Control by Reinforcement Learning: The fluidic pinball

Institut Pprime

CNRS, Université de Poitiers, ISAE-ENSMA

COWAVE ANR-17-CE22-0008

Thibaut Guégan
guegan.thibaut@gmail.com

Onofrio Semeraro
onofriosem@gmail.com

Laurent Cordier
laurent.cordier@univ-poitiers.fr

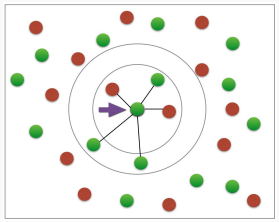
M. Alessandro Bucci
bucci.malessandro@gmail.com

Lionel Mathelin
mathelin@limsi.fr

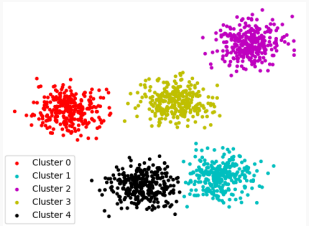
23 November 2020

Why Machine Learning?

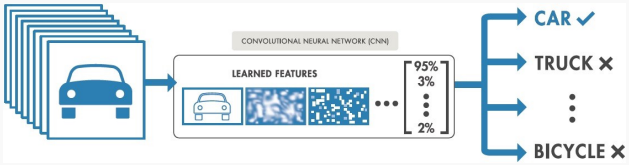
"Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed" Samuel A. - 1959



(a) K-NN



(b) Clustering



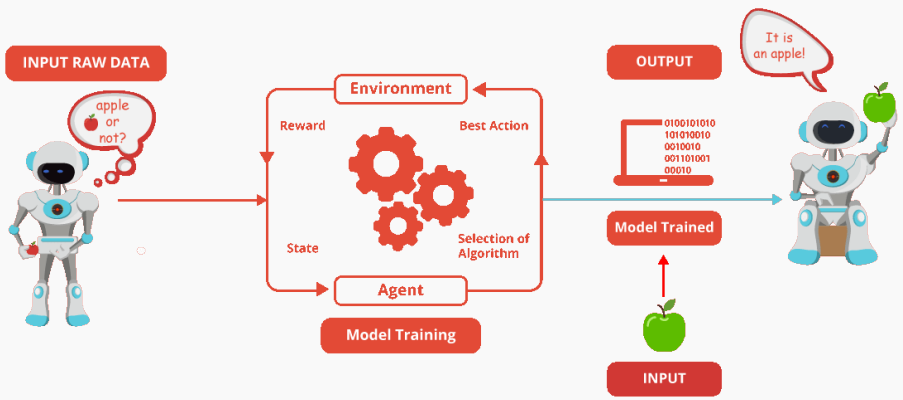
(c) CNN

"Multilayer feedforward networks are capable of approximating any measurable function to any desired degree of accuracy"

Hornik K., Stinchcombe M., White H. - 1989

Types of Machine Learning

- Supervised Learning:
  Need a training set of labeled data
  Example: classification, prediction
- Unsupervised Learning:
  Only need a raw set of data
  Example: SVD, clustering, autoencoders
- Reinforcement Learning:
  Look for the optimal policy to reach a goal
  Example: traffic light control, games, fluid mechanics



RL and Navier-Stokes: How?

NS equations can be seen as the environment:

- Inputs are observations of the flow field, drag, lift, etc.
- Outputs are the commands to send to actuators in order to maximize the reward
- The reward can be focused on reducing the drag, the lift, predicting the flow solution, etc.

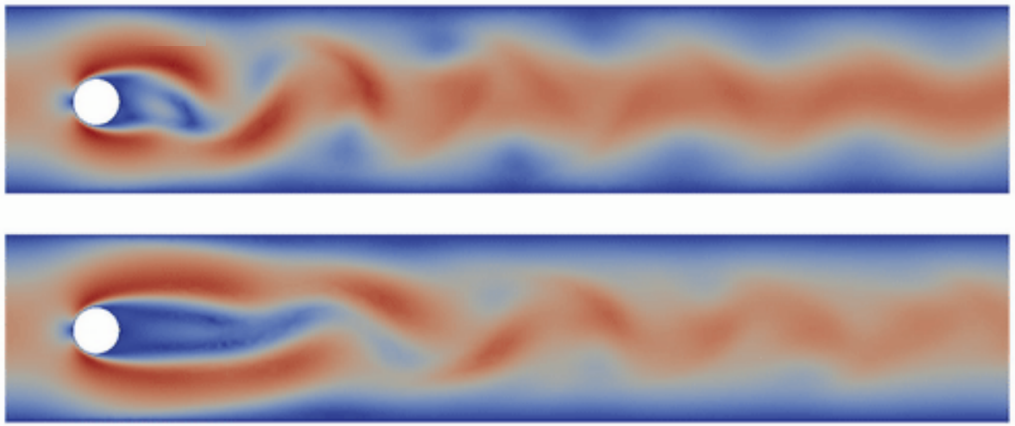


Figure: Control off (top) - Control on (bottom)
Rabault J. et al. (2019)

