Principle of Maximum Entropy in Decision Making (From IRL to RL and back)

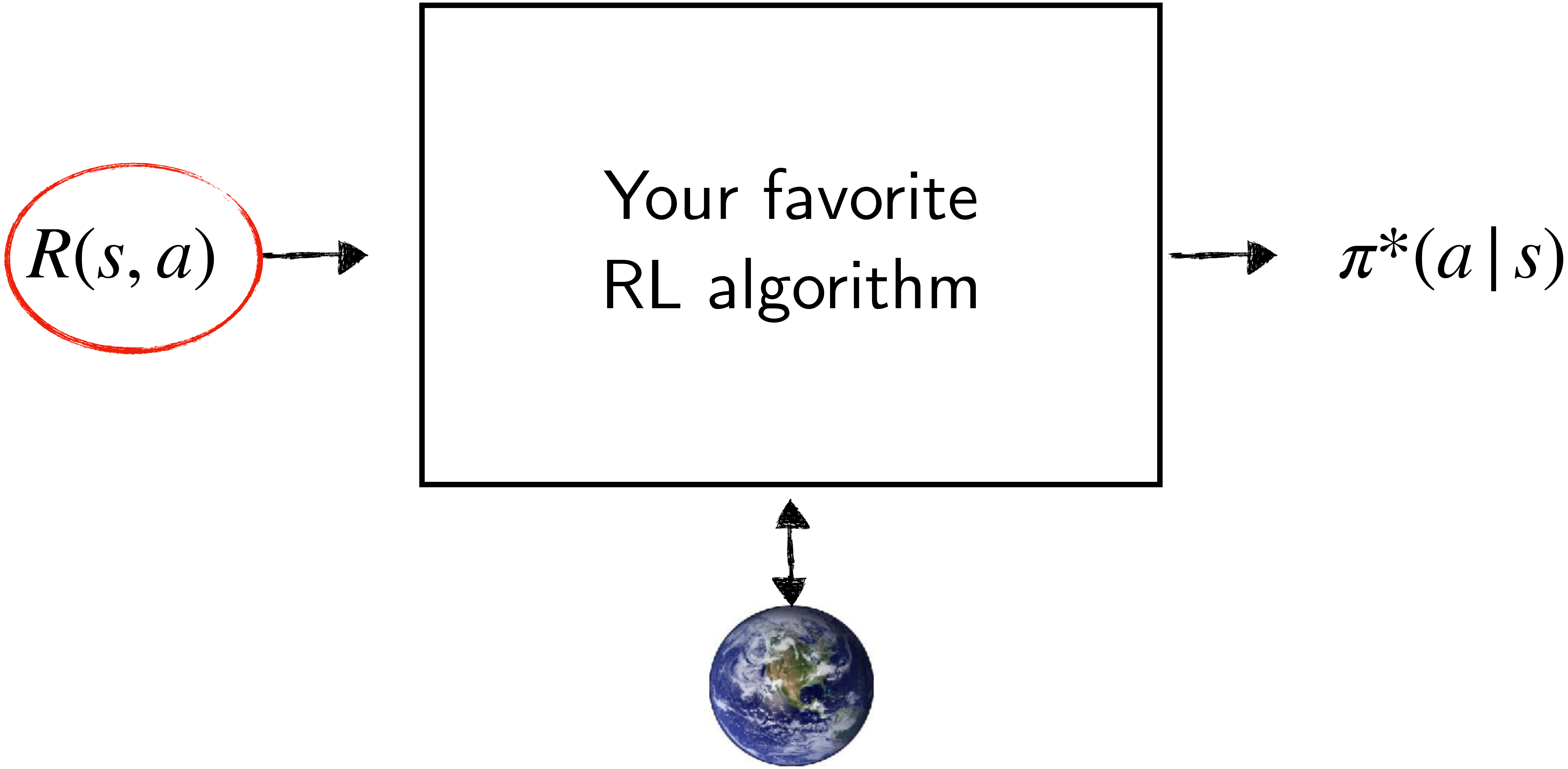
Sanjiban Choudhury

Cornell Bowers CIS
Computer Science

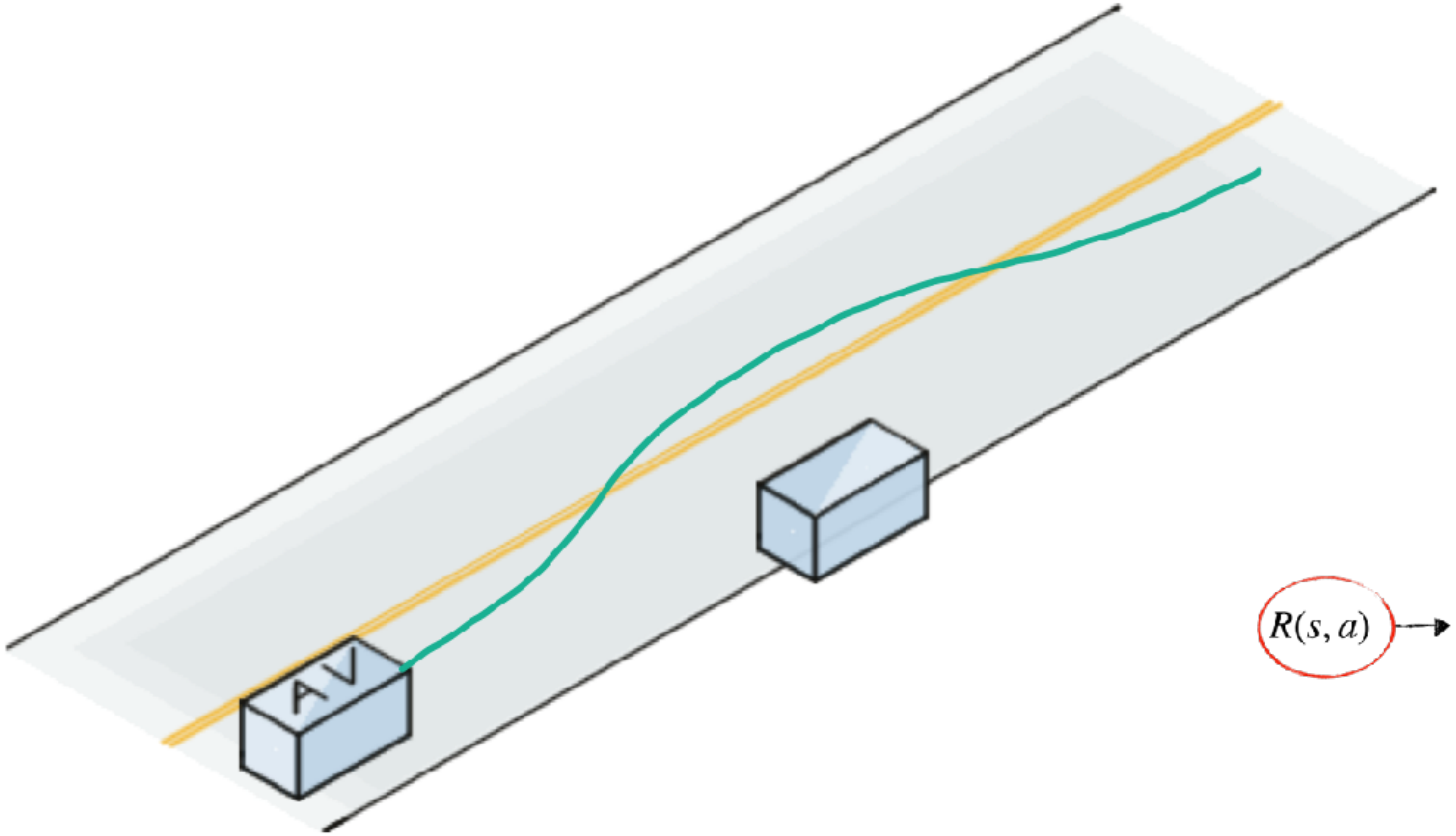
We know how to make a RL block!

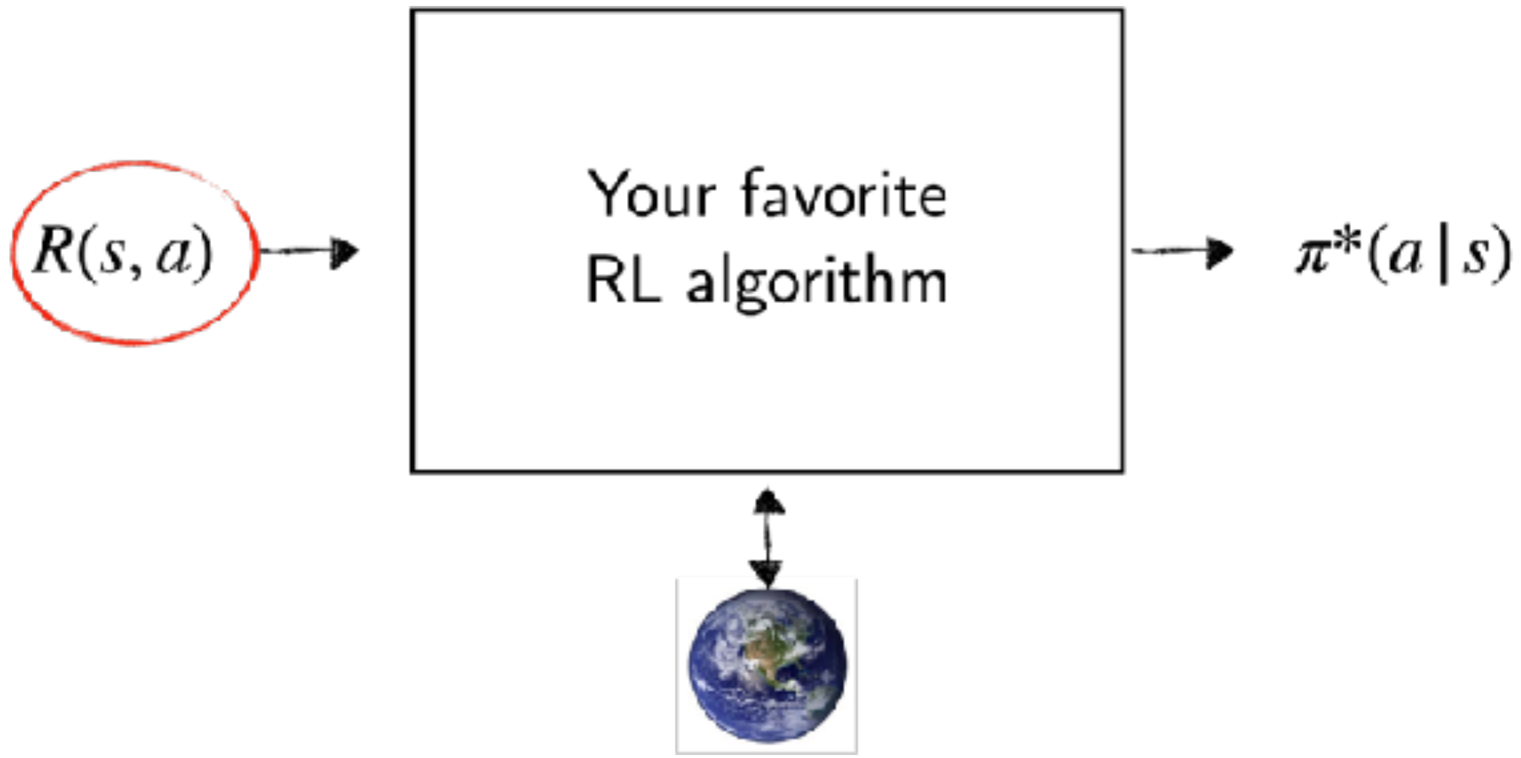
$R(s, a)$ → Your favorite RL algorithm → $\pi^*(a \mid s)$

But how do we design reward function??



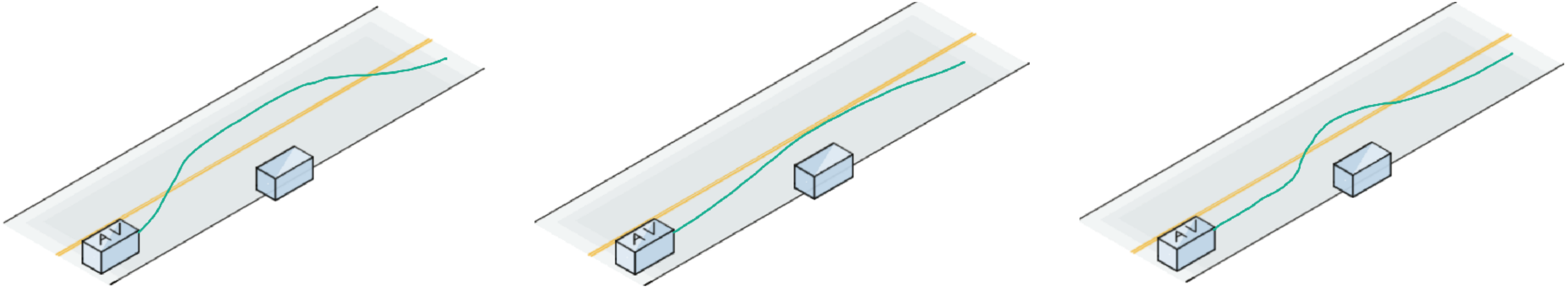
Designing R(s,a) for self-driving





Let's say we want a reward function that matches human like driving

But humans have a lot of variance in their motion!



Is there a reward function for which all these motions are optimal?

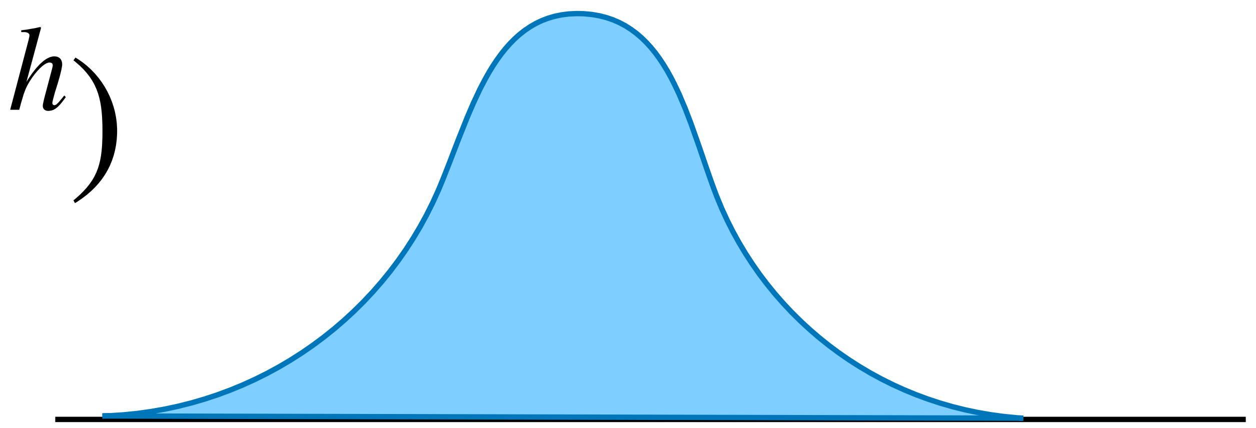
How do we imitate "real experts" who may be noisy / suboptimal?

Expert demonstrations are coming from some (unknown) distribution ..

