arXiv:2304.13426v1 [cs.LG] 26 Apr 2023

FLEX：面向非线性系统的自适应探索算法

Matthieu Blanke [1] **Marc Lelarge** [1]

摘要

基于模型的强化学习是一种强大的工具，但收集数据以拟合系统的准确模型可能代价高昂。因此，以样本高效的方式探索未知环境至关重要。然而，动力学的复杂性以及实际系统的计算限制，使这项任务充满挑战。在本文中，我们介绍了 FLEX，这是一种基于最优实验设计、面向非线性动力学的探索算法。我们的策略最大化下一步的信息量，从而形成一种自适应探索算法；该算法兼容通用的参数化学习模型，并且只需极少的资源。我们在涵盖不同设置的多个非线性环境中测试了该方法，其中包括时变动力学。考虑到探索旨在服务于利用目标，我们还在下游的基于模型的经典控制任务中测试了该算法，并将其与其他先进的基于模型和无模型方法进行了比较。FLEX 所取得的性能具有竞争力，且计算成本较低。

1. 引言

控制理论和基于模型的强化学习已在航空学、机器人学和能源系统等多个领域取得了一系列成果（Kirk, 1970年；Sutton & Barto，2018年）。为了让智能体找到有效的控制策略，环境的数学模型必须忠实地捕捉系统的动力学，因此必须使用数据进行拟合。然而，收集观测可能成本高昂：例如，对于飞机系统而言，进行实验会耗费大量能源和时间（Gupta 等，1976年）。在这方面，主动探索（或系统辨识）旨在激励系统，以收集具有信息量的数据，并以样本高效的方式学习系统的全局特性（Yang 等，2021年），且不依赖于任何控制任务。一旦完成这一与任务无关的探索阶段，就可以利用学习得到的模型来解决多个下游任务。

动力学可能很复杂，通常表现为状态的非线性函数。例如，空气摩擦力对于制定准确的控制律至关重要，但很难依据物理原理对其建模（Faessler et al., 2018年；De Simone et al., 2015）。虽然线性系统中的探索已得到充分研究（Goodwin & Payne, 1977），但高效地学习非线性动力学要困难得多。对于实际应用，嵌入式系统在内存和计算资源方面的严格限制又进一步加剧了这一难题（Tassa et al., 2012年）。此外，动力学本身或智能体对其建立的模型可能会随时间变化，因此探索策略必须随着数据流的收集而进行调整。因此，算法必须能够自适应运行、占用有限的内存存储，并且速度足够快，能够在真实系统中实现，同时还要具备足够的灵活性，以利用可能较为复杂的模型学习复杂动力学。

近期，人们提出了不同的方法来探索非线性环境，这些方法通过最大化特定类别学习模型的参数信息增益（或不确定性）来实现。对于具有线性参数化的模型，可以推导出精确计算方法（Schultheis 等，2020年），但这类模型的表达能力有限。还可以使用高斯过程计算不确定性（Buisson-Fenet 等人，2020年），或利用神经网络集成进行近似（Shyam 等，2019年；Sekar 等，2020年）；然而，前一种方法具有二次内存复杂度，后一种方法的计算成本较高，因此它们不太可能在实际系统中实现。在上述所有方法中，输入都通过分幕式地求解非凸优化问题进行规划，其中非线性动力学会在可能很大的时域内进行模拟。因此，规划依赖于非线性求解器，而非线性求解器可能运行得过慢，无法满足实时要求（Kleff 等，2021年）。此外，在较大时域上进行规划会使算法无法根据新收集的观测进行适应。结果是，智能体可能长时间信任错误的模型，并探索信息不足的状态。近期研究集中于为线性动力学推导快速且自适应的探索策略（Blanke 与 Lelarge,

[1]INRIA、DI ENS、PSL研究大学。通讯作者：Matthieu Blanke <matthieu.blanke@inria.fr>。

2022年）。然而，在探索复杂得多的系统时，仍要保持此类保证，依然是一项重大挑战，也是一个有待研究的问题。

贡献 本研究探讨非线性环境中的主动探索问题，尤其重视现实系统施加的约束。基于信息论和最优实验设计，我们定义了适用于通用参数化学习模型的探索目标，涵盖线性模型和神经网络。我们推导出该目标的在线近似，并提出快速且具有自适应性的探索算法FLEX。通过在多种非线性系统（包括时变环境）上的实验，验证了该方法的样本效率和自适应性，并将FLEX 的性能与若干基线方法进行了比较。我们还在下游利用任务中评估了该探索方法，并将 FLEX 与基于模型和无模型的探索方法进行了比较。

组织结构 我们首先在第 2 节介绍问题的数学形式化。在第3节中，我们聚焦于采用线性参数化的模型，并据此推导出一个信息论目标函数。在第 4 节中，我们介绍FLEX（算法2），这是一种能够在线自适应地最大化该目标函数的探索算法。在第 5 节中，我们将该方法扩展到通用的非线性模型。我们在第 6 节通过实验检验该方法。

2. 探索非线性环境

在非线性动力系统中，状态 $x \in \mathbb{R}^d$ 和输入 $u \in \mathbb{R}^m$ 由以下形式的方程支配

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f_\star(x, u), \tag{2.1}$$

其中，$f_\star$ 是用于建模动力学的非线性函数。该函数未知或部分未知，我们的目标是从数据中学习该函数，同时尽可能减少所需样本数量。实践中，我们观测到的是动力学 (2.1) 有限个离散且带噪声的观测：

$$x_{t+1} = x_t + \mathrm{d}t f_\star(x_t, u_t) + w_t, \quad 0 \le t \le T-1, \tag{2.2}$$

其中，$\mathrm{d}t$ 是已知的时间步，T 是观测数量，$x_t \in \mathbb{R}^d$ 是状态向量，$w_t \sim \mathcal{N}(0,\ \sigma^2 I_d)$ 是方差 σ^2已知的正态分布各向同性噪声，控制变量 $u_t \in \mathbb{R}^m$ 由智能体在约束条件 $\|u_t\|_2 \le \gamma$下选择。我们假设 f 是可微函数。我们交替使用 $f_\star(x,\ u)$ 和 $f_\star(z)$ 表示，其中 $z = (x\ u) \in \mathbb{R}^{d+m}$ 是状态-动作对。请注意，我们的问题可以在连续马尔可夫决策过程的框架下进行表述（Sutton & Barto，2018年），但我们发现式（2.2）更适合我们的方法。

算法 1主动探索

输入 学习模型 f，时间范围 T，时间步长 $\mathrm{d}t$，策略 $\pi \in \Pi_\gamma$，学习规则 $\hat{\theta}$
输出参数估计 θ_T
对于 $0 \le t \le T-1$ **执行**
  选择 $u_t = \pi(x_{0:t},\ u_{0:t-1}|\theta_t)$
  观察 $x_{t+1} = x_t + \mathrm{d}t f_\star(x_t,\ u_t) + w_t$
  更新 $\theta_{t+1} = \hat{\theta}(x_{0:t+1},\ u_{0:t+1})$
结束循环

序贯学习 动力学函数 $f_\star$ 通过过去的观测，利用参数收集于向量 $\theta \in \mathbb{R}^n$中的参数化函数 f进行学习。我们用映射 $f(z,\ \theta)$表示一般的参数化模型。在每个时间 t，观测到的轨迹都会产生一个估计值 $\theta_t = \theta(x_{0:t+1},\ u_{0:t})$，该估计值依据学习规则 θ得出，例如最大似然。在探索结束时，智能体会返回最终值 θ_T。尽管这一学习问题内容丰富且本身具有独立研究价值，我们仍将采用简单的有界记忆在线学习规则，并在本文中聚焦于以下决策过程。

