Of Moments and Matching:

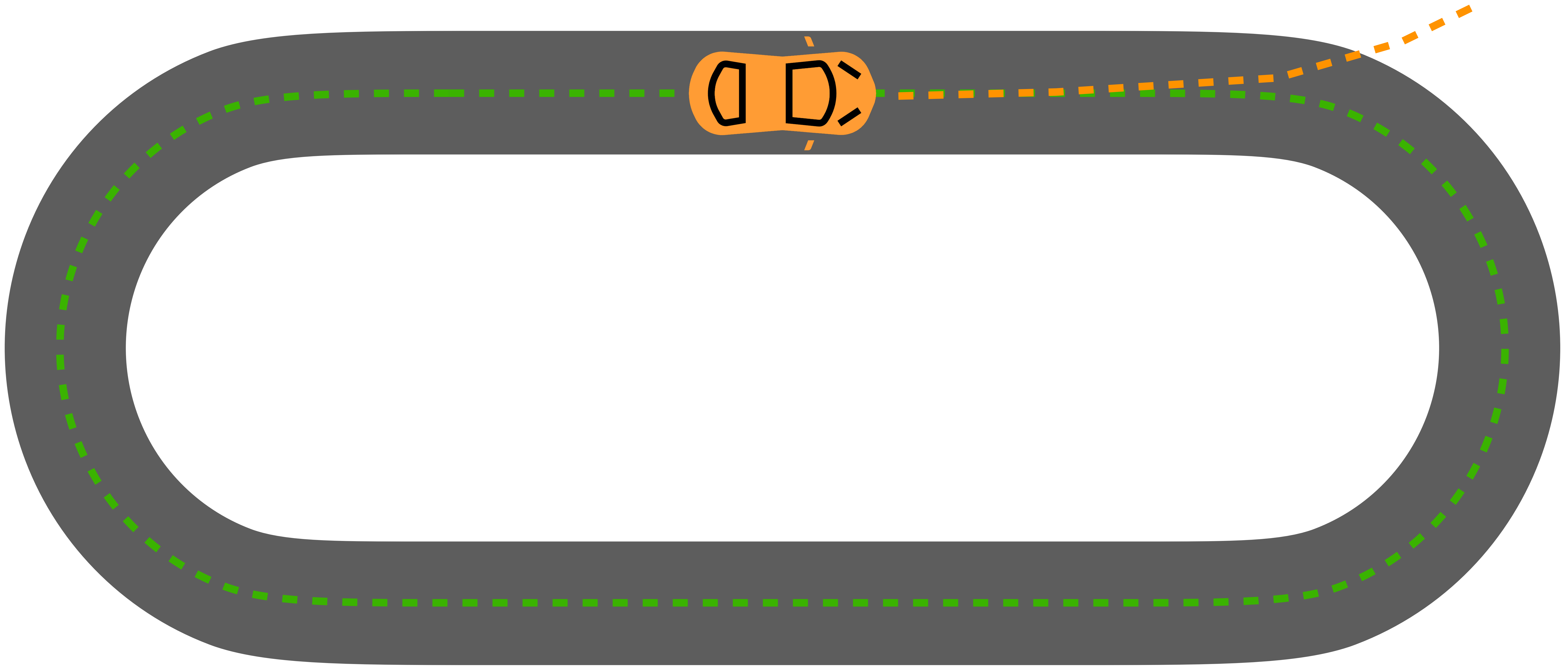
A Game-Theoretic Framework for Closing the Imitation Gap

