arXiv:2601.03132v1 [eess.SY] 6 Jan 2026

部分可观测马尔可夫决策过程中的最优控制有限记忆信念近似

Mintae Kim

摘要—**我们研究部分可观测（PO）随机最优控制（SOC）问题中的有限记忆信念近似。尽管信念状态足以用于部分可观测马尔可夫决策过程（POMDP）中的随机最优控制，但信念状态通常是无限维的，因此不切实际。我们将截断的输入输出（IO）历史解释为形成信念近似，并建立一种基于度量的理论，直接关联信息损失与控制性能。利用瓦瑟斯坦度量，我们推导出策略条件性能界，量化有限记忆沿典型闭环轨迹导致的价值劣化。我们的分析采用固定策略比较：在相同的闭环执行下评估两个代价泛函，并分离出在信念层级代价中以有限记忆近似替代真实信念所产生的影响。对于线性二次高斯（LQG）系统，我们提供信念失配的闭式形式评估，并通过实证验证所预测的机制，证明信念失配随记忆长度近似呈指数衰减，且由此产生的性能失配也相应变化。总体而言，这些结果以度量为基础，刻画了有限记忆信念近似在部分可观测情境下能够和不能够实现的效果。**

一、引言

在部分可观测随机最优控制（SOC）中，控制器无法直接观测系统状态。因此，决策必须依据过去的观测和控制输入，即输入输出历史。众所周知，最优控制可以用信念状态来表示，即给定输入输出历史时当前状态的后验分布；由此可将部分可观测马尔可夫决策过程 [1], [2], [3] 精确地表述为完全可观测的信念马尔可夫决策过程（信念MDP）。尽管这种表述是精确的，但即使对于简单的连续状态系统，信念通常也是无限维的，因此难以计算和存储 [4]。因此，部分可观测系统的实际控制器依赖有限记忆。一种常见架构是使用近期观测和输入的滑动窗口，并通过有限记忆策略选择控制动作 [3], [5]。在本文中，有限记忆指基于截断的输入输出历史进行操作，从而得到有限维信息状态，而非使用完整的输入输出历史。此类有限记忆架构广泛应用于经典控制和基于学习的控制，但其理论依据和性能分析仍不完善 [6], [7], [8], [9]。一个核心问题是，有限记忆何时能够有效替代信念状态，以及由此产生的信息损失会如何影响闭环性能。

作者供职于*Hybrid Robotics Lab*，加利福尼亚大学伯克利分校，美国加利福尼亚州伯克利，邮编 94720。电子邮箱：mintae.kim@berkeley.edu。代码和补充材料可在 https://github.com/mintaeshkim/fmba 获取。

部分可观测系统中的有限记忆策略已得到广泛研究 [3], [4], [5], [10]。具体而言， [5] 针对部分可观测马尔可夫决策过程中的有限记忆策略，在非一致的典型轨迹近似准则下确立了近似最优性结果，表明即使不对所有观测序列进行一致近似，也能实现较小的性能损失。在机器人学和基于学习的控制中，常见做法是将有限输入输出窗口（通常还包括当前观测）输入学习型控制器 [8], [9]。然而，从随机最优控制的角度来看，仍存在若干不足。有限记忆通常被视为对策略类的限制，而非对底层信念过程的近似 [5]。此外，现有结果很少在闭环轨迹上定量地建立信息损失与价值劣化之间考虑度量的关系，而基本局限往往也未被明确揭示。

本文针对部分可观测随机最优控制（POSOC），建立了一套基于度量的信念状态有限记忆近似理论。我们不将有限记忆仅仅视为对策略的限制，而是把截断的输入输出历史视为一种显式的有限记忆信念近似，并沿固定策略生成的轨迹，使用 Wasserstein-2 度量衡量其与真实信念之间的差异。这种以策略为条件的视角避免了对不太可能出现的历史提出一致的最坏情况要求，并得出了有限且易于解释的界 [5]。在适当的正则性条件下，我们证明，在相同的闭环执行下评估时，Wasserstein 信念失配能够控制真实信念代价泛函与其有限记忆对应项之间的性能差距；此外，我们将这种固定策略下的比较推广为最优值差距界。我们还刻画了有限记忆控制的基本局限，包括为重构信念而保留输入历史的必要性。最后，我们将该框架专门应用于 LQG 系统，在此类系统中，信念失配和性能失配均可通过闭式形式计算；并通过实证验证了本文的核心机制：截断输入输出历史会导致可测量的信念失配，而该失配能够定量解释性能下降。

II. 问题设定与预备知识

本节介绍本文研究的部分可观测随机最优控制问题，并定义所有对象和符号。

我们考虑一个由元组 $(\mathcal{X},\ \mathcal{U},\ \mathcal{Y},\ P,\ O,\ c,\ \gamma)$ 指定的无限时域折扣部分可观测马尔可夫决策过程，其中 $\mathcal{X} \subset \mathbb{R}^n$ 是状态空间， $\mathcal{U} \subset \mathbb{R}^m$ 是控制空间， $\mathcal{Y} \subset \mathbb{R}^p$ 是观测空间， $P(x' \mid x,\ u)$ 是转移核，$O(y \mid x)$ 是观测核， $c:\ \mathcal{X} \times \mathcal{U} \to \mathbb{R}$