序贯决策 智能体的决策采取策略 π 的形式：$(x_{0:t},\ u_{0:t-1}|\theta_t) \mapsto u_t$，在已知当前参数 θ_t 的情况下，将过去的轨迹映射为未来的输入。我们用 Π_γ 表示满足幅度约束 γ 的策略集合。序贯决策过程总结于算法 1 中，并在图 1 中进行了说明。主动探索的目标是选择输入，使轨迹对于使用 f 估计 $f_\star$ 尽可能具有信息量，具体如下所述。

问题 对于任意给定学习规则 θ 的动力学学习模型 f ˆ，以及固定数量的观测 T，目标是寻找一种探索策略 π，使探索结束时学习得到的模型尽可能接近 $f_\star$。形式化地，我们将参数 θ 的估计误差定义为

$$\varepsilon(\theta) = \|f(.,\theta) - f_\star\|_{L^2}, \tag{2.3}$$

并寻找一种策略，使其在时间 T 对 $f_\star$ 给出最佳估计：

$$\min_{\pi \in \Pi_\gamma} \ \mathbb{E}[\varepsilon(\theta_T)|\pi], \tag{2.4}$$

其中，期望取自随机动力学（2.1）以及策略可能引入的随机性。

实际考量 除样本效率目标函数 (2.4) 外，我们还非常重视从第 1 节的观测中总结出的以下三项实际要点。自适应性：与分幕式规划不同，后者的 $\pi(.|\theta)$ 在较长时间段内相对于 θ 保持不变；我们的策略则希望

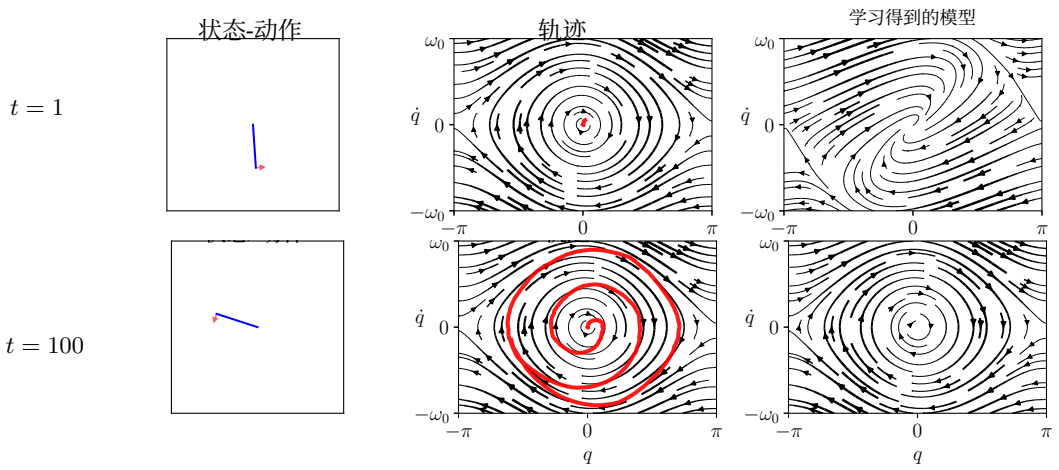


图 *1.* 摆的动力学主动探索示意图。在每个时间步，都会选择一个动作（**左**）。决策依据过去的轨迹（**中**）作出，以学习动力学模型 f（**右**），进而预测下一个状态。

能够在每个时间步适应新的观测。计算效率：评估策略 π 所需的计算资源应有限。灵活性：我们的策略应适用于广泛的模型 f，下面的示例将对此进行说明。

示例 2.1（阻尼摆）。扭矩 u 驱动的摆的角度 q 满足以下非线性微分方程：

$$\ddot{q} + \alpha\dot{q} + \omega_0^2 \sin q = bu. \tag{2.5}$$

这个二阶系统可以用二维状态变量 $x = (q,\ \dot{q})^\top \in \mathbb{R}^2$ 和非线性映射 $f_\star(q,\ \dot{q},\ u) = (\dot{q},\ -\alpha\dot{q} - \omega_0^2 \sin\ q + bu)^\top$来描述。

示例 2.2（非线性摩擦）。采用相同的符号，对于受到控制力和摩擦力作用的质量，其动力学可表示为（为简单起见，考虑一维系统） $f_\star(q,\ \dot{q},\ u) = (\dot{q},\ \text{friction}(q,\ \dot{q}) + bu)^\top$。众所周知，摩擦力很难建模。一种可能的做法是采用非线性函数 $\text{friction}(q,\ \dot{q}) = -\alpha|\dot{q}|q$（Zhang 等人，2014年）。

模型的选择至关重要，并取决于对系统的先验知识。在示例 2.1 中，如果已知系统的标量参数，那么学习动力学就归结为线性回归。相反，如果像示例 2.2 中那样，对系统结构知之甚少或一无所知，那么拟合更复杂的参数模型会更为理想。例如，神经网络近来的成功，以及它们表达复杂函数和在线训练的能力，使其成为探索任务的一个有前景的选择（Goodfellow 等，2016）。

重要的是，探索策略应依赖于学习得到的模型，并驱动系统进入不确定性较高的区域。相比之下，随机探索无法对非线性系统进行全局探索：通常，摩擦力会将系统拉向不动点，而学习环境则需要大振幅轨迹，因此激励需要具备时间相干性。

3. 探索的信息论观点

给定一个动力学模型，我们该如何选择输入，以便在相空间中高效移动至信息丰富状态？这一选择应由轨迹所提供的、关于学习模型的信息量度来指导。本节将转向信息论，并研究最简单的模型形式：线性模型不仅为我们提供了自然的学习规则，也提供了一种探索的信息论度量。随后，我们将利用这一准则来定义优化目标。

3.1. 线性模型

一类重要的模型是参数具有仿射依赖关系的函数类。在本文其余部分中，我们将“线性”一词用于表示仿射依赖关系。需要注意的是，“线性”一词指的是模型对参数的依赖关系，但状态-动作对 z 的依赖关系通常仍假设为非线性。

定义 3.1（线性模型）。参数中最一般的线性依赖形式为

$$f(z, \theta) = V(z) \times \theta + c(z), \tag{3.1}$$

其中，特征 $V(z) \in \mathbb{R}^{d\times n}$ 和 $c(z) \in \mathbb{R}^d$ 均与 θ无关。我们将 V的各行记为 $v^{(j)} \in \mathbb{R}^n,\ 1 \le j \le d$，并定义 $V_t := V(z_t)$。

线性模型这一类因多种原因至关重要。首先，动力学通常可以表示为若干定义明确的标量参数的仿射函数，因此可以通过普通

最小二乘法高效地进行估计（Tedrake，2022年）。此外，还可以从最优实验设计理论中推导出适用于线性模型的自然数学探索目标函数（见第 3.2 节）。最后，适用于线性模型的计算方法也可以推广到非线性模型，正如我们将在第 5 节中看到的那样。

示例 3.1（学习摆）。当示例 2.1 中的摆仅有参数 $\theta = (\omega_0^2,\ \alpha,\ b)^\top$ 未知时，可以通过定义以下内容，使用式 (3.1) 中定义的线性模型来学习其动力学，其中 $n = 3$

$$V(z) = \begin{pmatrix} 0 & 0 & 0 \\ -\sin\phi & -\dot{\phi} & u \end{pmatrix}, \quad c(z) = (\dot{\phi}, 0)^\top. \quad (3.2)$$

真实的动力学函数为 $f_\star(z) = V(z) \times \theta_\star + c(z)$，其中 $\theta_\star$ 为真实参数。

线性模型的一个重要情形是，参数按照按行依赖的形式以矩阵形式分离，如下所示。

定义 3.2（矩阵模型）。我们将以下形式的参数化称为矩阵模型

$$f(z, \Theta) = \Theta \times \phi(z), \quad (3.3)$$

其中 $\phi:\ \mathbb{R}^{d+m} \to \mathbb{R}^{n'}$ 是特征映射，$\Theta \in \mathbb{R}^{d\times r}$ 是参数矩阵。通过定义 θ 为 Θ 的向量化，且其大小为 $n = d \times n'$，并将特征定义为分块矩阵 $V = \mathrm{diag}(\phi^\top,\ \ldots,\ \phi^\top) \in \mathbb{R}^{d\times n}$，即可得到式 (3.1) 的结构。

