1. Motivation

- Nonstationary worlds require fast continuous adaptation.
- Multi-agent systems, any machine learning systems in the wild.
- A step towards continual, never-ending learning [1].

2. Background

Learning to learn for fast adaptation

Given a task description, a good adaptation rule must generate a model suitable for the task at hand:

- Adaptation via a gradient steps on a task-specific loss:
  \[ \phi_i = g_\theta(T_i) = \nabla \theta \mathcal{L}_{\phi_i}(T_i) \]
- At meta-training, search for a good parameter initialization:
  \[ \min_{\theta} \sum_{T \sim D} \mathcal{L}_{\phi_i}(T_i, \phi_i) \]

3. Adaptation as Inference

Meta-learning for RL. The data are trajectories: \( T: (\hat{z}, d_1, \ldots, z, d_M, r_M) \).

- Treat policy parameters, tasks, and all trajectories as random variables.
- In this view, adaptation = inference and meta-learning = learning a prior.
- Brings in compositionality of probabilistic modeling:
  - Different priors and inference algorithms \( \rightarrow \) new meta-learning methods (cf. [3]).
  - Different dependencies between the variables \( \rightarrow \) new adaptation methods.

4. Meta-learning for Continuous Adaptation

Real tasks are rarely i.i.d. There are often relationships that we can to exploit. Assuming that the tasks change over time consistently, we can learn to anticipate the changes and adapt to the temporal shifts.

Meta-learn on pairs of tasks by solving:

\[
\min_{\theta} \sum_{f_i} \mathcal{L}_{T_i, \phi_i}(\theta) \quad \text{where} \quad \\
\mathcal{L}_{T_i, \phi_i}(\theta) = \mathbb{E}_{T_i \sim D} \left[ \mathcal{L}_{T_i}(f_i, \phi_i, \pi) \right] \mathcal{L}_{T_i}(f_i, \phi_i, \pi)
\]

The algorithm

Meta-learning at training time:

- Sample a batch of task pairs, \([T_i, T_i+1]_K\].
- Rollout trajectories \( \tau_i \) for \( T_i \) (first task in each pair) using \( \pi_\phi \).
- Compute \( \phi(\tau_i, \theta, \alpha) \) and rollout \( \tau_{i+1} \) for each \( T_{i+1} \) using \( \pi_\phi \).
- Update \( \theta \) and \( \alpha \) using the stochastic gradient of the meta-loss.

Unbiased estimator of the gradient of the meta-loss

\[
\nabla_{\phi} \mathcal{L}_{T, \phi}(\theta, \alpha) = \mathbb{E}_{T_i \sim D} \left[ \mathcal{L}_{T_i}(\phi_i, \pi) \right] \nabla_{\phi} \mathcal{L}_{T_i}(\phi_i, \pi) + \nabla_{\phi} \sum_{j=1}^{K} \nabla_{\phi} \mathcal{L}_{T_{j+1}|T_{j}}(\phi_j, \pi_j)
\]

N.B.: The highlighted term was missing in the original derivation of the policy gradients for MAML-RL, which made the gradient estimators biased [2]. A general solution for such issues is developed in [4].

Adaptation at execution time:

- Interact with the environment using \( \pi_\theta \).
- Store all trajectories and importance weights, \( \gamma_i \), in the experience buffer.
- Before each episode, compute \( \theta \) using importance-corrected adaptation updates using trajectories from the buffer.

\[
\phi_i = \theta - \alpha \nabla \theta \mathcal{L}_{\phi_i}(f_i) + \sum_{j=1}^{K} \nabla_{\phi} \mathcal{L}_{T_{j+1}|T_{j}}(\phi_j, \pi_j)
\]

5. Environments & Setup

Iterated adaptation games

A multi-round game where an agent must adapt to opponents of increasing competence. The outcome of each round is either win, loss, or draw. Opponents are either pre-trained or also adapting.

6. Experiments

Nonstationary locomotion

Figure 1: Episode rewards for 7 consecutive episodes in 3 held-out nonstationary locomotion environments.

Multi-agent competition

Figure 2: Win rates for different adaptation strategies in iterated games with opponents pretrained via self-play. Competence of the opponents was increasing from round to round based on the precomputed policies at different stages of self-play.

Figure 3: Top: Evolution of a population of 1000 agents. Bottom: TrueSkill of the top-performing agents in the population.

Discussion

- \textbf{Limitations:} Gradient-based adaptation requires estimating second-order derivatives. This is computationally expensive and sample-inefficient (needs large batches).
- \textbf{Future work:} Adaptation + model-based RL.
- \textbf{Adaptation + curriculum learning/generation.}
- \textbf{Multi-step adaptation} (i.e., planning with tasks); better use of historical information.

- \textbf{References} [1] Ring et al. 94, 97, Mitchell et al. 15.

Acknowledgements

Harri Edwards, Jakob Foerster, Aditya Grover, Anurag Raveendran, Vishal Kumar, Yuhuai Wu, Carlos Florescu, anonymous reviewers, and the OpenAI team.

\textbf{Correspondence:} achedavid@cs.cmu.edu