

DETERMINING ACTION SELECTION POLICIES OF AN EXECUTION DEVICE

ABSTRACT

Disclosed herein are methods, systems, and apparatus for generating an action selection policy (ASP) for an execution device. One method includes obtaining an ASP in a current iteration; obtaining a respective first reward for each action in a current state; computing a first reward of the current state based on the respective first rewards for the actions and the ASP; computing a respective regret value of each action based on a difference between the respective first reward for the action and the first reward for the current state; computing an incremental ASP based on the respective regret value of the each action in the current iteration; computing a second reward for the current state based on the incremental ASP; determining an ASP in a next iteration based on the second reward of the current state; and controlling actions of the execution device according to the ASP.

FIG. 5