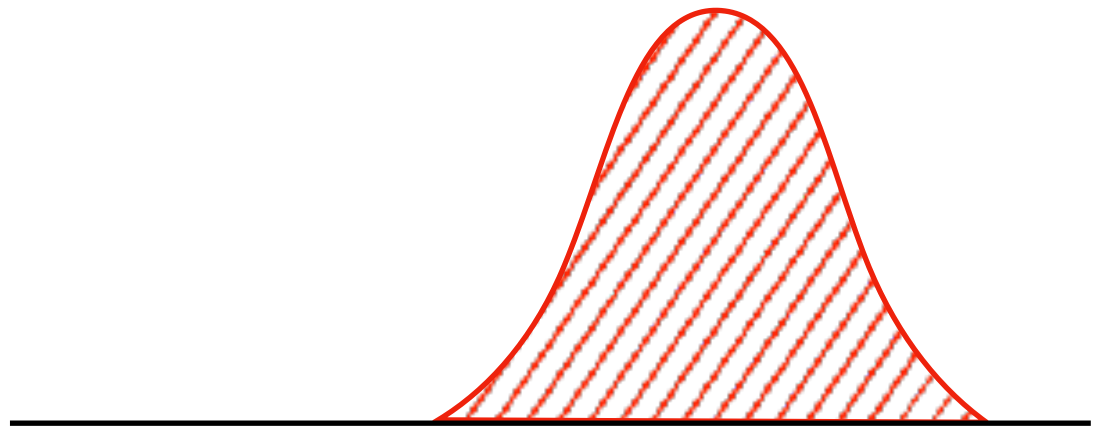
Can we learn this distribution?

The Distribution Matching Problem

$P_{expert}(\xi^h)$



(Unknown) expert distribution

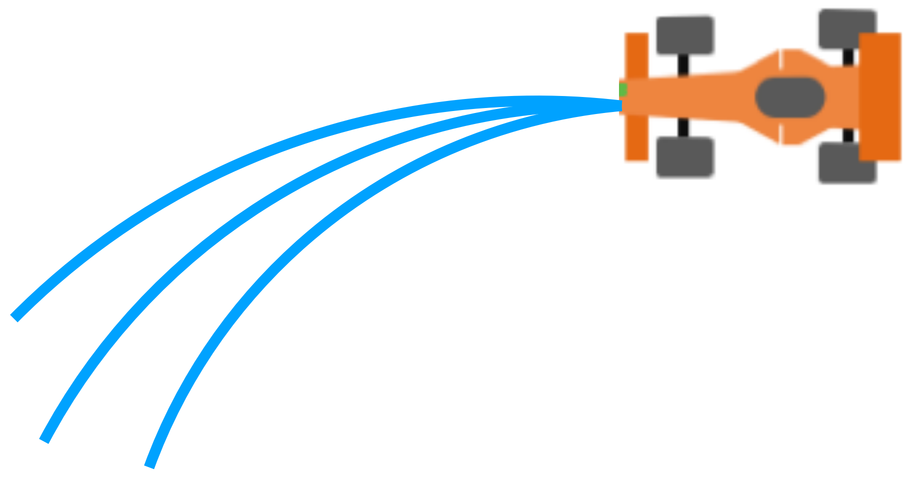


$P_\theta(\xi)$

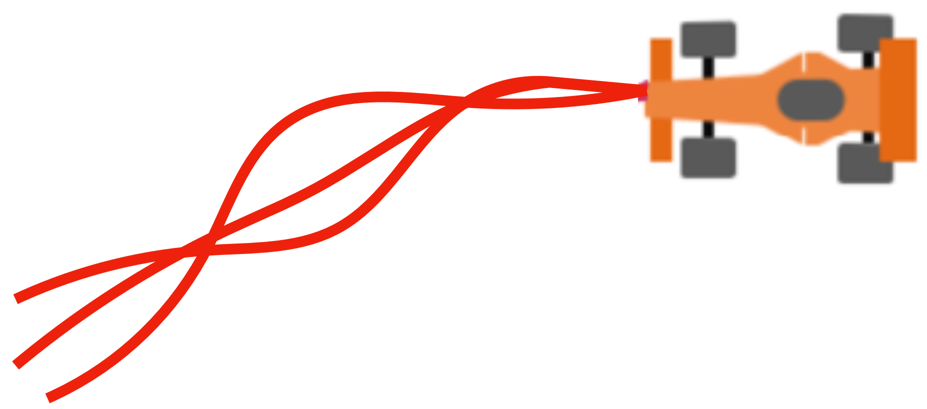
Learn distribution over trajectories

All we see are expert samples

Learner can also generate samples



What loss should we use?



What loss should we use?

What we actually care about is matching Performance Difference

$$J(\pi) = J(\pi^*)$$

$$\mathbb{E}_{\xi \sim P_\theta(\xi)} c(\xi) = \mathbb{E}_{\xi \sim P_{expert}(\xi)} c(\xi)$$

But we don't know the costs c(.)!!

What divergence do we care about?

What we actually care about is matching Performance Difference

$$J(\pi) = J(\pi^*)$$

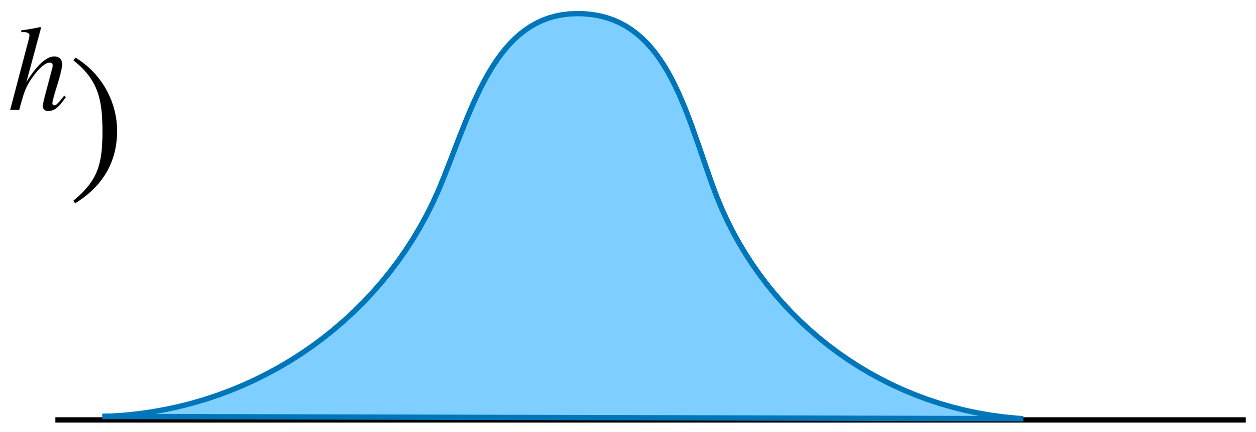
$$\mathbb{E}_{\xi \sim P_\theta(\xi)} c(\xi) = \mathbb{E}_{\xi \sim P_{expert}(\xi)} c(\xi)$$

But we don't know the costs c(.)

Costs are just weighted combination of features. What if we just matched all the expected features?

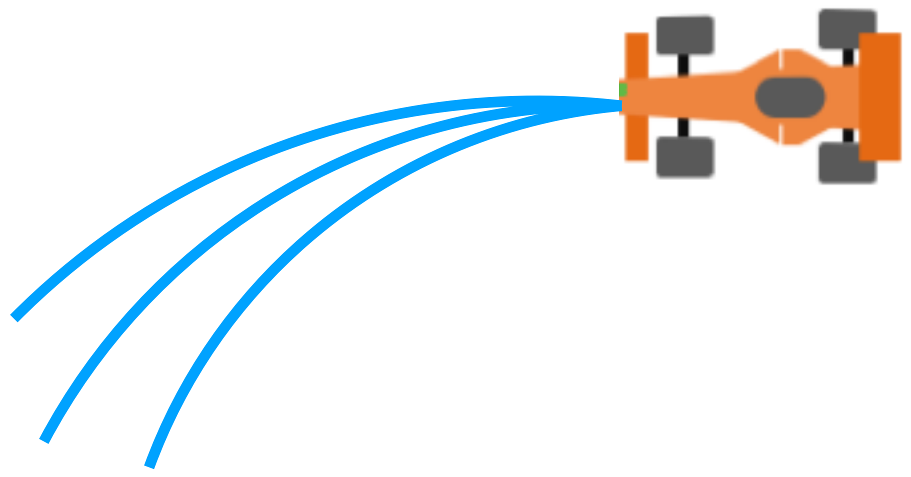
Proposal: Match cost features!

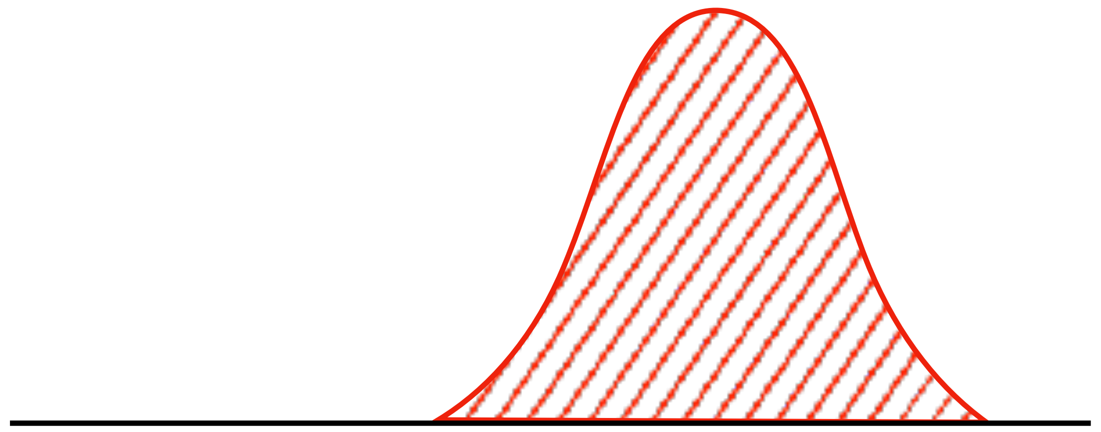
$P_{expert}(\xi^h)$



(Unknown) expert distribution

All we see are expert samples

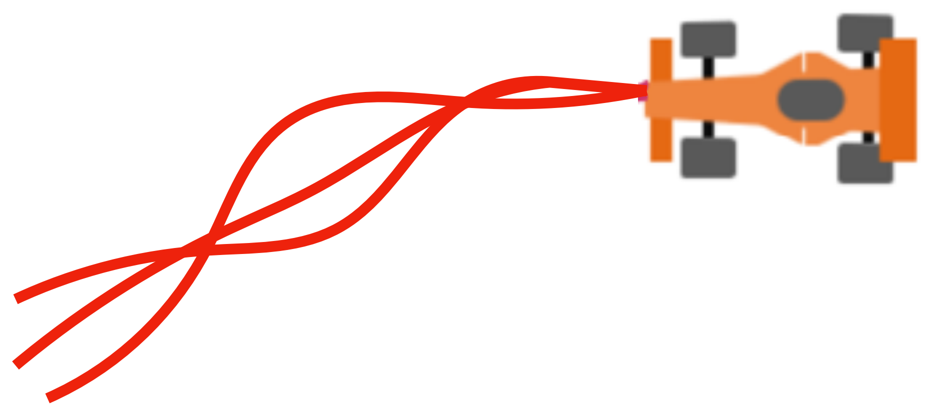




$P_\theta(\xi)$

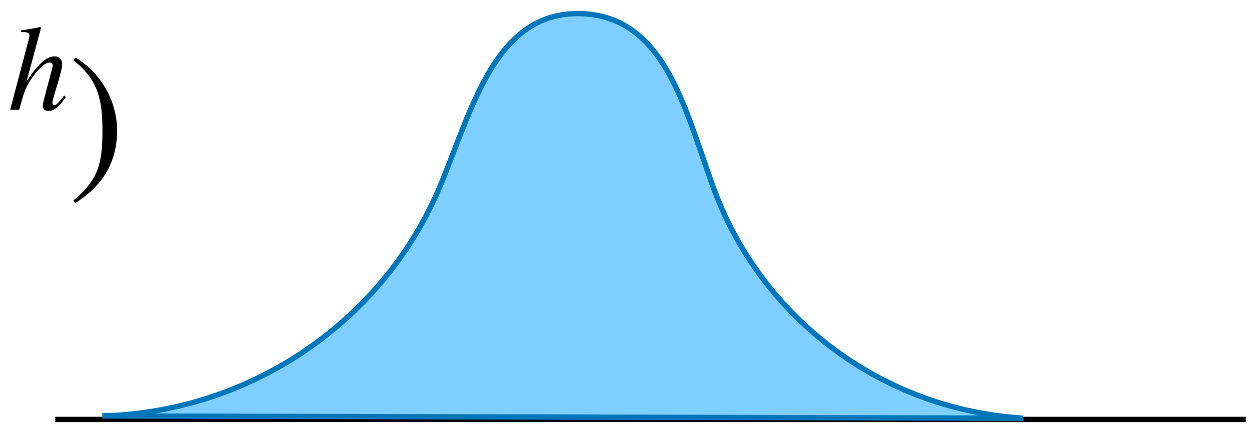
Learn distribution over trajectories

Learner can also generate samples

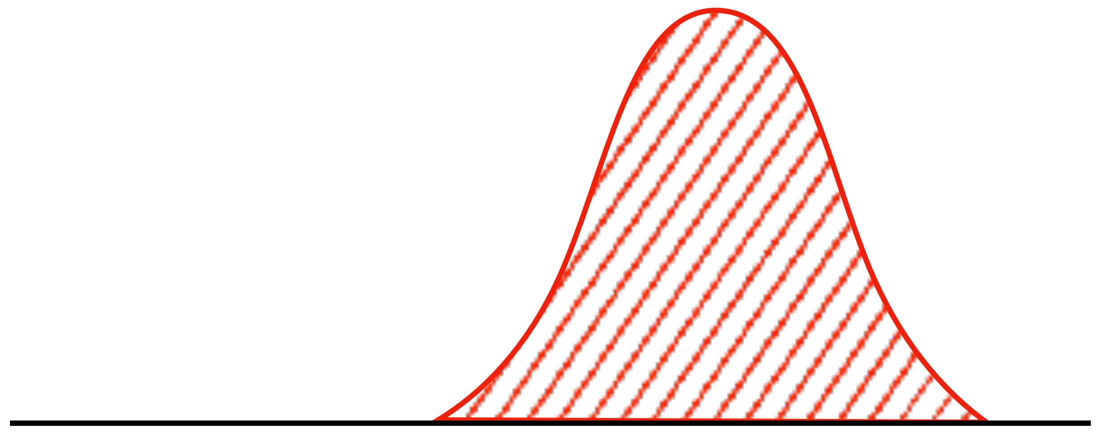


Proposal: Match cost features!

