The Pitfalls of Imitation Learning (when the action space is continuous)

Max Simchowitz, Daniel Pfrommer, Ali Jadbabaie

Carnegie Mellon University

MIT

Pre-training in Large Language Models

A **large language model** (**LLM**) is a type of machine learning model (source: Wikipedia)

We treat natural human language as an **expert demonstrator** which we aim to imitate. Here, the “observation” is the string of tokens thus far , and the “action” is the predicted next token.

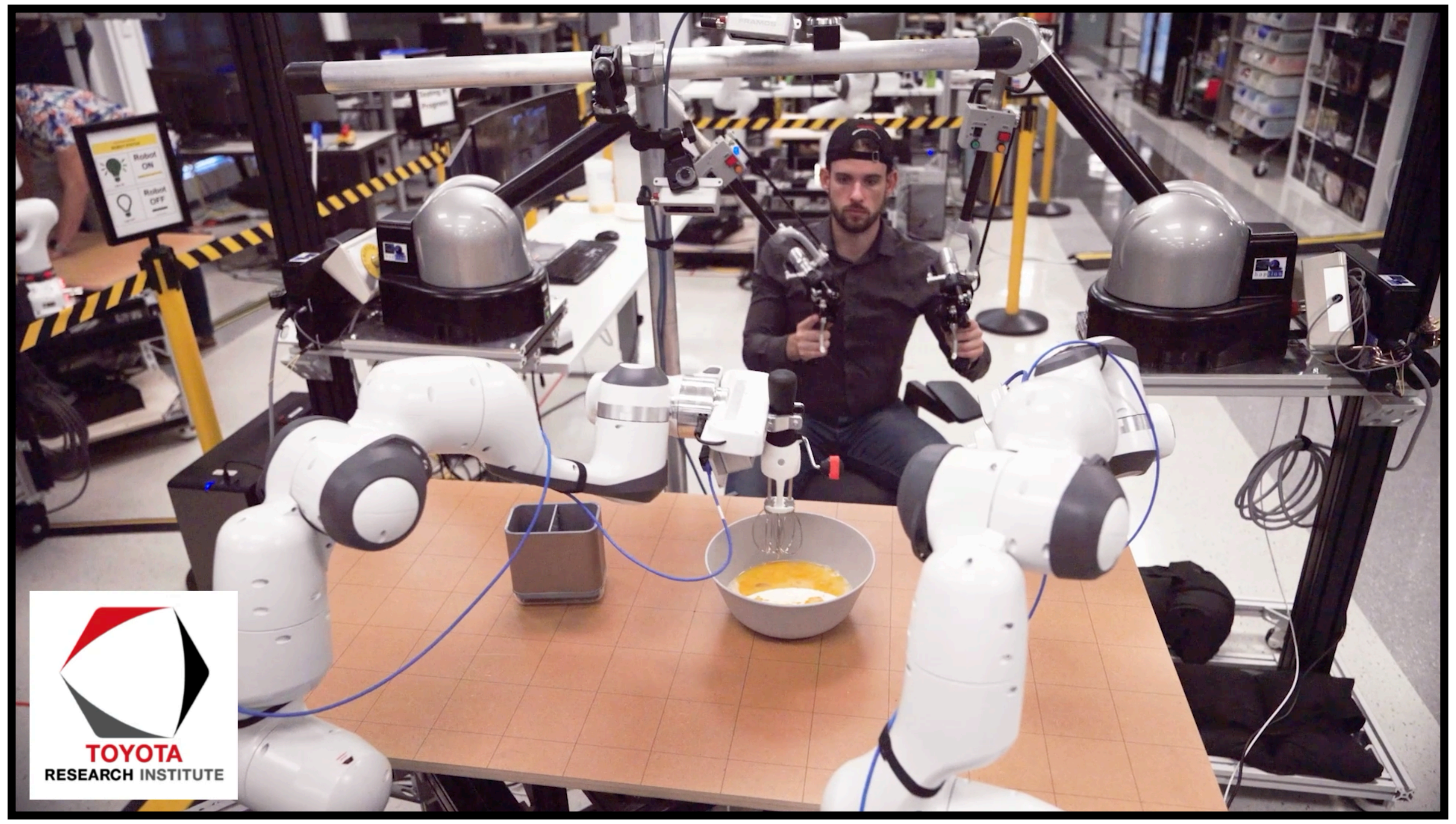
Pre-training in Large Robot Models

We treat use a human **expert demonstrator** which we aim to imitate. Our aim is to predict a "**next action**" (robot action) from **observation** (pixels, tactile sensing.)

Pre-training in Large Robot Models

- Will **scaling** solve robotic foundation models?
- Do we need **on-policy data** or can this be done entirely offline?
- How should we **design policies** that can scale?

Pre-training: Discrete v.s. Continuous?

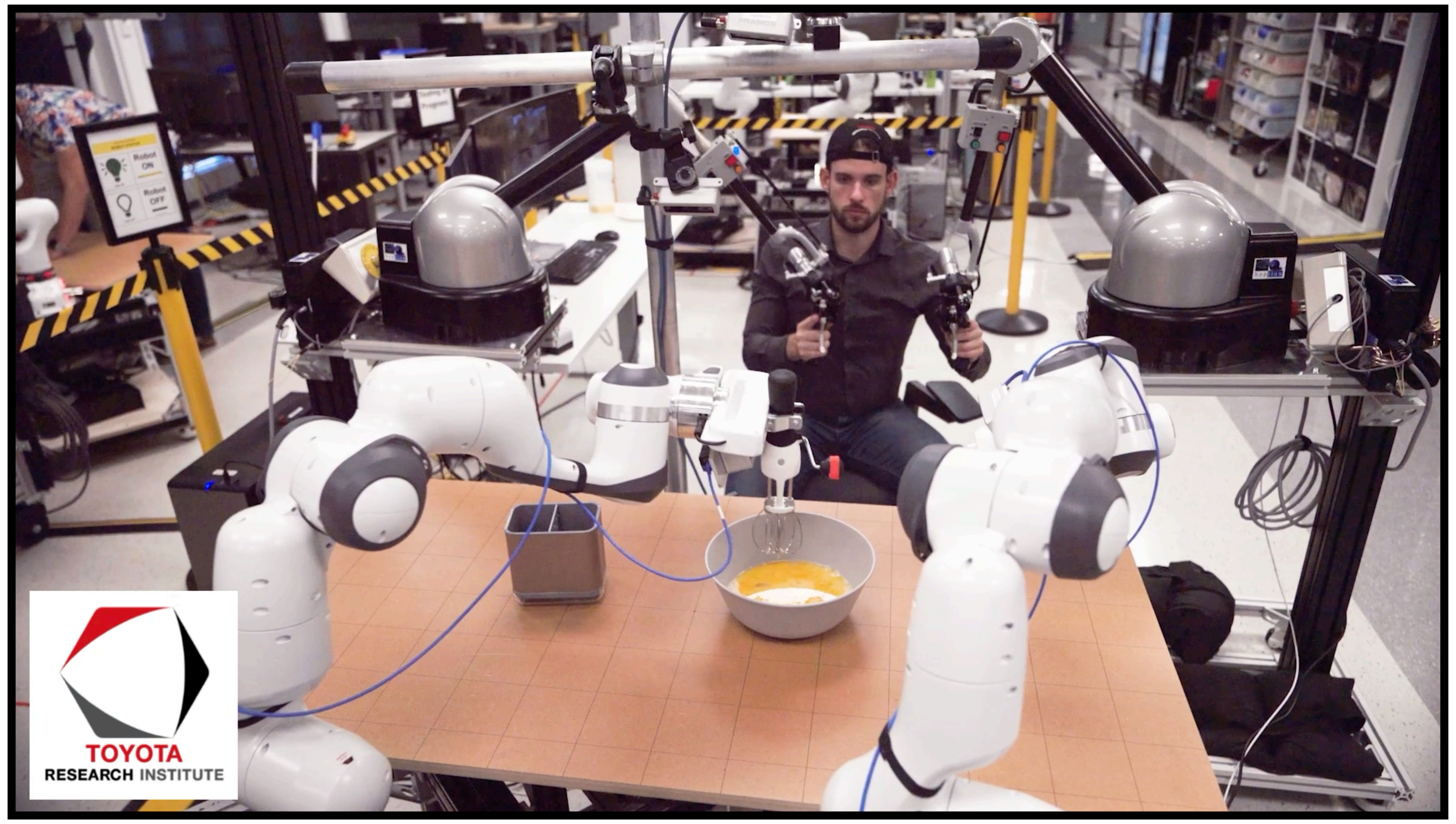


Language: **predict discrete tokens.**

Robotics: **predict continuous actions.**

Pre-training: Discrete v.s. Continuous?

GPT-4



Is there a **fundamental difference?**

Reinforcement Learning v.s. Continuous Control

GPT-4

Notation: states s, actions a

Dynamics: $s_{t+1} \sim P(s_t, a_t)$

Policy: $a_t \sim \pi(s_t)$

Semantics: $s_t = (w_1, \ldots, w_t)$, $a_t = w_{t+1}$

Notation: states x, actions u

Dynamics: $x_{t+1} = f(x_t, u_t) + (\text{noise})$

Policy: $u_t \sim \pi(x_t)$

Semantics: x, u **are continuous valued.**

Formalizing Imitation Learning

GPT-4

$$\textbf{Mininimize } \mathscr{R}_c(\hat{\pi}; \pi^\star) = \mathbb{E}_{\hat{\pi}}\Big[\sum_{h=1}^{H} c(x_t, u_t)\Big] - \mathbb{E}_{\pi^\star}\Big[\sum_{h=1}^{H} c(x_t, u_t)\Big]$$

"Horizon" H

error **cost under imitator** **cost under expert**

Example Algorithm: Behavior Cloning.

$$\textbf{Mininimize } \mathcal{R}_c(\hat{\pi}; \pi^\star) = \mathbb{E}_{\hat{\pi}}\left[\sum_{h=1}^{H} c(x_t, u_t)\right] - \mathbb{E}_{\pi^\star}\left[\sum_{h=1}^{H} c(x_t, u_t)\right]$$

error **cost under imitator** **cost under expert**

Algorithm: $\hat{\pi} \approx \arg\min_{\pi} \sum_{(x,u) \in \text{expert data}} \text{loss}(\pi, x, u)$

Goal: Train $\hat{\pi}$ to fit the expert data.

Example Algorithm: Behavior Cloning.

$$\textbf{Mininimize } \mathscr{R}_c(\hat{\pi}; \pi^\star) = \mathbb{E}_{\hat{\pi}}\left[\sum_{h=1}^{H} c(x_t, u_t)\right] - \mathbb{E}_{\pi^\star}\left[\sum_{h=1}^{H} c(x_t, u_t)\right]$$

error **cost under imitator** **cost under expert**

Algorithm: $\hat{\pi} \approx \arg\min_{\pi} \sum_{(x,u)\in \text{expert data}} \text{loss}(\pi, x, u)$

Example 1: $\text{loss}(\pi, x, u) = \|u - \pi(x)\|^2$ ($\pi^\star$ is deterministic)

Example 2: $\text{loss}(\pi, x, u) = \mathbf{1}_{\pi(x)=u}$ ($\pi^\star$ is discrete)

Example 3: $\text{loss}(\pi, x, u) = \log \pi(u \mid x)$ ($\pi^\star$ is discrete, or $\pi^\star(x)$ has density)

Example 4: $\text{loss}(\pi, x, u) =$ (**Score Matching**) (popular in robotics)

Example Algorithm: Behavior Cloning.

