



POLITECNICO
MILANO 1863

COMPATIBLE REWARD INVERSE REINFORCEMENT LEARNING

A. M. METELLI, M. PIROTTA AND M. RESTELLI

{albertomaria.metelli, marcello.restelli}@polimi.it
{matteo.pirota}@inria.fr



PROBLEM

- **Inverse Reinforcement Learning (IRL)** problem: recover a **reward function** explaining a set of expert's demonstrations.
- **Advantages** of IRL over *Behavioral Cloning* (BC):
 - Transferability of the reward.
- **Issues** with some IRL methods:
 - How to build the **features** for the reward function?
 - How to **select** a reward function among all the optimal ones?
 - What if **no access** to the environment?

CONTRIBUTIONS

1. We propose the **Compatible Reward Inverse Reinforcement Learning (CR-IRL)**:
 - CR-IRL is **model-free** since it requires *solely* a set of expert's demonstrations;
 - CR-IRL performs both **feature extraction** and **reward selection**.
2. We provide **empirical results** to show that the rewards recovered by CR-IRL allow learning the optimal policy **faster** than the original reward function.

COMPATIBLE REWARD INVERSE REINFORCEMENT LEARNING

Two-steps algorithm

1. **Feature extraction**: build an *approximation space* for the reward function using a *first-order condition* on the **policy gradient**.
2. **Reward selection**: select a reward function in the space exploiting a *second-order condition* on the **policy Hessian**.

FEATURE EXTRACTION

Goal: extract all the reward functions making the expert optimal.

- Parametric representation of the expert's policy π_θ estimated via Behavioral Cloning.

- **Optimality condition** for the Q-function:

$$\nabla_\theta J(\theta) = \int_S \int_A d_\mu^{\pi_\theta}(s, a) \nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}(s, a) da ds = \mathbf{0}$$

- Build the **Expert's Compatible Q Features** (ECO-Q) as:

$$\Phi = \text{null}(\nabla_\theta \log \pi_\theta^T D_\mu^{\pi_\theta})$$

Phase 1

- Build the **Expert's Compatible Reward Features** (ECO-R):

model-based - reversing *Bellman* equation:

$$\Psi = (\mathbf{I} - \gamma \mathbf{P} \pi_\theta) \Phi$$

model-free - using *Reward Shaping*:

$$\Psi = (\mathbf{I} - \tilde{\pi}_\theta) \Phi$$

Phase 2

REWARD SELECTION

Goal: select the reward function that:

1. is a maximum of $J(\theta)$;
2. penalizes the most deviations from the expert's policy.

- policy Hessian:

$$\mathcal{H}_\theta J(\theta, \omega) = \int_{\mathbb{T}} p_\theta(\tau) \left(\nabla_\theta \log p_\theta(\tau) \nabla_\theta \log p_\theta(\tau)^T + \mathcal{H}_\theta \log p_\theta(\tau) \right) \Psi(\tau) \omega d\tau$$

- **Second-order optimality criteria**:

- minimize the maximum eigenvalue of $\mathcal{H}_\theta J(\theta, \omega)$;
- minimize the trace of $\mathcal{H}_\theta J(\theta, \omega)$ s.t. $\mathcal{H}_\theta J(\theta, \omega) \preceq 0$.

