

随机动态系统的概率预测

一个上界、一个下界，和中间的空间

徐韬

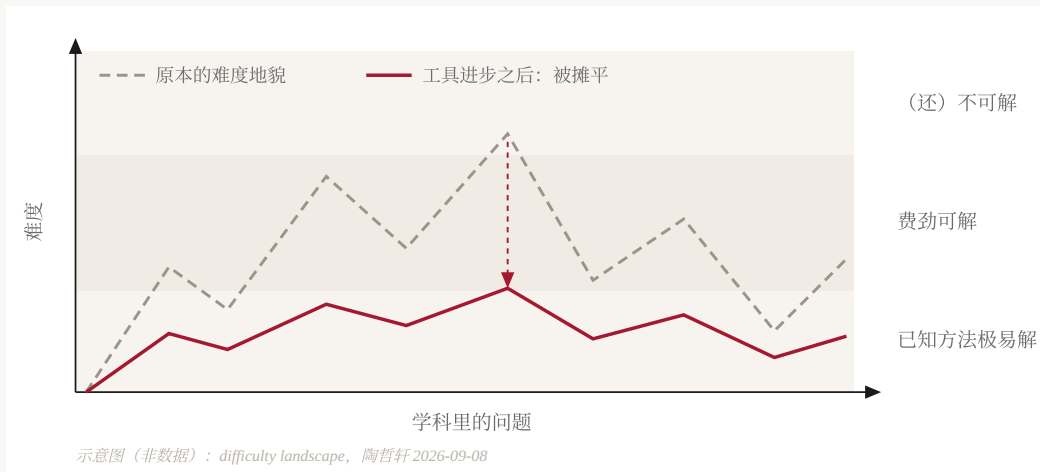
导师：何建平 教授

IWIN-FINS 实验室

上海交通大学自动化与感知学院

出发点：好问题比答案稀缺

AI 把解题变便宜了 —— 问出有意义的问题，反倒成了最稀缺的能力。



陶哲轩（2026-09-08）：每一次工具进步都在降低解题难度，代价是把难度地貌摊平——平到看不出形状，也就再难挑出有前景的问题。

- 问题可以无限生成（比如去算 π 的第 10^{10} 位），但绝大多数并不通向新洞见 —— 海很大，淡水照样会缺。
- "AI 做得到 / 做不到"之间没有一条清晰的线，于是"看出哪个问题有前景"本身成了稀缺资源。

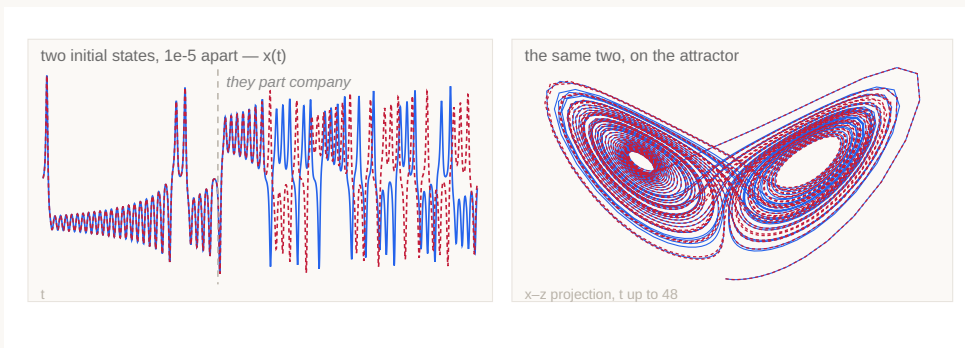
→ 那怎么问出有意义的问题？先看伟大的科学家问的是什么问题。

预测的极限

从"一个数"到"一个分布"

混沌给了一个完整的答案

混沌回答了"方程完全已知时，能预测多远"；可真实系统的方程，从来不会完全已知。



Lorenz (1963): 两个相距 10^{-5} 的状态先是同步，随后彻底分开。

- 它问的 —— 方程完全已知的确定性系统，能预测多远？
- 它做到的 —— 一个完整的答案：初值敏感，能预测多久由 Lyapunov 指数定。
- 它没问的 —— "方程已知"这个前提本身。