$P_{expert}(\xi^h)$



(Unknown) expert distribution

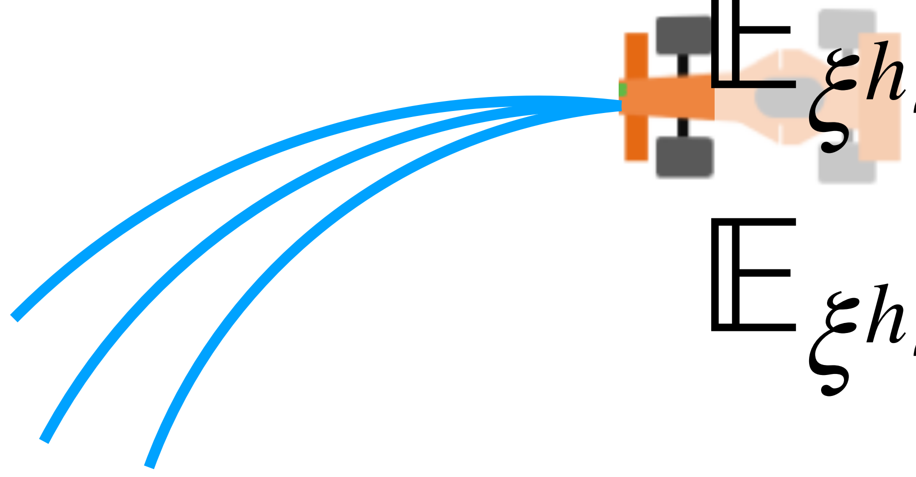


$P_\theta(\xi)$

Learn distribution over trajectories

All we see are expert samples

Learner can also generate samples

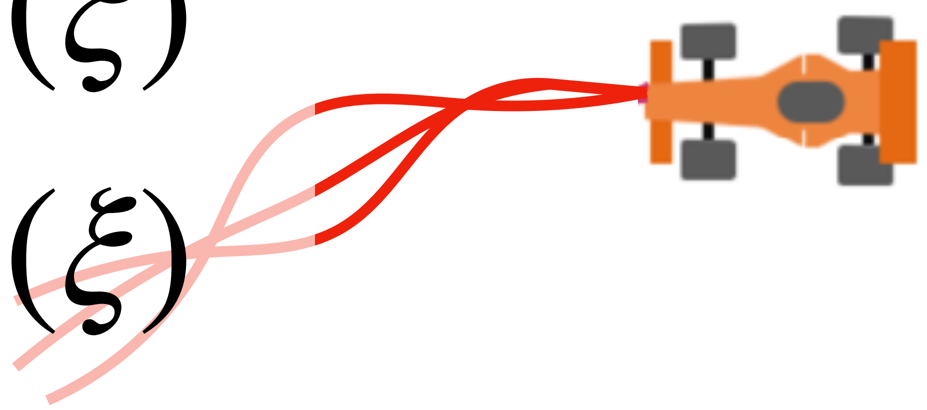


$$\mathbb{E}_{\xi^h \sim P_{expert}(.)} f_1(\xi^h) = \mathbb{E}_{\xi \sim P_\theta(.)} f_1(\xi)$$

$$\mathbb{E}_{\xi^h \sim P_{expert}(.)} f_2(\xi^h) = \mathbb{E}_{\xi \sim P_\theta(.)} f_2(\xi)$$

$$\vdots$$

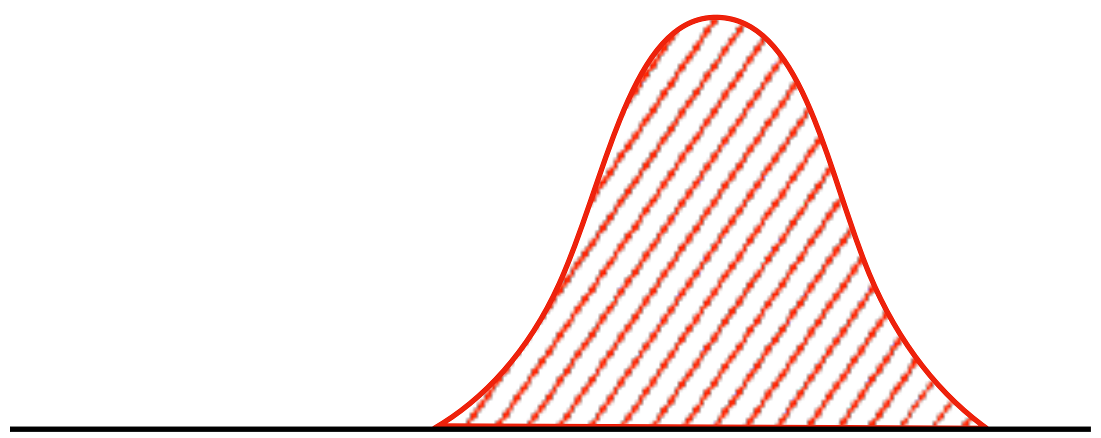
$$\mathbb{E}_{\xi^h \sim P_{expert}(.)} f_k(\xi^h) = \mathbb{E}_{\xi \sim P_\theta(.)} f_k(\xi)$$

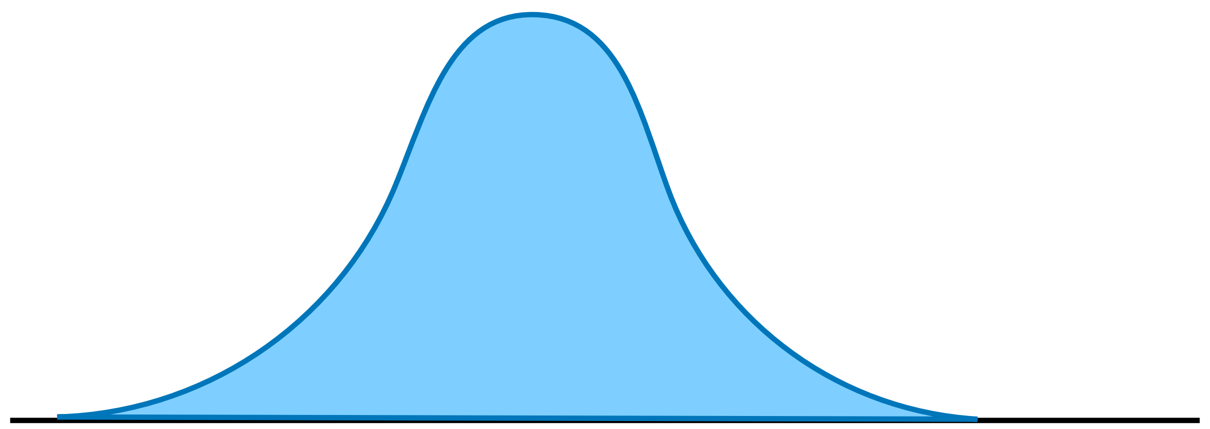


Moment Matching Constraint

Find $P_\theta(\xi)$

$$\mathbb{E}_{\xi \sim P_\theta(.)}\, f(\xi) = \mathbb{E}_{\xi^h \sim P(.)}\, f(\xi^h) \quad \forall f \in \mathcal{F}$$

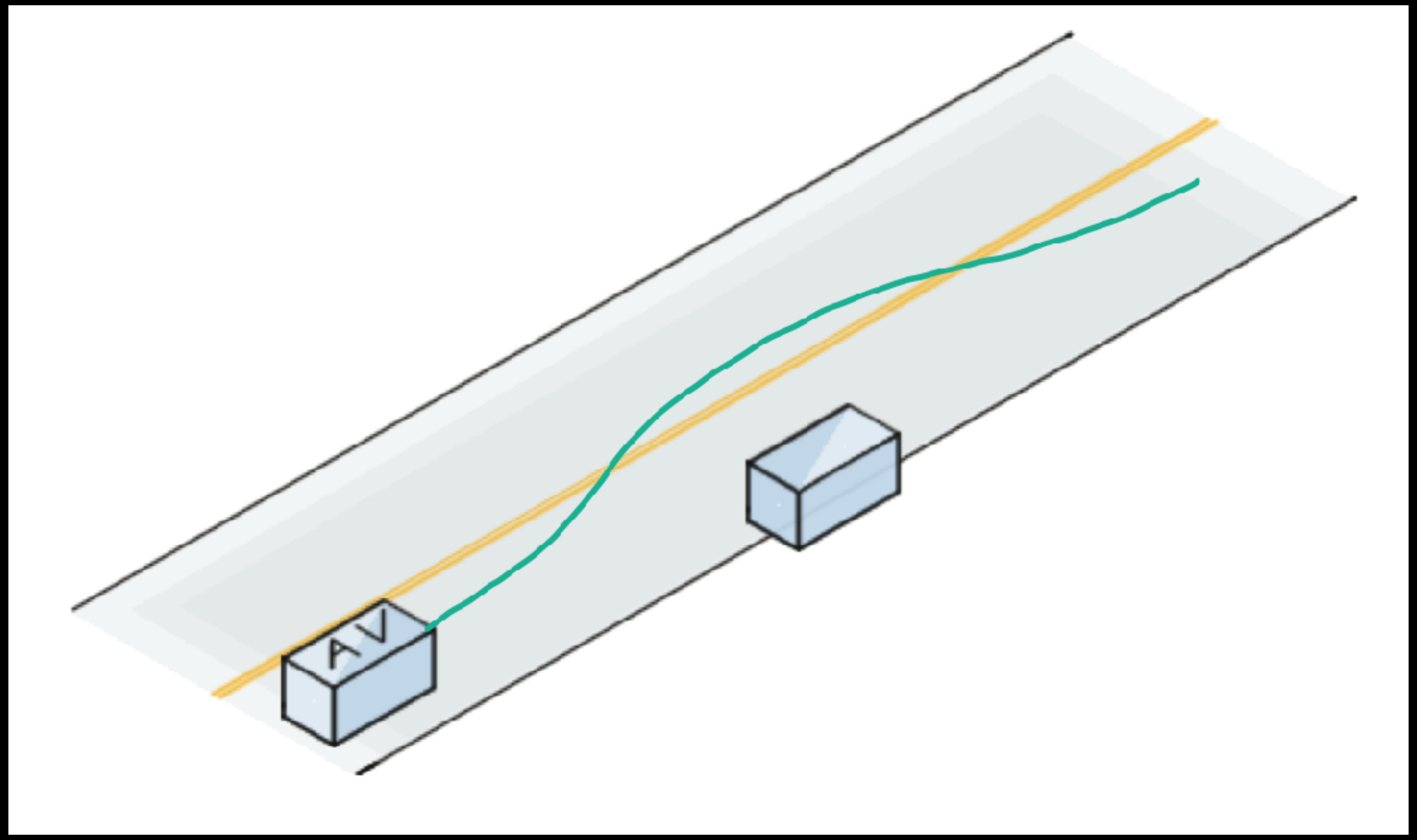




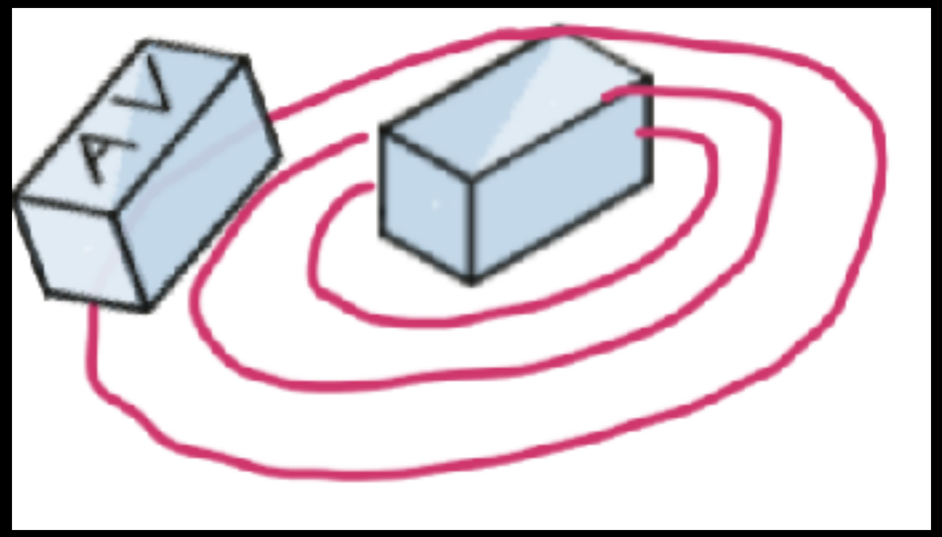
What are some features for this task?

Moments of
some features
of human trajectories

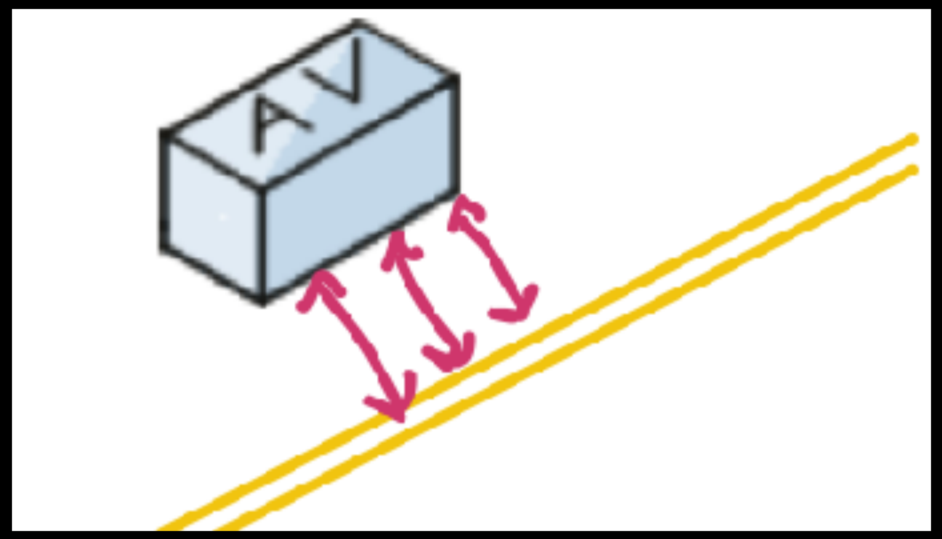
$$\mathbb{E}_{\xi^* \sim \pi^*} f(\xi^*)$$



Control Effort $f_1(\cdot)$



Proximity $f_2(\cdot)$



Boundary Violation $f_3(\cdot)$

Is there a unique solution to the moment matching problem?

Principle of Maximum Entropy to the rescue!

Information Theory and Statistical Mechanics

E. T. JAYNES
Department of Physics, Stanford University, Stanford, California
(Received September 4, 1956; revised manuscript received March 4, 1957)

Information theory provides a constructive criterion for setting up probability distributions on the basis of partial knowledge, and leads to a type of statistical inference which is called the maximum-entropy estimate. It is the least biased estimate possible on the given information; i.e., it is maximally noncommittal with regard to missing information. If one considers statistical mechanics as a form of statistical inference rather than as a physical theory, it is found that the usual computational rules, starting with the determination of the partition function, are an immediate consequence of the maximum-entropy principle. In the resulting "subjective statistical mechanics," the usual rules are thus justified independently of any physical argument, and in particular independently of experimental verification; whether or not the results agree with experiment, they still represent the best estimates that could have been made on the basis of the information available.

It is concluded that statistical mechanics need not be regarded as a physical theory dependent for its validity on the truth of additional assumptions not contained in the laws of mechanics (such as ergodicity, metric transitivity, equal *a priori* probabilities, etc.). Furthermore, it is possible to maintain a sharp distinction between its physical and statistical aspects. The former consists only of the correct enumeration of the states of a system and their properties; the latter is a straightforward example of statistical inference.

1. INTRODUCTION

THE recent appearance of a very comprehensive survey[1] of past attempts to justify the methods of statistical mechanics in terms of mechanics, classical or quantum, has helped greatly, and at a very opportune time, to emphasize the unsolved problems in this field. Although the subject has been under development for many years, we still do not have a complete and satisfactory theory, in the sense that there is no line of argument proceeding from the laws of microscopic mechanics to macroscopic phenomena, that is generally regarded by physicists as convincing in all respects. Such an argument should (a) be free from objection on mathematical grounds, (b) involve no additional arbi-

[1] D. ter Haar, Revs. Modern Phys. **27**, 289 (1955).