Gokul Swamy, Sanjiban Choudhury, Drew Bagnell, Steven Wu

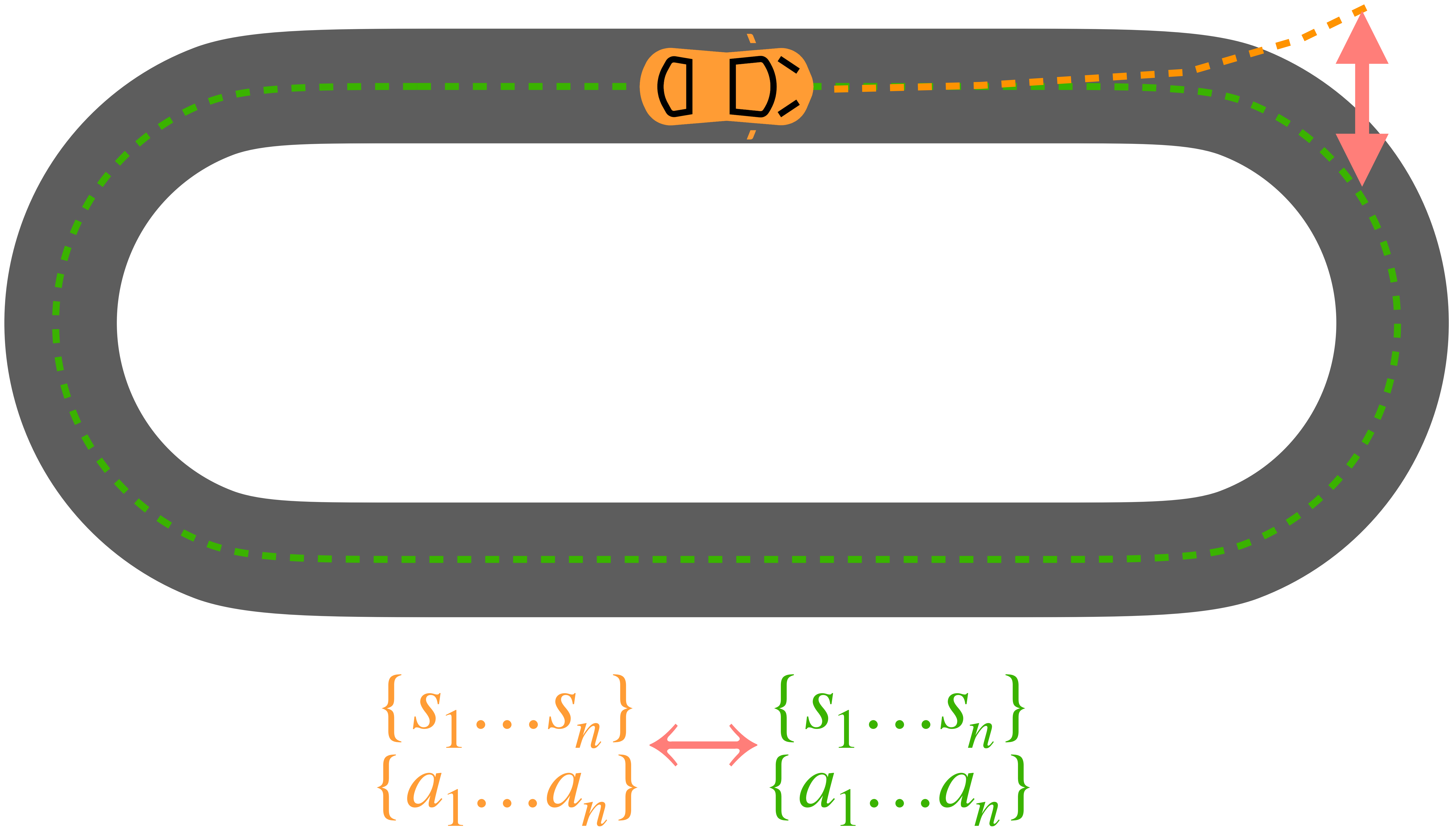
THE ROBOTICS INSTITUTE

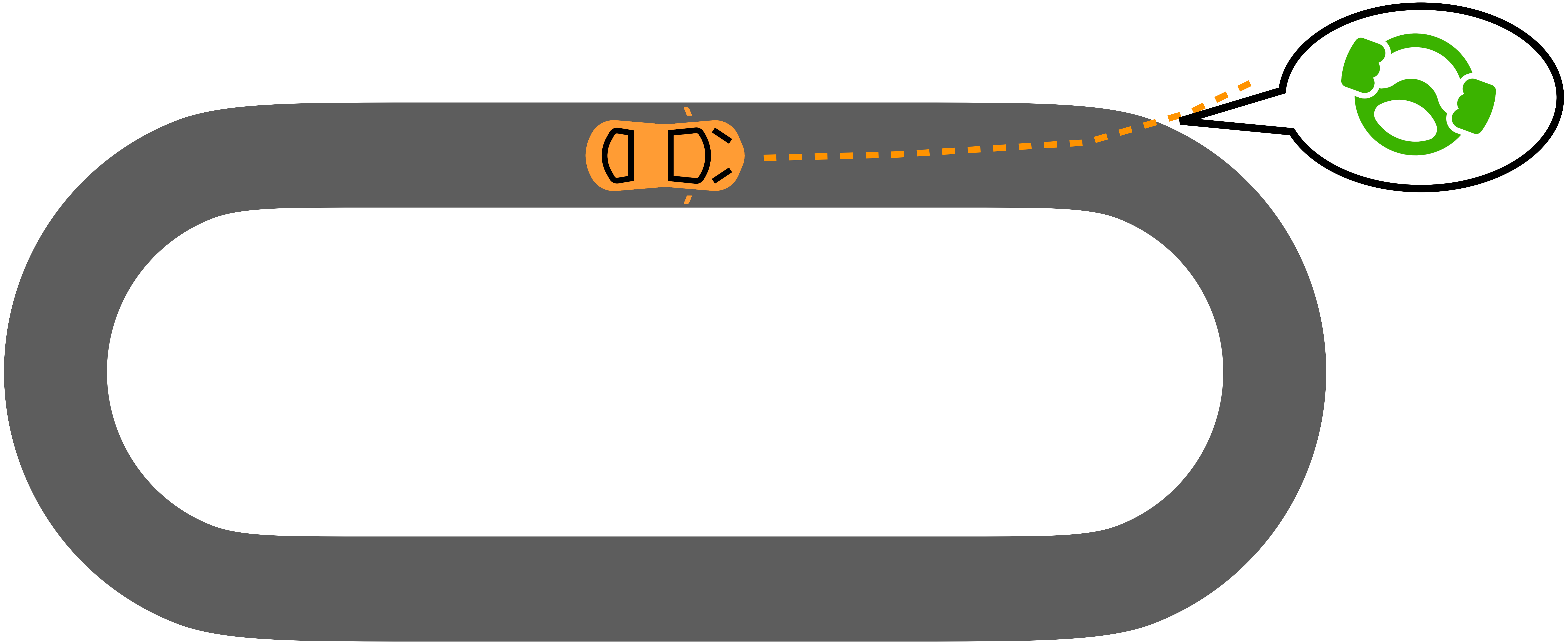
Aurora

isr



$$\{s_1 \ldots s_n\} \mapsto \{a_1 \ldots a_n\}$$


$\{s_1 \ldots s_n\}$
$\{a_1 \ldots a_n\}$
$\longleftrightarrow$
$\{s_1 \ldots s_n\}$
$\{a_1 \ldots a_n\}$