CFD configuration

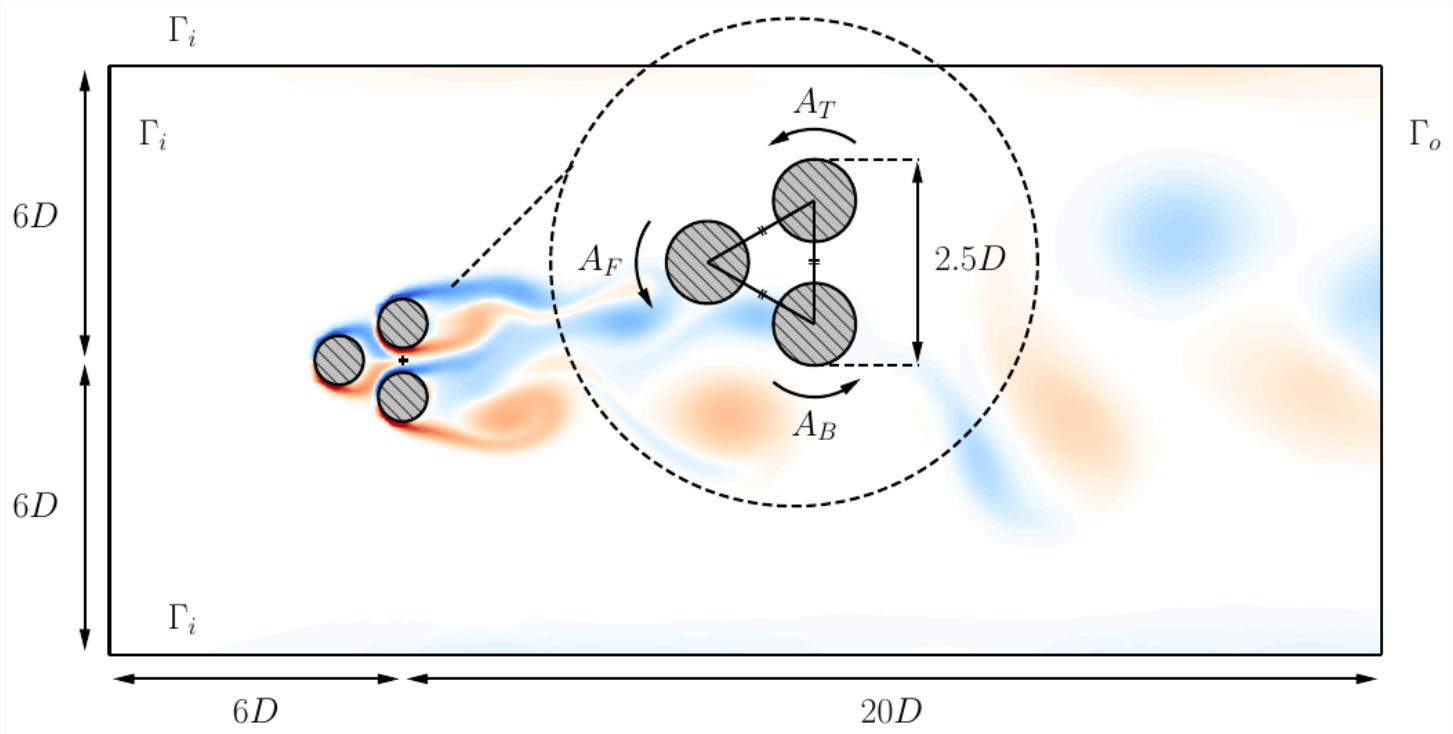


Figure: Each angular velocity is denoted by A_X, $Re = 100$

CFD configuration

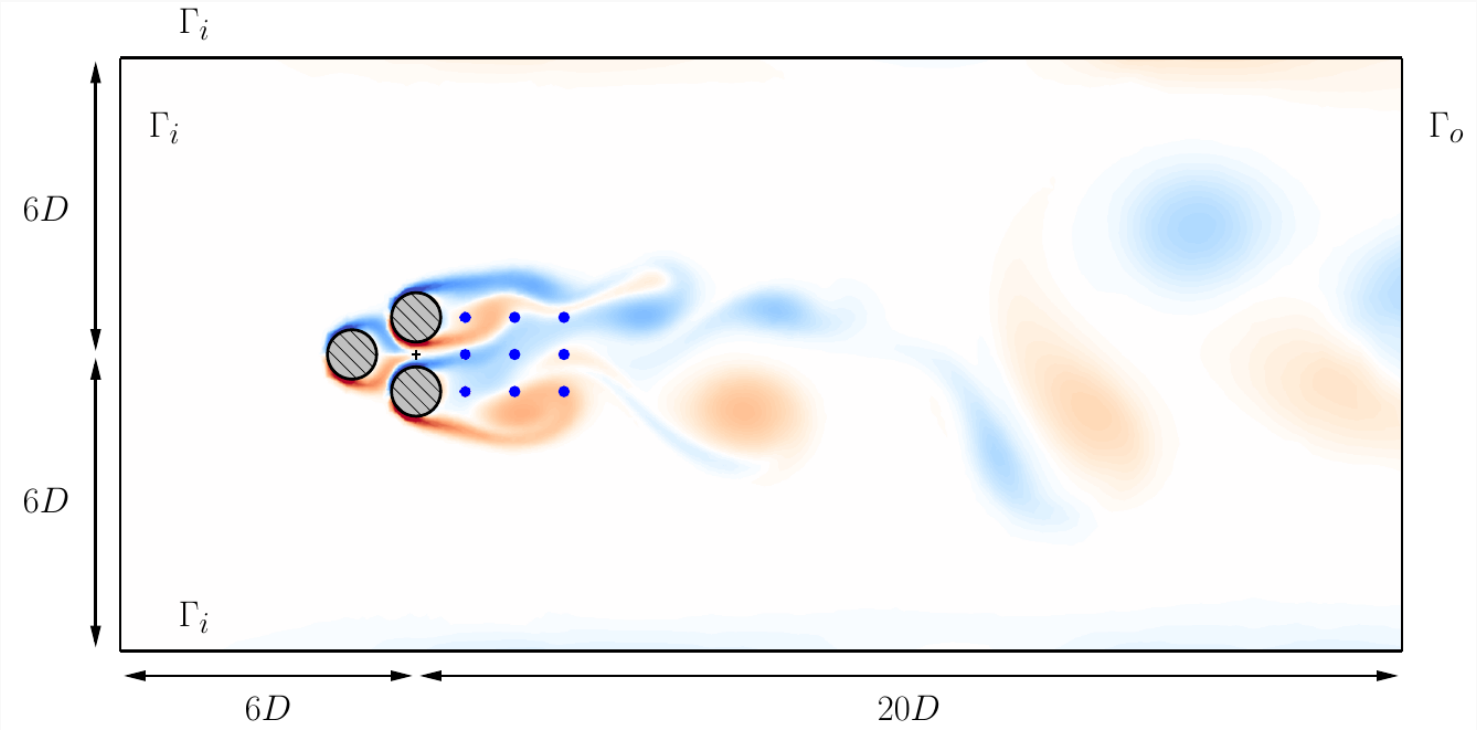


Figure: Pressure probes positioning

Reinforcement Learning setup

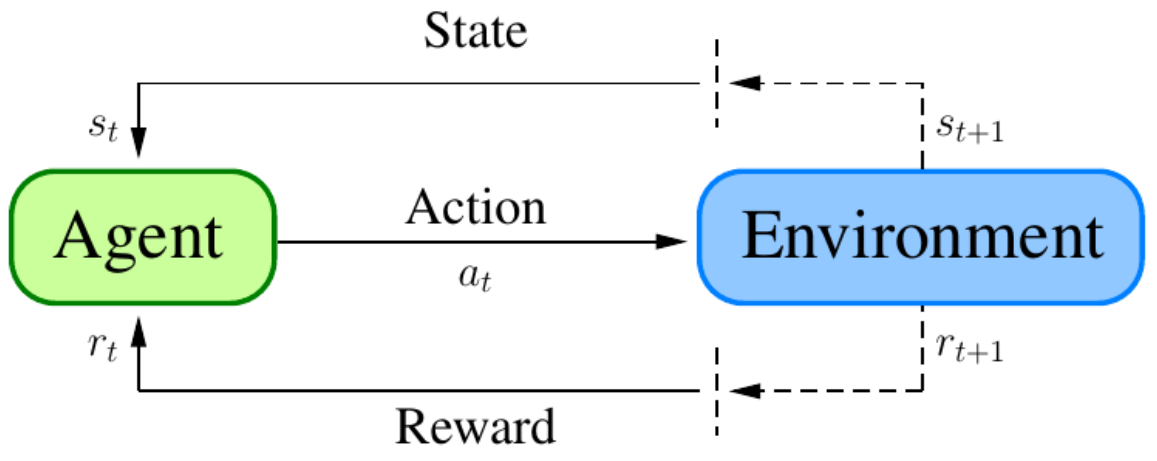


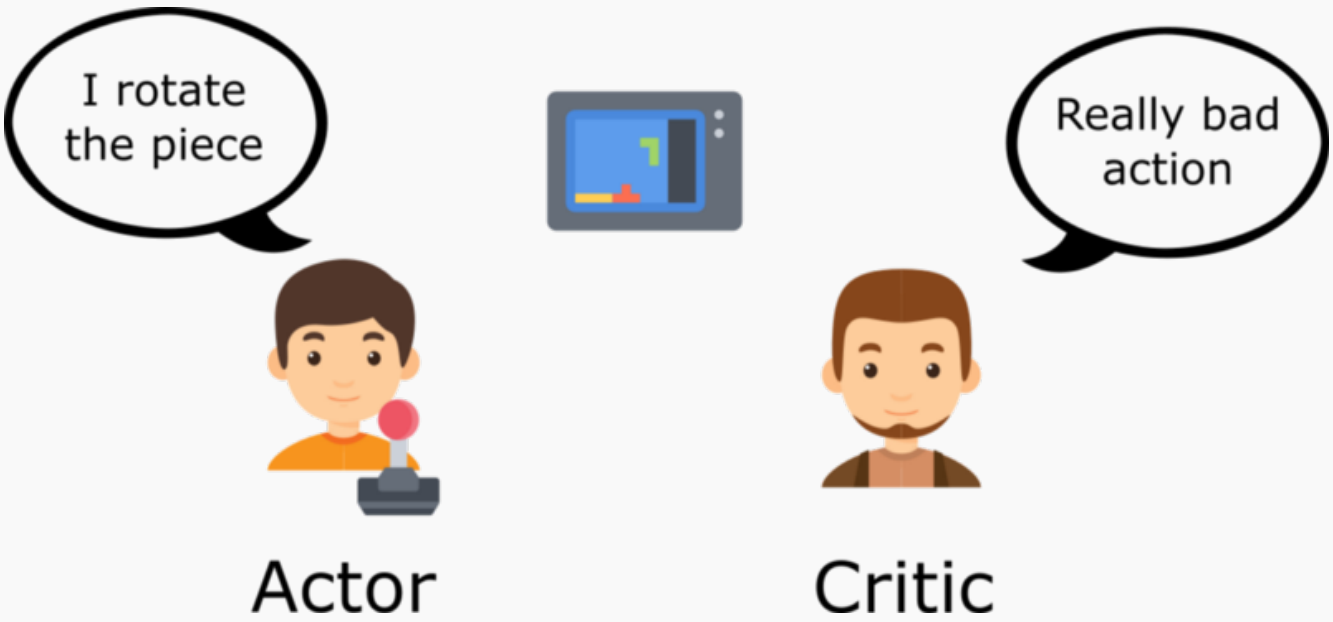
Figure: RL framework

- Actor-critic agent: TD3
- Environment: Fluidic pinball

- State (input): Pressure probes
- Action (output): Angular velocities
- Reward: $-F_D^2 - \alpha||a_t||_2$

The action range is set to $[-3, 3]$ rad/s

Actor-Critic principle



- The actor takes as input the state and outputs the best action.
- The critic evaluates the action by computing the value function.

Temporal Difference Learning

The action-value function is formally defined as:

$$Q^{\pi}(s,a) = \mathbb{E}\left[\sum_{k=t}^{\infty} \gamma^{k-t} r(s_k, a_k) | s_t = s, a_t = a\right]$$

All TDL methods are based on the fact that Q^{π} satisfies the Bellman equation:

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}[r(s_t, a_t) + \gamma Q^{\pi}(s_{t+1}, a_{t+1}) | s_t = s, a_t = a]$$

The classical TD(0) algorithm is then defined as:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(\overbrace{r(s_t, a_t) + \gamma Q(s_{t+1}, a_{t+1})}^{\text{TD target}} - Q(s_t, a_t))$$

Twin Delayed DDPG (TD3)

Deep Deterministic Policy Gradient (DDPG) works well but highly sensitive to hyperparameters

