Imitation Learning: The Big Picture

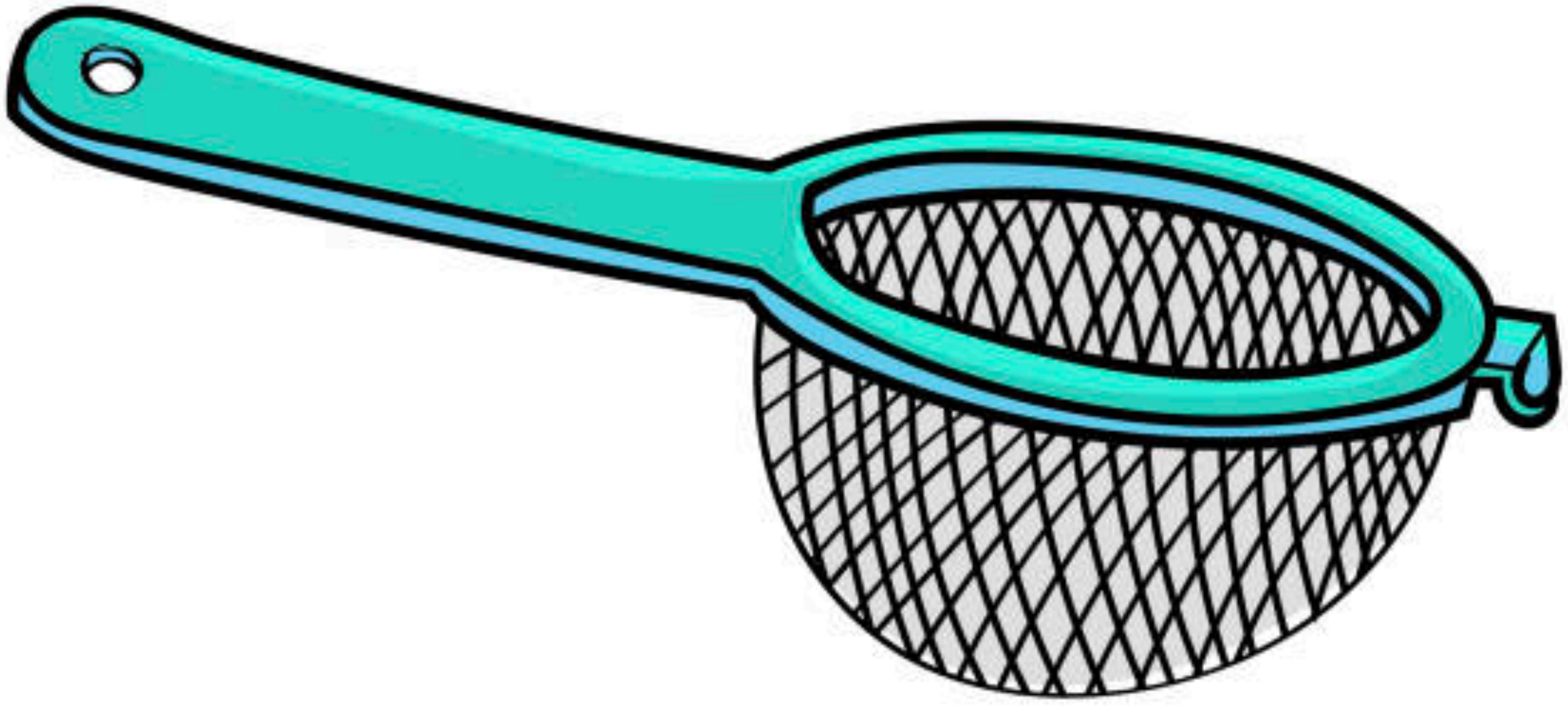
Sanjiban Choudhury

Cornell Bowers CIS
Computer Science

Imitation Learning ....

It's only a game!

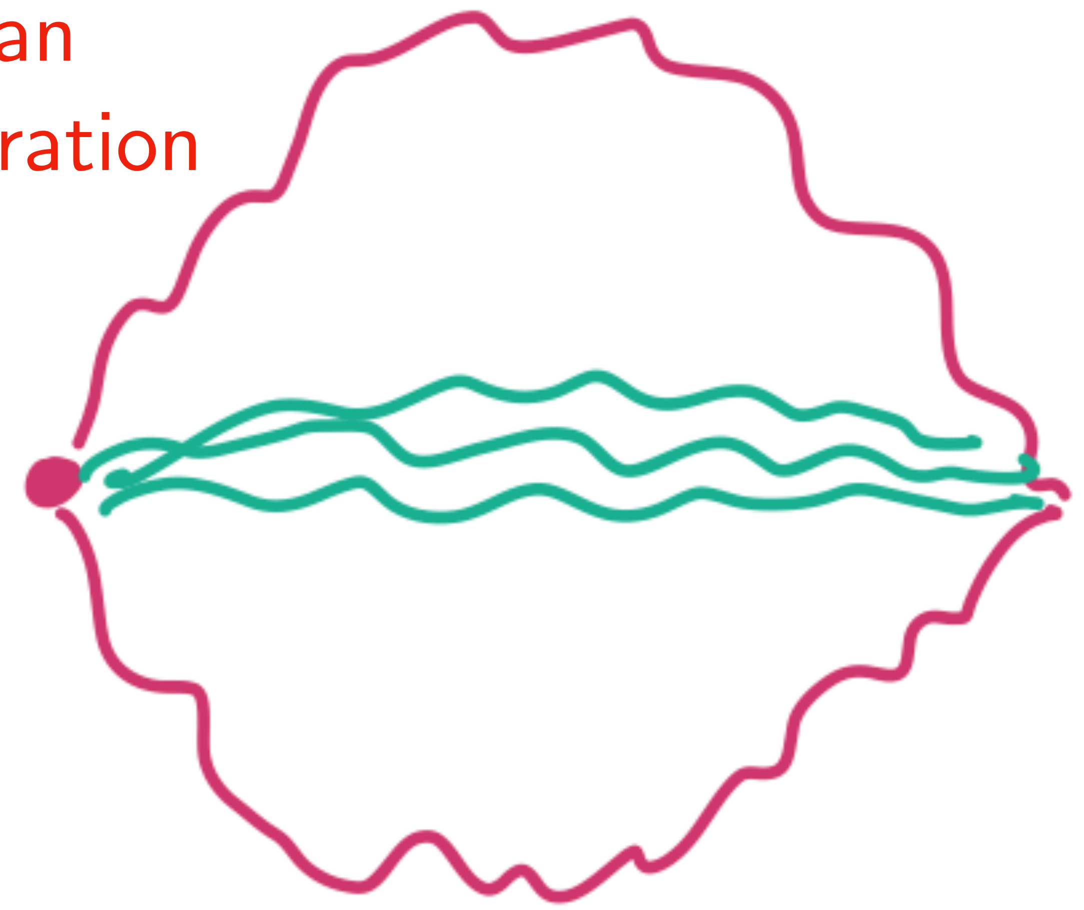
(And a rather simple one at that!)



But first ...

We are going to try to catch things we dropped in previous lectures

Maximum Entropy Inverse Optimal Control



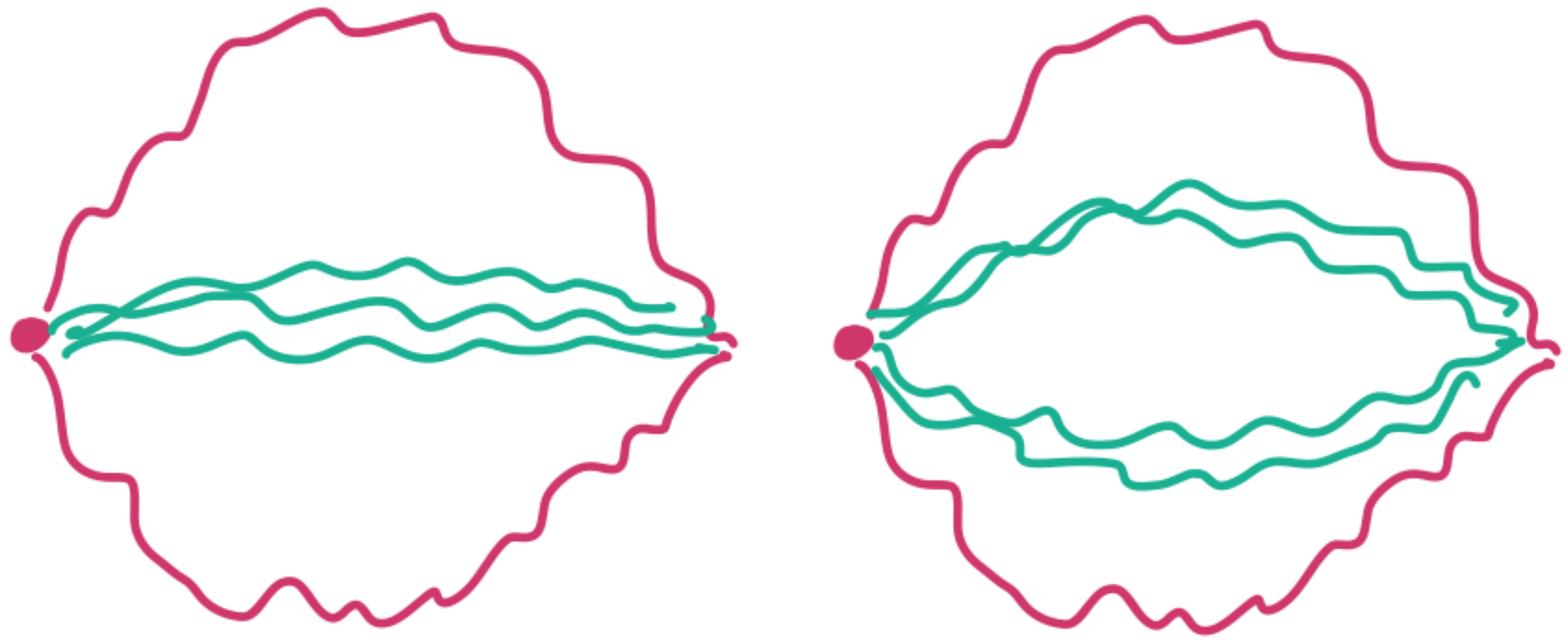
for $i = 1, \ldots, N$ # Loop over datapoints

$$\xi_i \sim \frac{1}{Z} \exp\left(-C_\theta(\xi, \phi_i)\right)$$ # Call planner!

$$\theta^+ = \theta - \eta[\nabla_\theta C_\theta(\xi_i^h, \phi_i) - \nabla_\theta C_\theta(\xi_i, \phi_i)]$$ # Update cost

