arXiv:2303.00694v1 [cs.LG] 1 Mar 2023

基于模型的强化学习中的懒惰之美：统一目标与算法

Anirudh Vemula [1] **Yuda Song** [2] **Aarti Singh** [2] **J. Andrew Bagnell** [1 2] **Sanjiban Choudhury** [3]

摘要

我们提出了一种新方法，用于应对基于模型的强化学习（MBRL）中的两项根本挑战：在学习得到的模型中反复寻找优质策略所带来的计算开销，以及模型拟合与策略计算之间的目标失配。我们提出的“懒惰”方法利用一种新颖的统一目标——模型中的优势性能差异，来刻画在真实动态下学习得到的策略与专家策略之间的性能差异。该目标表明，在探索分布下优化学习得到的模型中的期望策略优势，足以完成策略计算；与传统规划方法相比，这能显著提高计算效率。此外，该统一目标在模型拟合中采用价值矩匹配项，使其与模型在策略计算中的使用方式相一致。我们提出了两种无遗憾算法来优化这一目标，并通过模拟基准展示了相较于现有基于模型的强化学习方法所获得的统计与计算收益。

1. 引言

基于模型的强化学习（MBRL）方法在真实世界中的应用前景广阔，因为与无模型强化学习方法相比，它们通常所需的现实世界交互次数要少得多（Schrittwieser 等人，2020年；Hafner 等人，2023年）。其核心思想是：不同于无模型强化学习直接根据真实世界数据计算策略，我们可以执行以下迭代过程（Sutton & Barto，2018年）：利用学习得到的策略，基于迄今收集的数据拟合一个能够准确预测动力学的模型。随后，通过在学习得到的模型中进行最优规划，并利用规划得到的策略在真实世界中收集更多数据。重复这一过程，直到学得令人满意的策略。Ross & Bagnell (2012) 等理论研究表明，在某些条件下，例如能够获得良好的探索分布和足够丰富的模型类，这一过程可以以统计上高效的方式找到近最优策略；而其在实践中表现出色，也验证了这一点。

然而，上述流程面临两个主要挑战。在每次迭代中，策略计算步骤都依赖于求解一个计算开销很大的问题：在学习得到的模型中找到最优策略。这可能需要在模型中进行数量随任务时域呈指数增长的交互次数（Kearns 等人，1999年）。此外，包括 Ross & Bagnell（2012）；Jiang （2018年）；Vemula 等人（2020年）在内的以往研究表明，在学习得到的模型中进行最优规划，可能会产生利用模型不准确性的策略，从而阻碍快速学习和统计效率的提升。

第二个挑战涉及模型拟合与策略计算之间的目标失配，这是近期文献广泛研究的问题（Farahmand 等人，2017；Lambert 等人，2020年）。最小化预测误差这一模型拟合目标，与最大化基于模型得出的策略在真实世界中的性能这一目标并不一定相关。这导致用于拟合模型的目标与通过规划计算策略时模型的使用方式之间存在目标失配。当模型类不可实现时，这一问题会更加严重，也就是说，模型类中没有任何模型能够完美描述真实动力学；而这在真实世界任务中十分常见（Joseph 等人，2013）。

在这项工作中，我们提出了一种新的分解方法，用于分解真实动力学下学习得到的策略与专家策略之间的性能差异，并将其命名为模型中的优势性能差异。由此得到的统一目标，为现有基于模型的强化学习流程带来了两项重大改进。我们不再于每次迭代时计算学习得到的模型中的最优策略，而是在探索分布下优化模型中的期望策略优势；这种方法只需要在模型中进行与任务时域多项式相关的交互次数。对于

[1]Aurora Innovation，美国宾夕法尼亚州匹兹堡 [2]卡内基梅隆大学计算机科学学院，美国宾夕法尼亚州匹兹堡 [3]康奈尔大学计算机科学系，美国纽约州伊萨卡。通讯作者：Anirudh Vemula <vvanirudh@gmail.com>。

模型拟合方面，我们的新目标根据学习得到的模型中的价值函数，衡量预测的下一状态与观测到的下一状态之间的相似度。这确保模型在对策略计算至关重要的状态上得到准确更新，同时允许模型在无关状态上存在误差。因此，我们提出的统一目标鼓励基于模型的强化学习流程的两个步骤都采取“惰性”做法，从而同时解决**计算开销** 和**目标函数不匹配**这两项挑战。

本文的贡献如下：

- 我们提出了一个统一的基于模型的强化学习目标函数，该目标函数在策略计算方面具有更高的计算效率，并解决了模型拟合中的目标失配问题。
- 两种算法利用了所提出目标函数中的懒惰特性，其性能界比Ross 和 Bagnell（2012）更紧。
- 通过模拟基准测试进行的实证表明，与现有的基于模型的强化学习方法相比，我们提出的算法在统计和计算方面均有所提升。

2. 相关工作

基于模型的强化学习与最优控制已在文献中得到广泛研究，相关研究由来已久（Ljung, 1998; Morari 和 Lee, 1999; Sutton, 1991）。近期研究在低维状态空间任务（Levine 和 Abbeel, 2014; Chua 等人, 2018; Schrittwieser 等人, 2020）和高维状态空间任务（Hafner 等人, 2020; Wu 等人, 2022）中均取得了显著成果。研究者也开展了理论研究，分析基于模型的方法所能提供的性能保证和样本复杂度（Abbasi-Yadkori 和 Szepesv´ari, 2011; Ross & Bagnell, 2012; Tu 和 Recht, 2019; Sun 等人, 2019a）。然而，以往研究普遍要求在每次迭代中根据学习得到的模型计算最优策略，所用方法从值迭代（Azar 等人, 2013 ）到黑盒策略优化（ Kakade 等人 , 2020; Song 和 Sun, 2021）不等。本文表明，无须在学习得到的模型中计算最优策略，并提出了一种计算效率更高且不损害性能保证的替代方法。

使用探索分布的强化学习 在本研究以及此前的相关工作中，例如 Kakade 和 Langford (2002); Bagnell 等人 (2003年); Ross & Bagnell (2014年)，我们假设可以访问一种探索分布，从而利用任务的任何先验知识，快速学得优质策略。在本研究中，我们也在基于模型的强化学习框架内采用了类似的无模型策略搜索算法。近期混合强化学习领域的研究（Rajeswaran 等人，2017; Vecerik 等人，2017; Nair 等人，2018年; Hester 等人，2018年; Xie 等人，2021b; Song 等人，2022）研究了一个相关设定，其中既有离线数据集，也能在线访问环境并与之交互。然而，这些研究大多采用无模型设定，并要求离线数据集由专家策略收集；而我们的基于模型的设置只要求探索分布覆盖专家分布。

基于模型的强化学习中的目标失配近期研究（Farahmand 等人，2017; Lambert 等人，2020年; Voloshin 等人，2021年; Ey- senbach 等人，2021年）指出了基于模型的强化学习中的目标不匹配问题：训练目标（寻找最大似然估计模型）与真正的目标（在真实世界中寻找最优策略）并不一致。为了解决这一问题，多项研究提出在模型拟合过程中纳入价值感知目标。Farahmand 等人（2017）；Grimm 等人（2020年）；Voloshin 等人（2021年）提出寻找能够在预先定义的一组价值函数和策略上正确预测期望后继价值的模型。Modhe 等人（2021年）将学习得到的策略下的模型优势用作模型拟合目标，并通过规划计算策略。Ayoub 等人（2020年）提出了类似的方法，其中模型拟合采用价值目标回归目标，并利用乐观性，只选择与迄今收集的数据一致的模型。然而，他们的方法假设模型类具有可实现性，并且需要利用构造的模型集合求解一个乐观规划问题。与此不同，我们从第一性原理出发，为策略学习和模型学习提出了一个统一目标；该目标既具有价值感知能力，也可通过无遗憾算法进行优化。

3. 预备知识

我们假设真实世界遵循一个无限时域折扣马尔可夫决策过程（MDP）$(\mathcal{S},\ \mathcal{A},\ M^\star,\ \omega,\ c,\ \gamma)$，其中 $\mathcal{S}$ 是状态空间，$\mathcal{A}$ 是动作空间，$M^\star\colon \mathcal{S}\times\mathcal{A}\to\Delta(\mathcal{S})$ 是转移动力学，$c\colon \mathcal{S}\times\mathcal{A}\to[0,1]$ 是代价函数，γ 是折扣因子，$\omega\in\Delta(\mathcal{S})$ 是初始状态分布，其中 $\Delta(\mathcal{S})$ 定义了集合 $\mathcal{S}$ 上的概率分布集合。真实动力学 $M^\star$ 未知，但我们可以在真实世界中收集数据。我们假设代价函数 c 已知，但我们的结果可以扩展到 c 未知的情况。

对于任意策略 $\pi\colon \mathcal{S}\to\Delta(\mathcal{A})$，我们将 $D^h_{\omega,\pi}$ 记为时刻 h 的状态 - 动作分布，条件是从 ω 中采样一个初始状态，并在 $M^\star$ 中执行策略 π，直至时刻 $h-1$。这可以通过 $D_{\omega,\pi}=(1-\gamma)\sum_{h=1}^{\infty}\gamma^{h-1}D^h_{\omega,\pi}$推广，它表示无限时域上的状态-动作分布。类似地，我们用记号 $d_{\omega,\pi}$表示无限时域上的状态分布。我们将策略 π 在任意转移

算法1基于模型的强化学习元算法

要求： 迭代次数 T、模型类 $\mathcal{M}$、策略类 Π、探索分布 ν

1: 初始化模型 $M_1 \in \mathcal{M}$
2: 使用**ComputePolicy**计算策略 $\hat{}\pi_1$
3: **对** $t = 1, \ldots, T$ **do**
4: 在 $M^\star$ 中通过滚动执行 $\hat{}\pi_t$ 或采样来收集数据从 ν 中（等概率）并添加到数据集 $\mathcal{D}_t$
5: 拟合模型 $\hat{M}_{t+1}$ 到 $\mathcal{D}_t$，使用**拟合模型**
6: 使用**ComputePolicy**计算策略 $\hat{}\pi_{t+1}$
7: **结束循环**
8: **返回**策略序列 $\{\hat{}\pi_t\}_{t=1}^{T+1}$

函数 M 下的价值函数记为 $V_M^\pi(s)$，状态-动作价值函数记为 $Q_M^\pi(s,a) = c(s,a) + \mathbb{E}_{s' \sim M(s,a)} V_M^\pi(s')$，性能定义为 $J_M^\omega(\pi) = \mathbb{E}_{s\sim\omega}[V_M^\pi(s)]$。目标是找到使 π 最小的策略 $\pi^\star = \operatorname{argmin}_{\pi\in\Pi} J_{M^\star}^\omega(\pi)$。

与Ross & Bagnell (2012) 类似，我们的方法假设可以访问状态-动作探索分布 ν 并从中采样，同时能够保证：对于任何状态-动作分布接近 ν的策略，其遗憾都很小。如果 ν 接近$D_{\omega,\ \pi^\star}$，那么我们的方法就能保证近乎最优的性能。在实践中，良好的探索分布通常可以通过专家示范、领域知识，或我们希望系统遵循的期望轨迹获得。

3.1. 基于模型的强化学习框架

Ross & Bagnell （2012）提出的基于模型的强化学习框架被描述为一个元算法，见算法1。从一个探索分布开始，在每次迭代中，我们使用学习得到的策略和探索分布收集数据，基于迄今收集到的数据拟合模型，并使用新学习得到的模型计算策略。请注意，算法1中的模型拟合和策略计算过程目前尚未具体说明。

为了理解为何 算法1 会得到一个在真实世界中性能良好的策略 $M^\star$，让我们重新审视 Ross & Bagnell（2012）提出的目标函数。该目标函数源于在基于模型的强化学习分析中应用一项重要工具——（Kearns & Singh，2002）提出的模拟引理（见引理A.1）两次，以计算任意两个策略 $\hat{}\pi$、$\pi^\star$ 在真实世界 $M^\star$中的性能差异，而这正是我们希望优化的量。换言之，我们希望找到一个策略 $\hat{}\pi$，使其在 $M^\star$中的性能接近专家策略 $\pi^\star$在 $M^\star$中的性能。

引理 3.1（通过模型规划的性能差异）。 对于任意初始状态分布 ω，策略 $\hat{}\pi$，$\pi^\star$，$\hat{}$以及转移函数 $M, M^\star$，我们有：

$$(1-\gamma)[J_{M^\star}^\omega(\hat{\pi}) - J_{M^\star}^\omega(\pi^\star)] =$$

$$\underbrace{(1-\gamma)\mathop{\mathbb{E}}_{s\sim\omega}[V_{\hat{M}}^{\hat{\pi}}(s) - V_{\hat{M}}^{\pi^\star}(s)]}_{\textit{Performance difference in the Model}} \tag{1}$$

$$+ \underbrace{\mathop{\gamma\mathbb{E}}_{\substack{(s,a)\sim D_{\omega,\hat{\pi}} \\ s'\sim M^\star(s,a)}}[V_{\hat{M}}^{\hat{\pi}}(s') - \mathop{\mathbb{E}}_{s''\sim\hat{M}(s,a)}[V_{\hat{M}}^{\hat{\pi}}(s'')]]}_{\textit{Value difference on states visited by learned policy}} \tag{2}$$

$$+ \underbrace{\mathop{\gamma\mathbb{E}}_{\substack{(s,a)\sim D_{\omega,\pi^\star} \\ s'\sim M^\star(s,a)}}[\mathop{\mathbb{E}}_{s''\sim\hat{M}(s,a)}[V_{\hat{M}}^{\pi^\star}(s'')] - V_{\hat{M}}^{\pi^\star}(s')]}_{\textit{(Expert) Value difference on states visited by expert}} \tag{3}$$

上述引理告诉我们，性能差异可以分解为三项之和：第（1）项是两个策略在 $\hat{}$学习得到的模型 M中的性能差异；第（2）项和第（3）项则分别刻画了沿着从 $\hat{}\pi$和 $\pi^\star$ 在真实世界 $M^\star$ 中采样得到的轨迹，学习得到的模型与真实世界所导致的下一状态价值差异。通过确保学习得到的策略 $\hat{}\pi$ 在学习得到的模型 M 中实现 $\hat{}$低代价的最优规划 M，例如运行 $\hat{}$最优规划，使得

$$\mathop{\mathbb{E}}_{s\sim\omega}[V_{\hat{M}}^{\hat{\pi}}(s)] - \min_{\pi\in\Pi}\mathop{\mathbb{E}}_{s\sim\omega}[V_{\hat{M}}^{\pi}(s)] \le \epsilon_{oc} \tag{4}$$

如果模型的预测误差较低，则第（2）项和第（3）项可以变得很小。下面的推论通过对这些项应用 H¨older 不等式，将这一点形式化：

推论 3.1。 对于任意初始状态分布 ω、转移$\hat{}$函数 M、$M^\star$ 以及策略 $\pi^\star$，$\hat{}\pi$ 若其满足 （4），则有：

$$\begin{aligned}(1-\gamma)[J_{M^\star}^\omega(\hat{\pi}) - J_{M^\star}^\omega(\pi^\star)] &\le \epsilon_{oc} \\ &+ \gamma\hat{V}_{\max}\mathop{\mathbb{E}}_{(s,a)\sim D_{\omega,\hat{\pi}}}\left\|\hat{M}(s,a) - M^\star(s,a)\right\|_1 \\ &+ \gamma V_{\max}\mathop{\mathbb{E}}_{(s,a)\sim D_{\omega,\pi^\star}}\left\|\hat{M}(s,a) - M^\star(s,a)\right\|_1,\end{aligned}$$

where $\hat{V}_{\max} = \|V_{\hat{M}}^{\hat{\pi}}\|_\infty, V_{\max} = \|V_{\hat{M}}^{\pi^*}\|_\infty$.

