

CLAIMS

1. A computer-implemented method of an execution device for generating an action selection policy for completing a task in an environment that includes the execution device and one or more other devices, the method comprising:

identifying a plurality of possible actions in a state, wherein the state corresponds to a vector of information sets, and each information set in the vector of information sets comprises a sequence of actions taken by the execution device that leads to the state;

identifying a vector of current action selection policies in the state, wherein each current action selection policy in the vector of current action selection policies corresponds to an information set in the vector of information sets, and the action selection policy specifies a respective probability of selecting an action among the plurality of possible actions in the state;

computing a sampling policy based on the vector of current action selection policies in the state, wherein the sampling policy specifies a respective sampling probability corresponding to each of the plurality of possible actions in the state;

sampling an action among the plurality of possible actions in the state according to a sampling probability of the action specified in the sampling policy; and

updating the each current action selection policy in the vector of current action selection policies of the execution device in the state based on the action.

2. The method of claim 1, further comprising, in response to determining that a convergence condition is met, controlling operations of the execution device based on the action selection policy.

3. The method of claim 1 or 2, wherein updating the each current action selection policy in the vector of current action selection policies of the execution device in the state based on the action comprises performing Monte Carlo counterfactual regret minimization (MCCFR) based on the action.

4. The method of any preceding claim, wherein updating the each current action selection policy in the vector of current action selection policies of the execution device in the state based on the action comprises:

calculating a probability of a sampled terminal sequence of actions based on the sampling probability of the action, the sampled terminal sequence of actions including the action and a terminal state for completing a task;

calculating a sampled counterfactual value of the action based on the probability of the sampled terminal sequence of actions;

calculating a regret value of the action based on the sampled counterfactual value of the action; and

updating the each of the vector of current action selection policies of the execution device in the state based on the regret value of the action.

5. The method of any preceding claim, wherein the state corresponds to a public sequence that comprises one or more actions publically known by the execution device and the one or more other devices; and
the each information set in the vector of information sets comprises the public sequence.

6. The method of any preceding claim, wherein computing a sampling policy based on the vector of current action selection policies of the execution device in the state comprises:
computing the sampling probability corresponding to each of the plurality of possible actions in the state as a mean value of current action selection policies of each of the plurality of possible actions in the state over the vector of information sets.

7. The method of any one of claims 1-5, wherein computing a sampling policy based on the vector of current action selection policies of the execution device in the state comprises:
computing the sampling probability corresponding to each of the plurality of possible actions in the state based on current action selection policies of each of the plurality of possible actions in the state and respective reach probabilities of the vector of information sets.

8. The method of claim of claim 7, wherein computing the sampling probability corresponding to each of the plurality of possible actions in the state based on current action selection policies of each of the plurality of possible actions in the state and respective reach probabilities of the vector of information sets comprises:

computing the sampling probability corresponding to each of the plurality of possible actions in the state based on a sum of the current action selection policies of each of the plurality of possible actions in the state weighted by the respective reach probabilities of the vector of information sets.

9. The method of any one of claims 1 to 5, wherein computing a sampling policy based on the vector of current action selection policies of the execution device in the state comprises:

computing the sampling probability corresponding to each of the plurality of possible actions in the state based on average action selection policies of each of the plurality of possible actions in the state and respective reach probabilities of the vector of information sets.

10. The method of claim 9, wherein computing the sampling probability corresponding to each of the plurality of possible actions in the state based on average action selection policies of each of the plurality of possible actions in the state and respective reach probabilities of the vector of information sets comprises:

computing the sampling probability corresponding to each of the plurality of possible actions in the state based on a sum of the average action selection policies of each of the plurality of possible actions in the state weighted by the respective reach probabilities of the vector of information sets.

11. A system for performing a software-implemented application for generating an action selection policy for a software-implemented application that performs actions in an environment that includes an execution party supported by the application and one or more other devices, the system comprising:

one or more processors; and

one or more computer-readable memories coupled to the one or more processors and having instructions stored thereon that are executable by the one or more processors to perform the method of any of claims 1 to 10.

12. An apparatus for generating an action selection policy for a software-implemented application that performs actions in an environment that includes an execution party supported by the application and one or more other devices, the apparatus comprising multiple modules for performing the method of any one of claims 1 to 10.