→ 但真实系统不是一条已知的确定性方程 —— 所以问题本身得换。

一旦有噪声，预测就不再是一个数，而是一个分布——而分布要能比，就得先有评分规则。

- 有了分布，才谈得上**风险、尾部、稀有事**——决策真正看的就是这些。
- 评分规则用**对数得分**：一个预测密度、一个实际结果，得出一个数。

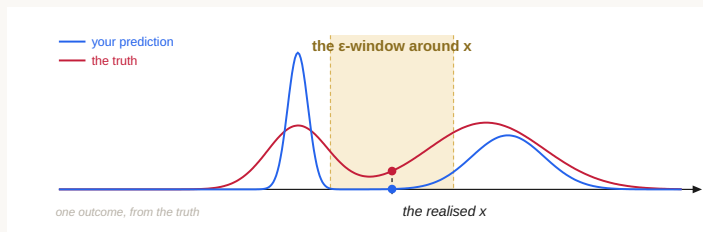
$$S(\hat{p}, x) = \log \hat{p}(x), \quad S(\hat{p}, p) := \mathbb{E}_{x \sim p} [\log \hat{p}(x)]$$

对数得分是标准做法：老老实实报出真实分布时，它的期望最大（“诚实最优”）。这一条先接受，不证。

→ 但标准得分有一个致命处。

从"一个点"到"一个窗口"

模型稍微不对，对数得分就掉到 $-\infty$ —— 而 $-\infty$ 的来源不是数据，是把分数压在"一个点"上。



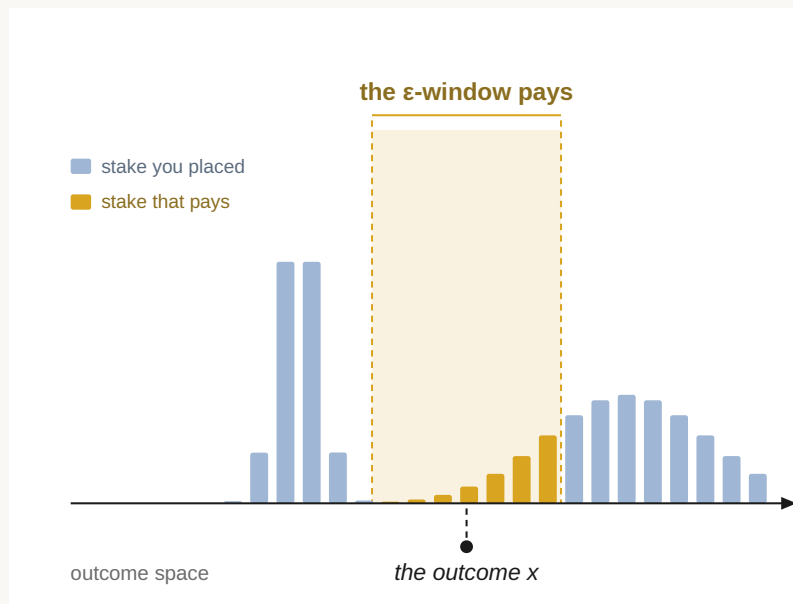
定义 1 (ϵ -对数得分 · Auto. 2025, §III-C)

(i) 得分 —— 对半径 $\epsilon \geq 0$ 和这次实际落到的 x ：读的是盒子里的概率质量，而不是某一点的密度：

$$\mathcal{L}_{\epsilon}(\hat{p}_{\mathbf{x}}, x) := \begin{cases} \log \hat{p}_{\mathbf{x}}(x), & \epsilon = 0 \\ \log \int_{\|s-x\|_{\infty} \leq \epsilon} \hat{p}_{\mathbf{x}}(s) \, ds, & \epsilon > 0 \end{cases} \quad (5)$$

