), \phi(Y_t) | h_t] \\
&\le \mathbb{E}\big[d_{\mathcal{S}}\big(\phi(S_t), S_t\big) + d_{\mathcal{S}}(S_t, Y_t) + d_{\mathcal{S}}\big(Y_t, \phi(Y_t)\big) | h_t\big] \\
&\le R + r + R = r + 2R.
\end{aligned} \tag{21}$$

因此，我们有 $\eta_t \le r + 2R$，且假设 6 也得到满足。此外，定理 ε_t 和 δ_t 中的 2可上界为

$$\varepsilon_t \le F_t^c(r + 2R) \quad \text{and} \quad \delta_t \le F_t^P(r + 2R) + r + 2R.$$

因此，我们可以用定理2中的界（15）来给出采用确定性等价策略时次优性的上界。

D. 学习/自适应控制中的确定性等价

1）系统模型：考虑一个参数化的马尔可夫决策过程 $\mathcal{M}_X(\theta)$，其状态空间为 $\mathcal{X}$，动作空间为 $\mathcal{A}$，具有时不变动力学$P_{X,\theta}$ 以及时不变的单步成本 ℓ_θ，其中参数$\theta \in \Theta$和根据某个概率分布 P_Θ 分布，且该分布独立于动力学中的噪声。我们假设 $\mathcal{X}$ 和 Θ 是度量空间，其度量分别为 $d_{\mathcal{X}}$ 和d_Θ。

控制器不知道参数 θ，但知道状态与动作历史 $h_t = (x_{1:t}, a_{1:t-1})$。上述模型可视为一个状态空间为$\mathcal{S} = \mathcal{X} \times \Theta$、观测空间为 $\mathcal{X} \times \mathbb{R}$ 的部分可观测马尔可夫决策过程，其中 $S_t = (X_t, \theta)$，且$Y_t = (X_t, \ell(X_{t-1}, A_{t-1}))$。我们取 $d_{\mathcal{S}}((x_1, \theta_1), (x_2, \theta_2)) = d_{\mathcal{X}}(x_1, x_2) + d_\Theta(\theta_1, \theta_2)$。注意，对应于该部分可观测马尔可夫决策过程的马尔可夫决策过程 $\mathcal{M}$ 等价于 $\mathcal{M}_X(\theta)$。我们用 $\pi^{\mathcal{M}}(x, \theta) = \pi^{\mathcal{M}_X(\theta)}(x)$ 表示该马尔可夫决策过程的最优策略。

我们假设对所有 $\theta \in \Theta$，模型 $\mathcal{M}_X(\theta)$ 满足可测选择条件（假设 1）。此外，存在非负有限常数 L^P 和 L^c，使得对于任意x, $x' \in \mathcal{X}$和 θ, $\theta' \in \Theta$，均有

$$\begin{aligned}
&d_{\mathrm{Was}}(P_{X,\theta}(\cdot|x,a), P_{X,\theta'}(\cdot|x',a)) \\
&\qquad \le L^P(d_{\mathcal{X}}(x,x') + d_\Theta(\theta,\theta')),
\end{aligned} \tag{22}$$

and

$$\big|\ell_\theta(x,a) - \ell_{\theta'}(x',a)\big| \le L^c(d_{\mathcal{X}}(x,x') + d_\Theta(\theta,\theta')). \tag{23}$$

2）确定性等价策略：与前面的例子一样，我们考虑 $\phi(s) = s$，但使用一般估计器 $\hat{\mathcal{E}}_t(h_t) = (x_t, \theta_t)$，其中 θ_t 是基于 h_t对 θ 的某个估计值，例如，最小均方误差估计器 $\theta_t = \mathrm{E}[\theta | h_t]$。于是，确定性等价策略为

$$\mu_t^{\mathcal{E}}(h_t) = \pi_t^{\mathcal{M}}(x_t, \hat{\theta}_t).$$

3）次优性界：我们假设条件1 和 2 均成立。此外，

$$\eta_t = \sup_{h_t \in \mathcal{H}_t} \mathbb{E}[d_{\mathcal{S}}(S_t, \mathcal{E}_t(H_t)) | h_t] = \sup_{h_t \in \mathcal{H}_t} \mathbb{E}[d_\Theta(\theta_t, \hat{\theta}_t) | h_t]. \tag{24}$$

因此，若假设3 成立，则定理 ε_t 和 δ_t 中的量可以用以下上界来界定：1

$$\varepsilon_t \le L^c \eta_t \quad \text{and} \quad \delta_t \le L^P \eta_t + \eta_{t+1}.$$

因此，根据定理7中的界（1），我们可以界定采用确定性等价策略时的次优性。

这些结果表明，确定性等价策略的性能取决于 η_t 参数估计的性能。如果 η_t 衰减得足够快，例如呈指数速度衰减，那么即使 $T \to \infty$，我们也能对（ $2\alpha_t$ 7）中的性能误差给出一致上界。

E. 事件触发通信控制

1) 系统模型：考虑一个由被控对象、与被控对象位于同一位置的传感器以及远程控制器组成的系统。令$X_t \in \mathcal{X}$ 和 $A_t \in \mathcal{A}$ 分别表示被控对象的状态和控制输入。被控对象的状态按照受控转移核 $P_{X,t}$ 演化： $\mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})$。

传感器观测当前状态 X_t，并决定是否传输该状态。令 $Y_t \in \mathcal{X} \cup \{\mathfrak{E}\}$ 表示远程控制器的观测，即：

$$Y_t = \begin{cases} \mathfrak{E} & \text{if sensor does not communicate} \\ X_t & \text{if sensor communicates} \end{cases}$$

其中，$\mathfrak{E}$ 表示空观测。

远程控制器根据一般的依赖历史的策略 $\mu = (\mu_1, \ldots, \mu_T)$ 生成动作 A_t，并产生单步成本 $c_t(x_t, a_t)$。

上述问题是一个分散式控制问题，其中必须确定传感器处的通信策略以及远程控制器处的控制策略。我们考虑通信策略固定为事件触发的通信策略 [34]——[36],其运行方式如下。假设远程控制器按照以下方式维护状态估计 $\hat{X}_{t|t} \in \mathcal{X}$：

$$\hat{X}_{t|t} = \begin{cases} Y_t & \text{if } Y_t \neq \mathfrak{E} \\ \hat{X}_{t|t-1}, & \text{if } Y_t = \mathfrak{E} \end{cases} \tag{25}$$