(Push down human cost) *(Push up planner cost)*

Maximum Entropy Inverse Optimal Control

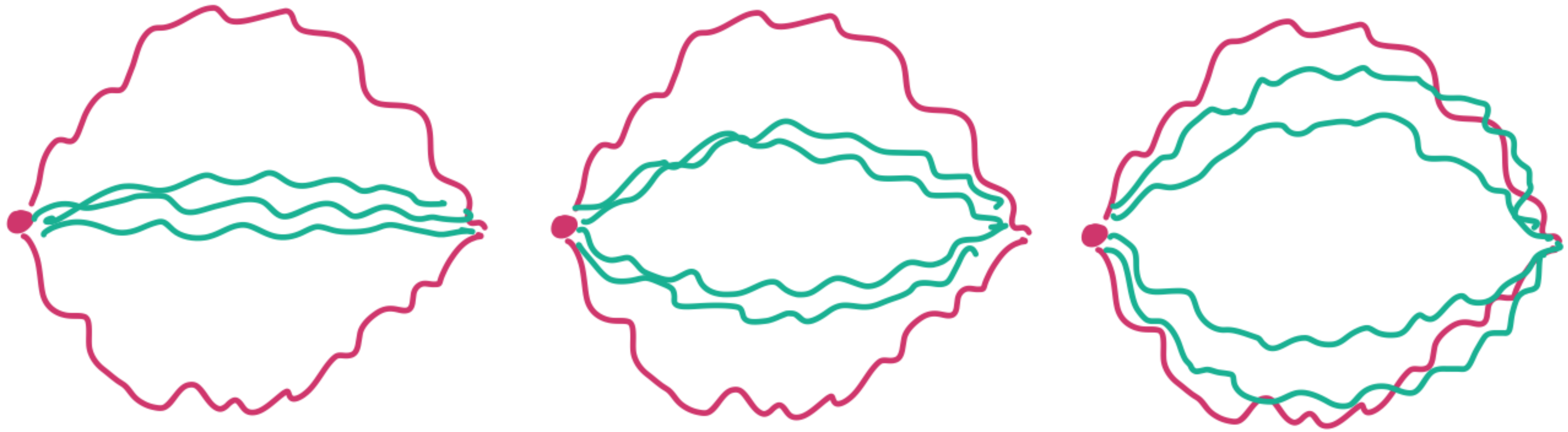


for $i = 1, \ldots, N$ # Loop over datapoints

$$\xi_i \sim \frac{1}{Z} \exp\left(-C_\theta(\xi, \phi_i)\right)$$ # Call planner!

$$\theta^+ = \theta - \eta[\underbrace{\nabla_\theta C_\theta(\xi_i^h, \phi_i)}_{\text{(Push down human cost)}} - \underbrace{\nabla_\theta C_\theta(\xi_i, \phi_i)}_{\text{(Push up planner cost)}}]$$ # Update cost

Maximum Entropy Inverse Optimal Control



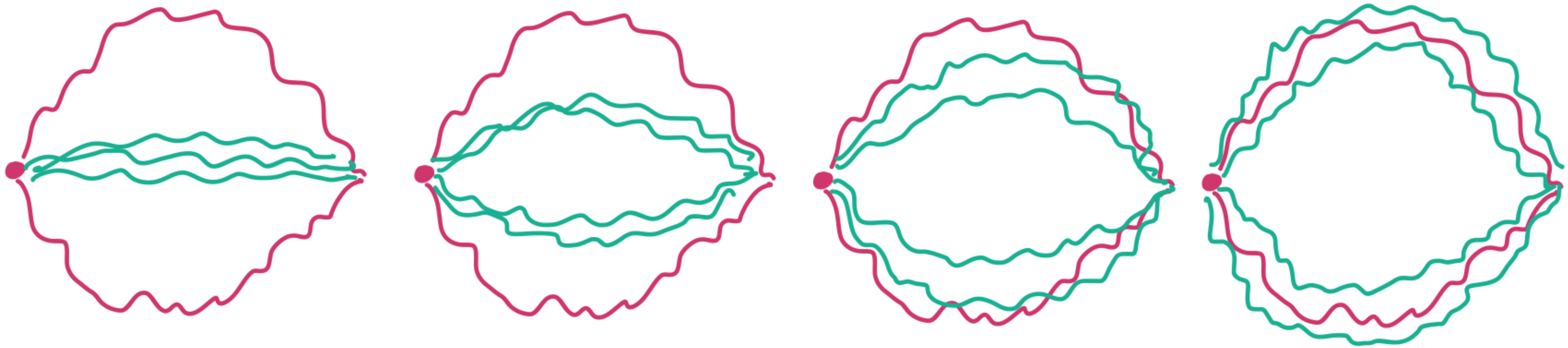
for $i = 1, \ldots, N$ # Loop over datapoints

$$\xi_i \sim \frac{1}{Z} \exp\left(-C_\theta(\xi, \phi_i)\right)$$ # Call planner!

$$\theta^+ = \theta - \eta[\nabla_\theta C_\theta(\xi_i^h, \phi_i) - \nabla_\theta C_\theta(\xi_i, \phi_i)]$$ # Update cost

(Push down human cost) *(Push up planner cost)*

Maximum Entropy Inverse Optimal Control



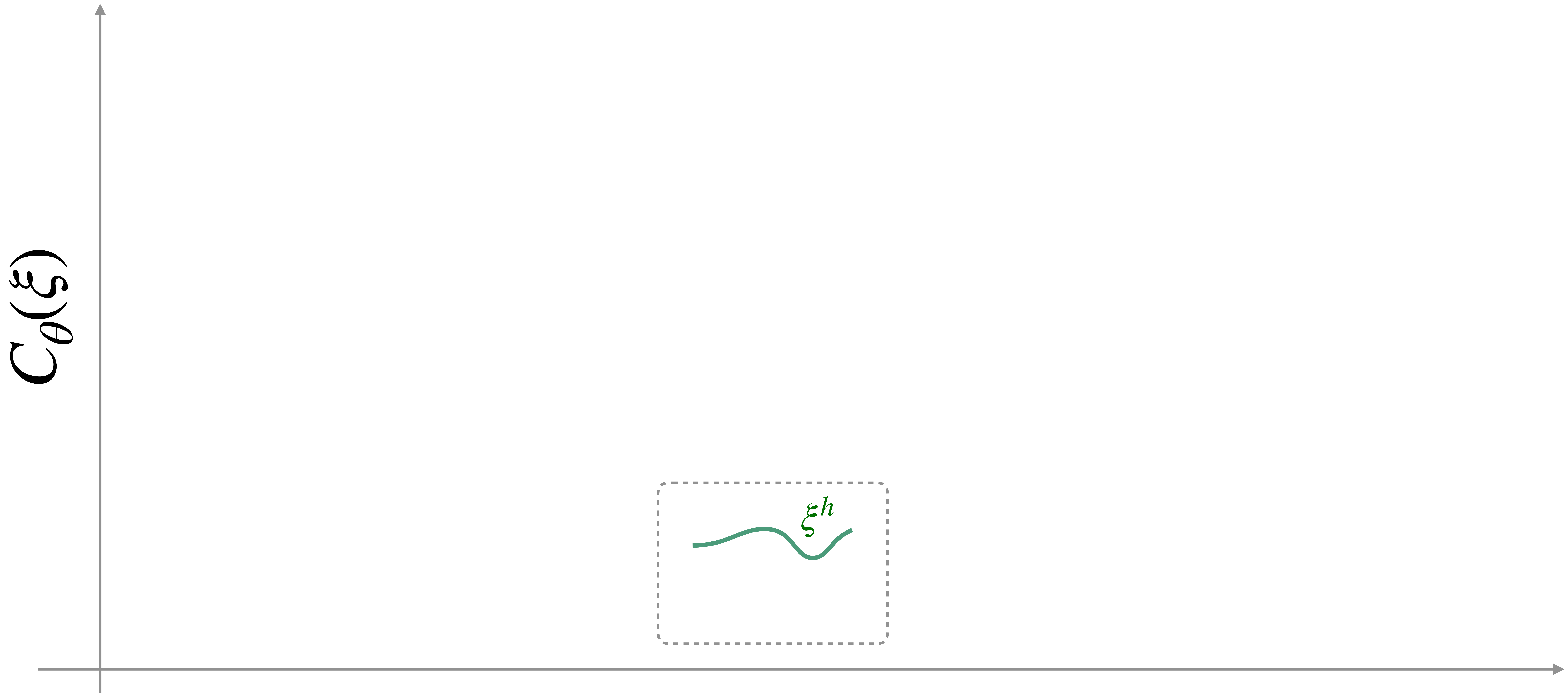
for $i = 1,\ldots,N$ # Loop over datapoints

$$\xi_i \sim \frac{1}{Z}\exp\left(-C_\theta(\xi,\phi_i)\right)$$ # Call planner!

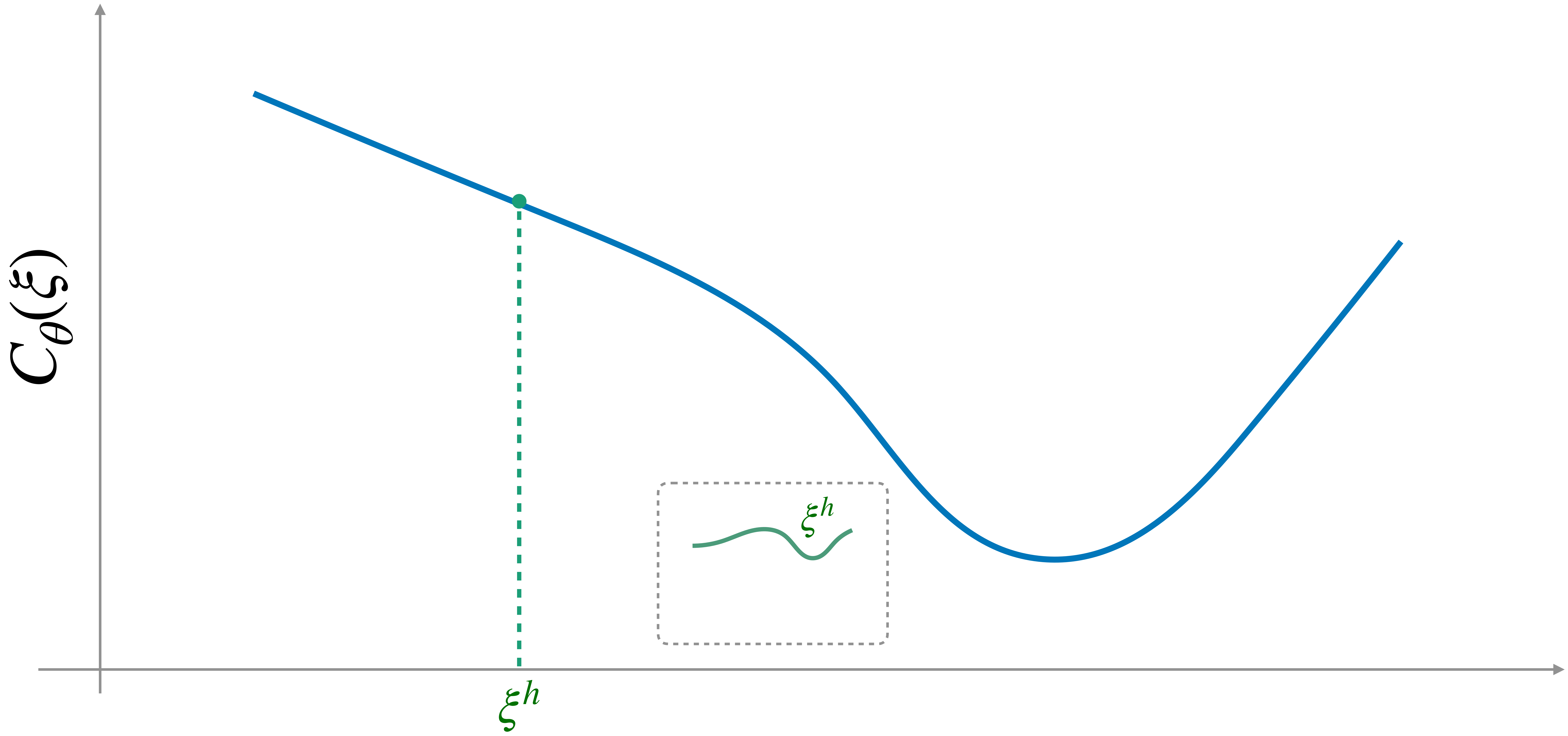
$$\theta^+ = \theta - \eta[\nabla_\theta C_\theta(\xi_i^h,\phi_i) - \nabla_\theta C_\theta(\xi_i,\phi_i)]$$ # Update cost

