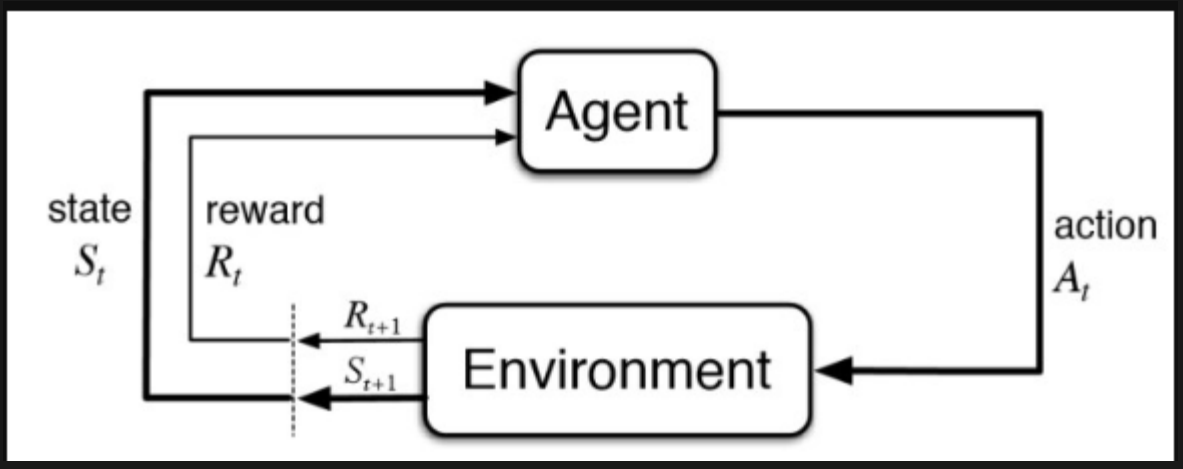
HYBRID BAYESOPT AND RL OPTIMIZATION IN EMUKIT

R244 - Chris Tomy - 2024/12/04

SETUP

Standard RL Setup: Markov Decision Process environment (S, A, P, R)



PARAMETERIZATIONS

- Environment is parameterized: $\theta = [g, t, \ldots]$
- Agent learns control policy: $\pi : S \rightarrow A$

MOTIVATION

Goal: minimize some cost function.

Example: $\min(\lambda_1 T + \lambda_2 F)$ where T is time, and F is fuel.

View this as $\max E[\sum r_i]$.

NAIVE APPROACH

- BO using a GP surrogate $\mathbf{f} \sim N(\mu, K)$
- Model relationship between hyperparameters $\vec{\theta}$ and the cost of trained policy $f(\vec{\theta})$.
- Computing $f(\vec{\theta})$: re-train and evaluate policy for each θ?

How do we do better than re-training a policy?

INSIGHT

For $g = -1 \rightarrow g = -1.1$, the policy may not change much.