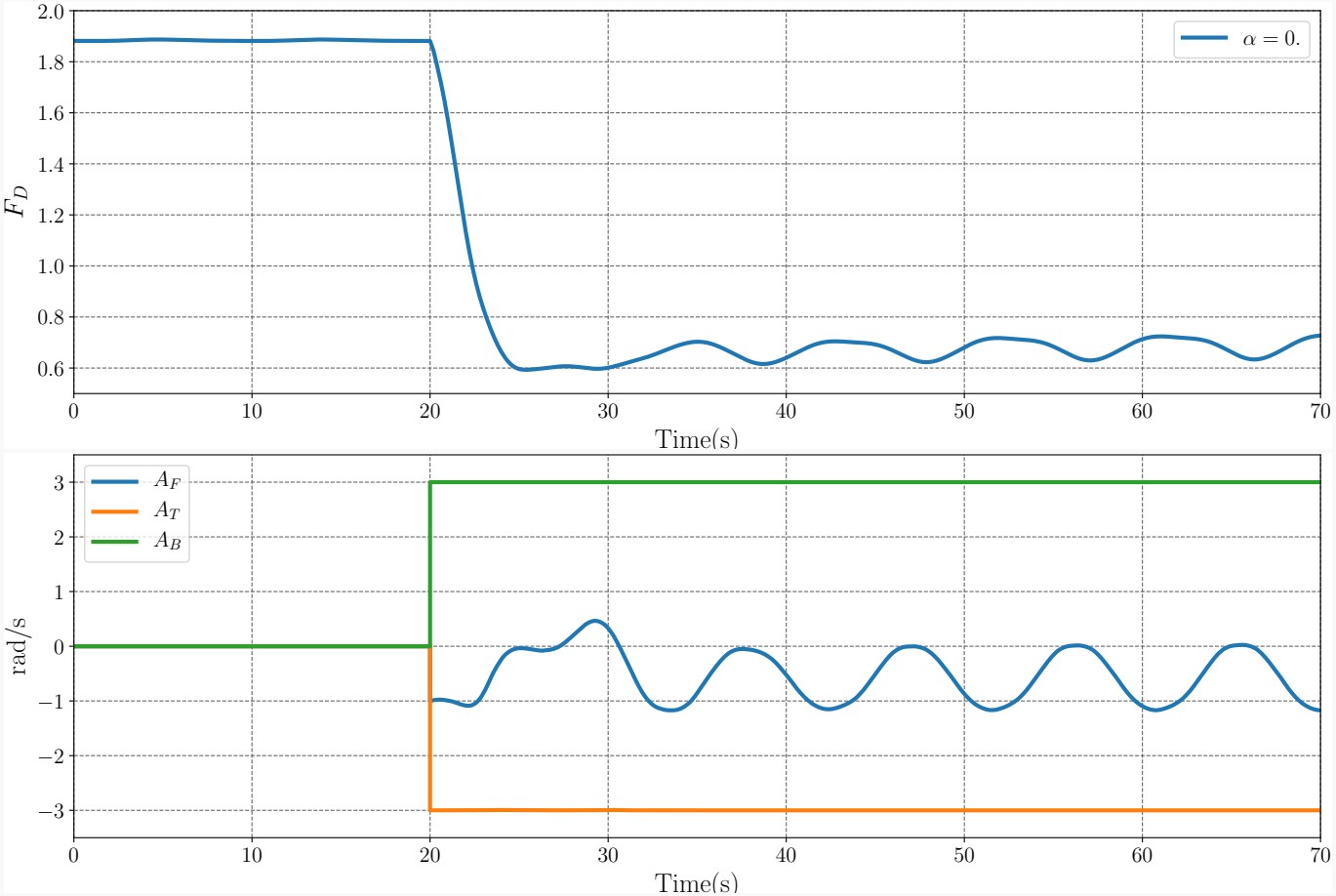
TD3 uses three tricks:

- Clipped Double-Q Learning
- “Delayed” Policy Updates
- Target Policy Smoothing

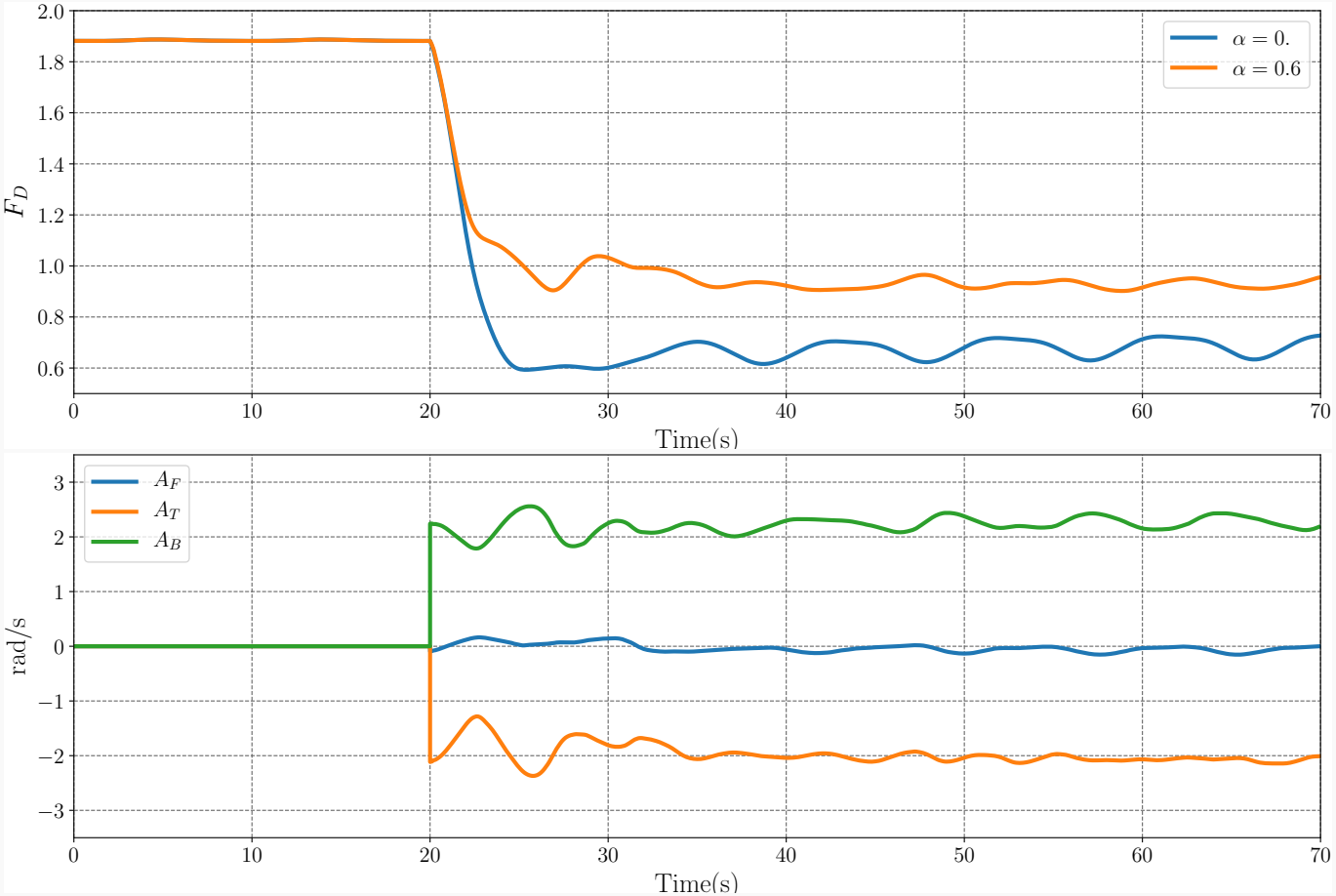
Quick facts:

- Off-policy algorithm
- Continuous action spaces only

9 probes, no penalization



9 probes, with penalization



Effect of the number of probes

Quick summary:

- 9 probes without penalization $\approx 65\%$ drag force reduction
- 9 probes with penalization $\approx 50\%$

Effect of the number of probes

Quick summary:

- 9 probes without penalization $\approx 65\%$ drag force reduction
- 9 probes with penalization $\approx 50\%$

Trying to reduce the number of probes:

- 3 probes with penalization $\approx 35\%$
- 1 probe with penalization $\approx 25\%$

Performances drops by a lot... Any solution?

Effect of the number of probes

Quick summary:

- 9 probes without penalization $\approx 65\%$ drag force reduction
- 9 probes with penalization $\approx 50\%$

