# D-optimal exploration of physical systems


Matthieu Blanke, Marc Lelarge

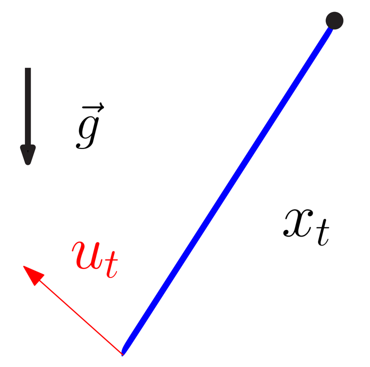
DI ENS, INRIA Paris

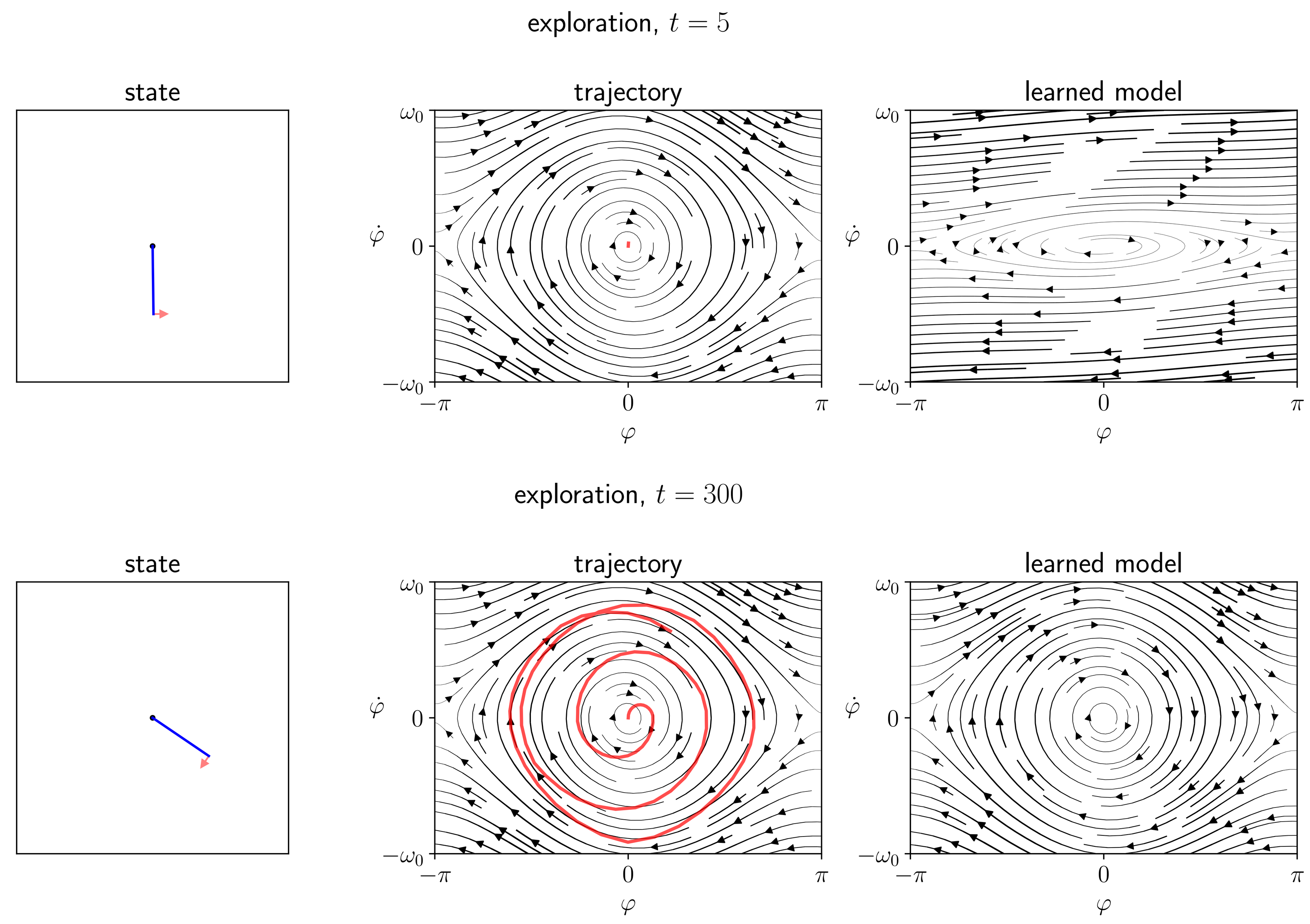

## Motivation

In order to be effective, control theory needs a faithful model of the controlled system. In most cases, parameters must be fitted using experimental data, but running experiments is coslty so the system must be explored efficiently.

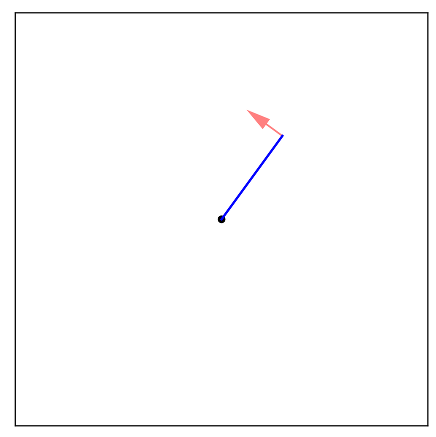
**Exploration**: an agent takes actions to move in an unknown environment in order to map it.

Environment with unknown dynamics




After **exploration** the learned model can then serve model-based **exploitation**: a downstream contrnol task is solved using the learned model.



## The problem

Nonlinear dynamics in $\mathbb{R}^d$:

$$\boxed{x_{t+1} = x_t + \mathrm{d}t f(x_t, u_t) + w_t, \quad 0 \leq t \leq T-1},$$

with noise $w_t \sim \mathcal{N}(0, \sigma^2 I_d)$, control variables $u_t \in \mathbb{R}^m$ subject to the power constraint $\|u_t\|^2 \leq \gamma^2$. A model $f_\theta$ is fitted by regression on the past observations:

$$\theta_t = \hat{\theta}\left(x_{0:t+1}, u_{0:t}\right)$$

We focus on online algorithms.