例 3.2（线性动力学）。考虑线性时不变系统的情形：$f_\star(x,\ u) = A_\star x + B_\star u$。一种自然的参数化形式是 $f(z, \Theta) = \Theta \times z$，其中参数为 $\Theta = (A\ B) \in \mathbb{R}^{d\times(d+m)}$。这是一个矩阵模型（3.3），其中 $\phi(z) = z$ 以及 $n' = d + m$。

示例 3.3（随机傅里叶特征）。在（Schultheis 等，2020）中，非线性动力学采用随机傅里叶特征模型进行建模，其形式为（3.3），其中 $\phi(z)$ 是一个随机特征。

3.2. 最优实验设计

设 y 为一个观测，其分布 $p(y|\pi,\ \theta)$ 取决于参数 $\theta \in \mathbb{R}^n$ 和决策变量（或设计）π。实验设计理论定量回答了以下问题：当采取决策 π 时，观测 y 对于估计 θ 提供了多少信息？信息被定义为费舍尔信息矩阵（Fedorov，2010）的标量函数，具体如下。

定义 3.3（费舍尔信息和信息增益）。我们将 $\ell(y,\ \theta) = \log\ p(y|\theta)$ 称为该分布的对数似然。观测 y 在某个参数值 θ 处的观测费舍尔信息矩阵（Gelman 等人，2004）为

$$I(y, \theta) = -\frac{\partial^2 \ell}{\partial \theta^2}(y, \theta) \quad \in \mathbb{R}^{n\times n}. \quad (3.4)$$

D-最优信息增益定义为

$$g(\pi|\theta) = \log\det\left(\mathbb{E}\left[I(y, \theta)\middle|\pi, \theta\right]\right). \quad (3.5)$$

示例 3.4（线性回归）。在线性回归中，$\pi \in \mathbb{R}^d$ 是回归量，$y = \pi^\top\theta + w \in \mathbb{R}^d$ 带有噪声 $w \sim \mathcal{N}(0,\ I_d)$，从而得到 $I(y,\ \theta) = \pi\pi^\top \in \mathbb{R}^{d\times d}$以及 $g(\pi) = \log\det \pi\pi^\top$。

该信息增益量化了设计为 π 的观测所提供的关于 θ 的信息，也可以将其解释为参数向量 θ 的置信椭球体积。还可以使用其他泛函代替 log det，从而得到其他最优性准则。D-最优性准则具有尺度不变性这一性质（Pukelsheim，2006年），并且适用于我们的在线设置，因为它提供了一个简单的秩一更新公式，正如我们将在第 4 节中看到的那样。D-最优设计最大化信息增益：

$$\max_{\pi} \quad g(\pi|\theta). \quad (3.6)$$

从贝叶斯视角来看，D-最优性会使 θ 上的期望后验分布的熵最小化，或者等价地，使当前先验与期望后验之间的互信息最大化（Chaloner & Verdinelli, 1995）。

3.3. 用于探索的最优信息输入

本节将最优实验设计框架应用于我们的动力学场景。在该场景中，高斯噪声假设和模型的线性结构使得精确计算成为可能。观测是轨迹 $y = (x_{0:t},\ u_{0:t-1})$，设计则是生成输入 $u_{0:t-1}$ 的策略 π。

定义3.4（Gram 矩阵）。一个重要的量是特征的 Gram 矩阵，定义如下：

$$M_t = \sum_{s=0}^{t-1} {V_s}^\top V_s \quad \in \mathbb{R}^{n\times n}. \quad (3.7)$$

命题 1. 假设轨迹的数据生成分布服从线性模型 (3.1)，为简单起见，令 $c = 0$。则 θ 的最大似然估计为

$$\hat{\theta}(y) = M_t^{-1} \sum_{s=0}^{t-1} {V_s}^\top x_{s+1} \quad (3.8)$$

而费舍尔信息矩阵为

$$I(y, \theta) = \frac{1}{2\sigma^2} M_t. \quad (3.9)$$

注意，I 并不显式依赖于 θ，而仅依赖于观测，因为模型关于参数是线性的。

由式（3.6）和命题 1 可知，在我们的设定中，D-最优输入满足以下最优控制问题：

$$\begin{aligned} \underset{(z_t)}{\text{maximize}} \quad & \log\det\left(\sum_{t=0}^{T-1} {V_t}^\top V_t\right) \\ \text{subject to} \quad & x_{t+1} = x_t + \mathrm{d}t f(x_t, u_t), \\ & \|u_t\|_2 \le \gamma^2, \quad 0 \le t \le T-1, \end{aligned} \tag{3.10}$$

为简单起见，我们忽略动力学中的噪声，并回顾 $V_t = V(z_t)$。

注 3.1（矩阵模型的最优设计）。对于形如 (3.3) 的模型，只需定义 $V_t := \phi(z_t)^\top$，即可轻松证明得到一个等价目标函数，而不是 $V_t = V(z_t)$。

3.4. 序贯学习

线性模型可以使用估计器 (3.8) 的递归最小二乘公式进行在线学习。为简单起见，假设 $d = 1, c = 0$，并记 $v_t = V(z_t)^\top \in \mathbb{R}^n$，在线学习的形式为

$$\theta_{t+1} = \theta_t - M_t^{-1} v_t ({v_t}^\top \theta_t - x_{t+1}) \tag{3.11a}$$

$$= \theta_t - H_t \nabla \ell_t(\theta_t). \tag{3.11b}$$

其中平方误差损失为 $\ell_t(\theta) = \frac{1}{2} \times \|{v_t}^\top \theta - x_{t+1}\|_2^2$，矩阵步长为 $H_t := M_t^{-1}$。方程 (3.11b) 清楚地表明，递归最小二乘是一个在线梯度下降步骤。学习的内存成本为 $\mathcal{O}(n^2)$，用于存储 M_t 和 θ_t。对于 $d > 1$，每一行都有一次更新，因此每个时间步有 d 次更新 t。

4. 使用 FLEX 进行自适应 D-最优探索

我们采用 D-最优性（3.10）作为探索策略的目标函数。然而，该问题是非凸的，提供数值解在计算上具有挑战性。此外，动力学约束是未知的，只能根据当前对动力学的认识进行近似，而这种认识会在每个时间步得到改进。尽管以往的方法选择了具有较大时间范围的分幕式优化（Wagenmaker 等人，2021年；Schultheis 等，2020年），但理想情况下，应以与输入收集相同的频率更新输入的选择。在本节中，我们介绍 FLEX（算法2），这是一种计算复杂度较低的自适应 D-最优探索算法。在本节中，我们使用第3节所定义的线性模型记法。我们将在第 5 节展示，这一形式体系如何扩展到任意的非线性模型。

4.1. 一步超前信息增益

由于我们追求最低的复杂度和自适应性，因此在时间 t，我们选择将计算工作仅用于选择下一个输入 u_t。我们希望定义输入 u_t 的信息量度量 F_t，并求解以下形式的一系列问题

$$\begin{aligned} u_t \in \quad & \underset{u \in \mathbb{R}^m}{\text{argmax}} \quad F_t(u) \\ & \text{subject to} \quad \|u\|_2 \le \gamma^2. \end{aligned} \tag{4.1}$$

如前所述，我们非常重视求解（4.1）所需的计算时间。

函数 F_t 应量化 u_t带来的信息，我们在第3节的信息论考量中推导出了其数学表达式。因此，我们希望问题序列（4.1）是问题（3.10）的近似。

可以按如下方式推导出贪心近似。在时刻 t，已知过去的轨迹 $z_{0:t}$，而选择 u_t会立即确定下一个状态 x_{t+1}。我们将 F_t定义为截断于 $t+1$处的预测信息增益。