Mininimize $\mathcal{R}_c(\hat{\pi}; \pi^\star) = \mathbb{E}_{\hat{\pi}}[\sum_{h=1}^{H} c(x_t, u_t)] - \mathbb{E}_{\pi^\star}[\sum_{h=1}^{H} c(x_t, u_t)]$

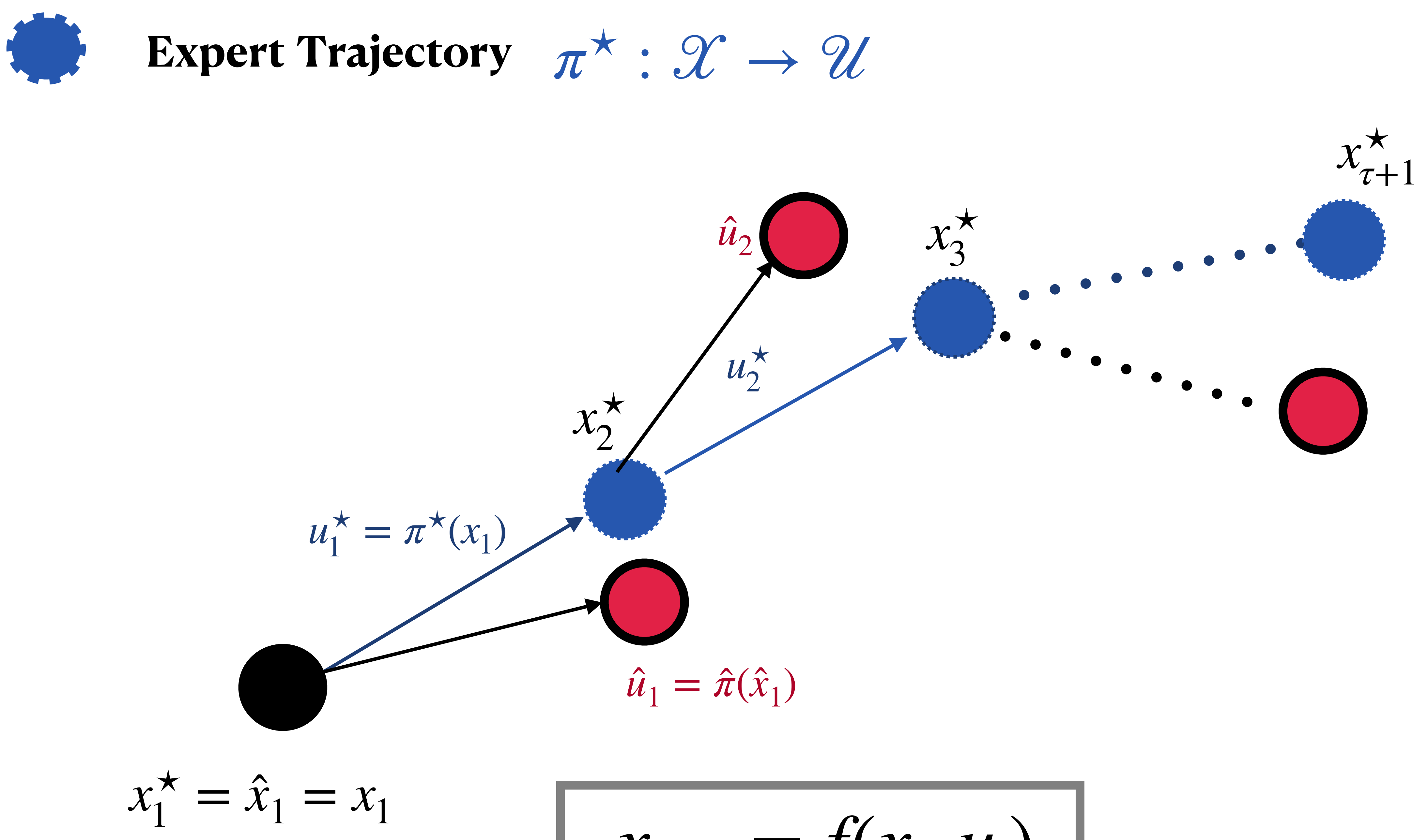
error **cost under imitator** **cost under expert**

Algorithm: $\hat{\pi} \approx \arg\min_{\pi} \sum_{(x,u) \in \text{expert data}} \text{loss}(\pi, x, u)$

Compare to $\mathcal{R}_{\text{expert}}(\hat{\pi}; \pi^\star) = \mathbb{E}_{\pi^\star}[\sum_{h=1}^{H} \text{loss}(\hat{\pi}, x_t, u_t)]$

trajectories **loss of imitator under expert distribution**

The Compounding Error Problem.



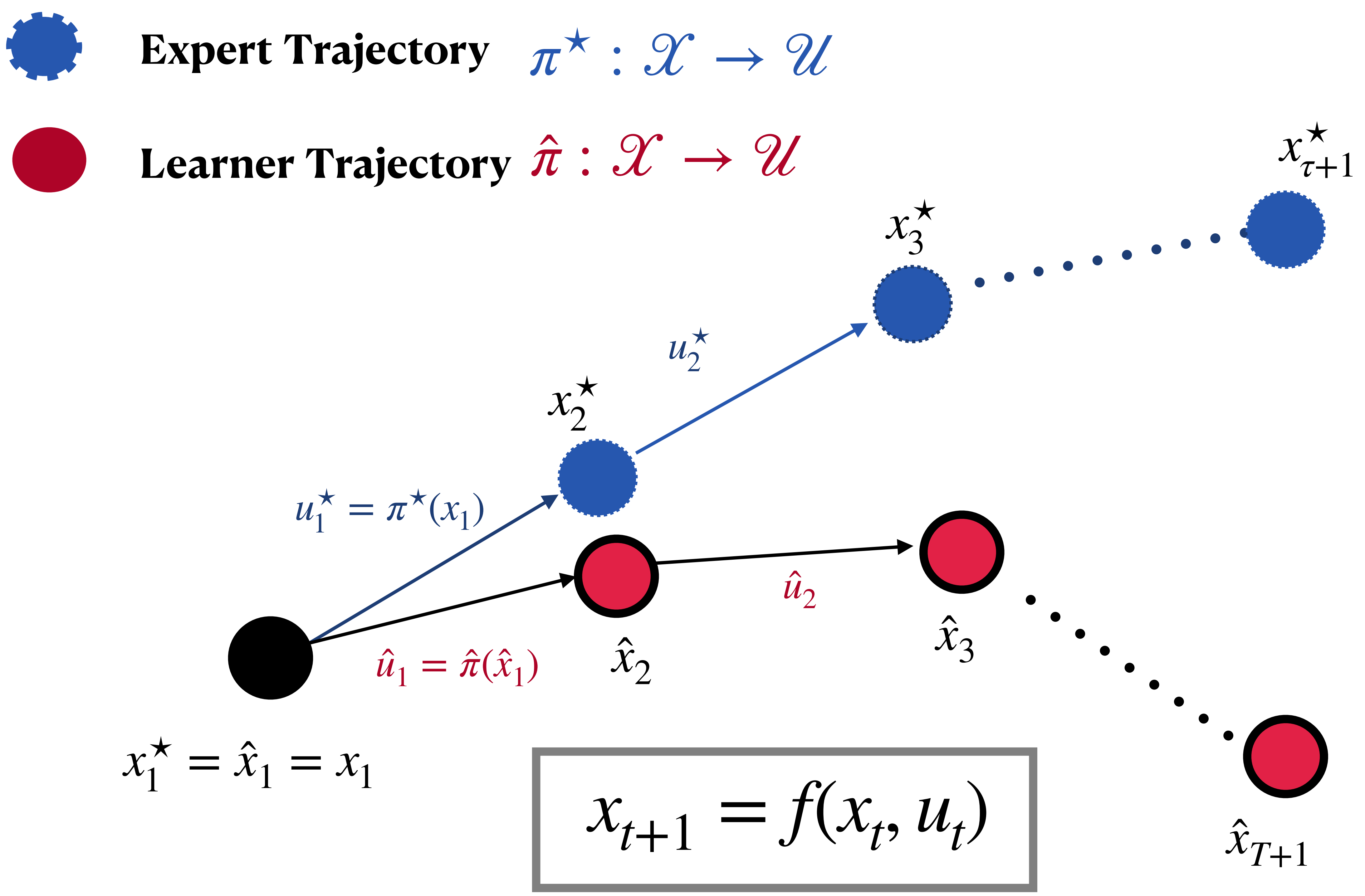
$$x_{t+1} = f(x_t, u_t)$$

$$\mathcal{R}_{\text{expert}}(\hat{\pi}; \pi^\star) = \mathbb{E}_{\pi^\star}\left[\sum_{h=1}^{H} \text{loss}(\hat{\pi}, x_t, u_t)\right]$$

The Compounding Error Problem.

Mininimize $\mathscr{R}_c(\hat{\pi}; \pi^\star) = \mathbb{E}_{\hat{\pi}}[\sum_{h=1}^{H} c(x_t, u_t)] - \mathbb{E}_{\pi^\star}[\sum_{h=1}^{H} c(x_t, u_t)]$

error **cost under imitator** **cost under expert**



Challenge A: Error accumulates over time steps, larger with larger **H**.

Challenge B: After error has accumulated, we are now **out of distribution.**

What is known?

Mininimize $\mathscr{R}_c(\hat{\pi}; \pi^\star) = \mathbb{E}_{\hat{\pi}}[\sum_{h=1}^{H} c(x_t, u_t)] - \mathbb{E}_{\pi^\star}[\sum_{h=1}^{H} c(x_t, u_t)]$

error **cost under imitator** **cost under expert**

Compare to $\mathscr{R}_{\text{expert}}(\hat{\pi}; \pi^\star) = \mathbb{E}_{\pi^\star}[\sum_{h=1}^{H} \text{loss}(\hat{\pi}, \pi^\star, u_t)]$

loss of imitator under expert distribution

What is known?

"Folklore Theorem" (DAGGER): Suppose that a function of $\mathrm{loss}(\pi, x, u) = \mathbf{1}_{\pi(x)=u}$ is the **zero-one loss**, and that $c(x, u)$ is bounded in [0,1]. Then,

$$\boxed{\mathscr{R}(\hat{\pi}; \pi^\star) \leq H \cdot \mathscr{R}_{\mathrm{expert}}(\hat{\pi}; \pi^\star)}$$

Beautiful Improvements due to Foster et al. '24 for the **Log Loss.**

"Compounding error is at most linear(ish) **in horizon"**

Limitations of Prior Work.

Warmup: Can we imitate in the zero-one loss?

Scalar Prediction Problem: $x \sim \mathrm{Uniform}([0,1])$, $u = \pi^\star(x)$

$$\boxed{\mathscr{R}_{\mathrm{expert},\{0,1\}}(\hat{\pi}, \pi^\star) = \mathbb{E}_{x \sim [0,1]}[\mathrm{I}\{\hat{\pi}(x) \neq \pi^\star(x)\}]}$$

Is this possible to do with non-vanishing error?

Warmup: Can we imitate in the zero-one loss?

Theorem: There exists a class of $\Pi = \{\pi\}$ such that, given n **examples** $(x, \pi^\star(x)), x \sim [0,1]$

A. Any learning algorithm suffers $\mathscr{R}_{\mathrm{expert},\{0,1\}}(\hat{\pi}, \pi^\star) = 1$ **with probability one**

B. Behavior cloning with $\mathrm{loss}(x, u, \pi) = (\pi(x) - u)^2$

$$\mathscr{R}_{\mathrm{expert},L_2}(\hat{\pi}, \pi^\star) = \mathbb{E}_{x\sim[0,1]}[\,|\,\hat{\pi}(x) - \pi^\star(x)\,|^2\,]^{1/2} = n^{-\omega(1)}$$

Proof Sketch: Consider $\pi(x) = \sum_{k\geq 1} \alpha_k 2^{-k}\cos(2\pi k x)$, $\alpha_k \in \{-1,1\}$. Getting small $\{0,1\}$ error requires perfect estimation of $\{a_k\}$ from finite data.

Warmup: Can we imitate in the zero-one loss?

Theorem: There exists a class of $\Pi = \{\pi\}$ such that, given n **examples** $(x, \pi^\star(x)), x \sim [0,1]$

A. Any learning algorithm suffers $\mathscr{R}_{\text{expert},\{0,1\}}(\hat{\pi}, \pi^\star) = 1$ **with probability one**

B. Behavior cloning with $\text{loss}(x, u, \pi) = (\pi(x) - u)^2$

$$\mathscr{R}_{\text{expert},L_2}(\hat{\pi}, \pi^\star) = \mathbb{E}_{x\sim[0,1]}[\,|\,\hat{\pi}(x) - \pi^\star(x)\,|^2\,]^{1/2} = n^{-\omega(1)}$$

Key Implication: The linear-in-horizon compounding error (DAGGER) is not applicable.

Results.

What is a “nice” imitation learning problem?

Property 1: Dynamics and expert are **deterministic** $x_{t+1} = f(x_t, u_t),\ \pi^\star(x_t)$ is **deterministic**.

Property 2: The dynamics and the expert are C^∞, and their first and second derivatives are bounded (i.e. **Lipschitz** and **smooth**).

(unimodal)

Property 3: The dynamics are “exponentially incrementally input-to-state stable” (**E-IISS**)

(okay … what does this mean?)