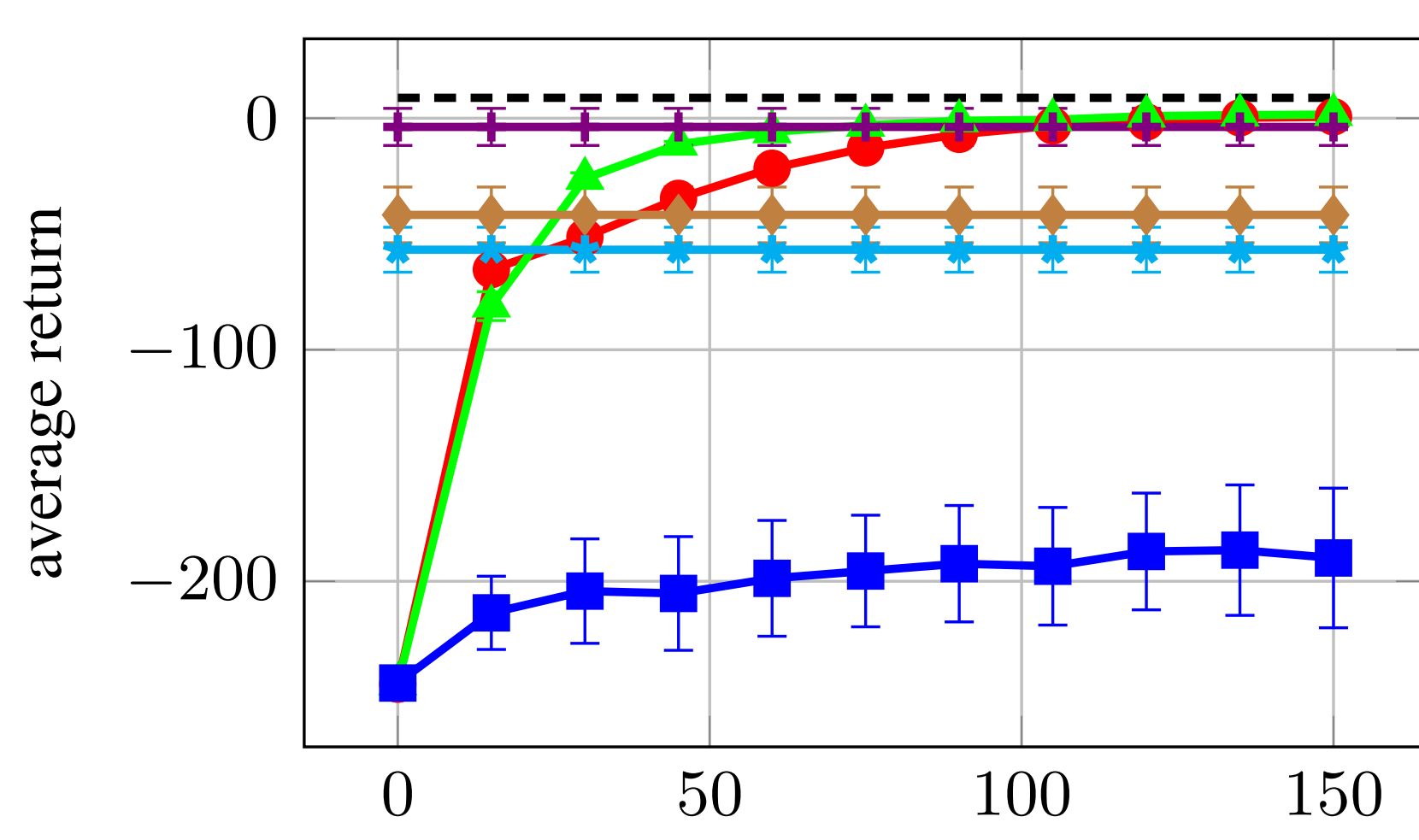
- **Second-order heuristic criterion**:

$$\min_{\omega} \omega^T \text{tr} \quad \text{s.t.} \quad \|\omega\|_2 = 1 \quad \rightarrow \quad \omega^* = \frac{\text{tr}}{\|\text{tr}\|_2} \quad \text{Phase 3}$$