其中，$\hat{X}_{1|0} = \mathbb{E}[X_1]$ 以及

$$\hat{X}_{t|t-1} = g(\hat{X}_{t-1|t-1}, A_{t-1}), \quad t > 1, \tag{26}$$

其中，$g: \mathcal{X} \times \mathcal{A} \to \mathcal{X}$ 是预先指定的更新函数。

在事件触发策略中，当满足以下不等式时，传感器进行传输

$$d_{\mathcal{X}}(X_t, \hat{X}_{t|t-1}) > r.$$

其中 r 是预先指定的常数。该策略确保 $d_{\mathcal{X}}(X_t, \hat{X}_{t|t}) \leq r$。因此，目标是在远程控制器处找到最佳控制策略。

一旦确定事件触发策略，上述模型就对应于一个部分可观测马尔可夫决策过程 $\mathcal{P}$，其状态为 $S_t = (X_t, X_{t|t-1})$，观测为 Y_t，动作为 A_t。令 $\mathcal{M}$ 表示相应的马尔可夫决策过程，在该过程中控制器可以访问 S_t。

我们考虑一般的 Borel 空间 $\mathcal{X}$，其状态空间为 $\mathcal{X}$，该空间由状态抽象函数 $\phi(x, \hat{x}) = x$ 构造；随机核 $\lambda^P(\cdot|x)$ 和 $\lambda^c(\cdot|x)$ 是定义在 (x, x) 上的狄拉克 δ 测度。于是，抽象马尔可夫决策过程的动力学 P_t 等于 $P_{X,t}$，单步成本 $\tilde{c}_t$ 等于 c_t。因此，马尔可夫决策过程 $\tilde{M} = \langle \mathcal{X}, \mathcal{A}, \{P_{X,t}\}_{t=1}^{T-1}, \{c_t\}_{t=1}^{T} \rangle$。我们假设 $\mathcal{M}$ 满足假设4，并且存在凹且非递减的函数 $F_t^P, F_t^c: \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$，$t \in \{1, \ldots, T\}$，使得对于任意 $x, x' \in \mathcal{X}$ 和 $a \in \mathcal{A}$，均有

$$d_{\text{Was}}(P_{X,t}(\cdot|x, a), P_{X,t}(\cdot|x', a)) \leq F_t^P(d_{\mathcal{X}}(x, x')) \tag{27}$$

and

$$\left|c_t(x, a) - c_t(x', a)\right| \leq F_t^c(d_{\mathcal{X}}(x, x')). \tag{28}$$

假设4 意味着对于马尔可夫决策过程 $\tilde{\mathcal{M}}$，存在一个最优策略，我们将其记为 $\pi^{\tilde{\mathcal{M}}}$。此外，可以验证（27）和（28）蕴含假设5。

2）确定性等价策略：在此示例中，我们考虑相对于抽象马尔可夫决策过程 $\mathcal{M}$ 的确定性等价策略。此外，我们将状态估计设为 $\mathcal{E}_t(h_t) = \hat{x}_{t|t}$，它根据历史使用（25）递归计算。于是，确定性等价策略定义如下：

$$\mu_t^{\mathcal{E}}(h_t) = \pi_t^{\tilde{\mathcal{M}}}(\hat{x}_{t|t}).$$

3）次优性界：我们已假设假设4并假设了足以推出假设5的充分条件。回顾一下，事件触发传感器传输策略确保 $d_{\mathcal{X}}(X_t, \hat{X}_{t|t}) \leq r$。因此，

$$\mathbb{E}[d_{\mathcal{X}}(\phi(S_t), \mathcal{E}(H_t))|h_t] = \mathbb{E}[d_{\mathcal{X}}(X_t, \hat{X}_{t|t})|h_t] \leq r.$$

因此，$\eta_t \leq r$ 且假设6 成立。此外，ε_t 和 δ_t 在定理2中可由上界界定为

$$\varepsilon_t \leq F_t^c(r) \quad \text{and} \quad \delta_t \leq F_t^P(r) + r.$$

因此，我们可以用界（15）对使用确定性等价策略的次优性给出上界，该界见定理2。这些界量化了带事件触发感知的确定性等价控制的次优性差距。

F. 非齐次多粒子系统的控制

1）系统模型：考虑一个由 n 个粒子组成的系统，其中每个粒子 i，$i \in \{1, \ldots, n\}$的状态为 $X_t^i \in \mathbb{R}$。系统的全局状态为 $X_t = (X_t^1, X_t^2, \ldots\ldots, X_t^n)^\intercal \in \mathbb{R}^n$。观测由下式给出

$$Y_t = X_t + N_t$$

其中 $N_t = (N_t^1, N_t^2, \ldots\ldots, N_t^n)^\intercal$ 是观测噪声，其中 $N_t^i \in [-r^i, r^i]$ 对某个 $r^i \geq 0$成立。

令 $M_t = \sum_{i=1}^n \alpha^i X_t^i$ 表示全局状态的加权均值，其中 α^i、$i \in \{1, \ldots, n\}$是非负权重，且权重之和为 1。我们假设每个粒子的动力学由下式给出：

$$X_{t+1}^i = \bar{f}(M_t, A_t, W_t) + f^i(X_t, A_t, W_t) \tag{29}$$

其中 $A_t \in \mathcal{A}$ 是时刻 t 的控制动作，且 $\{W_t\}_{t \geq 1}$、$W_t \in \mathcal{W}$是分布为 P_W的独立同分布过程。

假设7 我们作如下假设。

1) 函数 $\bar{f}$关于第一个参数是 $L^{\bar{f}}$-利普希茨连续的，即对所有 m，$m' \in \mathbb{R}$、$a \in \mathcal{A}$和 $w \in \mathcal{W}$，

$$|\bar{f}(m, a, w) - \bar{f}(m', a, w)| \leq L^{\bar{f}}|m - m'| \tag{30}$$

2) 存在常数 γ^i，使得 $\|f^i\|_\infty \leq \gamma^i$，对于 $i \in \{1, \ldots, n\}$。■

单步成本定义为

$$c(X_t, A_t) = \bar{\ell}(M_t, A_t) + \ell(X_t, A_t). \tag{31}$$

假设 8 我们作如下假设。

1) 函数 $\bar{\ell}$ 关于第一个自变量是 $L^{\bar{\ell}}$-利普希茨连续的，即对于所有 $m, m' \in \mathbb{R}$ 和 $a \in \mathcal{A}$，

$$|\bar{\ell}(m, a) - \bar{\ell}(m', a)| \leq L^{\bar{\ell}}|m - m'|. \tag{32}$$

2) 存在常数 β，使得 $\|\ell\|_\infty \leq \beta$。 ■

上述模型是一个部分可观测马尔可夫决策过程 $\mathcal{P}$，具有状态 $S_t = X_t$、观测 Y_t、动作 A_t 以及单步成本 $c(X_t, A_t)$。