What is a "nice" imitation learning problem?

Property 1: Dynamics and expert are **deterministic** $x_{t+1} = f(x_t, u_t),\ \pi^\star(x_t)$ is **deterministic**.

Property 2: The dynamics and the expert are C^∞, and their first and second derivatives are bounded (i.e. **Lipschitz** and **smooth**).

(unimodal)

Our lower bounds hold for "**simple**" imitator policies:

$$\hat{\pi}(x) = \text{mean}(\hat{\pi}(x)) + z$$

Lipschitz/smooth

independent of x

see later for **non-simple**

An Informal Statement

Theorem: Pick your favorite $k \in \mathbb{N}$. Then there exists a family of "**nice**" imitation learning problems of **problem dimension** 3 such that, given **n** example trajectories, there exists an algorithm for which

$$\boxed{\mathscr{R}_{\text{expert},L_1}(\hat{\pi}; \pi^\star) = \mathbb{E}_{\pi^\star}\left[\sum_{t=1}^{H} \|\pi_t^\star(x_t) - \hat{\pi}(x_t)\|\right] \leq n^{-k}}$$

Unlike $\{0,1\}$ loss, this can be **minimized.**

An Informal Statement

Theorem: Pick your favorite $k \in \mathbb{N}$. Then there exists a family of "**nice**" imitation learning problems of **problem dimension** 3 such that, given **n** example trajectories, there exists an algorithm for which $\mathscr{R}_{\text{expert},L_1}(\hat{\pi}; \pi^\star) \leq n^{-k}$

However, there exists a **1-Lipschitz, bounded** $c(\cdot,\cdot) \in [0,1]$ such that any learning algorithm returns **"simple"** policies $\hat{\pi}$ suffers

$$\boxed{\mathscr{R}_c(\hat{\pi}; \pi^\star) \geq \text{const} \cdot \min\left\{1, 2^H \cdot n^{-k}\right\}}$$

excess cost under imitator relative to expert

An Informal Statement

Theorem: There exists a family of "**nice**" imitation learning problems of problem dimension 3 such that, given **n** example trajectories

$$\mathcal{R}_{\text{expert},L_1}(\hat{\pi};\pi^\star) \leq n^{-k}$$

$$\mathcal{R}_c(\hat{\pi};\pi^\star) \geq \text{const} \cdot \min\left\{1, 2^H \cdot n^{-k}\right\}$$

Remark 1: Deployment error can be exponentially larger than expert-distribution error.

Remark 2: We will see: result depends on **imitator policy,** not learning algorithm . Applies to **behavior cloning, offline RL, inverse RL** (all without on-policy data).

Remark 3: We will see how to break our lower bound with "improper" policies.

What is a nice control system?

What is a "nice" imitation learning problem?

Property 1: Dynamics and expert are **deterministic** $x_{t+1} = f(x_t, u_t),\ \pi^\star(x_t)$ is **deterministic**.

Property 2: The dynamics and the expert are C^∞, and their first and second derivatives are bounded (i.e. **Lipschitz** and **smooth**).

(unimodal)

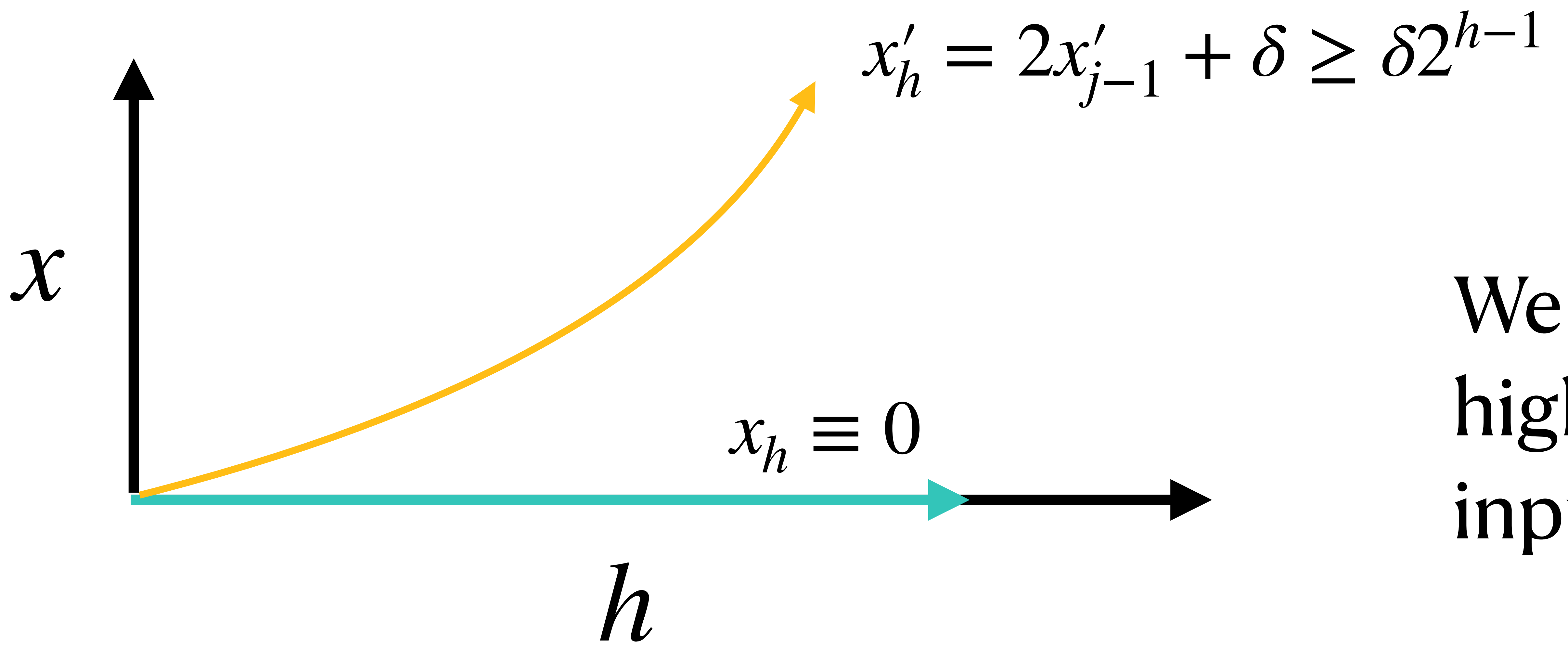
Property 3: The dynamics are "exponentially incrementally input-to-state stable" (**E-IISS**)

(okay … what does this mean?)

Instability in control systems

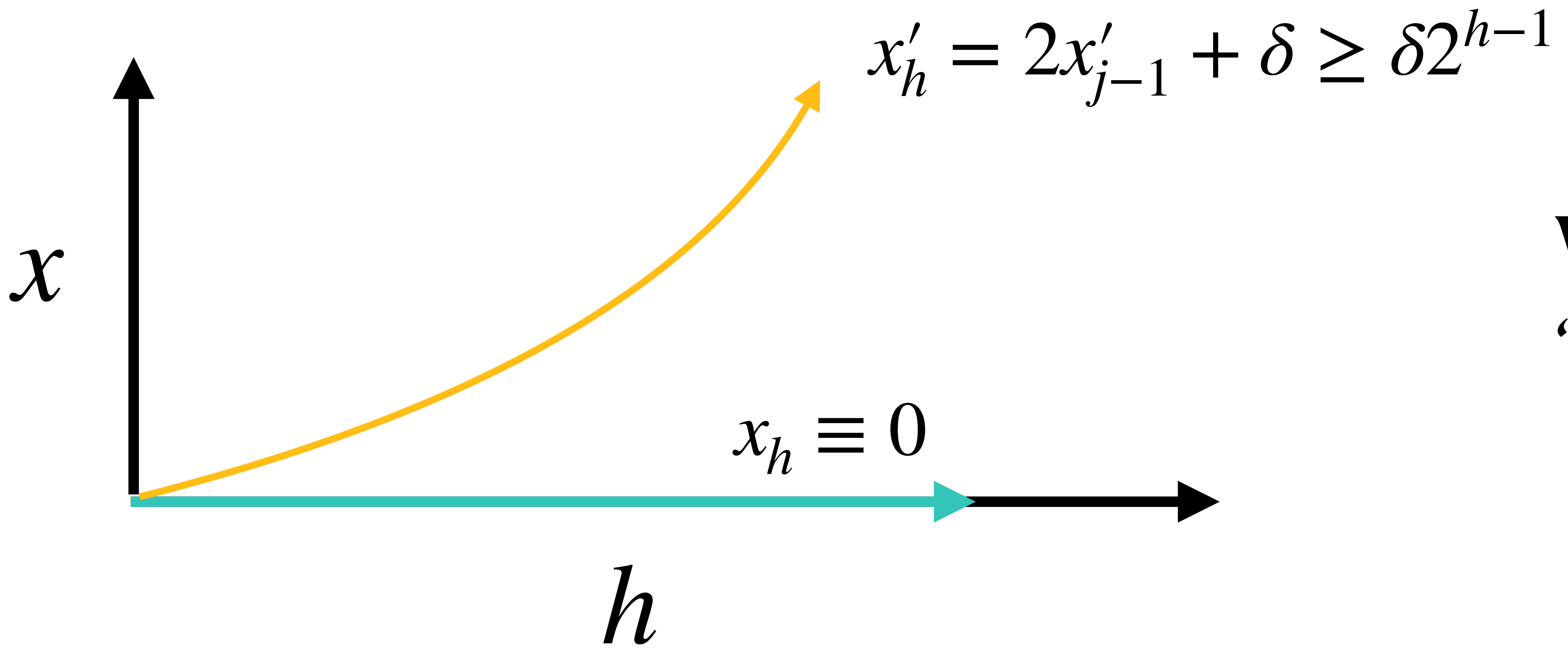
Consider the scalar, linear control system $f(x, u) = 2x + u$

Consider two trajectories: $(x_1, u_1, \ldots), u_i \equiv 0$ **and** $(x'_1, u'_1, \ldots),\ u_i \equiv \delta, x_1 = x'_1 = 0$



We call systems with such high sensitivity to their inputs "**unstable**"

Instability in control systems



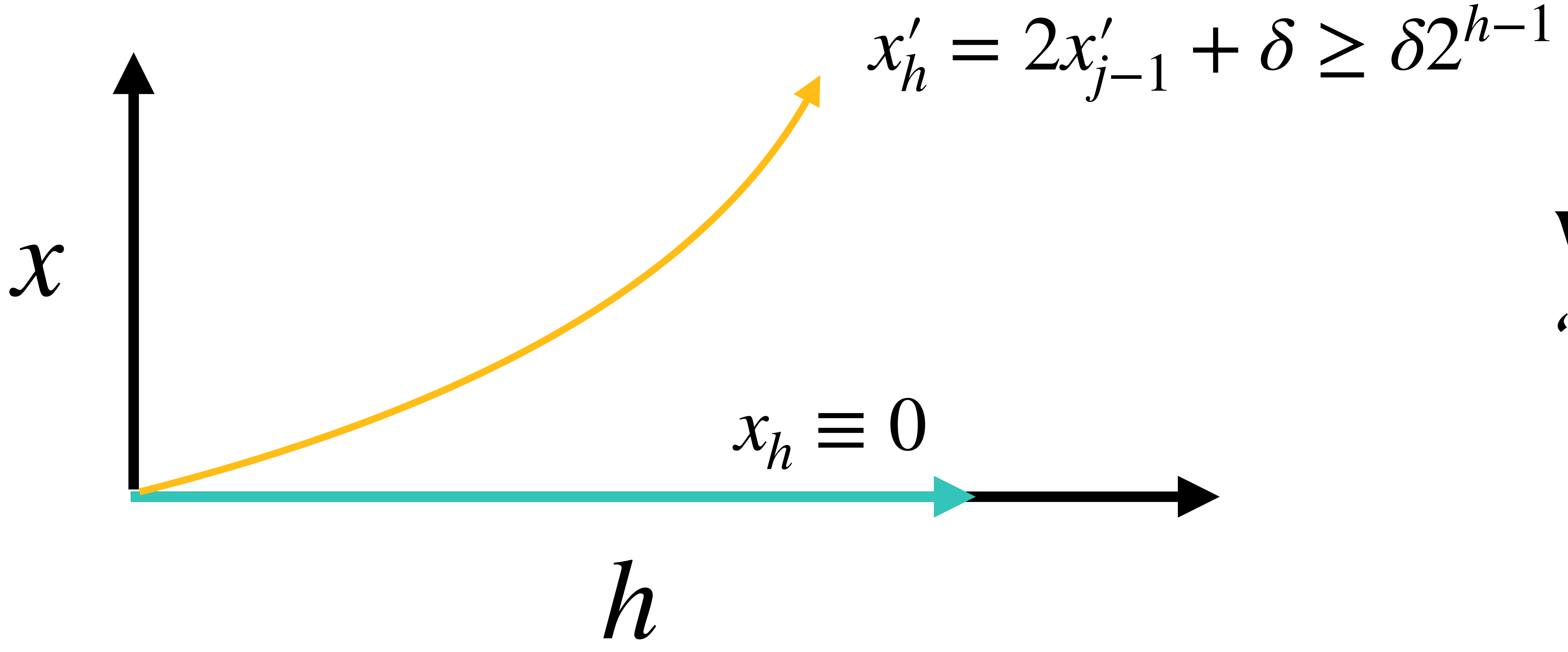
We call such a system "**unstable**"

Theorem (Informal): There exist imitation learning problems which satisfy **Property 1** (Determinism) and **Property 2** (Smoothness) but are **unstable** (violate property 3) for which **all learning algorithms** (no restriction) suffer, for $H \leq e^{\text{dimension}}$,

$$\mathscr{R}_{\text{expert},L_1}(\hat{\pi}; \pi^\star) \leq n^{-k}$$

$$\mathscr{R}(\hat{\pi}; \pi^\star) \geq \text{const} \cdot \min\left\{1, 2^H \cdot n^{-k}\right\}$$

Instability in control systems



We call such a system "**unstable**"

Unstable systems are real in aeronautics! Not so much in robotic manipulation...