The loaded die problem

What is the measure of uncertainty?

$$H(X) = -\sum_{X} P(X) \log P(X)$$

1. Decreasing in $P(X)$, such that if $P(X_1) < P(X_2)$, then $h(P(X_1)) > h(P(X_2))$.
2. Independent variables add, such that if X and Y are independent, then $H(P(X,Y)) = H(P(X)) + H(P(Y))$.

These are only satisfied for $-\log(\cdot)$. Think of it as a "surprise" function.

A Mathematical Theory of Communication

By C. E. SHANNON

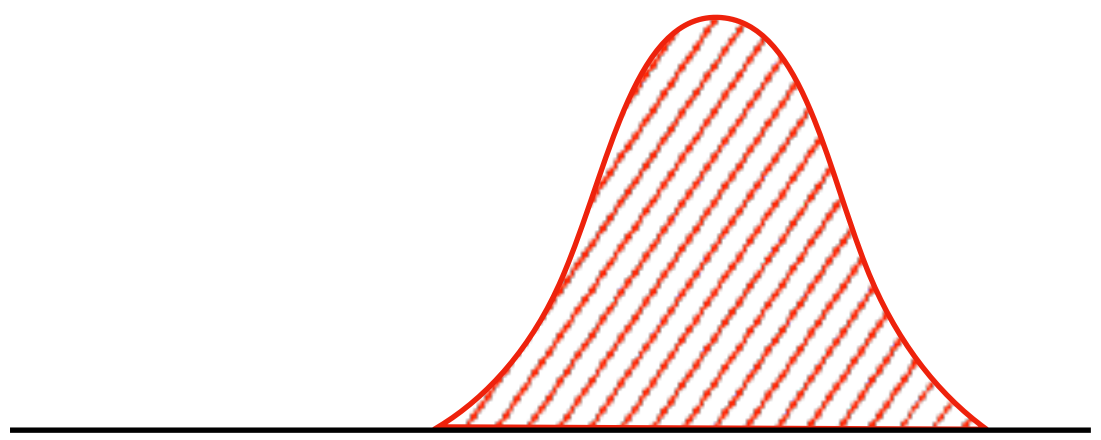
INTRODUCTION

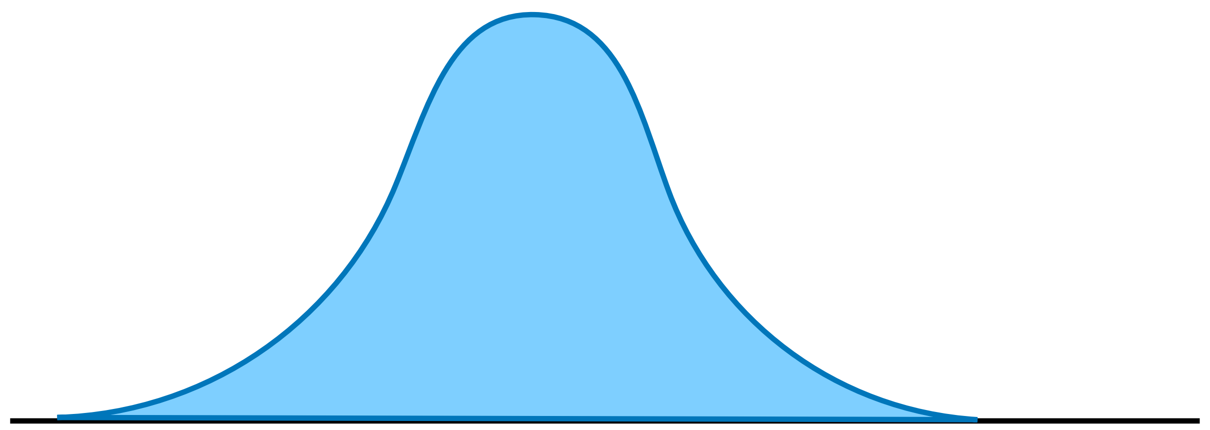
Maximum Entropy Moment Matching

Find $P_\theta(\xi)$

$$\max_{\theta} H(P_\theta(\xi))$$

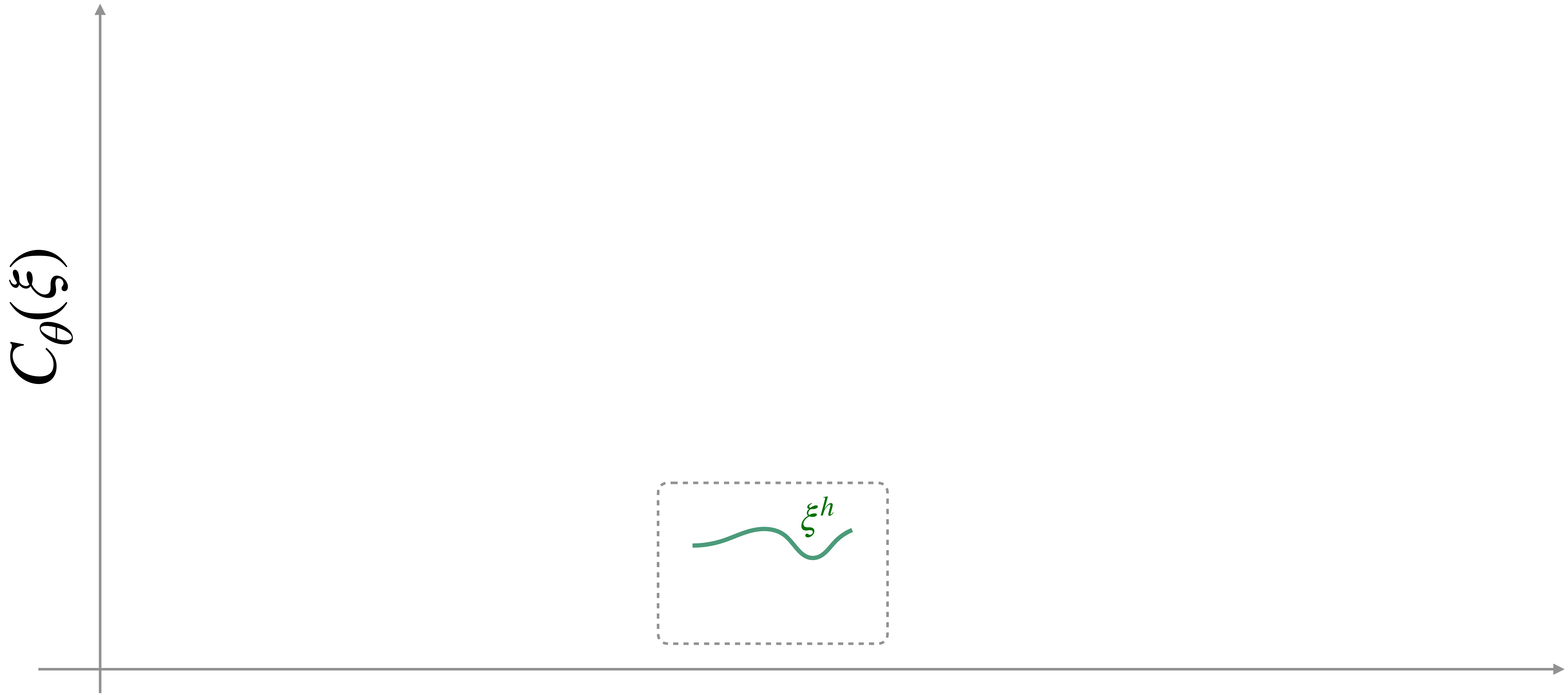
$$\mathbb{E}_{\xi \sim P_\theta(.)} f(\xi) = \mathbb{E}_{\xi^h \sim P(.)} f(\xi^h) \quad \forall f \in \mathscr{F}$$



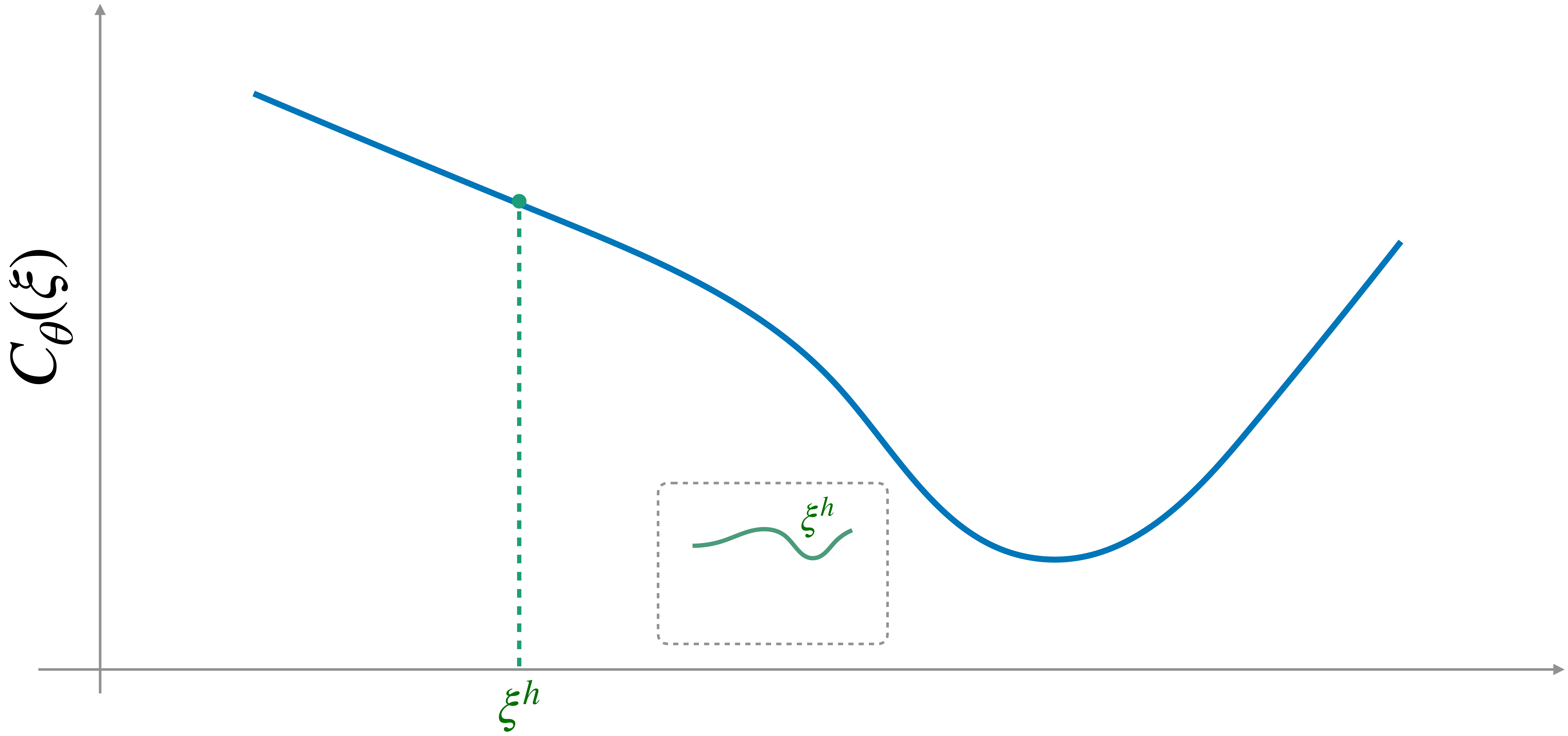


Let's derive!