定义4.1（预测信息增益）。令 $x(u) := x_t + \mathrm{d}t f(x_t, u, \theta_t)$ 为一步前瞻状态预测，且 $z(u) := (x(u), 0)$，我们定义

$$F_t(u) = G(z(u)) \tag{4.2}$$

with

$$G(z) = \log\det(M_t + V(z)^\top V(z)). \tag{4.3}$$

当 V的秩为一时，可以推导出一个更简单的公式。

引理4.1（行列式引理）。通过选择特征矩阵 V 中的第 $1 \le k \le d$ 行 $v := v^{(k)}$ 并近似 ${V}^\top V \simeq v v^\top$，信息增益的表达式可简化如下：

$$G(z) = \log\det M_t + v^\top(z) M_t^{-1} v(z). \tag{4.4}$$

在本文的其余部分，我们采用引理4.1的秩一近似。注意，行 v 的索引 k 可以随机抽取，也可以利用先验知识进行选择：如果模型的第 k 个分量对参数敏感，则 V 的第 k 行具有信息量。也可以为这一选择设计一个数值判据。

4.2. 计算 D-最优输入

对于线性系统，近期研究表明，贪心 D最优策略能够取得良好的探索表现（Blanke 与 Lelarge，2022）。然而，对于非线性系统，$v(z)$ 的非线性使最大化问题 (4.1) 颇具挑战。为得到更简单的优化问题，我们针对状态对模型进行线性化。直观地说，由于我们在 t 和 $t + \mathrm{d}t$ 之间进行规划，状态的相应变化很小，因此对映射 $z \mapsto v(z)$ 使用线性近似是合理的。

算法2快速线性化探索（FLEX）

输入模型 f、时域 T、时间步长 $\mathrm{d}t$、初始估计值 θ_0
输出参数估计 θ_T
对于 $0 \le t \le T-1$，**执行**
  根据 (4.6) 计算 Q_t, b_t
  选择 $u_t \in \underset{u^\top u \le \gamma^2}{\operatorname{argmax}} \; u^\top Q_t u - 2b_t^\top u$ (命题 2)
  观测 $x_{t+1} = x_t + \mathrm{d}t\, f_\star(x_t, u_t) + w_t$
  计算 $\ell_t(\theta) = \frac{1}{2}\| f(x_t, u_t, \theta) - \underbrace{(x_{t+1} - x_t)/\mathrm{d}t}_{\text{如 (5.6) 所示}} \|_2^2$
  更新 $\theta_{t+1} = \theta_t - H_t \nabla \ell_t(\theta_t)$
结束循环

命题 2. 将目标函数 (4.2) 关于 $\mathrm{d}t$ 进行一阶线性化，可得到优化问题 (4.1) 的如下近似：

$$\begin{aligned} \underset{u \in \mathbb{R}^m}{\text{maximize}} \quad & u^\top Q u - 2b^\top u \\ \text{subject to} \quad & \|u\|_2^2 \le \gamma^2, \end{aligned} \tag{4.5}$$

其中，Q 和 b 是根据 Gram 矩阵 $M := M_t \in \mathbb{R}^{n\times n}$、向量 v 以及导数 $D := \partial v/\partial x \in \mathbb{R}^{n\times d}$和 $B := \mathrm{d}t\, \partial f/\partial u \in \mathbb{R}^{d\times m}$计算得出，这些量均在 $\bar{z} := z(u=0)$处求值，具体如下

$$\begin{aligned} Q &= B^\top D^\top M^{-1} D B \quad \in \mathbb{R}^{m\times m}, \\ b &= -B^\top D^\top M^{-1} v \quad \in \mathbb{R}^m. \end{aligned} \tag{4.6}$$

注 4.1（线性动力学）。 当动力学如例 3.2 中那样为线性时，由例 3.2 和注 3.1 可知 $v(z) = z$，因此 $\partial v/\partial x = I_{n,\,d}$和 $M_t = \sum_{s=0}^{t-1} z_s z_s{}^\top \in \mathbb{R}^{d\times d}$。然后，应用命题 2 可得 $Q = B^\top M_t B$ 和 $b = B^\top M_t (I_d + \mathrm{d}tA) x_t$，由此我们恰好得到 (Blanke & Lelarge, 2022) 中的贪心最优设计算法。

我们的探索策略总结于算法2。需要注意的是，由于该策略在每个时间步都优化贪心目标函数，因此它本质上是自适应的，且不需要知道时间范围 T。以下结果表明，求解 (4.5) 的成本很低，从而确保我们的策略在计算上是高效的。

命题 3. 问题 (4.5) 可以通过标量求根和一次 $m \times m$ 矩阵特征值分解，以较低的成本进行数值求解。

5. 从线性模型到非线性模型

对于无法获得有关动力学结构的先验信息的情况（参见示例 2.2），将第 4 节中推导的策略推广到参数上不呈线性的更复杂模型是可取的。在本节中，我们将算法2扩展到一般的非线性参数模型。我们假设 f 关于 z 和 θ 是二阶可微的。

5.1. 线性化模型

第 4 节的推导基于模型对参数的线性依赖。对于非线性模型，一个自然的思路是对模型进行线性展开：假设参数向量接近收敛值 $\theta_\star$，我们可以将 f 关于 $\theta - \theta_\star$ 线性化到一阶：

$$f(z, \theta) \simeq f(z, \theta_\star) + \frac{\partial f}{\partial \theta}(z, \theta_\star) \times (\theta - \theta_\star). \tag{5.1}$$

在该线性近似成立的极限情况下，f 将是一个具有以下特征的线性模型

$$V(z) = \frac{\partial f}{\partial \theta}(z, \theta_\star) \quad \in \mathbb{R}^{d\times n}. \tag{5.2}$$

基于这一类比，我们可以推广 D 最优实验设计（MacKay, 1992），并希望将第 4 节的结果扩展到非线性模型。

5.2. 使用非线性模型进行在线探索

在我们的动力学框架中，我们预计随着收集到更多观测，近似式 (5.1) 会越来越准确，因此有必要将第 4 节提出的探索策略推广到非线性模型。线性化模型 (5.1) 的特征未知，因为雅可比矩阵 (5.2) 通常取决于未知参数 $\theta_\star$。不过，在轨迹的每个时间步，我们都可以用当前估计值 θ_t 近似 $\theta_\star$。定义

$$V_t := \frac{\partial f}{\partial \theta}(z_t, \theta_t), \tag{5.3}$$

我们将定义3.4中的 Gram 矩阵概念以及最优控制问题 (3.10) 推广到非线性模型。注意，由于 V_t 仅取决于 $(z_t,\ \theta_t)$，因此仍可沿轨迹在线计算 Gram 矩阵。同样，我们通过定义以下内容，将第 4 节提出的方法加以推广：

$$v(z) := \nabla_\theta f^{(k)}(z, \theta_t) \tag{5.4}$$

其中 $1 \le k \le d$ 按引理4.1选取。定义了这些量后，算法2便推广到了任意模型。

注 5.1（与线性模型的一致性）。 当模型如式 (3.1) 所示对参数恰好呈线性时，雅可比矩阵为 $\partial f/\partial \theta = V(z)$，因此式 (5.3) 与定义 3.1 中 V_t 的定义一致。因此，式 (5.3) 可视为将第 4 节的最优实验设计探索推广到非线性模型。

5.3. 计算视角

在算法2中，我们需要计算 $D = \partial v/\partial x$，这相当于计算 f 在 θ 和 z 中的导数：

$$\frac{\partial v}{\partial x} = \left(\frac{\partial^2 f^{(k)}}{\partial x_j \partial \theta_i} \right)_{\substack{1\le i\le n \\ 1\le j\le d}} \in \mathbb{R}^{n\times d}. \tag{5.5}$$