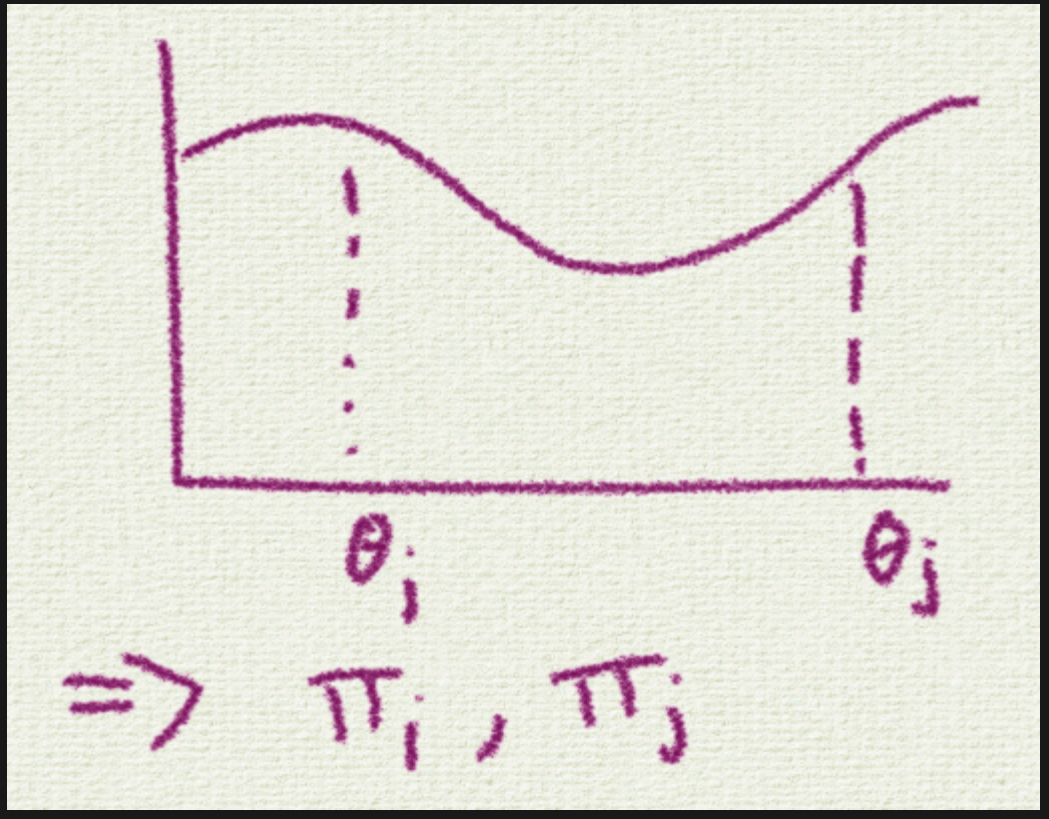
BETTER APPROACH

General idea: **adapt** the policy, don't re-learn.

POLICY FINE-TUNING

- Sound method: reduces to better parameter inits for the policy.
- Fine-tune: $\theta_{k+1} = \theta_k - \alpha \nabla J'$

- The larger $||\vec{\theta} - \vec{\theta'}||_2$ gets, the more difficult finetuning may become.
- Store all previous policies $\pi_1, \pi_2, \ldots$ and init from closest