是阶段成本，且 $\gamma \in (0,\ 1)$ 是折扣因子。[1]系统根据 $x_{t+1} \sim P(\cdot \mid x_t,\ u_t)$ 和 $y_t \sim O(\cdot \mid x_t)$ 演化，而控制器只能观测输入输出历史 $\zeta_t = (y_{0:t},\ u_{0:t-1})$，并据此选择 $u_t = \pi_t(\zeta_t)$。

与 ζ_t 相关联的信念状态定义为

$$b_t := \mathbb{P}(x_t \mid \zeta_t), \tag{1}$$

它是定义在 $\mathcal{X}$ 上的概率测度。请注意，信念被定义为概率测度，而非状态本身。信念通过贝叶斯滤波器 $b_{t+1} = \Phi(b_t,\ u_t,\ y_{t+1})$演化，其中 Φ表示由（$P,\ O$）诱导的信念更新算子。利用信念状态，POMDP 可以精确地约化为定义在信念空间$\mathcal{P}(\mathcal{X})$上的完全可观测 MDP，这通常称为信念*MDP*。这种约化依赖于以下事实：信念 b_t 是控制的充分统计量，具体而言，任何诱导出相同信念的 IO 历史，在任意控制序列下，对未来状态、观测和成本的条件分布都相同。特别地，给定信念 b_t 和输入u_t，下一状态的预测分布由 $\mathbb{P}(x_{t+1} \in \cdot \mid b_t,\ u_t) = \int P(\cdot \mid x,\ u_t)\ b_t(dx)$唯一确定；下一信念 b_{t+1} 则通过对（$b_t,\ u_t,\ y_{t+1}$）应用贝叶斯滤波器得到。因此，信念过程 $\{b_t\}$是一个受控马尔可夫过程，满足

$$\mathbb{P}(b_{t+1} \mid y_{0:t}, u_{0:t}) = \mathbb{P}(b_{t+1} \mid b_t, u_t). \tag{2}$$

此外，阶段成本可以精确地表示为 $\bar{c}(b_t,\ u_t) := \mathbb{E}_{x\sim b_t}[c(x,\ u_t)] = \int_{\mathcal{X}} c(x,\ u_t) b_t(dx)$，且在信念空间上定义的策略等价于依赖历史的策略。因此，信念MDP的建模不会引入近似，也不会损失最优性。

因此，一旦将信念视为系统状态，POSOC问题就化简为定义在 $\mathcal{P}(\mathcal{X})$ 上的全观测折扣MDP，并可应用标准的动态规划论证。特别地，最优价值函数为

$$V^\star(b_0) := \inf_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \bar{c}(b_t, u_t) \;\middle|\; b_0\right], \tag{3}$$

其中，下确界是在基于信念的策略 $u_t = \pi(b_t)$ 上取的。相应的贝尔曼算子为

$$(\mathcal{T}V)(b) = \inf_{u\in\mathcal{U}} \{\bar{c}(b,u) + \gamma\, \mathbb{E}[V(\Phi(b,u,y'))]\}, \tag{4}$$

其中，y' 服从由 $(b,\ u)$ 诱导的预测观测律。在完整的输入输出历史上定义的策略与在信念状态上定义的策略，是同一决策规则的等价表示；为简便起见，本文始终采用基于信念的表示。

在以下各节中，为了量化信念近似误差，我们将在瓦瑟斯坦度量空间上进行研究

$$\mathcal{P}_2(\mathcal{X}) := \left\{\mu \in \mathcal{P}(\mathcal{X}) \;\middle|\; \mathbb{E}_{x\sim\mu}[\|x\|^2] \in (0,\infty)\right\}, \tag{5}$$

配备了 2-瓦瑟斯坦距离 W_2。

[1]我们采用基于核的表述以保持一般性。基于随机微分方程的模型将通过离散化作为线性二次高斯特例进行讨论，并在第五节中明确处理。

对于记忆长度 H，我们定义截断的输入输出历史

$$\zeta_t^{(H)} := (y_{t-H:t}, u_{t-H:t-1}), \tag{6}$$

以及相应的有限记忆信念近似

$$\hat{b}_t^{(H)} := \mathbb{P}(x_t \mid \zeta_t^{(H)}). \tag{7}$$

这些对象完整地定义了本文其余部分所使用的信念近似框架。

III. 有限记忆信念近似

本节分析第二节中定义的有限记忆信念近似$\hat{b}_t^{(H)}$，并建立由使用截断输入输出历史所导致的策略条件信念近似误差的指数上界。在本节中，所有信念均取值于配备 2-瓦瑟斯坦距离 W_2 的 $\mathcal{P}_2(\mathcal{X})$。

对于固定策略 π，令 b_t^π 表示在 π 下由闭环轨迹诱导的真实信念过程，令$\hat{b}^{(H),\ \pi}_{t}$ 表示由相同截断输入输出历史诱导的有限记忆信念近似。两种信念定义在同一概率空间上，且仅在 σ-代数方面有所不同。

我们将策略条件有限记忆信念近似误差定义为

$$\varepsilon_H(\pi) := \sup_{t\ge 0} \mathbb{E}_\pi\left[W_2\left(b_t^\pi, \hat{b}_t^{(H),\pi}\right)\right]. \tag{8}$$