2) 确定性等价策略：我们考虑一个关注全局状态加权均值的抽象马尔可夫决策过程。˜抽象状态空间为 $\mathcal{S} = \mathrm{R}$，状态抽象函数 ϕ: $\mathrm{R}^n \to \mathrm{R}$ 由下式给出

$$\phi((x^1, x^2, \ldots, x^n)^\intercal) = \sum_{i=1}^{n} \alpha^i x^i$$

with

$$\phi^{-1}(m) = \Big\{(x^1, x^2, \ldots, x^n)^\intercal : \sum_{i=1}^{n} \alpha^i x^i = m\Big\}.$$

我们假设抽象状态空间上的度量为$d_{\tilde{\mathcal{S}}}(m,\ m') = |m - m'|$，并令 $\lambda^P(\cdot|m)$和 $\lambda^c(\cdot|m)$为在 $m\mathbf{1}_n$处的狄拉克分布，其中 $\mathbf{1}_n$ 为全 1 的 n 维向量。

因此，抽象单步成本为 b y

$$\tilde{c}(\tilde{s}, a) = c(\tilde{s}\mathbf{1}_n, a) = \bar{\ell}(\tilde{s}, a) + \ell(\tilde{s}\mathbf{1}_n),$$

抽象状态动态为

$$\tilde{S}_{t+1} = \bar{f}(\tilde{S}_t, A_t, W_t) + \sum_{i=1}^{n} \alpha^i f^i(\tilde{S}_t\mathbf{1}_n, A_t, W_t).$$

˜上述模型对应于一个抽象马尔可夫决策过程 $\mathcal{M}$。我们˜假设 $\mathcal{M}$ 满足假设4。

我们将最近一次观测的加权均值作为抽象状态的估计值，即

$$\mathcal{E}_t(h_t) = \sum_{i=1}^{n} \alpha^i y_t^i.$$

因此，确定性等价策略为

$$\mu_t^{\mathcal{E}}(h_t) = \pi_t^{\mathcal{M}}(\mathcal{E}(h_t)).$$

3) 次优性界：我们假设抽-˜象模型 $\mathcal{M}$ 满足假设 4。现在我们证明假设 5 得到满足。

引理 1 假设 *7* 和*8* 意味着假设*5* 得到满足，其中 F_t^c和 F_t^P定义如下

$$F_t^c(r) = L^{\bar{\ell}}r + 2\beta \quad \textit{and} \quad F_t^P(r) = L^{\bar{f}}r + 2\sum_{i=1}^{n} \alpha^i\gamma^i, \quad r \in \mathbb{R}.$$

证明见附录A。

此外，我们有 $\eta_t \le \sum_{i=1}^n \alpha^i r^i$，因此假设6 得到满足。因此，我们可以在定理 ε_t 和δ_t 2中界定如下

$$\varepsilon_t \le L^{\bar{\ell}}\bar{r} + 2\beta \quad \text{and} \quad \delta_t \le L^{\bar{f}}\bar{r} + 2\bar{\gamma}$$

其中 $\bar{r} = \sum_{i=1}^n \alpha^i r^i$ 和 $\bar{\gamma} = \sum_{i=1}^n \alpha^i\gamma^i$。

因此，我们可以利用定理2中的界（15）对采用确定性等价策略的次优性给出上界，该界可简化如下：

$$W_t^{\mathcal{P},\mu^{\mathcal{E}}}(h_t) - W_t^{\mathcal{P}}(h_t) \le 2\Big[(T-t+1)(L^{\bar{c}}\bar{r} + 2\beta) + (L^{\bar{f}}\bar{r} + 2\bar{\gamma})\sum_{\tau=t}^{T-1} \mathrm{Lip}(V_{\tau+1}^{\tilde{\mathcal{M}}})\Big]. \tag{33}$$

备注 4 这类模型可能出现在每个粒子都配有一个局部控制器，且该局部控制器确保 $\|f^i\|_\infty \le \gamma^i$ 的情形中。此外，注意，界（33）取决于 $\bar{r}$，即使某些 $\{r^i\}_{i=1}^n$ 很大，它也可能很小。

IV. 分析

本节将分析我们的主要结果。如前所述，定理1 是定理2˜在取 $\mathcal{S} = \mathcal{S}$ 和 $\phi(s) = s$时的一个特例。因此，我们仅给出定理2的证明。

我们先介绍一些关于策略无关信念和AIS理论的背景，然后给出关键引理及其证明。

A. 策略无关信念

考虑模型 μ 上的任意依赖历史的策略 $\mathcal{P}$，该模型定义于第II节。我们定义以下两种信念，它们在部分可观测马尔可夫决策过程中很常用：

- $b_{t|t}(\cdot|h_t)$表示控制器在策略 S_t 下给定历史 h_t 时对当前状态的后验分布μ，即对于 M_S 的任意博雷尔子集 $\mathcal{S}$，$b_{t|t}(M_S|h_t) = \mathbb{P}^\mu(S_t \in M_S|h_t)$。信念 $b_{t|t}(\cdot|h_t)$ 称为信念状态。众所周知，它不依赖于所选的依赖历史的策略 μ [2]，[3]。
- $b_{t+1|t}(\cdot,\ \cdot|h_t,\ a_t)$表示控制器在策略 μ下，给定历史 h_t 和动作 a_t 时，对下一状态 S_{t+1} 和下一观测 Y_{t+1}的后验分布。注意，对于 $\mathcal{S}$ 和 $\mathcal{Y}$的任意博雷尔子集 M_S 和 M_Y，我们有

$$b_{t+1|t}(M_S, M_Y|h_t, a_t) = \int_{\mathcal{S}} P_t(M_S, M_Y|s_t, a_t) b_{t|t}(ds_t|h_t).$$

由于信念状态 $b_{t|t}(\cdot|h_t)$不依赖于策略 μ的选择，由上述关系可知，对于 $b_{t+1|t}(\cdot,\ \cdot|h_t,\ a_t)$也同样如此。稍微滥用一下记号，我们仍将 $b_{t+1|t}$ 用于表示其在 $\mathcal{S}$ 或 $\mathcal{Y}$上的边缘分布。