(Push down human cost) *(Push up planner cost)*

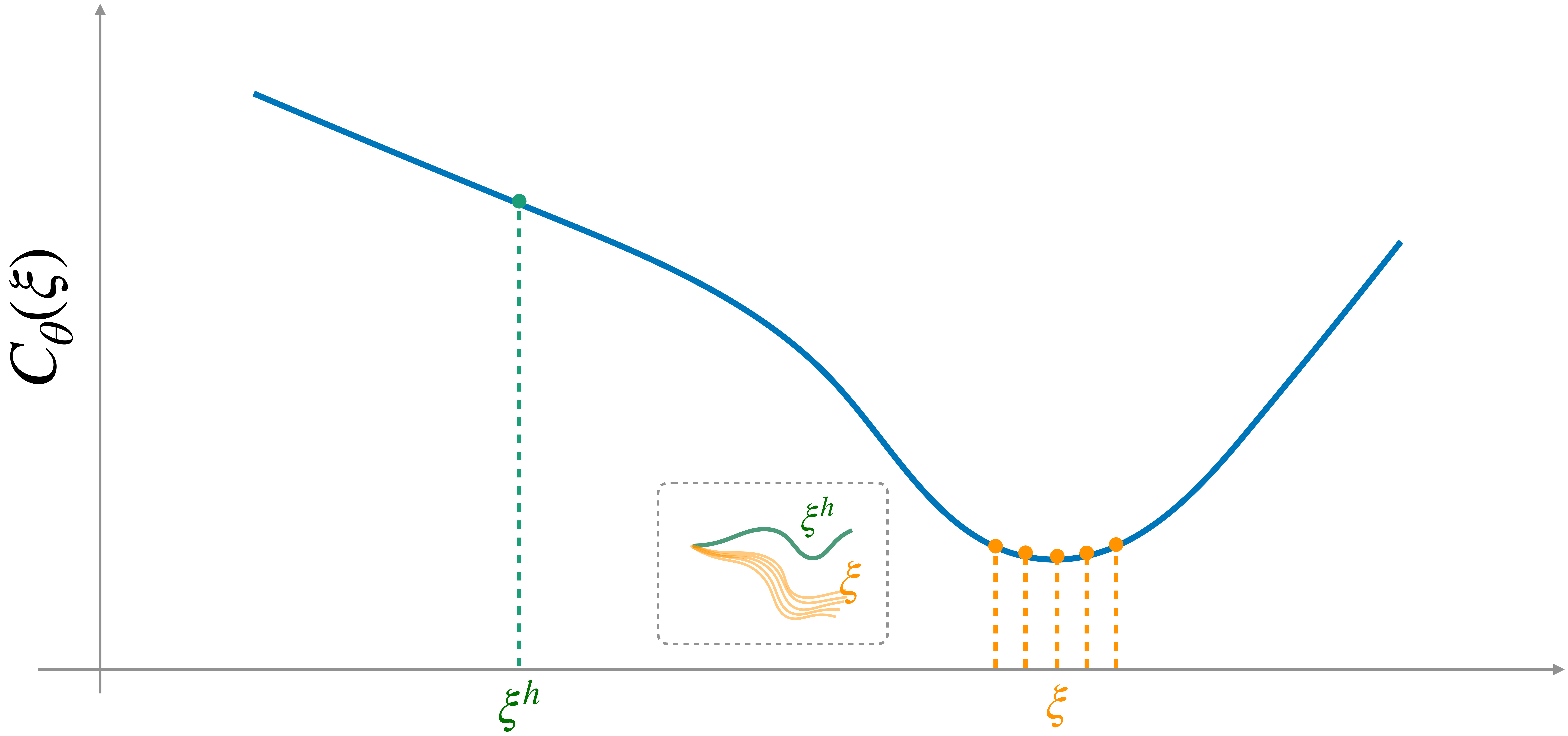
Animation from Sodhi et al. 2021



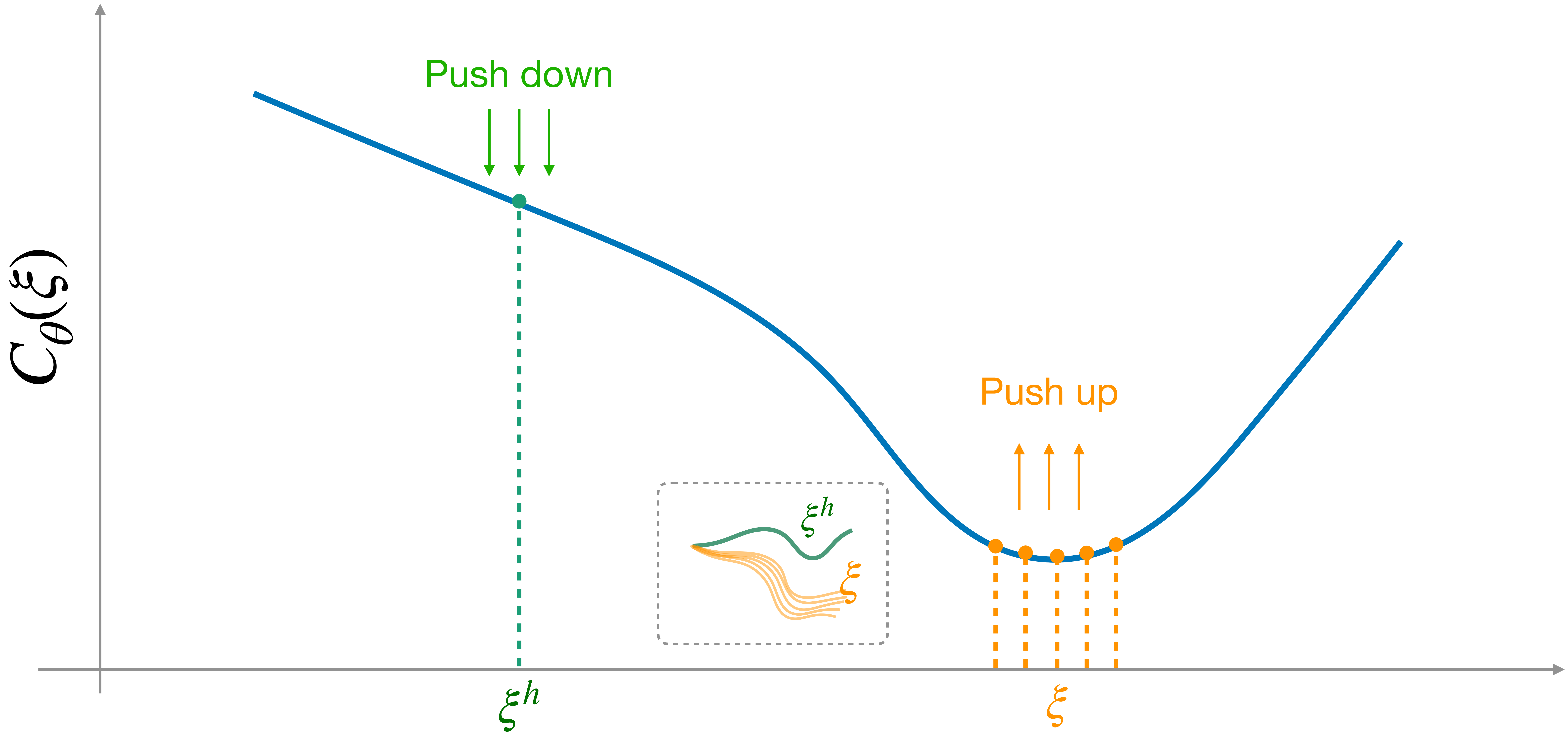
Animation from Sodhi et al. 2021



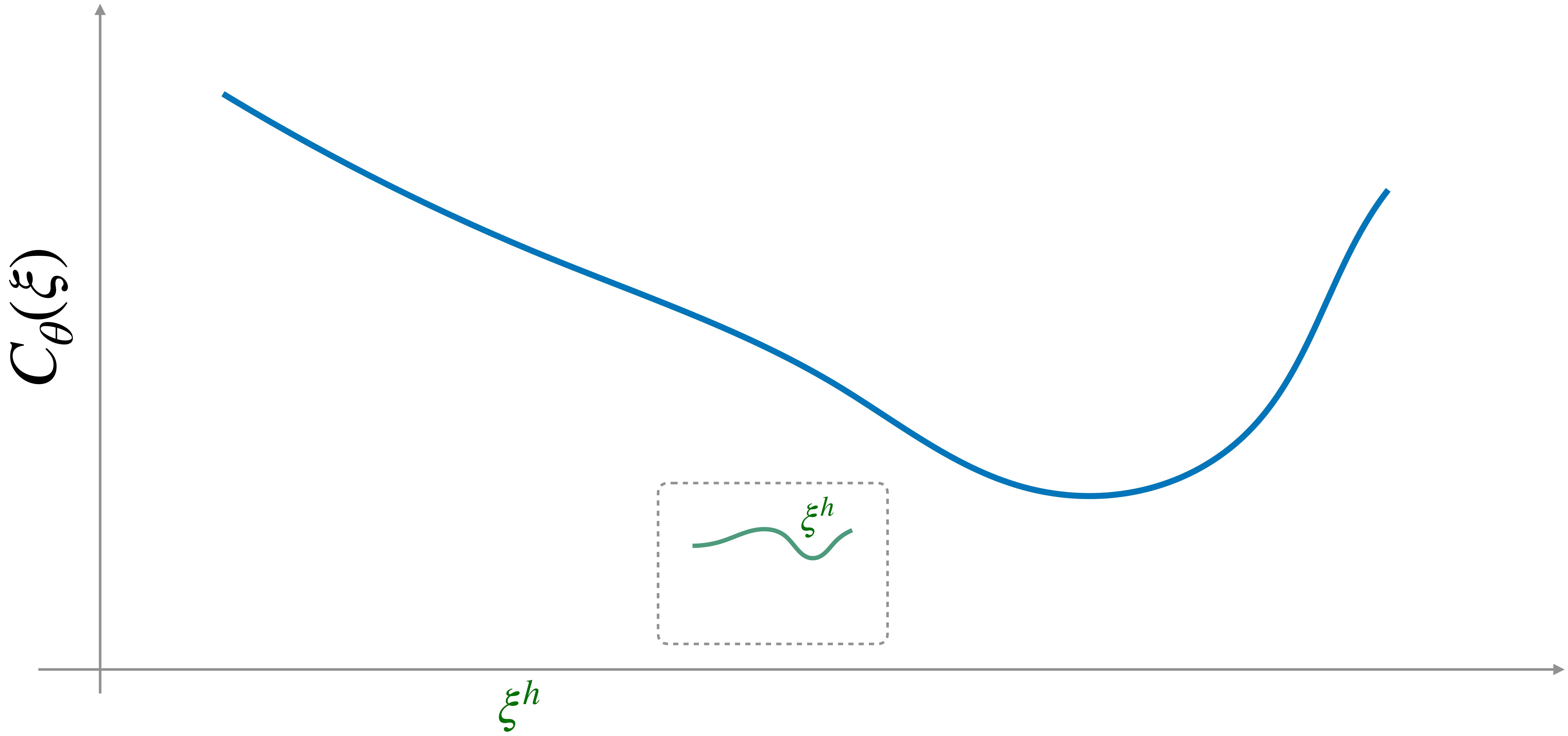
Animation from Sodhi et al. 2021



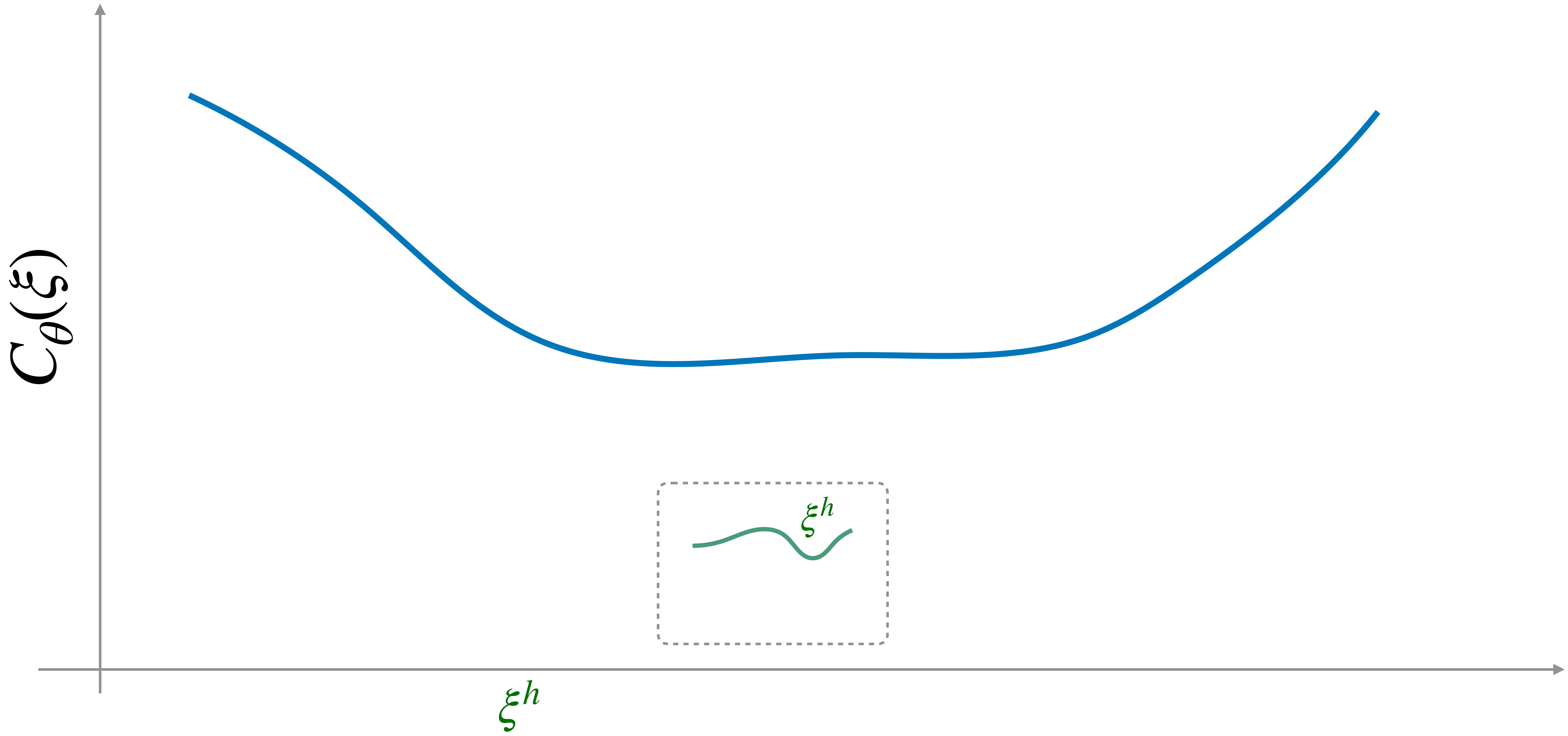
Animation from Sodhi et al. 2021



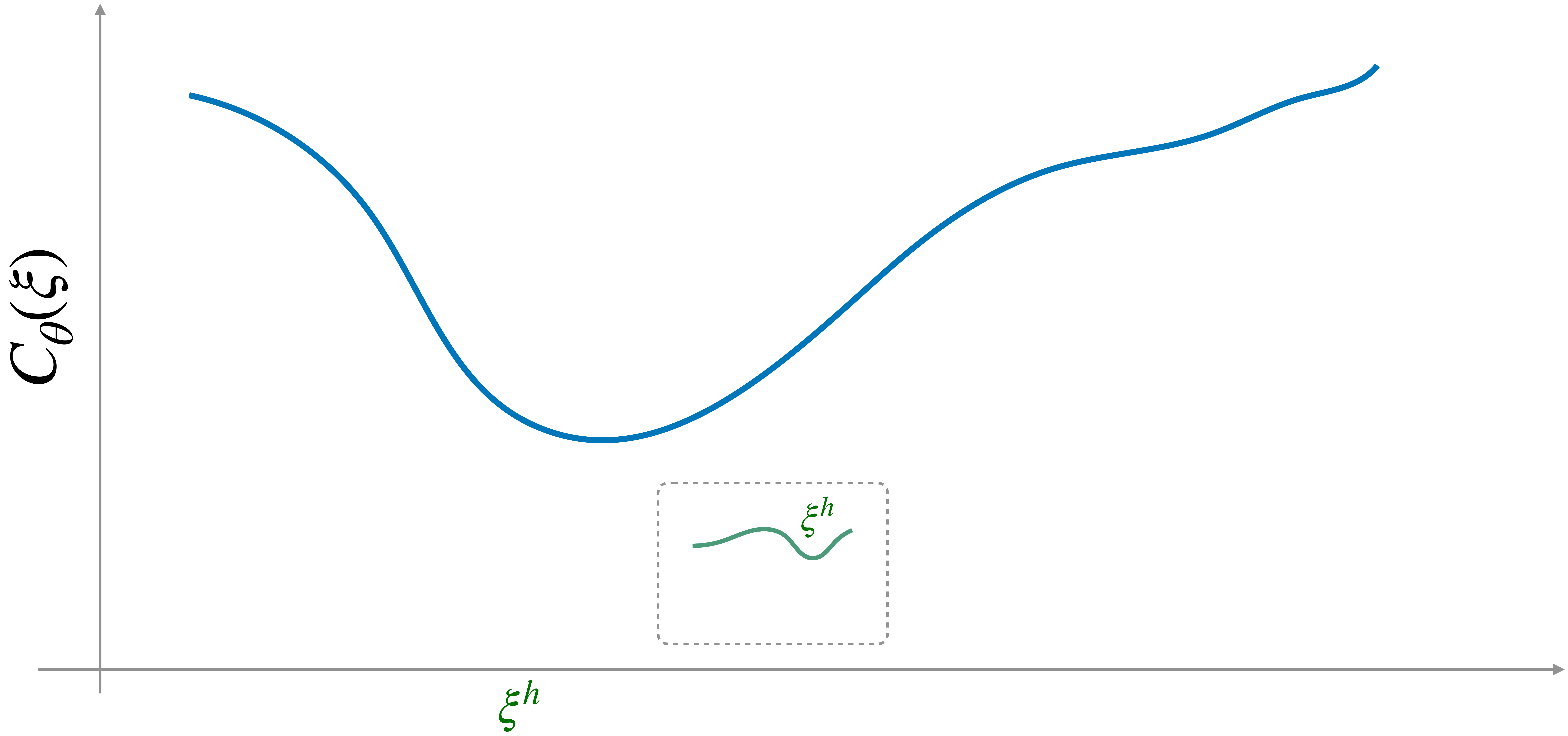
Animation from Sodhi et al. 2021



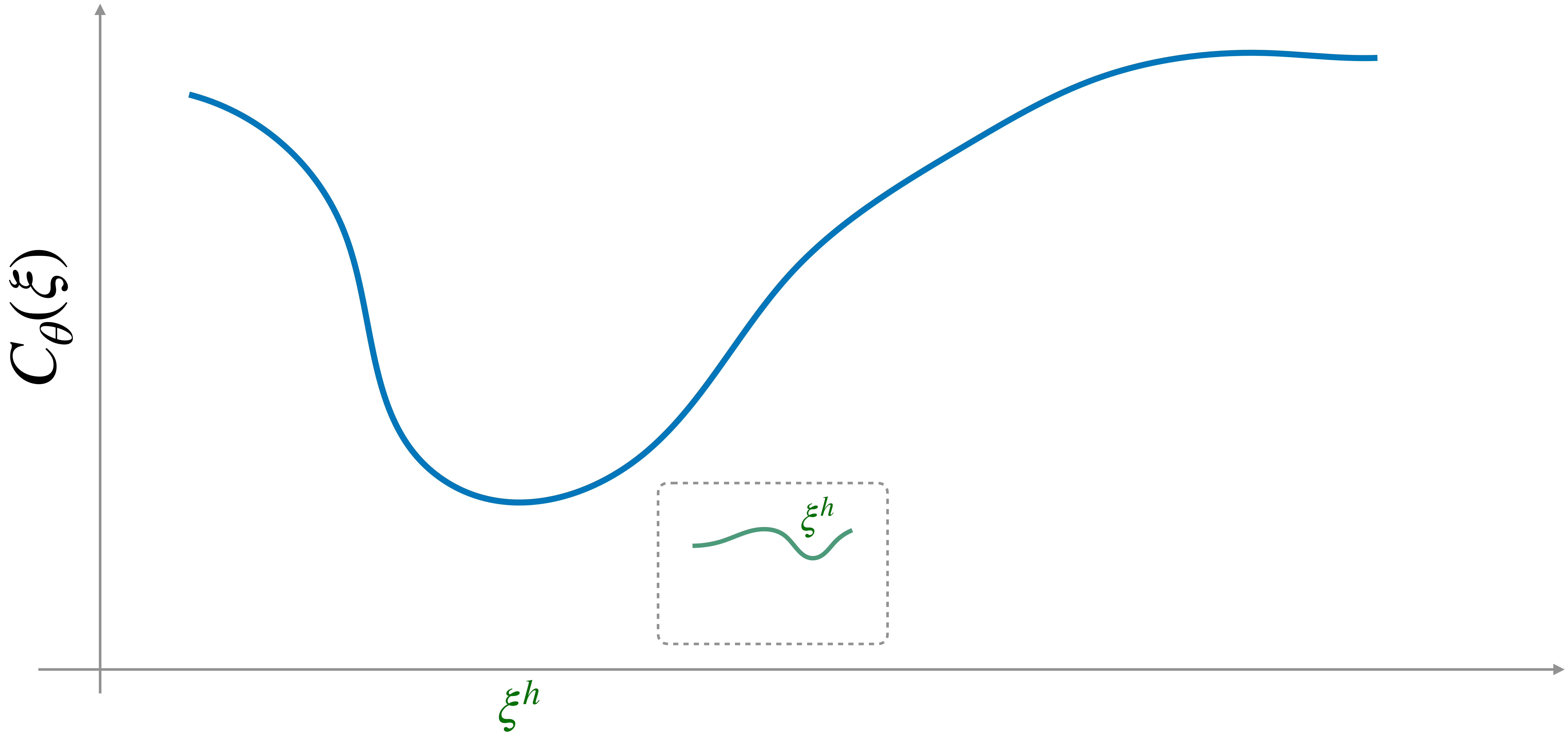
Animation from Sodhi et al. 2021