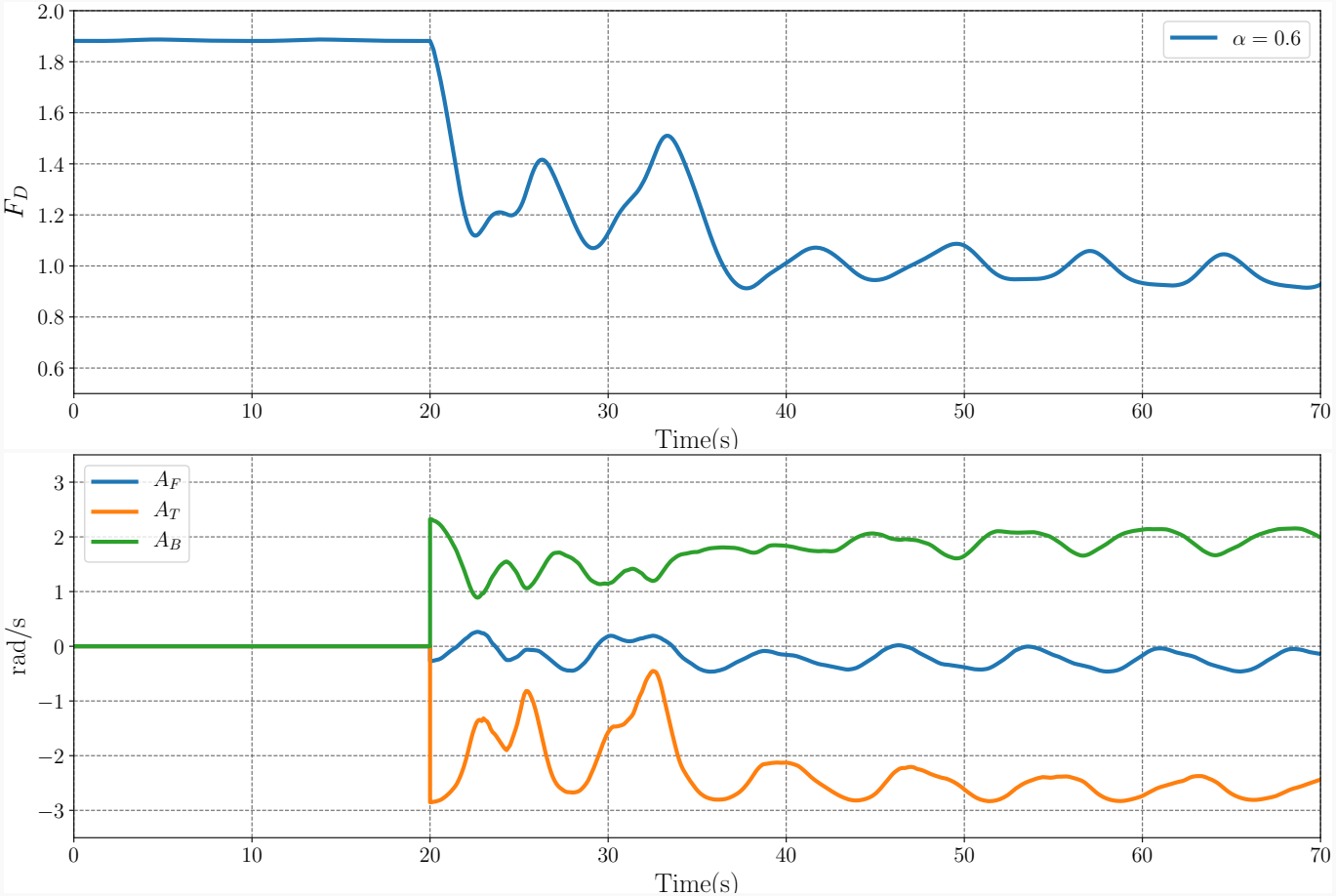
Trying to reduce the number of probes:

- 3 probes with penalization $\approx 35\%$
- 1 probe with penalization $\approx 25\%$