So what about "nice" systems?

Exponential Stability (E-IISS)

Definition (Angelis '08, Pfrommer '23)**:** We say that a control system f is **E**xponentially **I**ncremental **I**nput-to-**S**tate **S**table **(E-IISS)** if for any initial states x_1, x_1' and any sequences $u_1, \ldots, u_H$ and $u_1', \ldots, u_H'$ of control inputs, the resulting trajectories satisfy

$$\boxed{\|x_{h+1} - x_{h+1}'\| \leq C\rho^h \|x_1 - x_1'\| + C \sum_{j=1}^{h} \rho^{h-j} \|u_j - u_j'\|}$$

$C > 0,\ \rho \in (0,1)$

exponential forgetting of past states & inputs

Example: $x_1 = x_1' = 0$, and $u_h \equiv 0,\ u_h' \equiv \delta$. **Then,** $\|x_{h+1} - x_{h+1}'\| \leq \frac{C}{1-\rho} \cdot \delta = O(\delta)$

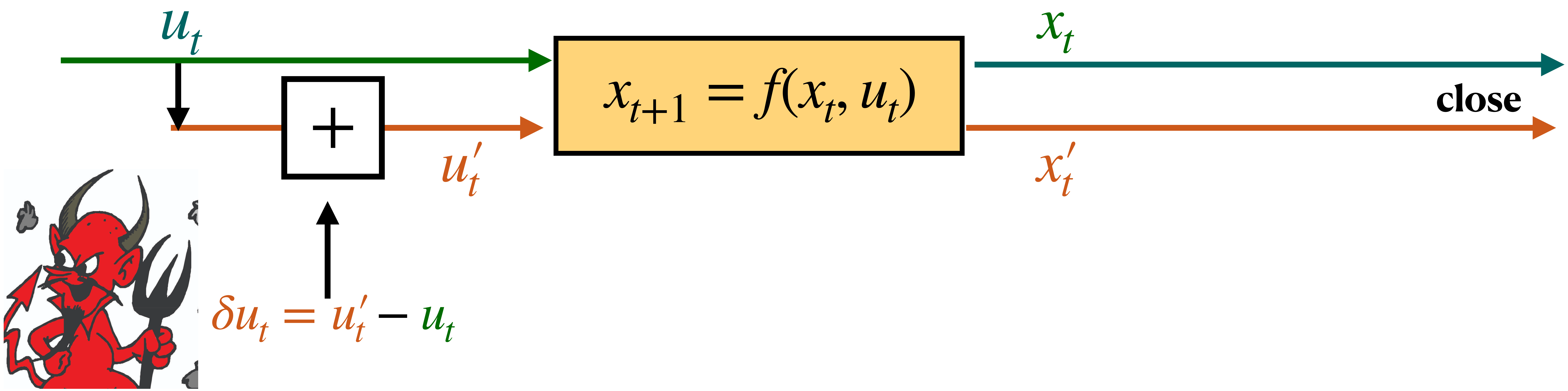
perturbations of inputs lead to bounded perturbations of states!

Open Loop Stable

Property 3: The dynamics $(x, u) \mapsto f(x, u)$ are **E-IISS**

$$\|x_{h+1} - x'_{h+1}\| \leq C\rho^h \|x_1 - x'_1\| + C\sum_{j=1}^{h} \rho^{h-j} \|u_j - u'_j\|$$

perturbations of inputs lead to bounded perturbations of states!

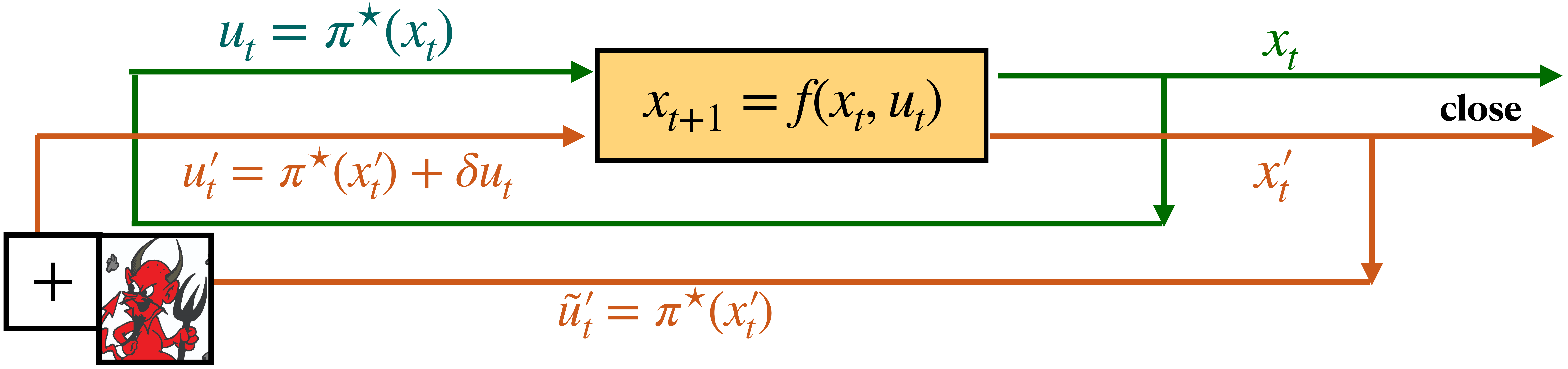


Closed Loop Stable

Property 3: The dynamics $(x, u) \mapsto f(x, u)$ and $(x, \delta u) \mapsto f(x, \pi^\star(x) + \delta u)$ are **E-IISS**

$$\|x_{h+1} - x'_{h+1}\| \le C\rho^h \|x_1 - x'_1\| + C \sum_{j=1}^{h} \rho^{h-j} \|u_j - u'_j\|$$

perturbations of inputs lead to bounded perturbations of states!



What is a "nice" imitation learning problem?

Property 1: Dynamics and expert are **deterministic** $x_{t+1} = f(x_t, u_t)$, $\pi^\star(x_t)$ is **deterministic**.

Property 2: The dynamics and the expert are C^∞, and their first and second derivatives are bounded (i.e. **Lipschitz** and **smooth**).

Property 3: The dynamics $(x, u) \mapsto f(x, u)$ and $(x, \delta u) \mapsto f(x, \pi^\star(x) + \delta u)$ are **E-IISS**

$$\|x_{h+1} - x'_{h+1}\| \leq C\rho^h \|x_1 - x'_1\| + C\sum_{j=1}^{h} \rho^{h-j} \|u_j - u'_j\|$$

perturbations of inputs lead to bounded perturbations of states!

"open and closed-loop" stability

The Theorem Statement

Property 3: The dynamics $(x, u) \mapsto f(x, u)$ and $(x, \delta u) \mapsto f(x, \pi^\star(x) + \delta u)$ are **E-IISS**

$$\|x_{h+1} - x'_{h+1}\| \leq C\rho^h \|x_1 - x'_1\| + C \sum_{j=1}^{h} \rho^{h-j} \|u_j - u'_j\|$$

perturbations of inputs lead to bounded perturbations of states!

$$\mathscr{R}_{\mathrm{expert},L_2}(\hat{\pi}; \pi^\star) \leq n^{-k}$$

$$\mathscr{R}(\hat{\pi}; \pi^\star) \geq \mathrm{const} \cdot \min\left\{1, 2^H \cdot n^{-k}\right\}$$

Wait…wait… how can this be?

Property 3: The dynamics $(x, u) \mapsto f(x, u)$ and $(x, \delta u) \mapsto f(x, \pi^\star(x) + \delta u)$ are **E-IISS**

$$\|x_{h+1} - x'_{h+1}\| \le C\rho^h \|x_1 - x'_1\| + C\sum_{j=1}^{h} \rho^{h-j} \|u_j - u'_j\|$$

perturbations of inputs lead to bounded perturbations of states!

$$\mathscr{R}_{\mathrm{expert},L_2}(\hat{\pi}; \pi^\star) \le n^{-k}$$

This says that the imitator is learning up to "small perturbations"

$$\mathscr{R}(\hat{\pi}; \pi^\star) \ge \mathrm{const} \cdot \min\left\{1, 2^H \cdot n^{-k}\right\}$$

Yet still, the error under deployment grows!

Proof via Linear Control.

Roadmap

1. Introduce linear control systems
2. Explain incremental instability for linear control systems
3. Explain the **tension** between **imitation and stability** in linear systems
4. Gesture to the general result.

Linear Dynamical Systems

Definition: A linear dynamical system is a dynamical map where $f(x, u)$ is **linear.**

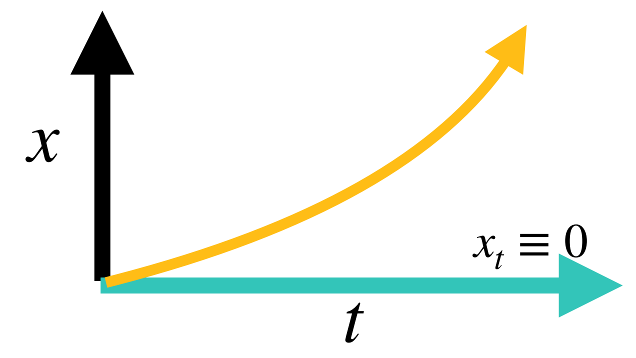
$$x_{t+1} = Ax_t + Bu_t$$

Lemma: Let $B = I$ be the identity. Then, a linear system is **E-ISSS** if and only if

$\rho(A) := \max\{\,|\mathrm{Re}(\lambda)| : \lambda \in \mathrm{spec}(A)\}$ is strictly less than **one.**

Proof Sketch: If you unroll the dynamics, you get powers of A^k. These decay exponentially if $\rho(A) < 1$, but **grow exponentially** if $\rho(A) > 1$

(exponentially large perturbation sensitivity)



$x_{t+1} = Ax_t + Bu_t$

Linear Feedback Controllers

Definition: A linear state feedback policy is linear memoryless policy $\pi(x) = Kx$.

Lemma: Consider closed-loop system $f^\pi(x, u) = f(x, \pi(x) + u)$ with linear dynamics and linear feedback policy. Then

1. $f^\pi(x, \delta u) := f(x, \pi(x) + \delta u) = (A + BK)x + B\delta u$
2. If $B = I$ is the identity, then f^π is E-ISSS if and only if $\rho(A + K) < 1$
3. If $B = I$ is the identity and $\rho(A + K) > 1$, **exponential perturbation sensitivity.**

$x_{t+1} = Ax_t + Bu_t$

Linear Feedback Controllers

Corollary: Let $A, K^\star, \hat{K}$ have $\rho(A) < 1$ and $\rho(A + K^\star) < 1$, but $\rho(A + \hat{K}) > 1$.