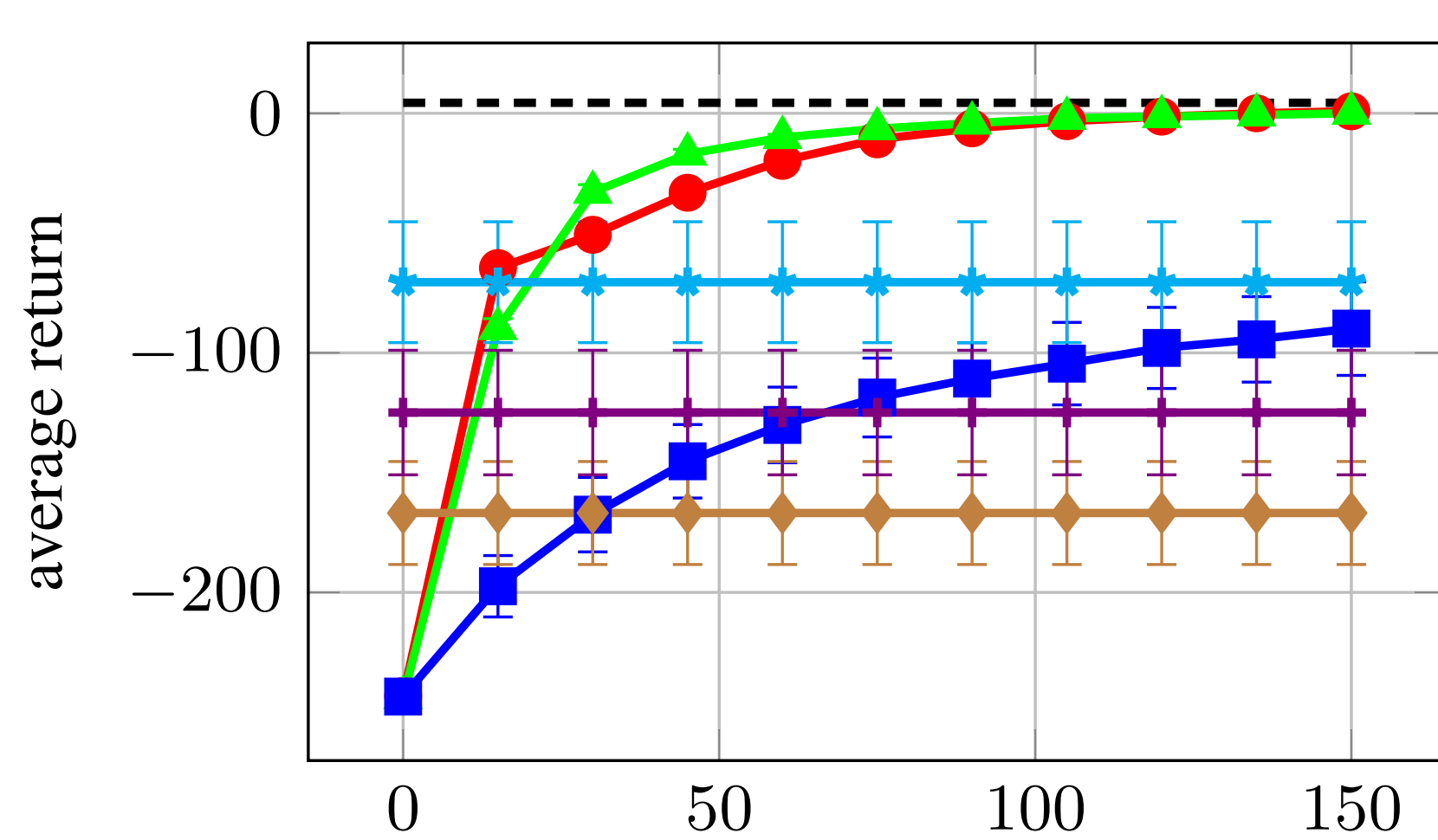
EXPERIMENTAL EVALUATION

TAXI

$\epsilon = 0$