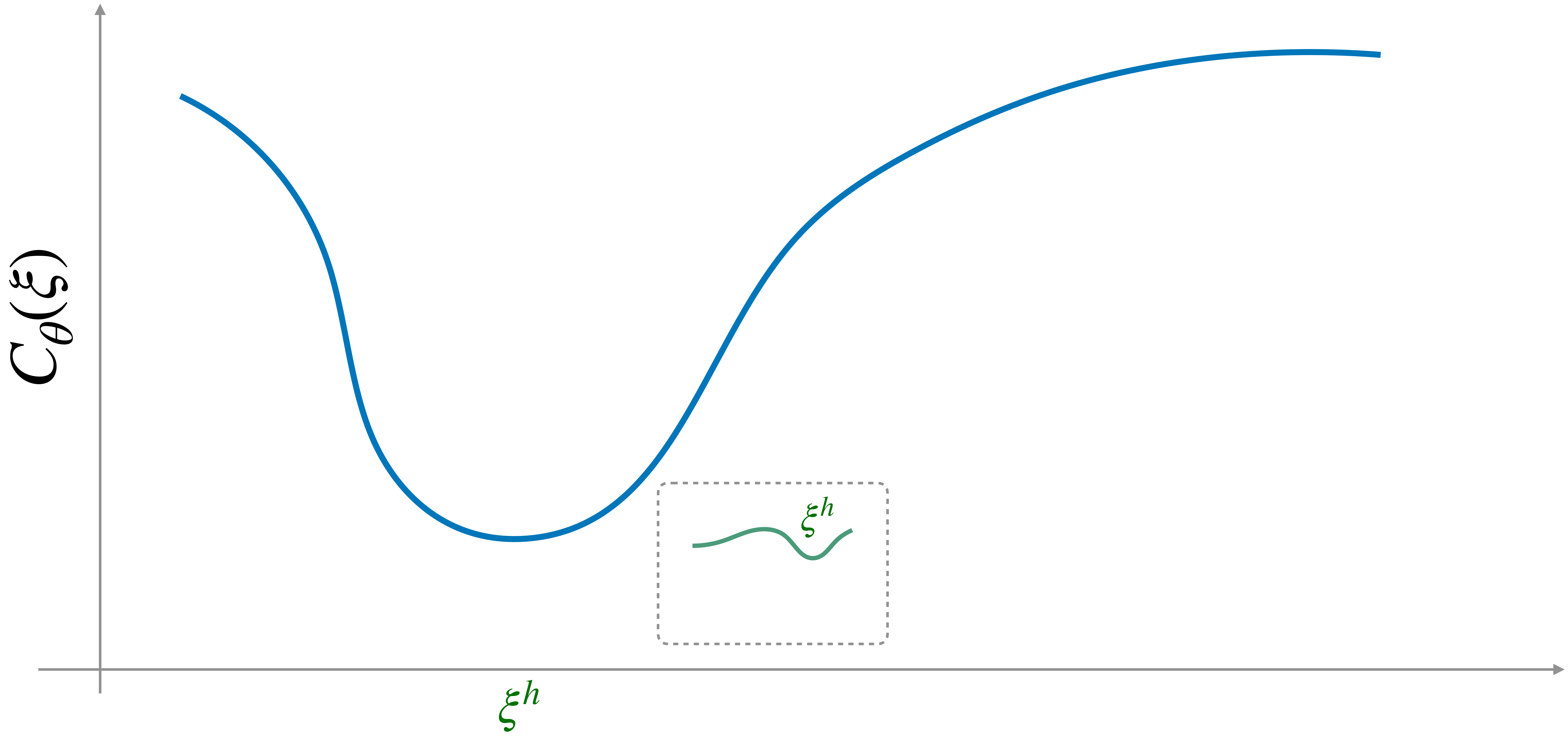
Animation from Sodhi et al. 2021



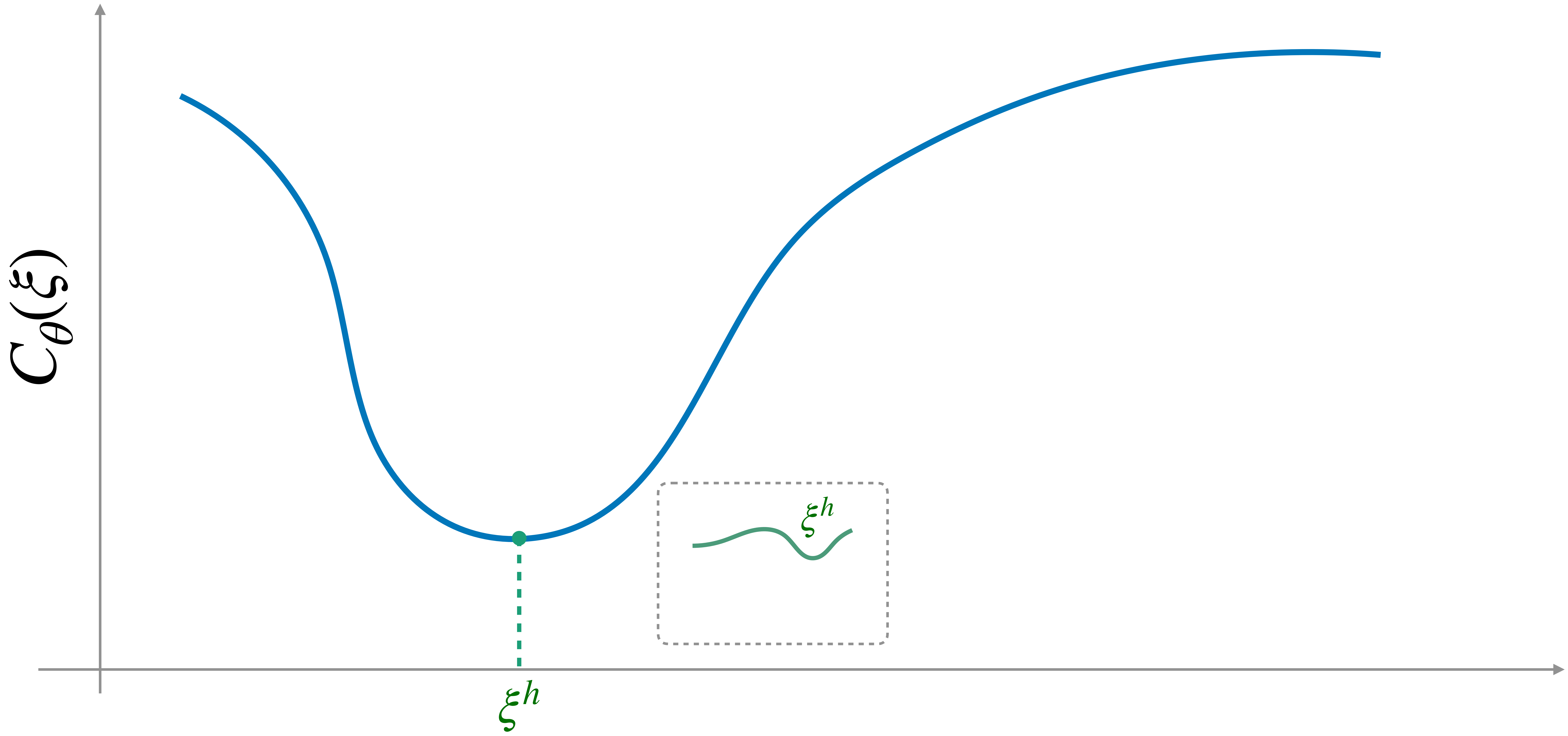
Animation from Sodhi et al. 2021



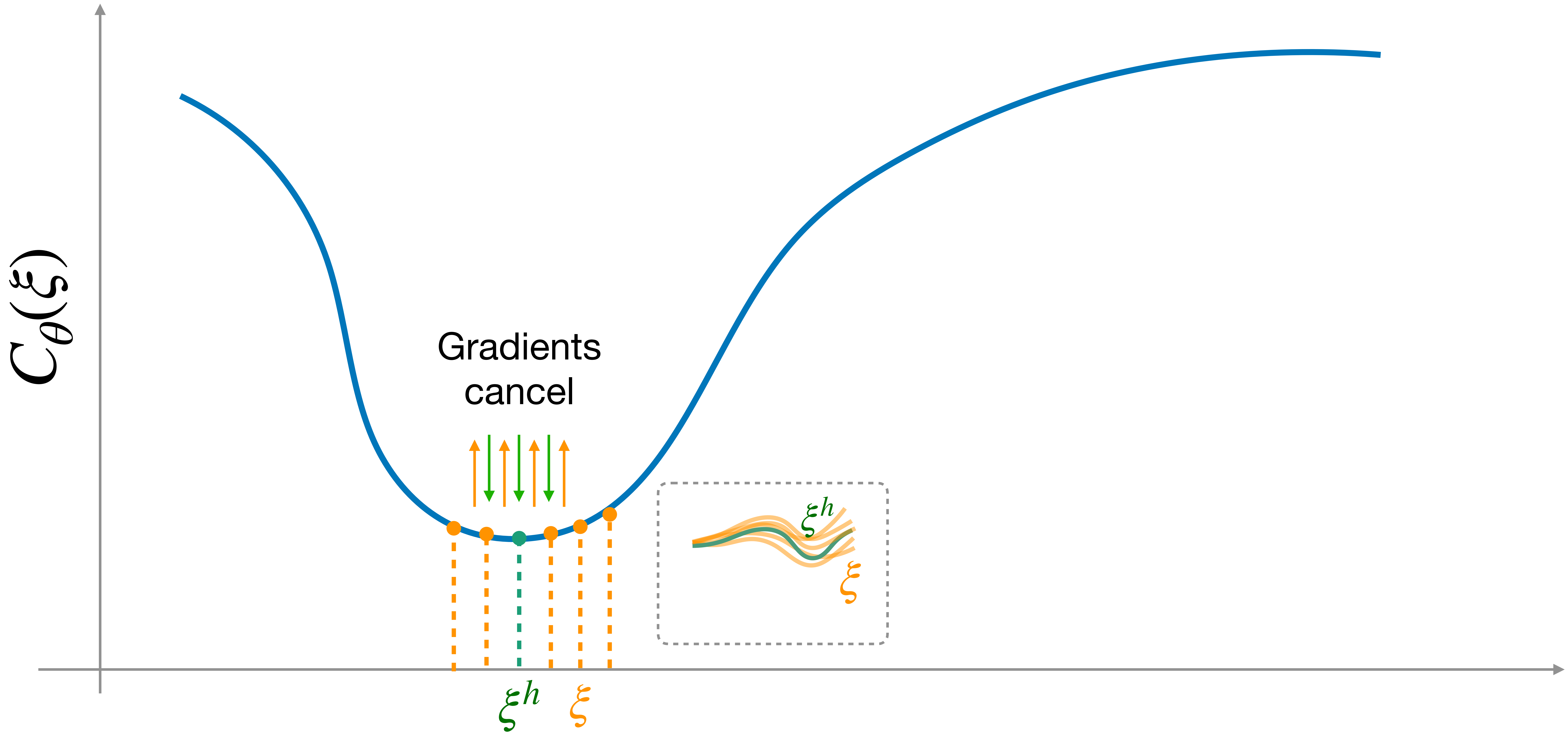
Animation from Sodhi et al. 2021



Animation from Sodhi et al. 2021



Animation from Sodhi et al. 2021



Okay...
But how do we sample from

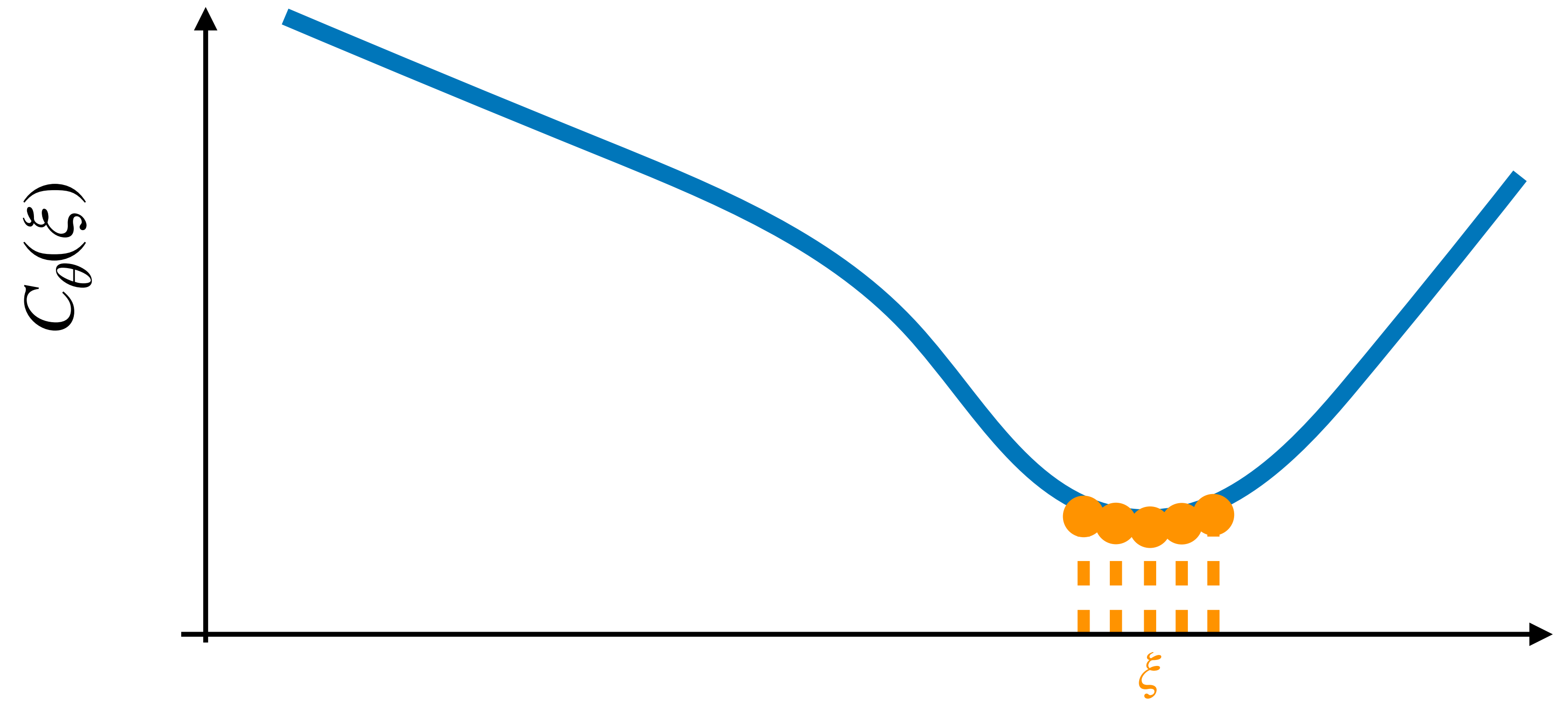
$$\xi \sim \frac{1}{Z} \exp\left(-C_\theta(\xi)\right)$$

The discrete case is easy!

Just call softmax()!