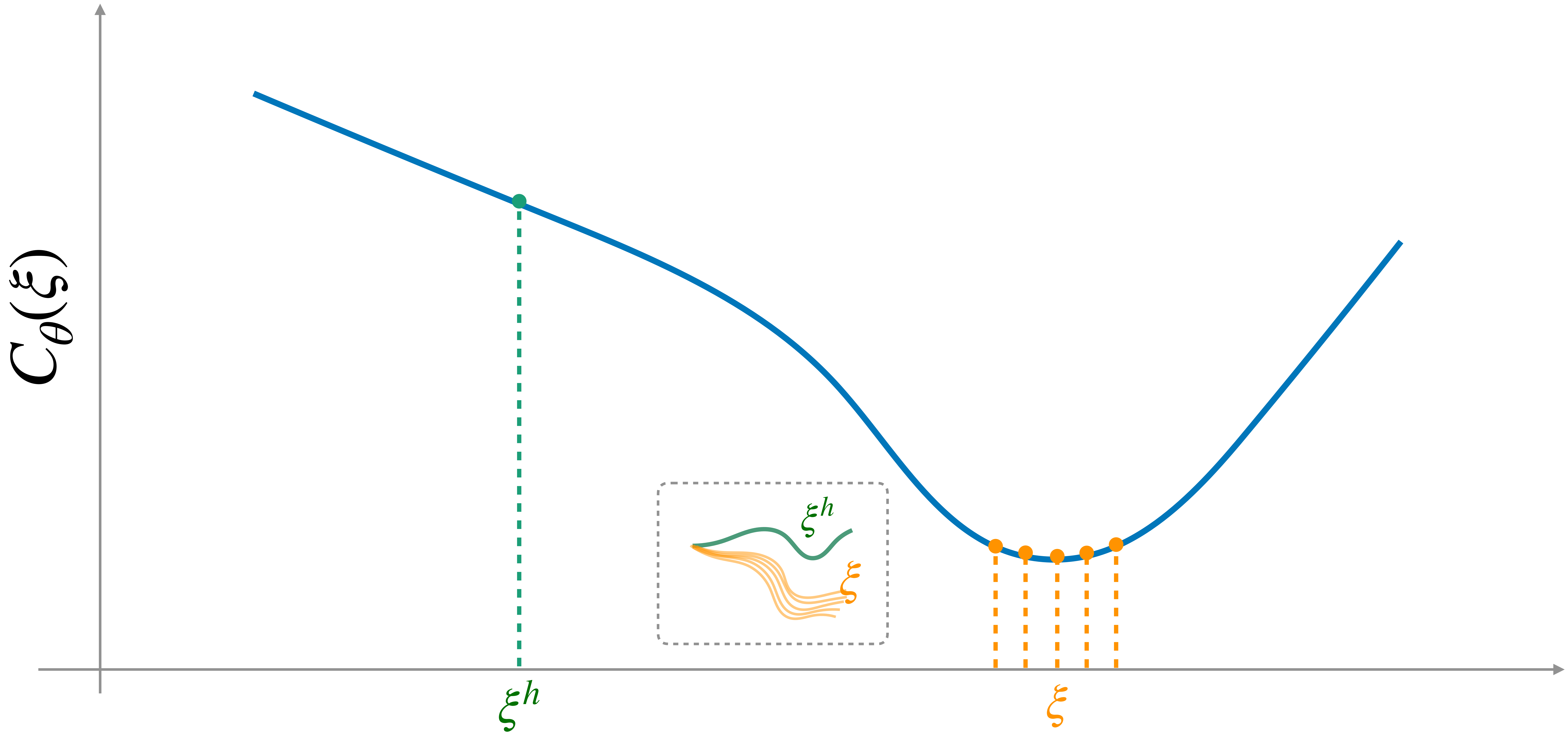
Animation from Sodhi et al. 2021



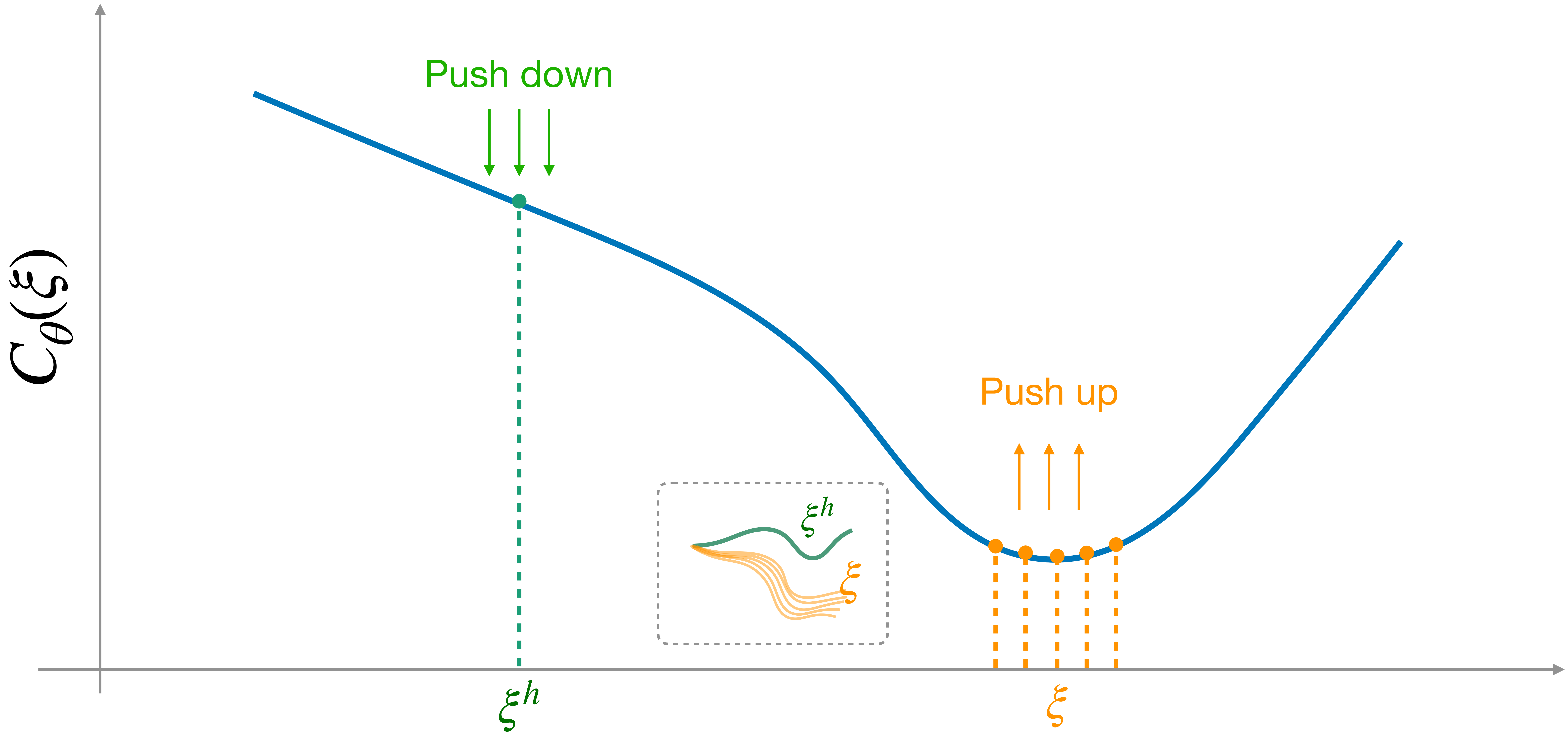
Animation from Sodhi et al. 2021



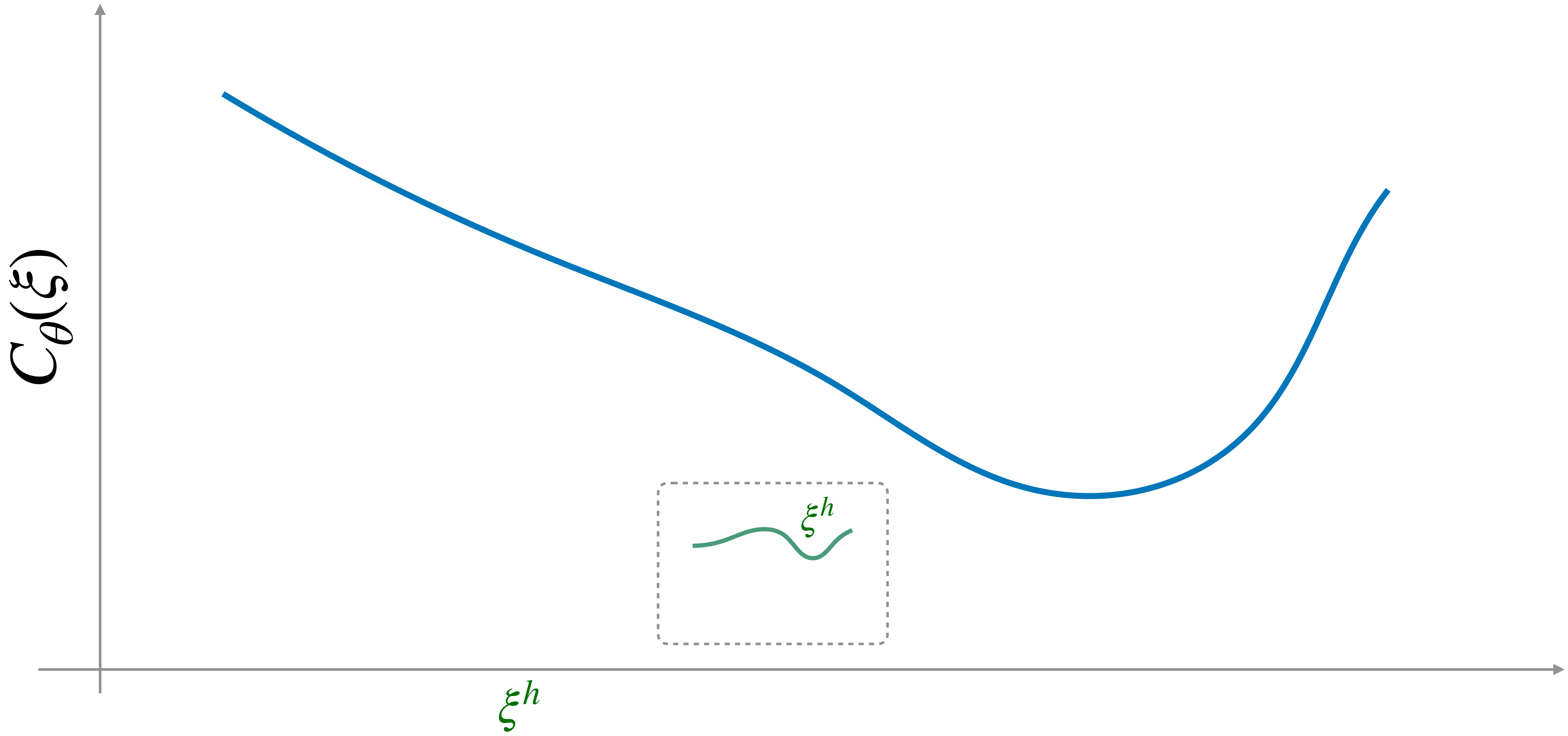
Animation from Sodhi et al. 2021



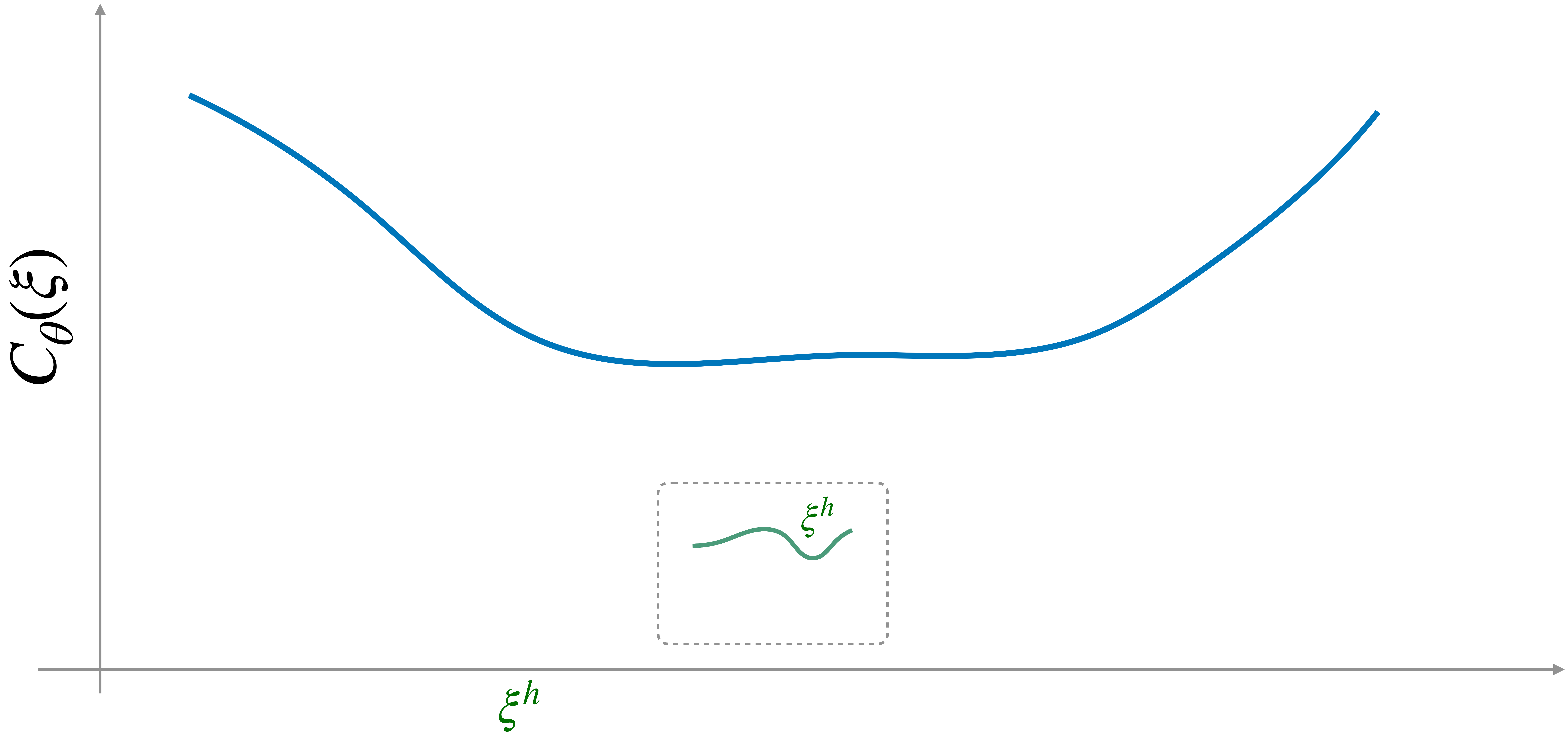
Animation from Sodhi et al. 2021



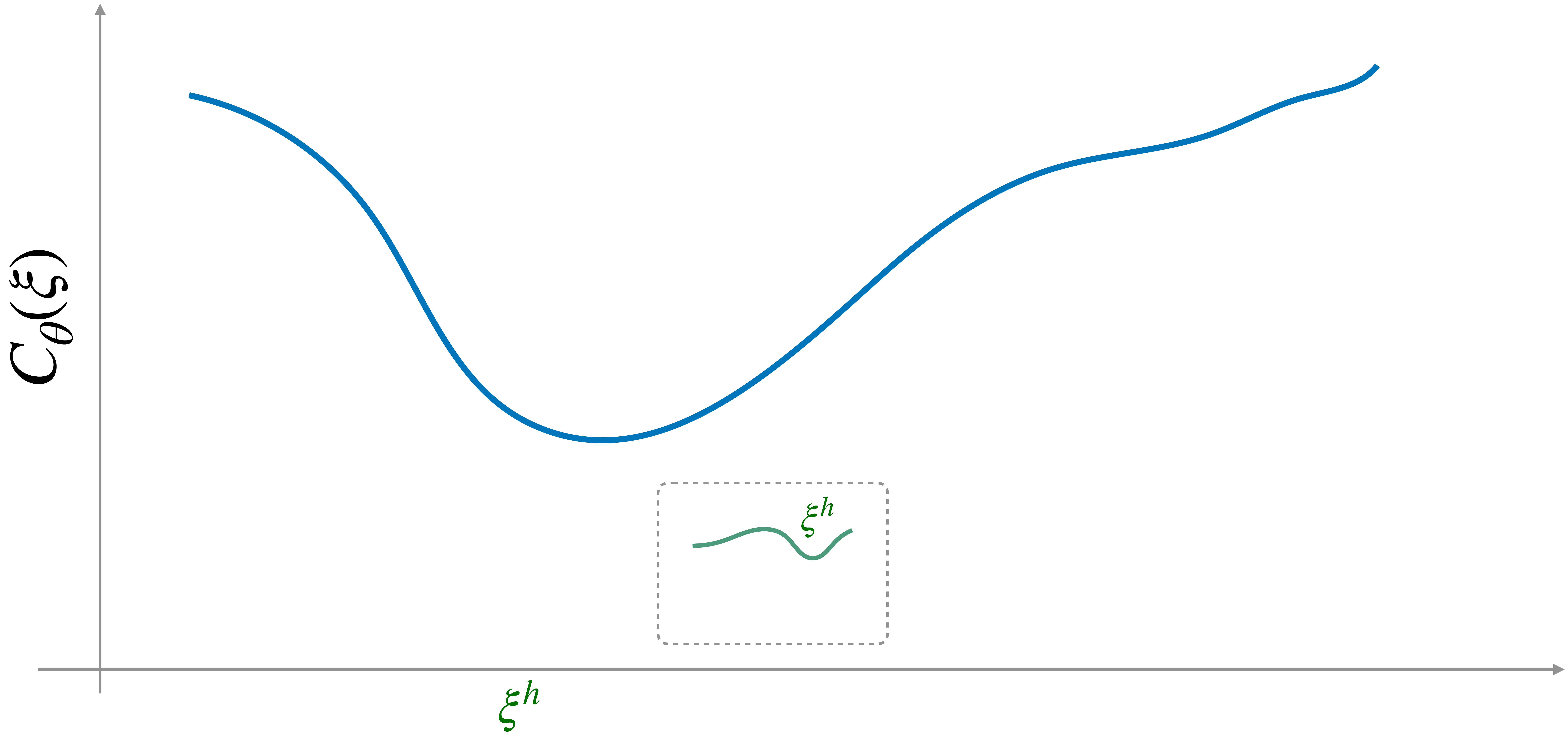
Animation from Sodhi et al. 2021



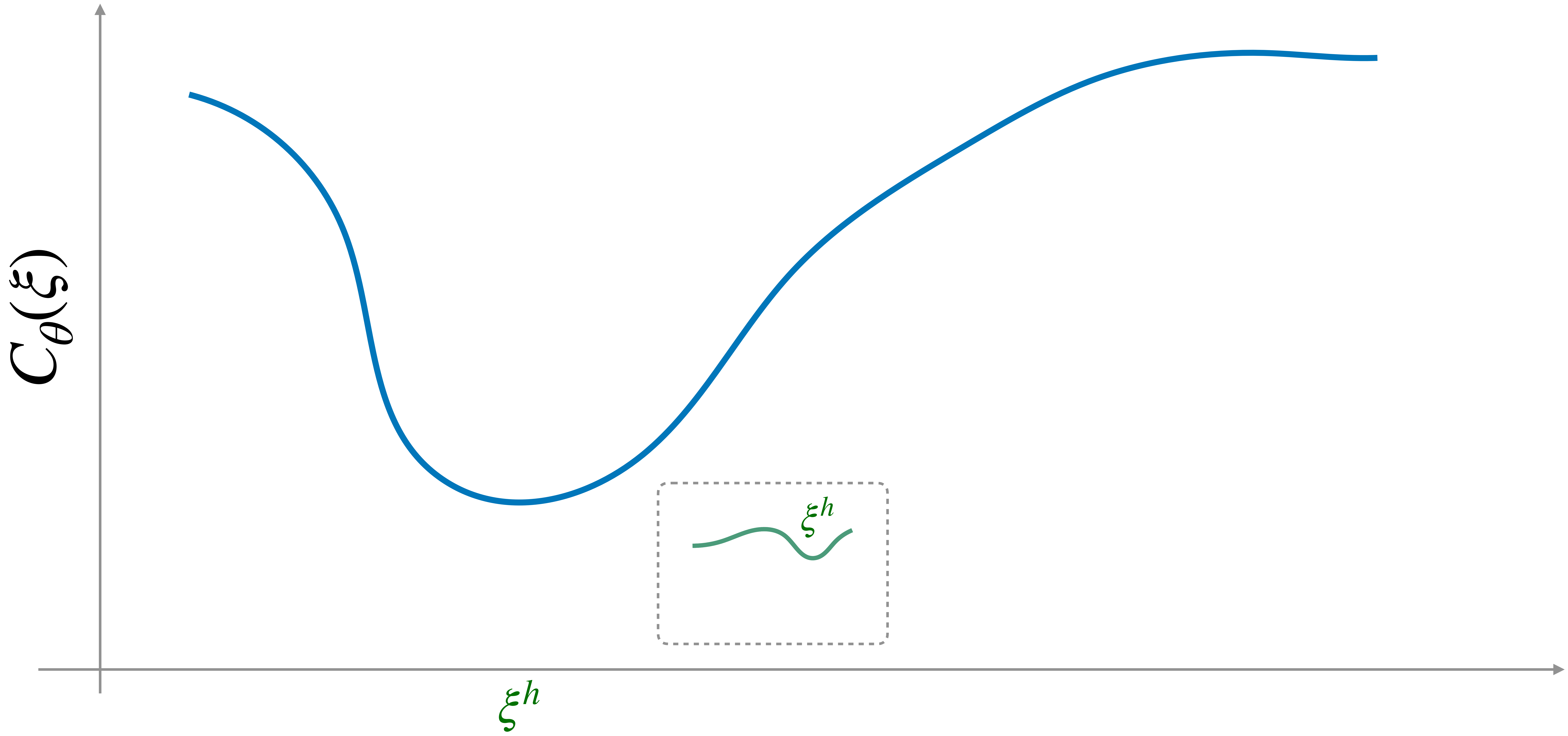
Animation from Sodhi et al. 2021



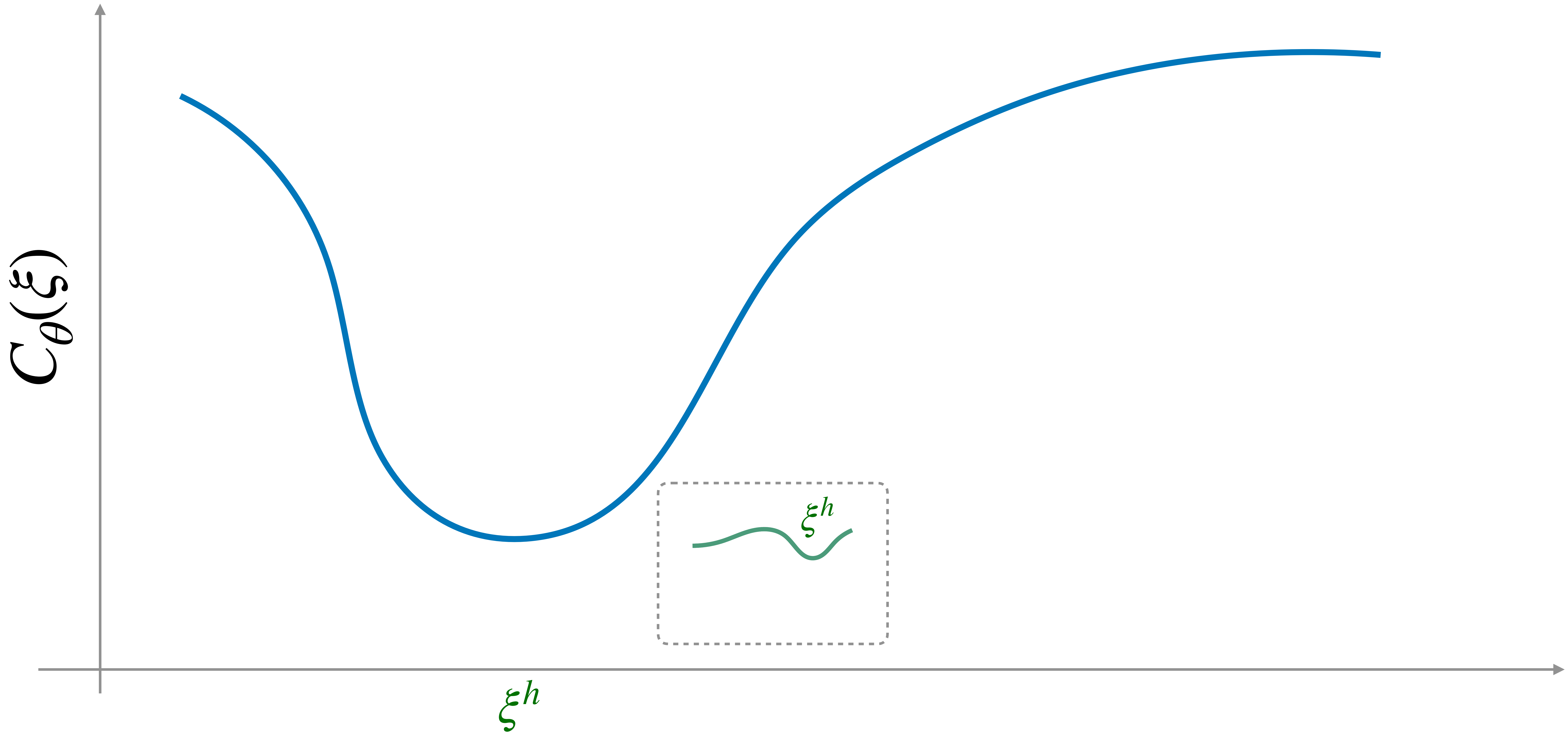
Animation from Sodhi et al. 2021



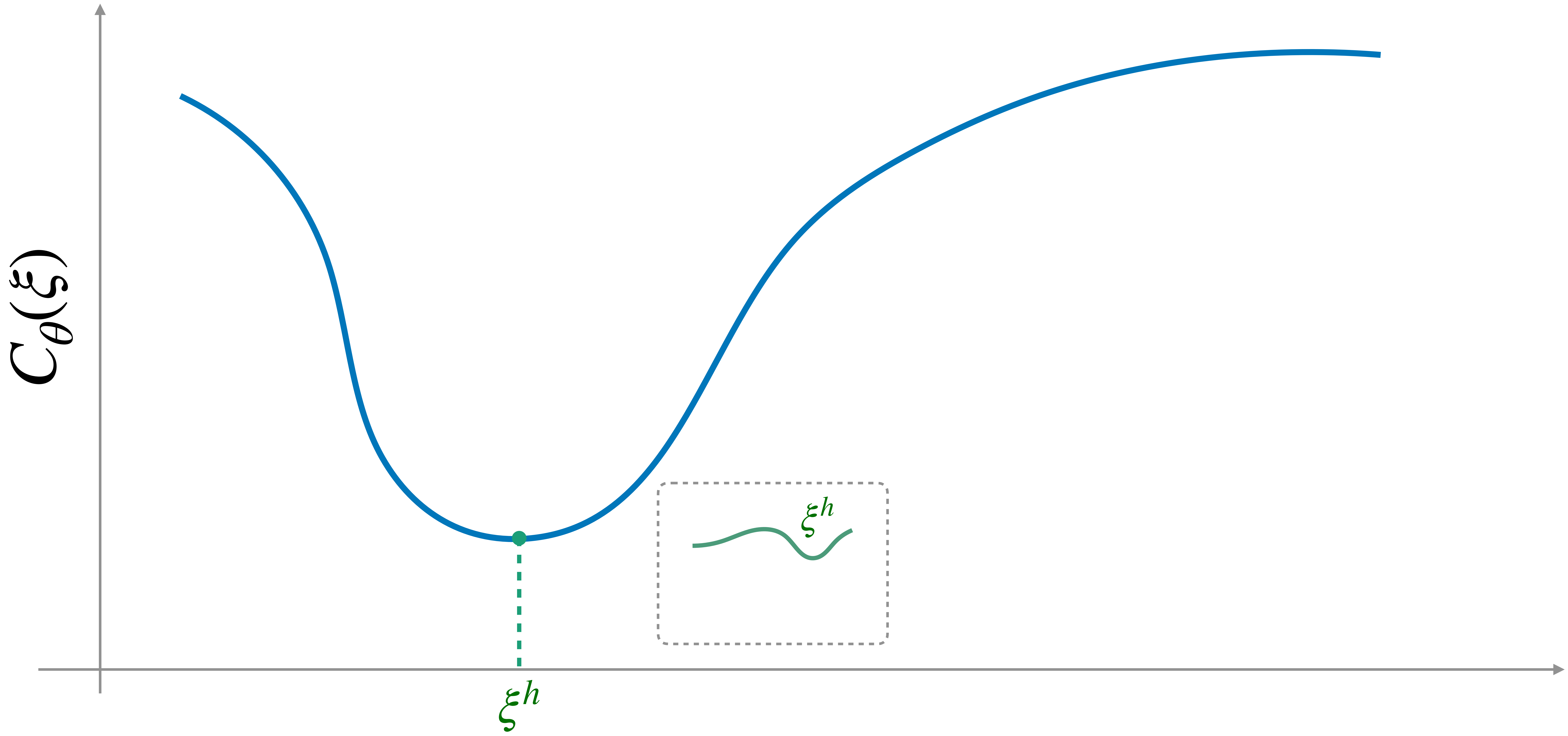
Animation from Sodhi et al. 2021