该定义仅沿由 π 诱导的闭环分布下实现的轨迹评估近似质量，从而避免对输入输出历史取一致上确界。一致近似 是一种自然但过于严格的方法，它控制所有可能历史下信念状态与其有限记忆近似之间的最坏偏差。在随机系统中，可行输入输出历史的空间十分庞大，其中许多轨迹在闭环反馈控制下发生的概率趋近于零。因此，一致近似提出了不必要的强要求，实际上要求即使沿着概率极低的样本路径也能准确近似。具体而言，一致近似隐含地要求信念更新在控制下具有路径稳定性，而在受控情形中通常无法保证这一点 [5], [10]。

为确保下文界中的常数有限且与策略无关，并排除与信息截断无关的退化行为，我们将讨论范围限定在稳定策略类 Π_{stab} 内；在该策略类下，所有信念过程的二阶矩都一致有界。

有限记忆截断的影响取决于在闭环运行下，信念更新遗忘远端信息的速度。

假设*1*（受控遗忘，Π_{stab}）：存在常数 $\rho \in (0,\ 1)$ 和 $C_\pi \in (0,\ \infty)$，使得对于任意$\pi \in \Pi_{\mathrm{stab}}$、任意初始信念 $\mu,\ \nu \in \mathcal{P}_2(\mathcal{X})$ 以及所有 $t \ge 0$,

$$\mathbb{E}[W_2(\Phi_t^\pi(\mu), \Phi_t^\pi(\nu)) \mid u_{0:t-1}] \le C_\pi\, \rho^t\, W_2(\mu,\nu), \tag{9}$$

其中，$\Phi_t^\pi(\cdot)$ 表示由 π生成的已实现输入输出序列驱动的 t 步信念更新。

假设1 表示在给定已实现输入序列的条件下，信念更新具有指数稳定性，并不要求信念更新具有逐点收缩性。

对于 $t \geq H$，将时刻 $t-H$ 的边界信念定义为

$$\tilde{b}^{\pi}_{t-H} := \mathbb{P}_{\pi}(x_{t-H} \mid y_{t-H}). \tag{10}$$

然后，通过在时刻 $t-H$ 用 b^{π}_{t-H} 初始化信念递推，并沿着已实现序列 $(u_{t-H:t-1}, y_{t-H+1:t})$ 应用信念更新算子，即可得到 $\hat{b}^{(H),\pi}_{t}$。

引理*1*（有限记忆信念表示）：对于所有$t \geq H$，有限记忆信念近似满足 $\hat{b}^{(H),\pi}_{t} = \Phi^{\pi}_{H}(\tilde{b}^{\pi}_{t-H})$，其中 $\Phi^{\pi}_{H}(\cdot)$ 表示由已实现 IO 序列驱动的 H 步信念更新。

证明：根据受控状态过程的马尔可夫性质和贝叶斯法则，以 $(y_{t-H:t}, u_{t-H:t-1})$ 为条件，等价于以 y_{t-H} 为条件来初始化时刻 $t-H$ 的信念，然后沿后缀 $(u_{t-H:t-1}, y_{t-H+1:t})$ 递归应用贝叶斯滤波器。∎

引理*2*（矩界蕴含瓦瑟斯坦距离界）：若 $\mu, \nu \in \mathcal{P}_2(\mathcal{X})$，由 (5) 可知，存在 $M \in (0,\infty)$ 使得 $\mathbb{E}_{x\sim\mu}[\|x\|^2] \leq M$ 且 $\mathbb{E}_{x\sim\nu}[\|x\|^2] \leq M$，则 $W_2(\mu,\nu) \leq 2\sqrt{M}$。

证明：设 $X \sim \mu$ 和 $Y \sim \nu$ 相互独立。于是，根据 W_2 度量的性质，$W_2(\mu,\nu)^2 \leq \mathbb{E}[\|X-Y\|^2] \leq 2\mathbb{E}[\|X\|^2] + 2\mathbb{E}[\|Y\|^2] \leq 4M$。∎

引理 *3*（遗忘蕴含有限记忆精度）：假设假设1对某个 $\pi \in \Pi_{\mathrm{stab}}$成立。那么存在一个常数 $C'_{\pi} \in (0,\infty)$，使得对所有 $H \geq 0$,

$$\varepsilon_H(\pi) \leq C'_{\pi}\, \rho^{H}. \tag{11}$$

证明：固定 $H \geq 0$ and $t \geq H$. 根据引理1，b^{π}_{t} 和 $\hat{b}^{(H),\pi}_{t}$ 分别沿着相同的已实现输入输出序列，对初始信念 b^{π}_{t-H} 和 $\tilde{b}^{\pi}_{t-H}$ 应用相同的 H 步信念更新得到。在该时间窗口内应用假设1，得到

$$\mathbb{E}_{\pi}\left[W_2\left(b^{\pi}_{t}, \hat{b}^{(H),\pi}_{t}\right)\right] \leq C_{\pi}\, \rho^{H}\, \mathbb{E}_{\pi}\left[W_2\left(b^{\pi}_{t-H}, \tilde{b}^{\pi}_{t-H}\right)\right]. \tag{12}$$

由于 $\pi \in \Pi_{\mathrm{stab}}$，两个信念 b^{π}_{t} 和 $\hat{b}^{(H),\pi}_{t}$ 的二阶矩均一致有界，因此引理2推出 $\sup_t \mathbb{E}_{\pi}[W_2(b^{\pi}_{t-H},\ b^{\pi}_{t-H})] \in (0,\infty)$。将此界吸收到常数中，即可得到对所有 $t \geq H$ 成立的所述不等式。对于 $t < H$，同样适用该矩界，且 $\rho^{H} \leq 1$ 可将其吸收到同一个常数中。对 $t \geq 0$ 取上确界，即完成证明。∎