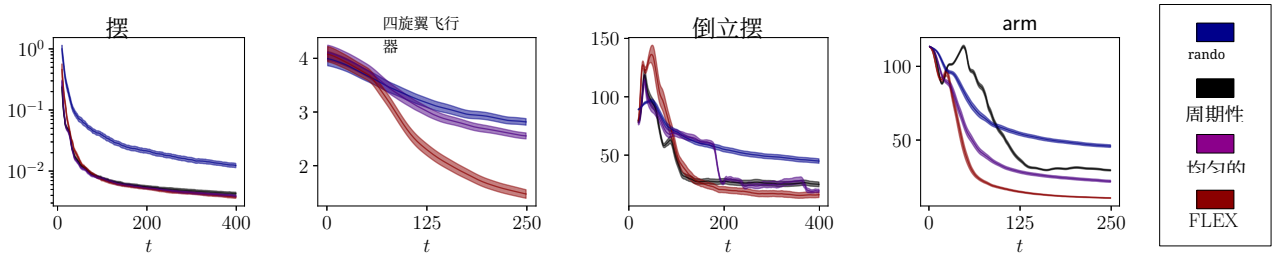


图 2. 不同环境下随时间变化的评估误差 t，取 100 次试验的平均值。

对于神经网络，可以通过自动微分计算该矩阵。求解 (4.5) 的成本不依赖于 n，因此，对于大型模型，计算 D 成为了计算瓶颈。使用自动微分时，后一个操作的复杂度为 $\mathcal{O}(nd)$。

5.4. 序贯学习

我们使用在线梯度下降训练非线性模型：

$$\theta_{t+1} = \theta_t - H_t \nabla \ell_t(\theta_t), \tag{5.6}$$

该方法扩展了（3.11b）。对于神经网络，采用自适应学习率（以及标量 H_t）进行在线学习已被证明是有效的（Bottou, 2012年；Kingma 和 Ba，2015年）。为了使学习更加平滑，可以对一个批次进行平均后再执行梯度更新，但代价是需要存储少量数据点。

6. 实验

我们开展了多项实验来验证我们的方法。我们的代码和演示视频可在 `https://github.com/MB-29/exploration` 获取。有关实验的更多细节请参见附录 C。

6.1. 探索基准

我们首先在经典控制中的各种非线性环境上测试我们的策略，这些环境涵盖 d 和 m 的不同取值。我们根据样本效率，将其性能与各种基线方法进行比较。智能体使用相同的学习模型，但采用不同的探索策略。随机探索会随机抽取输入。对于类摆环境，周期性先知基线会在一个特征模态上激励系统，从而产生大振幅的共振轨迹。一种称为“uniform”的基线最大化轨迹点在状态空间中的距离：它使用 $G(z) = \frac{1}{2}\sum_{s=0}^{t} \|x - x_s\|^2$ 优化目标函数（4.2）。

实验设置 学习模型对动力学包含不同程度的先验知识。摆使用线性模型，其他环境使用神经网络。命题 2 的雅可比矩阵通过自动微分计算。在每个时间步，模型都会使用在固定网格上计算的 (2.3) 进行评估。

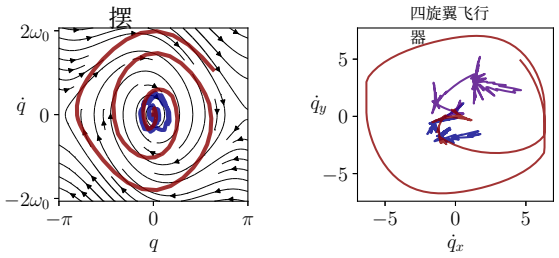


图 3. 相空间中的轨迹。

结果 结果见图 2。我们的算法具有样本高效性，在所有环境中都优于基线方法。我们还在图 3 中展示了使用我们的策略在相空间中获得的轨迹。FLEX 不仅能生成振幅较大的信息丰富的轨迹，更具体地说，它还会投入能量探索具有更高不确定性的区域，这一点不同于基线方法。尽管该策略以贪心方式优化信息量，但有趣的是，它生成的输入具有长期时间相干性。我们在演示视频中对此进行了展示。高维探索将在附录 B 中讨论。

6.2. 时变动力学跟踪

在实际系统中，动力学可能会随时间变化，因此自适应探索对于适应环境变化至关重要。我们通过探索一个时变系统来测试方法的自适应性，并将其与一个回合式智能体进行比较。该系统是一个以恒定周期 T 绕圆运动的中心斥力场，斥力场以一颗恒星为中心。智能体是一艘宇宙飞船，控制变量是加速度。只有在恒星附近，力场才会显著变化：宇宙飞船只有在靠近恒星时才能了解动力学信息，因此需要跟踪恒星。

实验设置 该模型学习力场的中心和半径，从而得到一种非线性参数化表示（见附录 C.2）。利用该模型，三个智能体学习动力学：随机策略、FLEX，以及一个通过在时间范围内反复求解 (3.10) 来规划 D-最优输入的回合式智能体，时间范围为 $T/20 = 50$。在每个时间步，模型都会在参数空间中根据其与当前真实系统参数的

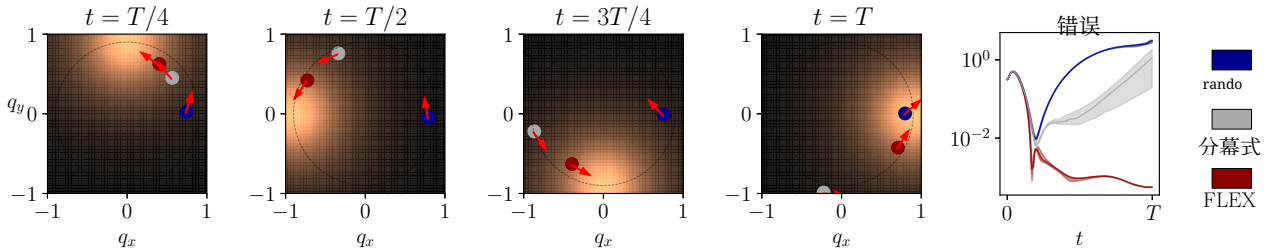


图 *4*。 时变力场系统。 **左图**。 智能体在位置空间中的轨迹（q_x， q_y）。颜色渐变表示力场，红色箭头表示智能体的动作。 **右图**。随时间变化的误差曲线，取 100 次试验的平均值。

距离进行评估。

结果 轨迹和误差曲线如图 4 所示。尽管回合式智能体最初会跟随星星，但由于在一段时间间隔上进行规划所导致的延迟，它最终无法跟踪动力学。另一方面，我们的自适应策略成功地探索了动力学。尽管在两种情况下，输入的规划都具有 T 中的线性时间复杂度，但我们观察到回合式智能体的速度降低了 100 倍。我们认为，这项关于该玩具模型的实验说明了自适应探索策略在现实场景中的相关性。

6.3. 从探索到利用

探索的最终目标是为基于模型的控制获得一个准确的系统模型。为了验证我们的方法在这一框架中的适用性，我们在基于模型的控制任务中评估探索期间学习得到的模型。我们将其与近期的主动探索算法 RHC 和 MAX（Schultheis 等，2020；Shyam 等，2019），以及无模型强化学习算法 SAC（Haarnoja 等，2018）进行比较。

实验设置 我们在 DeepMind 控制套件（Tunyasuvunakool 等，2020）中的摆和倒立摆任务上进行实验，并为其添加了噪声。在整个探索过程中，我们使用（2.3）以及基于模型的控制算法在摆起任务中取得的利用成本来评估学习得到的模型。实验细节见附录 C.3 和 Schultheis 等人（2020），其中还介绍了用于比较的算法的性能。当控制成本低于我们基于模拟任意选定的某个值时，即认为控制任务已解决。

结果 我们的结果呈现在表 1 和图 6 中。就探索而言，我们的算法具有样本高效性，因为它使基于模型的控制算法能够比其他基线方法更快地解决任务。其计算成本较低。

表 *1.* 完成摆起任务所需的观测数量（第一行）以及每次观测的平均计算时间（第二行），其中摆的数据位于上方各行，倒立摆的数据位于下方各行。

方法	RAND	MAX	SAC	RHC	FLEX
样本	$>$ 2k	2000	$>$ 2k	500	50
计算	1	100	2	8	4
样本	$>$ 2k	$>$ 2k	$>$ 2k	600	300
计算	1	20	1.5	2	1.6