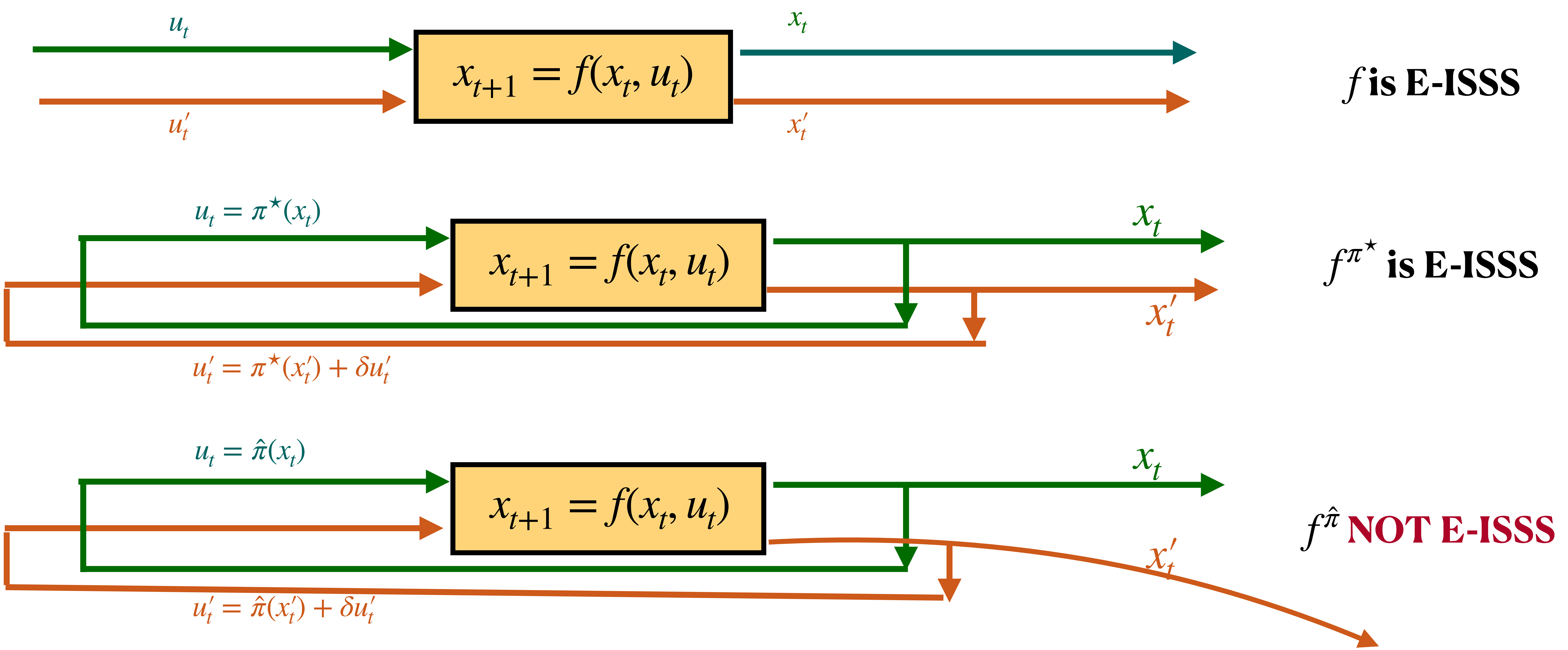
1. Open loop dynamics $f(x, u) = Ax + u$ is E-ISSS
2. Closed-loop dynamics $f^{\pi^\star}(x, u) = f(x, \pi^\star(x) + u)$ for $\pi^\star(x) = K^\star x$ is E-ISSS
3. Closed-loop dynamics $f^{\hat{\pi}}(x, u) = f(x, \hat{\pi}(x) + u)$ for $\hat{\pi}(x) = \hat{K}x$ can have **exponentially large perturbation sensitivity.**

Intuition: For the construction above, $f, f^{\pi^\star}$ are "**nice,**" but $\hat{\pi}$ is likely to have exponentially large compounding error.

$$x_{t+1} = Ax_t + Bu_t$$

Comparison of Stability

Corollary: Let $A, K^\star, \hat{K}$ have $\rho(A) < 1$ and $\rho(A + K^\star) < 1$, but $\rho(A + \hat{K}) > 1$.



$x_{t+1} = Ax_t + Bu_t$

The Challenging Pair

Key Lemma: There exists a pair of 2x2 matrix $(A_1, K_1^\star)$ and $(A_2, K_2^\star)$ with the following properties:

1. $\rho(A_i)$ and $\rho(A_i + K_i^\star)$ are both **strictly less than one** (E-ISS).

2. For any matrix $\hat{K}$ which can be "learned from imitation data," $\max_i \rho(A_i + \hat{K}) > 1$

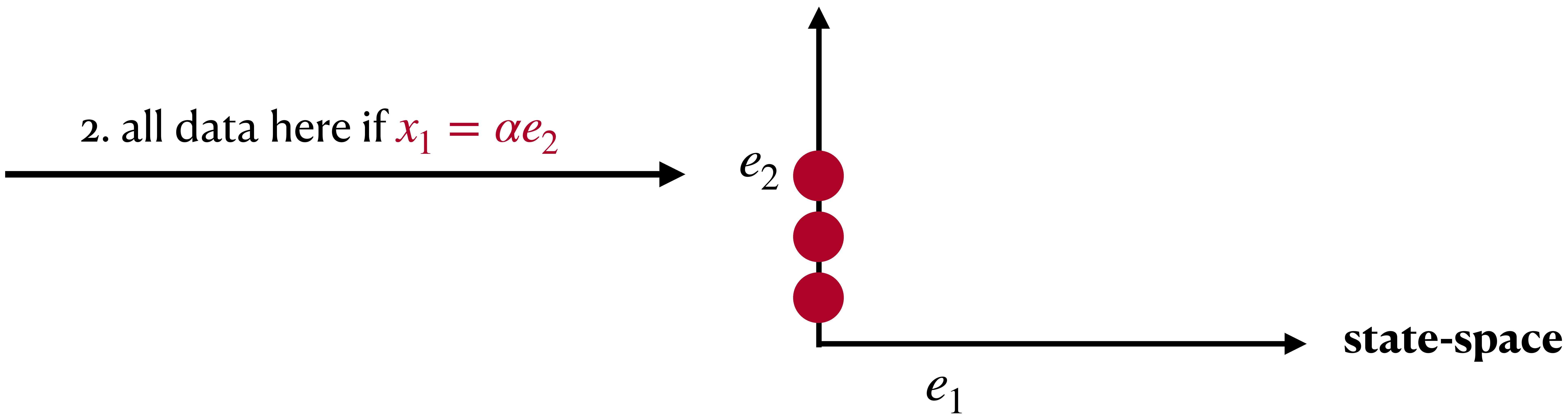
Intuition: $(A_i, K_i^\star)$ describe the **unknown** dynamics and expert, $\hat{K}$ is a linear imitator

Takeaway: Both systems + experts are closed loop stable, but **not the imitation policy**!

$x_{t+1} = Ax_t + Bu_t$

The Challenging Pair

Lemma: There exists a pair of 2x2 matrix $(A_1, K_1^\star)$ and $(A_2, K_2^\star)$ with the following properties:



1. $\rho(A_i)$ and $\rho(A_i + K_i^\star)$ are both **strictly less than one** (E-ISS).
2. The span of the vector $e_2 = (0,1)$ is an **invariant subspace** of $A_i + K_i^\star$

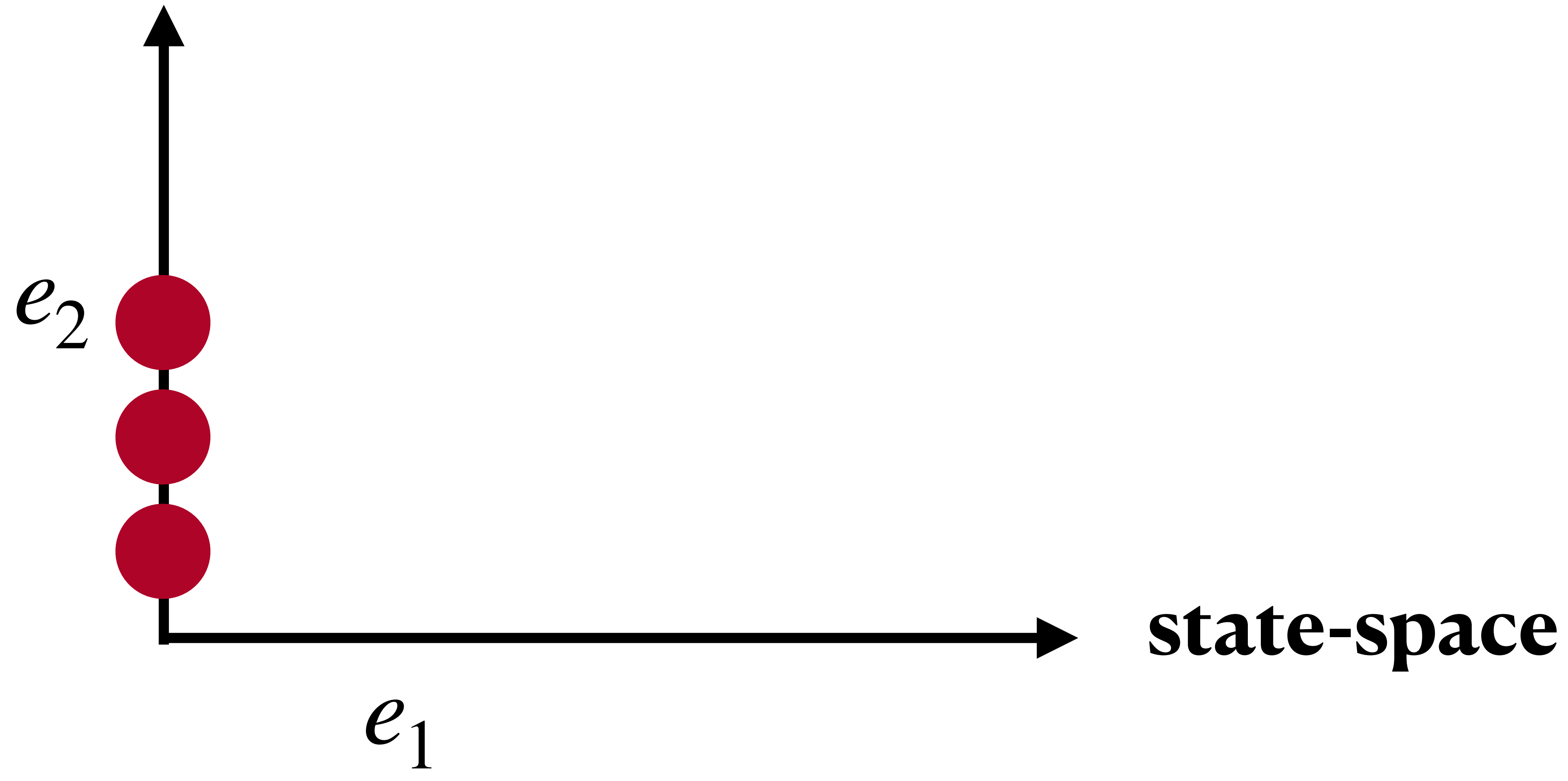
$x_{t+1} = Ax_t + Bu_t$

The Challenging Pair

Lemma: There exists a pair of 2x2 matrix $(A_1, K_1^\star)$ and $(A_2, K_2^\star)$ with the following properties:

2. all data here if $x_1 = \alpha e_2$

3. cannot identify index i

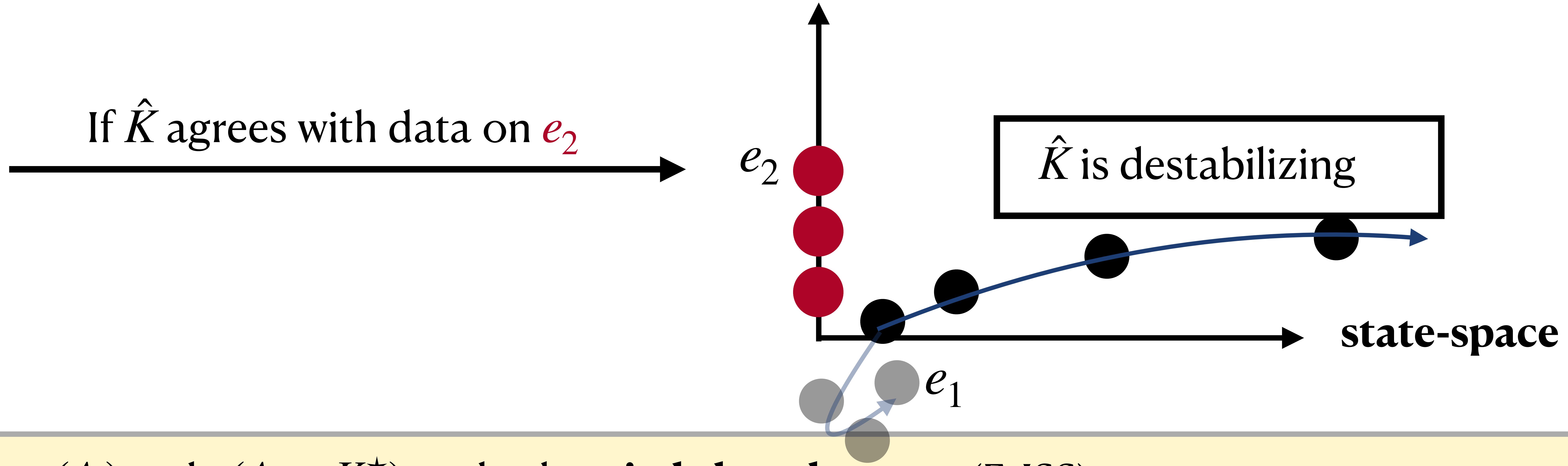


1. $\rho(A_i)$ and $\rho(A_i + K_i^\star)$ are both **strictly less than one** (E-ISS).
2. The span of the vector $e_2 = (0,1)$ is an **invariant subspace** of $A_i + K_i^\star$
3. $A_1 e_2 = A_2 e_2$ and $K_1^\star e_2 = K_2^\star e_2$

$x_{t+1} = Ax_t + Bu_t$

The Challenging Pair

Lemma: There exists a pair of 2x2 matrix $(A_1, K_1^\star)$ and $(A_2, K_2^\star)$ with the following properties:



1. $\rho(A_i)$ and $\rho(A_i + K_i^\star)$ are both **strictly less than one** (E-ISS).

2/3. Data from $e_2 = (0,1)$ **cannot distinguish** systems.

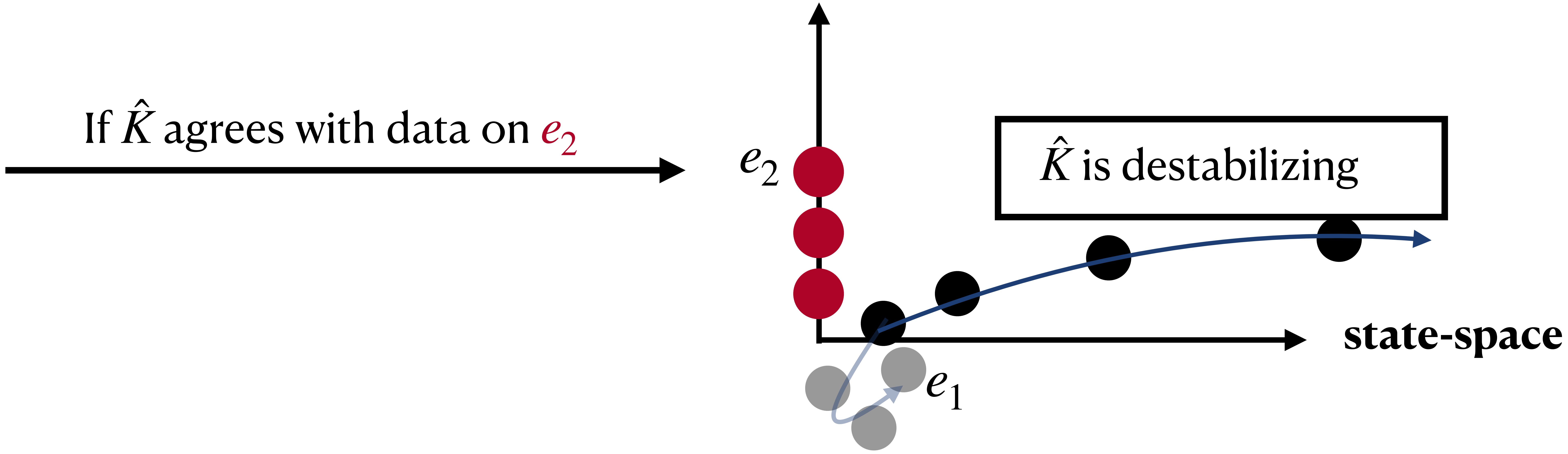
4. If $\hat{K}\, e_2 = K_i^\star\, e_2$, then $\hat{K}$ destabilizes one system: $\boxed{\max_i \rho(A_i + \hat{K}) > 1}$

$x_{t+1} = Ax_t + Bu_t$

The Challenging Pair

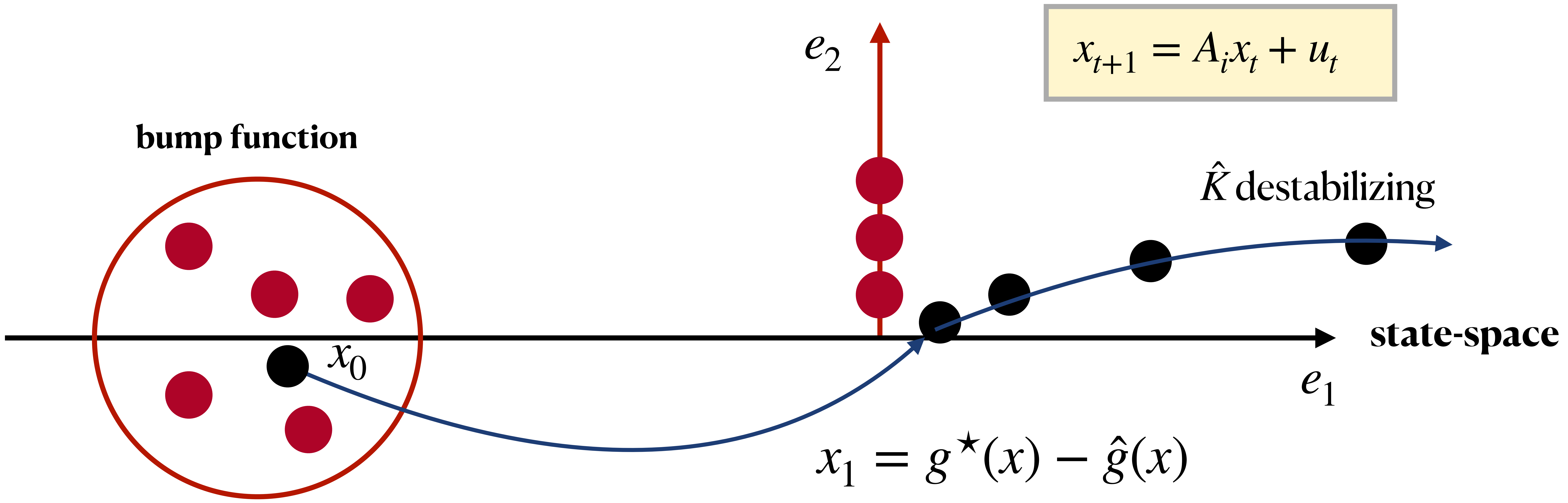
Corollary: Exists a pair of 2x2 matrix $(A_1, K_1^\star)$ and $(A_2, K_2^\star)$ such any linear policy $\hat{\pi}(x) = \hat{K}x$ either (a) disagrees with training data or (b) has exponentially sensitivity to e_1-perturbations for **one of** A_1, A_2.

If $\hat{K}$ agrees with data on e_2



Unfortunately, linear systems are too "**all-or-nothing**" for a lower bound.

Nonlinear Construction



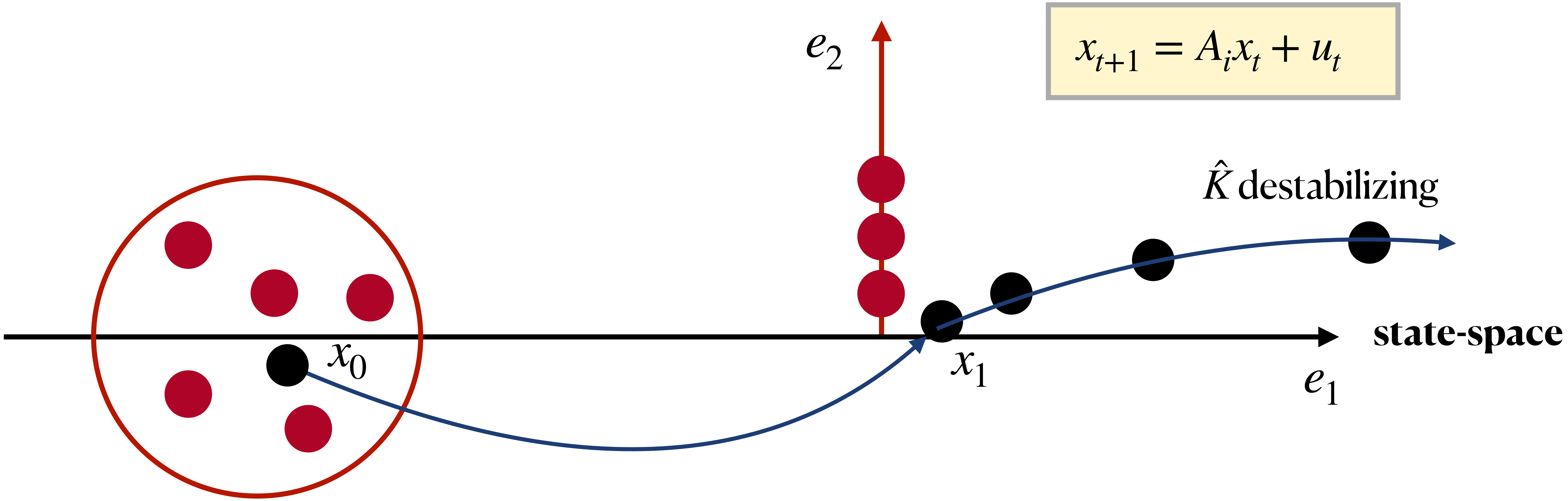
$\pi^\star(x) = g^\star(x)$

$f(x) = g^\star(x) - u$

$g^\star(x)$ can be learned up to a non-parametric rate

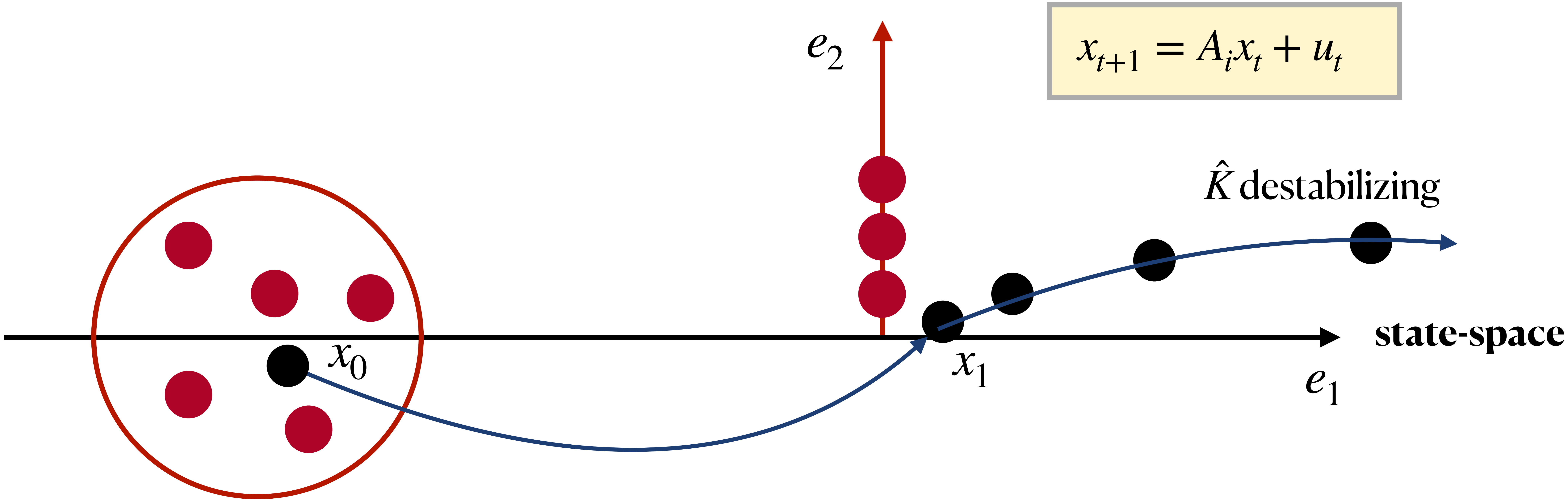
Key Idea: Embed the linear problem into a "nonlinear" problem that forces the learner in the e_1 direction, but only provides expert data in the e_2 direction.