B. 近似信息状态

AIS理论 [11] 提供了一个框架，可为一类部分可观测马尔可夫决策过程的近似解推导次优性界。该框架的核心思想是近似信息状态这一概念，我们将在下文给出其正式定义。我们的定义与 [11] 中的定义相似，但有两点不同。首先，[11]中的分析是在状态空间与观测空间有限的假设下进行的，而我们处理的是博雷尔空间。因此，我们加入一个可测选择假设，以确保由近似信息状态得到的近似动态规划具有良定义的解。其次，[11] 中的分析使用了通用的积分概率度量（IPMs）[37]。我们将讨论范围限定为一种特定的积分概率度量（Wasserstein-1 距离），因为我们的结果采用的正是这种形式。

以下讨论针对一般的部分可观测马尔可夫决策过程模型 $\mathcal{P}$，该模型定义于第II节。

定义2 给定序列 $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_T)$ 和 $\delta = (\delta_1, \ldots, \delta_{T-1}) \in \mathrm{R}_{\geq 0}^{T}$，过程 $\{Z_t\}_{t\geq 1}$、$Z_t \in \mathcal{Z}$称为 (ε, δ)-近似信息状态（*AIS*），如果存在

- 一系列历史压缩函数 $\{\sigma_{\text{近似信息状态}_t}\}_{t=1}^{T}$，其中 $\sigma_{\text{近似信息状态}_t}$：$\mathcal{H}_t \to \mathcal{Z}$ 满足 $Z_t = \sigma_{\text{近似信息状态}_t}(H_t)$
- 一系列成本近似器 $\{c_{\text{近似信息状态}_t}\}_{t=1}^{T}$，其中$c_{\text{近似信息状态}_t}$：$\mathcal{Z}\times\mathcal{A}\to\mathrm{R}$
- 一系列动力学近似器 $\{P_{\text{近似信息状态}_t}\}_{t=1}^{T-1}$，其中 $P_{\text{近似信息状态}_t}$：$\mathcal{Z}\times\mathcal{A}\to\Delta(\mathcal{Z})$

使得以下三个性质得到满足：

(AP1) 对性能评估近似充分：对于任意时刻 $t \in \{1,\ldots,T\}$ 以及任意 $h_t \in \mathcal{H}_t$和$a_t \in \mathcal{A}$，有

$$\left|\mathbb{E}[c_t(S_t,a_t)|h_t] - c_t^{\mathsf{AIS}}(\sigma_t^{\mathsf{AIS}}(h_t),a_t)\right| \leq \varepsilon_t$$

(AP2) 对预测自身近似充分：对于任意时刻 $t\in\{1,\ldots,T-1\}$ 以及任意 $h_t\in\mathcal{H}_t$和 $a_t\in\mathcal{A}$，定义 ν_t 在 $\mathcal{H}_t\times\mathcal{A}_t\to\Delta(\mathcal{Z})$ 上的随机核如下：对于 $\mathcal{Z}$ 的任意 Borel可测子集 M_Z，

$$\begin{aligned}\nu_t(M_Z|h_t,a_t) &= \mathbb{P}(Z_{t+1}\in M_Z|h_t,a_t)\\ &= \int_{\mathcal{Y}} \mathbb{1}\{\sigma_{t+1}^{\mathsf{AIS}}(h_t,a_t,y_{t+1})\in M_Z\} b_{t+1|t}(dy_{t+1}|h_t,a_t).\end{aligned}$$

那么，对于任意时刻 $t\in\{1,\ldots,T-1\}$，我们有

$$d_{\mathrm{Was}}\big(\nu_t(\cdot|h_t,a_t), P_t^{\mathsf{AIS}}(\cdot|\sigma_t^{\mathsf{AIS}}(h_t),a_t)\big) \leq \delta_t.$$

(M)可测选择：MDP $\mathcal{M}^{\mathsf{AIS}} := \langle \mathcal{Z}, \mathcal{A}, \{P_{\text{近似信息状态}_t}\}_{t=1}^{T-1}, \{c_t^{\mathsf{AIS}}\}_{t=1}^{T}, T\rangle$ 满足可测选择条件。

称元组 $(\sigma^{\mathsf{AIS}}, c^{\mathsf{AIS}}, P^{\mathsf{AIS}})$ 为一个*AIS*生成器，其中每个分量都是一个序列。

我们可以为 $\mathcal{M}^{\mathsf{AIS}}$ 写出一个动态规划，其中价值函数 $\{V_{\text{近似信息状态}_t}\}_{t=1}^{T+1}$, $V_{\text{近似信息状态}_t}$：$\mathcal{Z}\to\mathrm{R}$，定义如下。我们将 $V_{\text{近似信息状态}_{T+1}}(z) = 0$ 初始化为对所有 $z\in\mathcal{Z}$ 都成立，然后递归定义 $t\in\{T,T-1,\ldots,1\}$

$$V_t^{\mathsf{AIS}}(z_t) = \min_{a\in\mathcal{A}}\left\{c_t^{\mathsf{AIS}}(z_t,a) + \int_{\mathcal{Z}} V_{t+1}^{\mathsf{AIS}}(z') P_t^{\mathsf{AIS}}(dz'|z_t,a)\right\}. \tag{34}$$

可测选择条件 (M) 意味着存在一个可测选择器 π_t^{AIS}：$\mathcal{Z}\to\mathcal{A}$，$t\in\{1,\ldots,T\}$使得 $\pi_t^{\mathsf{AIS}}(z_t)$ 是右侧表达式 (34) 的一个最小化自变量，并且函数 V_t^{AIS}是可测的。根据马尔可夫决策过程理论中的标准结果 [31],，我们知道策略 $\pi^{\mathsf{AIS}} = (\pi_1^{\mathsf{AIS}},\ldots,\pi_T^{\mathsf{AIS}})$是 $\mathcal{M}^{\mathsf{AIS}}$的一个最优策略。

AIS理论的主要结果如下：

定理3 为部分可观测马尔可夫决策过程 $\mathcal{P}$定义如下依赖历史的策略 $\mu^{\mathsf{AIS}} = (\mu_1^{\mathsf{AIS}},\ldots,\mu_T^{\mathsf{AIS}})$：对于任意 $t\in\{1,\ldots,T\}$ 和任意 $h_t\in\mathcal{H}_t$，定义

$$\mu_t^{\mathsf{AIS}}(h_t) = \pi^{\mathsf{AIS}}(\sigma_t^{\mathsf{AIS}}(h_t)).$$