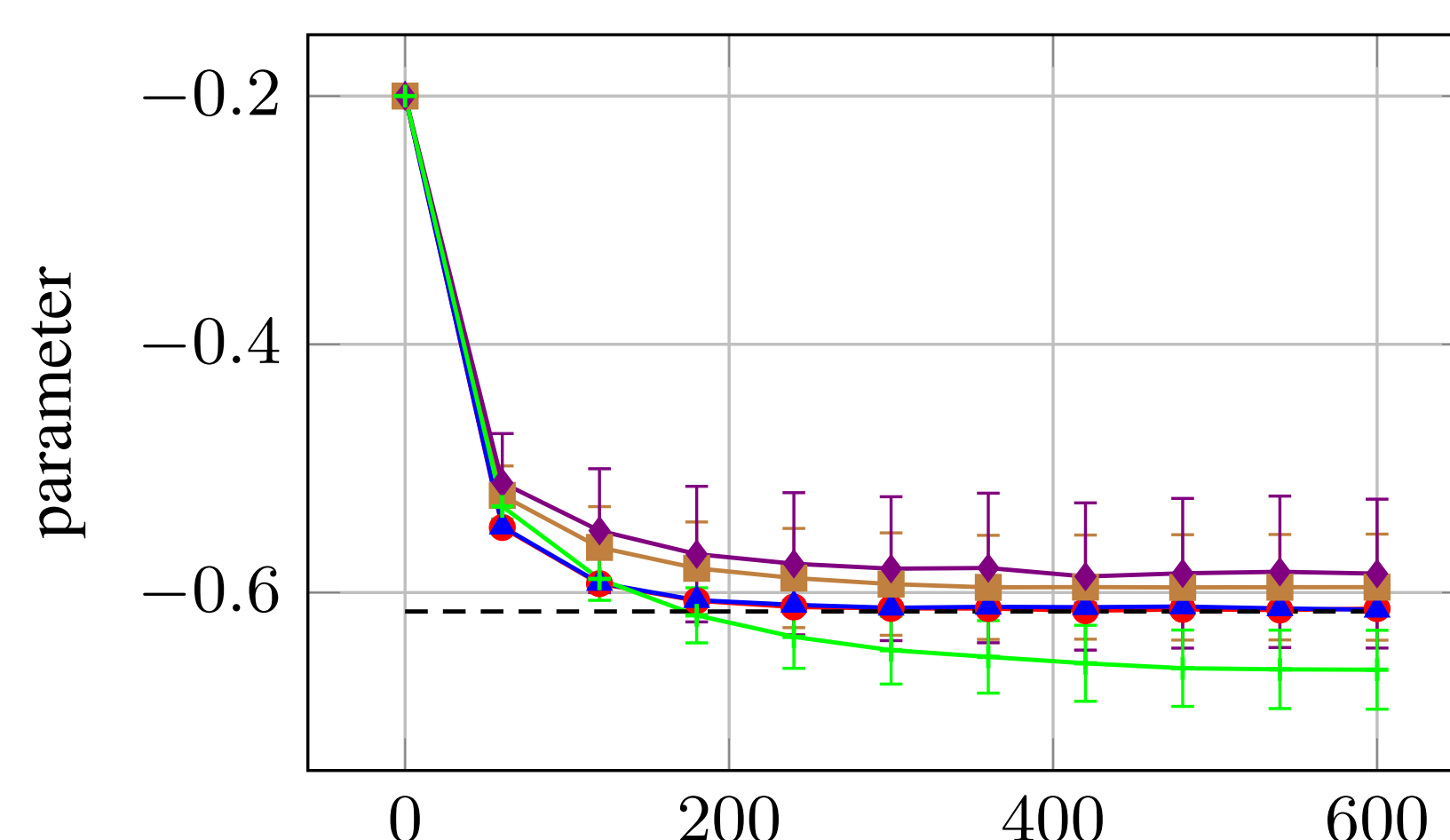
$\epsilon = 0.1$



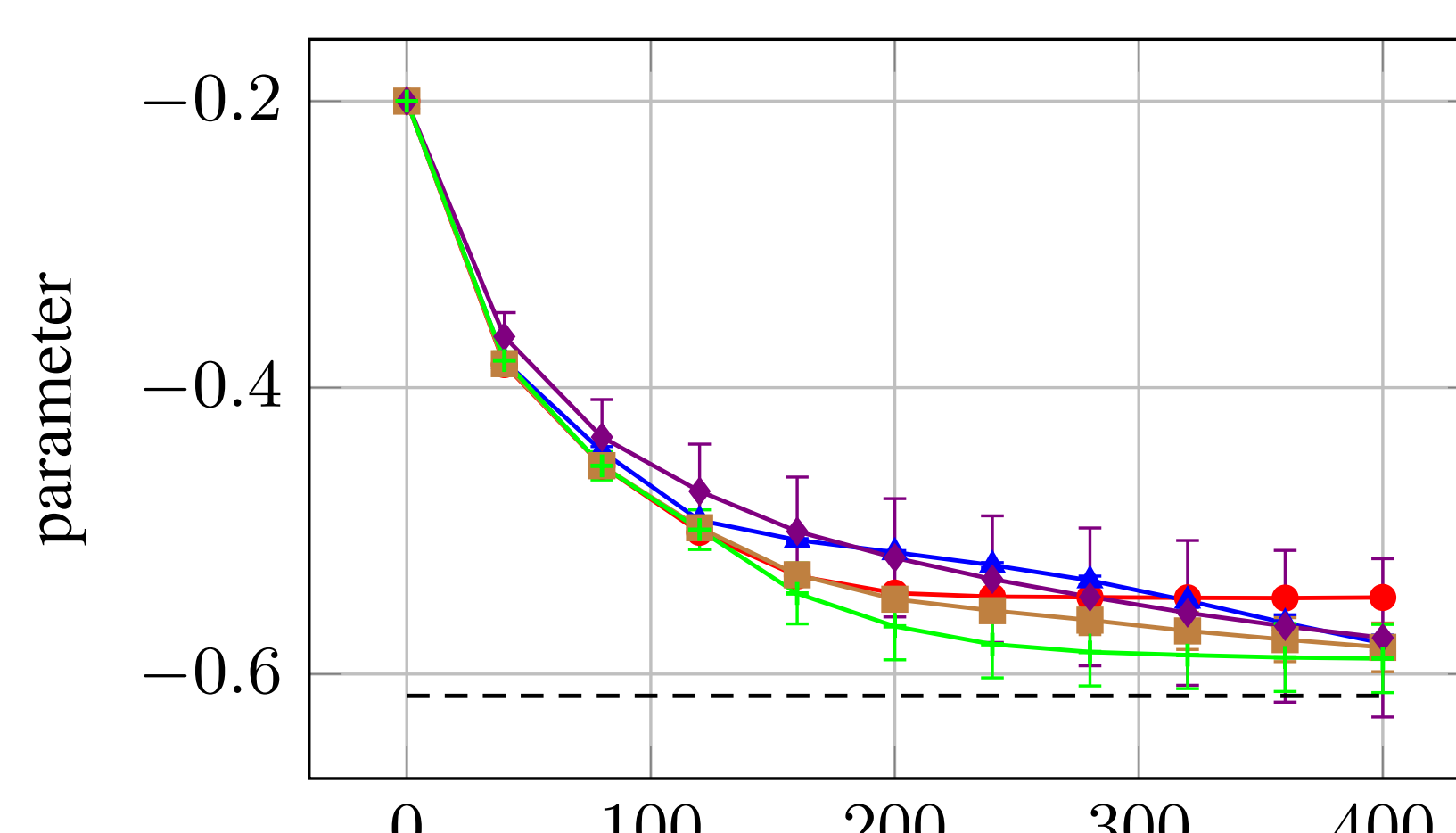
iteration
● Reward ▲ CR-IRL ■ ME-IRL
◆ LPAL × BC (ε = 0) ◆ BC (ε = 0.1)
--- Expert