7. 相关工作

纯探索任务最近在控制和强化学习领域都引起了广泛关注。人们提出了多种方法来学习动力学并对不确定性进行建模，包括高斯过程（Buisson-Fenet 等人，2020年）、随机傅里叶特征（Schultheis 等，2020年）以及神经网络集成（Shyam 等，2019年；Sekar 等，2020年）。Blanke 与 Lelarge，2022年研究了线性系统的在线探索策略。Mania 等人，2022年对非线性系统的辨识进行了理论研究。最优实验设计向神经网络的扩展分别由 MacKay，1992年针对静态系统提出，以及由 Cohn，1993年针对动力系统提出；后者采用离线算法，并侧重于 G-最优设计。

8. 结论

我们提出了一种基于 D-最优设计的在线自适应探索算法。实验结果表明，该算法在探索和利用方面均具有较高的样本效率、较低的计算成本，并且能够跟踪时变动力学。这些结果为其在实际系统中的应用提供了令人鼓舞的前景。

尽管在线学习规则并非我们工作的重点，但它会影响探索的质量。具体而言，智能体应能够使用有界内存进行学习，以满足嵌入式系统的计算要求。尽管我们采用了相当朴素的在线学习算法，但机器学习和控制领域的最新进展为在线学习动力学带来了希望（Min

等，2022年）。我们方法的计算成本主要来自导数的计算。自动微分是一个活跃的研究领域，我们相信该方向的进展能够降低这一成本。

将 FLEX 推广到更现实的部分可观测状态模型（Good- win & Payne，1977）将会很有意义。另一个研究方向是，在探索-利用权衡中将我们的探索目标用作探索奖励。

参考文献

Amos、B.、Jimenez、I.、Sacks、J.、Boots、B. 和 Kolter、J. Z.《用于端到端规划和控制的可微 MPC》。2018年。

Barto, A. G., Sutton, R. S., and Anderson, C. W. 能够解决复杂学习控制问题的类神经元自适应元件。*IEEE*系统、人类与控制论汇刊，(5):834–846，1983年。

Bitar, D., Kacem, N., and Bouhaddi, N. 对模态交互及其对周期性耦合摆链非线性动力学的影响进行研究。《国际机械科学杂志》，127:130–141，2017年。ISSN 0020-7403。doi: https://doi.org/10.1016/j.ijmecsci.2016.11.030。URL https://www.sciencedirect.com/science/article/pii/S0020740316309614。工程振动国际会议（ICoEV2015）特刊。

Blanke, M. 和 Lelarge, M. 线性动力系统的在线贪心识别。载于 *2022*年 *IEEE* 第*61*届决策与控制会议（*CDC*），第 5363–5368 页，2022年。doi：10.1109/CDC51059.2022.9993030。

Bottou, L. 随机梯度下降技巧。载于 神经网络：实用技巧，第 421–436 页。Springer，2012年。

Buisson-Fenet, M.、Solowjow, F. 和 Trimpe, S. 主动学习高斯过程动力学。载于 动力学与控制学习，第 5–15 页。PMLR，2020年。

Chaloner, K. 和 Verdinelli, I. 贝叶斯实验设计：综述。《统计科学》，10(3):273–304，1995年。ISSN 08834237。URL http://www.jstor.org/ stable/2246015。

Chen, J. 简单中的混沌：双摆导论。2008。

Cohn, D. 使用最优实验设计进行神经网络探索。神经信息处理系统进展，6，1993。

De Simone, M. C., Russo, S., 和 Ruggiero, A. 空气动力学对四旋翼动力学的影响。机械与交通系统最新研究进展，第 111–118 页，2015。

Faessler, M.、Franchi, A. 和 Scaramuzza, D. 考虑旋翼阻力的四旋翼飞行器动力学微分平坦性：用于高速轨迹的精确跟踪。《机器人学与自动化快报》，3(2):620–626，2018年。doi: 10.1109/LRA.2017.2776353。

Fedorov, V. 最优实验设计。*Wiley* 跨学科评论：计算统计学，2(5):581–589，2010年。

Gelman, A., Carlin, J. B., Stern, H. S., 和 Rubin, D. B.贝叶斯数据分析。Chapman and Hall/CRC，第 2 版，2004。

Goodfellow, I., Bengio, Y., 和 Courville, A. 深度学习。麻省理工学院出版社，2016。

Goodwin, G. 和 Payne, R. 动态系统辨识：实验设计与数据分析。发展心理学系列。Academic Press，1977。ISBN 9780122897504。URL https://books.google.fr/books?id=tu5QAAAAMAAJ。

Gupta, N.、Mehra, R. 和 Hall Jr, W. 将最优输入综合应用于飞机参数辨识，1976年。

Haarnoja, T.、Zhou, A.、Abbeel, P. 和 Levine, S. 软演员评论家：采用随机演员的离策略最大熵深度强化学习。载于 Dy, J. 和 Krause, A.（编），第 *35* 届国际机器学习会议论文集，机器学习研究论文集第 80 卷，第 1861–1870 页。PMLR，2018年7月10–15日。网址：https://proceedings.mlr.press/v80/haarnoja18b.html。

Kingma, D. P. 和 Ba, J. Adam：一种随机优化方法。载于 *ICLR*（海报），2015年。网址：http://dblp.uni-trier.de/db/conf/ iclr/iclr2015.html#KingmaB14。

Kirk, D. E. 最优控制理论：导论。Springer，2021年。

Kleff, S.、Meduri, A.、Budhiraja, R.、Mansard, N. 和 Righetti, L. 机械臂的高频非线性模型预测控制。载于 *2021*年 *IEEE*机器人与自动化国际会议（*ICRA*），第 7330–7336 页。IEEE，2021年。

MacKay, D. J. 用于主动数据选择的基于信息的目标函数。神经计算，4(4):590–604，1992 年。

Mania, H.、Jordan, M. I. 和 Recht, B. 带有保证的非线性系统辨识主动学习。机器学习研究期刊，23(32):1–30，2022年。网址：http://jmlr.org/papers/v23/20-807.html。

Min, Y.、Ahn, K. 和 Azizan, N. 通过衔接正交梯度下降与递归最小二乘实现单遍学习。载于*2022*年 *IEEE* 第 *61* 届决策与控制会议 *(CDC)*，第 4720–4725 页，2022年。doi：10.1109/ CDC51059.2022.9992939。

Paszke, A.、Gross, S.、Chintala, S.、Chanan, G.、Yang, E.、DeVito, Z.、Lin, Z.、Desmaison, A.、Antiga, L. 和 Lerer, A. PyTorch 中的自动微分。2017年。

Pukelsheim, F. 实验最优设计。SIAM，2006年。

Schultheis, M., Belousov, B., Abdulsamad, H., and Peters, J. 滚动时域好奇心。收录于Kaelbling, L. P., Kragic, D., and Sugiura, K.（编），机器人学习会议论文集，机器学习研究论文集第100卷，第1278–1288页。PMLR，2020年10月30日至11月1日。网址：https://proceedings.mlr.press/v100/schultheis20a.html。

Sekar, R., Rybkin, O., Daniilidis, K., Abbeel, P., Hafner, D., and Pathak, D. 通过自监督世界模型规划探索。收录于国际机器学习会议，第8583–8592页。PMLR，2020年。

Shyam, P.、Ja´skowski, W. 和 Gomez, F. 基于模型的主动探索。载于 Chaudhuri, K. 和 Salakhutdinov, R.（编），第*36*届国际机器学习会议论文集，机器学习研究论文集第97卷，第5779–5788页。PMLR，2019年6月9–15日。网址：https://proceedings.mlr.press/v97/shyam19a.html。