**Goal** Find a policy yielding inputs $(u_t)$ that drive the system towards a maximally informative trajectory, with small computational complexity.

## Exploration algorithm

**Policy** $\pi : (x_{0:t}, u_{0:t-1}; f_\theta) \mapsto u_t$ models our choice of the inputs.

**Algorithm 1** Online neural exploration

```
input neural model f_θ, policy π, time horizon T, time-step dt, learning rate η
output dynamics model f_θ
for 0 ≤ t ≤ T − 1 do
    choose u_t = π_t(x_{0:t}, u_{0:t−1}; f_θ)
    observe x_{t+1} = x_t + dt f(x_t, u_t)
    compute the loss
        ℓ_t = ‖f_θ(x_t, u_t) − (x_{t+1} − x_t)/dt‖²_2
    update θ ← θ − η∇ℓ_t(θ)
end for
```

## Input design

How to choose $u_t$ ? The theory of linearized optimal design suggests optimizing the Gram matrix of the covariates:

$$\max_{(z_s)} \log\det\left(\mathbb{E}[M_T]\right)$$

with $\quad M_t = \sum_{s=0}^{t-1} J_s^\top J_s$ and $J_t = \dfrac{\partial f_\theta}{\partial \theta}(x_t, u_t, \theta).$

We derive a tractable, greedy approximation of this objective yielding a quadratic optimization problem. The resulting exploration algorithm is is fast, online, and experiments show that it is sample efficient.