大多数基于模型的强化学习方法，包括 Ross& Bagnell (2012)，使用最大似然估计（MLE）算法 2来界定推论 31[1] 中的全变差损失项，并采用最优规划方法来满足方程（4）。结合算法1 与 算法 2以及方程（4），我们可以据此构建一个理解现有基于模型的强化学习方法的框架。

[1]我们还可以通过 Pinsker不等式，使用 KL 散度进一步界定总变差项。随后，在学习得到的模型下最大化观测数据的似然，将最小化 KL 散度。

算法 2 最大似然估计 **拟合模型**$(\mathcal{D}_t, \{\ell_i\}_{i=1}^{t-1})$

要求： 数据 $\mathcal{D}_t$，模型类 $\mathcal{M}$，先前的损失 $\{\ell_i\}_{i=1}^{t-1}$1：定义损失 $\ell_t(M) = \mathbb{E}_{(s,a,s')\sim\mathcal{D}_t} \log M(s'|s,\ a)$ˆ2：使用在线无遗憾算法计算模型 M_{t+1}，例如跟随领头者（FTL）（Hazan, 2019），

$$\hat{M}_{t+1} \leftarrow \operatorname*{argmin}_{M\in\mathcal{M}} \sum_{\tau=1}^{t} \ell_\tau(M).$$

3: **返回** $\hat{M}_{t+1}$

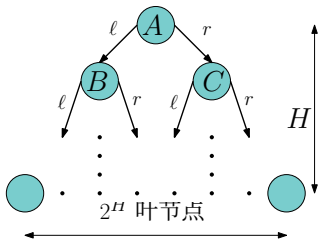

图 *1.* 图中展示了具有两个动作 ℓ 和 r的马尔可夫决策过程，以及真实动力学 $M^\star$。代价 $c(s,a) = \epsilon << 1$ 在任意 $s \neq B$ 下成立，而 $c(B,a) = 1$对任意动作 a均成立。因此，在 A 采取的动作至关重要。模型类 $\mathcal{M}$ 仅包含两个模型：M^{good} 能够正确捕捉 A 处的动力学，但在其他所有地方都会出错；而 M^{bad} 只在 A 处出错，却能在其他所有地方捕捉真实动力学。

3.2. 基于模型的强化学习中的挑战

我们从上一节中得出了两个重要观察，它们直接体现为基于模型的强化学习中的两个基本挑战。我们将借助图 1 来阐释这些挑战。首先，要确保学习得到的策略ˆπ_t 满足（4）中的条件算法1，就需要在每次迭代中，针对学习得到的模型ˆ M_t 执行最优规划。需要注意的是，最优规划可能需要在ˆ M_t中进行数量随有效任务时域 $\frac{1}{1-\gamma}$（Kearns 等人，1999年）呈指数增长的交互。例如，在图1中，求解最优策略需要 $\mathcal{O}(2^H)$ 次运算。我们将其称为**C1：计算开销**挑战。

其次，引理 3.1 表明，优化项（2）和（3）（以及（4））可以保证优化性能差异。然而，我们无法直接优化项（3），因为这需要访问模型中的专家价值函数 $V^{\pi^\star}$ˆ$_M$，但该函数未知。为避免这一问题，基于模型的强化学习方法在推论 3.1中使用模型预测误差为这些项构造上界。这个上界易于优化，但非常松，尤其是因为未知的缩放项$\|V^{\pi^\star}$ˆ$_M\|_\infty$ 可能大到 $\frac{1}{1-\gamma}$。我们将此称为基于模型的强化学习中的 **C2：目标失配**挑战。最小化预测误差的模型拟合目标并不能很好地近似项（2）和（3），而这两项能够通过后继状态的价值，捕捉转移 $(s,\ a)$ 在策略计算中的相对重要性。例如，在图 1中，M^{bad} 的预测误差低于 M^{good}，尽管它在关键状态 A 上出错，而该状态下预测后继状态与真实后继状态之间的价值差很大；相比之下，根据 M^{good}，引理 3.1中的目标函数，表现更好。

4. 模型中的优势性能差异

为克服上一节提出的挑战**C1**和**C2**，我们重新审视引理 3.1 中的性能差异，并为其引入一种新的分解，从而得到一个更易于优化的统一目标函数。这种分解是本文的主要贡献，我们将其命名为模型中的优势性能差异（PDAM）。详细证明见附录 A。

引理 4.1 (模型中的优势性能差异)。 对于任意初始状态分布 ω、策略 ˆπ、 $\pi^\star$、ˆ以及转移函数 M、 $M^\star$，我们有：

$$(1-\gamma)[J^\omega_{M^\star}(\hat{\pi}) - J^\omega_{M^\star}(\pi^\star)] =$$

$$\underbrace{\mathop{\mathbb{E}}_{s\sim d_{\omega,\pi^\star}}[V^{\hat{\pi}}_{\hat{M}}(s)] - \mathop{\mathbb{E}}_{a\sim\pi^\star(s)}[Q^{\hat{\pi}}_{\hat{M}}(s,a)]]}_{\textit{Disadvantage on states visited by expert}} \tag{5}$$

$$+ \underbrace{\gamma\mathop{\mathbb{E}}_{\substack{(s,a)\sim D_{\omega,\hat{\pi}}\\ s'\sim M^\star(s,a)}}[V^{\hat{\pi}}_{\hat{M}}(s') - \mathop{\mathbb{E}}_{s''\sim\hat{M}(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s'')]]}_{\textit{Value difference on states visited by learned policy}} \tag{6}$$

$$+ \underbrace{\gamma\mathop{\mathbb{E}}_{\substack{(s,a)\sim D_{\omega,\pi^\star}\\ s'\sim M^\star(s,a)}}[\mathop{\mathbb{E}}_{s''\sim\hat{M}(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s'')] - V^{\hat{\pi}}_{\hat{M}}(s')]}_{\textit{Value difference on states visited by expert}} \tag{7}$$

上述引理提出了一个用于联合模型与策略学习的新颖统一目标。我们可以做几点重要说明。首先，对项 (5) 进行界定并不需要计算学习得到的模型 M中的最优策略ˆ，这与项 (1) 不同。相反，我们需要一个策略，使其在 $M^\star$中沿 $\pi^\star$ 轨迹采样的状态上，相对于学习得到的模型中的最优策略具有较小的“劣势”。这一劣势项曾被广泛用作若干无模型方法中的策略搜索目标（Kakade 和 Langford, 2002；Bagnell 等人，2003年；Ross & Bagnell, 2014年）。若能获取覆盖D_ω的探索分布 ν，$\pi^\star$，计算这样的策略所需的计算量关于有效任务时域呈多项式级（Kakade，2003年）；相比之下，式 (4) 所需的计算量可能随任务时域呈指数级增长。换言之，在策略计算步骤中采取“懒惰”做法，我们便能解决挑战C1，同时仍优化ˆπ与 $\pi^\star$之间的性能差异。

算法3 最小化劣势 **ComputePolicy**(M_t)

要求： 探索分布 ν、学习得到的模型 M_t、策略类 Π。1：通过对状态进行代价敏感分类，找到 $\hat{\pi}_t \in \Pi$；这些状态从 $\hat{}$中采样，且位于 ν 中的 M_t （Ross & Bagnell，2014年），使得

$$\mathop{\mathbb{E}}_{s\sim\nu}\left[V^{\hat{\pi}_t}_{\hat{M}_t}(s) - \min_{a\in\mathcal{A}}[Q^{\hat{\pi}_t}_{\hat{M}_t}(s,a)]\right] \leq \epsilon_{po} \quad (8)$$

2: **返回** $\hat{\pi}_t$

其次，尽管 PDAM 在模型拟合项（引理 3.1 中的（6）和（7）看起来相似，但有一个关键差异：在新的目标函数中，我们只需要 $V^{\pi}_{\hat{M}}$，这可以通过在学习得到的模型中使用任何策略评估方法来计算（Sutton & Barto，2018年）。另一方面，引理 3.1 需要访问 $V^{\pi^\star}{}_{\hat{M}}$，其中 $\pi^\star$ 未知，因此我们必须在推论 3.1 中利用模型预测误差对目标函数进行上界估计。优化预测误差要求学习得到的模型在所有地方都同样准确地捕捉动力学。然而，统一目标PDAM为模型拟合提供了一种“懒惰”的替代方案：通过关注对策略计算至关重要的状态转移（即具备价值感知能力），解决挑战**C2**。

5. LAMPS：惰性基于模型的策略搜索

本节介绍我们的首个算法 LAMPS，它利用统一目标 PDAM解决挑战**C1**。算法3 沿着探索分布进行策略优化，这与以往的策略搜索方法类似（Kakade 和 Langford，2002年；Bagnell 等人，2003年；Ross & Bagnell，2014年）。注意，(8) 只要求在从探索分布中采样的状态上执行单步代价敏感分类 ν，因此算法3的计算效率很高。另一方面，最优规划要求在学习得到的策略下所有状态上最小化劣势 $d_{\omega,\,\hat{\pi}_t}$[2] 而学习得到的策略会随着更新而改变，导致对时域呈指数级依赖。

为了理解 算法3 如何帮助优化 PDAM，我们对项 (6) 和 (7) 应用 Hölder 不等式，从而得到 PDAM 的上界：

推论 5.1。对于任意初始状态分布 ω、转移函数 M，$M^\star$以及策略 π，$\pi^\star$满足 π 符合 (8)，

我们有：

$$\begin{aligned}(1-\gamma)[J^{\omega}_{M^\star}(\hat{\pi}) - J^{\omega}_{M^\star}(\pi^\star)] &\leq \epsilon_{po} \\ &+ \gamma\hat{V}_{\max}\mathop{\mathbb{E}}_{(s,a)\sim D_{\omega,\hat{\pi}}}\|\hat{M}(s,a) - M^\star(s,a)\|_1 \\ &+ \gamma\hat{V}_{\max}\mathop{\mathbb{E}}_{(s,a)\sim D_{\omega,\pi^\star}}\|\hat{M}(s,a) - M^\star(s,a)\|_1\end{aligned}$$

where $\hat{V}_{\max} = \|V^{\hat{\pi}}_{\hat{M}}\|_\infty$.

推论5.1 表明，只需对现有的基于模型的强化学习方法稍作修改：在算法 2 中使用基于最大似然估计的算法3 进行模型拟合，并在算法1的框架中进行策略计算。我们将这个新算法称为 LAMPS。请注意，通过使用模型预测误差对引理 4.1 中的模型拟合项进行上界估计，LAMPS 只能解决挑战 **C1**，但无法解决**C2**。

我们的算法只需要一个探索分布ν，该分布能够覆盖专家策略的状态-动作分布$D_{\omega,\pi^\star}$，如第3节所述。为刻画ν 对$D_{\omega,\pi^\star}$的覆盖程度，我们仿照 Ross & Bagnell（2012）定义覆盖系数 $\mathcal{C} = \sup_{s,a}\frac{D_{\omega,\pi^\star}(s,a)}{\nu(s,a)}$。现在，我们给出 LAMPS 的遗憾界；其证明见附录 A。

定理 5.1. 令 $\{\hat{\pi}_t\}_{t=1}^T$ 为 LAMPS 返回的策略序列。我们有：[4]

$$\frac{1}{T}\sum_{t=1}^{T} J^{\omega}_{M^\star}(\hat{\pi}_t) - J^{\omega}_{M^\star}(\pi^\star) \leq \tilde{O}\left(\mathcal{C}\epsilon_{po} + \frac{\mathcal{C}\hat{V}_{\max}}{1-\gamma}\left(\sqrt{\epsilon^{KL}_{model}} + \frac{1}{\sqrt{T}}\right)\right),$$

其中，$\hat{V}_{\max} = \|V^{\hat{\pi}}_{\hat{M}}\|_\infty$，$\mathcal{C} = \sup_{s,a}\frac{D_{\omega,\pi^\star}(s,a)}{\nu(s,a)}$ 是覆盖系数，ϵ_{po} 是策略优势误差，而 $\epsilon^{KL}_{model} = \min_{M\in\mathcal{M}}\mathbb{E}_{s,a\sim\tilde{\mathcal{D}}_T}KL(M(s,a)\ M^*(s,a))$ 是无关模型误差[5]。

相比之下，Ross &Bagnell （2012）的遗憾界为

$$\tilde{O}\left(\epsilon_{oc} + \frac{\mathcal{C}\max\{V_{\max},\hat{V}_{\max}\}}{1-\gamma}\left(\sqrt{\epsilon^{KL}_{model}} + \frac{1}{\sqrt{T}}\right)\right),$$

其中 $V_{\max} = \|V^{\pi^*}{}_{\hat{M}}\|_\infty$ 难以优化，因为 $\pi^\star$ 未知，且可能与有效时域 $\frac{1}{1-\gamma}$一样大。另一方面，我们的界只包含 $\hat{V}_{\max} = \|V^{\hat{\pi}}_{\hat{M}}\|_\infty$。

[2]要证明这一点，将性能差异引理（Kakade 和 Langford，2002年）应用于 引理 3.1中的项 (1)。

[3]请注意，离线和混合强化学习设置中也有覆盖概念（Xie 等人，2021a；Song 等人，2022），但其覆盖系数并不直接适用于基于模型的设置。[4]我们使用 $\tilde{O}$ 来省略各项中的对数依赖。[5]这里我们用KL（·，·）表示 Kullback–Leibler 散度，并将训练数据分布记为 $\mathcal{D}_T = \frac{1}{T}\sum_{t=1}^{T}\mathcal{D}_t$

算法 4 矩匹配**拟合模型**($\mathcal{D}_t$, $\{\ell_i\}_{i=1}^{t-1}$)

需要： 数据 $\mathcal{D}_t$、模型类 $\mathcal{M}$、先前的损失$\{\ell_i\}_{i=1}^{t-1}$、强凸正则化器 $\mathcal{R}$1：定义损失

$$\ell_t(M) = \mathop{\mathbb{E}}_{(s,a,s')\sim\mathcal{D}_t} \left| V_{\hat{M}_t}^{\hat{\pi}_t}(s') - \mathop{\mathbb{E}}_{s''\sim M(s,a)} [V_{\hat{M}_t}^{\hat{\pi}_t}(s'')] \right| \tag{9}$$

ˆ2: 使用在线无遗憾算法计算模型 M_{t+1}，例如跟随正则化领导者(Hazan，2019)，

$$\hat{M}_{t+1} \leftarrow \operatorname*{argmin}_{M\in\mathcal{M}} \sum_{\tau=1}^{t} \ell_\tau(M) + \mathcal{R}(M). \tag{10}$$

3: **返回** $\hat{M}_{t+1}$

该策略针对从算法3中采样的状态ˆ进行优化 ν。因此，我们预计在某些情况下，$V_{\max}$会远小于 $V_{\max}$[6]，从而在定理 5.1中得到更紧的遗憾界。因此，尽管 LAMPS 主要解决了挑战**C1** 并带来计算收益，但正如我们在第7节的实验中所示，我们在实践中也观察到了统计收益。另一个关键区别是，LAMPS 遗憾界中的策略优化误差包含覆盖系数 $\mathcal{C}$，这表明 LAMPS 对探索分布的质量 ν相对更加敏感。我们在附录C.2中给出了一个例子。

6. LAMPS-MM：通过价值矩匹配进行惰性基于模型的策略搜索

上一节介绍了算法 LAMPS。尽管它的计算效率高于现有的基于模型的强化学习方法，但并未充分发挥我们提出的统一目标PDAM的优势。LAMPS 优化的是推论 5.1 中给出的目标函数，该目标函数是统一目标的一个弱上界（如 第3节.2所述）。一个自然的问题是，是否存在一种算法，无需构造弱上界，就能直接优化统一目标。本节将介绍这样一种算法。其核心思想是将问题表述为矩匹配问题（Sun 等人，2019b；Swamy等人，2021年），让学习得到的模型利用价值矩，在期望意义上匹配真实动力学。

算法 4 在给定的数据集上，最小化真实后继状态与预测后继状态之间的价值矩差异

[6]请注意，在最坏情况下，ˆ $V_{\max}$ 也可能高达 $\frac{1}{1-\gamma}$。ˆ不过，我们在第7节展示了一项实验，其中 $V_{\max}$ 在实践中小于 $V_{\max}$。