那么，对于任意 $t\in\{1,\ldots,T\}$ 和 $h_t\in\mathcal{H}_t$，有

$$W_t^{\mathcal{P},\mu^{\mathsf{AIS}}}(h_t) - W_t^{\mathcal{P}}(h_t) \leq 2\alpha_t \tag{35}$$

其中

$$\alpha_t = \varepsilon_t + \sum_{\tau=t}^{T-1}\big[\delta_\tau \operatorname{Lip}(V_{\tau+1}^{\mathsf{AIS}}) + \varepsilon_{\tau+1}\big].$$

证明 该结果与 [11, 定理 9],相同；该定理是在假设 $\mathcal{S}$ 和 $\mathcal{A}$ 为有限集的条件下提出的，而我们研究的是博雷尔空间。正如前文所述，可测选择条件确保 V_t^{AIS}和 π_t^{AIS}定义良好且可测。在此假设下，近似界可通过与 [11, 定理 9]中完全相同的分析得出。■

C. 关键引理

我们的次优性界的核心思想是证明：抽象马尔可夫决策过程 $\mathcal{M}$ 的抽象状态估计函数 $\mathcal{E}_t$、逐步代价 $\tilde{c}_t$ （定义见（10））以及动力学 P_t （定义见（9））共同构成一个AIS生成器，其中选择适当的 ε 和 δ。 $(\mathcal{E},\tilde{c},P)$。

我们首先证明 $\mathcal{E},\tilde{c}$ 满足AIS的条件（AP1）。

引理2 在假设*5* 和*6*下，对于任意 $h_t\in\mathcal{H}_t$ 和$a_t\in\mathcal{A}$，我们有

$$\left|\mathbb{E}[c_t(S_t,a_t)|h_t] - \tilde{c}_t(\mathcal{E}_t(h_t),a_t)\right| \leq F_t^c(\eta_t).$$

证明

$$\begin{aligned}&\left|\mathbb{E}[c_t(S_t,a_t)|h_t] - \tilde{c}_t(\mathcal{E}_t(h_t),a_t)\right|\\ &\quad\leq \mathbb{E}\big[\left|c_t(S_t,a_t) - \tilde{c}_t(\mathcal{E}_t(h_t),a_t)\right| \mid h_t\big]\end{aligned} \tag{36}$$

现在，对于一个固定的实现 s_t，我们考虑内层项，

$$\begin{aligned}&|c_t(s_t,a_t) - \tilde{c}_t(\mathcal{E}_t(h_t),a_t)|\\ &\overset{(a)}{=} \left|\int_{\phi^{-1}(\mathcal{E}_t(h_t))}\big[c_t(s_t,a_t) - c_t(s',a_t)\big]\lambda_t^c\big(ds'|\mathcal{E}_t(h_t)\big)\right|\\ &\overset{(b)}{\leq} \int_{\phi^{-1}(\mathcal{E}_t(h_t))}\left|c_t(s_t,a_t) - c_t(s',a_t)\right|\lambda_t^c\big(ds'|\mathcal{E}_t(h_t)\big)\\ &\overset{(c)}{\leq} \int_{\phi^{-1}(\mathcal{E}_t(h_t))} F_t^c(d_{\tilde{\mathcal{S}}}(\phi(s_t),\phi(s')))\lambda_t^c\big(ds'|\mathcal{E}_t(h_t)\big)\\ &\overset{(d)}{=} F_t^c(d_{\tilde{\mathcal{S}}}(\phi(s_t),\mathcal{E}_t(h_t))).\end{aligned} \tag{37}$$

其中，(*a*)由 $\tilde{c}_t$的定义得出，(*b*)由詹森不等式得出，(*c*)由假设5得出，而(*d*)则由以下事实得出：对于任意 $s'\in\phi^{-1}(\mathcal{E}_t(h_t))$，$\phi(s') = \mathcal{E}_t(h_t)$，并且 $\lambda_t^c(\phi^{-1}(\mathcal{E}_t(h_t))|\mathcal{E}_t(h_t)) = 1$。将(37)代入(36)，我们得到

$$\begin{aligned}&\left|\mathbb{E}[c_t(S_t,a_t)|h_t] - \tilde{c}_t(\mathcal{E}_t(h_t),a_t)\right|\\ &\quad\leq \mathbb{E}[F_t^c(d_{\tilde{\mathcal{S}}}(\phi(S_t),\mathcal{E}_t(h_t)))|h_t]\\ &\quad\overset{(e)}{\leq} F_t^c\big(\mathbb{E}[d_{\tilde{\mathcal{S}}}(\phi(S_t),\mathcal{E}_t(h_t)) \mid h_t]\big)\\ &\quad\overset{(f)}{\leq} F_t^c(\eta_t)\end{aligned} \tag{38}$$

其中，(*e*)由詹森不等式和 F_t^c的凹性得出，而(*f*)由 η_t的定义得出。■

为了证明 $\mathcal{E}$，P 满足近似信息状态的条件(AP2)，我们将使用以下中间引理。

引理3 在假设 *5*下，对于任意 $s_t\in\mathcal{S}$，$\hat{s}_t\in\mathcal{S}$ 和$a_t\in\mathcal{A}$，我们有

$$d_{\mathrm{Was}}\big(P_{S,t}^{\phi}(\cdot|s_t,a_t), \tilde{P}_t(\cdot|\hat{s}_t,a_t)\big) \leq F_t^P\Big(d_{\tilde{\mathcal{S}}}\big(\phi(s_t),\hat{s}_t\big)\Big).$$

证明

$$
\begin{aligned}
&d_{\mathrm{Was}}\big(P^{\phi}_{S,t}(\cdot|s_t,a_t),\tilde{P}_t(\cdot|\hat{s}_t,a_t)\big)\\
&=d_{\mathrm{Was}}\big(P^{\phi}_{S,t}(\cdot|s_t,a_t),\int_{\phi^{-1}(\hat{s}_t)}P^{\phi}_{S,t}(\cdot|s',a_t)\lambda^P(ds'|\hat{s}_t)\big)\\
&\overset{(a)}{\leq}\int_{\phi^{-1}(\hat{s}_t)}d_{\mathrm{Was}}\Big(P^{\phi}_{S,t}(\cdot|s_t,a_t),P^{\phi}_{S,t}(\cdot|s',a_t)\Big)\lambda^P(ds'|\hat{s}_t)\\
&\overset{(b)}{\leq}\int_{\phi^{-1}(\hat{s}_t)}F^P_t\Big(d_{\tilde{\mathcal{S}}}\big(\phi(s_t),\phi(s')\big)\Big)\lambda^P(ds'|\hat{s}_t)\\
&=F^P_t\Big(d_{\tilde{\mathcal{S}}}\big(\phi(s_t),\hat{s}_t\big)\Big),
\end{aligned}
\tag{39}
$$