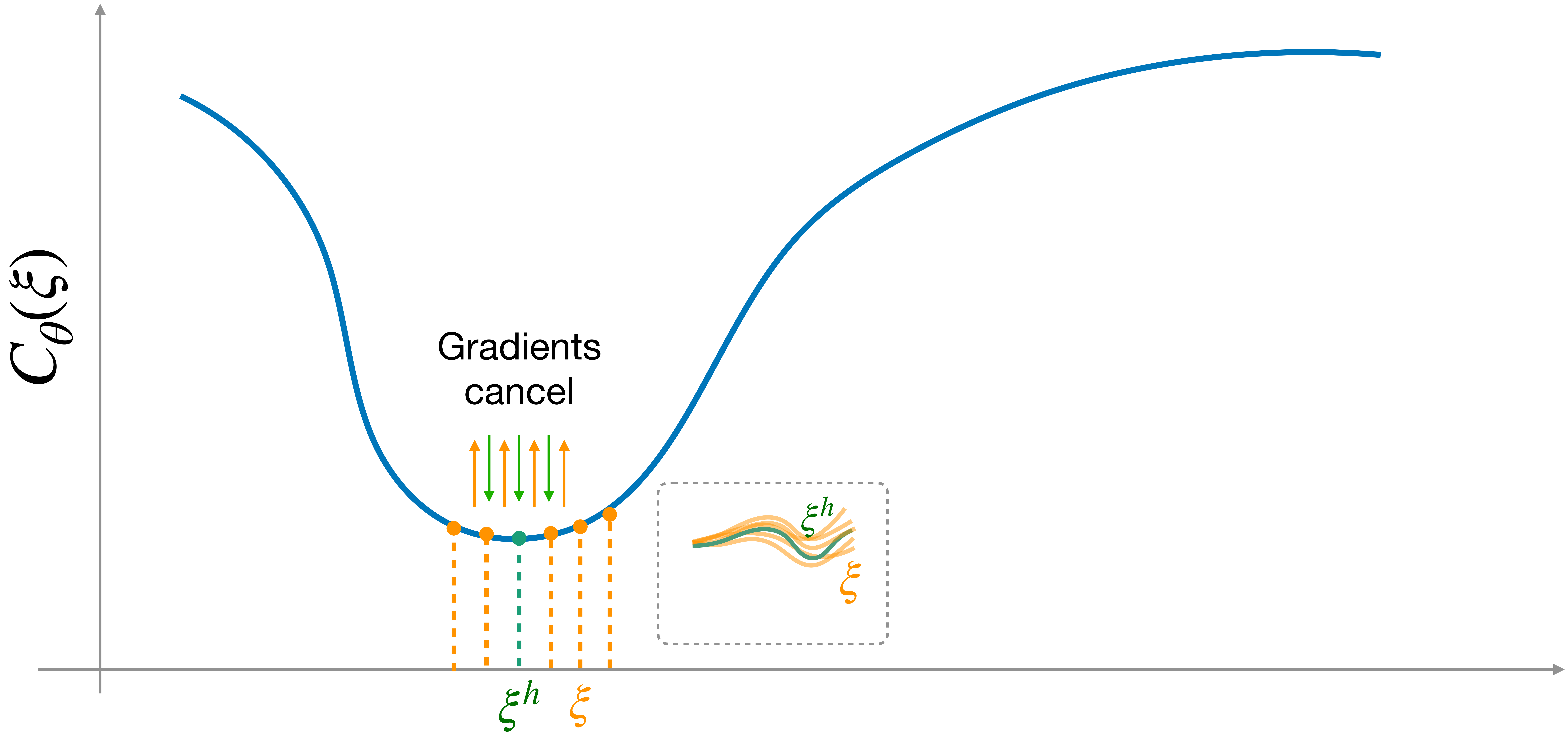
Animation from Sodhi et al. 2021



Animation from Sodhi et al. 2021



Animation from Sodhi et al. 2021



Okay...

But how do we sample from

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

Let's derive soft value iteration

Soft Actor Critic

Soft actor-critic

1. Q-function update

Update Q-function to evaluate current policy:

$$Q(\mathbf{s},\mathbf{a}) \leftarrow r(\mathbf{s},\mathbf{a}) + \mathbb{E}_{\mathbf{s}'\sim p_{\mathbf{s}},\ \mathbf{a}'\sim\pi}\left[Q(\mathbf{s}',\mathbf{a}') - \log\pi(\mathbf{a}'|\mathbf{s}')\right]$$

This converges to Q^{π}.