状态转移。目标函数 ℓ_t 在（9）中对（6）和（7）中的项给出了上界，并采用带有强凸正则化器的跟随正则化领导者（FTRL）方法$\mathcal{R}$（M）使我们能够获得无遗憾保证。我们将这种新的基于模型的强化学习算法称为 LAMPS-MM：它在算法 1 的框架下，使用基于价值矩匹配的算法 4 进行模型拟合，并使用算法3进行策略计算。算法 1 称为 LAMPS-MM。

LAMPS-MM 通过优化 PDAM，不会受到挑战 **C2**，因为模型拟合目标会聚焦于关键状态，在这些状态下，动作中的任何错误都可能导致巨大的价值差异，从而帮助学习有助于策略计算的模型。回到 图 1中的例子，LAMPS-MM 会选择 $M_{\text{好的}}$ 而非 M^{bad}，因为后者会在状态 ℓ_t (9) 处产生较高的损失 A。因此，LAMPS-MM 解决了挑战 **C1** 和 **C2**。如其更紧的遗憾界所示，这带来了更高的统计效率：

定理 6.1。 令 $\{\hat{}\pi_t\}_{t=1}^T$ 为返回策略序列 LAMPS-MM，则有：

$$\frac{1}{T}\sum_{t=1}^{T} J_{M^\star}^{\omega}(\hat{\pi}_t) - J_{M^\star}^{\omega}(\pi^\star) \leq \tilde{O}\left(\mathcal{C}\epsilon_{po} + \frac{\mathcal{C}}{1-\gamma}\left(\epsilon_{model}^{mm} + \frac{1}{\sqrt{T}}\right)\right),$$

其中，$\epsilon_{model}^{mm} = \min_{M\in\mathcal{M}} \frac{1}{T}\sum_{t=1}^{T} \ell_t(M)$ 为无关模型误差，ϵ_{po} 为策略优势误差，而 $\mathcal{C} = \sup_{s,a} \frac{D_{\omega}}{\nu_{(s,a)}}$ 为覆盖系数。

将上述遗憾界与 LAMPS 的遗憾界进行比较，见 定理 5.1，我们通过消除对 ˆ $V_{\max}$[7] 的依赖改进了该界。正如第 5 节所述，在最坏情况下，该项可能高达 $\frac{1}{1-\gamma}$。

尽管 LAMPS-MM 具有统计优势，但其实际实现较为困难，因为估计损失 ℓ_t (9) 需要在学习得到的模型中评估策略，而在大型马尔可夫决策过程中，这只能通过近似来实现。以往研究（Ayoub 等人，2020年；Grimm 等人，2020年；Farahmand 等人，2017）提出用于模型拟合的价值感知目标时，也遇到了类似困难。我们指出几种算法 4 在实际中可实现的情形。首先，对于有限马尔可夫决策过程，我们可以精确评估策略，从而避免这一困难。其次，对于具有二次代价的线性动力系统等马尔可夫决策过程，如果能够以闭式形式计算策略价值，我们便可以

[7]请注意，ϵ_{model}^{mm} 隐式依赖于价值函数。不过，如果模型类 $\mathcal{M}$ 足够丰富，我们可以预期这一误差会很小。例如，如果 $\mathcal{M}$ 包含 $M^\star$，那么这一误差将趋于零，从而得到一个不依赖于 ˆ $V_{\max}$的遗憾界；而定理 5.1中的遗憾界即使在可实现情形下，仍然依赖于 ˆ $V_{\max}$。

估计 ℓ_t，并使用基于梯度的优化方法，在优化问题（10）中找出模型类中的最佳模型。我们在第7节的实验中展示了这两种情形。还需注意的是，使用批量算法（如梯度下降）求解（10）需要汇总数据 $\mathcal{D}_t$ 以及价值函数 V^{π_t, M_t} 在算法1的各次迭代中。这种做法的优点在于，我们的方法不需要计算模型变化如何 M 影响 $\ell_t(M)$ 中所用价值估计的梯度（参见（9）），而这类梯度往往很难计算。

为完整起见，我们还在附录B.3中给出了 LAMPS-MM 的有限样本分析。请注意，我们略去了 LAMPS 的有限样本分析，因为其分析形式与 Ross & Bagnell (2012) 相同。

7. 实验

本节中，我们在五个不同领域开展实验，将所提出的算法与基线进行比较。对于基线，我们将 Ross& Bagnell (2012) 的方法作为对比，该方法使用最大似然估计进行模型拟合，并通过最优规划计算策略，我们将此基线称为 SysID。此外，我们还采用 MBPO（Janner 等人，2019），并设计了它的若干变体，使用与 SysID 类似的探索分布，以确保与所提出的算法进行公平比较。实验设置、代价函数和基线实现的详细信息见附录C[8]。

7.1. 直升机

在 Ross & Bagnell (2012) 提出的这一领域中，我们将 LAMPS 与 SysID 进行比较。任务目标是在存在噪声的情况下，使直升机在动力学未知时跟踪期望轨迹。对于两种方法，我们都使用从期望轨迹采样的探索分布 ν。对于 SysID 的最优规划，我们运行 iLQR（Li & Todorov，2004）直至收敛。对于 LAMPS 的算法3 实现，我们只运行一次 iLQR 迭代：将前向传播替换为期望轨迹，并在其上运行一次 LQR 反向传播，以计算策略。有关动力学、代价函数以及 SysID 和 LAMPS 实现的详细说明，请参阅附录C.1.1。

图 2（a）表明，在真实世界数据量和探索分布相同的情况下，LAMPS 学到的策略优于 SysID，这体现了统计收益。为检验我们的假设，即这是由于 LAMPS 的遗憾界更紧（当 $V_{\max} < V_{\max}$ 时），图 2（c）展示了两种方法学到的策略 π 在学习得到的模型 M 中的表现，以及专家在 M 中的表现。

[8]我们所有实验的代码均可在以下地址找到：https://github.com/vvanirudh/LAMPS-MBRL。

注意到，LAMPS 和 SysID 都能够优化 $V^{\hat{\pi}, M}$，但专家的表现 $V^{\pi^\star, M}$ 并未得到同等程度的优化，因此与 LAMPS 相比，SysID 的遗憾界更弱。图 2(b) 展示了 LAMPS 在计算方面的优势。我们绘制了各方法调用线性二次调节器求解器（最昂贵的操作）的次数，可以看出，仅针对探索分布进行优化，LAMPS 使用的计算量就显著少于 SYSID。

7.2. WideTree

我们采用了 Ay- oub 等人（2020年）提出的有限马尔可夫决策过程领域的一种变体，它与图 1 非常相似 $H = 3$。模型类由两个模型组成：M^{good} 和 $M^{\text{错误}}$，如图 1所述。动力学的详细说明请参阅图5和附录 C.1.2。我们比较 LAMPS-MM 和 SYSID。为了根据目前为止的数据，在每次迭代时计算应选择的最佳模型，我们使用 Hedge（Freund 与 Schapire，1997年）更新两个模型上的离散概率分布。图 2(d) 展示了在 SYSID 中使用基于最大似然估计的模型拟合损失、在 LAMPS-MM 中使用价值矩匹配损失时，该分布如何演变。由于 M^{bad} 除根节点外，在所有位置都与真实动力学相匹配，SYSID 最终收敛到一个总是选择坏模型而非好模型的分布；而 LAMPS-MM 会考虑状态转移对计算优质策略的作用，因此收敛到一个更常选择好模型的分布。

7.3. 线性动力系统

在这个实验中，我们采用有限时域内带二次代价的简单线性动力学系统作为研究对象（类似于线性二次调节），其价值函数可以通过闭式形式计算。真实世界中的动力学是时变线性的，而模型类仅包含时不变线性模型，因此无法感知这种变化。代价函数在每个时间步都会惩罚控制输入，而仅在最后一个时间步惩罚状态（中间状态不产生代价）。有关动力学、代价函数和模型拟合损失的详细说明，请参阅附录 C.1.3。图 2(e) 展示了 LAMPS-MM 和 SYSID 的结果。SYSID 试图在每个时间步匹配真实动力学，因此缓慢收敛至专家性能；而 LAMPS-MM 使用价值矩匹配目标，迅速发现只有最终状态会影响代价，并找到一个简单模型，使控制动作能够在时域结束时将状态归零。因此，我们看到，LAMPS-MM 通过感知价值，能够比传统的基于模型的强化学习方法获得显著的统计收益。

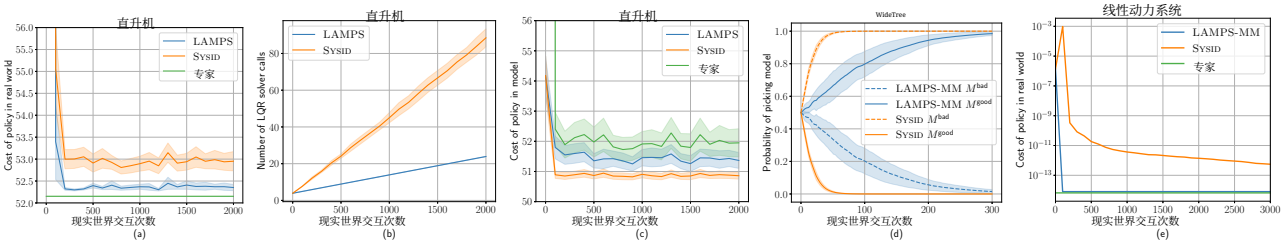

图 *2.* 我们在直升机、WideTree 和线性动力系统（LDS）领域上的实验结果。所有实验均采用 10 个随机种子进行，实线表示均值，阴影区域表示标准误差。在直升机和 LDS 实验中，每次迭代都会向数据集添加 100 个样本，这些样本以相同概率从探索分布和学习得到的策略轨迹中采样。

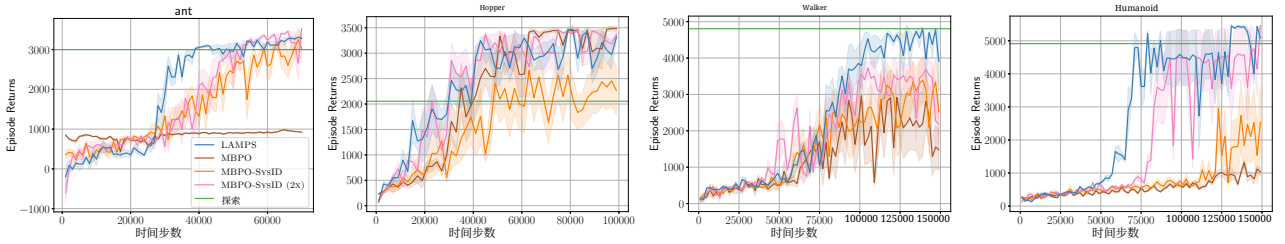

图 *3.* MuJoCo 运动基准上的实验结果。所有实验均采用 5 个随机种子进行，阴影区域表示标准误差。Humanoid 任务使用 50000 个探索样本，其他所有任务均使用 10000 个样本。MBPO 也使用 10000 个随机探索样本进行热启动。基线“探索”表示探索数据集的平均回合回报。

7.4. MuJoCo 运动基准

在这个实验中，我们在标准的稠密奖励 MuJoCo 基准上测试 LAMPS（Brockman 等人，2016）。所有基线方法均基于 MBPO 实现（Janner 等人，2019）。除了 MBPO，我们还设计了一个变体 MBPO-SysID，它也使用来自探索分布的数据进行模型拟合，类似于 SysID。MBPO 中的分支更新在 MBPO-SysID 中作为最优规划的高效替代方案。另一个变体 MBPO-SysID (2x) 将策略更新次数和与学习得到的模型的交互次数都增加一倍，相较于 MBPO-SysID。对于 LAMPS，我们采用与 MBPO-SysID 相同的模型拟合流程，并在策略更新时使用从探索分布中采样的状态，而非当前策略的访问分布。对于探索分布 ν，我们使用一个离线数据集，并在每次迭代时从中采样。有关超参数等实现细节，请参阅附录 C.3。

我们在图 3中展示了结果。与 MBPO 相比，LAMPS 和 MBPO-SYSID 都表现出更高的统计效率，这凸显了探索分布的优势（Ross & Bagnell，2012）。值得注意的是，在所有环境中，LAMPS 都能以少于 MBPO-SYSID 的现实世界交互次数，持续找到更优的策略；这一点在 Humanoid 环境中尤为明显，因为它是所用环境中最困难的。MBPO-SYSID（2X）的性能表明，即使其计算量是 LAMPS 的两倍，LAMPS 在所有实验中的表现仍然优于或不逊于它。这凸显了 LAMPS 在计算和统计效率方面的优势。

7.5. 迷宫

我们的最后一项实验以 PointMaze 环境（Fu 等人，2020年）为领域，考察 LAMPS 在稀疏奖励任务中的性能。我们仅使用离线数据集中的一小部分作为探索分布，因此覆盖范围有限，专家轨迹数量也较少。更多详情请参见附录 C.1.4。由于 LAMPS 在模型拟合和策略计算阶段都会使用探索分布，我们预计它会优于 MBPO-SYSID；后者仅在模型拟合时使用该分布，而稀疏奖励环境需要智能探索。图 4 证实了我们的假设：LAMPS 的表现显著优于 MBPO-SYSID。由于策略计算仅围绕探索分布进行，LAMPS 不会利用学习得到的模型在其他区域的不准确性，并能迅速收敛到良好的策略。

8. 讨论

在本工作中，我们为基于模型的强化学习引入了一种新的统一目标函数。所提出的目标函数旨在提高策略计算的计算效率，并缓解模型拟合中的目标函数不匹配问题。此外，我们提出了两种无遗憾算法，即 LAMPS 和

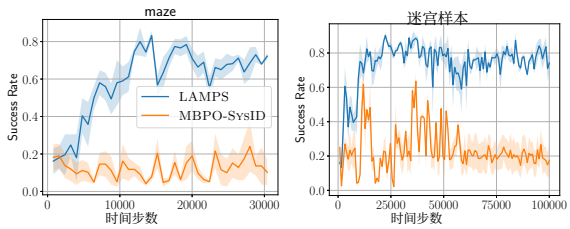

图 *4.* D4RL PointMaze（大型）环境中的结果。我们分别使用 10000 个探索样本（左）和 50000 个样本（右）。在这两种设置下，尽管两种方法都没有采用显式的探索方案（例如探索奖励），我们的算法仍展现出更强的探索能力。结果取 5 个随机种子的平均值，阴影区域表示标准误差。

LAMPS-MM，它们利用所提出的目标函数，并通过在模拟基准上获得的统计和计算收益展示了其有效性。

然而，值得注意的是，尽管 LAMPS 的实现相对简单，但 LAMPS-MM 可能难以应用于难以进行精确策略评估的大型马尔可夫决策过程。此外，两种算法都对探索分布 ν 的质量较为敏感，并且只能保证针对状态-动作分布接近 ν 的策略具有较小的遗憾。因此，如果我们对任务的先验知识足以帮助设计出良好的探索分布，就可以预期这些算法能够计算出良好的策略。

一项有趣的未来工作是将其扩展到潜在模型设定，即学习底层潜在状态的动力学。在这种设定下，典型的最大似然估计模型拟合目标直观上并不适用，因为我们无法观测底层状态，只能获取原始观测。我们还希望探究是否存在一种“双重稳健”版本，结合 SYSID 和 LAMPS 的优势，使我们能够利用良好的探索分布，或计算开销低廉的最优规划器。

参考文献

Abbasi-Yadkori, Y. 和 Szepesv´ari, C. 线性二次系统自适应控制的遗憾界。载于第*24*届学习理论年会论文集，第 1–26 页。JMLR研讨会与会议论文集，2011年。

Abbeel，P. 和 Ng，A. Y. 强化学习中的探索与学徒式学习。载于 Raedt，L. D. 和 Wrobel，S.（编），机器学习：第二十二届国际会议论文集（*ICML 2005*），德国波恩，*2005*年*8*月*7*日至*11*日，第 119 卷，*ACM* 国际会议论文集系列，第 1–8 页。ACM，2005。doi: 10.1145/1102351.1102352。网址 https://doi.org/10.1145/1102351.1102352。