■ ϵ 不是权宜之计 —— 它就是你能承担多少风险。

- 你摊开一份赌注。密度 $\hat{p}_{\mathbf{x}}$ 就是你在每个位置押的份额。
- 赔付规则。落在实现值 x 的 ϵ 邻域内的结果才赔 —— 赔那个窗口里的总份额， $R(x) = \int_{\|s-x\|_{\infty} \leq \epsilon} \hat{p}_{\mathbf{x}}(s) ds$ 。
- 对数效用把这份赔付变成我们的得分： $U(x) = \log R(x) = \mathcal{L}_{\epsilon}(\hat{p}_{\mathbf{x}}, x)$ 。



被追问 —— ϵ -log 分数在 $\epsilon > 0$ 时还 proper 吗？赔付规则和效用定下来之后，最优组合就是那个期望效用最大的组合 —— 它等不等于真实分布，并不重要。proper 是学习任务才需要的性质（那里的目标才是学到真实分布）。

预测的不是下一个状态，而是它的条件分布。

一个随机动态系统，是**你已知的规律**加上**你不知道的扰动**：

$$x_{t+1} = f(x_t, u_t) + w_t, \quad x_t \in \mathbb{R}^{d_x}, u_t \in \mathbb{R}^m$$

- f —— 动力学（你知道、或以为知道的那部分）
- u_t —— 控制输入（你选的那部分）
- w_t —— 过程噪声（你不知道的那部分）

概率预测要给的，是每一步的**条件分布** —— 已知到此刻为止的一切，下一个状态会落在哪里：

$$\hat{p}_{\mathbf{x}_k | \mathbf{x}_{1:k-1}}(\cdot \mid x_{1:k-1})$$

而你只看得到**一条轨迹**：机器真正跑过的那一条。

→ 那么，一个系统本身能预测到多好？

三个问题，对应一个闭式、一条保证、一张能用的地图。

1 一个系统本身，能预测到多好？

有天花板，而且是闭式 —— 只由容差半径、噪声熵、维度决定。

2 模型必然不准时，还能保证什么？

有一条地板 —— 对不确定的动力学给出最坏情形保证，代价还能提前算出来。

3 那这两条线怎么用？

任何预测器都活在中间的空间里；说清你离哪条线更近，就是可以用的语言。

这三问就是我的主线：把动态系统里的不确定性看清楚、绕过去、再用起来。

天花板

Part I — 一个系统本身有多可预测？

可预测性是系统本身的属性，不是方法的属性。

“能预测到多好”被问了几十年；而随机动态系统下的极限，还没有被从概率预测的视角理解。



Edward N. Lorenz

把可预测性提为一个问题

“Predictability: A problem partly solved”

1996

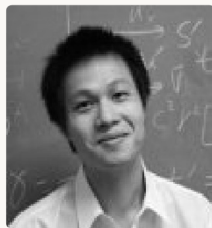


Andrew B. Nobel

序列预测的最优性（理论）

“On optimal sequential prediction for general processes”

2003



Chaoming Song

具体系统的极限（实证）

“Limits of predictability in human mobility”

2010

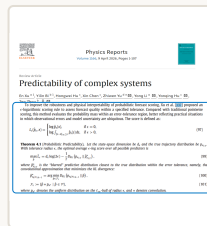


Richard Kleeman

信息论的视角

“Information Theory and Dynamical System Predictability”

2011



Tao Zhou 等

Physics Reports
综述（引用了本工作）

“Predictability of Complex Systems”

2026

→ 而这个极限，其实可以写成一个闭式。

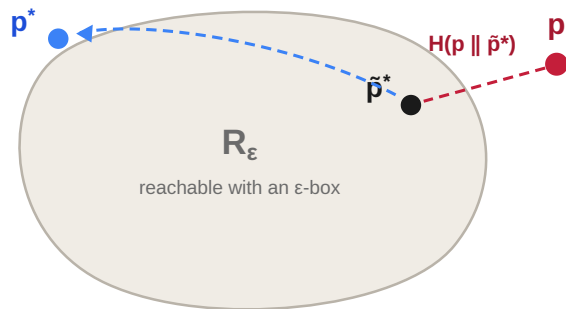
上界是一个公式， $\max$ 取遍每一步的预测器——再多数据也突破不了。

定理 1 (概率可预测性 · Auto. 2025, §IV)