Sutton, R. S. 和 Barto, A. G. 强化学习：导论。麻省理工学院出版社，2018年。

Tassa, Y.、Erez, T. 和 Todorov, E. 通过在线轨迹优化合成并稳定复杂行为。载于 *2012*年 *IEEE/RSJ* 国际智能机器人与系统会议，第4906–4913页。IEEE，2012年。

Tedrake, R. 欠驱动机器人学。2022年。网址 https://underactuated.csail.mit.edu。

Tunyasuvunakool, S., Muldal, A., Doron, Y., Liu, S., Bohez, S., Merel, J., Erez, T., Lillicrap, T., Heess, N. 和 Tassa, Y. dm control：用于连续控制的软件和任务。*SoftwareImpacts*，6:100022，2020年。ISSN 2665-9638。doi: https://doi.org/10.1016/j.simpa.2020.100022。网址 https://www.sciencedirect.com/science/article/pii/S2665963820300099。

Wagenmaker, A. J., Simchowitz, M. 和 Jamieson, K. 在线性动力系统中的任务最优探索，2021年7月18日至24日。网址 https://proceedings.mlr.press/v139/wagenmaker21a.html。

Yang, T., Tang, H., Bai, C., Liu, J., Hao, J., Meng, Z., Liu, P. 和 Wang, Z. 深度强化学习中的探索：一篇全面综述。*arXiv* 预印本 *arXiv:2109.06668*，2021年。

Zhang, D.、Qi, H.、Wu, X.、Xie, Y. 和 Xu, J. 四旋翼飞行器动力学建模与室内目标跟踪控制方法。《工程数学问题》，2014年，2014年。

A. 关键定义和近似

我们逐步总结 关于 FLEX 从线性模型到非线性模型的近似和定义 s.

表 2. 我们定义的重要量及其在线性和非线性模型中的近似总结。

	线性模型	非线性模型
假设	$f(z,\ \theta) = V(z) \times \theta + c(z)$ 如（3.1）所示	$f(z,\ \theta)$ 可微
学习	在线最小二乘法 = 最大似然估计量 $\theta_{t+1} = \theta_t - H_t \nabla \ell_t(\theta_t)$ 如（3.11b）所示	在线梯度下降 ≃最大似然估计量 $\theta_{t+1} = \theta_t - \eta_t \nabla \ell_t(\theta_t)$ 如（5.6）所示
特征映射	$V(z)$ 如定义 3.1 所示	$V(z) := \frac{\partial f}{\partial \theta}(z,\ \theta_\star)$ 如（5.2）所示 $\theta_\star \simeq \theta_t$
	$V_t := V(z_t)$，如定义 3.1 所示	$V_t := \frac{\partial f}{\partial \theta}(z_t,\ \theta_t)$ 如（5.3）所示
	$v := V^{(k)}$ 如引理4.1所示	
Gram 矩阵	$M_t := \sum_{s=0}^{t-1} V_s^\top V_s$ 如（3.7）所示	
信息矩阵	$I = M_t$ 如（3.9）所示	$I \simeq M_t$ 如 5.2 所示
D-最优信息增益	$G(z) := \log\det(M_t + V(z)^\top V(z))$ 如式（4.3）所示	
秩一近似	$G(z) \simeq \log\det\ M_t + v^\top(z) M_t^{-1} v(z)$ 如式（4.4）所示	
矩阵	$M := M_t$ $D := \partial v / \partial x$ $B := \mathrm{d}t\, \partial f / \partial u$ $Q := B^\top D^\top M^{-1} D B$ $b := -B^\top D^\top M^{-1} v$ 如（5.3）所示	,

B. 高维环境中的探索

我们提出了一个额外的实验，以展示我们算法在高维环境中的表现。我们将把该实验添加到提交稿中。

我们考虑的非线性系统由一串 N 个相互耦合的阻尼摆组成，摩擦未知。每个摆都与其两个最近邻相耦合。只有第一个摆受到驱动，其余摆的运动则源于连续的耦合（Bitar 等人，2017）。对该系统的探索包括确定摆上的摩擦力。

我们认为，该系统能够说明系统辨识的典型设置。例如，在机器人学中，实验旨在测量由大量关节构成的人形机器人的参数。此外，我们提出的系统同时带来了高维和欠驱动方面的挑战。事实上，状态空间的维度为 $d(N) = 2N$，且 N 个摆中只有一个受到驱动，运动会从一个邻居传播到另一个邻居。该实验使我们能够在高维这一具有挑战性的环境中，同时监测 FLEX 的样本效率和计算成本。

设置 该系统采用摩擦系数未知的线性模型进行建模，$\theta \in \mathbb{R}^N$ 这些摩擦系数由观测 $x_t \in \mathbb{R}^{2N}$获得。我们将样本复杂度定义为获得 10^{-2} 参数误差所需的样本数量。对于 N的不同取值，我们测量 FLEX 和 Random 的样本复杂度与计算时间。

结果 我们在表3中给出了结果。这些结果展现了算法在维度 d 增长时的表现。相对于 d 而言，我们的算法 FLEX 似乎能够达到线性样本复杂度，并且计算时间合理；相比之下，随机探索无法进行有效探索，导致超线性样本复杂度。我们的结果表明，尽管计算成本更高，FLEX 在高维环境中仍然保持较高的样本效率和竞争力。

表3。 FLEX 和 Random 算法对 N 耦合摆进行探索的性能。

N	2	5	10	20	50
d	4	10	20	40	100
样本复杂度，Random	20	100	200	500	> 1000
计算，Random	1	2	6	60	250
样本复杂度，FLEX	10	50	100	200	500
计算量，FLEX	3	5	10	80	350

C. 实验细节

C.1. 探索基准

C.1.1. 补充结果

FLEX 生成的输入的时间相干性如图5所示。谱密度显示能量集中在某一频率上，这表明存在一种时间结构，使倒立摆在接近共振频率处受到激励，从而产生信息丰富的轨迹。

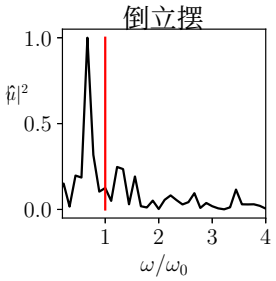


图5。 FLEX 在倒立摆环境中生成的输入的谱密度。

C.1.2. 环境

我们补充介绍所用环境和学习模型的更多细节。我们用 q_x和 q_y 表示平移变量，用 q_ϕ表示角度变量。

环境 1 （摆， $d = 2$， $m = 1$）。 其动力学由式（2.5）给出。

环境 2 （四旋翼飞行器， $d = 6, m = 2$）。 具有非线性摩擦的平面四旋翼飞行器遵循以下方程（Zhangetal., 2014年）

$$\begin{aligned} m\ddot{q_x} &= -(u_1 + u_2)\sin q_\phi - \alpha q_x|\dot{q_x}|\dot{q_x} \\ m\ddot{q_y} &= (u_1 + u_2)\cos q_\phi - \alpha q_y|\dot{q_y}|\dot{q_y} - mg \\ I\ddot{q_\phi} &= r(u_1 - u_2) \end{aligned} \tag{C.1}$$

其中 m 表示质量， I 表示转动惯量。

环境 3 （倒立摆， $d = 4$， $m = 1$）。 我们实现了 Barto 等人，1983 提供的动力学。

环境 4 （**机械臂** / 双摆， $d = 6$， $m = 2$）。 方程见（Chen，2008）。[$d = 4, m = 2$]

C.1.3. 基线方法

随机基线方法返回 $u_t \sim \sqrt[\gamma]{m}\mathcal{U}([-1,1]^m)$。均匀策略通过梯度下降最大化均匀性目标，每个时间步进行 100 次梯度更新。周期策略返回形式为 $u_t = \gamma\sin(\omega_0 t)$ 的输入，其中 ω_0 是系统的一个特征模态。