Ayoub，A.、Jia，Z.、Szepesvari，C.、Wang，M. 和 Yang，L. 基于模型的强化学习与价值目标回归。载于国际机器学习会议，第 463–474 页。PMLR，2020年。

Azar, M. G., Munos, R., 和 Kappen, H. J. 基于生成模型的强化学习样本复杂度的极小极大 PAC 界。机器学习，91(3)：325–349，2013年。

Bagnell, J. A., Kakade, S. M., Ng, A. Y., 和 Schneider, J. G. 通过动态规划进行策略搜索。载于 Thrun, S., Saul, L. K., 和 Sch¨olkopf, B.（编），神经信息处理系统进展*16* [神经信息处理系统，*NIPS 2003*，*2003*年*12*月*8*日至*13*日，温哥华和惠斯勒，不列颠哥伦比亚省，加拿大]，第831–838页。麻省理工学院出版社，2003年。网址 https://proceedings.neurips.cc/paper/2003/hash/3837a451cd0abc5ce4069304c5442c87-Abstract.html。

Bertsekas，D. P. 动态规划与最优控制，第*3*版。Athena Scientific，2005。ISBN 1886529264。网址 https://www.worldcat.org/oclc/314894080。

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., 和 Zaremba, W. OpenAI Gym。*arXiv* 预印本 *arXiv:1606.01540*，2016。

Cesa-Bianchi, N., Conconi, A., 和 Gentile, C. 关于在线学习算法的泛化能力。*IEEE* 信息论汇刊，50(9):2050–2057，2004。

Chua, K.、Calandra, R.、McAllister, R. 和 Levine, S. 使用概率动力学模型，通过少量试验实现深度强化学习。载于神经信息处理系统进展，第 4754–4765 页，2018 年。

Eysenbach, B.、Khazatsky, A.、Levine, S. 和 Salakhutdinov, R. 不再失配：基于模型的强化学习中的联合模型-策略优化。*arXiv* 预印本 *arXiv:2110.02758*，2021年。

Farahmand, A. M.、Barreto, A. 和 Nikovski, D. 面向基于模型的强化学习的价值感知损失函数。载于 Singh, A. 和 Zhu, X. J.（编），第*20*届国际人工智能与统计会议论文集，*AISTATS 2017*，*2017*年*4*月*20*日至*22*日，美国佛罗里达州劳德代尔堡，第54卷，机器学习研究论文集，第1486–1494页。PMLR，2017年。网址：http://proceedings.mlr.press/v54/farahmand17a.html。

Freund, Y. 和 Schapire, R. E. 在线学习的决策理论泛化及其在提升算法中的应用。计算机与系统科学杂志，55(1):119–139，1997年。doi: 10.1006/jcss.1997.1504。网址：https://doi.org/10.1006/jcss.1997.1504。

Fu, J., Kumar, A., Nachum, O., Tucker, G., 和 Levine, S. D4rl：深度数据驱动强化学习数据集。*arXiv* 预印本 *arXiv:2004.07219*，2020年。

Grimm, C., Barreto, A., Singh, S., 和 Silver, D. 基于模型的强化学习的价值等价原则。神经信息处理系统进展，33:5541–5552，2020年。

Hafner, D., Lillicrap, T., Norouzi, M., 和 Ba, J. 使用离散世界模型精通 Atari。*arXiv* 预印本 *arXiv:2010.02193*，2020年。

Hafner, D., Pasukonis, J., Ba, J., 和 Lillicrap, T. 通过世界模型精通多样化领域。*arXiv* 预印本 *arXiv:2301.04104*，2023年。

Hazan, E. 在线凸优化导论。*CoRR*，abs/1909.05207，2019。网址http://arxiv.org/abs/1909.05207。

Hester, T., Vecerik, M., Pietquin, O., Lanctot, M., Schaul, T., Piot, B., Horgan, D., Quan, J., Sendonaris, A., Osband, I., 等。基于演示的深度Q学习。载于*AAAI*人工智能大会论文集，第32卷，2018年。

Janner, M., Fu, J., Zhang, M., 和 Levine, S. 何时信任你的模型：基于模型的策略优化。神经信息处理系统进展，32，2019。

Jiang, N. 使用不完美模型的PAC强化学习。载于McIlraith, S. A.和Weinberger, K. Q.（编），第三十二届*AAAI*人工智能大会论文集（*AAAI-18*）、第*30*届人工智能创新应用大会（*IAAI-18*）及第*8*届*AAAI*教育进展研讨会（*EAAI-18*），美国路易斯安那州新奥尔良，*2018*年*2*月*2*日至*7*日人工智能，第3334–3341页。AAAI出版社，2018年。网址：https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16052。

Joseph, J. M.、Geramifard, A.、Roberts, J. W.、How, J. P.和Roy, N. 使用设定错误的模型类进行强化学习。载于*2013*年*IEEE*机器人与自动化国际会议，德国卡尔斯鲁厄，*2013*年*5*月*6*日至*10*日，第939–946页。IEEE，2013年。doi：10.1109/ ICRA.2013.6630686。网址：https://doi.org/10.1109/ICRA.2013.6630686。

Kakade, S., Krishnamurthy，A., Lowrey, K., Ohnishi, M., 和 Sun，W. 在线非线性控制的信息论遗憾界。神经信息处理系统进展，33：15312–15325，2020年。

Kakade，S. M. 强化学习的样本复杂度。伦敦大学大学学院（英国），2003年。

Kakade，S. M. 和 Langford，J. 近似最优的近似强化学习。载于 Sammut，C. 和 Hoffmann，A. G.（编），机器学习：第十九届国际会议论文集（*ICML 2002*），新南威尔士大学，悉尼，澳大利亚，*2002*年*7*月*8*日至*12*日，第267–274页。摩根·考夫曼，2002。

Kearns，M. 和 Singh，S. 多项式时间内的近似最优强化学习。机器学习，49(2)：209–232，2002。

Kearns，M. J.、Mansour，Y. 和 Ng，A. Y. 通过可复用轨迹在大型部分可观测马尔可夫决策过程中进行近似规划。载于 Solla，S. A.、Leen，T. K. 和 M¨uller，K.（编），神经信息处理系统进展 *12*，*[NIPS*会议，丹佛，科罗拉多州，美国，*11*月*29*日至*12*月 *4*，*1999]*，第1001–1007页。麻省理工学院出版社，1999年。网址 http://papers.nips.cc/paper/1664-approximate-planning-in-large-po m基于可复用轨迹

Lambert, N. O., Amos, B., Yadan, O., 和 Calandra, R. 基于模型的强化学习中的目标不匹配。载于 Bayen, A. M., Jadbabaie, A., Pappas, G. J., Parrilo, P. A., Recht，B., Tomlin, C. J., 和 Zeilinger, M. N.（编），第二届动力学与控制学习年度会议论文集，*L4DC 2020*，在线会议，美国加利福尼亚州伯克利，*2020*年*6*月*11*日至*12*日，第机器学习研究论文集120卷，第761–770页。PMLR，2020年。网址：http://proceedings.mlr.press/v120/lambert20a.html。

Levine, S. 和 Abbeel，P. 在未知动力学下通过引导式策略搜索学习神经网络策略。神经信息处理系统进展，27，2014年。

Li, W. 和 Todorov, E. 非线性生物运动系统的迭代线性二次调节器设计。收录于 Ara´ujo, H., Vieira, A., Braz, J., Encarnaç˜ao, B. 和 Carvalho, M.（编），*ICINCO 2004*，第一届控制、自动化与机器人信息学国际会议论文集´葡萄牙塞图巴尔，*2004*年*8*月*25*日至*28*日，第222–229页。INSTICC出版社，2004。

Ljung, L. 系统辨识。载于信号分析与预测，第163–173页。Springer，1998。

Modhe, N., Kamath, H., Batra, D., and Kalyan, A. 面向基于模型的强化学习的模型优势优化。*arXiv*预印本 *arXiv:2106.14080*，2021年。

Morari, M. 和 Lee, J. H. 模型预测控制：过去、现在与未来。《计算机与化学工程》，23(4-5):667–682，1999年。Nair, A.、McGrew, B.、Andrychowicz, M.、Zaremba, W. 和 Abbeel, P. 通过示范克服强化学习中的探索问题。载于 *2018*年*IEEE*机器人与自动化国际会议（*ICRA*），第6292–6299页。IEEE，2018年。Pineda, L.、Amos, B.、Zhang, A.、Lambert, N. O. 和 Calandra, R. MBRL-lib：用于基于模型的强化学习的模块化库。 *arXiv*预印本 *arXiv:2104.10159*，2021年。Rajeswaran, A.、Kumar, V.、Gupta, A.、Vezzani, G.、Schulman, J.、Todorov, E. 和 Levine, S. 通过深度强化学习和示范学习复杂的灵巧操作。 *arXiv*预印本 *arXiv:1709.10087*，2017年。Ross, S. 和 Bagnell, J. A. 面向基于模型的强化学习的无先验系统辨识。载于第*29*届国际机器学习会议论文集，第1905–1912页，2012。Ross, S. 和 Bagnell, J. A. 通过交互式无悔学习进行强化学习与模仿学习。 *CoRR*，abs/1406.5979，2014年。网址 http://arxiv.org/abs/1406.5979。Schrittwieser, J.、Antonoglou, I.、Hubert, T.、Simonyan, K.、Sifre, L.、Schmitt, S.、Guez, A.、Lockhart, E.、Hassabis, D.、Graepel, T. 等。通过使用学习得到的模型进行规划，掌握Atari、围棋、国际象棋和将棋。《自然》，588(7839)：604–609，2020年。Song, Y. 和 Sun, W. PC-MLP：通过策略覆盖引导探索实现基于模型的强化学习。载于 机器学习国际会议，第9801–9811页。PMLR，2021年。

Song, Y.、Zhou, Y.、Sekhari, A.、Bagnell, J. A.、Krishnamurthy, A. 和 Sun, W. 混合强化学习：同时使用离线数据和在线数据可以提高强化学习的效率。*arXiv*预印本 *arXiv:2210.06718*，2022年。

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. 情境决策过程中的基于模型的强化学习：PAC界及相较于无模型方法的指数级改进。载于学习理论会议，第2898–2933页。PMLR，2019a。

Sun, W., Vemula, A., Boots, B., 和 Bagnell, D. 仅凭观察即可实现可证明高效的模仿学习。载于 Chaudhuri, K. 和 Salakhutdinov, R.（编），《第*36*届国际机器学习会议论文集》，*ICML 2019*，*2019*年*6*月*9*日至*15*日，美国加利福尼亚州长滩，第97卷，《机器学习研究论文集》，第6036–6045页。PMLR，2019b。网址 http://proceedings.mlr.press/v97/sun19b.html。Sutton, R. S. Dyna：一种集学习、规划和反应于一体的架构。*ACM SIGART*公报，2(4):160–163，1991。Sutton, R. S. 和 Barto, A. G.《强化学习：导论》。麻省理工学院出版社，2018。Swamy, G., Choudhury, S., Bagnell, J. A., 和 Wu, S. 论矩与匹配：用于弥合模仿差距的博弈论框架。载于国际机器学习会议，第10022–10032页。PMLR，2021。Tu, S. 和 Recht, B. 基于模型的方法与无模型方法在线性二次型调节器上的差距：渐近视角。载于学习理论会议，第3036–3083页。PMLR，2019。Vecerik, M., Hester, T., Scholz, J., Wang, F., Pietquin, O., Piot, B., Heess, N., Roth¨orl, T., Lampe, T., 和 Riedmiller, M. 利用示范，在稀疏奖励的机器人学问题上进行深度强化学习。*arXiv*预印本 *arXiv:1707.08817*，2017。Vemula, A., Oza, Y., Bagnell, J. A., 和 Likhachev, M. 使用不精确模型进行规划与执行，并提供可证明的保证。载于 Toussaint, M., Bicchi, A., 和 Hermans, T.（编），《机器人学：科学与系统第十六届》，线上活动 / 美国俄勒冈州科瓦利斯，*2020*年*7*月*12*日至*16*日，2020。doi: 10.15607/RSS.2020.XVI.001。网址 https://doi.org/10.15607/RSS.2020.XVI.001。Voloshin, C., Jiang, N., 和 Yue, Y. 极小极大模型学习。载于国际人工智能与统计会议，第1612–1620页。PMLR，2021。Wu, P., Escontrela, A., Hafner, D., Goldberg, K., 和 Abbeel, P. Daydreamer：用于实体机器人学习的世界模型。*arXiv*预印本 *arXiv:2206.14176*，2022。

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., 和 Agarwal, A. 离线强化学习中的贝尔曼一致悲观主义。《神经信息处理系统进展》，34:6683–6694，2021a。

Xie, T., Jiang, N., Wang, H., Xiong, C., 和 Bai, Y. 策略微调：弥合样本高效的离线与在线强化学习之间的差距。神经信息处理系统进展，34:27395–27407，2021b。

A. 证明

A.1. 第3.1节的证明

模拟引理有助于比较任意策略 π在两个模型之间的性能，例如学习得到的ˆ模型 M 和真实模型 M^*:

ˆ**引理A.1** (模拟引理)。 对于任意初始分布 ω、策略 π 和转移函数 M、 $M^\star$，有

$$\begin{aligned} J^{\omega}_{M^\star}(\pi) - J^{\omega}_{\hat{M}}(\pi) &= \mathbb{E}_{s\sim\omega}[V^{\pi}_{M^\star}(s) - V^{\pi}_{\hat{M}}(s)] \\ &= \frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim D_{\omega,\pi}}\left[\mathbb{E}_{s'\sim M^\star(s,a)}[V^{\pi}_{\hat{M}}(s')] - \mathbb{E}_{s''\sim \hat{M}(s,a)}[V^{\pi}_{\hat{M}}(s'')]\right] \end{aligned} \tag{11}$$

证明。 第一个等式由 $J^{\omega}_{M}(\pi)$ 的定义得出，该定义见第3节。为证明第二个等式，我们建立如下递推关系:

$$\begin{aligned} &\mathbb{E}_{s\sim\omega}[V^{\pi}_{M^\star}(s) - V^{\pi}_{\hat{M}}(s)] \\ &= \mathop{\mathbb{E}}_{s\sim\omega,a\sim\pi(s)}[c(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim M^\star(s,a)}[V^{\pi}_{M^\star}(s')] - c(s,a) - \gamma \mathop{\mathbb{E}}_{s''\sim\hat{M}(s,a)}[V^{\pi}_{\hat{M}}(s'')]] \\ &= \gamma \mathop{\mathbb{E}}_{s\sim\omega,a\sim\pi(s)}[\mathop{\mathbb{E}}_{s'\sim M^\star(s,a)}[V^{\pi}_{M^\star}(s')] - \mathop{\mathbb{E}}_{s''\sim\hat{M}(s,a)}[V^{\pi}_{\hat{M}}(s'')]] \\ &= \gamma \mathop{\mathbb{E}}_{s\sim\omega,a\sim\pi(s)}[\mathop{\mathbb{E}}_{s'\sim M^\star(s,a)}[V^{\pi}_{M^\star}(s')] - \mathop{\mathbb{E}}_{s'\sim M^\star(s,a)}[V^{\pi}_{\hat{M}}(s')] + \mathop{\mathbb{E}}_{s'\sim M^\star(s,a)}[V^{\pi}_{\hat{M}}(s')] - \mathop{\mathbb{E}}_{s''\sim\hat{M}(s,a)}[V^{\pi}_{\hat{M}}(s'')]] \\ &= \gamma \mathop{\mathbb{E}}_{s'\sim d^1_{\omega,\pi}}[V^{\pi}_{M^\star}(s') - V^{\pi}_{\hat{M}}(s')] + \gamma \mathop{\mathbb{E}}_{(s,a)\sim D^0_{\omega,\pi}}\left[\mathbb{E}_{s'\sim M^\star(s,a)}[V^{\pi}_{\hat{M}}(s')] - \mathbb{E}_{s''\sim\hat{M}(s,a)}[V^{\pi}_{\hat{M}}(s'')]\right] \end{aligned}$$

