Project: Parallel Reinforcement Learning

team07: Sam DePaolo, Michael Kielstra, Manquing Liu, Xiaohan Wu

February 19, 2022

An *n-armed bandit* is, for our purposes, a process where an agent can input some number from 1 to n and will receive a randomized reward with mean μ_n and standard deviation σ_n. These values are fixed but are kept secret from the agent. The mechanism is named by analogy to a "one-armed bandit," and in future we will refer interchangeably to inputting the value k, taking action k, and pulling handle k. The problem, the basis of the field of reinforcement learning (RL), is to maximize total reward over some large number of handle pulls.

Epsilon-greedy (or ϵ-greedy) exploration is one algorithm developed for situations such as these. The agent begins by pulling each handle exactly once. It then, with probability ϵ, pulls a random handle, and, with probability $1 - \epsilon$, pulls the handle which has given it the best mean rewards in the past. This ensures that the agent will rarely get stuck pulling a handle that normally gives a bad reward but gave an excellent one once by chance, while also allowing it to mostly pull the handle that gives it the best reward.

Multiple agents working on the same bandit can pool their knowledge in *parallel RL*.[1] This is beneficial in the context of a "black-box bandit," where we are allowed to execute the code as much as we want but cannot decompile and inspect it. (We might imagine, for instance, that we are conducting a security audit of proprietary software.) We propose to develop both a simple black-box bandit and a parallel ϵ-greedy exploration scheme to determine the optimum actions on that bandit, using distributed-memory programming to allow agents to synchronize data among themselves at regular intervals. We would be able to benchmark such metrics as the wall clock time to achieve a certain average reward per handle pull using n CPUs, and thereby test on a grand scale the extent to which parallelism is worth exploring in RL.

[1] `http://personal.denison.edu/~kretchmar/pubs/SCI2002.pdf` is our main reference.