Nonlinear Construction



Key Technical Tool: Because **simple policies** have smooth means, we can analyze them as "local linear controllers" by Taylor approximation.

Nonlinear Construction



Core Insight: For smooth '**simple**' **policies,** tension between **fidelity to expert data** (imitation) and **stabilization of unseen dynamical modes.**

Connecting Stability + Dynamic Programming

The Q function in Deterministic Control

Definition: for dynamics f, policy π, and cost c, the Q **function** is

$$Q_t^{f,\pi,c}(x, u) := \sum_{t'=t}^{H} c(x_{t'}, u_{t'}) \qquad \text{s.t. dynamics obey } (f, \pi), \quad x_{t'} = x, u_{t'} = u$$

"cost-to-go"

The Q function in Deterministic Control

Definition: for dynamics f, policy π, and cost c, the Q function is $Q_t^{f,\pi,c}(x, u)$.

Theorem (Performance Difference):

$$\mathcal{R}_c(\hat{\pi}; \pi^\star) := \mathbb{E}_{\hat{\pi}}[\textstyle\sum_{h=1}^{H} c(x_t, u_t)] - \mathbb{E}_{\pi^\star}[\textstyle\sum_{h=1}^{H} c(x_t, u_t)]$$

$$= \mathbb{E}_{\pi^\star}[\textstyle\sum_{h=1}^{H} Q_t^{f,\hat{\pi},c}(x_t, \hat{\pi}(x_t)) - Q_t^{f,\hat{\pi},c}(x_t, \pi^\star(x_t))$$

expectation under expert distribution **Q function of the learner**

The Q function in Deterministic Control

Definition: for dynamics f, policy π, and cost c, the Q function is $Q_t^{f,\pi,c}(x, u)$.

Theorem (Performance Difference):

$$\mathscr{R}_c(\hat{\pi}; \pi^\star) := \mathbb{E}_{\hat{\pi}}[\sum_{h=1}^{H} c(x_t, u_t)] - \mathbb{E}_{\pi^\star}[\sum_{h=1}^{H} c(x_t, u_t)]$$
$$= \mathbb{E}_{\pi^\star}[\sum_{h=1}^{H} Q_t^{f,\hat{\pi},c}(x_t, \hat{\pi}(x_t)) - Q_t^{f,\hat{\pi},c}(x_t, \pi^\star(x_t))$$

policy of the learner

policy of expert

The Q function in Deterministic Control

Theorem (Performance Difference):

$$\mathcal{R}_c(\hat{\pi}; \pi^\star) = \mathbb{E}_{\pi^\star}[\sum_{h=1}^{H} Q_t^{f,\hat{\pi},c}(x_t, \hat{\pi}(x_t)) - Q_t^{f,\hat{\pi},c}(x_t, \pi^\star(x_t))$$

Corollary: If $Q^{f,\hat{\pi},c}$ is **Lipschitz** in u: $|Q^{f,\hat{\pi},c}(x,u) - Q^{f,\hat{\pi},c}(x,u')| \le L\|u - u'\|$, then

$$\mathcal{R}_c(\hat{\pi}; \pi^\star) \le L \cdot \mathbb{E}_{\pi^\star}[\sum_{h=1}^{H} \|\pi^\star(x_t) - \hat{\pi}(x_t)\|]$$

The Q function in Deterministic Control

Corollary: If $Q^{f,\hat{\pi},c}$ is **Lipschitz** in u: $|Q^{f,\hat{\pi},c}(x,u) - Q^{f,\hat{\pi},c}(x,u')| \le L\|u-u'\|$, then

$$\mathscr{R}_c(\hat{\pi};\pi^\star) \le LH \cdot \mathbb{E}_{\pi^\star}[\textstyle\sum_{h=1}^{H} \|\pi^\star(x_t) - \hat{\pi}(x_t)\| = \boxed{L \cdot \mathscr{R}_{\text{expert},L_1}(\hat{\pi};\pi^\star)}$$

Lipschitz $Q^{f,\hat{\pi},c}$ ensures linear-in-L compounding error!

(see also Swamy et al. '21)

The Q function in Deterministic Control

Corollary: If $Q^{f,\hat{\pi},c}$ is **Lipschitz** in u: $|Q^{f,\hat{\pi},c}(x,u) - Q^{f,\hat{\pi},c}(x,u')| \le L\|u-u'\|$, then

$$\mathcal{R}_c(\hat{\pi};\pi^\star) \le LH \cdot \mathcal{R}_{\mathrm{expert},L_1}(\hat{\pi};\pi^\star)$$

1. Low compounding error is guaranteed by insensitive $Q^{f,\hat{\pi},c}$
2. Large compounding error requires highly sensitive $Q^{f,\hat{\pi},c}$
3. **Our Result** (Re-Interpretation): Even if $(f,\pi^\star)$ are open/closed-loop stable, it is **hard** to both **imitate** $\pi^\star$ and ensure $Q^{f,\hat{\pi},c}$ is insensitive to perturbation

The Q function in Deterministic Control

Corollary: If $Q^{f,\hat{\pi},c}$ is **Lipschitz** in u: $|Q^{f,\hat{\pi},c}(x,u) - Q^{f,\hat{\pi},c}(x,u')| \le L\|u - u'\|$, then

$$\mathscr{R}_c(\hat{\pi};\pi^\star) \le LH \cdot \mathscr{R}_{\mathrm{expert},L_1}(\hat{\pi};\pi^\star)$$

Takeaway for RL: Assumptions on the class of Q functions might not be fundamental! Instead, we need to operate from first principles from the dynamics and (as we will see...) policy classes!

The Q function in Deterministic Control

Corollary: If $Q^{f,\hat{\pi},c}$ is **Lipschitz** in u: $|Q^{f,\hat{\pi},c}(x,u) - Q^{f,\hat{\pi},c}(x,u')| \leq L\|u - u'\|$, then

$$\boxed{\mathcal{R}_c(\hat{\pi};\pi^\star) \leq LH \cdot \mathcal{R}_{\text{expert},L_1}(\hat{\pi};\pi^\star)}$$

Theorem (Pfrommer, **S**,J '25): If $\mathscr{C} = \{c\}$ is a sufficiently expressive set of cost functions, then uniform **Lipschitzness of** $Q^{f,\hat{\pi},c}$ **over** $c \in \mathscr{C}$ **is equivalent to incremental stability of** $(f,\hat{\pi})$

Weirdness of Continuous Action Spaces

(and the power of **non-simple policies**)

We need new notions of 'coverage'

Theorem (Super Informal)**:** If the **expert trajectories** are sufficiently "**anti-concentrated**" in the sense that they have lower bounded "**local variance**", **then we can imitate without compounding error.**

Note: The expert always have "**perfect coverage**" of itself!

Takeaway: We need "metric," not just "probabilistic" notions of coverage in continuous action spaces!

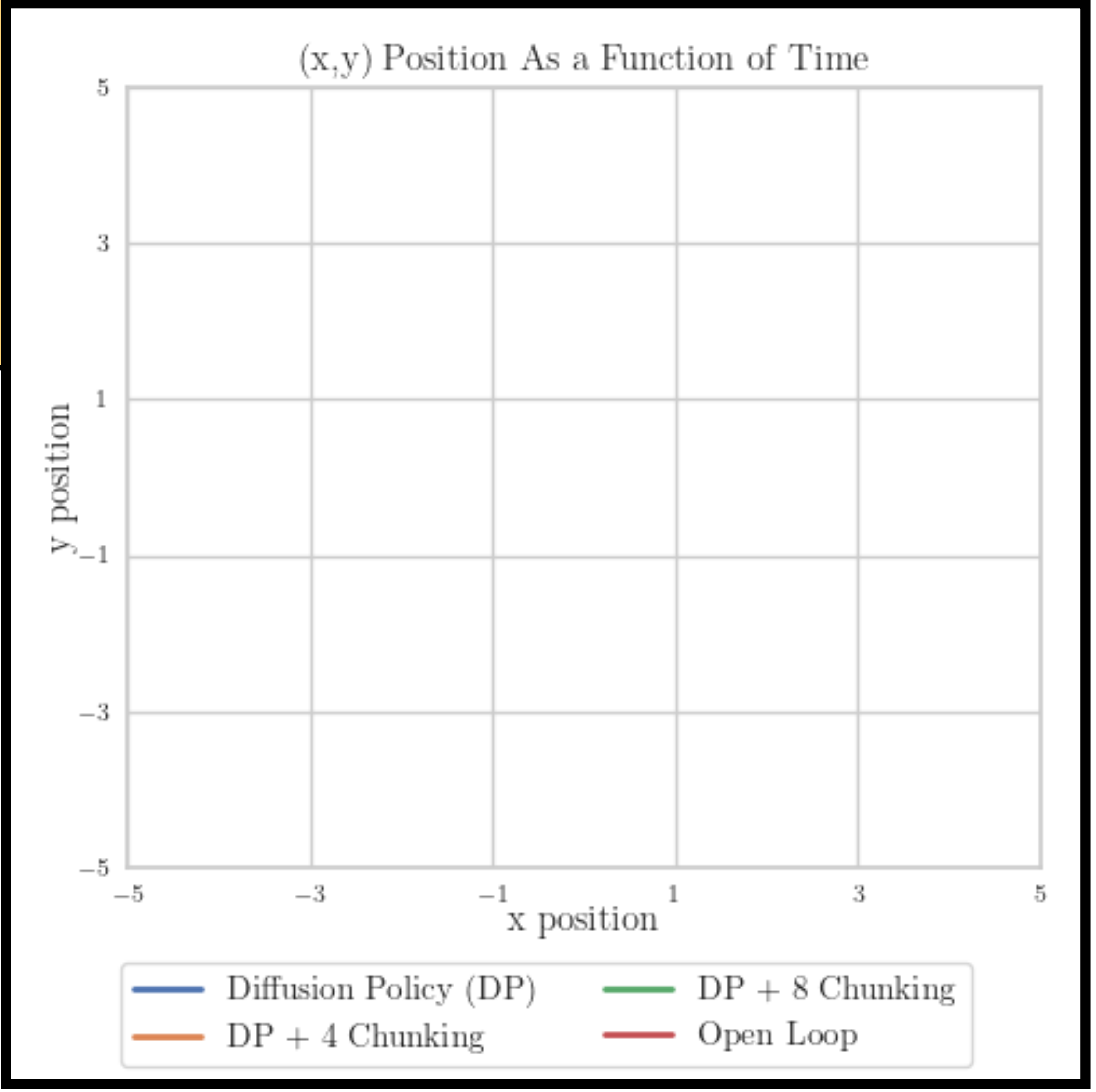
Algorithmic takeaway: We prove in forthcoming work that adding some exploration during data collection avoids compounding error, even if **open-loop unstable.**

Improper policies can be more powerful!

