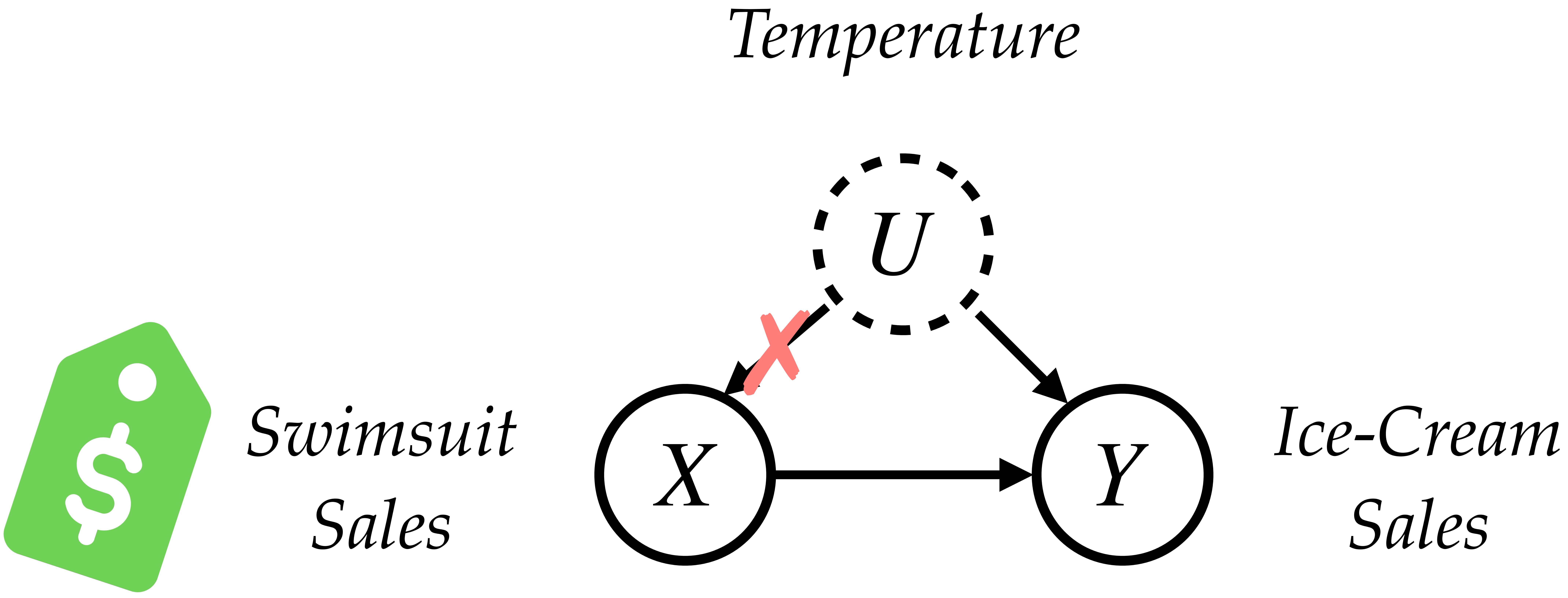
Causal Confounds in Sequential Decision Making

Gokul Swamy

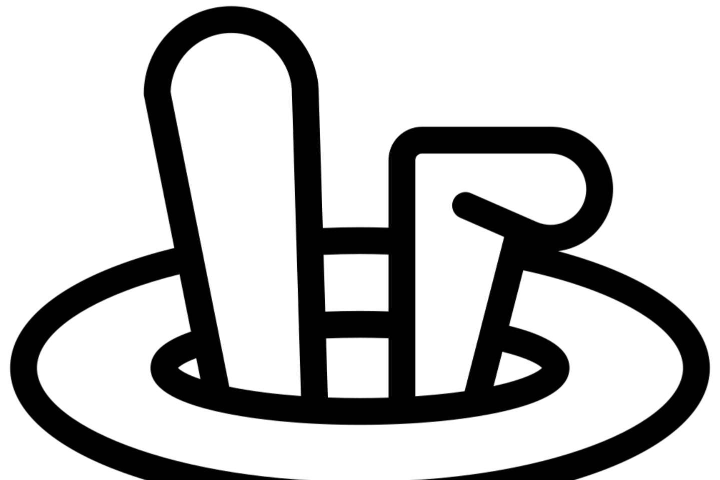


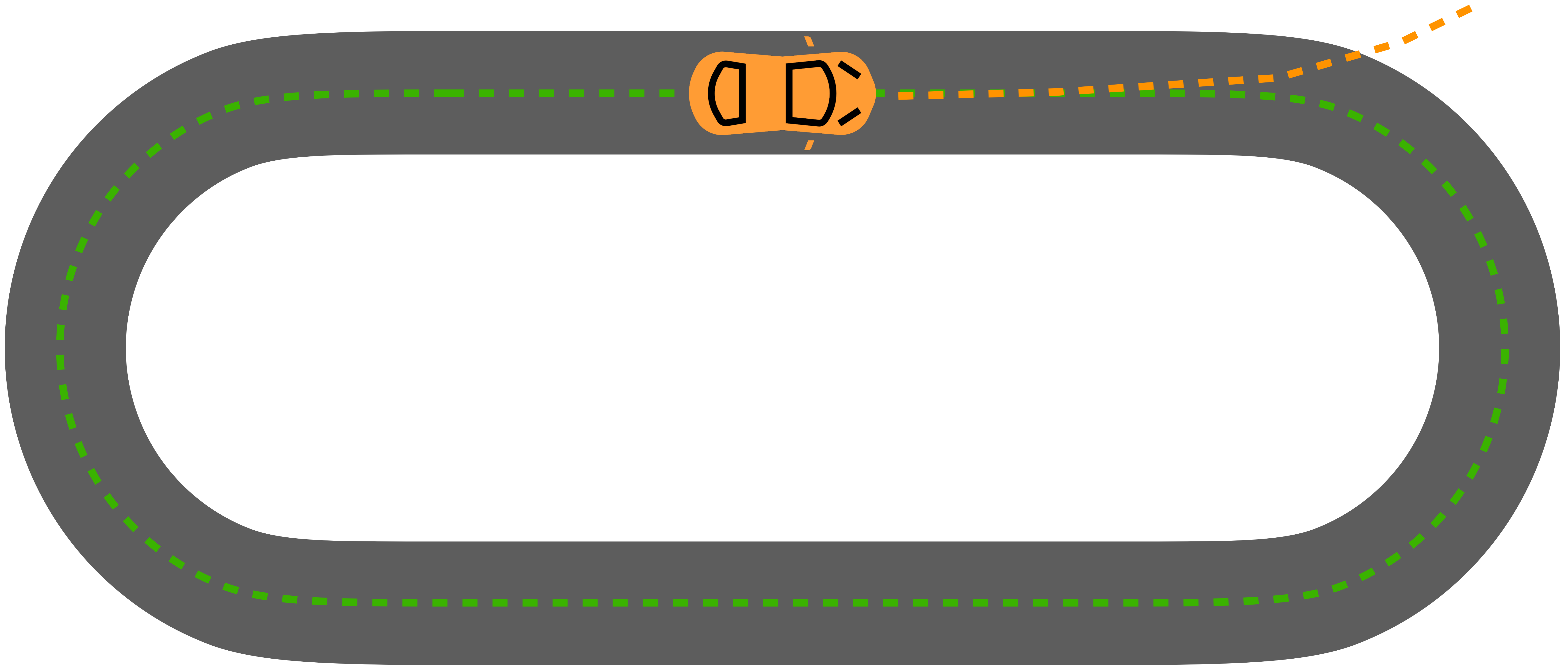

(joint work with Sanjiban Choudhury, Drew Bagnell, Steven Wu)



Temperature
U
Swimsuit Sales
X
Y
Ice-Cream Sales

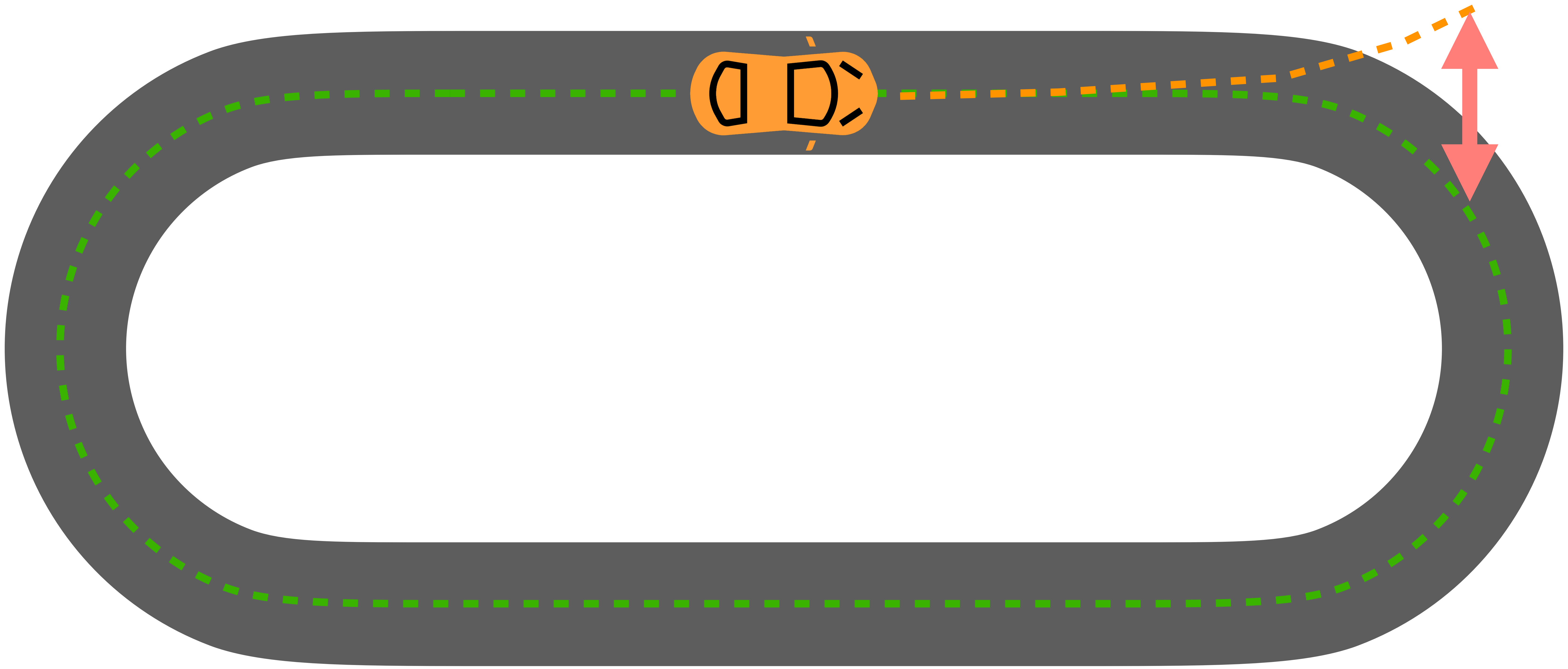
Interventions happen via interaction with the environment in sequential decision making.





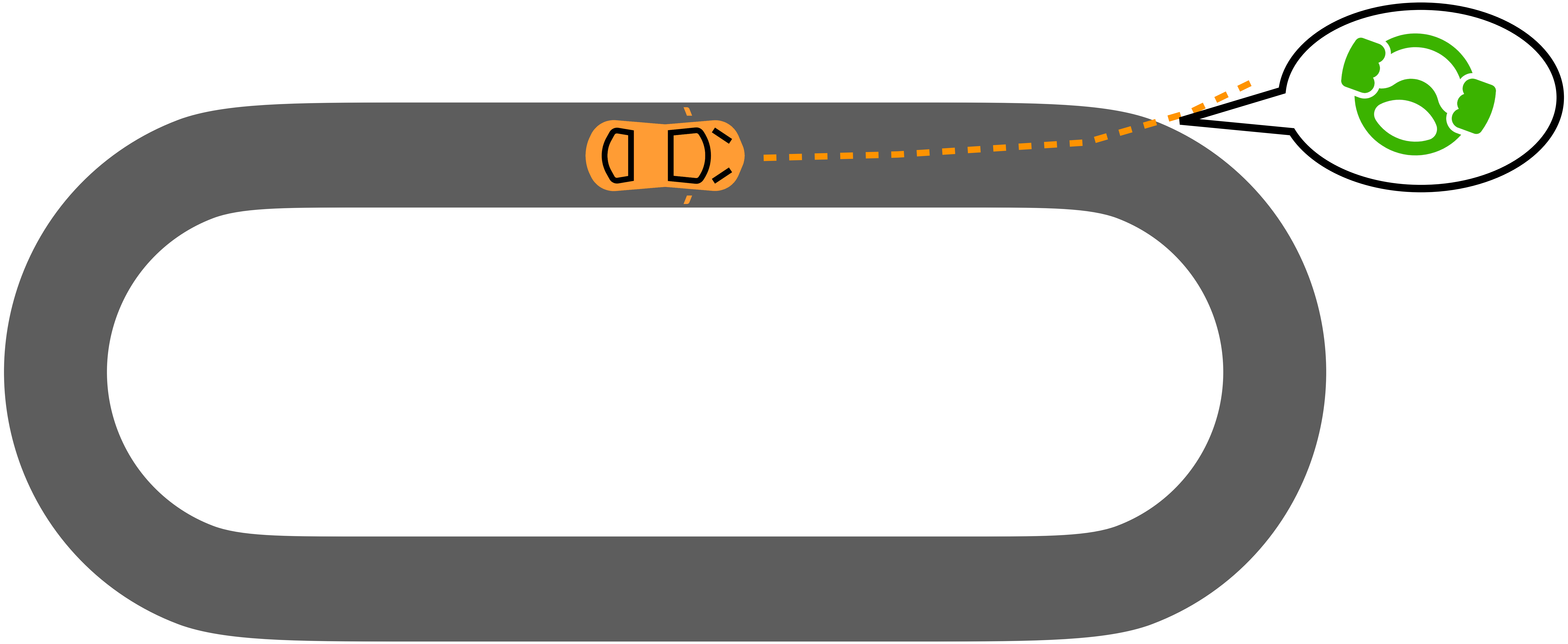
$$\{s_1 \ldots s_n\} \mapsto \{a_1 \ldots a_n\}$$