$$\xi \sim \frac{1}{Z} \exp\left(-C_\theta(\xi)\right)$$

What about a continuous trajectories?



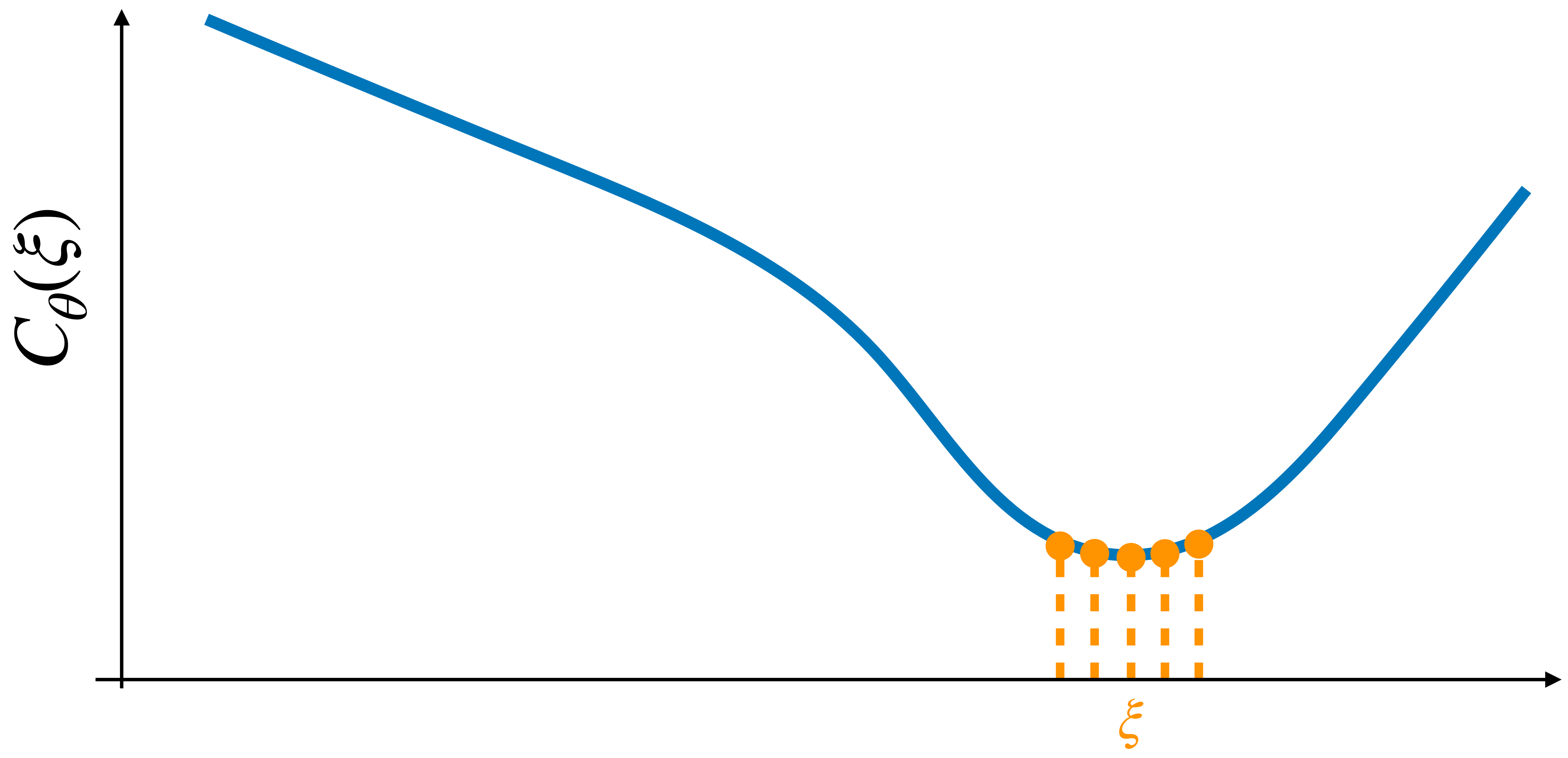
Activity!

Think-Pair-Share!

Think (30 sec): Let's say you had access to the (convex) function $C_\theta(\xi)$? How can we generate samples from exp(-$C_\theta(\xi)$)?

Pair: Find a partner

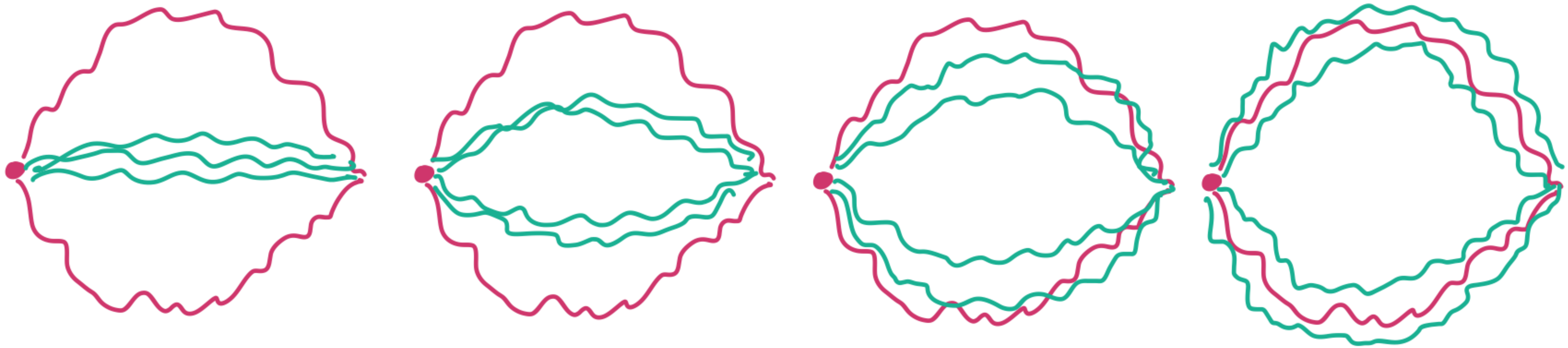
Share (45 sec): Partners exchange ideas



How can we use LQR / iLQR to sample from

$$\xi \sim \frac{1}{Z} \exp\left(-C_\theta(\xi)\right)?$$