$$\begin{aligned} & \max_{\hat{p}_{\mathbf{x}_k | \mathbf{x}_{1:k-1}} \in \mathcal{P} \forall k \in [1, T]} \bar{\mathcal{L}}_{\epsilon}(\hat{p}_{\mathbf{x}_{1:T}}, p_{\mathbf{x}_{1:T}}) \\ &= \mathbb{I}_{\epsilon > 0} \cdot d_x \log(2\epsilon) - \frac{1}{T} H(p_{\mathbf{x}_{1:T}} \| \tilde{p}_{\mathbf{x}_{1:T}}^*) \end{aligned} \quad (10)$$

- $d_x \log(2\epsilon)$ —— 容差项，每维计一次
- $H(p \| \tilde{p}^*)$ —— 够不到真值的代价（图里那段虚线）

- p — the truth (out of reach)
- $\tilde{p}^*$ — closest member of the reachable set
- p^* — the predictor that attains the ceiling



the space of predictive densities

决定天花板的是三个量

天花板由容差半径、噪声的微分熵与系统维度决定 —— 你的模型不在其中。

容差半径

ϵ

轨迹的微分熵

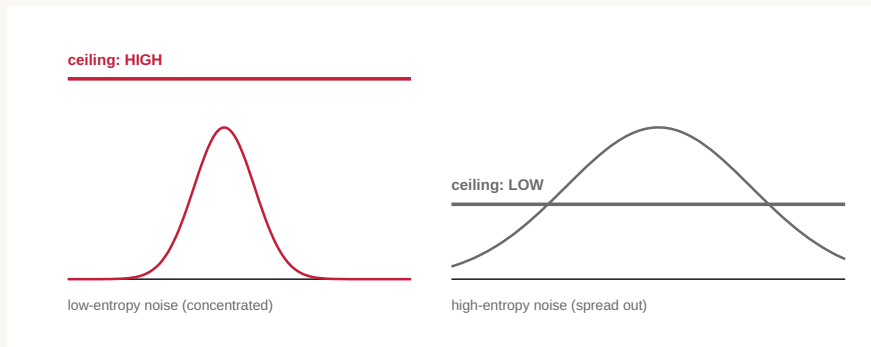
$H_d(p_{\mathbf{x}_{1:T}})$

状态维度

d_x

这三样东西，出现在天花板的一个熵上界里（式 (13) [Auto. 2025]）：

$$\mathbb{I}_{\epsilon>0} d_x \log(2\epsilon) - \frac{1}{T} H_d(p_{\mathbf{x}_{1:T}}) \quad (13)$$



- 天花板随噪声熵线性下降；等号成立当且仅当每一步的真值都可达 —— 那是最好的情形，不是一般情形。

不必是最优的那个 —— 手上这个预测器的期望得分也是闭式。

定理 3 (轨迹得分的逼近 · Auto. 2025, §V)

只要密度连续, 整条轨迹的得分也有闭式, 误差 $O(\epsilon)$:

$$\left| \bar{\mathcal{L}}_{\epsilon}(\hat{p}_{\mathbf{x}_{1:T}}, p_{\mathbf{x}_{1:T}}) - \mathbb{I}_{\epsilon>0} d_x \log(2\epsilon) + \frac{1}{T} H(p_{\mathbf{x}_{1:T}} \| \hat{p}_{\mathbf{x}_{1:T}}) \right| = O(\epsilon)$$

- 这是有限 ϵ 下的 $O(\epsilon)$, 不是渐近结论
- 前提是 $p_{\mathbf{x}_{1:T}}$ 与每个 $\hat{p}_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}$ 连续 ($\mathcal{P}_c$), 不是"任意预测器"
- 注意这里出现的是给定预测器的 $\hat{p}$, 不是最优的 $\tilde{p}^*$