Behavioral Cloning



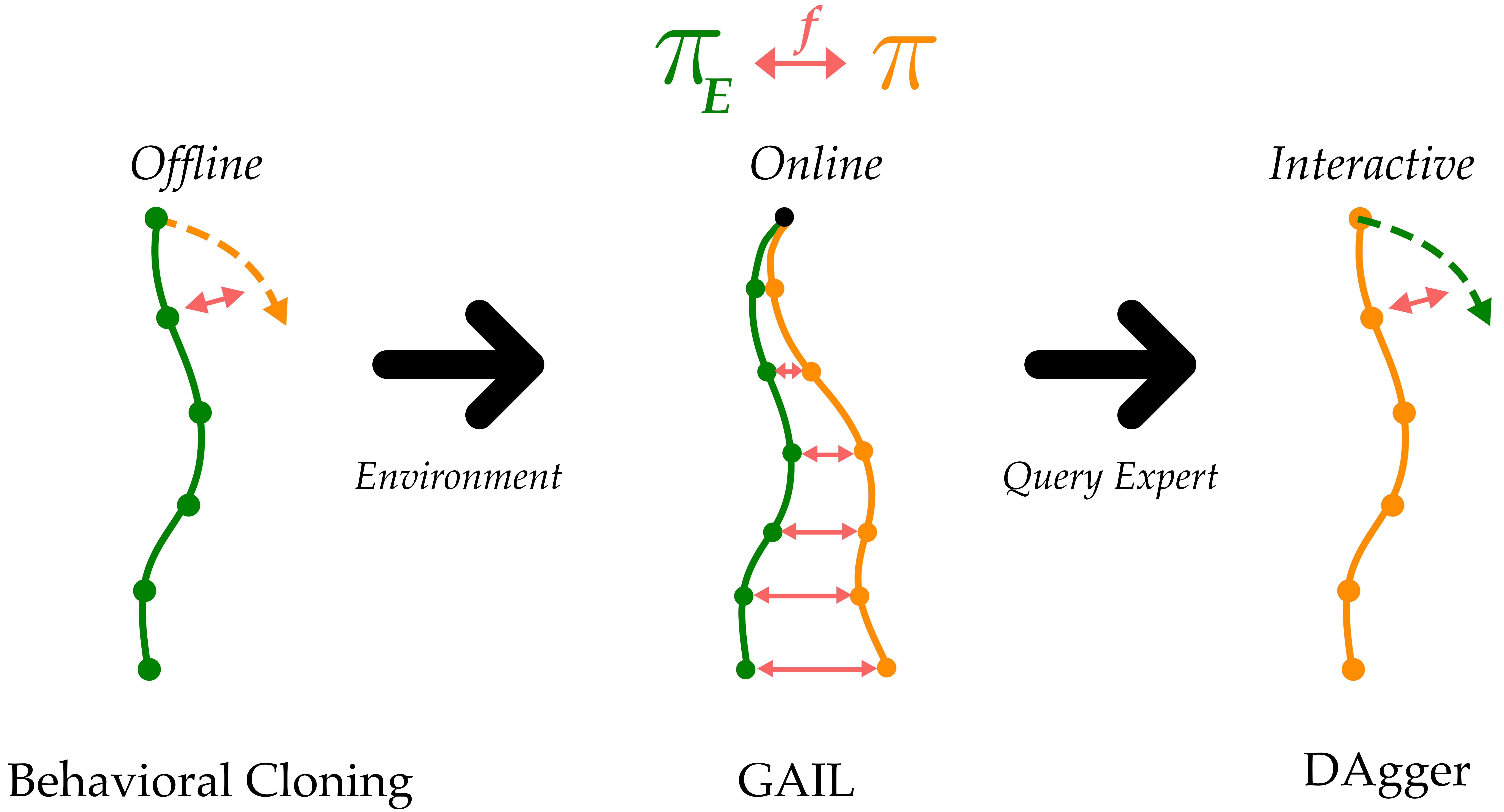
$$\{s_1 \ldots s_n\} \quad \{s_1 \ldots s_n\}$$
$$\{a_1 \ldots a_n\} \longleftrightarrow \{a_1 \ldots a_n\}$$

MaxEnt IRL / GAIL

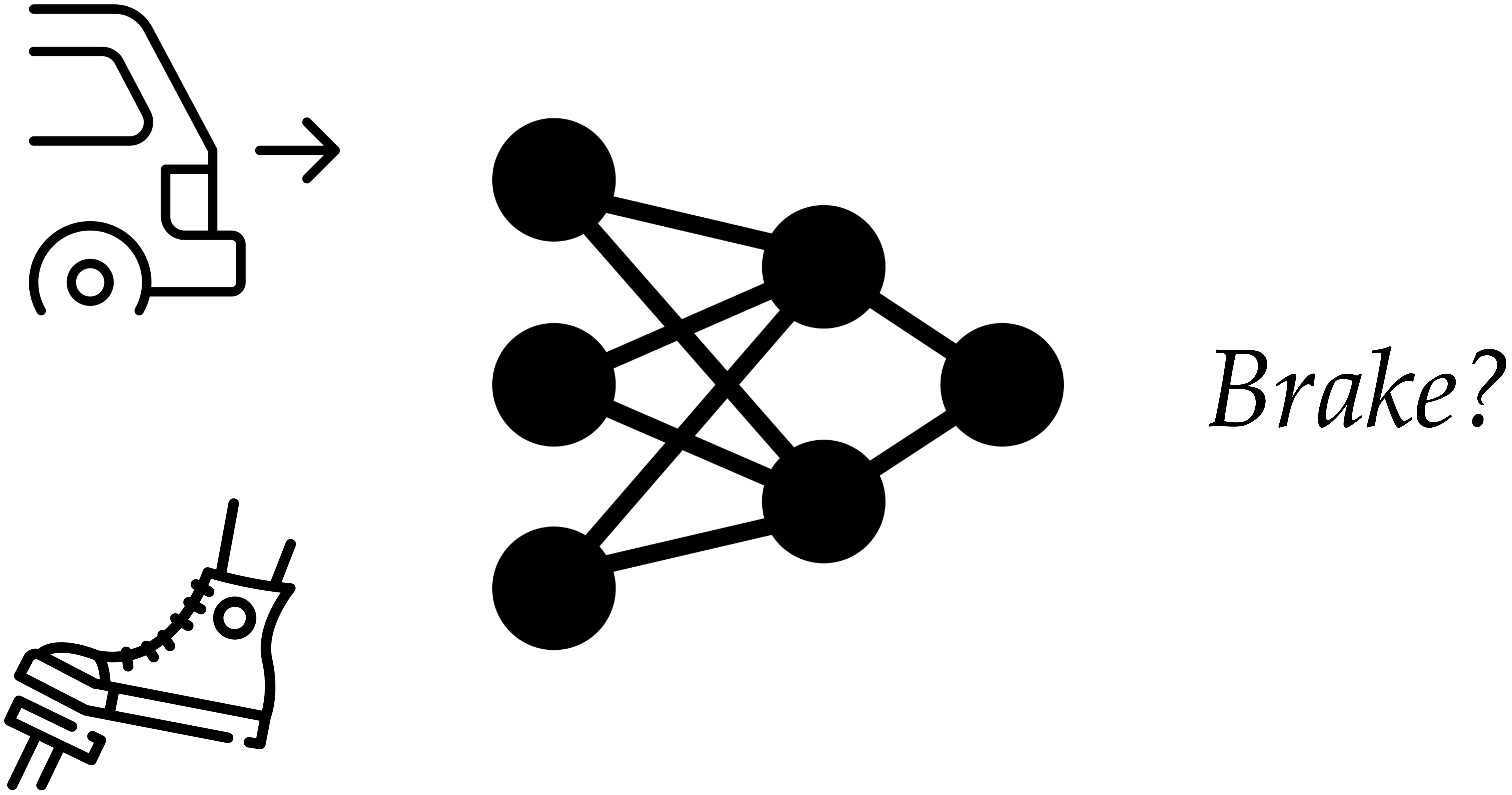


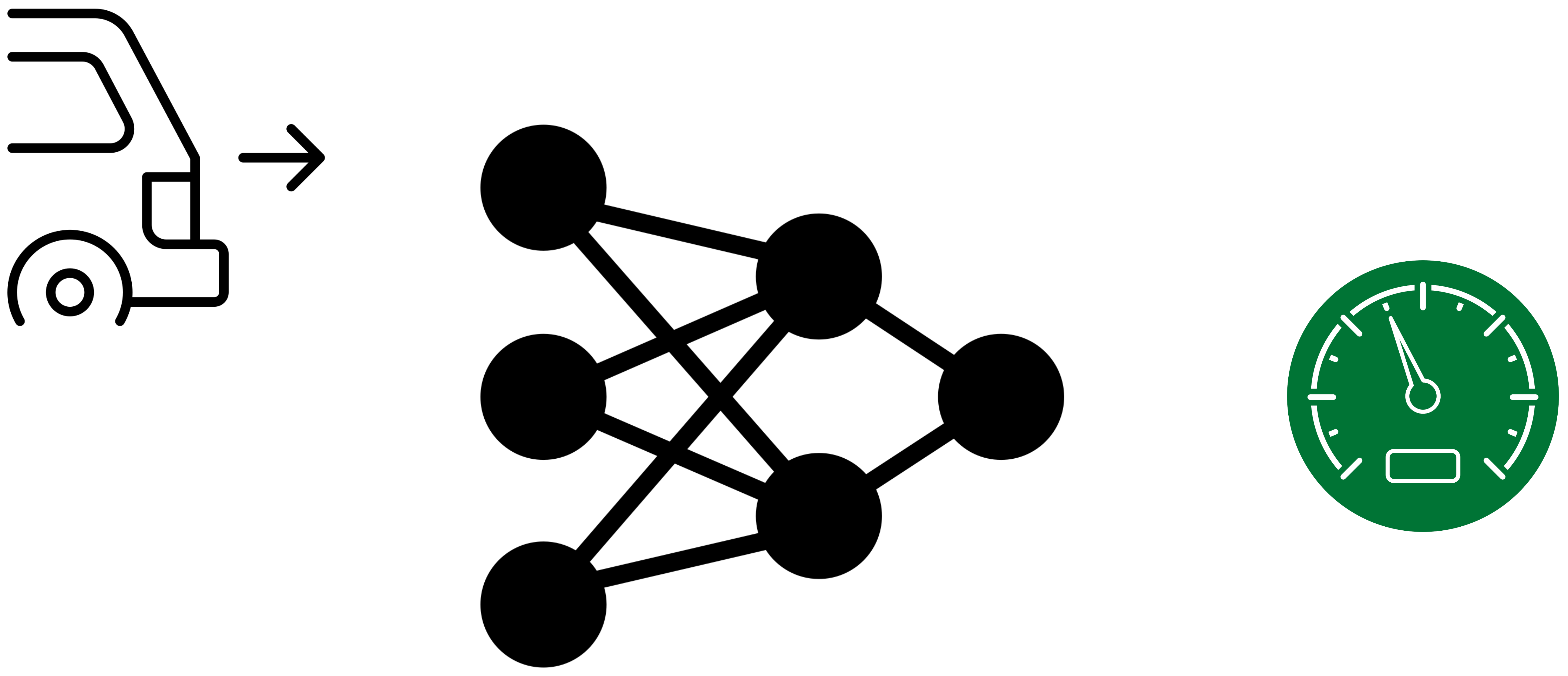
$$\{s_1 \ldots s_n\} \mapsto \{a_1 \ldots a_n\}$$