在讨论性能保证之前，我们先说明，有限记忆必须同时保留观测历史和输入历史。否则，即使在简单的受控系统中，信念重构也会失败。

命题*1*（输入历史的必要性）：存在这样的 POSOC系统：任何仅依赖有限观测窗口（$y_{t-H:t}$）的控制器，都无法在不依赖过去输入（$u_{t-H:t-1}$）的情况下，唯一确定后验分布 $\mathbb{P}(x_t \mid y_{t-H:t})$。

第四节仅利用量 $\varepsilon_H(\pi)$ 和上述已证明的指数衰减，将信息损失转化为性能损失界，而无需假设贝尔曼算子或价值函数具有任何全局正则性。

IV. 基于有限记忆信念近似的策略的性能保证

本节通过在固定策略下评估的策略条件信念失配，界定有限记忆信念近似所导致的性能损失。我们比较在同一策略所诱导的闭环执行下评估的两个代价泛函，它们仅在信念层级阶段代价中的信念参数上有所不同。[2] 这样就无需比较两条不同的闭环轨迹，也无需假设贝尔曼算子或价值函数具有任何全局正则性。

固定记忆长度 $H \geq 0$ 和基于信念的策略 π。过程 $\{b^{\pi}_{t}\}$ 和 $\{\hat{b}^{(H),\pi}_{t}\}$ 定义在同一概率空间上，并通过同一已实现输入输出序列耦合。如脚注所述，在本节中，我们沿同一已实现输入序列 $u_t = \pi(b^{\pi}_{t})$评估两个代价泛函，并且只改变代价 $\bar{c}(\cdot,\ u_t)$中的信念参数。

我们将 π 下的真实信念代价泛函定义为

$$J(\pi) := \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^{t} \bar{c}(b^{\pi}_{t}, u_t)\right], \tag{13}$$

并将相同 π 及输入 $u_t = \pi(b^{\pi}_{t})$下的有限记忆信念近似代价泛函定义为

$$\hat{J}_H(\pi) := \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^{t} \bar{c}\left(\hat{b}^{(H),\pi}_{t}, u_t\right)\right]. \tag{14}$$

具体而言，$J(\pi)$ 和 $\hat{J}_H(\pi)$ 是在由 π 诱导的、关于 $(x_{0:\infty}, y_{0:\infty}, u_{0:\infty})$ 的相同分布下进行评估的；二者的唯一区别在于，将 $\bar{c}(\cdot, u_t)$ 内的 b^{π}_{t} 替换为 $\hat{b}^{(H),\pi}_{t}$。

回顾在 π 下由策略条件误差定义的有限记忆信念失配，见式 (8)。本节余下部分将说明信念失配 $\varepsilon_H(\pi)$ 如何转化为性能差距。第三节仅用于将 $\varepsilon_H(\pi)$ 上界化为 H 的函数。

为了在固定的闭环执行下将信念失配转化为定量的性能界，我们施加温和的策略条件正则性条件，以确保矩有限，并保证信念层级代价具有局部光滑性。

假设 *2*（策略条件状态正则性）：对于固定的可容许策略 π，存在常数 $M_{\pi} \in (0,\infty)$，使得对于测度 b^{π}_{t} 和 $\hat{b}^{(H),\pi}_{t}$ 均有

$$\sup_{t\geq 0} \mathbb{E}_{x\sim b^{\pi}_{t}}\left[\|x\|^2\right] \leq M_{\pi}, \quad \sup_{t\geq 0} \mathbb{E}_{x\sim \hat{b}^{(H),\pi}_{t}}\left[\|x\|^2\right] \leq M_{\pi}. \tag{15}$$

假设 *3*（二次增长和代价的光滑性）：存在常数 $K_c > 0$ 和 $K_g > 0$，使得对于所有 $(x,u) \in \mathcal{X} \times \mathcal{U}$,

$$|c(x,u)| \leq K_c\left(1 + \|x\|^2 + \|u\|^2\right), \tag{16}$$

$$\|\nabla_x c(x,u)\| \leq K_g\left(1 + \|x\| + \|u\|\right). \tag{17}$$

[2]由信念近似诱导的策略始终劣于由真实信念诱导的策略。在相同策略和输入下进行成本比较，是实际比较 $J^{\star}$ 与 $\hat{J}_H(\pi^{\star}_{H})$ 的一个中间步骤。

假设 2 确保在由 π 诱导的固定闭环下，以下常数均为有限值。假设 3 与二次代价相容，且不要求 c 在 x 上为全局 Lipschitz。

引理 *4*（信念层面的成本敏感性，在 W_2）：假设固定策略π满足假设 2 和 3。则存在有限常数 $L_\pi \in (0, \infty)$，使得对于任意 $u \in \mathcal{U}$以及任意满足$\mathbb{E}_{x\sim b}[\|x\|^2] \le M_\pi$ 和 $\mathbb{E}_{x\sim \tilde{b}}[\|x\|^2] \le M_\pi$的 $b,\ b \in \mathcal{P}_2(\mathcal{X})$，

$$\left|\bar{c}(b,u) - \bar{c}(\tilde{b},u)\right| \le L_\pi \left(1 + \|u\|^2\right) W_2(b,\tilde{b}). \tag{18}$$