其中(a) 由Wasserstein距离的凸性可得 [38, 定理4.8] 以及(b) 由假设5可得。 ■

接下来，我们定义一个随机核 ψ_t: $\mathcal{H}_t\times\mathcal{A}_t\to\Delta(\mathcal{S})$, 它类似于(AP2)中定义的 ν_t 。对于任意 $h_t\in\mathcal{H}_t$和 $a_t\in\mathcal{A}$ 以及 $\mathcal{S}$的Borel可测子集 $M_{\tilde{S}}$,

$$
\begin{aligned}
&\hat{\psi}_t(M_{\tilde{S}}|h_t,a_t)=\mathbb{P}(\mathcal{E}_{t+1}(H_{t+1})\in M_{\tilde{S}}|h_t,a_t)\\
&=\int_{\mathcal{Y}}\mathbb{1}\{\mathcal{E}_{t+1}(h_t,a_t,y_{t+1})\in M_{\tilde{S}}\}b_{t+1|t}(dy_{t+1}|h_t,a_t)
\end{aligned}
\tag{40}
$$

这是 S_{t+1} $=\mathcal{E}_{t+1}(H_{t+1})$在给定 h_t, a_t条件下的条件概率分布。

我们还定义随机核 $\tilde{\psi}_t$: $\mathcal{H}_t\times\mathcal{A}_t\to\Delta(\tilde{\mathcal{S}})$, 它用于下一个引理的证明。对于任意 $h_t\in\mathcal{H}_t$和 $a_t\in\mathcal{A}$ 以及 $\mathcal{S}$的Borel可测子集 $M_{\tilde{S}}$,

$$
\begin{aligned}
&\tilde{\psi}_t(M_{\tilde{S}}|h_t,a_t)=\mathbb{P}(\tilde{S}_{t+1}\in M_{\tilde{S}}|h_t,a_t)\\
&=\int_{\mathcal{S}}\mathbb{1}\{\phi(s_{t+1})\in M_{\tilde{S}}\}b_{t+1|t}(ds_{t+1}|h_t,a_t)\\
&=\int_{\mathcal{S}}P^{\phi}_{S,t}(M_{\tilde{S}}|s_t,a_t)b_{t|t}(ds_t|h_t,a_t),
\end{aligned}
\tag{41}
$$

这是在给定 h_t, a_t的条件下， $\phi(S_{t+1})$的条件概率分布。

下述引理表明 $\mathcal{E}_t$, $\tilde{P}_t$ 满足（AP2）。

引理4 在假设 *5* 和 *6*下，对于任意 $h_t\in\mathcal{H}_t$ 和$a_t\in\mathcal{A}$，有

$$
d_{\mathrm{Was}}(\hat{\psi}_t(\cdot|h_t,a_t),\tilde{P}_t(\cdot|\mathcal{E}_t(h_t),a_t))\leq F^P_t(\eta_t)+\eta_{t+1},
$$

其中 $\hat{\psi}_t(\cdot|h_t,a_t)$ 是在 $\mathcal{S}$ 上定义的概率分布，见 (40)。

证明 令$\hat{s}_t=\mathcal{E}_t(h_t)$。根据三角不等式，有

$$
\begin{aligned}
&d_{\mathrm{Was}}(\hat{\psi}_t(\cdot|h_t,a_t),\tilde{P}_t(\cdot|\hat{s}_t,a_t))\\
&\quad\leq d_{\mathrm{Was}}(\hat{\psi}_t(\cdot|h_t,a_t),\tilde{\psi}_t(\cdot|h_t,a_t))\\
&\qquad+d_{\mathrm{Was}}(\tilde{\psi}_t(\cdot|h_t,a_t),\tilde{P}_t(\cdot|\hat{s}_t,a_t)).
\end{aligned}
\tag{42}
$$

现在我们分别考虑这两项。式（42）中的第一项可按如下方式进行界定。

$$
\begin{aligned}
&d_{\mathrm{Was}}(\hat{\psi}_t(\cdot|h_t,a_t),\tilde{\psi}_t(\cdot|h_t,a_t))\\
&\overset{(a)}{=}\inf_{(\hat{S},\tilde{S})\sim\Gamma(\hat{\psi}_t,\tilde{\psi}_t)}\mathbb{E}\big[d_{\tilde{\mathcal{S}}}(\hat{S},\tilde{S})\big]\\
&\overset{(b)}{\leq}\mathbb{E}[d_{\tilde{\mathcal{S}}}\big(\mathcal{E}_{t+1}(H_{t+1}),\phi(S_{t+1})\big)|h_t,a_t]\\
&=\mathbb{E}[\mathbb{E}[d_{\tilde{\mathcal{S}}}\big(\mathcal{E}_{t+1}(H_{t+1}),\phi(S_{t+1})\big)|H_{t+1}]|h_t,a_t]\\
&\leq\mathbb{E}[\eta_{t+1}|h_t,a_t]=\eta_{t+1}
\end{aligned}
\tag{43}
$$

其中，在（a）中， $\Gamma(\psi_t,\psi_t)$ 表示所有边缘分布为 $\psi_t(\cdot|h_t,a_t)$ 和 $\psi_t(\cdot|h_t,a_t)$ 的联合测度集合；在（b）中，我们利用了这样一个事实：给定 h_t,a_t 后， $\mathcal{E}_{t+1}(H_{t+1})$ 和 $\phi(S_{t+1})$ 的边缘分布分别为 $\psi_t(\cdot|h_t,a_t)$ 和 $\psi_t(\cdot|h_t,a_t)$; 因此，给定（h_t, a_t）后， （$\mathcal{E}_{t+1}(H_{t+1})$ $\phi(S_{t+1})$）上的联合分布属于 $\Gamma(\psi_t,\psi_t)$。