LINEAR QUADRATIC GAUSSIAN REGULATOR

$\sigma^2 = 1.0$



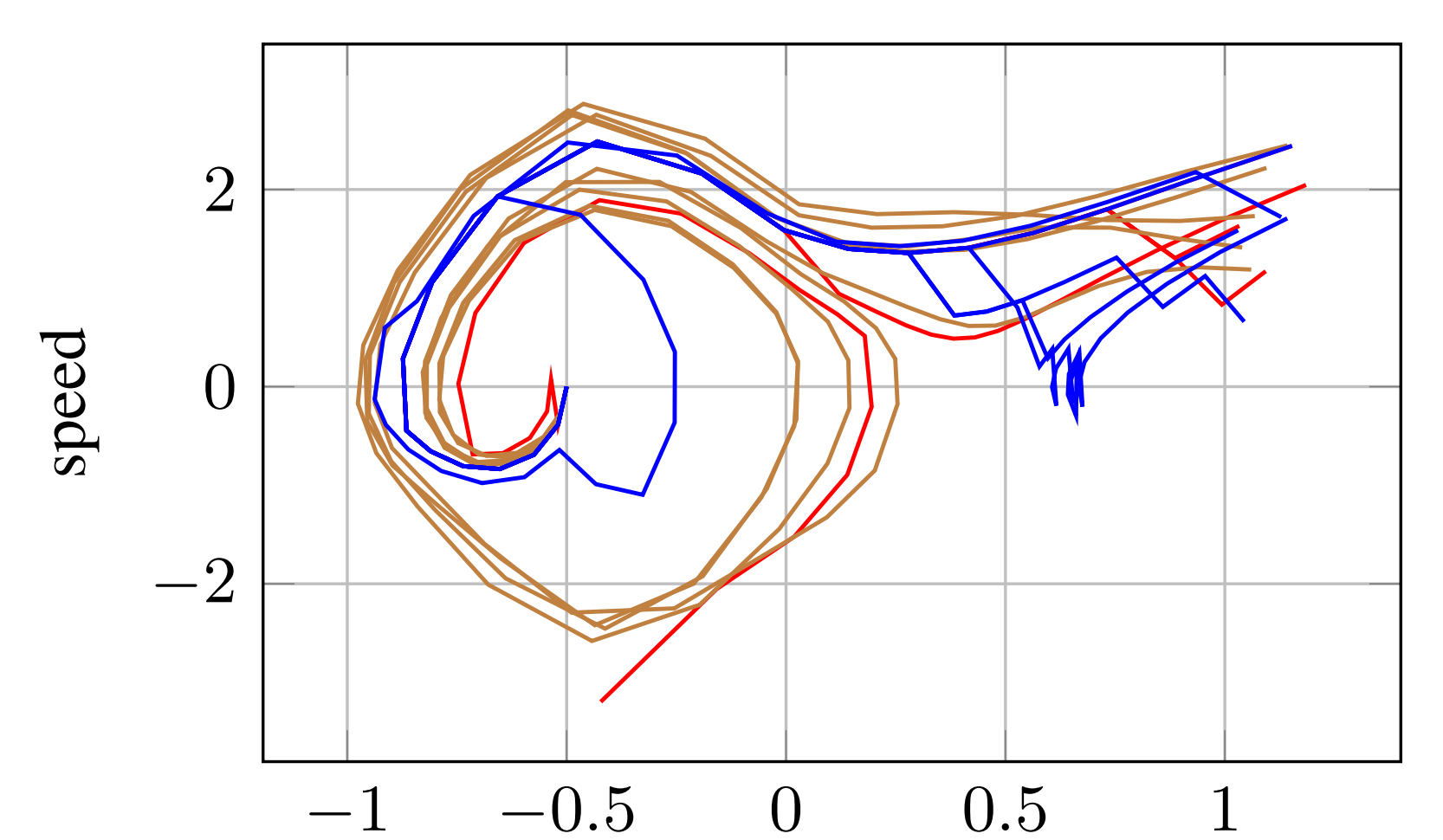
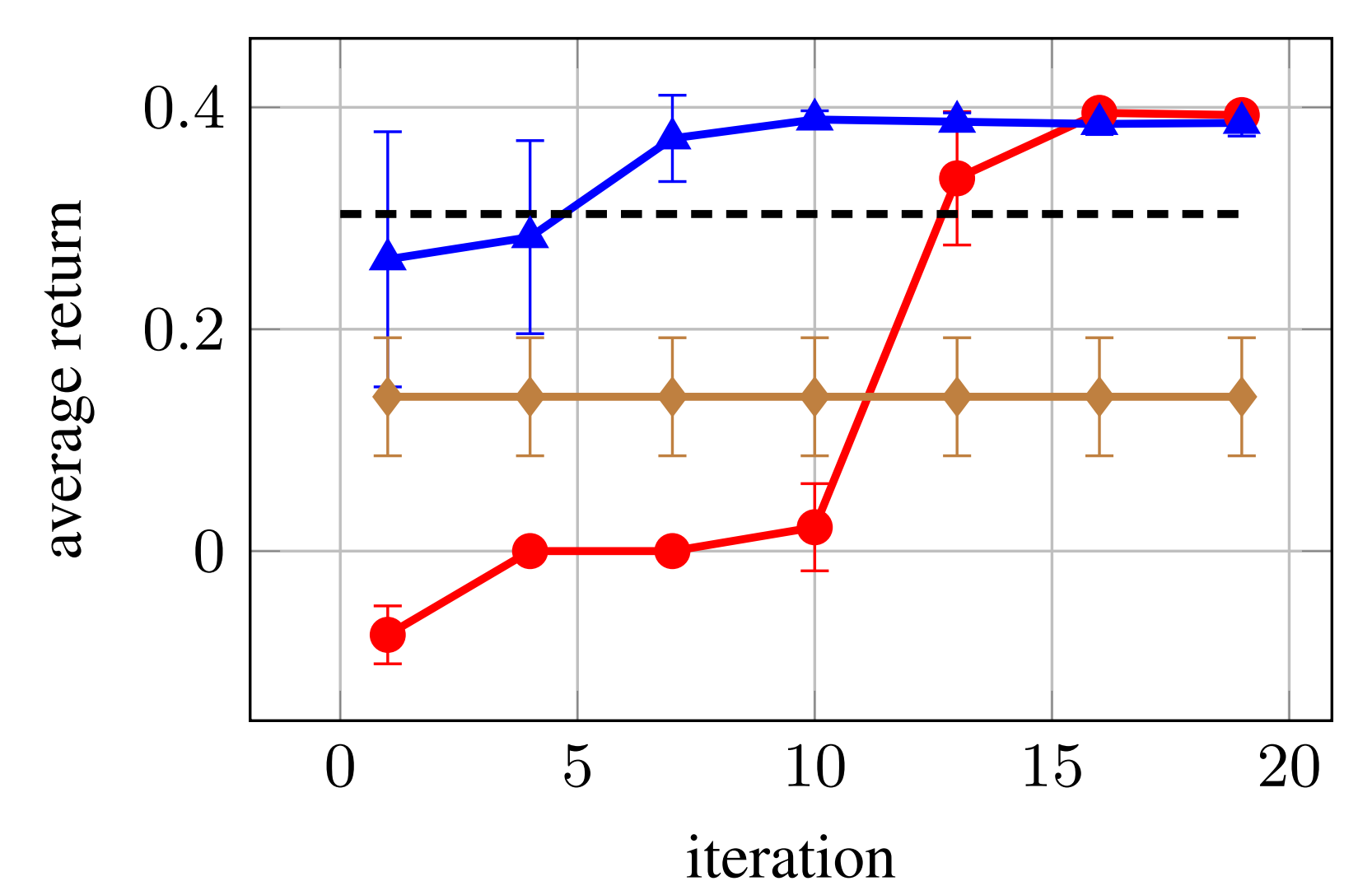
$\sigma^2 = 0.01$



iteration
● Reward ▲ Advantage ◆ GIRL-abs-val
× GIRL-square ▲ CR-IRL --- Expert

CAR ON THE HILL

$\epsilon = 0.1$



position
▲ CR-IRL ◆ BC
--- Expert ● Reward