证明：固定满足所述二阶矩界的 $u \in \mathcal{U}$和 $b,\ b \in \mathcal{P}_2(\mathcal{X})$。令 $(X,\ X)$ 为 $(b,\ b)$ 的任意耦合。定义 $X_\lambda := X + \lambda(X - X)$，其中 $\lambda \in [0,1]$。根据微积分基本定理，

$$c(X,u) - c(\tilde{X},u) = \int_0^1 \nabla_x c(X_\lambda,u)^\top (X - \tilde{X})\, d\lambda. \tag{19}$$

取绝对值并应用柯西-施瓦茨不等式可得

$$|c(X,u) - c(\tilde{X},u)| \le \left(\int_0^1 \|\nabla_x c(X_\lambda,u)\|\, d\lambda\right) \|X - \tilde{X}\|. \tag{20}$$

取期望并应用柯西-施瓦茨不等式可得

$$\mathbb{E}\Big[|c(X,u) - c(\tilde{X},u)|\Big] \le \left(\mathbb{E}\left[\left(\int_0^1 \|\nabla_x c(X_\lambda,u)\|\, d\lambda\right)^2\right]\right)^{1/2} \left(\mathbb{E}\|X - \tilde{X}\|^2\right)^{1/2}. \tag{21}$$

根据假设 3，

$$\|\nabla_x c(X_\lambda,u)\| \le K_g \left(1 + \|X_\lambda\| + \|u\|\right). \tag{22}$$

此外，$\|X_\lambda\| \le \|X\| + \|X - X\|$ 意味着

$$1 + \|X_\lambda\| + \|u\| \le 1 + \|\tilde{X}\| + \|X - \tilde{X}\| + \|u\|. \tag{23}$$

根据詹森不等式，

$$\left(\int_0^1 \|\nabla_x c(X_\lambda,u)\|\, d\lambda\right)^2 \le \int_0^1 \|\nabla_x c(X_\lambda,u)\|^2\, d\lambda, \tag{24}$$

因此

$$\mathbb{E}\left[\left(\int_0^1 \|\nabla_x c(X_\lambda,u)\|\, d\lambda\right)^2\right] \le K_g^2\, \mathbb{E}\left[\int_0^1 \left(1 + \|X_\lambda\| + \|u\|\right)^2 d\lambda\right]. \tag{25}$$

利用 $(a+b+c)^2 \le 3(a^2+b^2+c^2)$ 以及关于 $\|X_\lambda\|$ 的界，

$$\left(1 + \|X_\lambda\| + \|u\|\right)^2 \le 3\left(1 + \|u\|^2 + \|X_\lambda\|^2\right) \le 3\left(1 + \|u\|^2 + 2\|\tilde{X}\|^2 + 2\|X - \tilde{X}\|^2\right). \tag{26}$$

因此，

$$\mathbb{E}\left[\left(\int_0^1 \|\nabla_x c(X_\lambda,u)\|\, d\lambda\right)^2\right] \le 3K_g^2\left(1 + \|u\|^2\right) + 6K_g^2\, \mathbb{E}[\|\tilde{X}\|^2] + 6K_g^2\, \mathbb{E}[\|X - \tilde{X}\|^2]. \tag{27}$$

由于 $X \sim b$ 和 $\tilde{X} \sim \tilde{b}$ 满足 $\mathbb{E}[\|X\|^2] \le M_\pi$ 和 $\mathbb{E}[\|\tilde{X}\|^2] \le M_\pi$，我们还可以得到

$$\mathbb{E}[\|X - \tilde{X}\|^2] \le 2\mathbb{E}[\|X\|^2] + 2\mathbb{E}[\|\tilde{X}\|]^2 \le 4M_\pi. \tag{28}$$

结合 (27) 和 (28) 可得一致界

$$\mathbb{E}\left[\left(\int_0^1 \|\nabla_x c(X_\lambda,u)\|\, d\lambda\right)^2\right] \le \tilde{L}_\pi^2\left(1 + \|u\|^2\right), \tag{29}$$

其中可取 $\tilde{L}_\pi := K_g\sqrt{3 + 24M_\pi}$。代入柯西-施瓦茨界可得

$$\mathbb{E}\Big[|c(X,u) - c(\tilde{X},u)|\Big] \le \tilde{L}_\pi \left(1 + \|u\|^2\right)^{1/2} \left(\mathbb{E}[\|X - \tilde{X}\|^2]\right)^{1/2}. \tag{30}$$

由于 $\left|\bar{c}(\quad b\ u\ -\bar{c}\quad u\qquad X\ u\ -c\quad u\qquad 1+,)\right.$ $\left.(\tilde{b},)\right| \le \mathbb{E}\Big[|c(,)\,(\ \tilde{X},)|\Big]$ 且 $(\ )$，我们得到

$$\left|\bar{c}(b,u) - \bar{c}(\tilde{b},u)\right| \le \tilde{L}_\pi \left(1 + \|u\|^2\right) \left(\mathbb{E}\|X - \tilde{X}\|^2\right)^{1/2}. \tag{31}$$