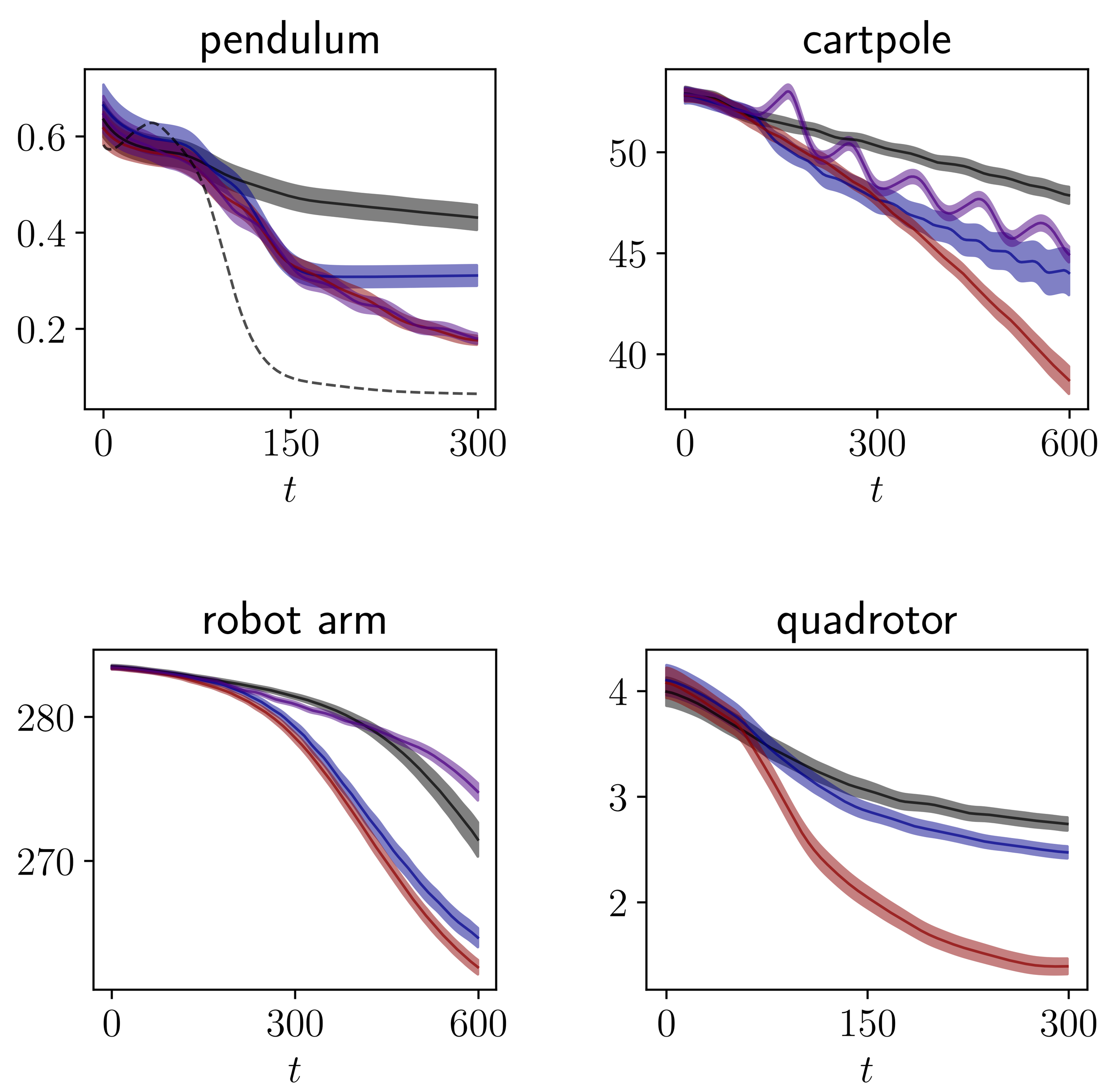
## Results

**Experiments** We test our exploration algorithm on several environments from classical control. The dynamics are initially unknown and are learned online. The models $f_\theta$ include neural networks. Our D-optimal policy is compared with baselines.

**Baselines** Random inputs, maximally uniform trajectory in the state space, periodic inputs.

**Results** Our D-optimal policy is sample efficient. It yields large amplitude trajectories that are informative for the underlying model.

**$L^2$ error against time**




**D-optimal trajectories in the phase space**