因此，我们建立了时刻 0 的性能差异与时刻 1 的性能差异之间的递推关系，其中状态是在时刻 1 遵循 π 后从状态分布中采样得到的。我们可以求解这一递推关系的无限时域情形，从而得到引理结论。

□

通过模型规划的性能差异（PDPM）引理如下:

引理 A.2 （PDPM（引理 3.1 重述））。对于任意初始状态分布 ω、策略ˆπ， $\pi^\star$以及转移函数ˆM， $M^\star$有:

$$\begin{aligned} J^{\omega}_{M^\star}(\hat{\pi}) - J^{\omega}_{M^\star}(\pi^\star) = \mathbb{E}_{s\sim\omega}[V^{\hat{\pi}}_{M^\star}(s) - V^{\pi^\star}_{M^\star}(s)] =& \mathbb{E}_{s\sim\omega}[V^{\hat{\pi}}_{\hat{M}}(s) - V^{\pi^\star}_{\hat{M}}(s)] \\ &+ \frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim D_{\omega,\hat{\pi}}}[\mathop{\mathbb{E}}_{s'\sim M^\star(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s')] - \mathop{\mathbb{E}}_{s''\sim\hat{M}(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s'')]] \\ &+ \frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim D_{\omega,\pi^\star}}[\mathop{\mathbb{E}}_{s''\sim\hat{M}(s,a)}[V^{\pi^\star}_{\hat{M}}(s'')] - \mathop{\mathbb{E}}_{s'\sim M^\star(s,a)}[V^{\pi^\star}_{\hat{M}}(s')]] \end{aligned}$$

证明。在等式左侧添加并减去相应项，可得

$$\begin{aligned} &\mathbb{E}_{s\sim\omega}[V^{\hat{\pi}}_{M^\star}(s) - V^{\pi^\star}_{M^\star}(s)] = \\ &\mathbb{E}_{s\sim\omega}\left[(V^{\hat{\pi}}_{\hat{M}}(s) - V^{\pi^\star}_{\hat{M}}(s)) + (V^{\hat{\pi}}_{M^\star}(s) - V^{\hat{\pi}}_{\hat{M}}(s)) + (V^{\pi^\star}_{\hat{M}}(s) - V^{\pi^\star}_{M^\star}(s))\right] \end{aligned}$$

应用 t 将模拟引理应用于上方期望中的第二项和第三项，得到结果 t

$$\begin{aligned} \mathbb{E}_{s\sim\omega}[V^{\hat{\pi}}_{M^\star}(s) - V^{\pi^\star}_{M^\star}(s)] = \mathbb{E}_{s\sim\omega}[V^{\hat{\pi}}_{\hat{M}}(s) - V^{\pi^\star}_{\hat{M}}(s)] \\ + \frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim d_{\omega,\hat{\pi}},s'\sim M^\star(s,a),s''\sim\hat{M}(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s') - V^{\hat{\pi}}_{\hat{M}}(s'')] \\ + \frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim d_{\omega,\pi^\star},s'\sim M^\star(s,a),s''\sim\hat{M}(s,a)}[V^{\pi^\star}_{\hat{M}}(s'') - V^{\pi^\star}_{\hat{M}}(s')] \end{aligned}$$

□

推论 A.1 (推论 3.1 重述)。 对于任意初始状态分布 ω， $\pi^\star$以及转移函数 M， $M^\star$，令 $\hat{}\pi$为$\hat{}$在 M 中按(4)返回的最优控制策略，则有：

$$\begin{aligned}\mathbb{E}_{s\sim\omega}[V^{\hat{\pi}}_{M^\star}(s) - V^{\pi^\star}_{M^\star}(s)] \leq \epsilon_{oc} + \\ \frac{\gamma \hat{V}_{\max}}{1-\gamma}\mathbb{E}_{(s,a)\sim D_{\omega,\hat{\pi}}}\left\|\hat{M}(s,a) - M^\star(s,a)\right\|_1 + \\ \frac{\gamma V_{\max}}{1-\gamma}\mathbb{E}_{(s,a)\sim D_{\omega,\pi^\star}}\left\|\hat{M}(s,a) - M^\star(s,a)\right\|_1,\end{aligned}$$

其中 $\hat{} V_{\max} = \|V^{\hat{}\pi}_{\hat{}M}\|_{\infty}$， $V_{\max} = \|V^{\pi^\star}{}_{\hat{}M}\|_\infty$。

证明。 根据引理 3.1，我们有：

$$\begin{aligned}\mathbb{E}_{s\sim\omega}[V^{\hat{\pi}}_{M^\star}(s) - V^{\pi^\star}_{M^\star}(s)] = \mathbb{E}_{s\sim\omega}[V^{\hat{\pi}}_{\hat{M}}(s) - V^{\pi^\star}_{\hat{M}}(s)] \\ + \frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim d_{\omega,\hat{\pi}}, s'\sim M^\star(s,a), s''\sim\hat{M}(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s') - V^{\hat{\pi}}_{\hat{M}}(s'')] \\ + \frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim d_{\omega,\pi^\star}, s'\sim M^\star(s,a), s''\sim\hat{M}(s,a)}[V^{\pi^\star}_{\hat{M}}(s'') - V^{\pi^\star}_{\hat{M}}(s')],\end{aligned}$$

然后由(4)界定第一项，并由霍尔德不等式界定第二项 y

$$\frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim d_{\omega,\hat{\pi}}, s'\sim M^\star(s,a), s''\sim\hat{M}(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s') - V^{\hat{\pi}}_{\hat{M}}(s'')] \leq \frac{\gamma}{1-\gamma}\|V^{\hat{\pi}}_{\hat{M}}\|_\infty \mathbb{E}_{s,a\sim D_{\omega,\hat{\pi}}}\left\|\hat{M}(s,a) - M^\star(s,a)\right\|_1,$$

同理，对第三项应用霍尔德不等式，即完成证明。 □

A.2. 第 4 节的证明

本节将给出第 4 节的证明。我们先从模型中的优势性能差异（PDAM）引理开始：

引理 A.3 （PDAM（重述引理 4.1））。 给定任意初始状态分布 ω、策略$\hat{}\pi$、 $\pi^\star$以及转移函数$\hat{}M$， $M^\star$有：

$$\begin{aligned}J^\omega_{M^\star}(\hat{\pi}) - J^\omega_{M^\star}(\pi^\star) = \mathbb{E}_{s\sim\omega}[V^{\hat{\pi}}_{M^\star}(s) - V^{\pi^\star}_{M^\star}(s)] = \\ \frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim D_{\omega,\hat{\pi}}}[\mathop{\mathbb{E}}_{s'\sim M^\star(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s')] - \mathop{\mathbb{E}}_{s''\sim\hat{M}(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s'')]] + \\ \frac{\gamma}{1-\gamma}\mathop{\mathbb{E}}_{(s,a)\sim D_{\omega,\pi^\star}}[\mathop{\mathbb{E}}_{s''\sim\hat{M}(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s'')] - \mathop{\mathbb{E}}_{s'\sim M^\star(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s')]] + \\ \frac{1}{1-\gamma}\mathop{\mathbb{E}}_{s\sim d_{\omega,\pi^\star}}[V^{\hat{\pi}}_{\hat{M}}(s) - \mathop{\mathbb{E}}_{a\sim\pi^\star(s)}[Q^{\hat{\pi}}_{\hat{M}}(s,a)]]\end{aligned}$$

证明。 让我们从左侧开始，并将其重新表述如下：

$$\mathbb{E}_{s\sim\omega}[V^{\hat{\pi}}_{M^\star}(s) - V^{\pi^\star}_{M^\star}(s)] = \mathbb{E}_{s\sim\omega}[V^{\hat{\pi}}_{\hat{M}}(s) - V^{\pi^\star}_{M^\star}(s)] + \mathbb{E}_{s\sim\omega}[V^{\hat{\pi}}_{M^\star}(s) - V^{\hat{\pi}}_{\hat{M}}(s)]$$

上面的第二项我们很熟悉，它就是模拟引理在公式（11）中的左侧。因此，我们可以应用模拟引理得到：

$$\mathbb{E}_{s\sim\omega}[V^{\hat{\pi}}_{M^\star}(s) - V^{\pi^\star}_{M^\star}(s)] = \mathbb{E}_{s\sim\omega}[V^{\hat{\pi}}_{\hat{M}}(s) - V^{\pi^\star}_{M^\star}(s)] + \frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim D_{\omega,\hat{\pi}}}\left[\mathbb{E}_{s'\sim M^\star(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s')] - \mathbb{E}_{s''\sim\hat{M}(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s'')]\right] \tag{12}$$

现在只剩下第一项，可以将其简化为：

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{s\sim\omega}[V^{\hat{\pi}}_{\hat{M}}(s) - V^{\pi^\star}_{M^\star}(s)] \\
&= \mathop{\mathbb{E}}_{s\sim\omega}[V^{\hat{\pi}}_{\hat{M}}(s) - \mathop{\mathbb{E}}_{a\sim\pi^\star(s)} Q^{\hat{\pi}}_{\hat{M}}(s,a) + \mathop{\mathbb{E}}_{a\sim\pi^\star(s)} Q^{\hat{\pi}}_{\hat{M}}(s,a) - V^{\pi^\star}_{M^\star}(s)] \\
&= \mathop{\mathbb{E}}_{s\sim\omega}[V^{\hat{\pi}}_{\hat{M}}(s) - \mathop{\mathbb{E}}_{a\sim\pi^\star(s)} Q^{\hat{\pi}}_{\hat{M}}(s,a)] \\
&\quad + \mathop{\mathbb{E}}_{s\sim\omega}\Bigg[\mathop{\mathbb{E}}_{a\sim\pi^\star(s)}[c(s,a) + \gamma \mathop{\mathbb{E}}_{s''\sim\hat{M}(s,a)} V^{\hat{\pi}}_{\hat{M}}(s'')] \\
&\qquad - \mathop{\mathbb{E}}_{a\sim\pi^\star(s)}[c(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim M^\star(s,a)} V^{\pi^\star}_{M^\star}(s')]\Bigg] \\
&= \mathop{\mathbb{E}}_{s\sim\omega}[V^{\hat{\pi}}_{\hat{M}}(s) - \mathop{\mathbb{E}}_{a\sim\pi^\star(s)} Q^{\hat{\pi}}_{\hat{M}}(s,a)] \\
&\quad + \gamma \mathop{\mathbb{E}}_{(s,a)\sim D^0_{\omega,\pi^\star}}[\mathop{\mathbb{E}}_{s''\sim\hat{M}(s,a)} V^{\hat{\pi}}_{\hat{M}}(s'') - \mathop{\mathbb{E}}_{s'\sim M^\star(s,a)} V^{\pi^\star}_{M^\star}(s')] \\
&= \mathop{\mathbb{E}}_{s\sim\omega}[V^{\hat{\pi}}_{\hat{M}}(s) - \mathop{\mathbb{E}}_{a\sim\pi^\star(s)} Q^{\hat{\pi}}_{\hat{M}}(s,a)] \\
&\quad + \gamma \mathop{\mathbb{E}}_{(s,a)\sim D^0_{\omega,\pi^\star}}\Bigg[\mathop{\mathbb{E}}_{s''\sim\hat{M}(s,a)} V^{\hat{\pi}}_{\hat{M}}(s'') - \mathop{\mathbb{E}}_{s'\sim M^\star(s,a)} V^{\hat{\pi}}_{\hat{M}}(s') \\
&\qquad + \mathop{\mathbb{E}}_{s'\sim M^\star(s,a)} V^{\hat{\pi}}_{\hat{M}}(s') - \mathop{\mathbb{E}}_{s'\sim M^\star(s,a)} V^{\pi^\star}_{M^\star}(s')\Bigg] \\
&= \mathop{\mathbb{E}}_{s\sim\omega}[V^{\hat{\pi}}_{\hat{M}}(s) - \mathop{\mathbb{E}}_{a\sim\pi^\star(s)} Q^{\hat{\pi}}_{\hat{M}}(s,a)] \\
&\quad + \gamma \mathop{\mathbb{E}}_{(s,a)\sim D^0_{\omega,\pi^\star}}[\mathop{\mathbb{E}}_{s''\sim\hat{M}(s,a)} V^{\hat{\pi}}_{\hat{M}}(s'') - \mathop{\mathbb{E}}_{s'\sim M^\star(s,a)} V^{\hat{\pi}}_{\hat{M}}(s')] \\
&\quad + \gamma \mathop{\mathbb{E}}_{(s,a)\sim D^0_{\omega,\pi^\star}}[\mathop{\mathbb{E}}_{s'\sim M^\star(s,a)} V^{\hat{\pi}}_{\hat{M}}(s') - \mathop{\mathbb{E}}_{s'\sim M^\star(s,a)} V^{\pi^\star}_{M^\star}(s')] \\
&= \mathop{\mathbb{E}}_{s\sim\omega}[V^{\hat{\pi}}_{\hat{M}}(s) - \mathop{\mathbb{E}}_{a\sim\pi^\star(s)} Q^{\hat{\pi}}_{\hat{M}}(s,a)] \\
&\quad + \gamma \mathop{\mathbb{E}}_{(s,a)\sim D^0_{\omega,\pi^\star}}[\mathop{\mathbb{E}}_{s''\sim\hat{M}(s,a)} V^{\hat{\pi}}_{\hat{M}}(s'') - \mathop{\mathbb{E}}_{s'\sim M^\star(s,a)} V^{\hat{\pi}}_{\hat{M}}(s')] \\
&\quad + \gamma \mathop{\mathbb{E}}_{s'\sim d^1_{\omega,\pi^\star}}[V^{\hat{\pi}}_{\hat{M}}(s') - V^{\pi^\star}_{M^\star}(s')]
\end{aligned}
$$

将上述递推关系求解至无限时域，可得：

$$
\begin{aligned}
&\mathbb{E}_{s\sim\omega}[V^{\hat{\pi}}_{\hat{M}}(s) - V^{\pi^\star}_{M^\star}(s)] = \\
&\quad \frac{\gamma}{1-\gamma} \mathop{\mathbb{E}}_{(s,a)\sim D_{\omega,\pi^\star}}[\mathop{\mathbb{E}}_{s''\sim\hat{M}(s,a)} V^{\hat{\pi}}_{\hat{M}}(s'') - \mathop{\mathbb{E}}_{s'\sim M^\star(s,a)} V^{\hat{\pi}}_{\hat{M}}(s')] \\
&\quad + \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d_{\omega,\pi^\star}}[V^{\hat{\pi}}_{\hat{M}}(s) - \mathop{\mathbb{E}}_{a\sim\pi^\star(s)}[Q^{\hat{\pi}}_{\hat{M}}(s,a)]]
\end{aligned}
$$

通过结合 这一结果与我们之前使用模拟引理在（12）中得到的结果，我们即可完成证明。 □

现在，我们使用探索分布 ν 和覆盖系数 C展示结果：

推论 A.2. 设 ν 为探索分布， C 为覆盖系数。对于任意初始状态分布

ˆω、策略 ˆπ、 $\pi^\star$ 以及转移函数 M、 $M^\star$，我们有：

$$
\begin{aligned}
&J^{\omega}_{M^\star}(\hat{\pi}) - J^{\omega}_{M^\star}(\pi^\star) = \mathbb{E}_{s\sim\omega}[V^{\hat{\pi}}_{M^\star}(s) - V^{\pi^\star}_{M^\star}(s)] \leq \\
&\frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim D_{\omega,\hat{\pi}}}\Big[\mathop{\mathbb{E}}_{s'\sim M^\star(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s')] - \mathop{\mathbb{E}}_{s''\sim \hat{M}(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s'')]\Big] + \\
&\frac{\gamma\mathcal{C}}{1-\gamma}\mathop{\mathbb{E}}_{(s,a)\sim\nu}\Big[\mathop{\mathbb{E}}_{s''\sim \hat{M}(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s'')] - \mathop{\mathbb{E}}_{s'\sim M^\star(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s')]\Big] + \\
&\frac{\mathcal{C}}{1-\gamma}\mathop{\mathbb{E}}_{s\sim\nu}[V^{\hat{\pi}}_{\hat{M}}(s) - \mathop{\mathbb{E}}_{a\sim\nu(\cdot|s)}[Q^{\hat{\pi}}_{\hat{M}}(s,a)]]
\end{aligned}
$$

