Actor-Expert: A Framework for using Action-Value Methods in Continuous Action Spaces

Motivation

Value-based methods are difficult to use in continuous action spaces, because of having to solve $\text{argmax}_a Q(s, a)$ at each time step.

Past approaches include:

- Constraining the action-value function to an easily maximizable form (Wire-Fitting, PICNN, NAF)
- Solving approximate $\text{argmax}_a Q(s, a)$ at each step (PICNN, QT-OPT)

However, these approaches may not learn action-value function accurately or select greedy actions accurately.

Overview

- We propose **Actor-Expert framework** for value-based methods in continuous action spaces, that decouples action-selection (Actor) from the action-value representation (Expert).
- Our Actor-Expert framework is analogous to Actor-Critic, but the Expert estimates the optimal value function, while the Actor aids in action-selection for both exploration and providing Q-learning target.
- We provide an instance of the Actor-Expert, that uses **Conditional Cross Entropy Method** to learn the greedy action from the Expert, and provide a two-timescale analysis to validate asymptotic behavior.

Conditional Cross Entropy Method

We extend the Cross Entropy Method (CEM) to be conditioned on states. Conditional CEM maintains a distribution over actions, starting with a wide distribution given state, i.e. $\pi(\cdot|S_t)$.

At each step, the goal is to iteratively minimize the KL-divergence to the uniform distribution over actions where the objective function$(Q(S_t, \cdot))$ is greater than some threshold. This target distribution can be approximated with an empirical distribution, by sampling and keeping the top-percentile action samples.

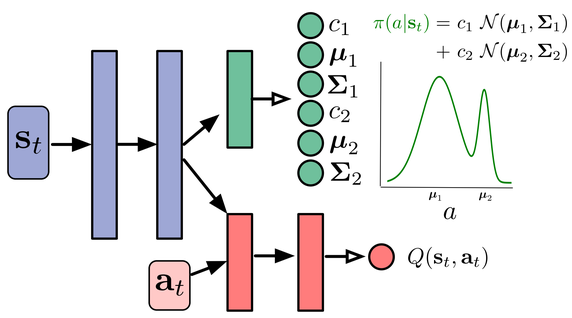


Figure: Actor-Expert with Conditional CEM, with a bimodal distribution. The policy $\pi(\cdot|S_t)$ is a conditional mixture model, with coefficients c_i, means μ_i and diagonal covariances Σ_i

Actor-Expert with Conditional CEM

High-level Algorithm:

Algorithm 1: Actor-Expert (with Conditional CEM)

Initialize Actor parameters $\mathbf{w}$ and Expert parameters θ.
for *t=1, 2, ...* **do**
  Observe S_t, sample $A_t \sim \pi_{\mathbf{w}}(\cdot|S_t)$
  Observe R_{t+1}, S_{t+1}
  Obtain maximum action a' from Actor $\pi_{\mathbf{w}}(\cdot|S_{t+1})$
  Update expert θ, using Q-learning with a'
  Obtain **empirical distribution** $\hat{I}(S_t) = \{a_1^*, \ldots, a_h^*\}$ based on $a_1, \ldots, a_N$
  ▷ Increase likelihood for high-value actions
  $\mathbf{w} \leftarrow \mathbf{w} + \alpha_t \sum_{j \in \hat{I}(S_t)} \nabla_{\mathbf{w}} \ln \pi_{\mathbf{w}}(a_j^*|S_t)$

Two Methods of Obtaining the Empirical Distribution:

Algorithm 2: Quantile Empirical Distribution [**AE**]

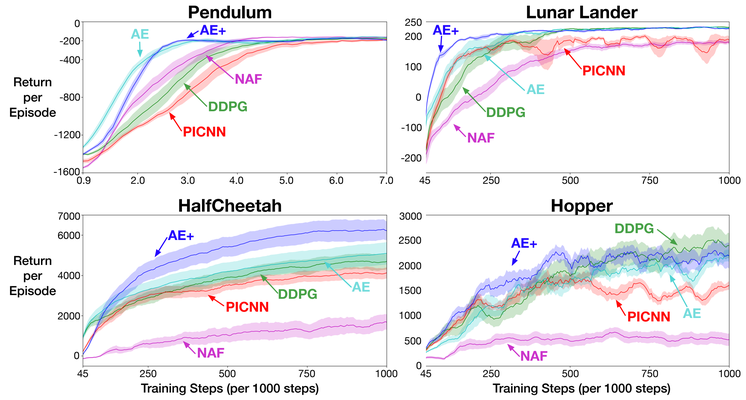
Sample N actions $a_i \sim \pi_{\mathbf{w}}(\cdot|S_t)$
Evaluate and sort in descending order:
$Q_\theta(S_t, a_{i_1}) \geq \ldots \geq Q_\theta(S_t, a_{i_N})$
▷ get top $(1-\rho)$ quantile, e.g. $\rho = 0.2$
return $\hat{I}(S_t) = \{a_{i_1}, \ldots, a_{i_h}\}$ (where $h = \lceil \rho N \rceil$)