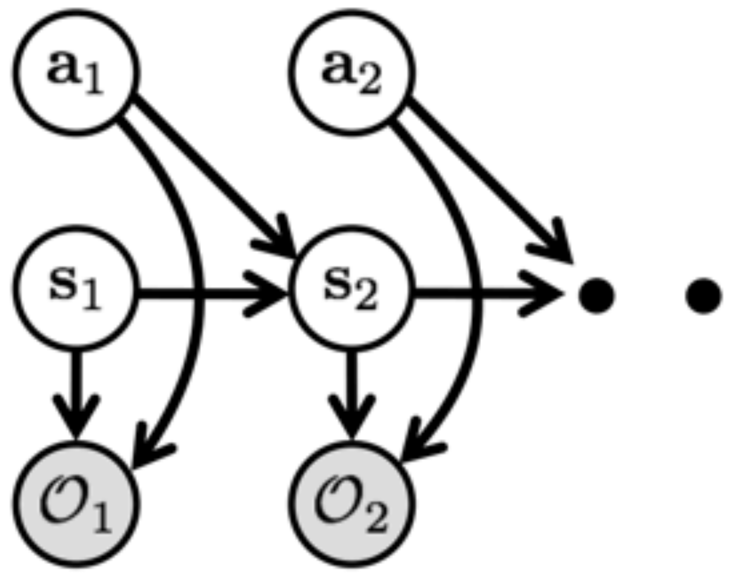
2. Update policy

Update the policy with gradient of information projection:

$$\pi_{\text{new}} = \arg\min_{\pi'} \mathrm{D}_{\mathrm{KL}}\left(\pi'(\cdot|\mathbf{s}) \,\middle\|\, \frac{1}{Z}\exp Q^{\pi_{\text{old}}}(\mathbf{s},\cdot)\right)$$

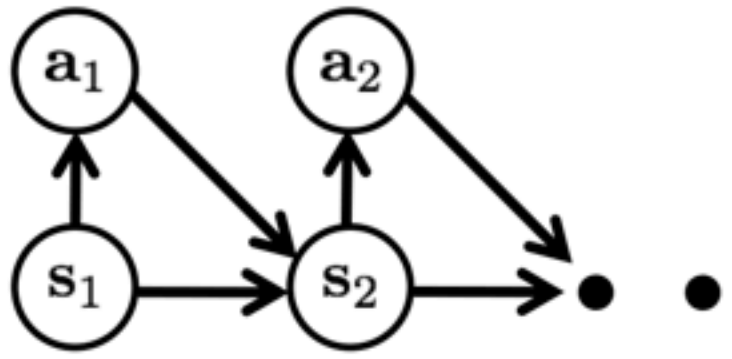
In practice, only take one gradient step on this objective

3. Interact with the world, collect more data



update messages

fit variational distribution

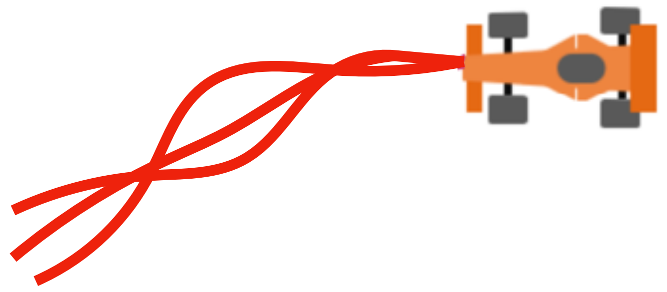


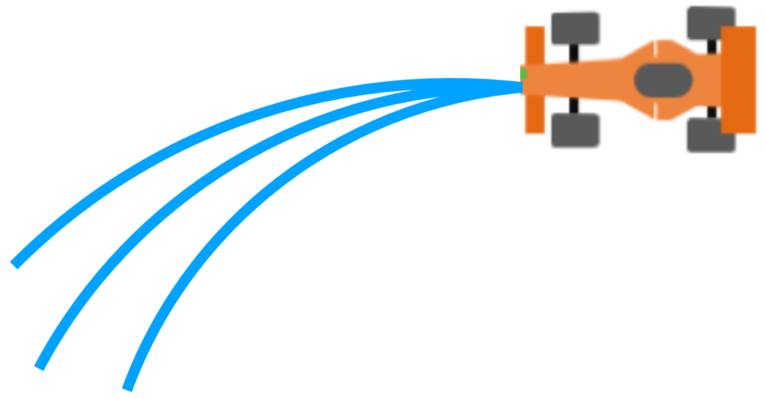
Haarnoja, Zhou, Hartikainen, Tucker, Ha, Tan, Kumar, Zhu, Gupta, Abbeel, L. **Soft Actor-Critic Algorithms and Applications**. ‘18

Credit S.Levine.

Max Entropy Inverse Reinforcement Learning

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

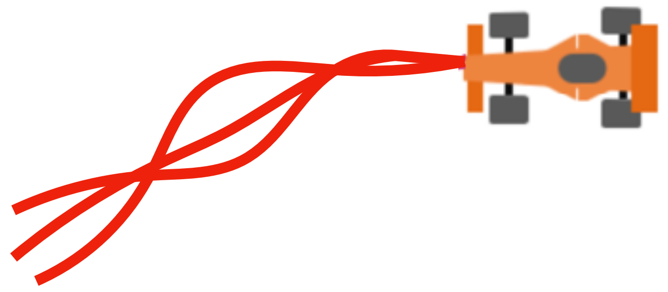


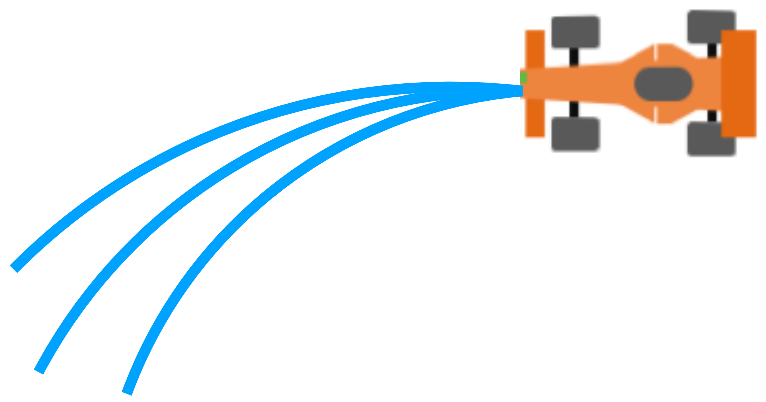


Entropy

The Entropy Regularized Game

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





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

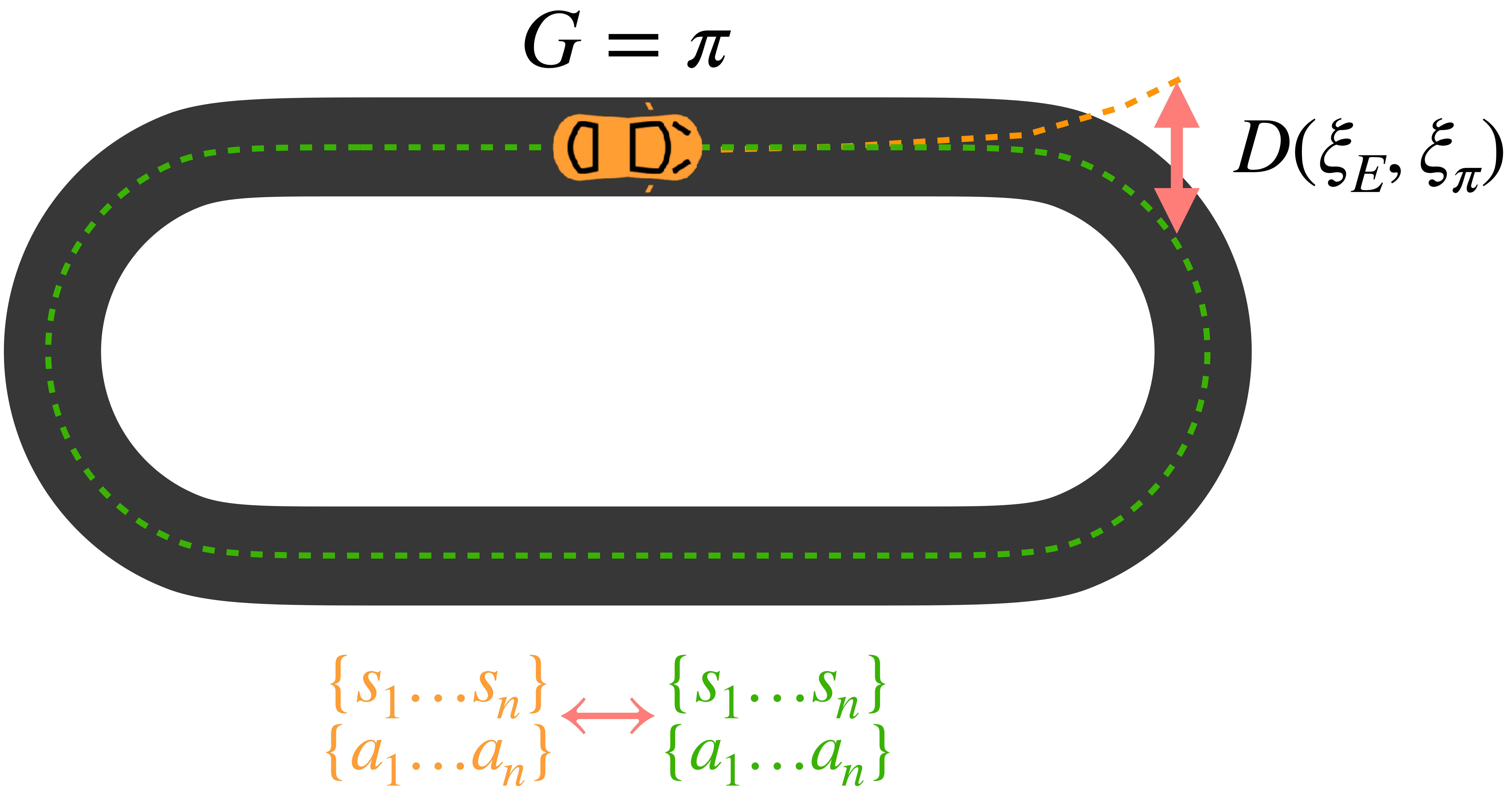
Inverse Reinforcement Learning without Reinforcement Learning

Gokul Swamy



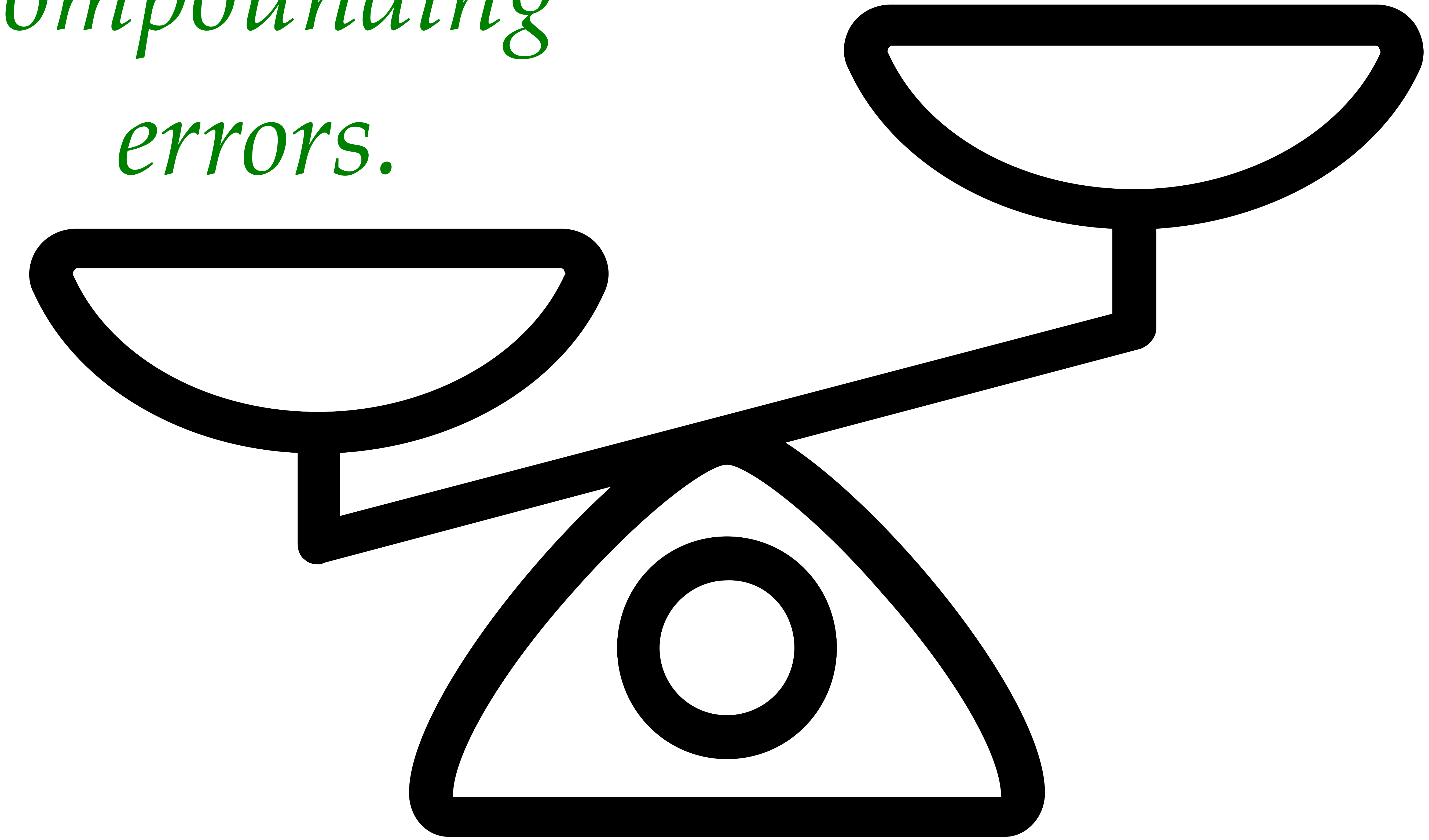

(with Sanjiban Choudhury, Drew Bagnell, and Steven Wu)

Inverse Reinforcement Learning for Imitation



Requires repeatedly solving an RL problem.

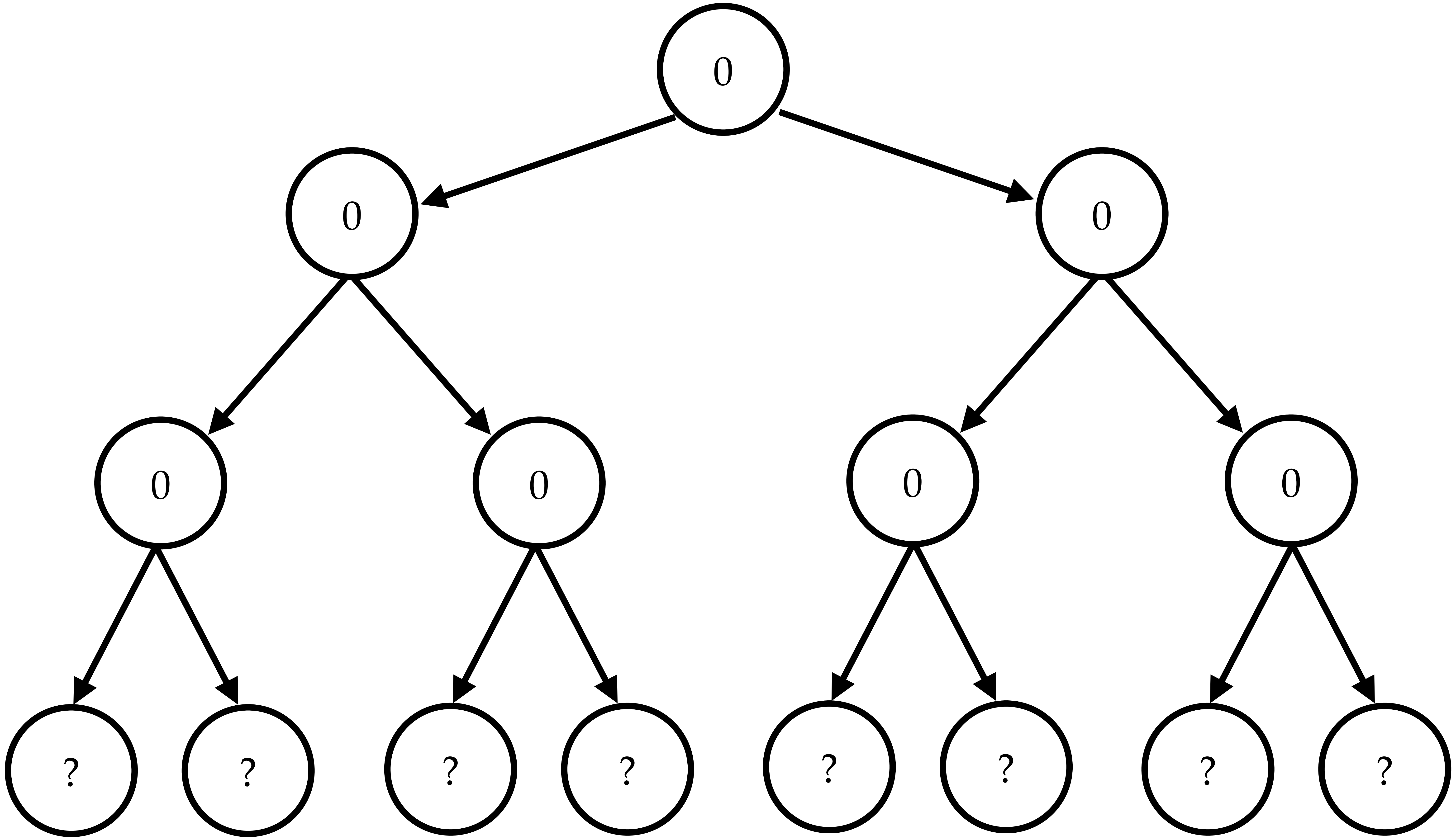
Robust to compounding errors.