天花板是怎么达到的 [Auto. 2025]: $\tilde{p}_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}^* = \arg \min_{\tilde{p} \in \mathcal{R}_{\epsilon}} H(p_{\mathbf{x}_k|\mathbf{x}_{1:k-1}} \| \tilde{p})$ 是可达集里离真值最近的成员 (式 11); 达到天花板的预测器是它与 ϵ -盒的反卷积, $p^* = \tilde{p}^* * / * p_{\mathbf{u}_{\epsilon}}$ (式 12)。这是定理 1 的 (ii)(iii), 正文因篇幅没有展开。

够 —— 你只看得到一条轨迹，它仍然收敛到期望得分。

定理 4 (单条轨迹上的收敛 · Auto. 2025, §V)

假设过程噪声独立同分布、密度为 $p_{\mathbf{w}}$ ，且预测器知道动力学 —— $\hat{p}_{\mathbf{x}_k | \mathbf{x}_{1:k-1}}(x_k | x_{1:k-1}) = \hat{p}_{\mathbf{w}}(x_k - f(x_{k-1}))$ ：

(i) 期望坍缩成一步：

$$\bar{\mathcal{L}}_{\epsilon}(\hat{p}_{\mathbf{x}_{1:T}}, p_{\mathbf{x}_{1:T}}) = \mathcal{L}_{\epsilon}(\hat{p}_{\mathbf{w}}, p_{\mathbf{w}})$$

(ii) 一条轨迹就够 —— 单条实际路径上的得分依概率收敛到该期望：

$$\lim_{T \rightarrow \infty} \bar{\mathcal{L}}_{\epsilon}(\hat{p}_{\mathbf{x}_{1:T}}, x_{1:T}) \stackrel{P}{=} \bar{\mathcal{L}}_{\epsilon}(\hat{p}_{\mathbf{x}_{1:T}}, p_{\mathbf{x}_{1:T}})$$

(iii) 而且是蒙特卡洛速度 —— 只要 $\mathbb{E}_{\mathbf{w}} \mathcal{L}_{\epsilon}(\hat{p}_{\mathbf{w}}, \mathbf{w})^2 < \infty$ ：

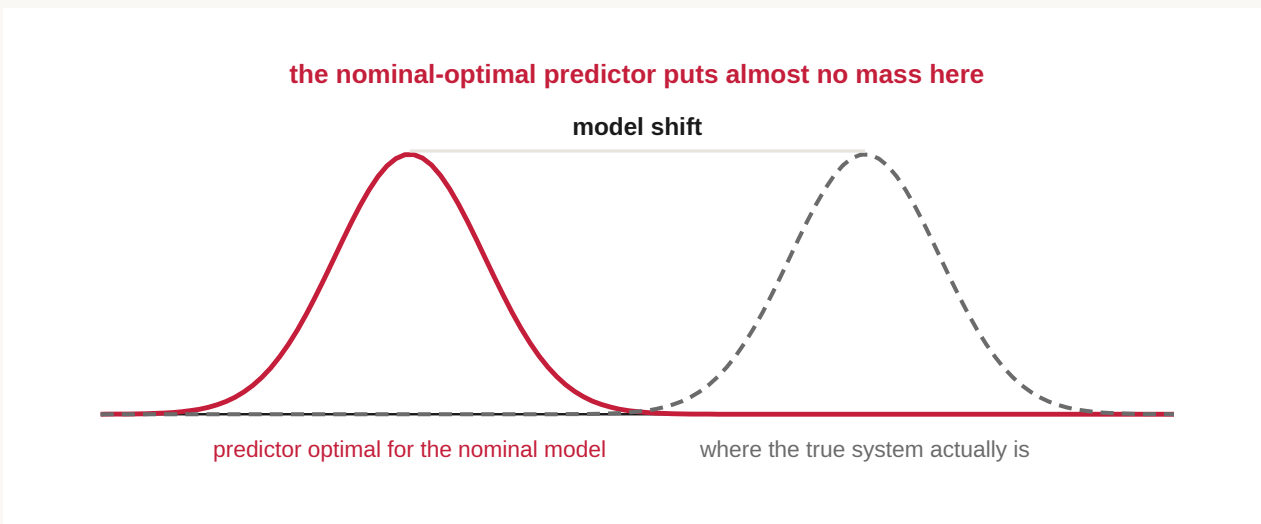
$$\Pr \left\{ \left| \bar{\mathcal{L}}_{\epsilon}(\hat{p}_{\mathbf{x}_{1:T}}, x_{1:T}) - \bar{\mathcal{L}}_{\epsilon}(\hat{p}_{\mathbf{x}_{1:T}}, p_{\mathbf{x}_{1:T}}) \right| \geq \delta \right\} = O\left(\frac{1}{\sqrt{T}}\right) \quad \forall \delta > 0$$