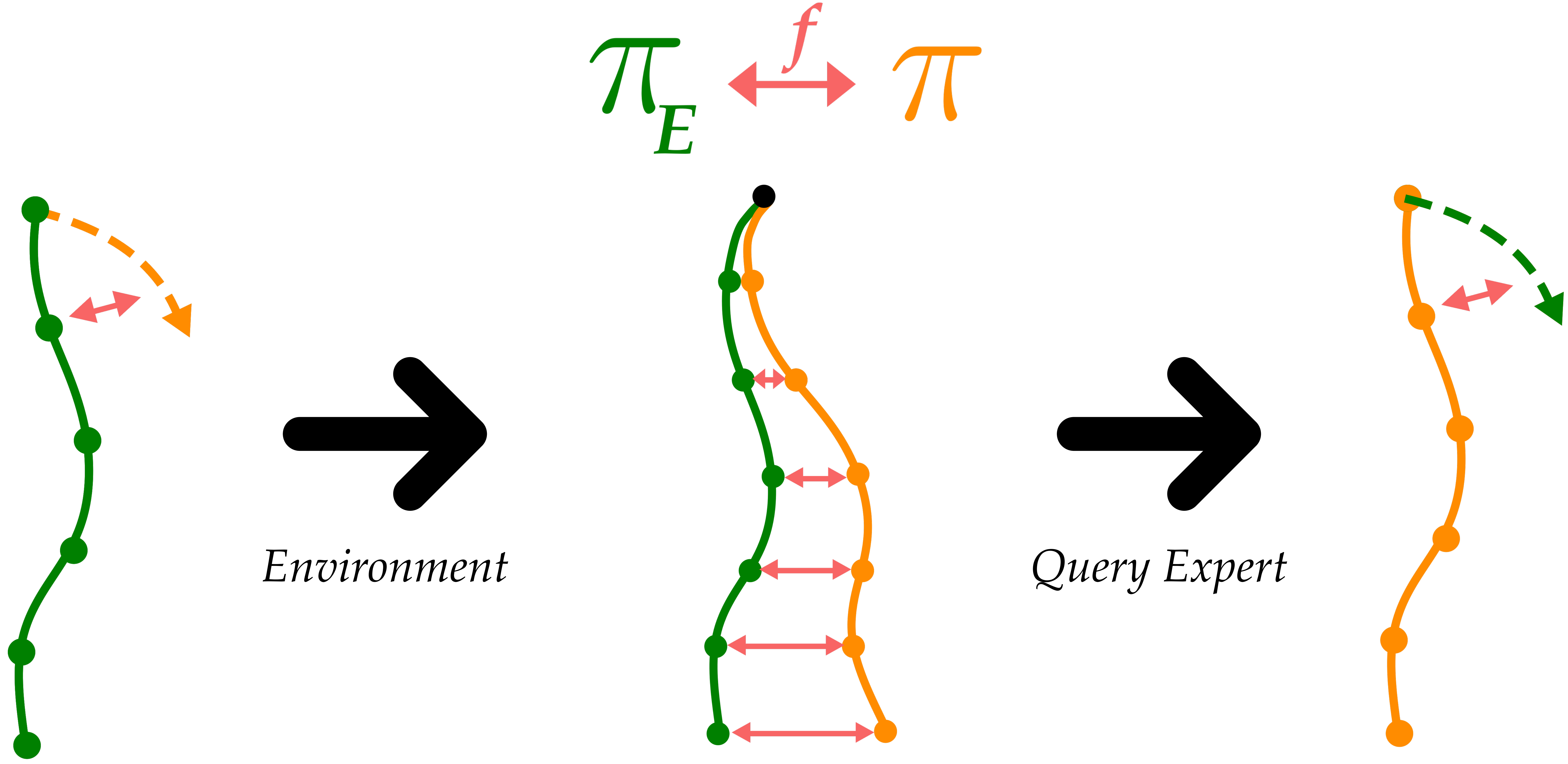


$$\{s_1 \ldots s_n\} \mapsto \{a_1 \ldots a_n\}$$