MaxEnt with ILQR



for $i = 1, \ldots, N$ # Loop over datapoints

$$\xi_i \sim \frac{1}{Z} \exp\left(-C_\theta(\xi, \phi_i)\right)$$ # Call iLQR sampler

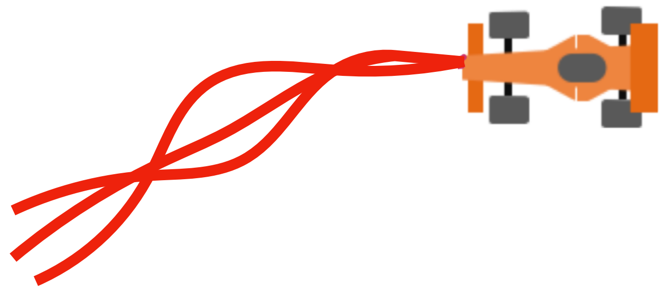
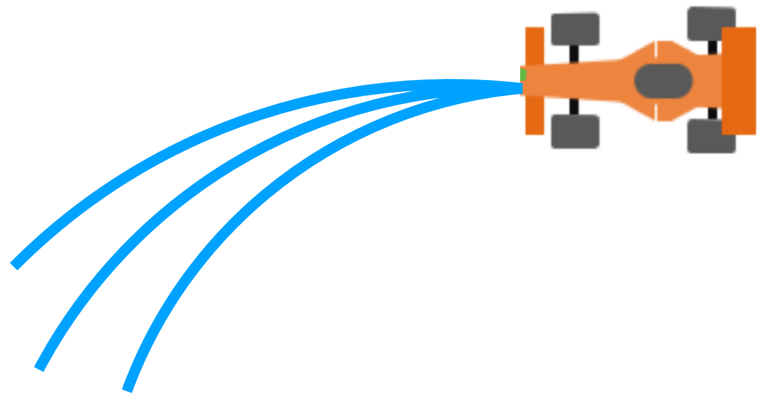
$$\theta^+ = \theta - \eta[\nabla_\theta C_\theta(\xi_i^h, \phi_i) - \nabla_\theta C_\theta(\xi_i, \phi_i)]$$ # Update cost

(Push down human cost) *(Push up planner cost)*

How do we “lift” MaxEntIOC from trajectories to policies?