Algorithm 3: Optimized Quantile Empirical Distr. [**AE+**]

For each a_i, do n steps of gradient ascent from $Q_\theta(S_t, a_i)$
return Quantile Empirical Distribution$(\{a_1^*, \ldots, a_N^*\})$

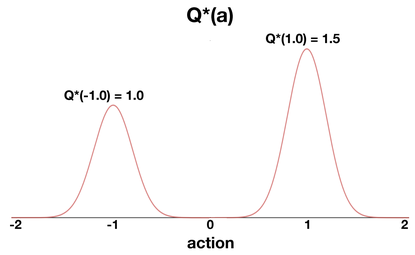
Experiments: Benchmark Domains

ActorExpert (AE) and ActorExpert+(AE+) performs similarly or better than other baseline methods in standard benchmark domains.

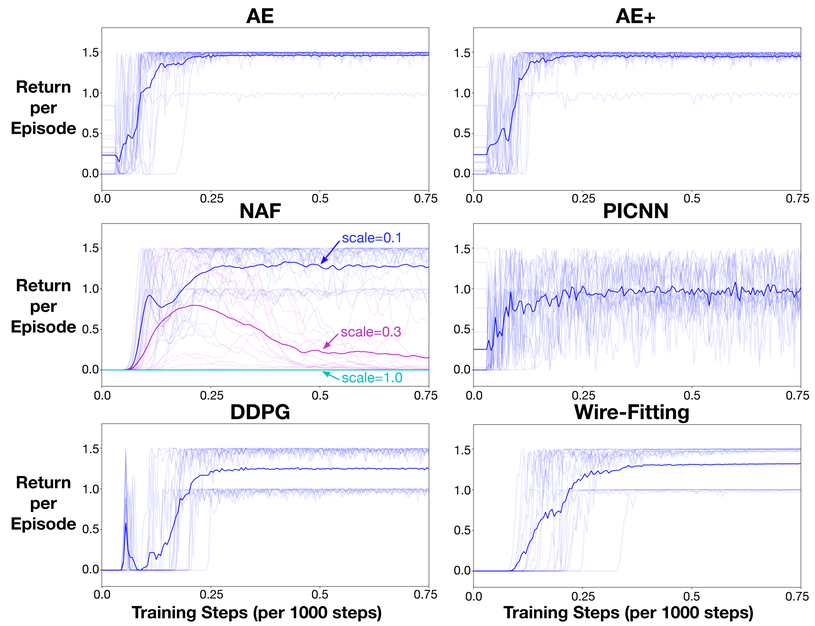


Experiments: Toy Bandit Domain

The optimal action-value function is bimodal, and methods that constrain the action-value function suffer while AE and AE+ do not.



- Constrained action-value function may try to fit both peaks, finding worse greedy action. (NAF)
- Solving approximate $\text{argmax}_a Q(s, a)$ may not be as robust. (PICNN)
- Random external exploration may lead to suboptimal greedy action. (DDPG, Wire-Fitting)



Conclusion and Future Work

- Like the Actor-Critic framework, we hope Actor-Expert framework can facilitate use of value-based methods in continuous action spaces.
- Under this framework, we can start a more systematic comparison between the advantages of value-based methods and policy gradient methods.

Sungsu Lim[1], Ajin Joseph[1], Lei Le[2], Yangchen Pan[1], Martha White[1]

sungsu@ualberta.ca, ajoseph@ualberta.ca, leile@iu.edu, pan6@ualberta.ca, whitem@ualberta.ca

RLAI & amii

[1] UNIVERSITY OF ALBERTA

[2] INDIANA UNIVERSITY BLOOMINGTON