[SCBW, '21]

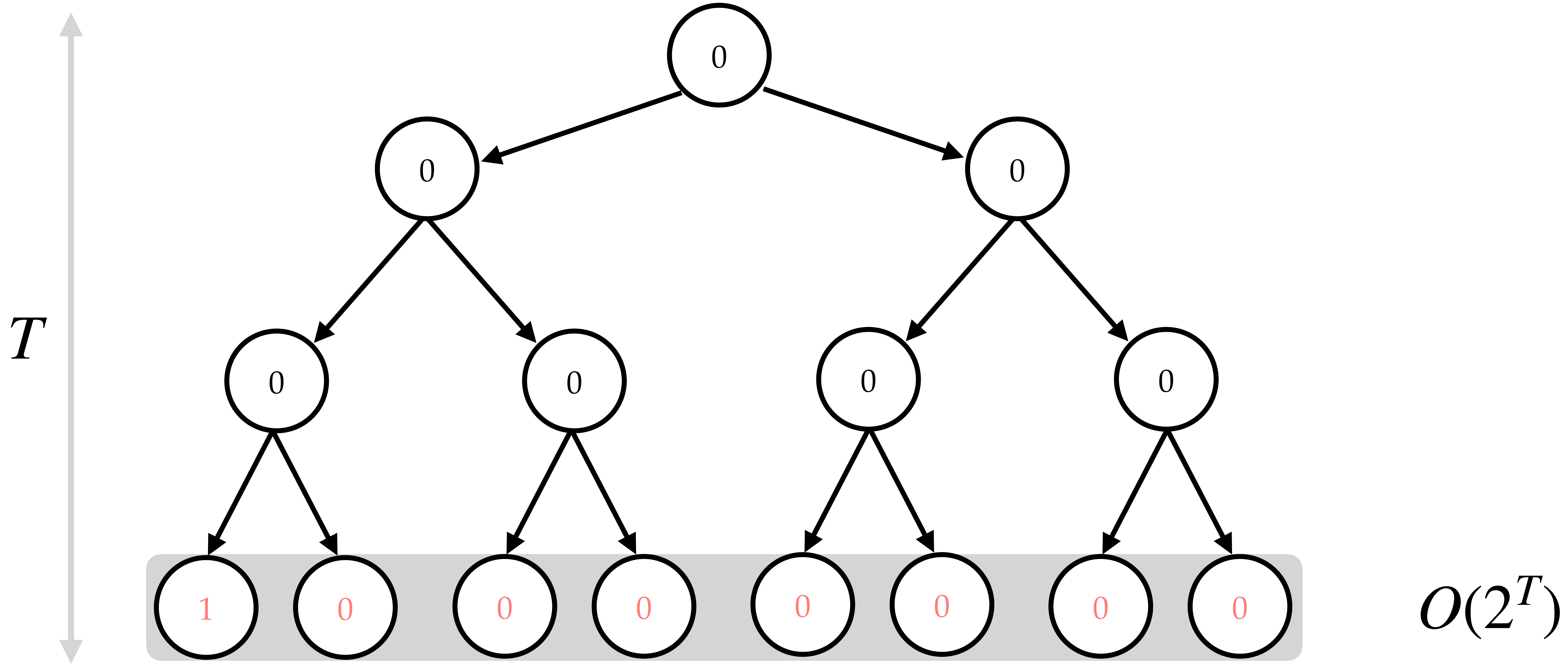
RL makes IRL Inefficient

$\pi_E \overset{f}{\leftrightarrow} \pi$



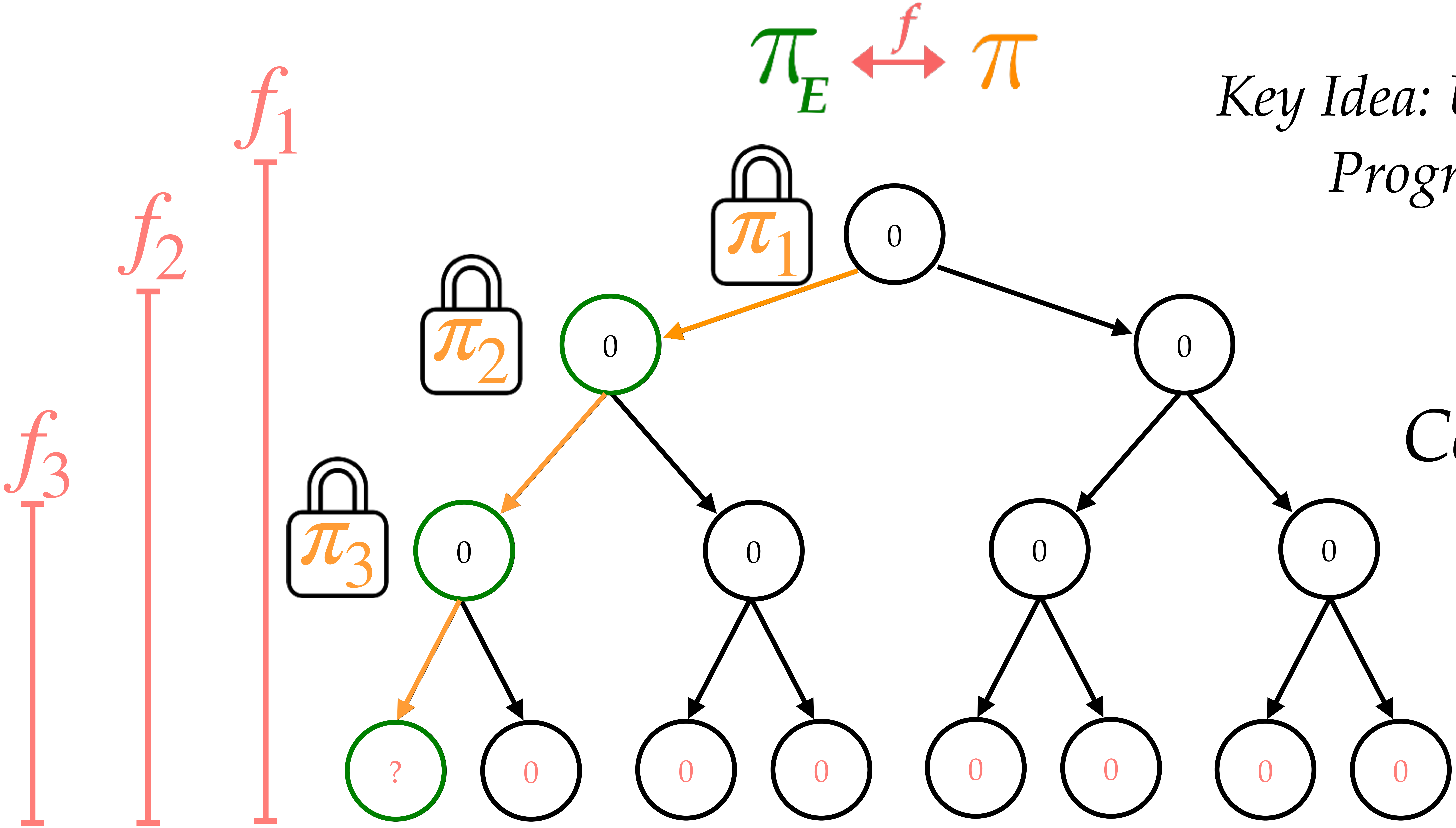
RL makes IRL Inefficient

$\pi_E \overset{f}{\longleftrightarrow} \pi$



🔑 ***Insight***: *We can reset the learner to states from the expert demonstrations to reduce unnecessary exploration.*

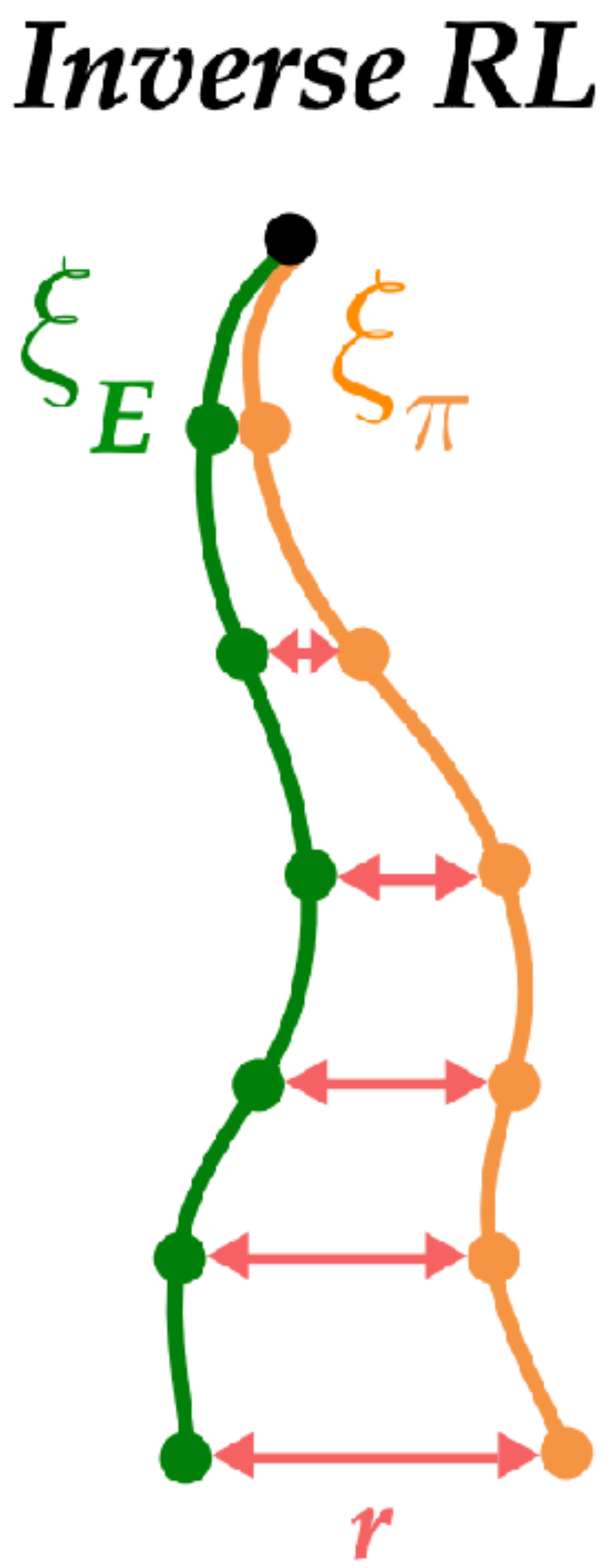
Speeding up IRL with Expert Resets

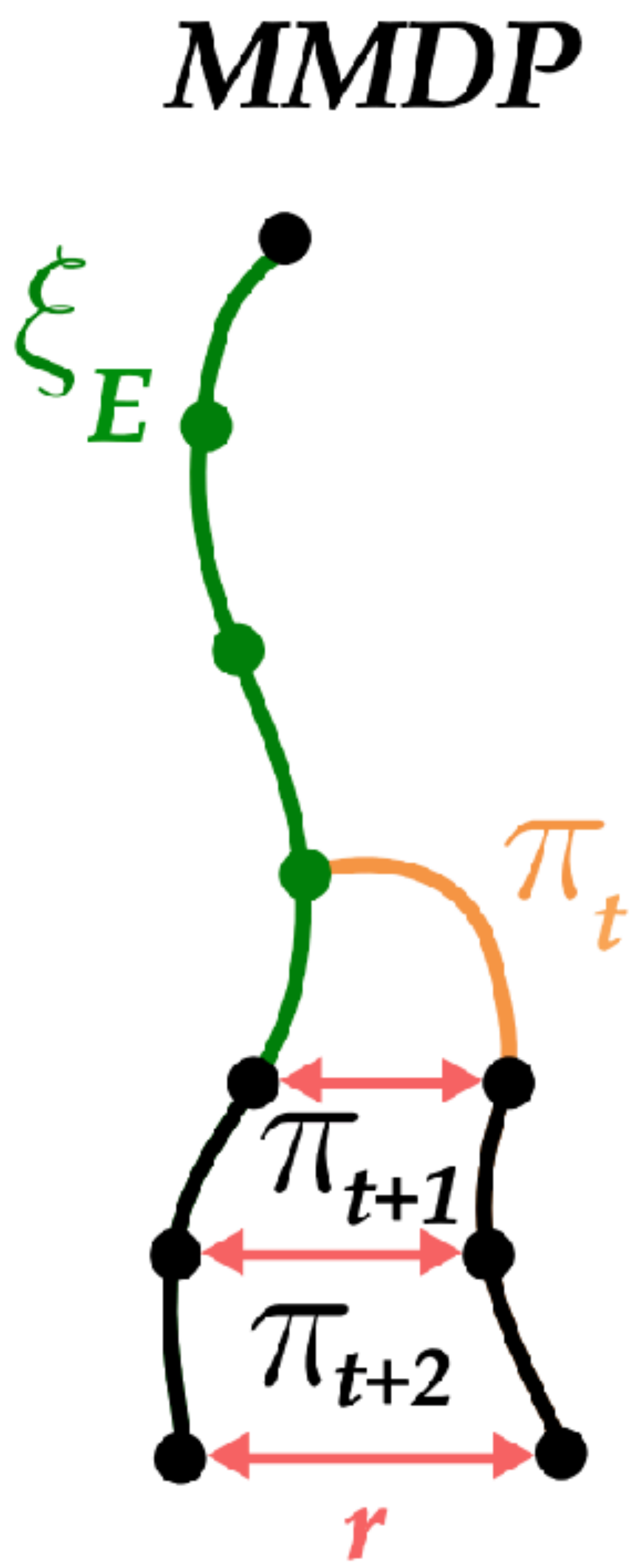


Key Idea: Use Dynamic Programming

$O(T^2)$ *Complexity!*

🔑 *Contribution: Poly-time Algorithms for IRL*





Expert Resets Speed Up IRL