式（42）的第二项可按如下方式界定。

$$
\begin{aligned}
&d_{\mathrm{Was}}(\tilde{\psi}_t(\cdot|h_t,a_t),\tilde{P}_t(\cdot|\hat{s}_t,a_t))\\
&\quad=d_{\mathrm{Was}}\Big(\int_{\mathcal{S}}P^{\phi}_{S,t}(\cdot|s_t,a_t)b_{t|t}(ds_t|h_t),\tilde{P}_t(\cdot|\hat{s}_t,a_t)\Big)\\
&\quad\overset{(a)}{\leq}\int_{\mathcal{S}}d_{\mathrm{Was}}\big(P^{\phi}_{S,t}(\cdot|s_t,a_t),\tilde{P}_t(\cdot|\hat{s}_t,a_t)\big)b_{t|t}(ds_t|h_t)\\
&\quad\overset{(b)}{\leq}\int_{\mathcal{S}}F^P_t\big(d_{\tilde{\mathcal{S}}}(\phi(s_t),\hat{s}_t)\big)b_{t|t}(ds_t|h_t)\\
&\quad=\mathbb{E}[F^P_t\big(d_{\tilde{\mathcal{S}}}(\phi(S_t),\mathcal{E}_t(h_t))|h_t\big])\\
&\quad\overset{(c)}{\leq}F^P_t\big(\mathbb{E}\big[d_{\tilde{\mathcal{S}}}(\phi(S_t),\mathcal{E}_t(h_t))|h_t\big]\big)\\
&\quad\leq F^P_t(\eta_t).
\end{aligned}
\tag{44}
$$

其中(a) 由 Wasserstein 距离的凸性 [38,定理 4。8], 以及(b) 由引理3得出，而(c)则由詹森不等式以及 F^P_t. ■的凹性得出。

D. 定理*2*的证明

在假设5 和6下，引理2 和4保证近似信息状态满足条件（AP1）和（AP2），其中$\varepsilon_t=F^c_t(\eta_t)$ $\delta_t=F^P_t(\eta_t)$ $+\eta_{t+1}$。假设4 保证近似信息状态满足条件（M）。因此，根据定理3，即可得出该结果。

V. 结论

本文提出了部分可观测马尔可夫决策过程（POMDP）中控制策略的确定性等价原理的一种推广。我们的方法将完全可观测马尔可夫决策过程中的最优状态反馈策略应用于状态估计，而不局限于最小均方误差（MMSE）等特定类型的估计器。我们建立了理论性能界，以刻画其次优程度。具体而言，我们利用近似信息状态（AIS）框架 [11]量化估计误差对控制性能的影响，并推导出以系统动力学和单步成本函数的光滑性表示的界。

为说明我们结果的实际意义，我们考察了若干示例，表明当状态估计误差较小时，确定性等价策略能够达到接近最优的表现。这意味着，在精确最优策略难以通过计算求得的场景中，确定性等价策略可作为一种实用且高效的替代方案，充分利用现有状态估计值，在保持计算可处理性的同时实现可靠的决策。

参考文献

[1] B. Bozkurt、A. Mahajan、A. Nayyar 和 Y. Ouyang，《部分可观测系统中的广义确定性等价策略》，载于 IEEE决策与控制会议。IEEE，2025年12月。[2] K. J. ˚Astr¨om，《不完全状态信息下马尔可夫过程的最优控制》，数学分析与应用杂志，第 10 卷，第 1 期，第 174–205 页，1965年2月。[3] R. D. Smallwood 和 E. J. Sondik，《有限时域内部分可观测马尔可夫过程的最优控制》，《运筹学》，第 21 卷，第 5 期，第 1071–1088 页，1973年10月。[4] C. H. Papadimitriou 和 J. N. Tsitsiklis，《马尔可夫决策过程的复杂性》，《运筹学数学》，第 12 卷，第 3 期，第 441–450 页，1987年。[5] G. Shani、J. Pineau 和 R. Kaplow，《基于点的部分可观测马尔可夫决策过程求解器综述》，《自主智能体与多智能体系统》，第 27 卷，第 1 期，第 1–51 页，2013年。[6] D. Burago、M. De Rougemont 和 A. Slissenko，《部分可观测马尔可夫决策过程的复杂性》，《理论计算机科学》，第 157 卷，第 2 期，第 161–183 页，1996年。[7] C. Lusena、J. Goldsmith 和 M. Mundhenk，《部分可观测马尔可夫决策过程的不可近似性结果》，《人工智能研究杂志》，第 14 卷，第 83–103 页，2001年。[8] C. C. White III 和 W. T. Scherer，《部分可观测马尔可夫决策过程的有限记忆次优设计》，《运筹学》，第 42 卷，第 3 期，第 439–455 页，1994年。[9] B. Van Roy、Q. Dong 和 L. Tang，《简单智能体，复杂环境：利用智能体状态实现高效强化学习》，《机器学习研究杂志》，第 24 卷，第 170 期，第 1–54 页，2023年。[10] A. Sinha 和 A. Mahajan，《部分可观测马尔可夫决策过程中的基于智能体状态的策略：超越信念状态马尔可夫决策过程》，载于决策与控制会议，2024年12月，第 6722–6735 页。[11] J. Subramanian、A. Sinha、R. Seraj 和 A. Mahajan，《部分可观测系统中的近似规划与强化学习的近似信息状态》，《机器学习研究杂志》，第 23 卷，第 12 期，第 1–83 页，2022年。[12] C. McDonald 和 S. Y¨uksel，《部分观测随机控制中对错误先验的鲁棒性及受控滤波器稳定性》，SIAM控制与优化杂志，第 60 卷，第 2 期，第 842–870 页，2022年。[13] A. Kara 和 S. Yuksel，《部分可观测马尔可夫决策过程中有限记忆反馈策略的近似最优性》，《机器学习研究杂志》，第 23 卷，第 11 期，第 1–46 页，2022年。[14] N. Golowich、A. Moitra 和 D. Rohatgi，《通过滤波器稳定性在部分可观测系统中进行规划与学习》，载于 ACM计算理论研讨会，系列：STOC 2023。纽约，NY，美国：计算机协会，2023年，第 349–362 页。[15] Q. Liu、A. Chung、C. Szepesv´ari 和 C. Jin，《部分可观测强化学习何时不再令人头疼？》，载于学习理论会议。机器学习研究会论文集，2022年，第 5175–5220 页。[16] W. Lee、N. Rong 和 D. Hsu，《是什么让某些部分可观测马尔可夫决策过程问题易于近似？》，《神经信息处理系统进展》，第 20 卷，2007年。[17] J. Guo、Z. Li、H. Wang、M. Wang、Z. Yang 和 X. Zhang，《低秩部分可观测马尔可夫决策过程中具有可处理规划的可证明高效表示学习》，载于国际机器学习会议。机器学习研究会论文集，2023年，第 11 967–11 997 页。[18] M. Belly、N. Fijalkow、H. Gimbert、F. Horn、G. A. P´erez 和 P. Vandenhove，《Revelations：具有 ω-正则目标的可判定类部分可观测马尔可夫决策过程》，载于 AAAI人工智能会议论文集，第 39 卷，第 25 期，2025年，第 26 454–26 462 页。[19] H. Theil，《计量经济模型与福利最大化》，《经济档案》，第 72 卷，第 60–83 页，1954年。[20] ——，《关于动态规划中确定性等价的一则札记》，《计量经济学》，第 346–349 页，1957年。[21] H. A. Simon，《具有二次准则函数的不确定性动态规划》，《计量经济学：计量经济学会杂志》，第 74–81 页，1956年。[22] Y. Bar-Shalom 和 E. Tse，《随机控制中的双重效应、确定性等价与分离》，IEEE自动控制汇刊，第 19 卷，第 5 期，第 494–500 页，1974年。[23] M. S. Derpich 和 S. Y¨uksel，《双重效应、确定性等价与分离再探：一个反例及对最优性的宽松刻画》，《IEEE自动控制汇刊》，第 68 卷，1266，2022 2，1259 期，第 – 页。[24] Y. Li 和 D. Bertsekas，《半线性动态规划：分析、算法与确定性等价性质》，arXiv 预印本 arXiv:2501.04668，2025年。