证明。 引理 4.1给出：

$$
\begin{aligned}
&J^{\omega}_{M^\star}(\hat{\pi}) - J^{\omega}_{M^\star}(\pi^\star) = \mathbb{E}_{s\sim\omega}[V^{\hat{\pi}}_{M^\star}(s) - V^{\pi^\star}_{M^\star}(s)] \\
=&\frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim D_{\omega,\hat{\pi}}}\Big[\mathop{\mathbb{E}}_{s'\sim M^\star(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s')] - \mathop{\mathbb{E}}_{s''\sim \hat{M}(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s'')]\Big] \\
+&\frac{\gamma}{1-\gamma}\mathop{\mathbb{E}}_{(s,a)\sim D_{\omega,\pi^\star}}\Big[\mathop{\mathbb{E}}_{s''\sim \hat{M}(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s'')] - \mathop{\mathbb{E}}_{s'\sim M^\star(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s')]\Big] \\
+&\frac{1}{1-\gamma}\mathop{\mathbb{E}}_{(s,a)\sim D_{\omega,\pi^\star}}[V^{\hat{\pi}}_{\hat{M}}(s) - Q^{\hat{\pi}}_{\hat{M}}(s,a)],
\end{aligned}
$$

那么令 ν 为探索分布，我们有：

$$
\begin{aligned}
&J^{\omega}_{M^\star}(\hat{\pi}) - J^{\omega}_{M^\star}(\pi^\star) = \mathbb{E}_{s\sim\omega}[V^{\hat{\pi}}_{M^\star}(s) - V^{\pi^\star}_{M^\star}(s)] \\
\leq&\frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim D_{\omega,\hat{\pi}}}\Big[\mathop{\mathbb{E}}_{s'\sim M^\star(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s')] - \mathop{\mathbb{E}}_{s''\sim \hat{M}(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s'')]\Big] \\
+&\mathcal{C}\frac{\gamma}{1-\gamma}\mathop{\mathbb{E}}_{(s,a)\sim D_e}\Big[\mathop{\mathbb{E}}_{s''\sim \hat{M}(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s'')] - \mathop{\mathbb{E}}_{s'\sim M^\star(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s')]\Big] \\
+&\mathcal{C}\frac{1}{1-\gamma}\mathop{\mathbb{E}}_{(s,a)\sim D_e}[V^{\hat{\pi}}_{\hat{M}}(s) - Q^{\hat{\pi}}_{\hat{M}}(s,a)],
\end{aligned}
$$

其中第一项由 $\mathcal{C} \geq 1$得到，后两项则通过重要性采样得到。 □

A.3. 第 第 5 节 的证明

为简洁起见，以下结果将以专家分布 $D_{\omega,\ \pi^*}$ 的形式陈述。我们将在推论 A.4 中证明，该结果可扩展到我们只能访问探索分布 ν的情况。

推论 A.3(推论 ˆry 5.1 restate)**.** 对于任意初始状态分布 ω， $\pi^\star$以及转移函数 M， $M^\star$，我们有 e,

$$
\begin{aligned}
J^{\omega}_{M^\star}(\hat{\pi}) - J^{\omega}_{M^\star}(\pi^\star) &= \mathbb{E}_{s\sim\omega}[V^{\hat{\pi}}_{M^\star}(s) - V^{\pi^\star}_{M^\star}(s)] \\
&\leq \frac{\gamma\hat{V}_{\max}}{1-\gamma}\mathbb{E}_{(s,a)\sim D_{\omega,\hat{\pi}}}||\hat{M}(s,a) - M^\star(s,a)||_1 \\
&+ \frac{\gamma\hat{V}_{\max}}{1-\gamma}\mathop{\mathbb{E}}_{(s,a)\sim D_{\omega,\pi^\star}}||\hat{M}(s,a) - M^\star(s,a)||_1 \\
&+ \frac{1}{1-\gamma}\mathop{\mathbb{E}}_{s\sim d_{\omega,\pi^\star}}[V^{\hat{\pi}}_{\hat{M}}(s) - \mathop{\mathbb{E}}_{a\sim\pi^\star(s)}[Q^{\hat{\pi}}_{\hat{M}}(s,a)]]
\end{aligned}
$$

证明。 根据引理 4.1，我们有：

$$
\begin{aligned}
&J^{\omega}_{M^\star}(\hat{\pi}) - J^{\omega}_{M^\star}(\pi^\star) = \mathbb{E}_{s\sim\omega}[V^{\hat{\pi}}_{M^\star}(s) - V^{\pi^\star}_{M^\star}(s)] = \\
&\frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim D_{\omega,\hat{\pi}}}\Big[\mathop{\mathbb{E}}_{s'\sim M^\star(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s')] - \mathop{\mathbb{E}}_{s''\sim \hat{M}(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s'')]\Big] + \\
&\frac{\gamma}{1-\gamma}\mathop{\mathbb{E}}_{(s,a)\sim D_{\omega,\pi^\star}}\Big[\mathop{\mathbb{E}}_{s''\sim \hat{M}(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s'')] - \mathop{\mathbb{E}}_{s'\sim M^\star(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s')]\Big] + \\
&\frac{1}{1-\gamma}\mathop{\mathbb{E}}_{s\sim d_{\omega,\pi^\star}}[V^{\hat{\pi}}_{\hat{M}}(s) - \mathop{\mathbb{E}}_{a\sim\pi^\star(s)}[Q^{\hat{\pi}}_{\hat{M}}(s,a)]].
\end{aligned}
$$

对前两项应用霍尔德不等式，即可完成证明。 □

定理 A.1(定理 5.1 重述)。 令 $\{\hat{\pi}_t\}_{t=1}^T$ 为 LAMPS返回策略序列，则有：

$$
\frac{1}{T}\sum_{t=1}^{T} J^{\omega}_{M^\star}(\hat{\pi}_t) - J^{\omega}_{M^\star}(\pi^\star) \le \tilde{O}\left(\epsilon_{po} + \frac{\hat{V}_{\max}}{1-\gamma}\left(\sqrt{\epsilon^{KL}_{model}} + \frac{1}{\sqrt{T}}\right)\right),
$$

其中 $\hat{V}_{\max} = \|V^{\hat{\pi}}_{\hat{M}}\|_\infty$，且 $\epsilon^{KL}_{model} =$最小值$_{M\in\mathcal{M}}\mathbb{E}_{s,\ a\sim\bar{\mathcal{D}}_T}\mathsf{KL}(M(s,a)\ M^*(s,a))$是无关模型误差 *or.*

证明。 与 Ross& Bagnell(2012) 类似，本证明旨在建立运行 算法 2 所得的模型误差保证。首先，根据推论 5.1，我们有

$$
\begin{aligned}
&\sum_{t=1}^{T} J^{\omega}_{M^\star}(\hat{\pi}_t) - J^{\omega}_{M^\star}(\pi^\star) \\
\le&\sum_{t=1}^{T}\frac{\gamma\hat{V}_{\max}}{1-\gamma}\mathbb{E}_{(s,a)\sim D_{\omega,\hat{\pi}_t}}||\hat{M}_t(s,a) - M^\star(s,a)||_1 + \frac{\gamma\hat{V}_{\max}}{1-\gamma}\mathop{\mathbb{E}}_{(s,a)\sim D_{\omega,\pi^\star}}||\hat{M}_t(s,a) - M^\star(s,a)||_1 \\
&+\sum_{t=1}^{T}\frac{1}{1-\gamma}\mathop{\mathbb{E}}_{s\sim d_{\omega,\pi^\star}}[V^{\hat{\pi}_t}_{\hat{M}_t}(s) - \mathop{\mathbb{E}}_{a\sim\pi^\star(s)}[Q^{\hat{\pi}_t}_{\hat{M}_t}(s,a)]] \\
\le&\sum_{t=1}^{T}\frac{\gamma\hat{V}_{\max}}{1-\gamma}\mathbb{E}_{(s,a)\sim D_{\omega,\hat{\pi}_t}}||\hat{M}_t(s,a) - M^\star(s,a)||_1 + \frac{\gamma\hat{V}_{\max}}{1-\gamma}\mathop{\mathbb{E}}_{(s,a)\sim D_{\omega,\pi^\star}}||\hat{M}_t(s,a) - M^\star(s,a)||_1 + T\epsilon_{po},
\end{aligned}
$$

其中最后一项 一行通过运行 算法3 得到。为了界定模型误差，回顾最大似然估计模型损失函数 ：

$$
\ell_t(M) = \mathbb{E}_{s,a,s'\sim\mathcal{D}_t}\log M(s' \mid s,a),
$$

然后按照 算法 2 运行跟随领袖 T 轮，可得：

$$
\begin{aligned}
\sum_{t=1}^{T}\ell_t(\hat{M}_t) &\le \min_{M\in\mathcal{M}}\sum_{t=1}^{T}\ell_t(M) + O(\log(T)) \\
\sum_{t=1}^{T}\ell_t(\hat{M}_t) + 2\mathbb{E}_{s,a\sim\mathcal{D}_t}\mathbb{E}_{s'\sim M^*(s,a)}\log(M^*(s' \mid s,a)) &\le \min_{M\in\mathcal{M}}\sum_{t=1}^{T}\ell_t(M) + 2\mathbb{E}_{s,a\sim\mathcal{D}_t}\mathbb{E}s' \sim M^*(s,a)\log(M^*(s' \mid s,a)) + O(\log(T)) \\
\sum_{t=1}^{T}2\mathbb{E}_{s,a\sim\mathcal{D}_t}\mathsf{KL}(\hat{M}_t(s,a), M^*(s,a)) &\le \min_{M\in\mathcal{M}}\sum_{t=1}^{T}2\mathbb{E}_{s,a\sim\mathcal{D}_t}\mathsf{KL}(M_t(s,a), M^*(s,a)) + O(\log(T)) \\
2\sum_{t=1}^{T}\mathbb{E}_{s,a\sim\mathcal{D}_t}\mathsf{KL}(\hat{M}_t(s,a), M^*(s,a)) &\le 2T\epsilon^{KL}_{model} + O(\log(T)),
\end{aligned}
$$

回顾一下收益 $\mathcal{D}_t = \frac{1}{2} D_{\omega,\ \pi_t} + \frac{1}{2} D_{\omega,\ \pi^*}$。然后根据Pinsker不等式和Jensen不等式，可得：

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}_{s,a\sim\mathcal{D}_t} \|\hat{M}^t(s,a) - M^*(s,a)\|_1 &\le \sum_{t=1}^{T} \sqrt{2\mathbb{E}_{s,a\sim\mathcal{D}_t} \mathsf{KL}(\hat{M}_t(s,a), M^*(s,a))} \\ &\le T\sqrt{\frac{1}{T}\sum_{t=1}^{T} 2\mathbb{E}_{s,a\sim\mathcal{D}_t} \mathsf{KL}(\hat{M}_t(s,a), M^*(s,a))} \\ &\le 2T\sqrt{\epsilon_{model}^{KL}} + \tilde{O}(\sqrt{T}).\end{aligned}$$

因此，我们有

$$\begin{aligned}&\sum_{t=1}^{T} \frac{\gamma \hat{V}_{\max}}{1-\gamma} \mathbb{E}_{(s,a)\sim D_{\omega,\hat{\pi}^t}} ||\hat{M}_t(s,a) - M^\star(s,a)||_1 + \frac{\gamma \hat{V}_{\max}}{1-\gamma} \mathop{\mathbb{E}}_{(s,a)\sim D_{\omega,\pi^\star}} ||\hat{M}_t(s,a) - M^\star(s,a)||_1 \\ \le &\frac{\gamma \hat{V}_{\max}}{1-\gamma} \left\{ 2T\sqrt{\epsilon_{model}^{KL}} + \tilde{O}(\sqrt{T}) \right\},\end{aligned}$$

最后，两边乘以 $\frac{1}{T}$，即可完成证明。 □

最后，我们证明这些结果可以轻松扩展到探索分布设置。

推论 A.4. 令 $\{\hat{\pi}_t\}_{t=1}^T$ 为 LAMPS 返回的策略序列，我们有：

$$\frac{1}{T}\sum_{t=1}^{T} J_{M^\star}^{\omega}(\hat{\pi}_t) - J_{M^\star}^{\omega}(\pi^\star) \le \tilde{O}\left(\mathcal{C}\epsilon_{po} + \frac{\mathcal{C}\hat{V}_{\max}}{1-\gamma} \left(\sqrt{\epsilon_{model}^{KL}} + \frac{1}{\sqrt{T}} \right) \right),$$

其中 $\hat{V}_{\max} = \|V_M^{\hat{\pi}}\|_\infty$, $\mathcal{C} = \sup_{s,a} \frac{D_\omega}{\pi^*(s,a)\nu(s,a)}$ 以及 $\epsilon_{model}^{KL} = \min_{M\in\mathcal{M}} \mathbb{E}_{s,a\sim\bar{\mathcal{D}}_T} \mathsf{KL}(M(s,a)\ M^*(s,a))$ 是无关模型误差。

证明。 我们从推论A.2出发，可得：

$$\begin{aligned}&J_{M^\star}^{\omega}(\hat{\pi}) - J_{M^\star}^{\omega}(\pi^\star) = \mathbb{E}_{s\sim\omega}[V_{M^\star}^{\hat{\pi}}(s) - V_{M^\star}^{\pi^\star}(s)] \\ \le &\mathcal{C}\frac{\gamma}{1-\gamma} \mathbb{E}_{(s,a)\sim D_{\omega,\hat{\pi}}} [\mathop{\mathbb{E}}_{s'\sim M^\star(s,a)} [V_{\hat{M}}^{\hat{\pi}}(s')] - \mathop{\mathbb{E}}_{s''\sim\hat{M}(s,a)} [V_{\hat{M}}^{\hat{\pi}}(s'')]] \\ &+\mathcal{C}\frac{\gamma}{1-\gamma} \mathop{\mathbb{E}}_{(s,a)\sim D_e} [\mathop{\mathbb{E}}_{s''\sim\hat{M}(s,a)} [V_{\hat{M}}^{\hat{\pi}}(s'')] - \mathop{\mathbb{E}}_{s'\sim M^\star(s,a)} [V_{\hat{M}}^{\hat{\pi}}(s')]] \\ &+\mathcal{C}\frac{1}{1-\gamma} \mathop{\mathbb{E}}_{(s,a)\sim D_e} [V_{\hat{M}}^{\hat{\pi}}(s) - Q_{\hat{M}}^{\hat{\pi}}(s,a)],\end{aligned}$$

其中第一项由 $\mathcal{C} \ge 1$得到，最后两项则来自重要性比率。接下来令

$$\ell_t(M) = \mathbb{E}_{s,a,s'\sim\mathcal{D}_t} \log M(s' \mid s,a),$$

并令 $\hat{\pi}$ 满足

$$\mathbb{E}_{s\sim d_e} \left[V_{\hat{M}_t}^{\pi_t}(s) - \mathbb{E}_{a\sim\pi^*(s)} [Q_{\hat{M}}^{\pi_e}(s,a)] \right] \le \epsilon_{po},$$