对所有耦合 (X,X) 取下确界，得到

$$\left|\bar{c}(b,u) - \bar{c}(\tilde{b},u)\right| \le \tilde{L}_\pi \left(1 + \|u\|^2\right) W_2(b,\tilde{b}). \tag{32}$$

令 $L_\pi := L_\pi$，证毕。 ■

引理 *5*（固定策略的性能差异 *h):*

再次假设固定策略π 满足假设 2 和 3，并进一步假设闭环输入 $u_t = \pi(b_t^\pi)$ 满足 $\sup_{t\ge 0} \mathbb{E}_\pi[\|u_t\|^2] \in (0,\infty)$。进一步假设存在常数 $U_\pi \in (0,\infty)$，使得$\|u_t\|^2 \le U_\pi$ 对所有 $t \ge 0$几乎必然成立。输入有界假设在大多数最优控制问题中是合理的。于是，存在有限常数 $C_\pi \in (0,\infty)$，使得对所有 $H \ge 0$，

$$\left|J(\pi) - \hat{J}_H(\pi)\right| \le \frac{C_\pi}{1-\gamma}\, \varepsilon_H(\pi). \tag{33}$$

证明：根据 $J(\pi)$ 和 $J_H(\pi)$ 的定义（见式 (13) 和 (14)），以及三角不等式，

$$\left|J(\pi) - \hat{J}_H(\pi)\right| \le \sum_{t=0}^{\infty} \gamma^t\, \mathbb{E}_\pi\left[\left|\bar{c}(b_t^\pi,u_t) - \bar{c}(\hat{b}_t^{(H),\pi},u_t)\right|\right]. \tag{34}$$

根据引理 4，

$$\left|\bar{c}(b_t^\pi,u_t) - \bar{c}(\hat{b}_t^{(H),\pi},u_t)\right| \le L_\pi \left(1 + \|u_t\|^2\right) W_2\left(b_t^\pi, \hat{b}_t^{(H),\pi}\right). \tag{35}$$

取期望可得

$$\mathbb{E}_\pi\left[\left|\bar{c}(b_t^\pi,u_t) - \bar{c}(\hat{b}_t^{(H),\pi},u_t)\right|\right] \le L_\pi\, \mathbb{E}_\pi\left[\left(1 + \|u_t\|^2\right) W_2\left(b_t^\pi, \hat{b}_t^{(H),\pi}\right)\right]. \tag{36}$$

根据输入一致有界性，我们有

$$\mathbb{E}_\pi\left[\left(1 + \|u_t\|^2\right) W_2\left(b_t^\pi, \hat{b}_t^{(H),\pi}\right)\right] \le (1 + U_\pi)\, \mathbb{E}_\pi\left[W_2\left(b_t^\pi, \hat{b}_t^{(H),\pi}\right)\right] \le (1 + U_\pi)\, \varepsilon_H(\pi). \tag{37}$$

结合上述不等式并对 $t \geq 0$ 求和，得到

$$\left|J(\pi) - \hat{J}_H(\pi)\right| \leq L_\pi \left(1 + U_\pi\right) \varepsilon_H(\pi) \sum_{t=0}^{\infty} \gamma^t = \frac{C_\pi}{1-\gamma} \varepsilon_H(\pi), \tag{38}$$

其中 $C_\pi := L_\pi (1 + U_\pi)$。 ■

现在，我们将固定策略失配界推广为信念最优控制器与最优有限记忆控制器之间的最优值差界。令 $\pi^\star$ 表示信念MDP的一个最优策略，令 $\pi_H^\star$ 表示在有限记忆信念近似策略中、且关于 $\zeta_t^{(H)}$ 可测的一个最优策略。

定理*1*（基于信念失配的性能界）：假设对于 $\pi^\star$，假设 2 和 3 成立，并且进一步假设输入一致有界，

$$\sup_{t \geq 0} \mathbb{E}_{\pi^\star}[\|u_t\|^2] \in (0, \infty) \tag{39}$$

对于闭环输入 $u_t = \pi^\star(b_t^{\pi^\star})$。那么对于每个 $H \geq 0$,

$$0 \leq J^\star - J_H^\star \leq \frac{C_{\pi^\star}}{1-\gamma} \varepsilon_H(\pi^\star), \tag{40}$$

其中 $J^\star := J(\pi^\star)$ 且 $J_H^\star := \hat{J}_H(\pi_H^\star)$。

证明：记 Π_H 为有限记忆策略类，最小化 $\hat{J}_H(\cdot)$ 有

$$\hat{J}_H(\pi_H^\star) \leq \hat{J}_H(\pi^\star). \tag{41}$$

因此，

$$J^\star - J_H^\star = J(\pi^\star) - \hat{J}_H(\pi_H^\star) \leq J(\pi^\star) - \hat{J}_H(\pi^\star). \tag{42}$$

将引理 5 应用于 $\pi = \pi^\star$，得到

$$J(\pi^\star) - \hat{J}_H(\pi^\star) \leq \frac{C_{\pi^\star}}{1-\gamma} \varepsilon_H(\pi^\star), \tag{43}$$

这就证明了该结论。 ■

推论*1*（指数衰减与遗忘）：
假设定理1的条件成立，并且还存在常数 $C'_{\pi^\star} \in (0, \infty)$ 和 $\rho \in (0,1)$，使得

$$\varepsilon_H(\pi^\star) \leq C'_{\pi^\star} \rho^H \tag{44}$$

对所有 $H \geq 0$ 均成立。因此，对所有 $H \geq 0$,

$$J^\star - J_H^\star \leq \frac{C_{\pi^\star} C'_{\pi^\star}}{1-\gamma} \rho^H. \tag{45}$$