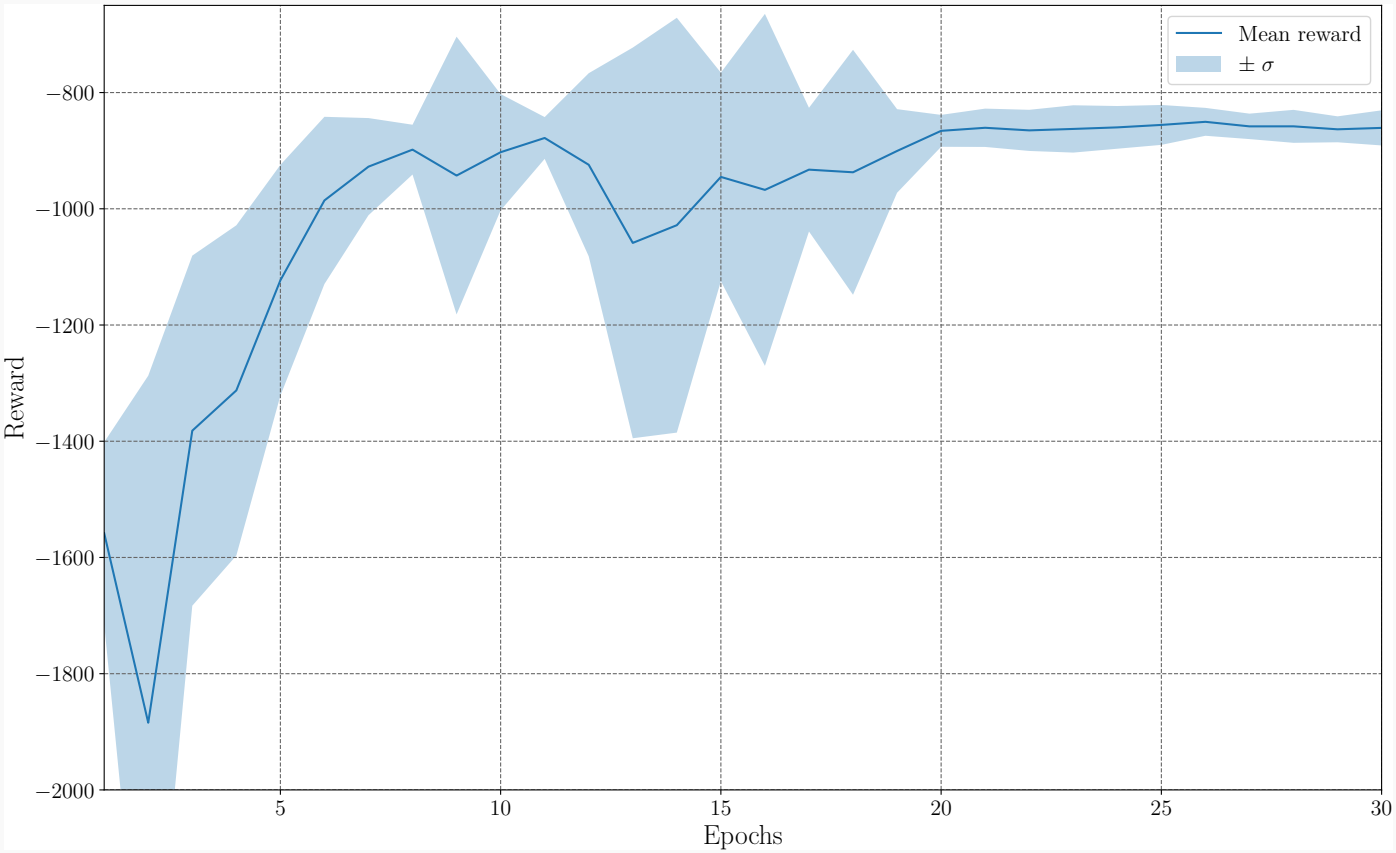
Performances drops by a lot... Any solution?

Takens's embedding theorem

3 probes, with penalization



Standard deviation, latest case



Summing up

- One of the first efforts
- Arbitrary placement of probes

Summing up

- One of the first efforts
- Arbitrary placement of probes

Pros:

- Can be trained on various Reynolds numbers
- Fast training $\approx$ 4 vortex shedding per epochs and 20 epochs at most

Summing up

- One of the first efforts
- Arbitrary placement of probes

Pros:

- Can be trained on various Reynolds numbers
- Fast training $\approx$ 4 vortex shedding per epochs and 20 epochs at most

Cons:

- Large time delay (credit assignment problem)

Summing up

- One of the first efforts
- Arbitrary placement of probes

Pros:

- Can be trained on various Reynolds numbers
- Fast training $\approx$ 4 vortex shedding per epochs and 20 epochs at most

Cons:

- Large time delay (credit assignment problem)

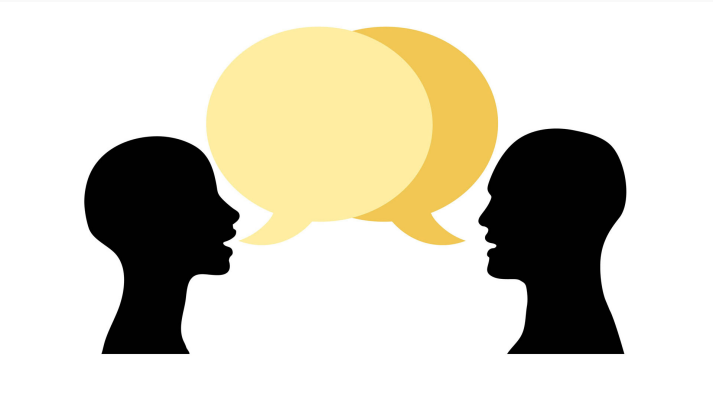
The experimental open cavity is the next milestone

On demand attendee meetings

Want to have a talk?

Availability:

- Tuesday 24 from 03:00 PM to 05:00 PM - CET
  08:00 AM to 10:00 AM - CST



Thank you for your attention!

Any questions?