地板

Part II — 模型错了，还能保证什么？

对模型最优 $\neq$ 对系统可用

对着手里的模型最优的预测器，会在真实系统上崩掉。



→ 那还能保证什么？

把模型的不确定与噪声的不确定放进同一个集合，再对它取最坏情形。

定义 2 (条件锥矩不确定集 · TAC. 2026, §IV-B)

给定 k 、 z 与标称数据 $(\bar{f}_k, \bar{\mu}_k, \bar{\Sigma}_k)$ ，集合 $\mathcal{I}_k(z)$ 收集所有你排除不掉的模型 $\mathbf{x}_{k+1} = f_k(z) + \mathbf{w}_k$ —— 三条锥约束、四个旋钮 $\gamma_0, \dots, \gamma_3$ ：

- 动力学贴近标称： $\|f_k(z) - \bar{f}_k(z)\|_2^2 \leq \gamma_0(z)$
- 噪声均值贴近标称： $\|\mathbb{E}(\mathbf{w}_k) - \bar{\mu}_k\|_{\bar{\Sigma}_k^{-1}}^2 \leq \gamma_1$
- 噪声协方差被夹住： $\gamma_3 \bar{\Sigma}_k \preceq \mathbb{E}(\mathbf{w}_k - \bar{\mu}_k)(\mathbf{w}_k - \bar{\mu}_k)^\top \preceq \gamma_2 \bar{\Sigma}_k$

$$(\mathbf{P}_0) : \sup_{\mathcal{F} \in \mathfrak{F}} \inf_{\mathcal{P} \in \mathfrak{P}} J_0^{\mathcal{F}, \mathcal{P}}(z_0) \quad (4)$$

- 外层 $\sup$ 取遍预测器族 $\mathcal{F}$ （函数），内层 $\inf$ 取遍系统核族 $\mathcal{P}$ （概率测度）—— 两侧都是无穷维，难就难在这。

从无穷维到能算，靠的是四步等价改写，不是松弛。

定理 1 (Bellman 方程 · TAC. 2026, §V-A)

取 $V_T^*(z) = 0$ ，鲁棒值函数满足

$$V_k^*(z) = \sup_{\hat{p}_k \in \mathcal{F}} \inf_{\rho_k \in \mathcal{I}_k(z)} \int_{\mathcal{X}} \mathcal{L}(\hat{p}_k, x) + V_{k+1}^*(x, \pi_{k+1}(x)) \, d\rho_k(x) \quad (7)$$

- 精确递推，不是松弛；内层仍在概率测度上 —— 让它有限维是另一步

四步：① 化成锥规划 $\mathbf{D}_1$ ；② 对偶化 + 最优性条件（定理 2）：最优预测器几乎处处落在指数族里， $\hat{p}_k(x) \stackrel{a.s.}{\propto} \exp\{-x^\top \theta_1 x - x^\top \theta_2 - V_{k+1}^*(x, \pi_{k+1}(x))\}$ —— “找解”变成“找两个参数”；③ 参数化 $\mathbf{P}_1 \rightarrow \mathbf{P}_2$ ；④ 化为凸-非凹的 minimax（式 12）。