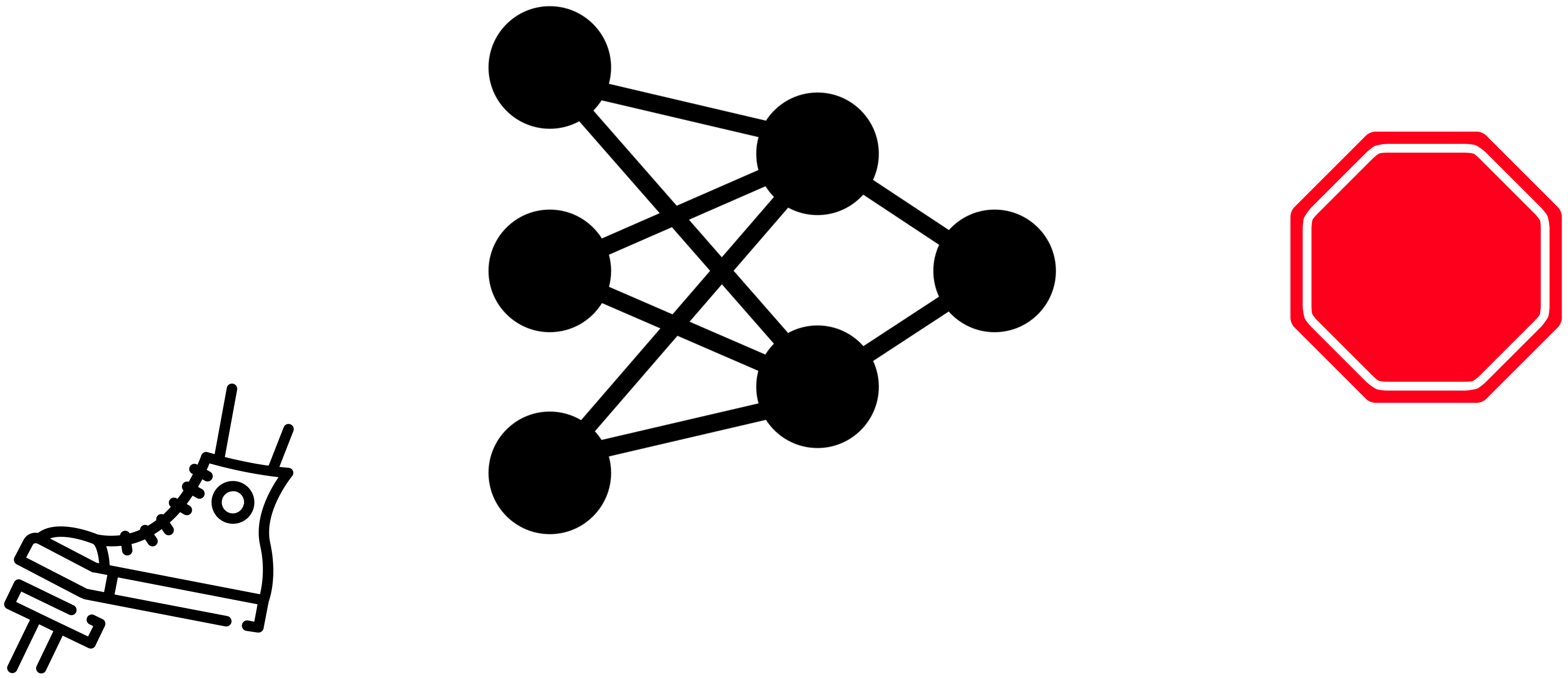
DAgger

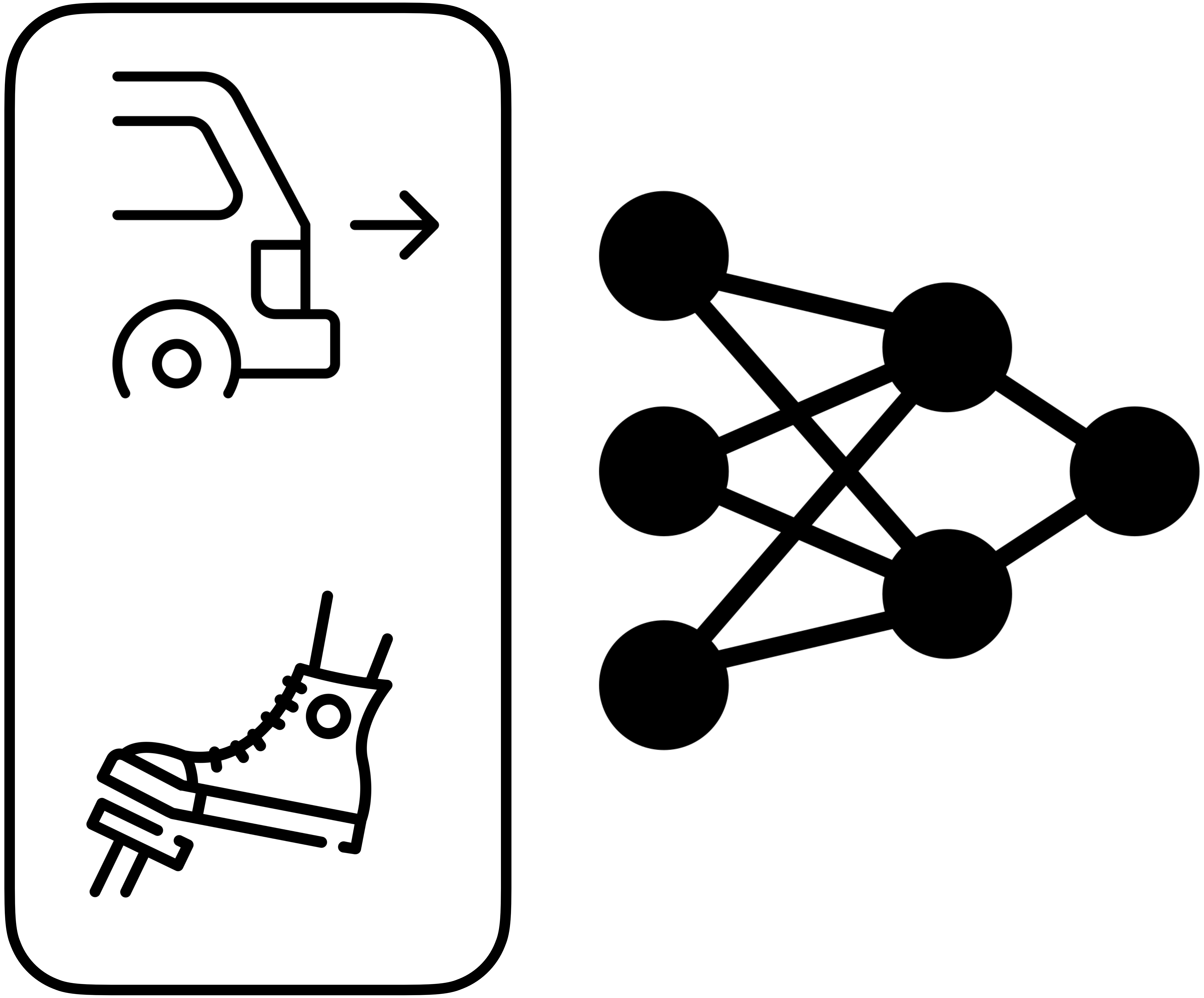


$\pi_E \xleftrightarrow{f} \pi$
Offline
Online
Interactive
Environment
Query Expert
Behavioral Cloning
GAIL
DAgger


Brake?



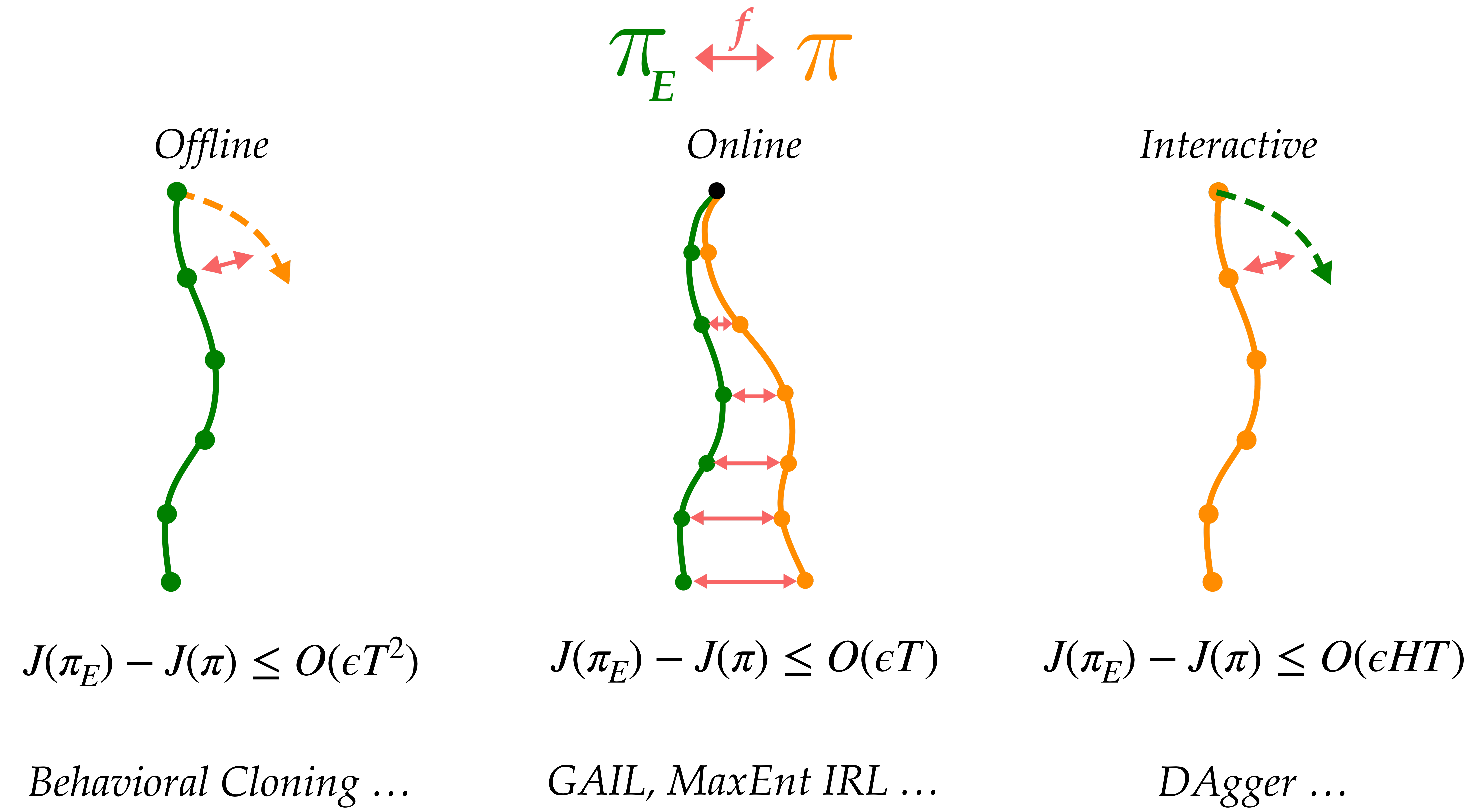




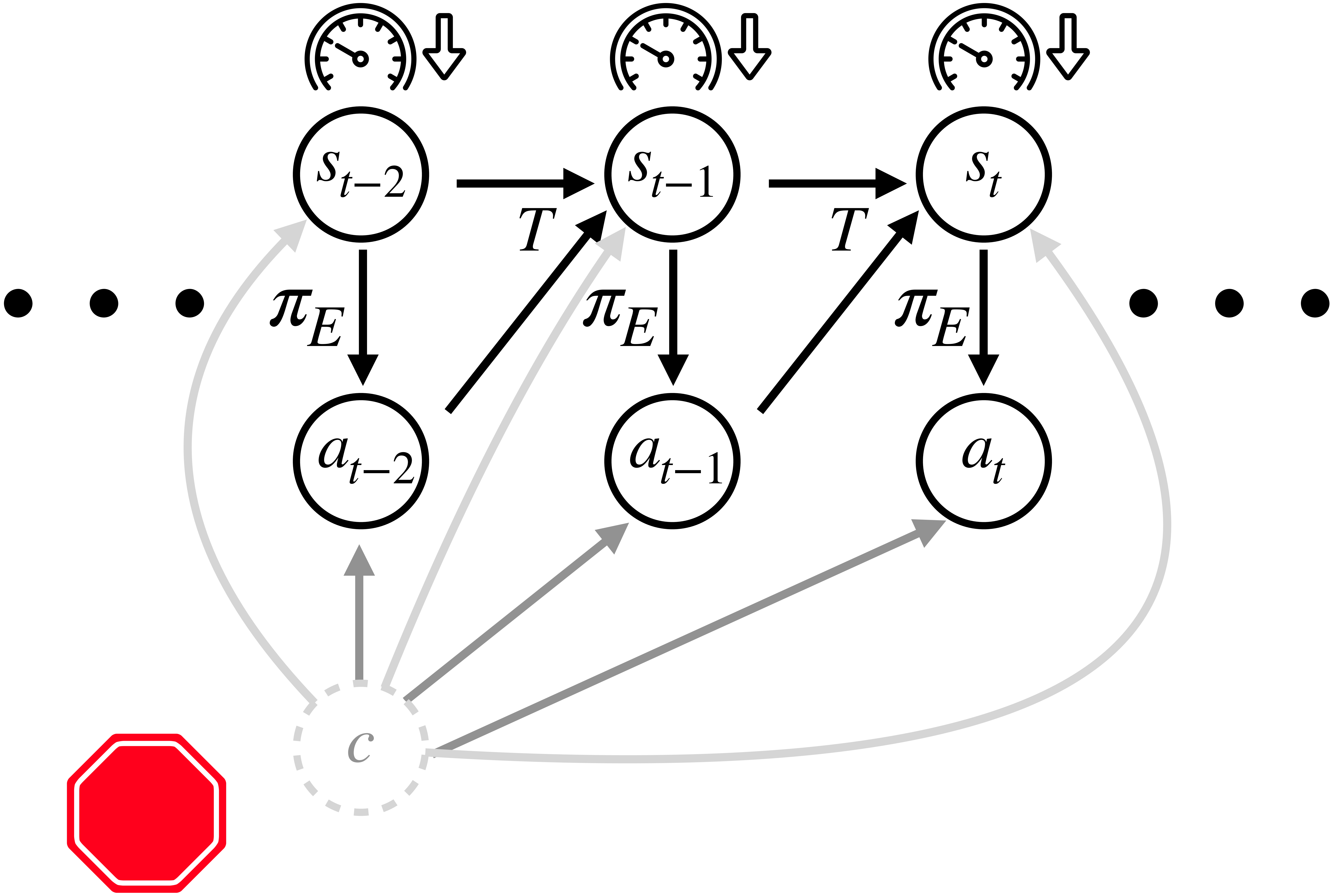
Q: Would DAgger fix this problem?

A: Yes, it's just covariate shift?

	Offline	Online	Interactive
Covariate Shift	✘	✔	✔
Hidden Context			
TCN			

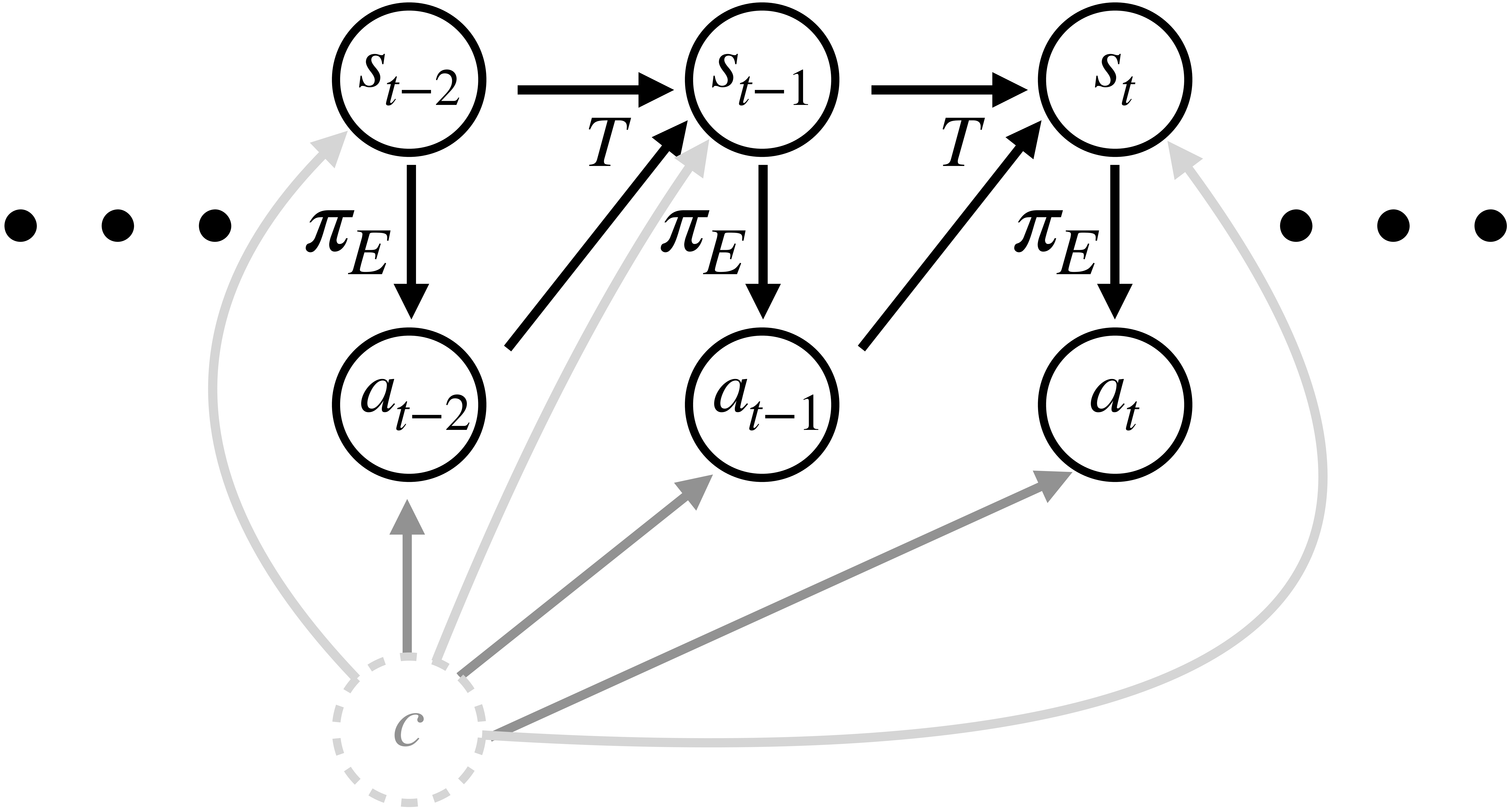

$\pi_E \xleftrightarrow{f} \pi$
Offline
Online
Interactive
$J(\pi_E) - J(\pi) \leq O(\epsilon T^2)$
$J(\pi_E) - J(\pi) \leq O(\epsilon T)$
$J(\pi_E) - J(\pi) \leq O(\epsilon HT)$
Behavioral Cloning …
GAIL, MaxEnt IRL …
DAgger …