[25] P. Whittle，“风险敏感确定性等价原理”，应用概率期刊，第23卷，第A期，第383–388页，1986年。[26] M. Hardt与B. Recht，模式、预测与动作：机器学习基础。普林斯顿大学出版社，2022年。[27] D. Bertsekas，动态规划与最优控制：第一卷。Athena Scientific，2012年。[28] B. Wolfe、M. R. James与S. Singh，“近似预测状态表示”，载于国际自主智能体与多智能体系统会议，2008年，第363–370页。[29] W. Hamilton、M. M. Fard与J. Pineau，“利用压缩预测状态进行高效学习与规划”，机器学习研究期刊，第15卷，第1期，第3395–3439页，2014年。[30] P. S. Castro、P. Panangaden与D. Precup，“完全可观测与部分可观测马尔可夫决策过程中的等价关系”，载于国际人工智能联合会议，2009年，第1653–1658页。[31] O. Hern´andez-Lerma与J. B. Lasserre，离散时间马尔可夫控制过程：基本最优性准则。施普林格科学与商业媒体，2012年。[32] K. Hinderer，“马尔可夫性决策过程中价值函数的利普希茨连续性”，运筹学数学方法，第62卷，第1期，第3–22页，2005年。[33] C. Gelada、S. Kumar、J. Buckman、O. Nachum与M. G. Bellemare，“DeepMDP：学习用于表示学习的连续潜在空间模型”，载于国际机器学习会议。机器学习研究会论文集，2019年，第2170–2179页。[34] A. Molin与S. Hirche，“网络化控制系统的次优事件触发控制”，第277–289页，2014年。[35] M. Mazo与P. Tabuada，“无线传感器/执行器网络上的分散式事件触发控制”，*IEEE*自动控制汇刊，第56卷，第10期，第2456–2461页，2011年。[36] W. P. Heemels、K. H. Johansson与P. Tabuada，“事件触发控制与自触发控制导论”，载于*IEEE*决策与控制会议。IEEE，2012年12月，第3270–3285页。[37] A. M¨uller，“马尔可夫决策过程的价值函数如何依赖于转移概率？”运筹学数学，第22卷，第4期，第872–885页，1997年。[38] C. Villani，最优传输：旧与新。施普林格，2009年。

附录 A 证明 引理 1

我们分别证明这两部分。

1）证明（13）：任意取 $s,\ s' \in \mathcal{S}$。令 $m = \phi(s)$ $m' = \phi(s') \in \mathcal{S} = \mathbb{R}$。于是，根据三角不等式，有

$$\begin{aligned}|c(s,a) - c(s',a)| &\le |\bar{\ell}(m,a) - \bar{\ell}(m',a)| + |\ell(s,a) - \ell(s',a)| \\ &\le L^{\bar{\ell}}|m - m'| + 2\beta := F_t^c(|m - m'|).\end{aligned}$$

2）证明（12）：任意取 $(m,\ a) \in \mathcal{S} \times \mathcal{A}$。令$M$ $(w) = \bar{f}\ (m,\ a,\ w) + \sum_{i=1}^{n} \alpha^i f^i\ (m\mathbf{1}_n,\ a,\ w)$，且 $M'\ (w') = \bar{f}\ (m',\ a,\ w') + \sum_{=1}^{n} \alpha^i f^i\ (m'\mathbf{1}_n,\ a,\ w')$。式（12）的左侧是随机变量$M$（$W$）与 M'（W'）之间的Wasserstein距离，其中 W和 W'是具有相同分布、边缘分布为 P_W的随机变量。令Γ 表示 W 与 W' 之间所有边缘分布为 P_W的联合耦合。那么，

$$\begin{aligned}d_{\text{Was}}(M(W), M'(W')) &= \inf_{(W,W')\sim\Gamma} \mathbb{E}[|M(W) - M'(W')|] \\ &\le \mathbb{E}[|M(W) - M'(W)|]\end{aligned}$$

其中，在最后一个方程中，我们选择了一种特定的耦合方式 $W = W'$。

现在注意到

$$
\begin{aligned}
&\mathbb{E}[|M(W) - M'(W)|] \\
&= \mathbb{E}\Bigg[\Bigg|\bar{f}(m, a, W) - \bar{f}(m', a, W) \\
&\quad + \sum_{i=1}^{n} \alpha^i \big(f^i(m\mathbf{1}_n, a, W) - f^i(m'\mathbf{1}_n, a, W)\big)\Bigg|\Bigg] \\
&\leq \mathbb{E}\big[|\bar{f}(m, a, W) - \bar{f}(m', a, W)|\big] \\
&+ \sum_{i=1}^{n} \alpha^i \mathbb{E}\Bigg[\Bigg|\big(f^i(m\mathbf{1}_n, a, W) - f^i(m'\mathbf{1}_n, a, W)\Bigg|\Bigg] \\
&\leq L^{\bar{f}}|m - m'| + 2\sum_{i=1}^{n} \alpha^i \gamma^i \coloneqq F_t^P(|m - m'|)
\end{aligned}
$$

其中，最后一个不等式由假设7成立。