The Entropy Regularized Game

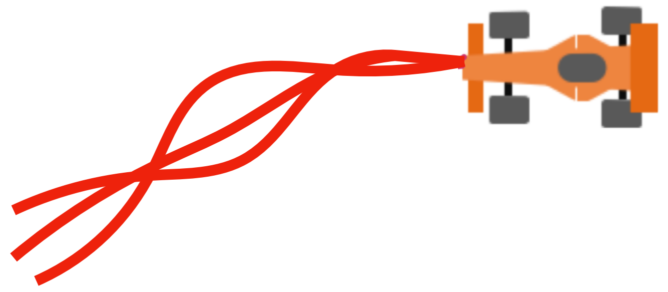
$$\max_{\phi}\min_{\theta}\ \mathbb{E}_{s_t,a_t\sim\pi_\theta}[C_\phi(s_t,a_t)]\ -\mathbb{E}_{s_t^*,a_t^*\sim\pi^*}[C_\phi(\xi)]\ -\beta H(\pi_\theta)$$

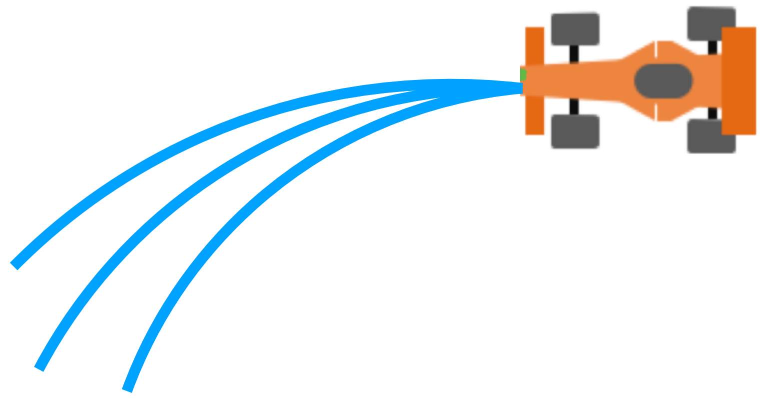
   

Entropy

The Entropy Regularized Game

$$\max_{\phi} \min_{\theta} \; \mathbb{E}_{s_t, a_t \sim \pi_\theta}[C_\phi(s_t, a_t)] \; - \mathbb{E}_{s_t^*, a_t^* \sim \pi^*}[C_\phi(\xi)] \; - \beta H(\pi_\theta)$$





Entropy

for $i = 1, \ldots, N$ # Loop over episodes

$$\pi_\theta = \arg\min_{\pi} \mathbb{E}_{s_t, a_t \sim \pi}[C_\phi(s_t, a_t)] - \beta H(\pi)$$ # Soft Actor Critic

$$\phi^+ = \phi + \eta[\nabla_\theta \mathbb{E}_{s_t, a_t \sim \pi_\theta}[C_\phi(s_t, a_t)] - \nabla_\theta \mathbb{E}_{s_t^*, a_t^* \sim \pi^*}[C_\phi(\xi)]]$$

Update cost

Two Core Ideas

Data

"What is the distribution of states?"

Loss

"What is the metric to match to human?"

Two Core Ideas

Data

"What is the distribution of states?"

Loss

"What is the metric to match to human?"

Easy | **Medium** | **Hard**

Setting

Expert is realizable $\pi^E \in \Pi$

Non-realizable expert but full expert support

Non-realizable expert + limited expert support

As $N \to \infty$, drive down $\epsilon = 0$ (or Bayes error)

Even as $N \to \infty$, behavior cloning $O(\epsilon C T)$ where C is conc. coeff

Even as $N \to \infty$, behavior cloning $O(\epsilon T^2)$

Just Behavior Cloning

Interactive Simulator

Interactive Expert

Solution

Nothing special. Collect lots of data and do Behavior Cloning

Requires interactive simulator (MaxEntIRL) to match distribution $\Rightarrow O(\epsilon T)$

Requires interactive expert (DAGGER / EIL) to provide labels $\Rightarrow O(\epsilon T)$

Two Core Ideas

Data

"What is the distribution of states?"

Loss

"What is the metric to match to human?"

Hints of a Big Picture ....

What we really want to solve is:

$$\min_{\pi} J(\pi) - J(\pi^*)$$

Loss

"What is the metric to match to human?"

Hints of a Big Picture ....

What we get from PDL:

$$\min_{\pi} \mathbb{E}_{s\sim d_\pi}[Q^*(s,\pi(s)) - Q^*(s,\pi^*(s))]$$

Loss

✓ *"What is the metric to match to human?"*

Difference in Q values!

Hints of a Big Picture ….

What we really want to solve is:

$$\min_{\pi} \mathbb{E}_{s \sim d_\pi}[Q^*(s, \pi(s)) - Q^*(s, \pi^*(s))]$$

$$Q^*(s, a)$$

Loss

"What is the metric to match to human?"

Difference in Q values!

But Q* is latent!

Hints of a Big Picture ....

Estimate Q* from demonstrations, interventions, preferences, ..
and even E-stops!

Demonstrations

Interventions

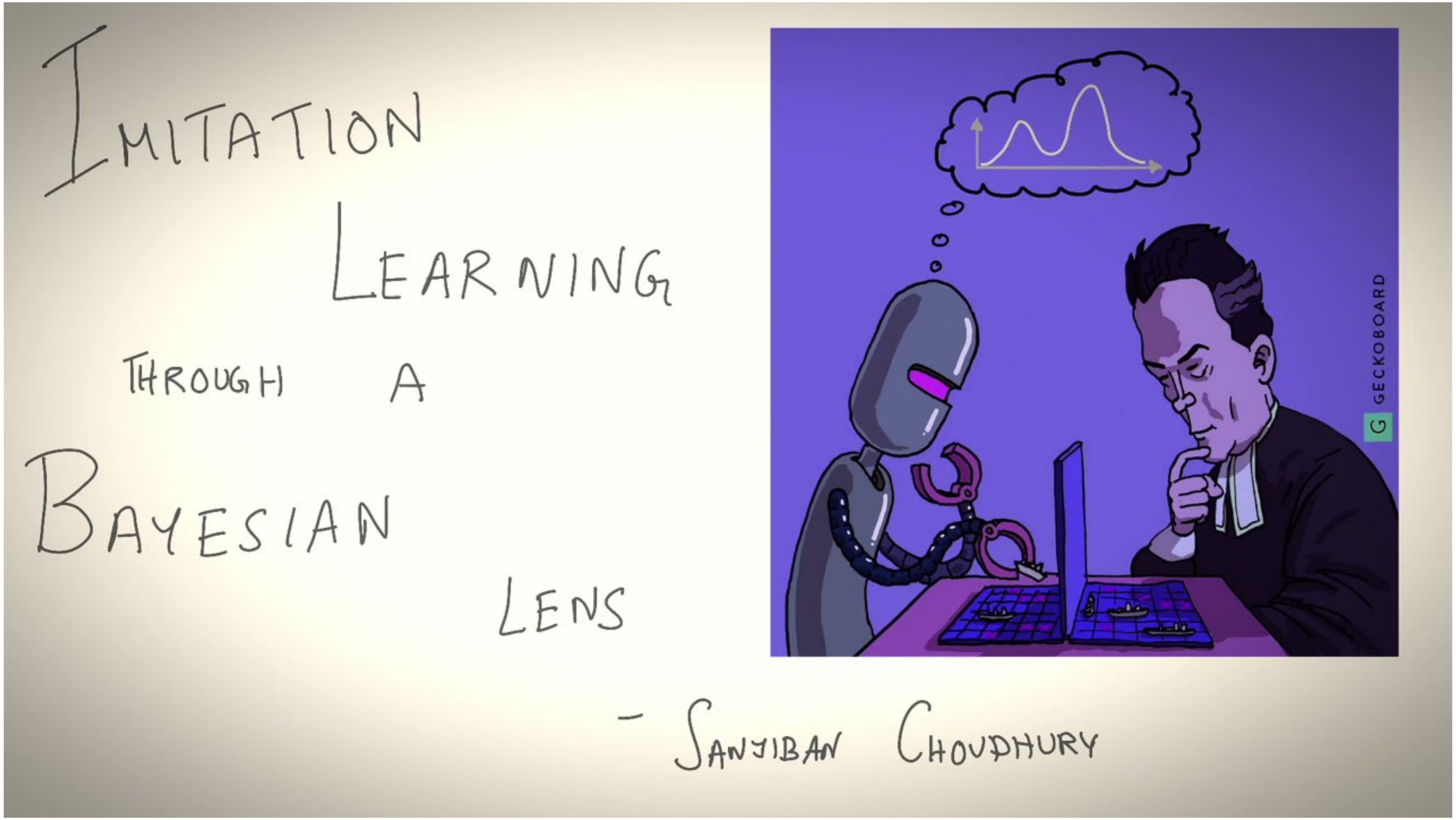
Preferences

E-stops

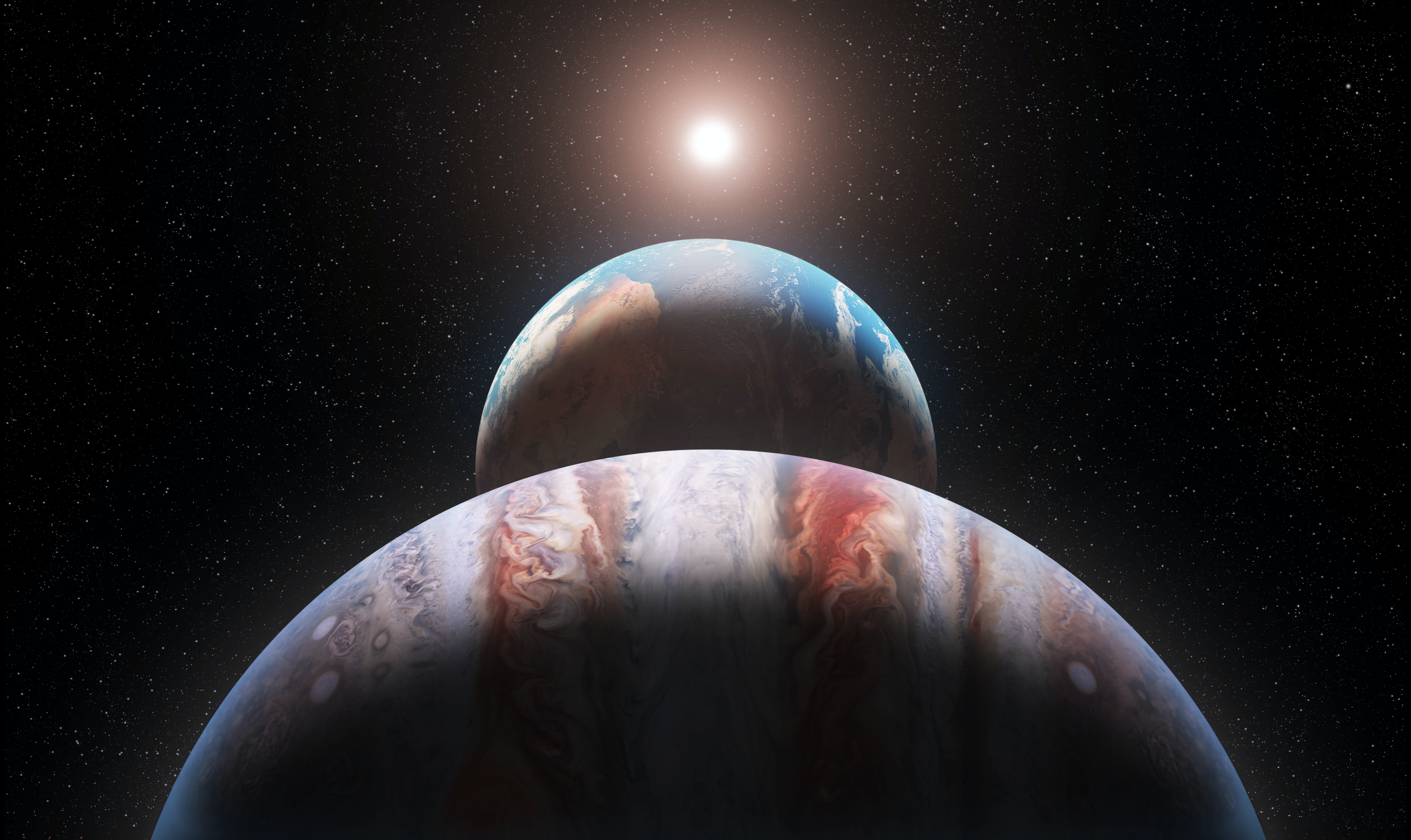
$$\mathscr{L}(Q^*_\theta)$$

Loss

Imitation Learning from a Bayesian Lens



The BIG Picture!