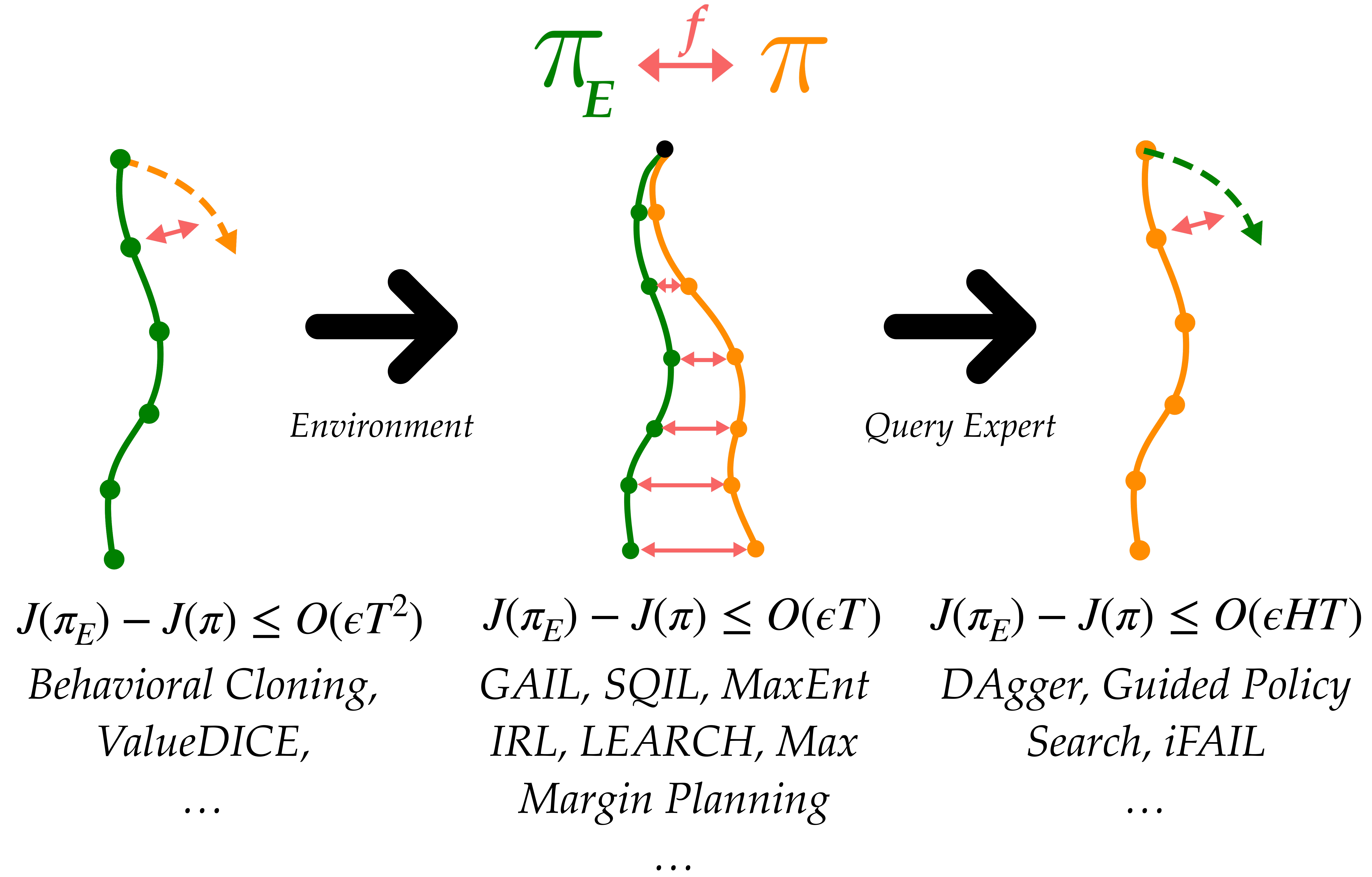
Q1: How well can we expect a learned policy to do in each of these three settings?