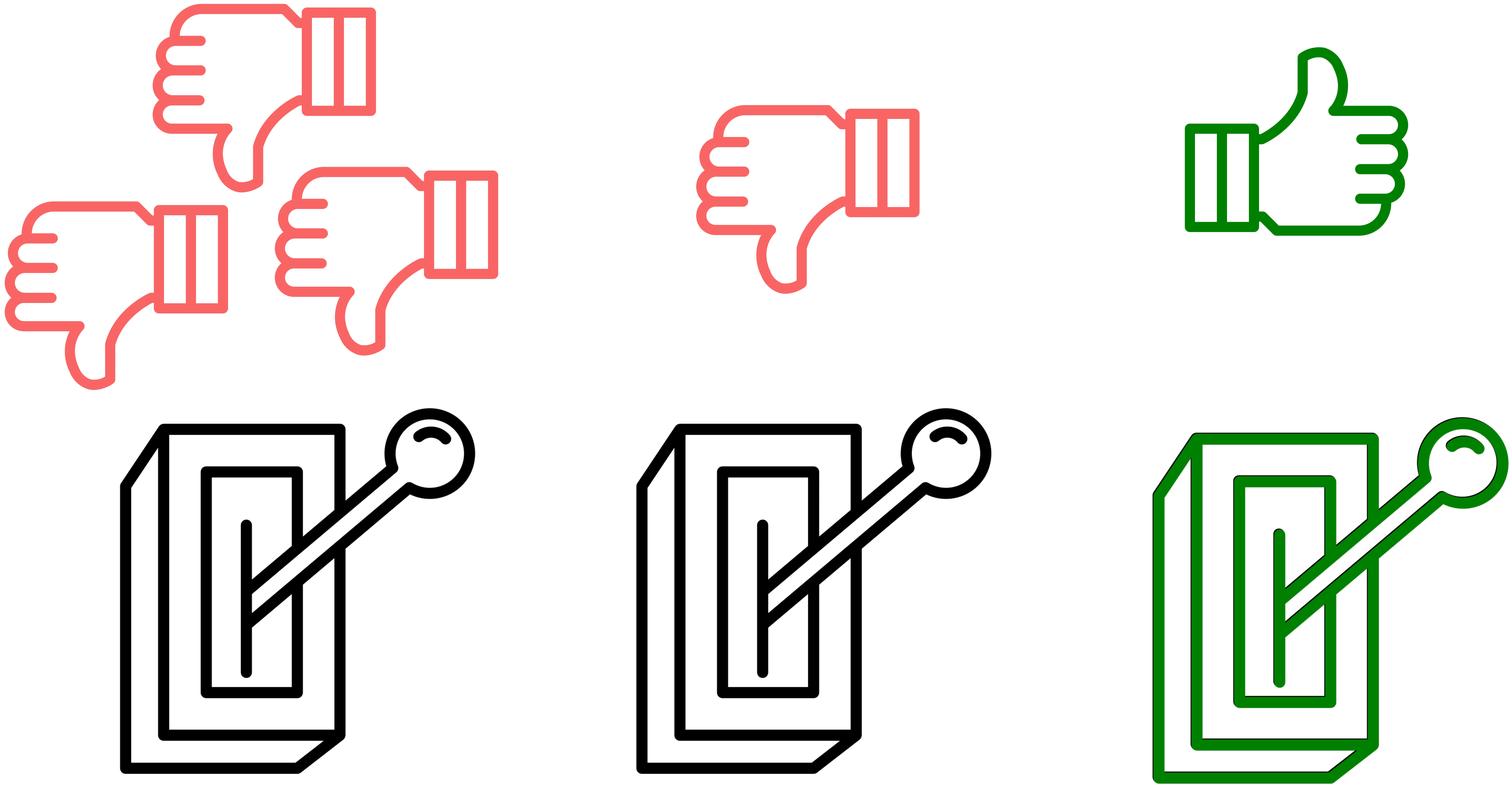
"Hence, a system trained with multiple frames would merely predict a steering angle equal to the current rate of turn as observed through the camera. This would lead to catastrophic behavior in test mode. The robot would simply turn in circles."
— Muller et al., 2006

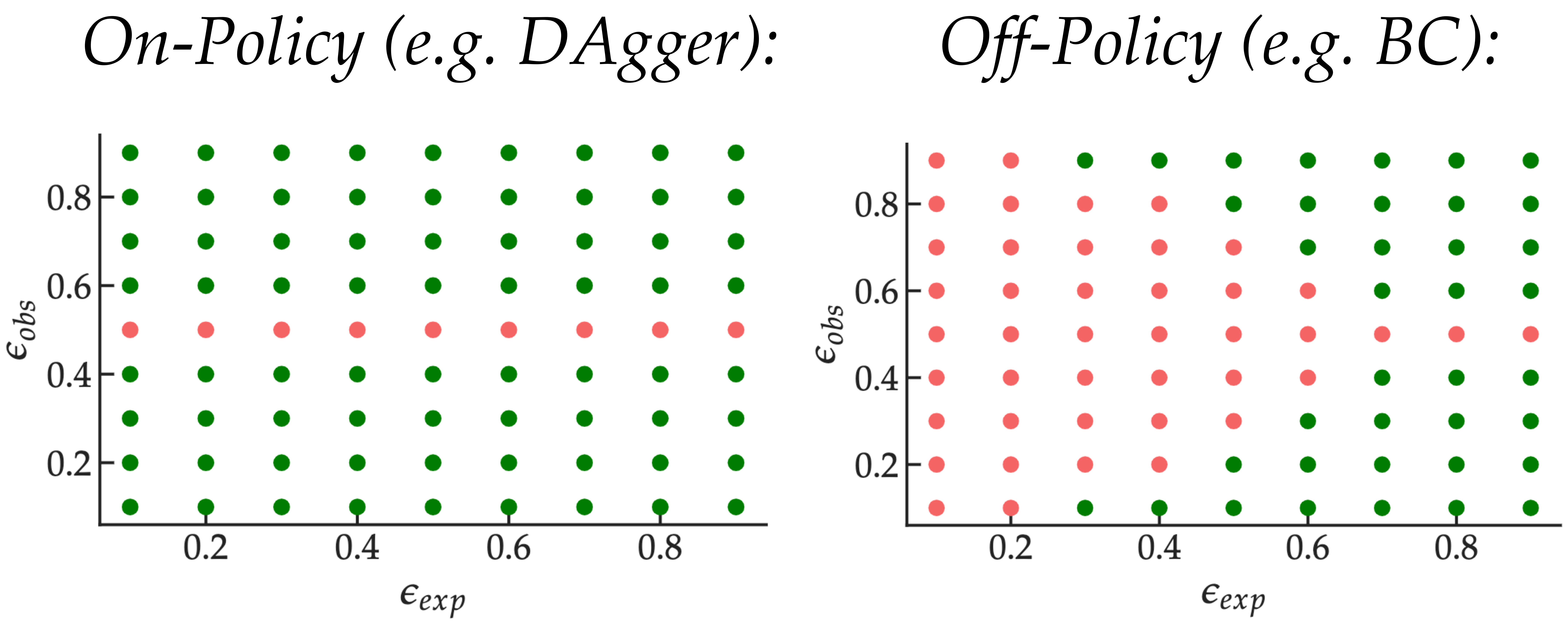


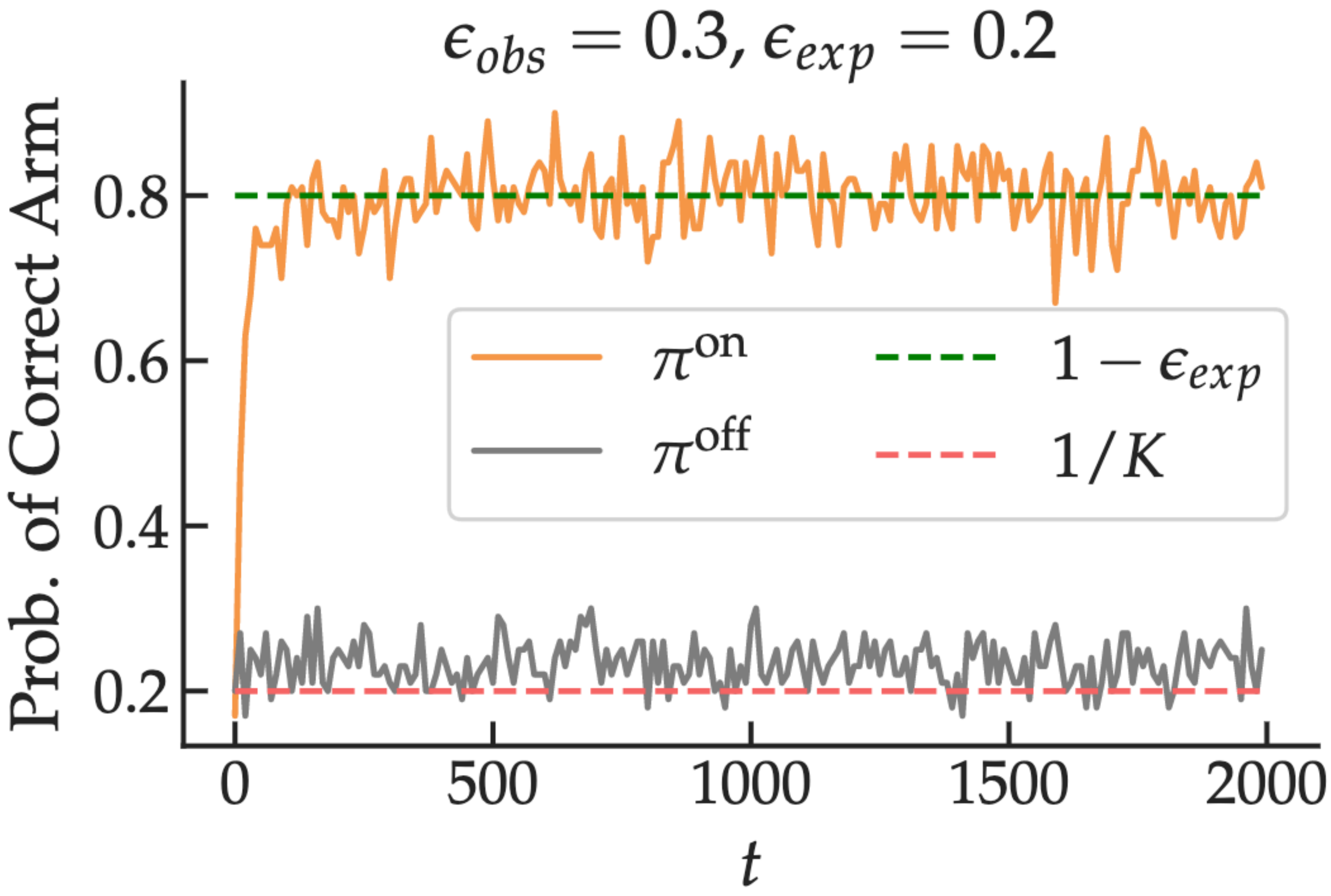
s_{t-2}
s_{t-1}
s_t
T
T
π_E
π_E
π_E
a_{t-2}
a_{t-1}
a_t
c

	MDP	POMDP
State	s_t	$p(s_t, c \mid s_1, a_1 \ldots s_{t-1}, a_{t-1})$
Policy	$\pi(\,\cdot \mid s_t)$	$\pi(\,\cdot \mid s_1, a_1 \ldots s_{t-1}, a_{t-1})$