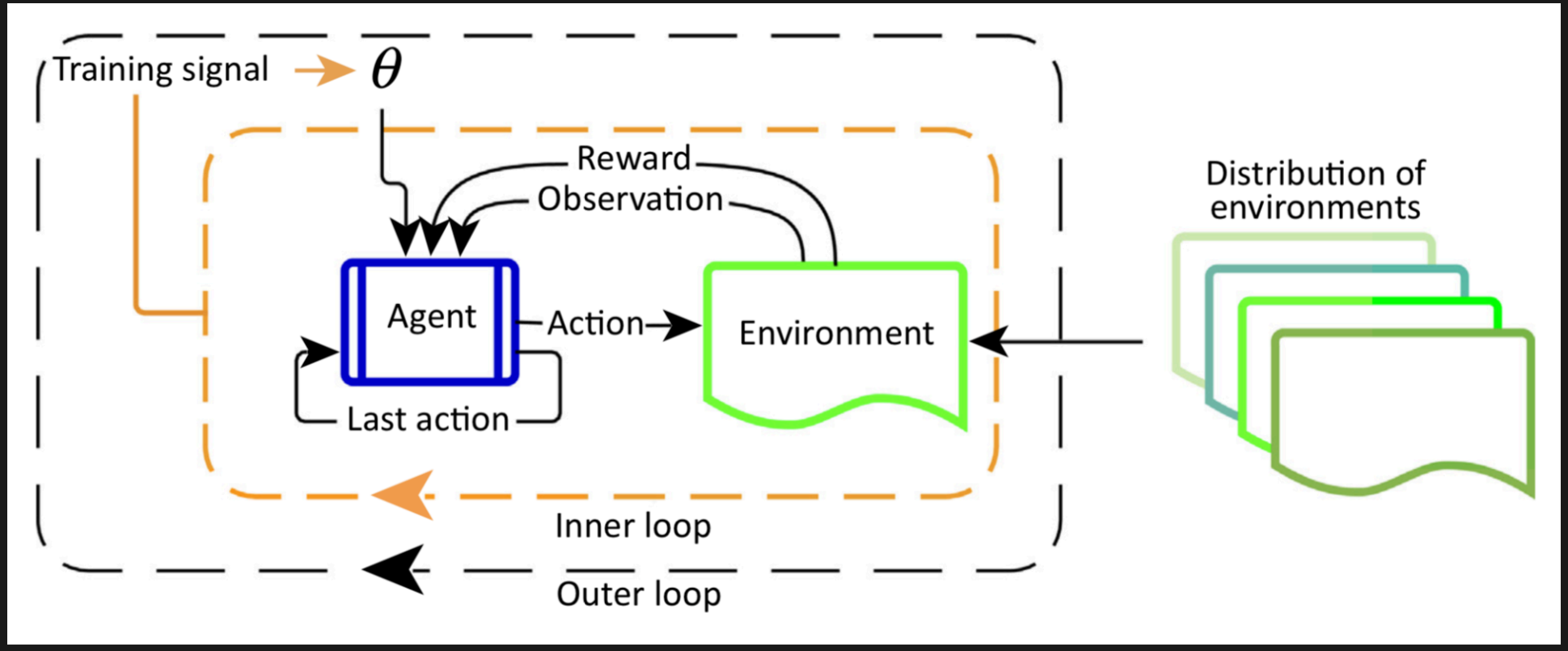


IDEAL CASE

- Perfectly adaptable policy $\pi^{\star}$
- Just run an episode for each $\vec{\theta}$

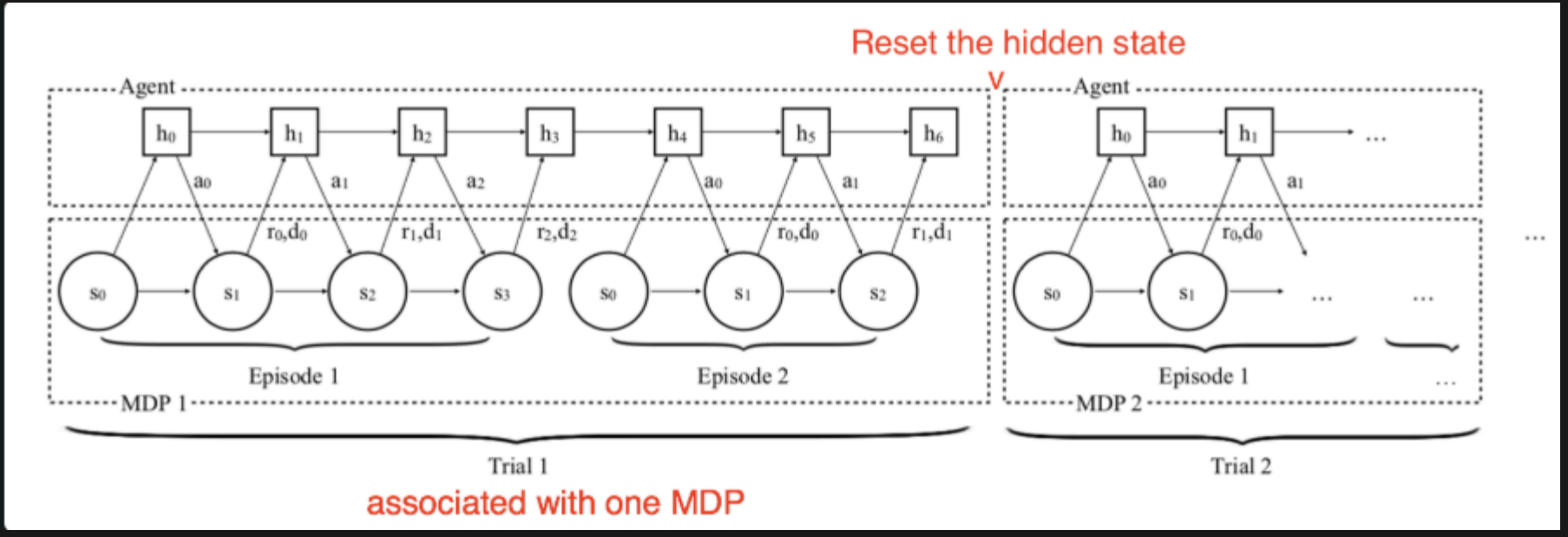
META RL

RL to learn adaptable policies.