其中 π_e 是探索策略，重复定理 定理 5.1 证明中的论证即可完成证明 f. □

A.4. 第 6 节证明

再一次，我们证明专家分布版本，以得到更简洁的结果。

定理 A.2(Theo 重述 6.1）。 设 $\{\hat{}\pi_t\}_{t=1}^T$ 为 LAMPS-MM返回的策略序列，则有 :

$$\frac{1}{T}\sum_{t=1}^{T} J^{\omega}_{M^\star}(\hat{\pi}_t) - J^{\omega}_{M^\star}(\pi^\star) \le \tilde{O}\left(\epsilon_{po} + \frac{1}{1-\gamma}\left(\epsilon^{mm}_{model} + \frac{1}{\sqrt{T}}\right)\right),$$

其中 $\epsilon^{mm}_{model} =$ 最小值$_{M\in\mathcal{M}} \frac{1}{T}\sum_{t=1}^{T} \ell_t(M)$ 是无关模型误差。

证明。 为简便起见，我们只证明专家分布版本。同样，我们从引理 4.1开始:

$$\begin{aligned}
&J^{\omega}_{M^\star}(\hat{\pi}) - J^{\omega}_{M^\star}(\pi^\star) = \mathbb{E}_{s\sim\omega}[V^{\hat{\pi}}_{M^\star}(s) - V^{\pi^\star}_{M^\star}(s)] \\
=&\frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim D_{\omega,\hat{\pi}}}\Big[\mathop{\mathbb{E}}_{s'\sim M^\star(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s')] - \mathop{\mathbb{E}}_{s''\sim \hat{M}(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s'')]\Big] \\
+&\frac{\gamma}{1-\gamma}\mathop{\mathbb{E}}_{(s,a)\sim D_{\omega,\pi^\star}}\Big[\mathop{\mathbb{E}}_{s''\sim \hat{M}(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s'')] - \mathop{\mathbb{E}}_{s'\sim M^\star(s,a)}[V^{\hat{\pi}}_{\hat{M}}(s')]\Big] \\
+&\frac{1}{1-\gamma}\mathop{\mathbb{E}}_{(s,a)\sim D_{\omega,\pi^\star}}[V^{\hat{\pi}}_{\hat{M}}(s) - Q^{\hat{\pi}}_{\hat{M}}(s,a)],
\end{aligned}$$

与前面的证明类似,

$$\begin{aligned}
&\sum_{t=1}^{T} J^{\omega}_{M^\star}(\hat{\pi}^t) - J^{\omega}_{M^\star}(\pi^\star) \\
\le&\sum_{t=1}^{T}\frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim D_{\omega,\hat{\pi}_t}}\Big[\mathop{\mathbb{E}}_{s'\sim M^\star(s,a)}[V^{\hat{\pi}_t}_{\hat{M}_t}(s')] - \mathop{\mathbb{E}}_{s''\sim \hat{M}_t(s,a)}[V^{\hat{\pi}_t}_{\hat{M}_t}(s'')]\Big] \\
+&\sum_{t=1}^{T}\frac{\gamma}{1-\gamma}\mathop{\mathbb{E}}_{(s,a)\sim D_{\omega,\pi^\star}}\Big[\mathop{\mathbb{E}}_{s''\sim \hat{M}_t(s,a)}[V^{\hat{\pi}_t}_{\hat{M}_t}(s'')] - \mathop{\mathbb{E}}_{s'\sim M^\star(s,a)}[V^{\hat{\pi}_t}_{\hat{M}_t}(s')]\Big] \\
+&\sum_{t=1}^{T}\frac{1}{1-\gamma}\mathop{\mathbb{E}}_{(s,a)\sim D_{\omega,\pi^\star}}[V^{\hat{\pi}_t}_{\hat{M}_t}(s) - Q^{\hat{\pi}_t}_{\hat{M}_t}(s,a)] \\
\le&\sum_{t=1}^{T}\frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim D_{\omega,\hat{\pi}_t}}\Big[\mathop{\mathbb{E}}_{s'\sim M^\star(s,a)}[V^{\hat{\pi}_t}_{\hat{M}_t}(s')] - \mathop{\mathbb{E}}_{s''\sim \hat{M}_t(s,a)}[V^{\hat{\pi}_t}_{\hat{M}_t}(s'')]\Big] \\
+&\sum_{t=1}^{T}\frac{\gamma}{1-\gamma}\mathop{\mathbb{E}}_{(s,a)\sim D_{\omega,\pi^\star}}\Big[\mathop{\mathbb{E}}_{s''\sim \hat{M}_t(s,a)}[V^{\hat{\pi}_t}_{\hat{M}_t}(s'')] - \mathop{\mathbb{E}}_{s'\sim M^\star(s,a)}[V^{\hat{\pi}_t}_{\hat{M}_t}(s')]\Big] \\
+&T\epsilon_{po}
\end{aligned} \tag{13}$$

现在我们看到 How 模型学习部分实际上如何帮助我们界定前两项。回顾我们在(9)

$$\ell_t(\hat{M}_t) = \mathbb{E}_{s,a\sim\mathcal{D}_t}\left|\mathop{\mathbb{E}}_{s'\sim M^\star(s,a)}[V^{\hat{\pi}_t}_{\hat{M}_t}(s')] - \mathop{\mathbb{E}}_{s''\sim M(s,a)}[V^{\hat{\pi}_t}_{\hat{M}_t}(s'')]\right|,$$

使用损失对式（13）中的各项进行上界估计 $\ell_t(\hat{}M_t)$ 我们得到

$$\begin{aligned}
&\sum_{t=1}^{T} J^{\omega}_{M^\star}(\hat{\pi}^t) - J^{\omega}_{M^\star}(\pi^\star) \\
&\le \frac{2\gamma}{1-\gamma}\sum_{t=1}^{T}\ell_t(\hat{M}_t) + T\epsilon_{po}
\end{aligned}$$

对损失序列 $\ell_t(\hat{}M_t)$使用跟随正则化领导者，可得:

$$\begin{aligned}
\sum_{t=1}^{T}\ell_t(\hat{M}_t) &\le \min_{M\in\mathcal{M}}\sum_{t=1}^{T}\ell_t(M) + O(\sqrt{T}) \\
&\le T\epsilon^{mm}_{model} + O(\sqrt{T})
\end{aligned}$$

算法5 矩匹配 **拟合模型** 与有符号损失($\mathcal{D}_t, \{\ell_i^{sn}\}_{i=1}^{t-1}$)

要求： 迄今为止从学习得到的策略中收集的数据 $\mathcal{D}_t^{learned}$，迄今为止从探索分布中收集的数据 $\mathcal{D}_t^{exp}$，模型类 $\mathcal{M}$，先前的损失 $\{\ell_i^{sn}\}_{i=1}^{t-1}$ 1：按如下方式定义损失 $\ell_t^{sn}(M)$,

$$\ell_t^{sn}(M) = \mathop{\mathbb{E}}_{(s,a,s')\sim\mathcal{D}_t^{learned}}\left[V_{\hat{M}_t}^{\hat{\pi}_t}(s') - \mathop{\mathbb{E}}_{s''\sim M(s,a)}[V_{\hat{M}_t}^{\hat{\pi}_t}(s'')]\right] + \mathop{\mathbb{E}}_{(s,a,s')\sim\mathcal{D}_t^{exp}}\left[\mathop{\mathbb{E}}_{s''\sim M(s,a)}[V_{\hat{M}_t}^{\hat{\pi}_t}(s'')] - V_{\hat{M}_t}^{\hat{\pi}_t}(s')\right] \tag{14}$$

ˆ 2：使用适用于凸损失的在线无遗憾算法计算模型 M_{t+1}，例如跟随正则化领导者，

$$\hat{M}_{t+1} \leftarrow \mathop{\mathrm{argmin}}_{M\in\mathcal{M}} \sum_{\tau=1}^{t} \ell_\tau^{sn}(M) + \mathcal{R}(M).$$

3：**返回** $\hat{M}_{t+1}$

将其代入上式可得

$$\begin{aligned}&\sum_{t=1}^{T} J_{M^\star}^{\omega}(\hat{\pi}^t) - J_{M^\star}^{\omega}(\pi^\star)\\ &\leq \frac{2\gamma}{1-\gamma}(T\epsilon_{model}^{mm} + O(\sqrt{T})) + T\epsilon_{po}\end{aligned}$$

再次将两边乘以 $\frac{1}{T}$，并利用 $\gamma \leq 1$, –即可完成证明。 □

B. 补充分析

本节将介绍正文中暂未给出的若干结果。

B.1. 使用 Hedge 的预热论证

为了理解数据聚合中无遗憾结果背后的直觉，我们现在考虑一个简化版本：假设我们有一个包含有限个模型的模型类 $\mathcal{M}$，并将模型的数量记为 N。对于每个模型 $M \in \mathcal{M}$，记 π^* 模型中的一个策略为 π ˆM，该策略的劣势非正。现在考虑我们所提出的采用 Hedge 的无遗憾算法 ˆ：在每次迭代 t中，我们首先根据当前权重采样一个模型 M_t，然后使用

$\pi_t = \pi^{M_t}$ M $\ell_t\ M\ = I\{\mathbb{E}_{d^{\pi_t}}\|M\ s\ a\ - M^*\ s\ a\ \|_1 +$ ˆ进行 $\mathbb{E}_{d^{\pi^*}}\|M\ s\ a\ - M^*\ s\ a\ \|_1 > 0\}$ rollout。接着，对于每个模型 ˆ，我们计算损失 (ˆ) ˆ (,) (,) ˆ (,) (,)。可以将这种损失理解为：模型在当前轨迹分布上是否会犯错。假设模型类 $\mathcal{M}$ 是可实现的。那么我们有：

$$\begin{aligned}R(T) &\leq \sum_{t=1}^{T} \mathbb{E}_{\hat{M}_t} \frac{\gamma V_{\max}}{1-\gamma}\mathbb{E}_{d^{\pi_t}}\|\hat{M}(s,a) - M^*(s,a)\|_1 + \frac{\gamma V_{\max}}{1-\gamma}\mathbb{E}_{d^{\pi^*}}\|\hat{M}(s,a) - M^*(s,a)\|_1\\ &\leq \sum_{t=1}^{T} \mathbb{E}_{\hat{M}_t}\ell_t(\hat{M}_t)\\ &\leq O(\sqrt{T\log(N)}).\end{aligned}$$

ˆ 但请注意，这种方法在计算上效率较低（因为我们需要在每一轮中为每个 $M \in \mathcal{M}$ 计算损失）。

B.2. 与有符号损失的比较

在本节中，我们介绍 LAMPS-MM 的一种替代算法，该算法使用带符号版本的损失，而不是

平方损失 (9)。我们在 算法5 中介绍这一替代算法，其中我们运行 算法1，并结合 算法3 和算法5。为简单起见，下面给出专家分布上的遗憾结果：

定理B.1。 令 $\{\hat{}\pi_t\}_{t=1}^T$ 为算法5返回策略的序列，则有：

$$\frac{1}{T}\sum_{t=1}^{T} J^{\omega}_{M^\star}(\hat{\pi}_t) - J^{\omega}_{M^\star}(\pi^\star) \le O\left(\epsilon_{po} + \frac{1}{1-\gamma}\left(\epsilon^{sn}_{model} + \frac{1}{\sqrt{T}}\right)\right),$$

其中 $\epsilon^{sn}_{model} =$ 最小值$_{M\in\mathcal{M}} \frac{1}{T}\sum_{t=1}^{T} \ell^{sn}_t(M)$ 是无关模型误差。

证明。 证明与定理 6.1的证明基本相同。我们有

$$\begin{aligned}
&\sum_{t=1}^{T} J^{\omega}_{M^\star}(\hat{\pi}^t) - J^{\omega}_{M^\star}(\pi^\star) \\
\le &\sum_{t=1}^{T} \frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim D_{\omega,\hat{\pi}_t}}\Big[\mathop{\mathbb{E}}_{s'\sim M^\star(s,a)}[V^{\hat{\pi}_t}_{\hat{M}_t}(s')] - \mathop{\mathbb{E}}_{s''\sim \hat{M}_t(s,a)}[V^{\hat{\pi}_t}_{\hat{M}_t}(s'')]\Big] \\
&+ \sum_{t=1}^{T} \frac{\gamma}{1-\gamma}\mathop{\mathbb{E}}_{(s,a)\sim D_{\omega,\pi^\star}}\Big[\mathop{\mathbb{E}}_{s''\sim \hat{M}_t(s,a)}[V^{\hat{\pi}_t}_{\hat{M}_t}(s'')] - \mathop{\mathbb{E}}_{s'\sim M^\star(s,a)}[V^{\hat{\pi}_t}_{\hat{M}_t}(s')]\Big] \\
&+ T\epsilon_{po} \\
= &\sum_{t=1}^{T} \ell^{sn}_t(\hat{M}_t) + T\epsilon_{po}.
\end{aligned}$$

使用跟随正则化领导者可得：

$$\begin{aligned}
\sum_{t=1}^{T} \ell^{sn}_t(\hat{M}_t) &\le \min_{M\in\mathcal{M}} \sum_{t=1}^{T} \ell^{sn}_t(M) + O(\sqrt{T}) \\
&\le T\epsilon^{mm}_{model} + O(\sqrt{T})).
\end{aligned}$$

在两边取 $\frac{1}{T}$后，证明完成。 □

我们注意到，在这里两个损失函数都能为我们带来 $\tilde{O}(\frac{1}{\sqrt{T}})$ 遗憾率。

B.3. LAMPS-MM的有限样本分析

在本节中，我们使用在线到批量技术（算法 4 （Cesa-Bianchi 等人，2004）对 $\mathcal{M}$进行有限样本分析。首先，我们引入一个新的函数类。该函数类由模型类 $\mathcal{X} = \mathcal{S}\times\mathcal{A}\times\mathcal{V}$构造，并将状态、动作和价值函数三元组作为输入。记 $\mathcal{V}$ 为函数类，

$$\mathcal{H} = \left\{h : \mathcal{X}\to\mathbb{R} \mid \exists M\in\mathcal{M} \quad \text{s.t.} \quad \forall (s,a,v)\in\mathcal{X}, h(s,a,v) = \int M(s' \mid s,a)v(s')\, d(s')\right\}.$$

记随机变量 $x_t = (s_t,\ a_t,\ v_t)$，我们注意到随机变量 y_t 的生成方式如下

$$y_t = v_t(s'_t), \quad s'_t \sim M^*(s_t,a_t).$$

记 $\mathcal{F}_t = \{(X_1,\ Y_1).....,\ (X_{t-1},\ Y_{t-1})\}$，并且在每一轮中，我们使用损失函数

$$\hat{\ell}_t(h) = |h(s_t,a_t,v_t) - y_t| + |h(\tilde{s}_t,\tilde{a}_t,v_t) - \tilde{y}_t)|,$$

其中 $(s,a)\ \sim \mathcal{D}_t, s'\sim M^*(s,a)$ 和 $(\tilde{s},\tilde{a})\ \sim \mathcal{D}_{\pi^*}, \tilde{s}'\sim M^*(\tilde{s},\tilde{a})$。此外，定义

$$Z_t = (\hat{\ell}(\hat{h}_t) - \ell(\hat{h}_t)) - (\hat{\ell}(h^*) - \ell(h^*)).$$

注意， Z_t 是关于滤过 $\mathcal{F}_t$适应的鞅差序列，使得

$$\mathbb{E}[Z_t \mid \mathcal{F}_t] = 0.$$

与此同时，我们还有 $|Z_t| \leq \frac{4}{1-\gamma}$。根据 引理 B.1，我们以概率 $1-\delta$ ，

$$\sum_{t=1}^{T} Z_t \leq \sqrt{\frac{32T\log(1/\delta)}{(1-\gamma)^2}},$$

然后对 $\mathcal{H}$应用并集界，对于任意 h，我们有：

$$\sum_{t=1}^{T} Z_t \leq \sqrt{\frac{32T\log(|\mathcal{H}|/\delta)}{(1-\gamma)^2}}.$$

根据 引理 B.2和可实现性，我们有

$$\begin{aligned}\sum_{t=1}^{T} \ell_t(\hat{h}_t) &\leq R(T) + \sum_{t=1}^{T} \ell_t(h^*) + Z_t \\ &\lesssim \frac{1}{1-\gamma}\sqrt{T\log(|\mathcal{H}|/\delta)},\end{aligned}$$