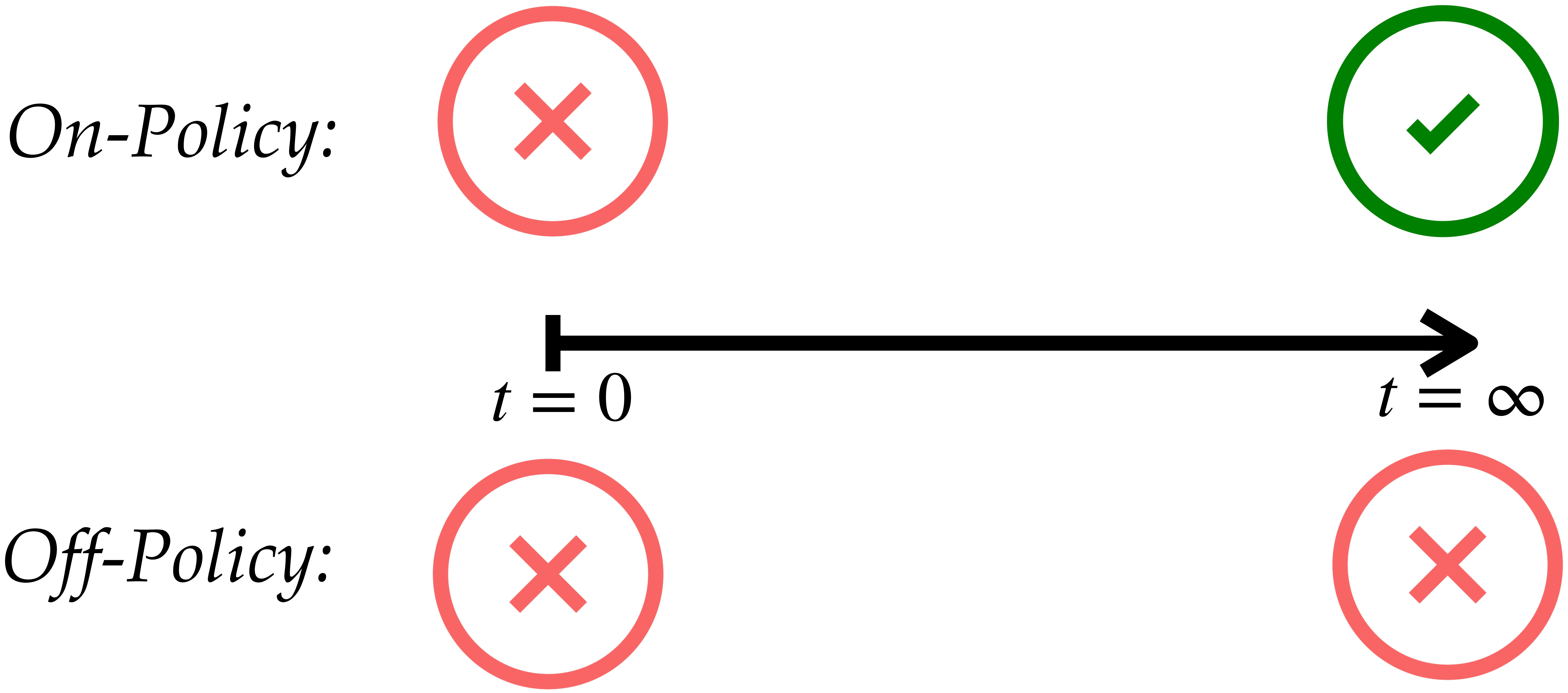

s_{t-2}
s_{t-1}
s_t
T
T
π_E
π_E
π_E
a_{t-2}
a_{t-1}
a_t
c




On-Policy (e.g. DAgger):
Off-Policy (e.g. BC):
ϵ_{obs}
0.2
0.4
0.6
0.8
ϵ_{exp}



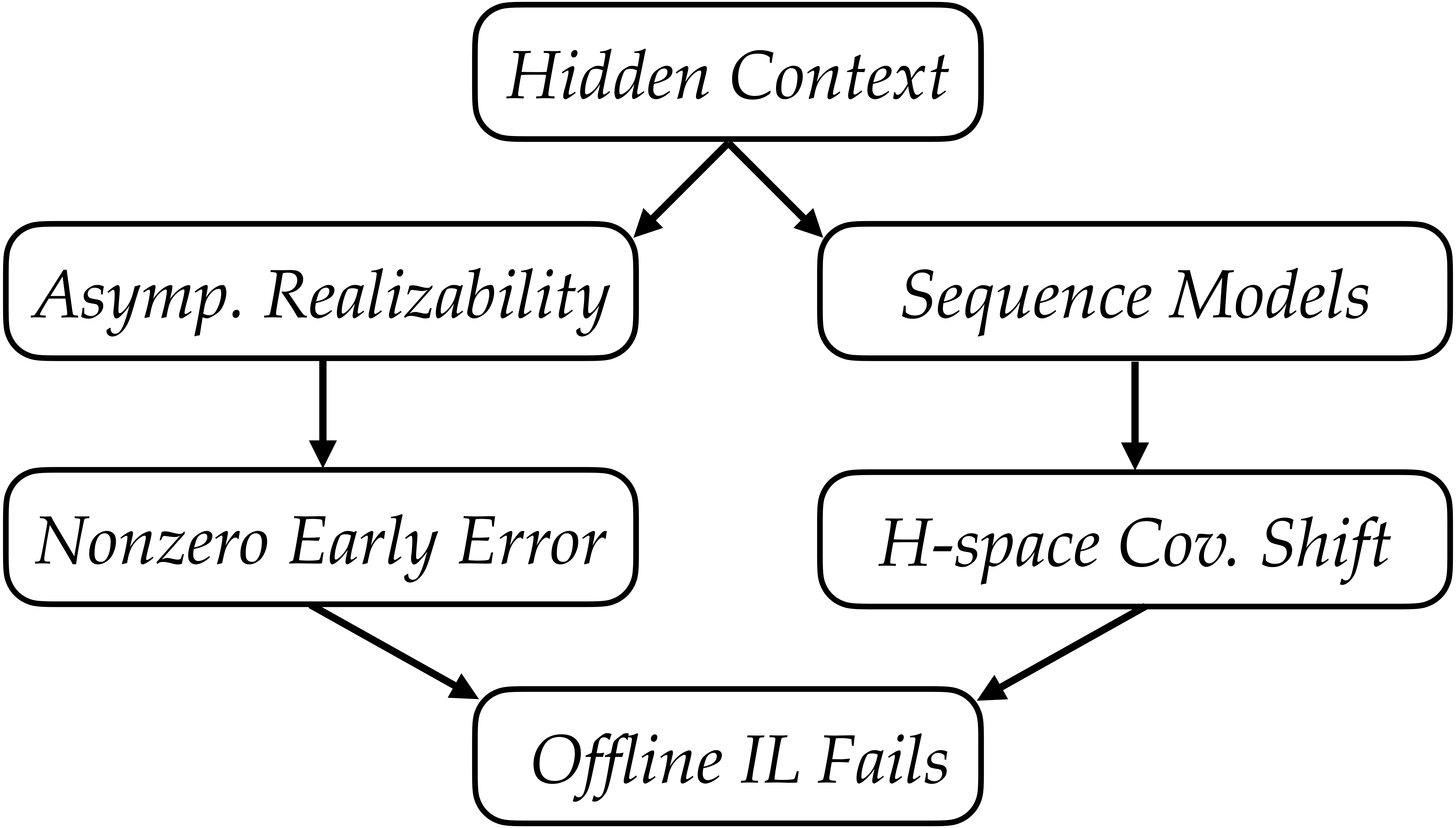
$\epsilon_{obs} = 0.3, \epsilon_{exp} = 0.2$
Prob. of Correct Arm
0.2
0.4
0.6
0.8
0
500
1000
1500
2000
t
π^{on}
π^{off}
$1 - \epsilon_{exp}$
$1/K$

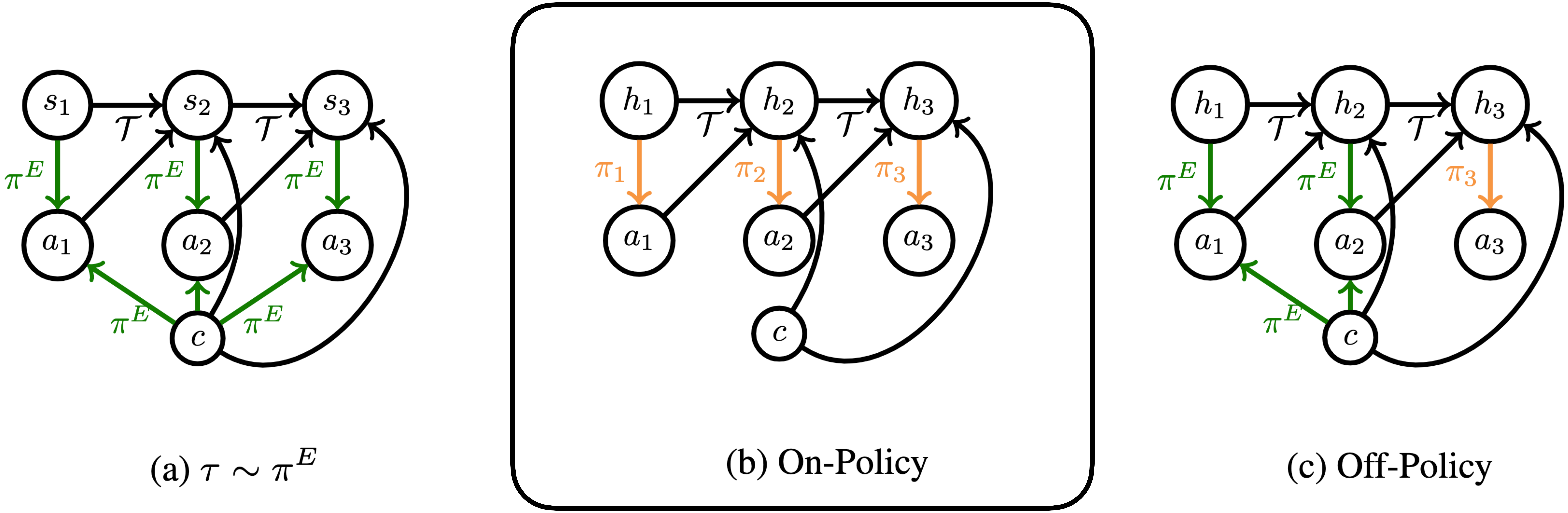

On-Policy:
Off-Policy:
$t = 0$
$t = \infty$

Train-time: $\pi(a_t \mid h_t) \approx p(a_t^E \mid s_1^E, a_1^E, \ldots, s_t^E)$

Test-time: $p(a_t^E \mid s_1, a_1, \ldots, s_t)$

It's just covariate shift in the space of histories!


Hidden Context
Asymp. Realizability
Sequence Models
Nonzero Early Error
H-space Cov. Shift
Offline IL Fails



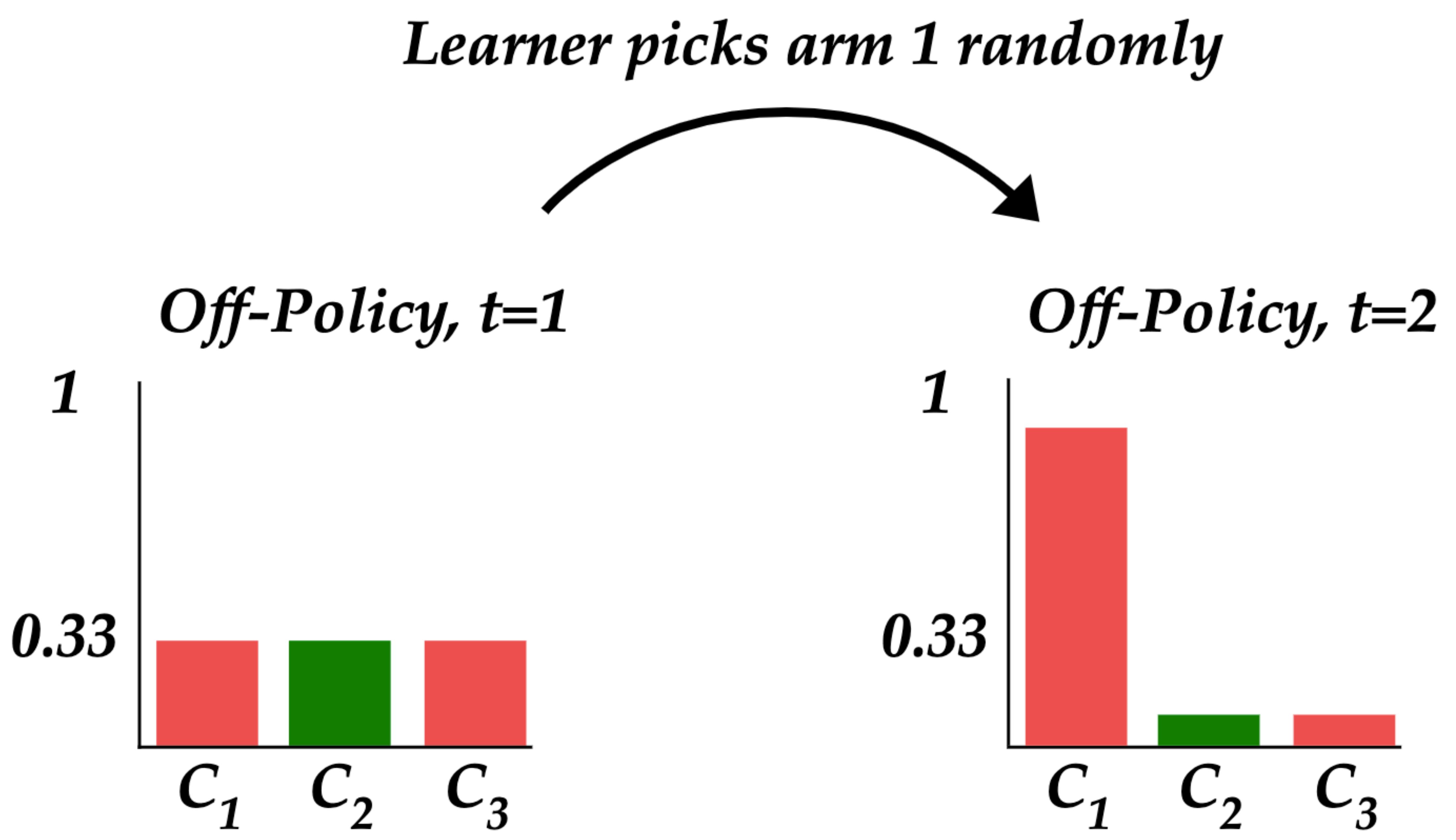
(a) $\tau \sim \pi^E$

(b) On-Policy

(c) Off-Policy

$$p_{\text{on}}(c, h_t) \propto p(\tau; \pi) \propto p(c)p(s_1)\prod_{i=1}^{t-1}\mathcal{T}(s_{i+1} \mid s_i, a_i, c)$$

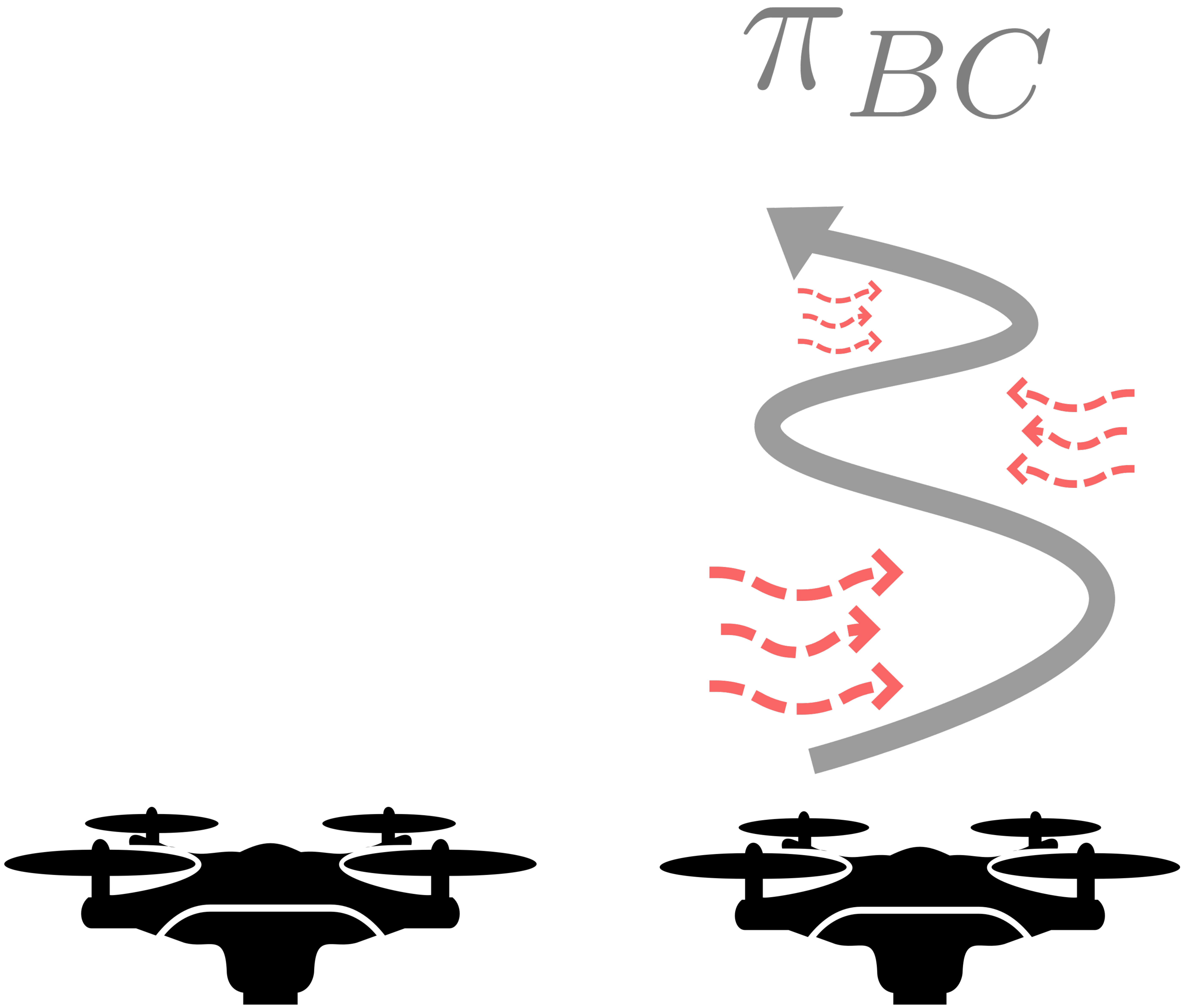
$$p_{\text{off}}(c, h_t) \propto p(\tau; \pi^E) \propto p(c)p(s_1)\prod_{i=1}^{t-1}{\color{green}\pi^E(a_i \mid c, s_i)}\mathcal{T}(s_{i+1} \mid s_i, a_i, c)$$