$$\min_{\pi} \quad J(\pi_E) - J(\pi)$$

$$\min_{\pi} \max_{f} f(\pi_E) - f(\pi)$$



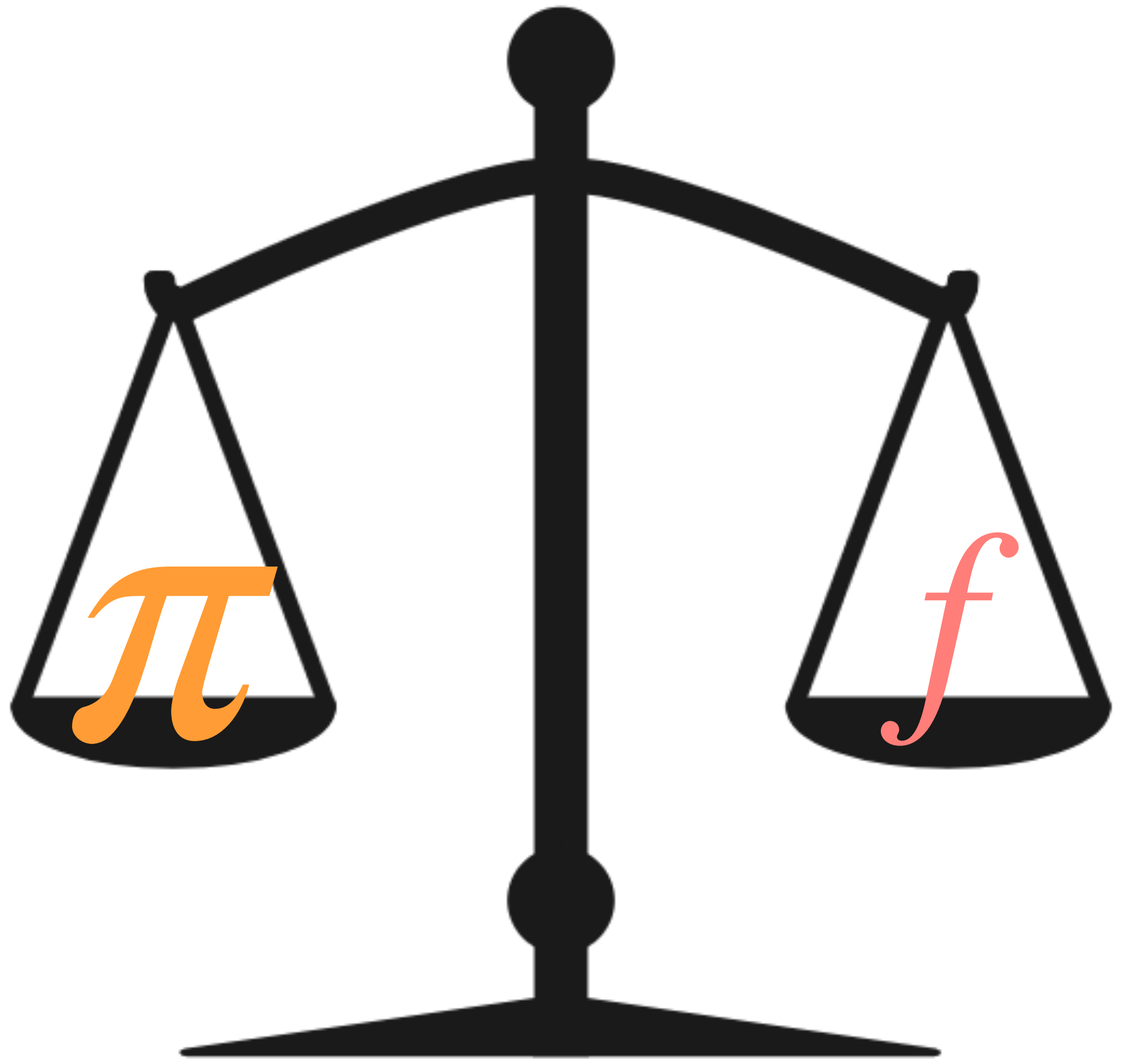
***Key Insight**: Each of these classes of approaches corresponds to solving a game with a different class of discriminators. Stronger feedback leads to more powerful discriminators and a tighter performance bound.*



$\pi_E \xleftrightarrow{f} \pi$
Environment
Query Expert
$J(\pi_E) - J(\pi) \leq O(\epsilon T^2)$
Behavioral Cloning, ValueDICE, …
$J(\pi_E) - J(\pi) \leq O(\epsilon T)$
GAIL, SQIL, MaxEnt IRL, LEARCH, Max Margin Planning …
$J(\pi_E) - J(\pi) \leq O(\epsilon HT)$
DAgger, Guided Policy Search, iFAIL …

Q2: How can we efficiently find a performant policy in each of these settings?

A:

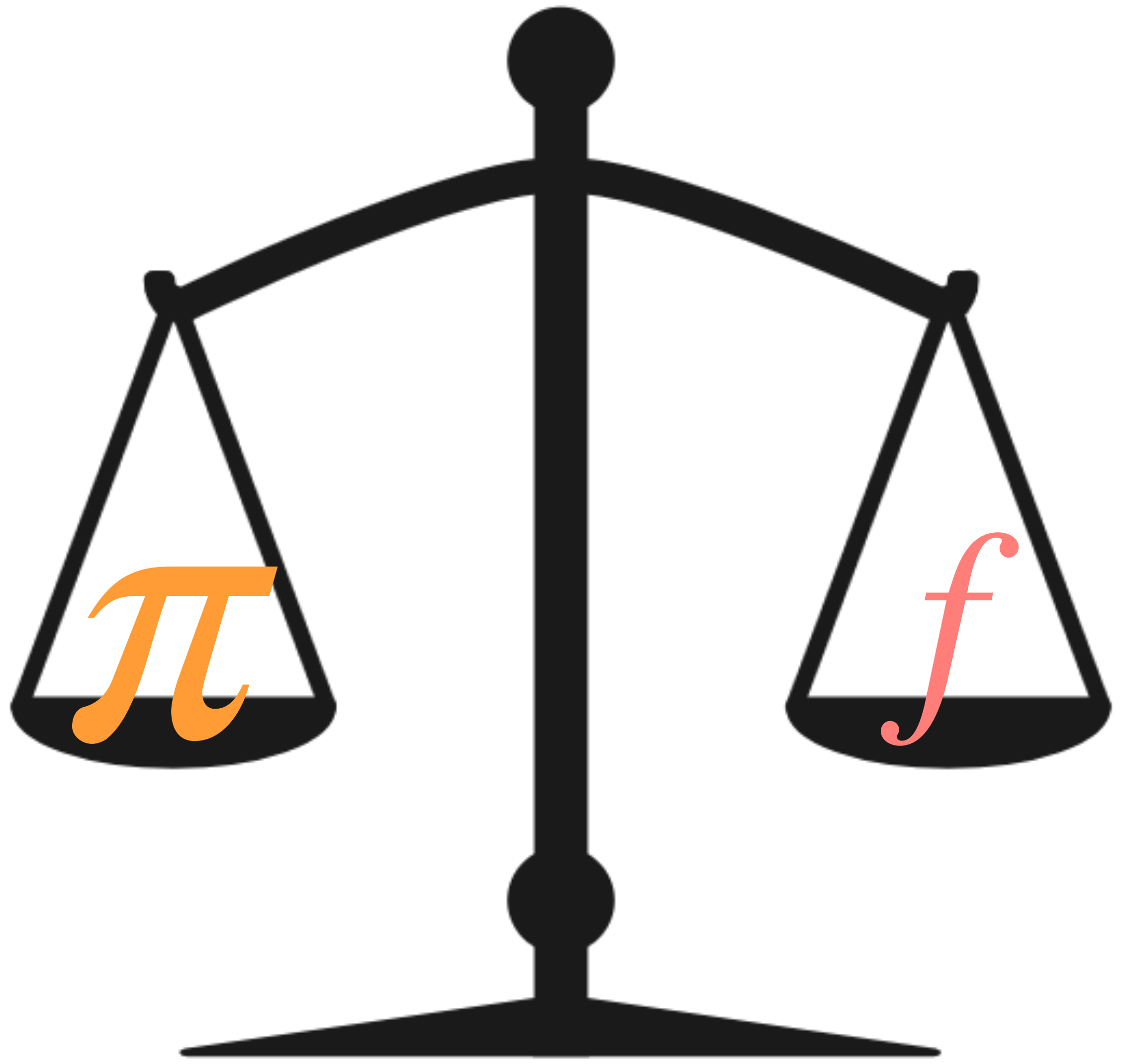


= *No Regret on* π vs. *Best Response on* f

= *Best Response on* π vs. *No Regret on* f

Q2: How can we efficiently find a performant policy in each of these settings?

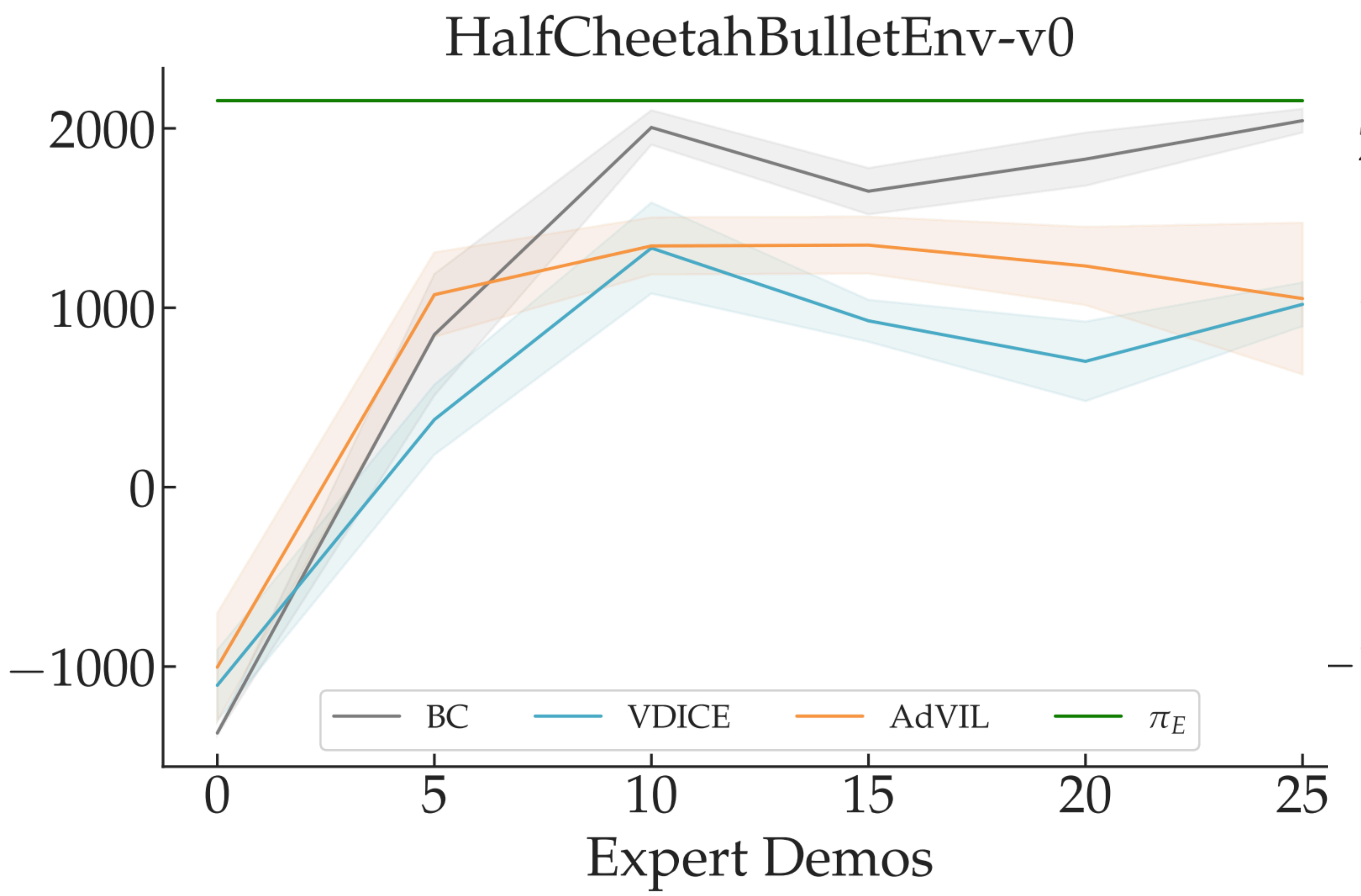
A:



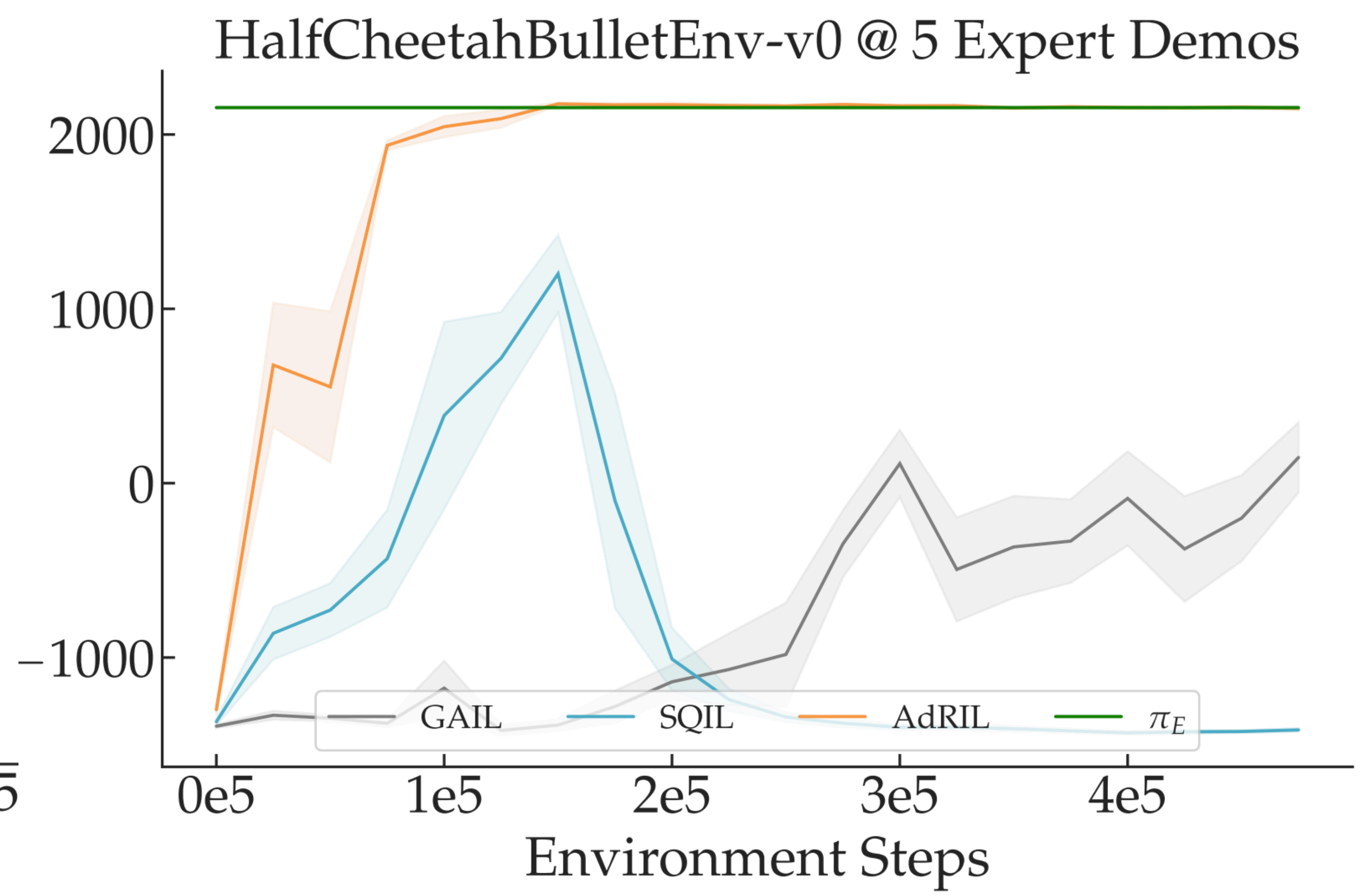
= ________ *on* π vs. *Best Response on* f