由于 $R(T) = O(\sqrt{T})$ 也成立。为简单起见，我们进一步假设 $\epsilon_{po} \leq 0$，于是得到以下遗憾界：

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} J^{\omega}_{M^\star}(\hat{\pi}_t) - J^{\omega}_{M^\star}(\pi^\star) &\leq \frac{1}{T}\sum_{t=1}^{T} \ell_t(\hat{h}_t) \\ &\leq \tilde{O}\left(\frac{T^{1/2}\log(|\mathcal{H}|/\delta)^{1/2}}{(1-\gamma)}\right),\end{aligned}$$

转换为样本复杂度后，通过取

$$T = \tilde{O}\left(\frac{\log(|\mathcal{H}|/\delta)}{(1-\gamma)^2\epsilon^2}\right),$$

我们以概率 $1-\delta$ 得到，

$$\frac{1}{T}\sum_{t=1}^{T} J^{\omega}_{M^\star}(\hat{\pi}_t) - J^{\omega}_{M^\star}(\pi^\star) \leq \epsilon.$$

这里补充几点说明。首先，这一结果无法直接与传统的基于模型的强化学习的样本复杂度进行比较，因为 $\mathcal{H}$ 大小的紧确界取决于具体实例。其次，这一结果并不与 第 6 节 中提到的困难相矛盾，因为通过构造$\mathcal{H}$，已经隐式地解决了从学习得到的模型中采样的问题。

B.4. 辅助引理

引理 B.1（霍夫丁–阿祖马不等式）。 假设 $X_1, \ldots, X_T$ 是一个鞅差序列，其中 $|X_t| \leq R$。那么，对于所有 $\epsilon > 0$ 以及所有正整数 T，我们有

$$P\left(\sum_{t=1}^{T} X_i \geq \epsilon\right) \leq \exp\left(\frac{-\epsilon^2}{2TR^2}\right).$$

引理 B.2（在线到批量转换）。 考虑一个函数类为 $\mathcal{H}$ 的序贯函数估计问题。令 $\mathcal{X}$ 为输入空间， $\mathcal{Y}$ 为目标空间。假设每个输入和目标 （x_t、 y_t）均独立同分布地生成，其中

ˆ$x_t \sim \rho\left(x_1, y_1, \ldots, x_{t-1}, y_{t-1}\right)$， $y \sim p^*(\cdot \mid x_t)$。令 h_t 为在线学习算法 A 的返回值，该算法接收输入ˆ ˆˆ$\{x_1, y_1, \ell_1, \ldots, x_{t-1}, y_{t-1}, \ell_{t-1}\}$，其中 ℓ_t 是第 t 轮损失函数 ℓ_t 的经验版本。定义

$$Z_t = (\hat{\ell}(\hat{h}_t) - \ell(\hat{h}_t)) - (\hat{\ell}(h^*) - \ell(h^*)),$$

其中 $h^* = \text{最小值}_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell_t(h)$，我们有

$$\sum_{t=1}^{T} \ell_t(\hat{h}_t) \leq R(T) + \sum_{t=1}^{T} \ell_t(h^*) + Z_t,$$

其中 $R(T)$ 是运行 A 至第 T 轮时的遗憾。

C. 实验详情

C.1. 环境详情

本节将详细介绍我们所使用的环境第7节，尤其是直升机和 WideTree 等非标准基准。

C.1.1. 直升机

直升机领域最初由 Abbeel & Ng (2005) 提出，也被 Ross & Bagnell (2012) 使用。本文聚焦于 悬停 任务。该环境具有20维状态空间和4维动作空间。

系统动力学是非线性的，并以质量和惯性量作为参数，构成一个20维向量。模型类为 $\mathbb{R}^{20}$，对应于用于定义动力学的参数向量。代价函数为

$$c(x, u) = \sum_{h=1}^{H} x_h^\top Q x_h + u_h^\top R u_h + x_H^\top Q_H x_H,$$

也就是说，我们对任何偏离原点的情况以及所付出的控制 effort 进行惩罚。

为了理解为何对期望轨迹进行一次反向传播就等同于 算法3，让我们重新审视 算法3中的目标函数：

$$\mathop{\mathbb{E}}_{s \sim \nu}\left[V_{\hat{M}_t}^{\hat{\pi}_t}(s) - \min_{a \in \mathcal{A}}[Q_{\hat{M}_t}^{\hat{\pi}_t}(s, a)]\right] \leq \epsilon_{po}$$

在上述目标函数中，ν 是探索分布，在本例中，它就是使ˆ轨迹保持悬停的期望轨迹（$\{s_{hover}, u_{hover}, \ldots, s_{hover}\}$）。$M_t$ 是我们进行优化所针对的模型，而上述目标函数表明，我们需要找到一个策略 ˆπ_t 使其仅在期望轨迹上与最优策略一样好。因此，为计算这一策略，我们ˆ对 M_t 的非线性动力学在期望轨迹附近进行线性化（前向传播），然后针对线性化后的动力学计算最优线性二次调节器（反向传播）。这样便得到策略ˆπ_t，它仅沿期望轨迹达到最优。请注意，这只需一次反向传播；而要实现（4），则需要进行多次迭代线性二次调节，包括多次反向传播。这凸显了 算法3 相较于传统最优规划方法的计算优势。

C.1.2. WIDETREE

这个马尔可夫决策过程是 图 1中所示过程的一种变体。它在 图5中进行了描述。

我们使用 Hedge（算法 4，采用 Freund 与 Schapire，1997年）实现该算法，同时维护一个离散分布 $(p,\ 1-p)$，该分布定义在两个模型 M^{good} 和 M^{bad}上。我们使用 $\epsilon = 0.9$ 进行 Hedge 更新。

C.1.3. 线性动力系统

在本实验中，任务是控制一个线性动力系统，其中真实动力学随时间变化，但模型类仅包含时不变线性动力学模型。该系统具有 5 维状态、1 维控制输入，并且我们

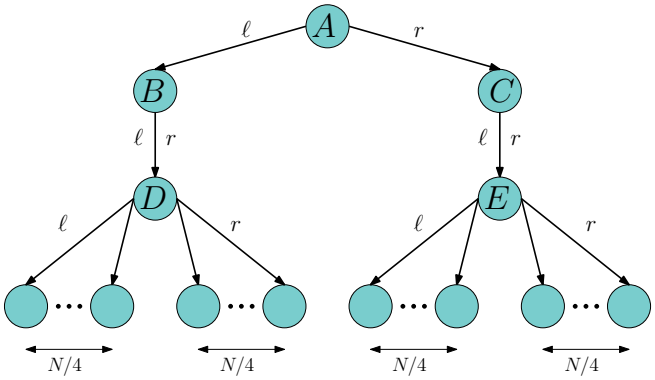

图5。 实验中使用的 Widetree 域。我们有一个包含 $N+5$ 个状态的马尔可夫决策过程，其中 N 个是终止状态（或叶节点）。每个状态都有两个动作 ℓ 和 r。在状态 B 和 C 下，两个动作分别都会转移到同一个状态 D 和 E。在状态 D 和 E 下，两个动作都会以均匀概率将状态随机转移到 $N/4$ 个终止状态中的一个。图中展示了真实动力学 $M^\star$。对于任意动作 a，在任意 $s \neq B$ 和 c（B, a）下，代价 c（s, a）$= \epsilon << 1$。因此，在 A 处采取的动作至关重要。模型类 $\mathcal{M}$ 仅包含两个模型： M^{good} 在 A 处正确捕捉了动力学，但在其他所有地方都会出错；而 M^{bad} 仅在 A 处出错，却在其他所有地方都捕捉到了真实动力学。

需要在 100 个时间步的时域内对其进行控制。真实动力学按照 $x_{t+1} = A_t x_t + B_t u_t$ 演化，其中，

$$A_t = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$$

当 t 为偶数时，且

$$A_t = \begin{bmatrix} 1.5 & 0 \\ 0 & 1.5 \end{bmatrix}$$

当 t 为奇数时。模型类 $\mathcal{M}$ 由如下形式的线性动力学模型组成： $\{x_{t+1} = Ax_t + Bu_t\}$ ，因此无法精确建模真实动力学，属于非特定模型类。

代价函数为二次函数，形式如下：

$$c(\mathbf{x}, \mathbf{u}) = \sum_{t=0}^{99} u_t^T u_t + x_{100}^T x_{100}$$

其中 $\mathbf{x}$,$\mathbf{u}$ 表示完整的状态和控制轨迹。请注意，代价函数只会在每个时间步对控制输入进行惩罚，而仅在最后一个时间步对状态进行惩罚。

给定一个模型 $(A, B) \in \mathcal{M}$，我们可以利用有限时域离散黎卡提方程解（Bertsekas，2005）以闭式形式计算最优策略及其价值函数。价值函数用矩阵表示为 $P_t \in \mathbb{R}^{5\times5}$ 其中$V_t(x) = x^T P_t x$ 表示从时间 t 到时域结束的后续代价。因此，我们可以为任意模型构造损失$M = (A, B)$ ，参见 LAMPS-MM 的 算法 4 ，如下：

$$\ell_t(A, B) = \mathop{\mathbb{E}}_{(x_h, u_h, x_{h+1})\sim\mathcal{D}_t} (x_{h+1}^T P_{h+1} x_{h+1} - (Ax_h + Bu_h)^T P_{h+1}(Ax_h + Bu_h))^2$$

而 SYSID 的最大似然估计损失则为，

$$\ell_t(A, B) = \mathop{\mathbb{E}}_{(x_h, u_h, x_{h+1})\sim\mathcal{D}_t} \|x_{h+1} - (Ax_h + Bu_h)\|_2^2$$

C.1.4. 迷宫

对于迷宫环境，我们采用 D4RL 中的 PointMaze（大型）任务（Fu 等人，2020年）。我们在图 6中展示了该任务的可视化结果。原始离线数据集包含 4000000 个样本，而我们的实验仅使用 10000 和 50000 个样本。



图 6.要点 迷宫（大型）环境。图片取自https://sites.google.com/view/d4rl/home .

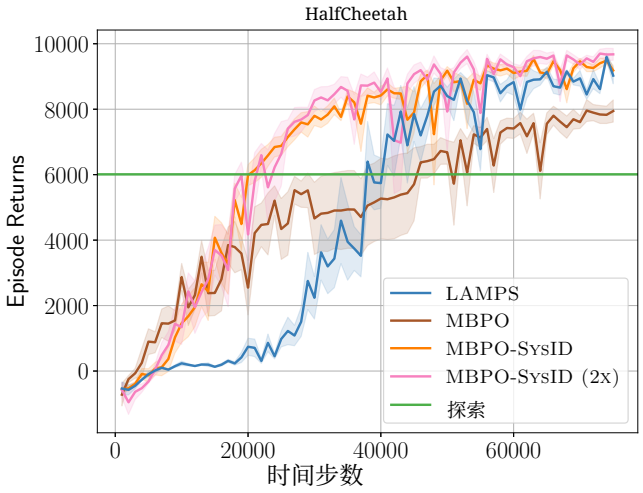

图 7. HalfCheetah基准上的结果。结果为 5 个随机种子的平均值，阴影区域表示标准误差。我们使用了 10000 个探索样本。请注意，在这种情况下，LAMPS 需要更多样本才能达到 MBPO-SYSID 的渐近性能。

C.2. 额外的 MuJoCo 实验

本节中，我们展示了另一项 MuJoCo 实验。在这种情况下，我们的算法起初表现不及 MBPO-SysID，但在数据充足时达到了相同性能。我们推测，主要原因是此处的探索分布质量不高，这表明，与算法3 相比，MBPO-SysID 对探索分布的质量不那么敏感，正如第 5 节所述。

C.3. 实现细节和超参数

本节介绍我们在第7节和第7节所述实验中使用的实现细节和超参数。如正文所述，我们的 MBPO 实现基于 Pineda 等人(2021年)，MBPO-SysID 和 LAMPS 也同样如此。在模型训练中，所有基线都使用与原始 MBPO 相同的模型集成，而 MBPO-SysID 和 LAMPS 则使用数据聚合数据和探索数据来训练模型。在策略训练中，MBPO-SysID 使用与 MBPO 相同的分支更新，并采用软演员-评论家（SAC）；LAMPS 则采用不同的目标函数，在演员更新步骤中用劣势项替换 Q-value（注意，这意味着演员更新也仅基于探索分布）。详细算法见算法 6。

我们使用默认的超参数 适用于大多数情况，但为了完整起见，我们仍将其列出。请注意，这些超参数 e

算法 6 Deep LAMPS

要求： 迭代次数 T、模型集成 M_θ、策略 π_ϕ、价值函数 q_ψ、探索数据集 ν、策略展开步数 H、模型展开 E、模型展开时域 k。

初始化数据聚合 $\mathcal{D} = \emptyset$。

对 $t = 1, \ldots, T$ **do**

使用最大似然估计训练模型集成 M_θ，并通过以下方式进行：从 ν 按 1 2 概率采样，以及按 1–2 概率从 $\mathcal{D}$采样。

对于 $h = 1, \ldots, H$ **do**

通过滚动执行 π_ϕ，在 $M^\star$ 中收集数据

初始化模型缓冲区 $\mathcal{D}_{model} = \emptyset$.

对于 $e = 1, \ldots, E$ **do**

从 ρ中均匀采样状态 s，使用 π_θ 展开 k 步，并将轨迹添加到 $\mathcal{D}_{model}$

使用 $\mathcal{D}_{model}$更新软 Q价值函数 q_ψ。

使用以下内容更新策略π_ϕ

$$\mathcal{J}(\pi_\phi, \nu, q_\psi) = \mathbb{E}_{s, a\sim\nu}\left[\mathbb{E}_{\tilde{a}\sim\pi_\phi(s)} \log(\pi_\phi(\tilde{a} \mid s)) - (q_\psi(s, \tilde{a}) - q_\psi(s, a))\right]$$

结束循环

结束循环

结束循环

返回 策略序列 $\{\hat{\pi}_t\}_{t=1}^{T+1}$

所有基线都使用相同的超参数，但 MBPO-SYSID (2X) 使用的数量是超参数末尾标注的数值的两倍（$*$）。

表 *1.* HalfCheetah 的超参数

	值
探索样本量	10000
集成规模	7
集成精英数量	5
模型学习率	0.001
模型批量大小	256
学习模型中的展开步数（$*$）	400
展开长度	$1 \rightarrow 1$
策略更新次数（$*$）	20
策略类型	随机高斯策略

表2. Ant 的超参数

	值
探索样本量	10000
集成规模	7
集成精英数量	5
模型学习率	0.0003
模型批量大小	256
学习模型中的展开步数（$*$）	400
展开长度	$1 \rightarrow 25$
策略更新次数（$*$）	20
策略类型	随机高斯策略

表3. Hopper 的超参数

	值
探索样本量	10000
集成规模	7
集成精英数量	5
模型学习率	0.001
模型批量大小	256
学习模型中的展开步数（$*$）	400
展开长度	$1 \rightarrow 15$
策略更新次数（$*$）	40
策略类型	随机高斯策略

表4. Humanoid 的超参数

	值
探索样本量	10000
集成规模	7
集成精英数量	5
模型学习率	0.0003
模型批量大小	256
学习模型中的展开步数（$*$）	400
展开长度	$1 \rightarrow 25$
策略更新次数（$*$）	20
策略类型	随机高斯策略

表5。 Walker 的超参数

	值
探索样本量	10000
集成规模	7
集成精英数量	5
模型学习率	0.001
模型批量大小	256
学习模型中的展开步数 (*)	400
展开长度	$1 \rightarrow 1$
策略更新次数 (*)	20
策略类型	随机高斯策略

表 6. PointMaze 的超参数

	值
探索样本量	10000/50000
集成规模	7
集成精英数量	5
模型学习率	0.001
模型批量大小	256
学习模型中的展开步数（$*$）	400
展开长度	$1 \rightarrow 1$
策略更新次数（$*$）	20
策略类型	确定性策略