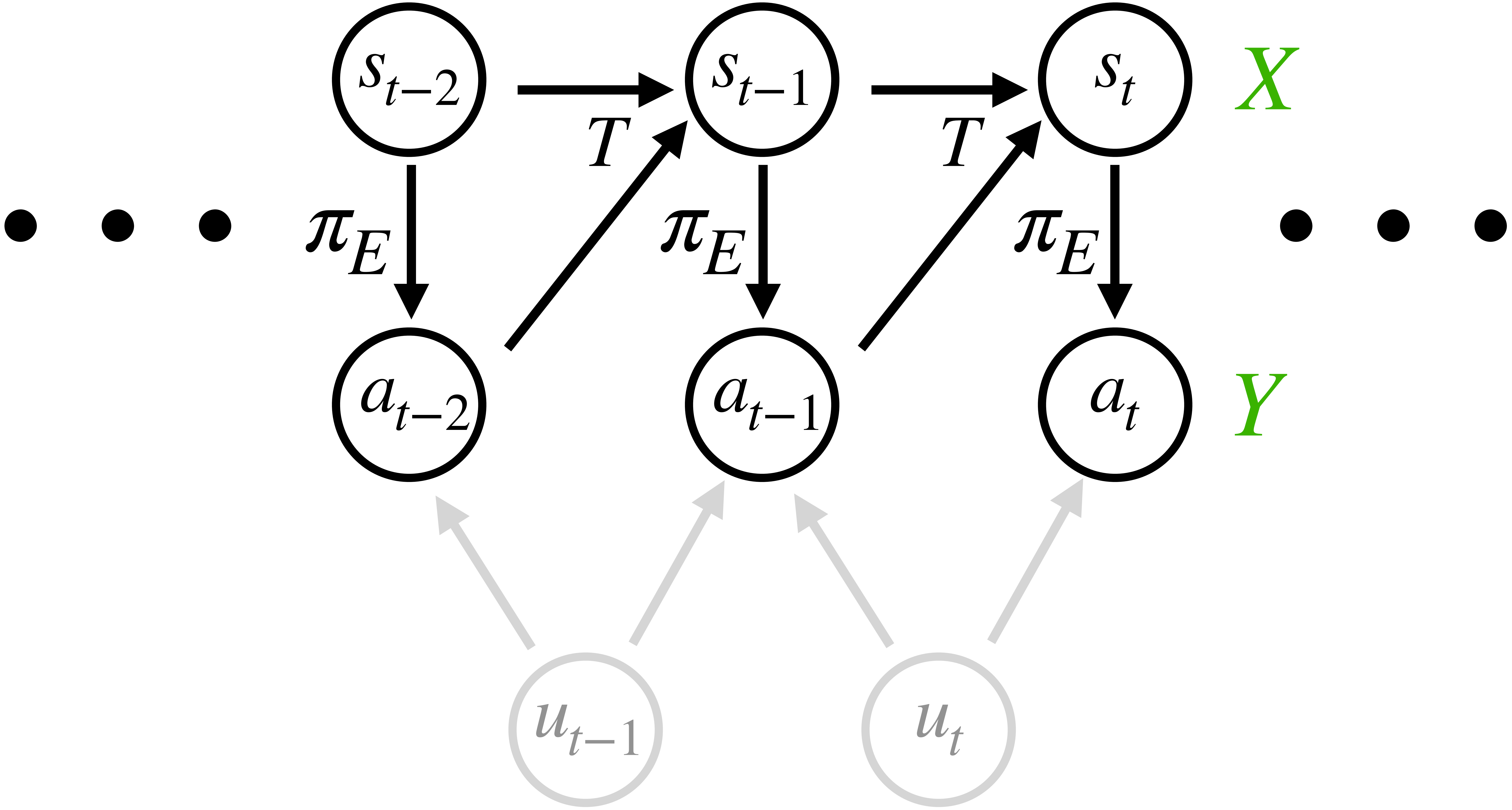


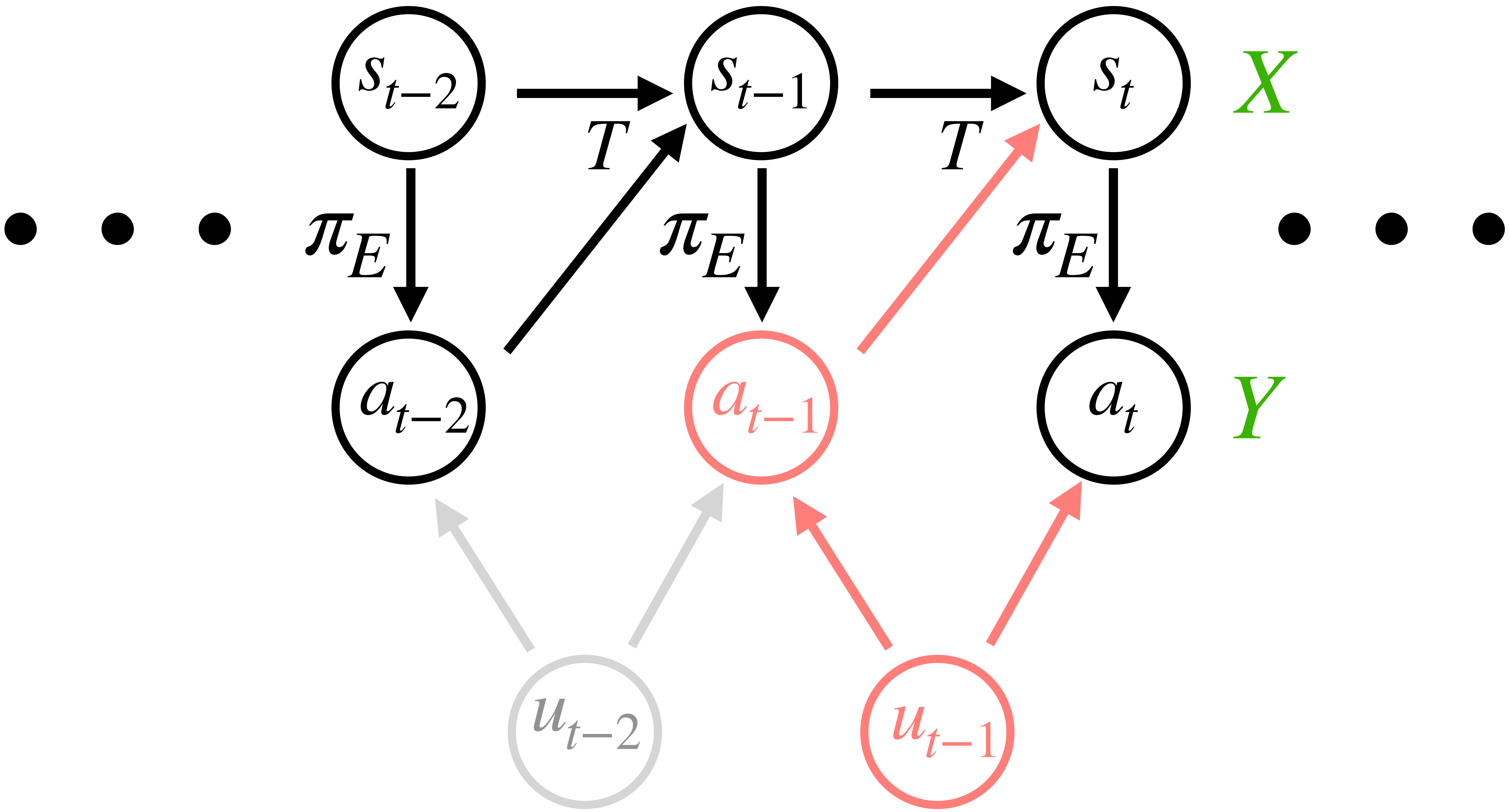
Theorem (informal): Off-policy learners have a value difference to the expert bounded by the sum of their errors (tight) while on-policy learners have one dependent on their asymptotic error.

	Offline	Online	Interactive
Covariate Shift	✘	✔	✔
Hidden Context	✘	✔ w/ History	✔ w/ History
TCN			

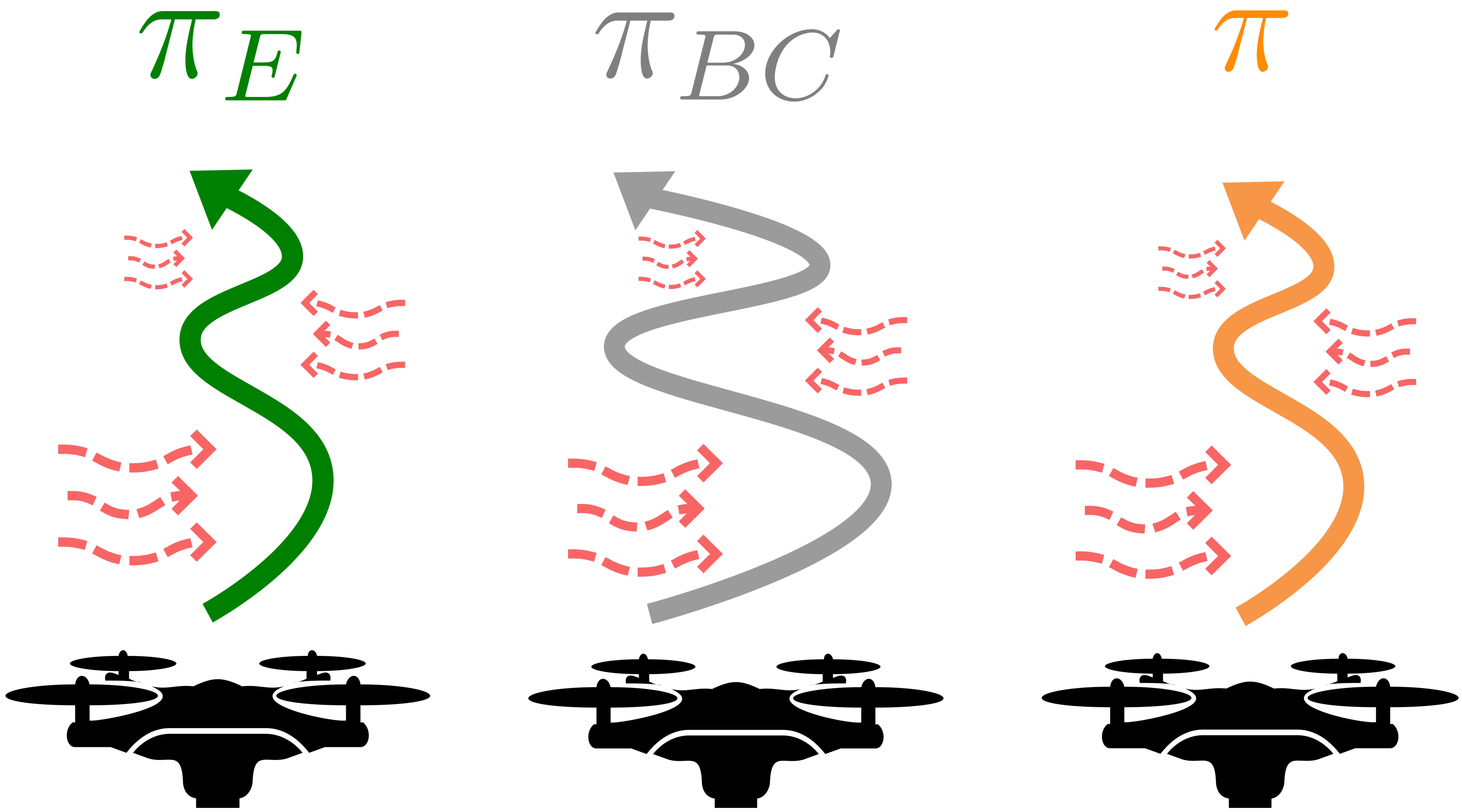
“Actually, since we were fitting a model to a time-series, samples tend to be correlated in time […] Thus, when leaving out a sample in cross validation, we actually left out a large window (16 seconds) of data around that sample, to diminish this bias.”
— Ng et al., 2003