C.1.4. 模型

我们使用以下学习 第 6.1 节中的模型，宽度为 8，包含一个隐藏层和 tanh非线性，使用 ADAM优化器（Kingma 和 Ba，2015）训练，批量大小为 100。

摆 我们使用示例 2.1 的线性模型，并通过普通最小二乘法对其进行学习。

四旋翼飞行器 我们使用神经网络学习摩擦力，学习率为 $\eta = 0.02$。

倒立摆 我们使用观测值 $\xi = (q_x,\ \dot{q}_x,\ \cos\ q_\phi,\ \sin\ q_\phi,\ \dot{q}_\phi)$ 对 $f(z,\ \theta)\ = a_\theta(\xi)\ + u \times b_\theta(\xi)$ 进行参数化，其中 a_θ 和 b_θ 由使用学习率 $\eta = 0.1$ 训练的神经网络给出。

机械臂 我们使用神经网络学习 $f(.,\ u = 0)$，将其表示为 $\xi = (\cos\ q_{\phi_1},\ \sin\ q_{\phi_1},\ \dot{q}_{\phi_1},\ \cos\ q_{\phi_2},\ \sin\ q_{\phi_2},\ \dot{q}_{\phi_2})$ 的函数，学习率为 $\eta = 0.05$。

C.2. 时变动力学跟踪

宇宙飞船在平面上的状态表示为 $x = (q_x\ \dot{q}_x\ q_y\ \dot{q}_y)^\top$。恒星中心的坐标随时间变化

$$\Big(\kappa_x(t), \kappa_y(t)\Big) = \Big(\cos(2\pi t/T), \sin(2\pi t/T)\Big). \tag{C.2}$$

动力学形式如下

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\dot{q}_x\\ \dot{q}_y\end{pmatrix} = -\frac{1}{1 + \frac{1}{\rho^2}\Big((q_x - \kappa_x)^2 + (q_y - \kappa_y)^2\Big)}\frac{q}{\|q|_2} \tag{C.3}$$

智能体了解动力学及其参数 κ_x，κ_y 和 ρ，并通过在线梯度下降学习这些参数，使用 ADAM 优化器，学习率为 $\eta = 0.01$。

C.3. 从探索到利用

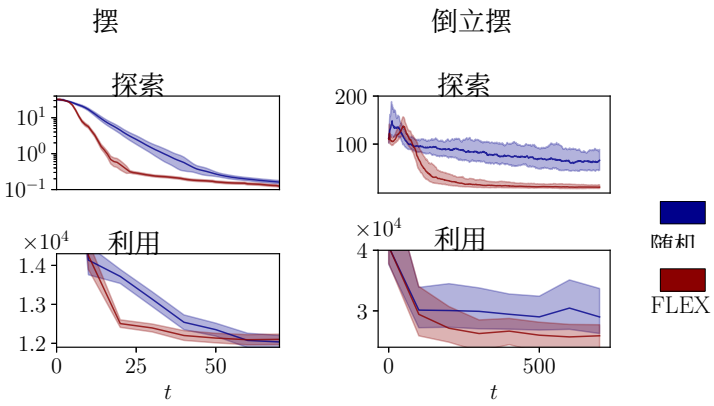


图 6。在摆和倒立摆小车。

我们在动力学中加入噪声：$\sigma = 0.001$ 对于摆，$\sigma = 0.05$ 对于倒立摆。

我们对摆使用线性模型，对倒立摆使用与 C.1 节中相同架构的神经网络模型。动力学被学习为 cos q_ϕ and sin q_ϕ 的函数，而在倒立摆中则被学习为 q_x and $\dot{q}_x$ 的函数。

由于我们的模型是用 PyTorch（Paszke 等人，2017年）实现的，因此我们使用 mpc 软件包和 iLQR 算法进行利用（Amos 等人，2018年）。二次代价为

$$C = 100(1 - \cos q_\phi)^2 + 0.1 \sin^2 q_\phi + 0.1\dot{q}_\phi^2 + 0.001u^2 \tag{C.4}$$

用于摆和

$$C = 100q_x^2 + 100(1 - \cos q_\phi)^2 + 0.1 \sin^2 q_\phi^2 + 0.1\dot{q}_x^2 + 0.1\dot{q}_\phi + 0.001u^2 \tag{C.5}$$

用于倒立摆。将成本值与其他方法进行比较时，数量级才是关键，因为每次实验所使用的利用控制算法都不同。

我们在一台笔记本电脑上测量计算时间，并对 FLEX 和随机策略分别运行 100 次后取平均值，再与 (Schultheis 等，2020) 中的数值进行比较；计算时将随机策略的计算时间设为 1。这里同样只有数量级才重要。

D. 证明

D.1. 命题 1 的证明

证明。 数据生成　分布在已知参数 θ 的条件下，可通过概率链式法则计算　　：

$$p(y|\theta) = (\frac{1}{\sqrt{2\pi\sigma^2}})^t \exp\left(-\frac{1}{2\sigma^2}\sum_{t=0}^{t-1}\|V(z_s)\times\theta - x_{s+1}\|_2^2\right). \tag{D.1}$$

由此可得（省略与 θ相关的常数）

$$\begin{aligned}\ell(y,\theta) &= -\frac{1}{2\sigma^2}\sum_{t=0}^{t-1}\|V(z_t)\times\theta - x_{t+1}\|_2^2. \\ &= -\frac{1}{2\sigma^2}\sum_{j=1}^{d}\sum_{t=0}^{t-1}\left\|v^{(j)}(z_t)\times\theta - x_{t+1}\right\|_2^2.\end{aligned} \tag{D.2}$$

求两次导数并取相反数，可得

$$\begin{aligned}I(y,\theta) &= \frac{1}{2\sigma^2}\sum_{j=1}^{d}\sum_{t=0}^{t-1}v^{(j)}(z_t)v^{(j)}(z_t)^\top \\ &= \frac{1}{2\sigma^2}\sum_{t=0}^{t-1}V(z_s)^\top V(z_s) \\ &= \frac{1}{2\sigma^2}M_t.\end{aligned} \tag{D.3}$$

□

D.2. 命题 2 的证明

证明。 在 $\bar{x} := x_t + \mathrm{d}t\, f(x_t,\ 0)$ 的邻域内进行一阶展开，

$$v(x) = v(\bar{x}) + \frac{\partial v}{\partial x}(\bar{x})\times(x-\bar{x}) + \mathcal{O}\left(\|x-\bar{x}\|^2\right), \tag{D.4}$$

此外， $x(u)$ 在 $u = 0$ 附近的一阶展开为

$$x(u) = \bar{x} + \mathrm{d}tBu + o\left(\mathrm{d}t\right) \tag{D.5}$$

其中， u 的阶为 γ。将展开式（D.5）代入展开式（D.4），得到 $v(z(u))$的一阶展开。省略加性常数后，我们得到所述结果。 □

D.3. 命题 3 的证明

证明。 参见（Blanke 与 Lelarge，2022年）。由于探索旨在产生大振幅轨迹，我们关注最大范数的输入，并假设优化问题中存在等式约束。严格不等式可以通过无约束最小化直接处理。记 $u_\star$ 为（4.5）的一个最小值点，记 $\{\alpha_i\}$ 为 Q的特征值，记 $u_\star^{(i)}$和 b_i 为 $u_\star$ 和 b 在相应正交归一基下的坐标。根据拉格朗日乘子定理，存在一个非零标量 μ，使得 $Qu_\star - b = -\mu\, u_\star$，其中可以对 μ 进行缩放，使 $Q + \mu I_m$ 为非奇异矩阵。对最优条件求逆并展开等式约束，可得

$$u_\star^{(i)} = b_i/(\alpha_i + \mu) \tag{D.6a}$$

$$\sum_i \frac{{b_i}^2}{(\alpha_i + \mu)^2} = \gamma^2. \tag{D.6b}$$

最小化器 $u_\star$ 通过求解（D.6b）确定 μ。 □