HOW DOES META-RL WORK?

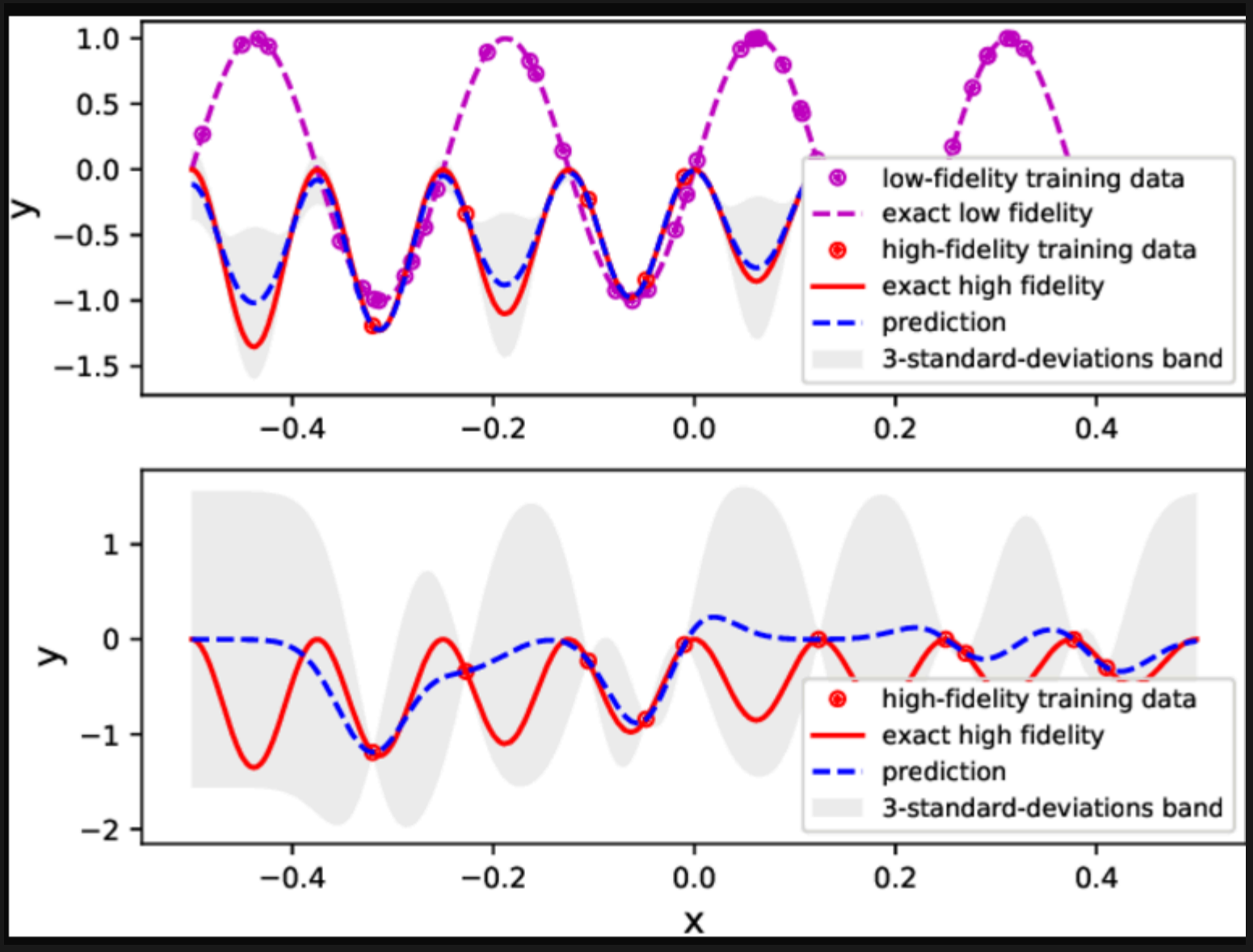
- LSTM hidden state: "memory".
- No gradient updates