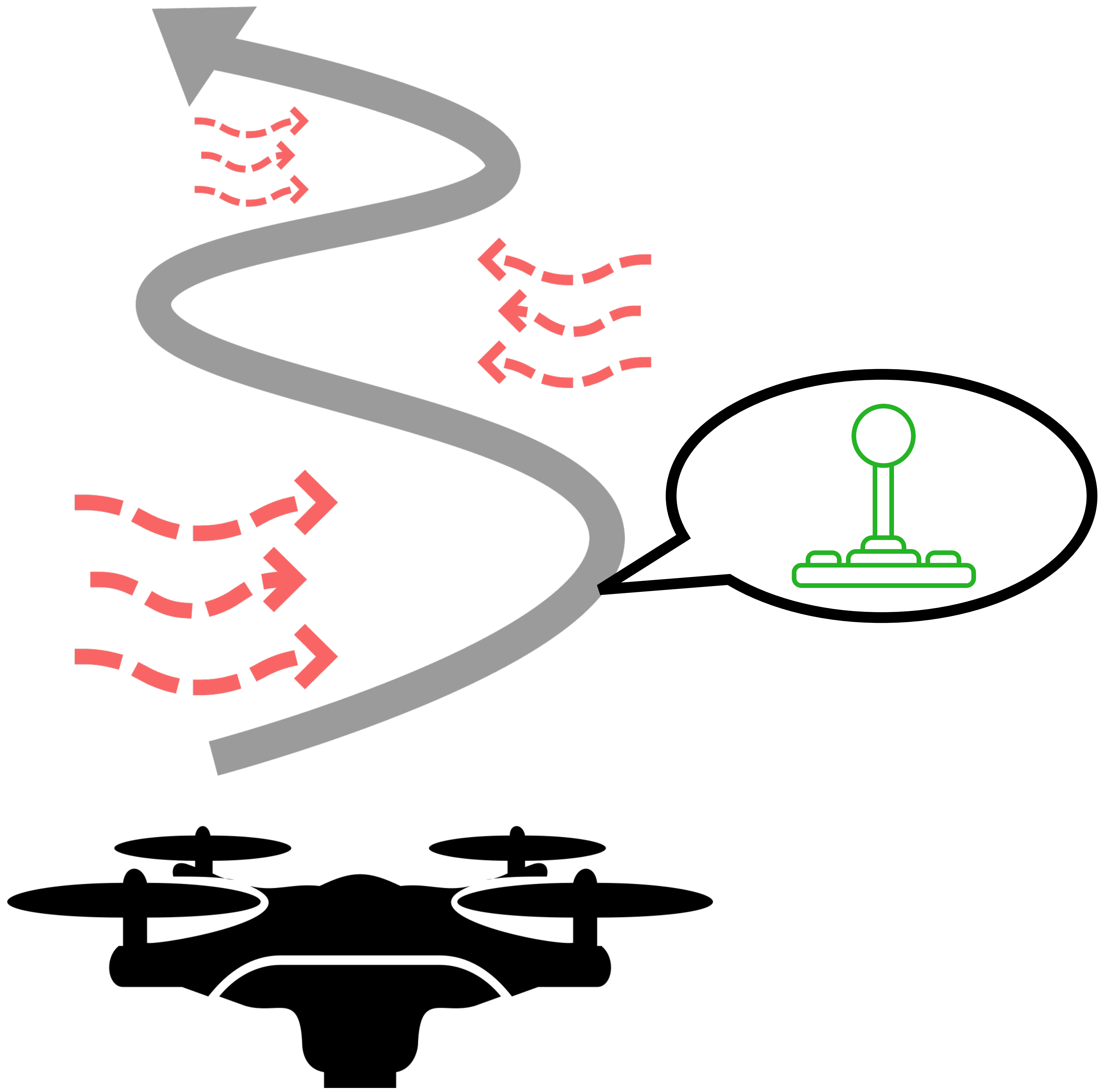

π_{BC}

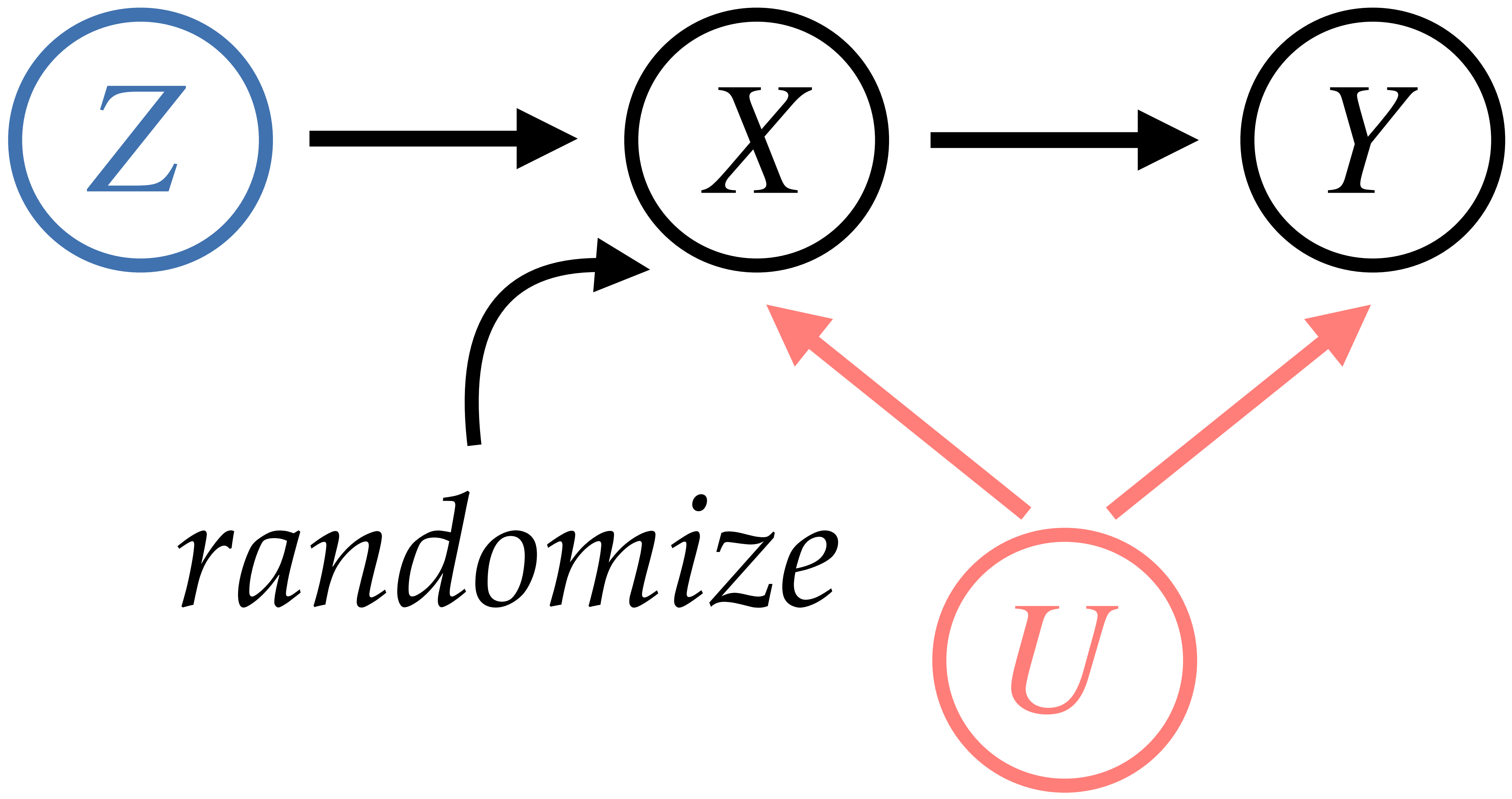

s_{t-2}
s_{t-1}
s_t
X
T
T
π_E
π_E
π_E
a_{t-2}
a_{t-1}
a_t
Y
u_{t-1}
u_t



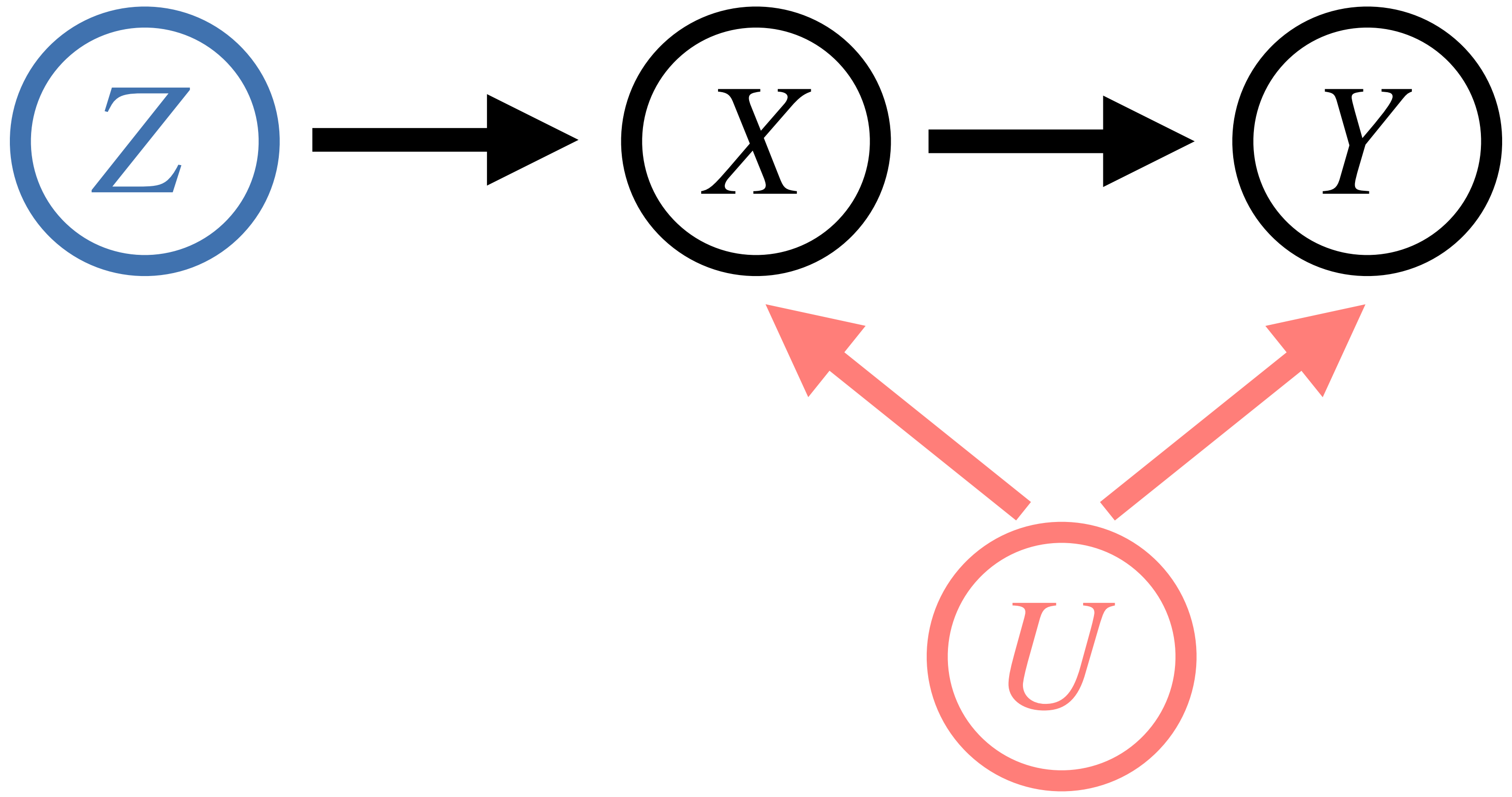
s_{t-2}
s_{t-1}
s_t
X
T
T
π_E
π_E
π_E
a_{t-2}
a_{t-1}
a_t
Y
u_{t-2}
u_{t-1}


π_E
π_{BC}
π



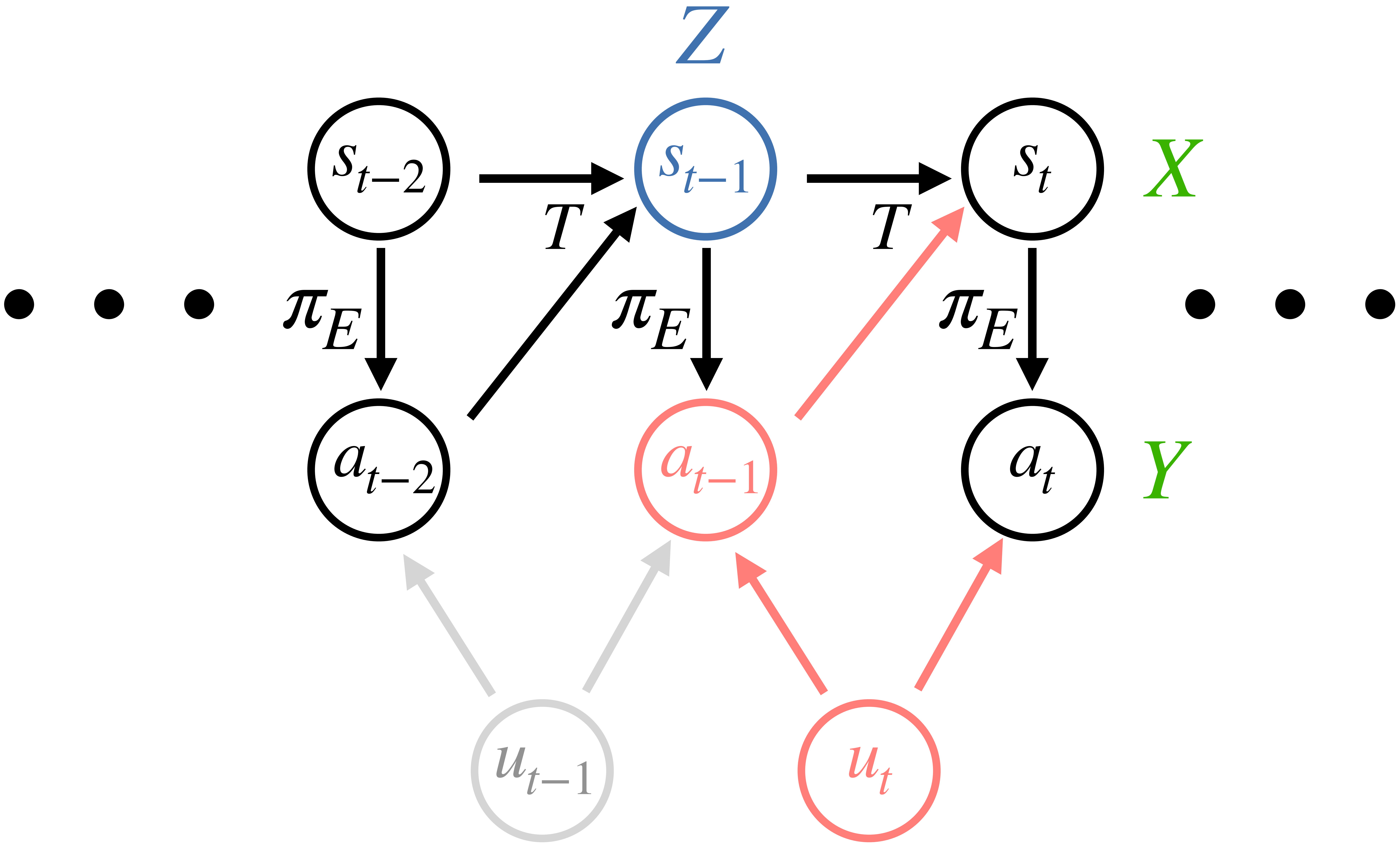


Key Idea: *We can condition on instrument Z to counter the effect of confounder U on X.*

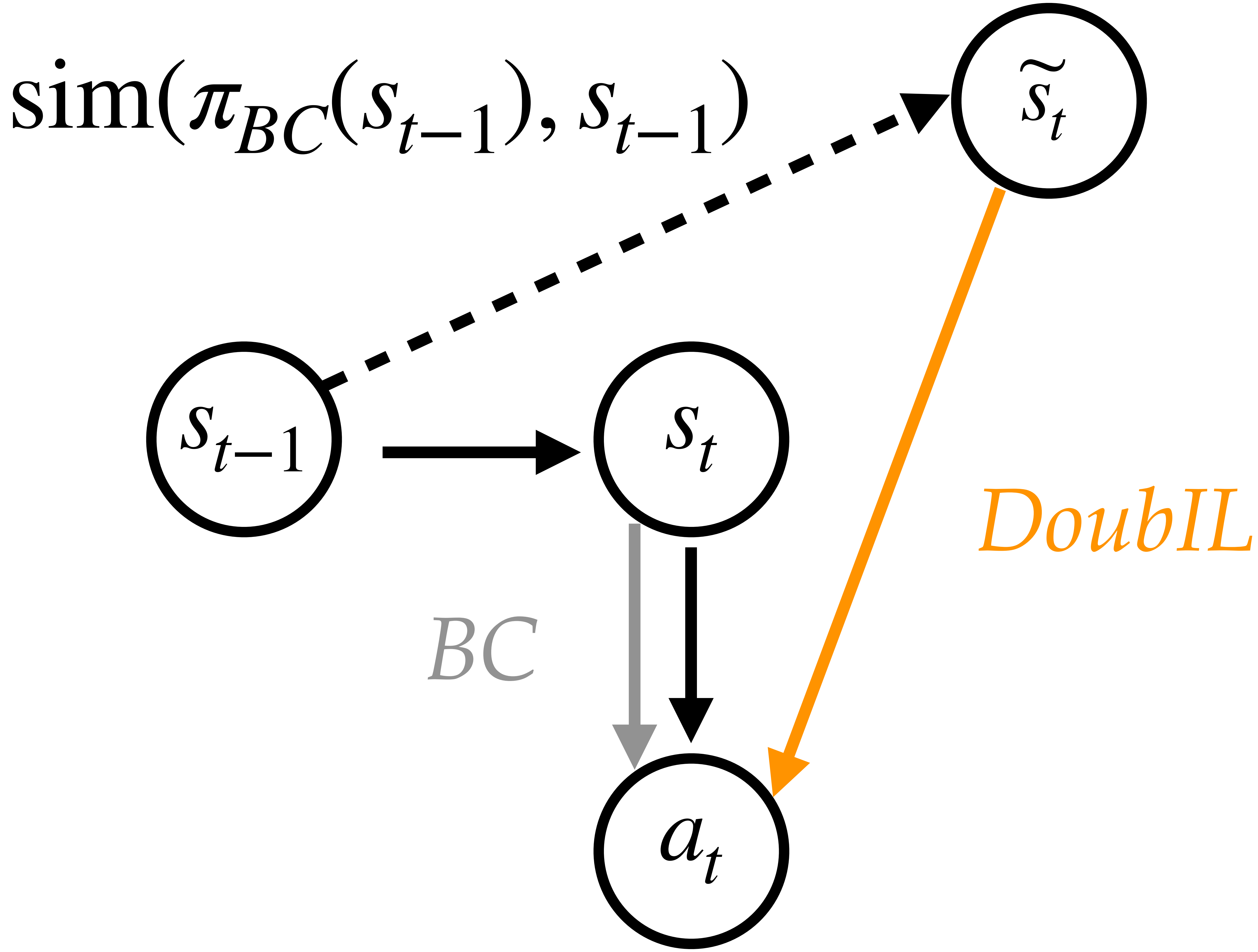


$$X = g(Z, U)$$
$$Y = h(X) + U$$
$$\mathbb{E}[U] = 0$$

$$0 = \mathbb{E}[U] = \mathbb{E}[U \mid z] = \mathbb{E}[Y - h(X) \mid z]$$
$$\Rightarrow \mathbb{E}[Y \mid z] = \mathbb{E}[h(X) \mid z], \forall z$$
$$\Rightarrow \min_h \mathbb{E}_z[(\mathbb{E}[Y \mid z] - \mathbb{E}[h(X) \mid z])^2]$$
$$\Leftrightarrow \min_h \max_f \mathbb{E}_z[2(Y - h(X))f(Z) - f^2(Z)]$$