化到欧氏空间，两条放松各得一解

论文把它等价地化到欧氏空间，从两个方向各得一个闭式解——而且“离最优多远”能离线算出来。

Noise-DRPP —— 把不确定集缩小

- 假设动力学精确 ($\bar{\nu}_k = \nu_k$)，只对噪声的不确定性鲁棒——你信 f 、不信噪声时选它

定理 3 (*Noise-DRPP · TAC. 2026, §VI-B*)

$$\hat{p}_k^* \sim \mathcal{N}(\bar{\nu}_k + \bar{\mu}_k, \gamma_2 \bar{\Sigma}_k)$$

代价：无需优化。

Eig-DRPP —— 限制预测器

- 保留完整的不确定集（含 f ），改为约束预测协方差 ($\hat{Q}_k = Q_k$)——你怀疑模型本身时选它

定理 5 (*Eig-DRPP · TAC. 2026, §VII-B*)

$$\hat{p}_k^* \sim \mathcal{N}(\bar{\nu}_k + \bar{\mu}_k, Q_k \hat{\Lambda}_k^* Q_k^\top)$$

代价： d_x 个凸子问题，取最优。

- 两个解离全局最优还有多远，由**最优性间隙**（定理 6）给出，而且**可以离线算出**——只与不确定集的参数有关。

这个间隙有闭式，而且只用到不确定集的参数。

定理 6 (*Eig-DRPP 的最优性间隙* · TAC. 2026, §VII-B)

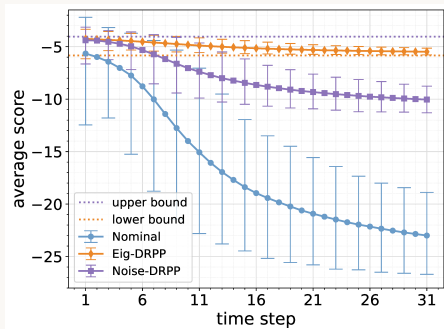
(ii) 到全局最优的间隙，给定 $\gamma_0(z) \leq \Gamma_0$:

$$V_k^*(z) - V_k^{\mathcal{F}^{\text{Eig}}}(z) \leq \sum_{t=k}^{T-1} -\frac{1}{2} \left\{ \underbrace{d_x + \log \det(\gamma_2 \bar{\Sigma}_t)}_{\text{(a) from the upper bound}} - \underbrace{\sum_{i=1}^{d_x} \left[\log(\hat{\lambda}_{i,t}^*) + \gamma_2 \frac{\lambda_{i,t}}{\hat{\lambda}_{i,t}^*} \right] + \frac{2\sqrt{\Gamma_0 \gamma_1 \lambda_{j_t^*,t}} + \Gamma_0}{\hat{\lambda}_{j_t^*,t}}}_{\text{(b) from the lower bound}} \right\} \quad (17)$$

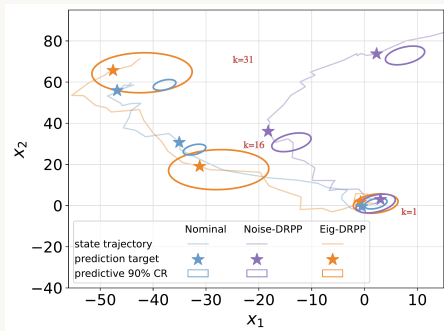
- **(a)** 来自上界（定理 4 [TAC. 2026]），**(b)** 来自下界——间隙是两者之差，而且可以离线算出。
- 所以间隙不是“差多少不知道”，而是一张标好价的价目表。