= *Best Response on* π vs. ________ *on* f

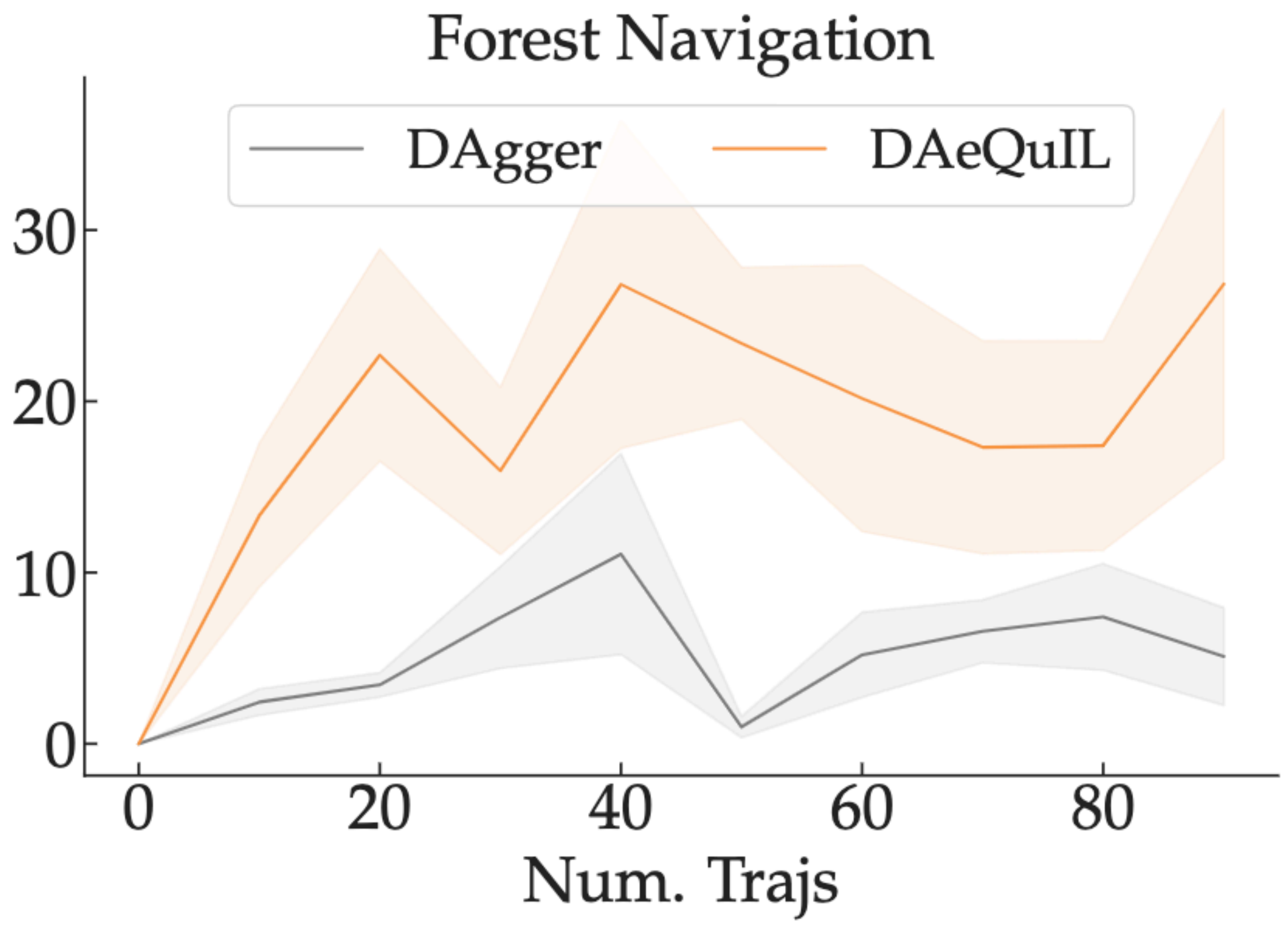
AdVIL



AdRIL



DAeQuIL



Of Moments and Matching:

A Game-Theoretic Framework for Closing the Imitation Gap

Gokul Swamy, Sanjiban Choudhury, Drew Bagnell, Steven Wu

https://gokul.dev/mmil/