证明： 将 $\varepsilon_H(\pi^\star) \leq C'_{\pi^\star} \rho^H$ 代入定理1即可得到该界。 ■

推论1表明，当 $\varepsilon_H(\pi^\star)$ 在 H 中呈指数衰减时，有限记忆最优性差距会在 H 中呈指数衰减。第三节通过稳定策略下信念更新的受控遗忘，给出了这种指数衰减成立的充分条件。

V. 线性二次高斯特例与数值结果

本节使用一个线性二次高斯系统，对第三节和第四节的理论结果进行实证验证。我们的目标是检验截断的输入输出历史是否会导致信念失配，以及这种失配是否如理论所预测的那样能够解释性能差距。尽管线性二次高斯系统中的信念计算可通过卡尔曼递推以闭式形式完成，但我们在此仅将这种易处理性视为实验优势。卡尔曼滤波器能够精确评估真实信念、有限记忆信念近似，以及二者之间的瓦瑟斯坦差异。有限记忆信念近似通过丢弃过去的信息构造，因此随着 H 减小，会造成信息损失。

我们考虑部分可观测线性随机系统

$$x_{t+1} = Ax_t + Bu_t + w_t, \qquad w_t \sim \mathcal{N}(0, \Sigma_w), \tag{46}$$

$$y_t = Cx_t + v_t, \qquad v_t \sim \mathcal{N}(0, \Sigma_v), \tag{47}$$

其高斯先验为 $x_0 \sim \mathcal{N}(m_0, P_0)$，且具有二次代价，

$$c(x_t, u_t) = x_t^\top Q x_t + u_t^\top R u_t, \tag{48}$$

并使用式 (13) 和式 (14) 中定义的折扣无限时间范围目标函数来评估性能。

真实信念 $b_t = \mathcal{N}(m_t, P_t)$ 由卡尔曼滤波器获得，控制策略固定为线性二次高斯控制器

$$u_t = -K m_t, \tag{49}$$

其中 K 是无限时间范围的线性二次调节器增益。在本节中，闭环执行始终由该策略生成，只有信念级代价中的信念参数会改变，这与理论分析中研究的固定策略比较完全一致。

有限记忆信念近似 $\hat{b}_t^{(H)}$ 通过与引理1一致的窗口重启构造实现。对于每个 $t \geq H$，信念递推会在时间 $s = t - H$ 使用边界信念 b_s 重新初始化，该边界信念仅取决于单个观测 y_s。在实现中，b_s 通过使用 y_s，基于固定先验 (m_0, P_0) 执行卡尔曼测量更新而获得。随后，仅使用截断输入输出 $(u_{s:s+H-1}, y_{s+1:s+H})$ 将信念向前传播 H 步，从而得到 $\hat{b}_t^{(H)}$。

由于 b_t 和 $\hat{b}_t^{(H)}$ 均为高斯分布，因此信念失配采用闭式形式的 Wasserstein-2 距离计算，

$$W_2^2(b_t, \hat{b}_t^{(H)}) = \|m_t - \hat{m}_t^{(H)}\|^2 + \operatorname{Tr}\Big(P_t + \hat{P}_t^{(H)} - 2\big(P_t^{1/2} \hat{P}_t^{(H)} P_t^{1/2}\big)^{1/2}\Big). \tag{50}$$

策略条件误差 $\varepsilon_H(\pi)$ 通过对多次滚动执行进行蒙特卡洛平均来估计。

该系统实例化为一个线性二次高斯双积分器，

$$A = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} \frac{1}{2}\Delta t^2 \\ \Delta t \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 \end{bmatrix}, \tag{51}$$

因此只能观测到位置。所有实验均使用固定的控制器，并遍历记忆长度 $H \in \{0, 1, 2, 5, 10, 20, 50, 100\}$，时间范围为 $T = 1000$，随机种子为 50 个。