实验：对抗动力学下谁站得住

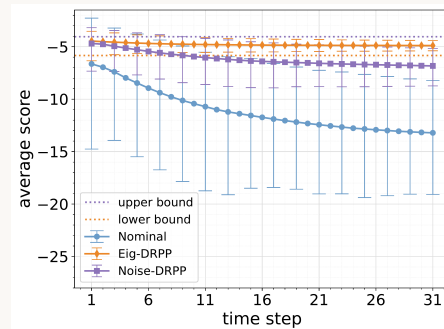
对抗动力学下标称预测器崩掉 —— DRPP 守得住。



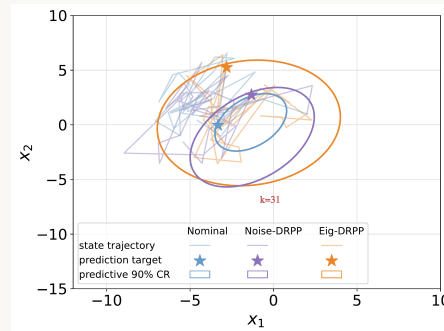
对抗，零控制 —— 标称预测器崩掉。



90% 区域，零控制 —— 星星落在标称椭圆之外。



对抗，LQR —— 同样的故事。

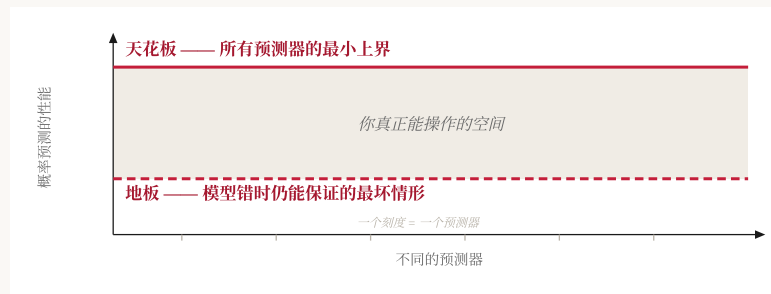


90% 区域，LQR —— 同样。

合起来

一个上界、一个下界，和中间的空间

每个预测器都活在两条线之间



每一个预测器都活在一条打不破的天花板与一条不该越过的地板之间。

- **天花板** [Auto. 2025] 给所有预测器封顶；**地板** [TAC. 2026] 是模型错时还能守住的底线。

被追问 —— 两条线用的分数不同，凭什么画在同一根轴上？下半场优化的是普通对数得分 $\log \hat{p}(x)$ ，上半场的天花板是 ϵ -对数得分。但定理 3 [Auto. 2025] 把 ϵ -对数得分写成对数得分：两者只差一个常数 $d_x \log(2\epsilon)$ 和一个 $O(\epsilon)$ 误差。常数不影响排序，所以这两个目标对预测器的排序，在 $O(\epsilon)$ 意义下是一致的。

→ 那夹在中间的是什么？

剩下的问题很具体：用更多信息，把上下界之间的 gap 一点点收窄。

- **更多信息** —— 上下文、控制输入、**轨迹数据**：天花板那侧能不能靠"看得更多"抬起来，地板那侧能不能靠"知道得更准"逼近最优？
- **更弱的前提** —— 非线性与**学习型预测器**；不确定集从**有限样本**里估出来，而不是先验给定。
- **更高的维度** —— 现在的结果对 d_x 是显式的，但真正高维的系统还没被验证。

这三条都还没有答案 —— 谁在做其中一条，会后可以找我聊。

你打不破天花板。你也不该越过地板。
剩下的，是如何用更多信息把这个 gap 一点点缩小。

两篇论文

幻灯片和代码都可以分享，会后找我就行。

Probabilistic Predictability of Stochastic Dynamical Systems

Automatica, 2025 — Xu, Li, He

[10.1016/j.automatica.2025.112160](https://doi.org/10.1016/j.automatica.2025.112160)

给出天花板 —— 闭式 (10)，以及单条轨迹上的收敛（定理 4）。

Distributionally Robust Probabilistic Prediction for Stochastic Dynamical Systems

IEEE TAC, 2026 — Xu, He

[10.1109/TAC.2026.3687495](https://doi.org/10.1109/TAC.2026.3687495)

给出地板 —— Noise-DRPP（定理 3）、Eig-DRPP，以及它们的最优性间隙（定理 5–6）。



徐韬 · Zerken@sjtu.edu.cn · www.tao-xu.com