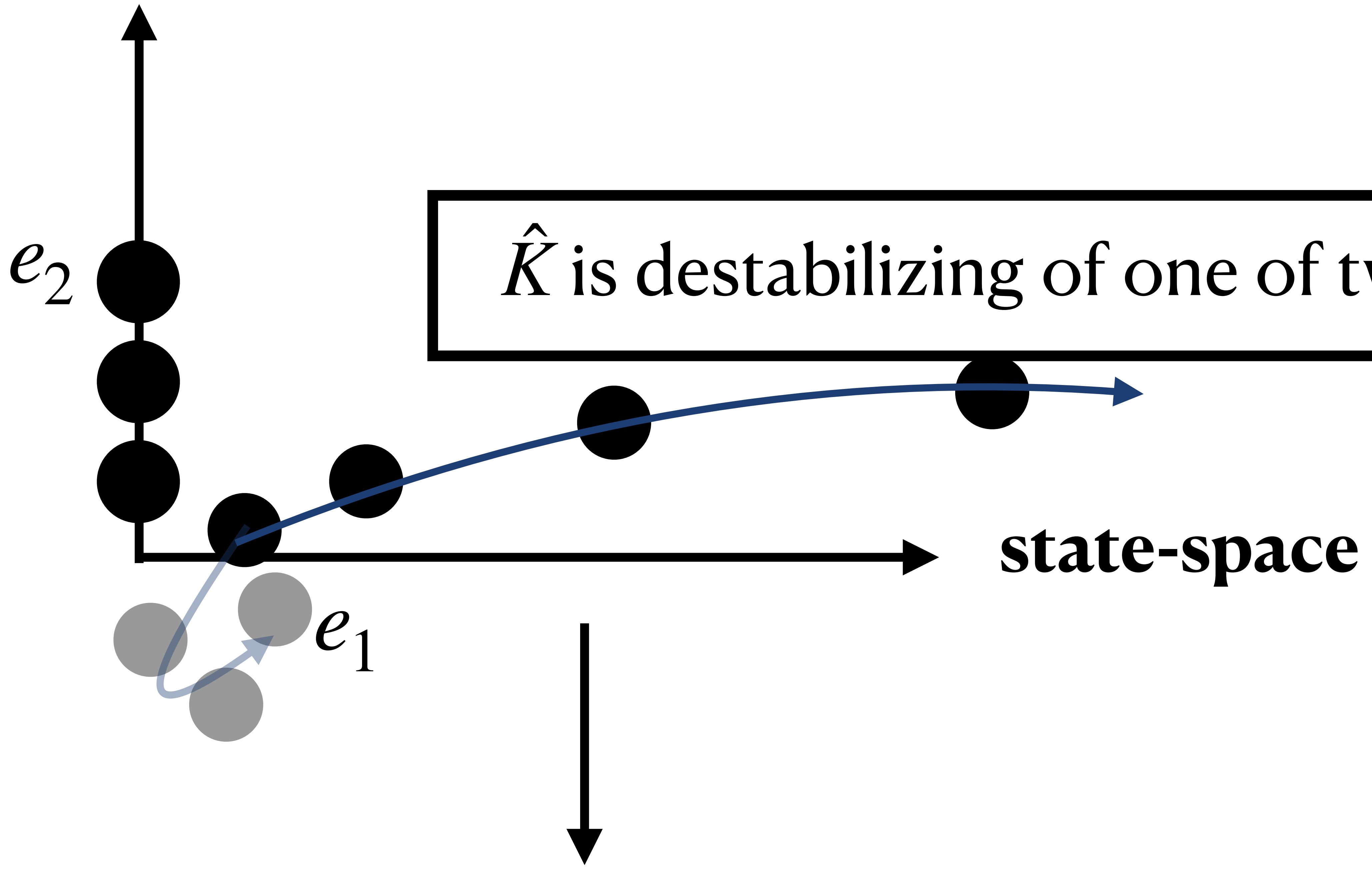
Theorem (Super Informal, forthcoming)**:** Under certain conditions, open-loop "**chunks**" of actions can result in bounded compounding error!

Longer chunks = **reduced compounding error!**

See also Block et al '24.



Food for thought: Stylizing Instability



Scalar Dynamics $x_{t+1} = \xi \rho x_t + u_t,\ \xi \in \{-1,1\}$ **unknown**, $\rho > 1$

unstable

Stylizing Instability

Scalar Dynamics $x_{t+1} = \xi\rho x_t + u_t$, $\xi \in \{-1,1\}$ **unknown**, $\rho > 1$

unstable

Observation: There is no linear feedback policy $\pi(x) = kx$ which stabilizes for both choices of ξ.

Proof: Under $\pi(x) = kx$, we have $x_{t+1} = (k+\xi\rho)x_t$

$\exists\xi$: **magnitude** > 1

Stylizing Instability

Scalar Dynamics $x_{t+1} = \xi \rho x_t + u_t,\ \xi \in \{-1,1\}$ **unknown**, $\rho > 1$ **unstable**

Observation: There is no linear feedback policy $\pi(x) = kx$ which stabilizes for both choices of k.

Corollary: There exists no **smooth, deterministic** policy which locally stabilizes.

Proof: Taylor Expansion and argue about linear approximation.

Stylizing Instability

Scalar Dynamics $x_{t+1} = \xi \rho x_t + u_t$, $\xi \in \{-1,1\}$ **unknown**, $\rho > 1$
unstable

Observation: There is no linear feedback policy $\pi(x) = kx$ which stabilizes for both choices of k.

Corollary: There exists no **simple** policy $\hat{\pi}(x) = \text{mean}(\hat{\pi}(x)) + z$ which locally stabilizes.

Lipschitz/smooth **independent of x**

Proof: Taylor Expansion and argue about linear approximation + noise.

Beyond Simple Policies

Scalar Dynamics $x_{t+1} = \xi\rho x_t + u_t,\ \xi \in \{-1,1\}$ **unknown**, $\rho > 1$

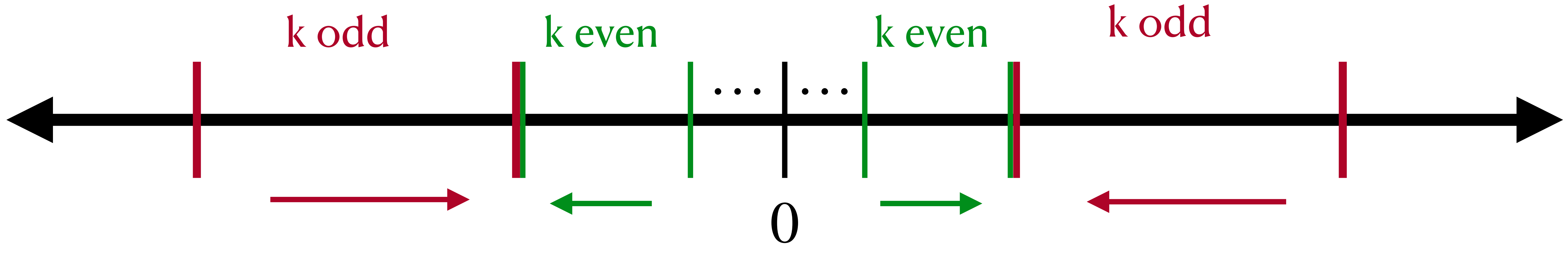
Observation: There is a very "simple", but **time-varying linear** policy which stabilizes the dynamics to in two times steps!

Proof: $\pi(x,t) = \begin{cases} \rho x & t \text{ even} \\ -\rho x & t \text{ odd} \end{cases}$

Concentric Stabilization

Scalar Dynamics $x_{t+1} = \xi\rho x_t + u_t$, $\xi \in \{-1,1\}$ **unknown**, $\rho > 1$

Observation: There is a deterministic, non-time varying **but non-smooth** policy which stabilizes around 0.



Proof: $\pi(x) = \begin{cases} \rho x & k \text{ even} \\ -\rho x & k \text{ odd} \end{cases}$ $\quad |x| \in ((2\rho^2)^{-k}, (2\rho^2)^{-(k-1)}]$

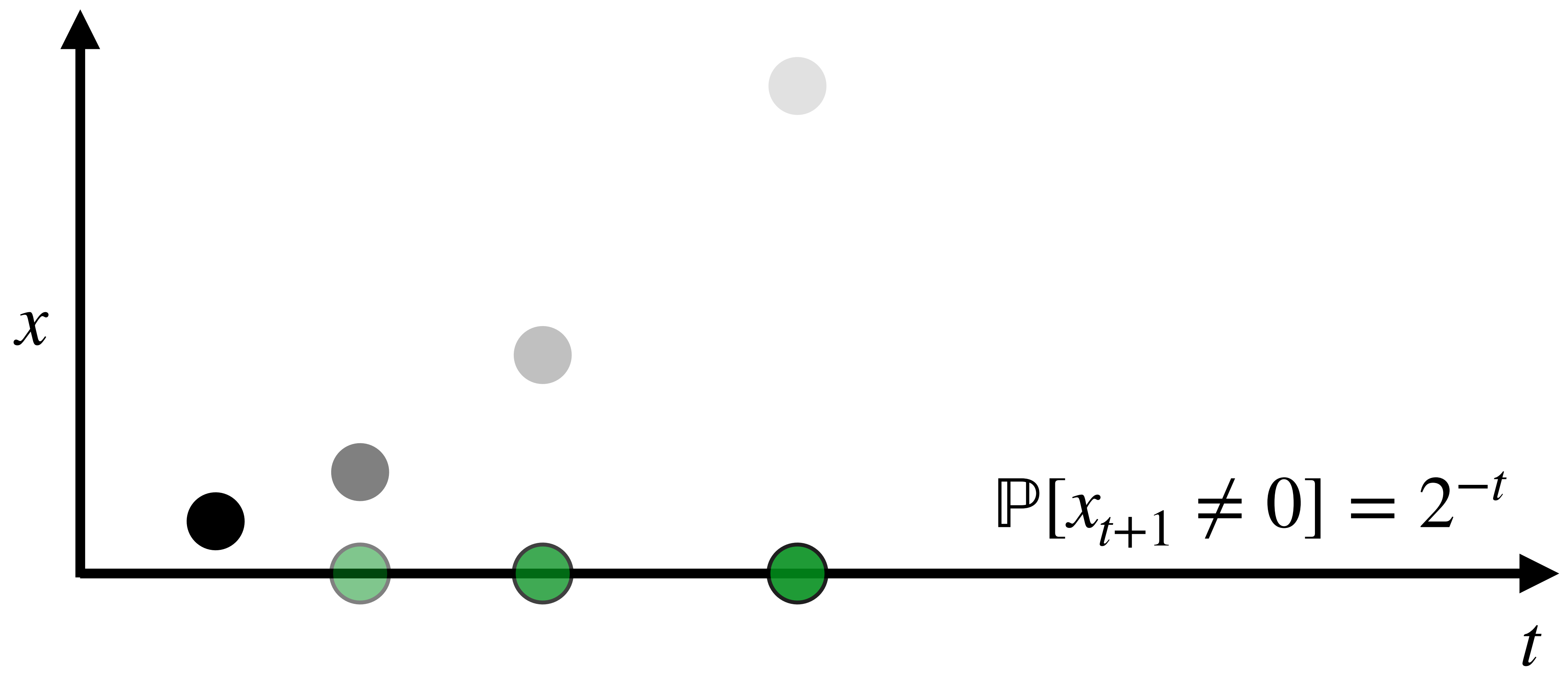
non-smooth

Benevolent Gambler's Ruin

Scalar Dynamics $x_{t+1} = \xi\rho x_t + u_t$, $\xi \in \{-1,1\}$ **unknown**, $\rho > 1$

Observation: There is a **stochastic, bi-modal** policy (i.e. **not-simple)** which stabilizes to the origin with high-probability.

$$\pi(x) = \begin{cases} \rho x & \text{w.p. } 1/2 \\ -\rho x & \text{w.p } 1/2 \end{cases}$$



Benevolent Gambler's Ruin

Scalar Dynamics $x_{t+1} = \xi\rho x_t + u_t$, $\xi \in \{-1,1\}$ **unknown**, $\rho > 1$

Observation: There is a **stochastic, bi-modal** policy (i.e. **not-simple)** which stabilizes to the origin with high-probability.

$$\pi(x) = \begin{cases} \rho x & \text{w.p. } 1/2 \\ -\rho x & \text{w.p } 1/2 \end{cases}$$

game: learner vs. "nature"

randomization over uncertainty in dynamics

Benevolent Gambler's Ruin

Scalar Dynamics $x_{t+1} = \xi \rho x_t + u_t, \;\; \xi \in \{-1, 1\}$ **unknown**, $\rho > 1$

Observation: There is a **stochastic, bi-modal** policy (i.e. **not-simple)** which stabilizes to the origin with high-probability.

Diffusion Policy, Chi et. al '23

game: learner vs. "nature"

randomization over uncertainty in dynamics

Surprising Takeaway: Stochastic, multi-modal policies can yield benefits, even for imitating deterministic policies.

What are the fundamental benefits of generative models for solving optimal control tasks?

Surprising Takeaway: Stochastic, multi-modal policies can yield benefits, even for imitating deterministic policies.

… for you RL theorists:

Takeaway 2: Re-think our assumptions on the class of Q functions!

Takeaway 3: Re-thinking coverage for continuous action spaces!

Takeaway 4: Re-think policy parametrization for scaling robot learning!