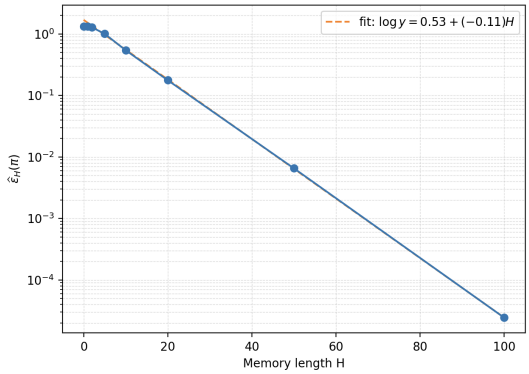


图 1：信念失配 $\varepsilon_H(\pi)$ 与记忆长度 H 的关系（y 轴为对数坐标）。近似线性的衰减证实了闭环运行下的指数遗忘。

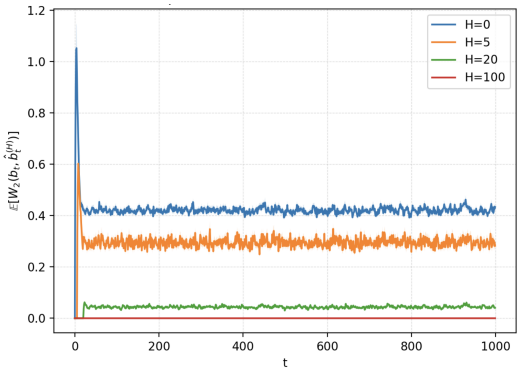


图 3：所选记忆长度下信念失配 $W_2(b_t, \hat{b}_t^{(H)})$ 的时间变化曲线。曲线显示 50 个随机种子下的均值 $\pm$ 标准误。

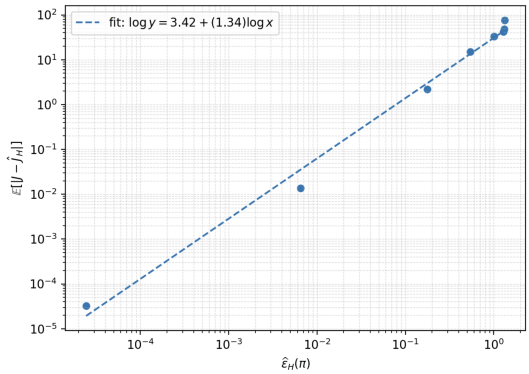


图 2：固定策略下，成本失配与信念失配的关系（双对数坐标）。观察到的线性缩放支持理论界 $|J(\pi) - \hat{J}_H(\pi)| \propto \varepsilon_H(\pi)$。

图 1 绘制了估计的信念失配 $\varepsilon_H(\pi)$ 随记忆长度 H 的变化。与引理 3 一致，失配随 H 近似呈指数衰减，在对数坐标轴上表现为线性趋势。

图 2 考察了信念失配与固定策略下性能差距之间的关系。正如引理 5 所预测的，成本差距随 $\varepsilon_H(\pi)$ 近似线性增长，这在对数坐标轴上呈现为线性趋势。

为便于诊断，图 3 展示了代表性 H 取值下信念失配随时间的演变。失配在早期瞬态阶段最大，之后趋于稳定，这说明有限记忆主要影响滤波器随时间累积信息的能力。

总体而言，线性二次高斯实验具体验证了本文的核心机制。截断的输入输出历史会导致可测量的信念失配，进而解释了闭环控制下的性能劣化。这里采用线性二次高斯并非为了简化信念近似，而是为了提供一个能够以闭式形式精确刻画信息损失、信念误差和成本劣化的场景。

VI. 结论

本文通过将截断的输入输出历史视为诱导有限记忆信念近似，研究了有限记忆部分可观测随机最优控制。通过衡量 Wasserstein 度量中的信息损失，并在固定闭环执行下评估性能，我们建立了信念失配与价值劣化之间的关系。在受控遗忘条件下，信念近似误差随记忆长度呈指数衰减，从而得到性能差距的显式指数界。我们的分析表明，有限记忆应被视为对底层信息状态的近似，而不应仅仅看作对策略类的限制。我们还指出了有限记忆控制的基本局限，包括必须保留输入历史，以及一般部分可观测系统中不可避免的指数级记忆需求。对 LQG 系统的特化分析表明，即使信念计算易于处理，这些效应依然存在；数值实验则验证了理论界能够捕捉到观测到的尺度变化规律。

参考文献

[1] R. D. Smallwood 和 E. J. Sondik，“有限时间范围内部分可观测马尔可夫过程的最优控制”，*Operations research*，第 21 卷，第 5 期，第 1071–1088 页，1973 年。[2] L. P. Kaelbling、M. L. Littman 和 A. R. Cassandra，“部分可观测随机领域中的规划与行动”，*Artificial intelligence*，第 101 卷，第 1-2 期，第 99–134 页，1998 年。[3] N. Saldi、S. Y¨uksel 和 T. Linder，“具有 Borel 空间的马尔可夫决策过程有限近似的渐近最优性”，*Mathematics of Operations Research*，第 42 卷，第 4 期，第 945–978 页，2017 年。[4] N. Saldi、T. Linder 和 S. Y¨uksel，离散时间随机控制的有限近似。Springer，2018 年。[5] A. Kara 和 S. Yuksel，“部分可观测马尔可夫决策过程中的有限记忆反馈策略的近似最优性”，*Journal of Machine Learning Research*，第 23 卷，第 11 期，第 1–46 页，2022 年。[6] 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”，*Operations Research*，第 42 卷，第 3 期，第 439–455 页，1994 年。[7] J. Subramanian 和 A. Mahajan，“部分可观测系统的近似信息状态”，载于*2019 IEEE* 第 *58* 届决策与控制会议 *(CDC)*，IEEE，2019 年，第 1629–1636 页。[8] M. Kim、J. Cai 和 K. Sreenath，“Roverfly：四旋翼载荷系统的鲁棒且多用途隐式混合控制”，*arXiv* 预印本 *arXiv:2509.11149*，2025 年。[9] J. Cai、V. Sangli、M. Kim 和 K. Sreenath，“仿鸟扑翼机器人的基于学习的轨迹跟踪”，载于 *2025* 年美国控制会议 *(ACC)*，IEEE，2025 年，第 430–437 页。[10] A. D. Kara 和 S. Y¨uksel，“部分可观测最优随机控制：正则性、最优性、近似与学习”，载于*2024* 年 *IEEE* 第 *63* 届决策与控制会议 *(CDC)*，IEEE，2024 年，第 6709–6721 页。