RESEARCH QUESTIONS

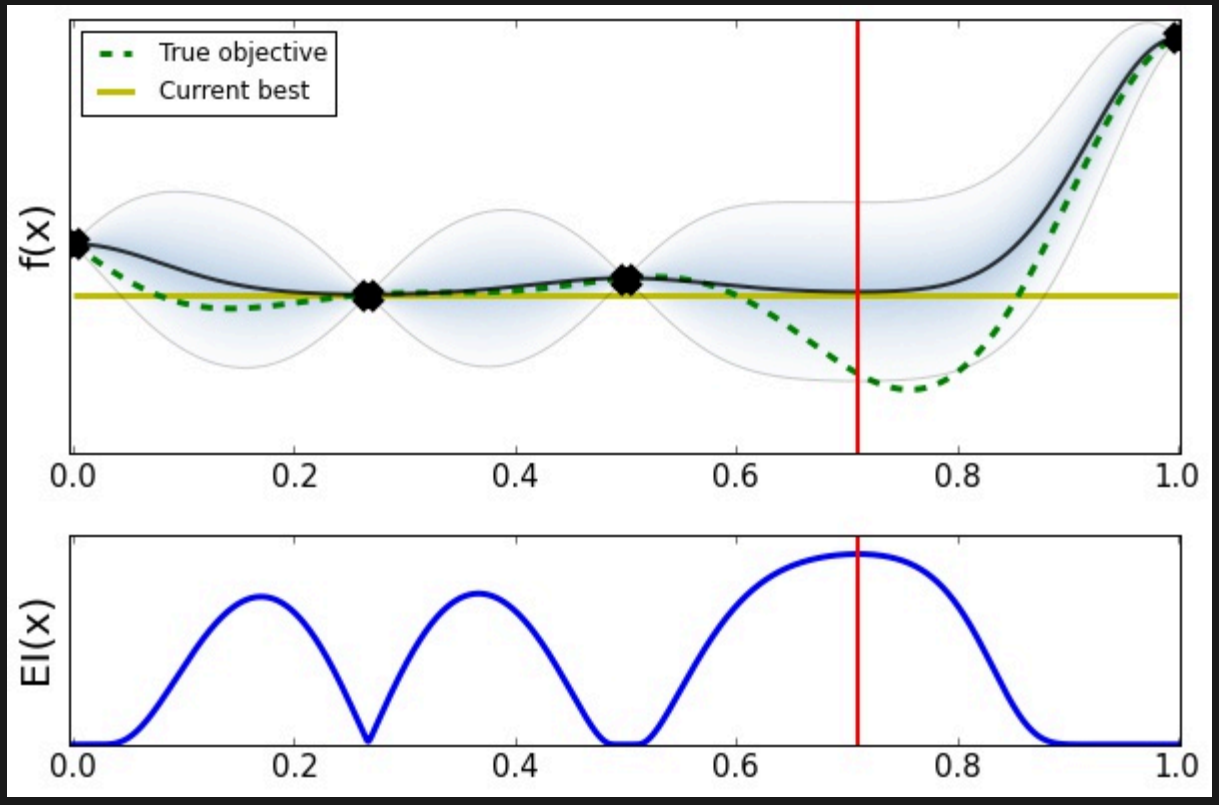
- Can we use the surrogate to better perform policy adaptation?
- Can we tweak the acquisition function to make adaptation training easier?
- How much accuracy do we lose with these adaptation methods?
- Multi-fidelity policy adaptation methods?

Multi-fidelity



EMUKIT

- Toolkit for BO
- Interfaces to define your parameter space, GP, acquisition, etc.
- No built-in RL support



WORK PLAN

- OpenAI Gym environment
- Implement naive hyperparameter search over environments
- Learn lots of RL methods
- Benchmark improvement with finetuning, MetaRL, etc.

Tech: OpenAI Gym, Emukit, PyTorch.

Thanks!