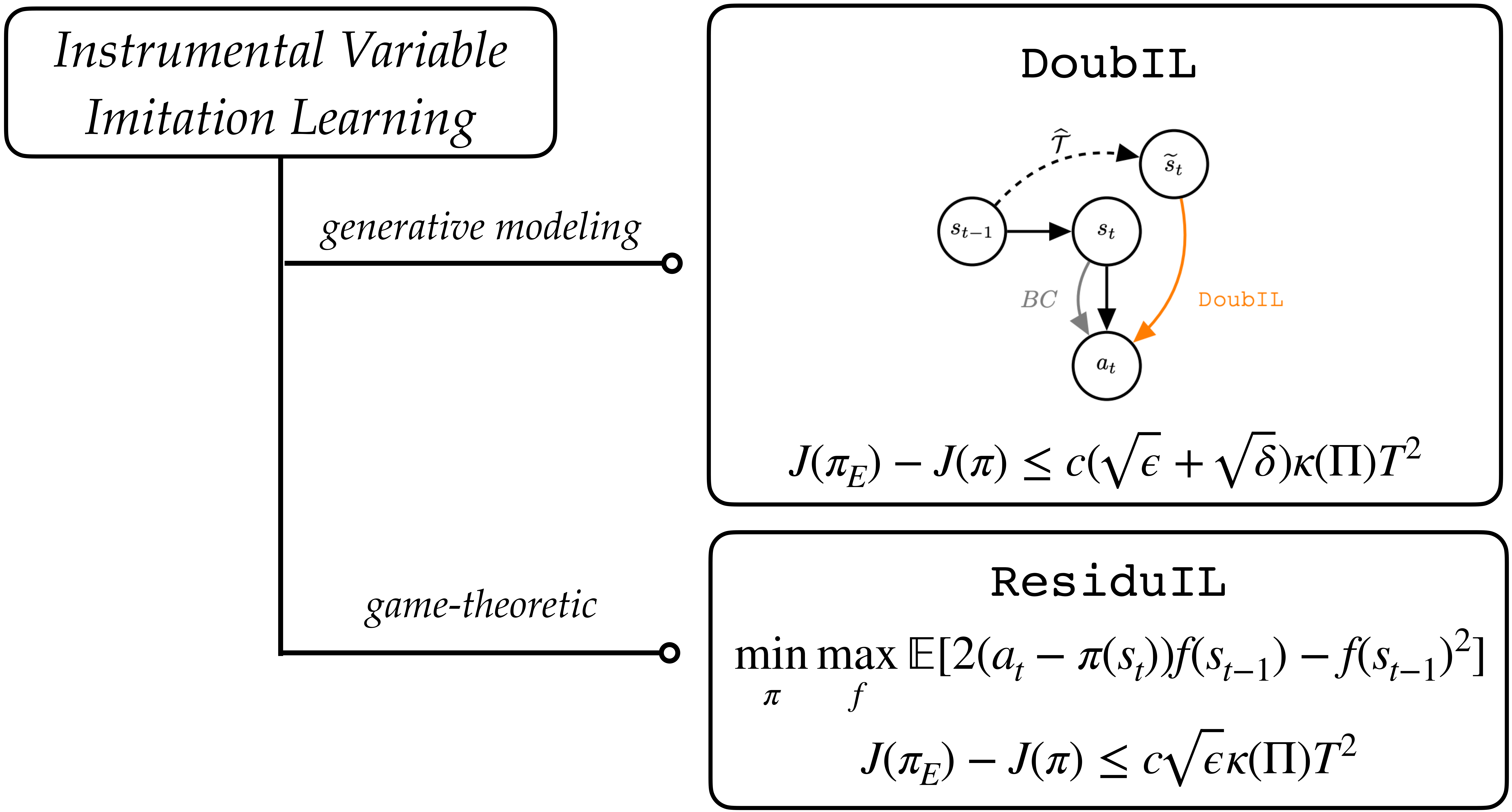
Z
s_{t-2}
s_{t-1}
s_t
X
T
T
π_E
π_E
π_E
a_{t-2}
a_{t-1}
a_t
Y
u_{t-1}
u_t



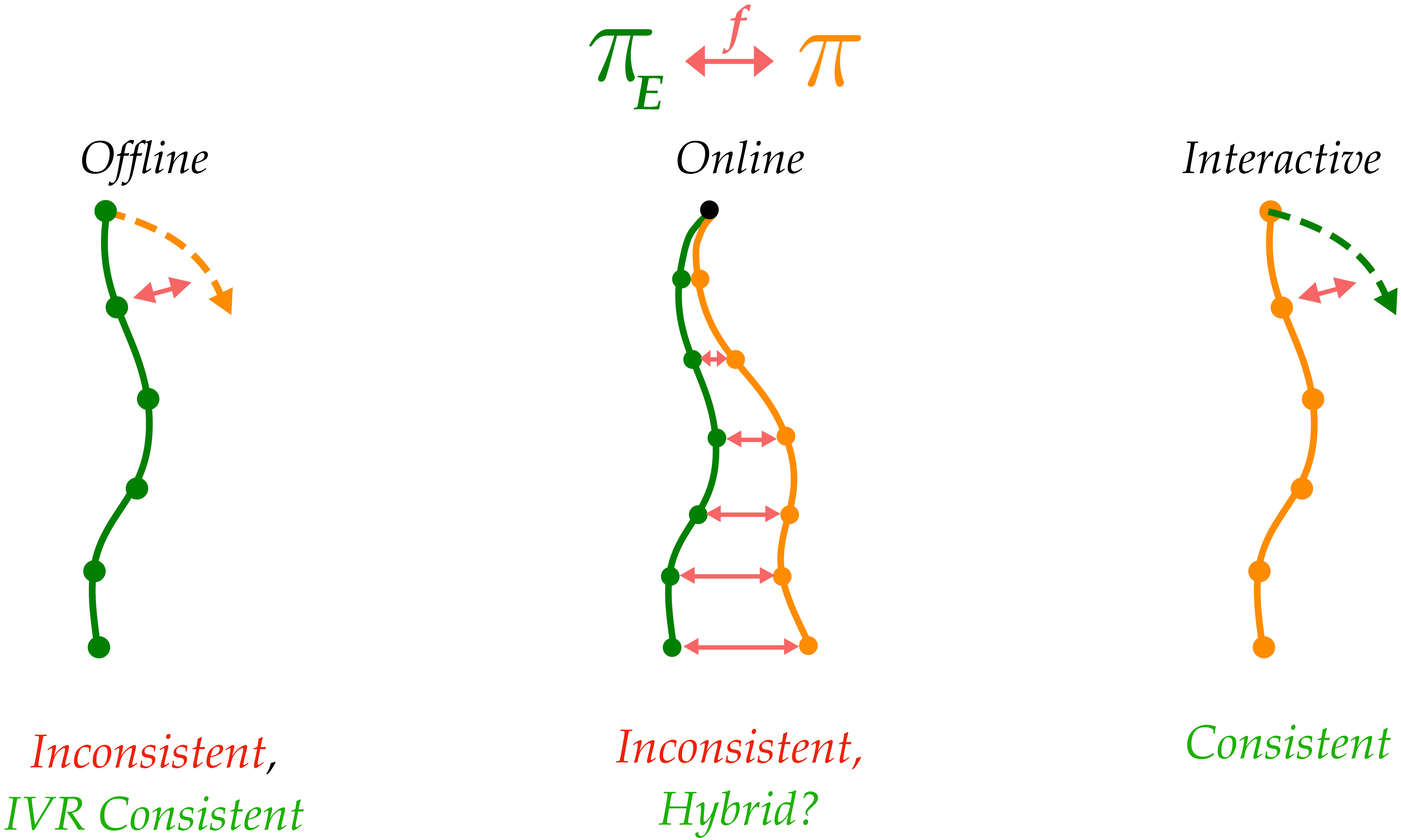
$$J(\pi_E) - J(\pi) \leq c(\sqrt{\epsilon} + \sqrt{\delta})\kappa(\Pi)T^2$$

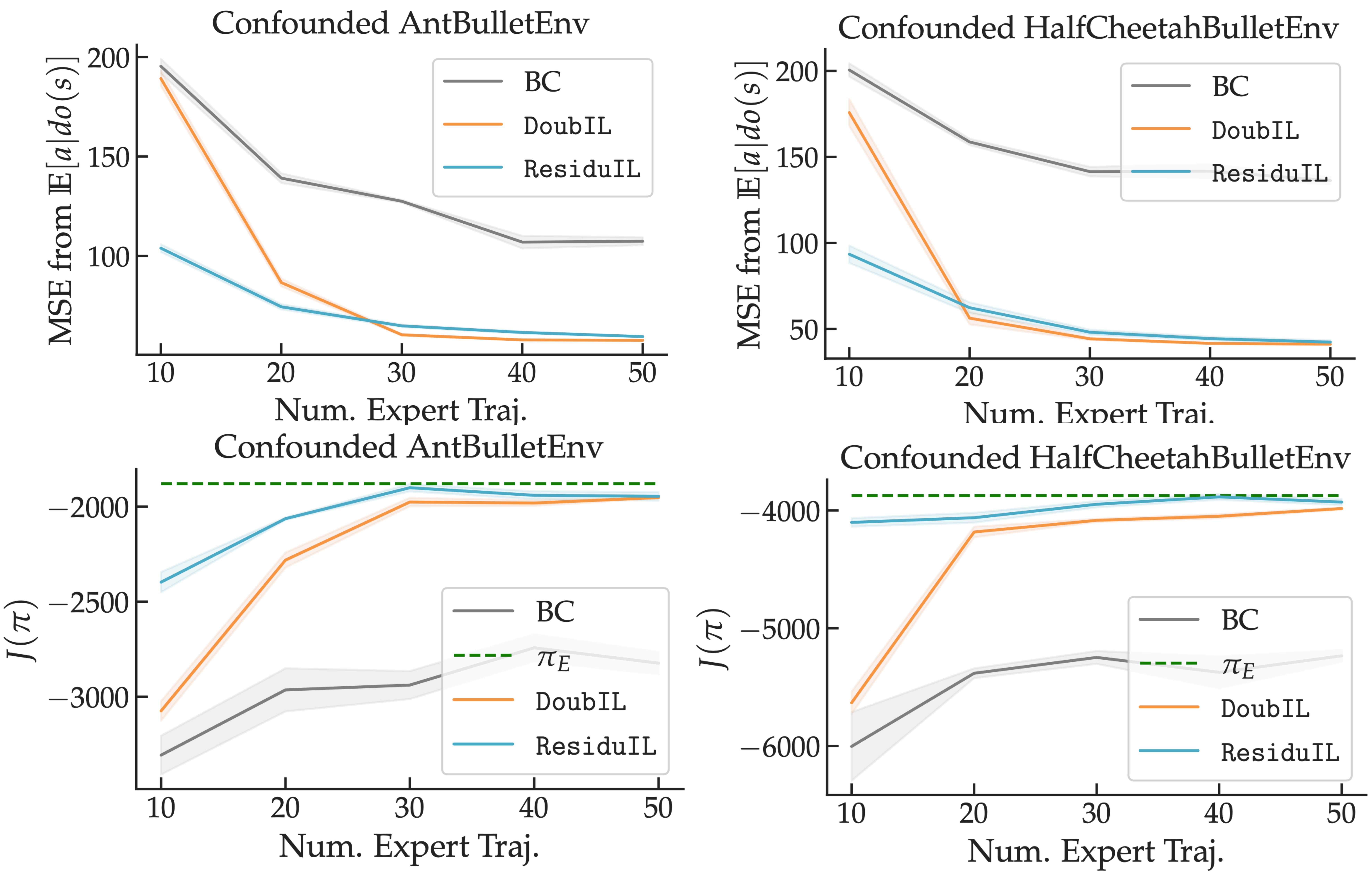
$$\min_{\pi} \max_{f} \mathbb{E}[2(a_t - \pi(s_t))f(s_{t-1}) - f(s_{t-1})^2]$$

$$J(\pi_E) - J(\pi) \leq c\sqrt{\epsilon\kappa(\Pi)}T^2$$

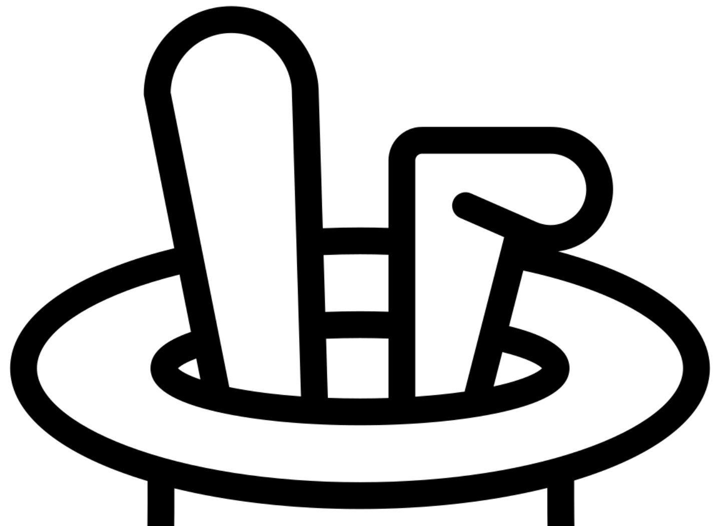


Instrumental Variable Imitation Learning
generative modeling
game-theoretic
DoubIL
$\widehat{\mathcal{T}}$
$\widetilde{s}_t$
s_{t-1}
s_t
a_t
BC
DoubIL
$J(\pi_E) - J(\pi) \leq c(\sqrt{\epsilon} + \sqrt{\delta})\kappa(\Pi)T^2$
ResiduIL
$\min_{\pi} \max_{f} \mathbb{E}[2(a_t - \pi(s_t))f(s_{t-1}) - f(s_{t-1})^2]$
$J(\pi_E) - J(\pi) \leq c\sqrt{\epsilon}\kappa(\Pi)T^2$


$\pi_E \overset{f}{\longleftrightarrow} \pi$
Offline
Online
Interactive
Inconsistent,
IVR Consistent
Inconsistent,
Hybrid?
Consistent


Confounded AntBulletEnv
MSE from $\mathbb{E}[a|do(s)]$
200
150
100
10
20
30
40
50
Num. Expert Traj.
BC
DoubIL
ResiduIL
Confounded HalfCheetahBulletEnv
MSE from $\mathbb{E}[a|do(s)]$
200
150
100
50
10
20
30
40
50
Num. Expert Traj.
BC
DoubIL
ResiduIL
Confounded AntBulletEnv
$J(\pi)$
−2000
−2500
−3000
10
20
30
40
50
Num. Expert Traj.
BC
π_E
DoubIL
ResiduIL
Confounded HalfCheetahBulletEnv
$J(\pi)$
−4000
−5000
−6000
10
20
30
40
50
Num. Expert Traj.
BC
π_E
DoubIL
ResiduIL

	Offline	Online	Interactive
Covariate Shift	✘	✔	✔
Hidden Context	✘	✔ w/ History	✔ w/ History
TCN	✔ w/ IVR	✔ w/ IVR	✔



Interventions happen via interaction with the environment in sequential decision making.

Thanks!

https://gokul.dev/

gswamy@cmu.edu