The Road Ahead!

Active Imitation Learning

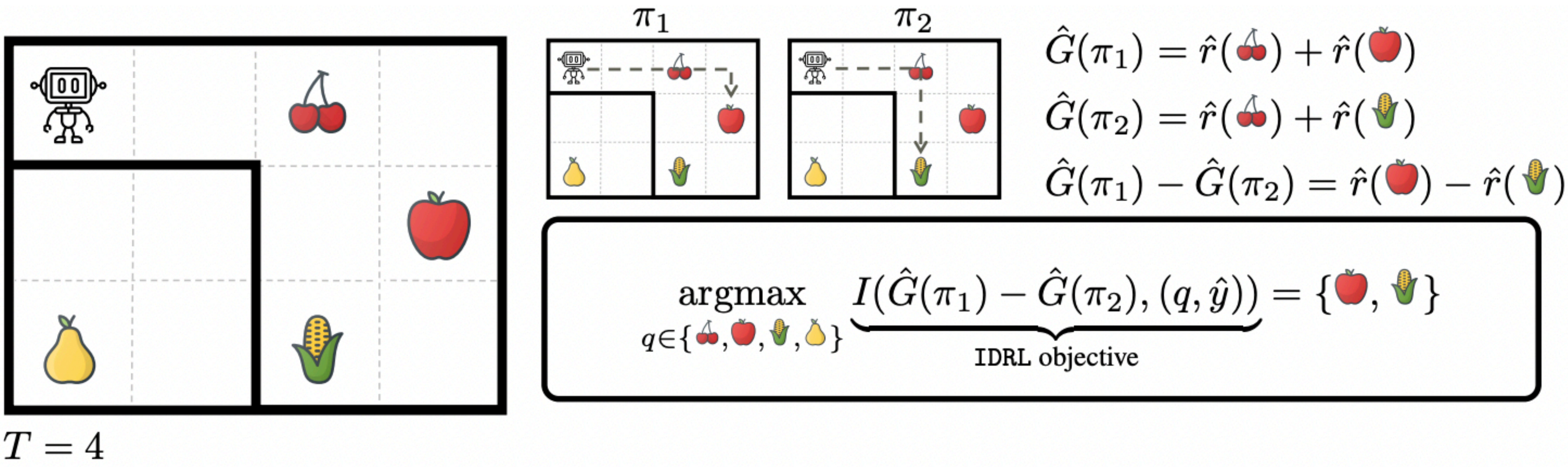
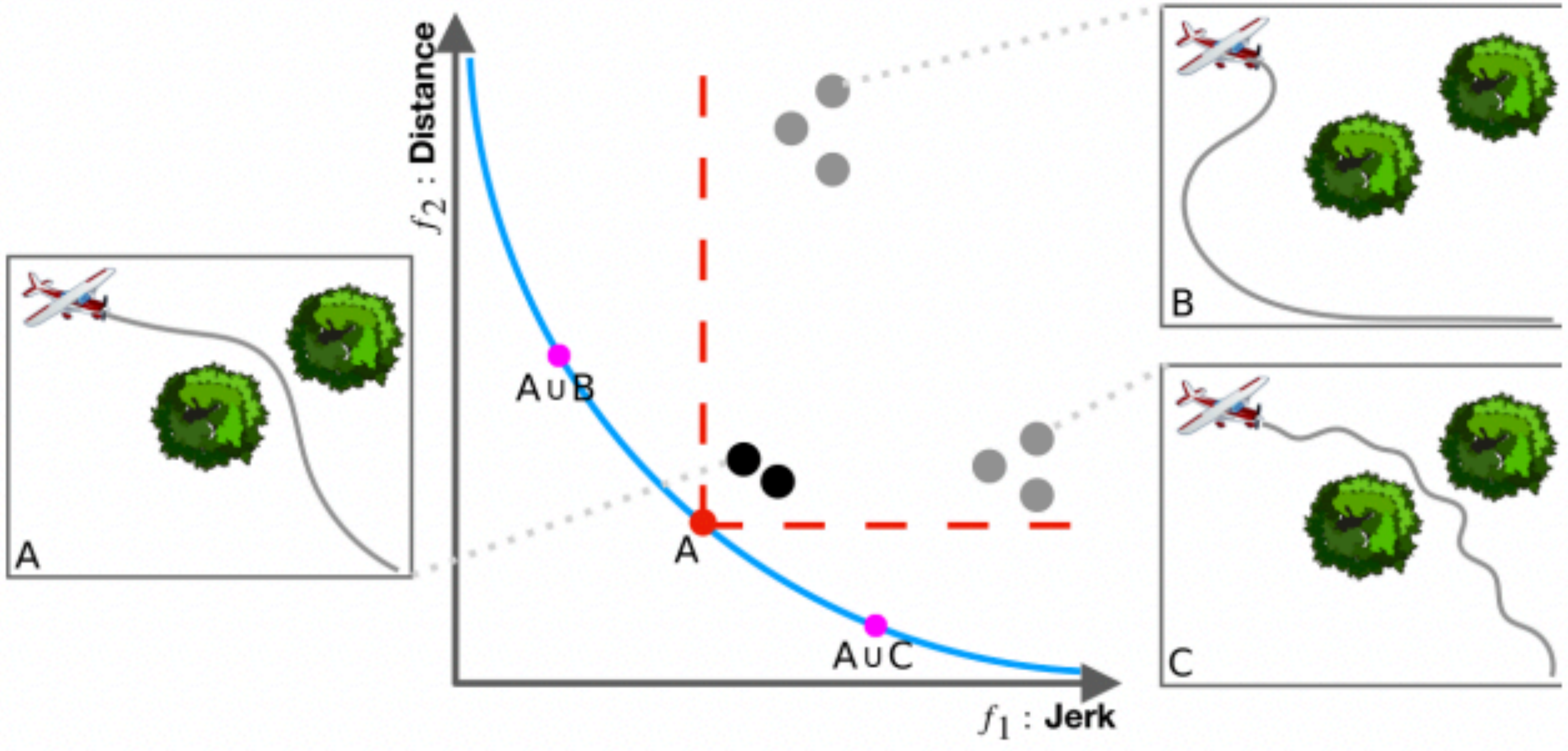


Figure 1: The robot wants to collect food for a human. It can only move 4 timesteps in the gridworld, cannot pass through the black walls, and collecting more food is always better. The robot does not know the human's preferences, but it can ask for food ratings. Common active learning methods aim to learn the reward uniformly well, and would query all items similarly often. In contrast, IDRL considers only the two plausibly optimal policies π_1 and π_2. Since both policies collect the cherry, and do not collect the pear, the robot only needs to learn about the apple and the corn. IDRL can solve the task with 2 queries instead of 4.

Lindner, D., Turchetta, M., Tschiatschek, S., Ciosek, K., & Krause, A. (2021). Information Directed Reward Learning for Reinforcement Learning. Advances in Neural Information Processing Systems, 34, 3850-3862.

Learning from Suboptimal Experts

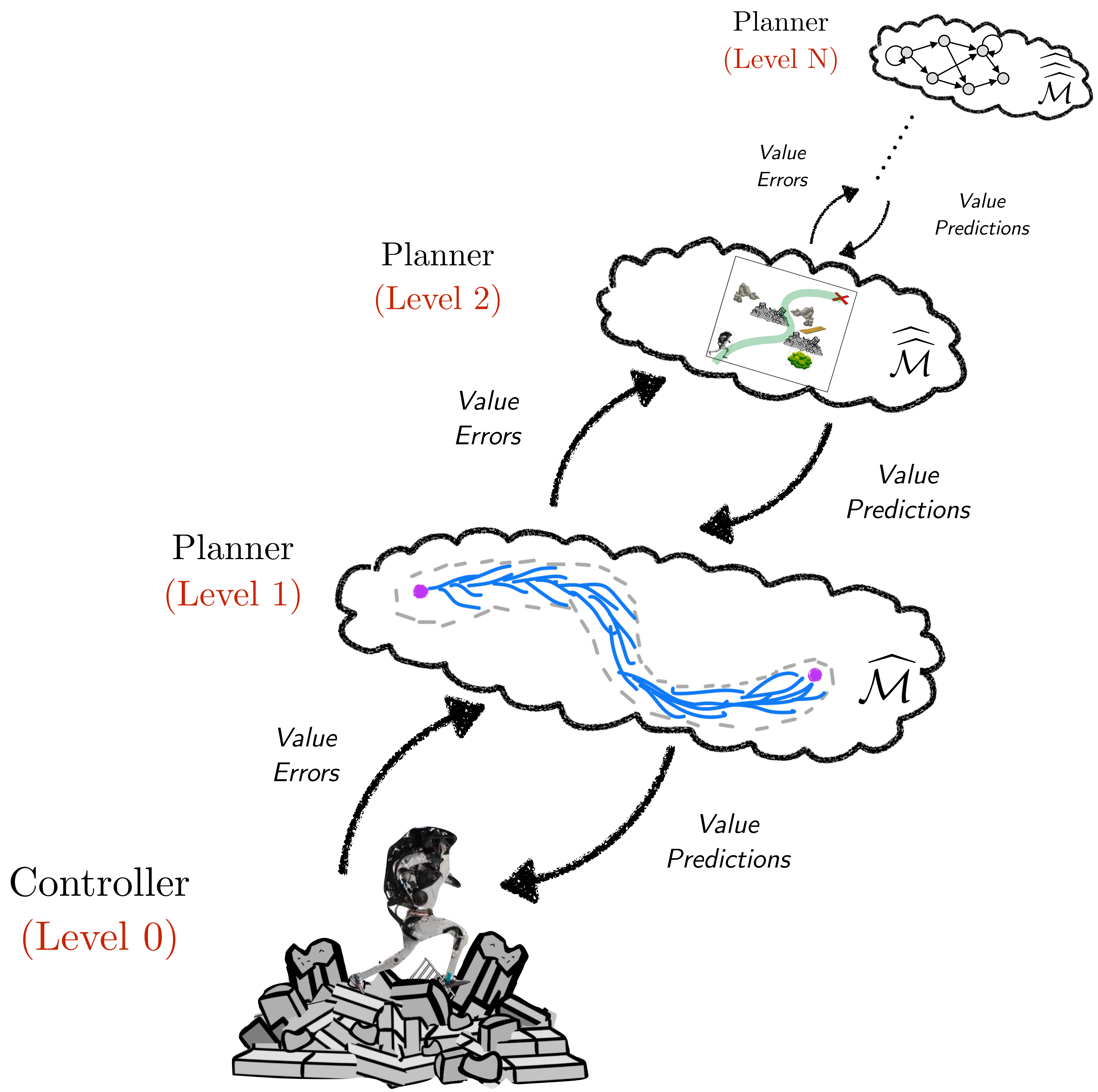


Learn policies that *outperform* expert for any choice of cost function

Towards Uniformly Superhuman Autonomy via Subdominance Minimization

Brian D. Ziebart [1] Sanjiban Choudhury [2] Xinyan (Shane) Yan [2] Paul Vernaza [2]

Hierarchical Imitation Learning



Inverse Optimal Heuristic Control for Imitation Learning

Nathan Ratliff, Brian Ziebart, Kevin Peterson, J. Andrew Bagnell, Martial Hebert, Anind K. Dey
Robotics Institute, MLD, CSD
Carnegie Mellon University
Pittsburgh, PA 15213

Siddhartha Srinivasa
Intel Research
Pittsburgh, PA 15213

Hierarchical Imitation and Reinforcement Learning

Hoang M. Le[1] **Nan Jiang**[2] **Alekh Agarwal**[2] **Miroslav Dudík**[2] **Yisong Yue**[1] **Hal Daumé III**[3 2]