R12/Sarthak Mittal/200050129

March 22, 2023

This paper proposes a new algorithm called **Neural Thompson Sampling (NeuralTS)** that uses **deep neural networks (DNNs)** for both exploration and exploitation. The posterior distribution that is used for the reward is centred at the NN approximation, and variance is taken from neural tangent features. The algorithm achieves **cumulative regret** of $\mathcal{O}(\sqrt{T})$ under the assumption that the reward function is bounded. The **empirical evaluation** on benchmarks with comparisons to several baselines supports theoretical guarantees.

Stochastic multi-armed bandits have been studied extensively in sequential decision-making. The **contextual bandit** is a variant used for recommendation, advertising, robotic control and healthcare. A **trade-off** between exploration and exploitation becomes necessary for maximizing the reward. The base idea of TS is: (1) compute the **posterior** of each arm being optimal for the present context (2) **sample** an arm from this distribution. Though recent works explore using NNs for contextual bandits, the **guarantees on regret** for TS are **limited** to simple models and under restrictive reward function.

The authors consider contextual K-armed bandit and use the observed K contextual vectors to minimize the pseudo regret by estimating the unknown reward using a **fully connected NN (FCNN)**. The algorithm maintains a **Gaussian** distribution for each arm's reward, samples the reward of each arm from the **posterior** distribution and then pulls **greedily**. The posterior is **updated** using observed reward and NN output.

The authors assume an unknown reward function h and an R-sub-Gaussian martingale difference noise sequence for the regret analysis. The theory of **Neural Tangent Kernel (NTK)** is crucial in their analysis as it connects DNNs to kernel methods. They define an **effective dimension** of the NTK matrix and show that it can be **upper bounded** if the contexts are nearly on some low-dimensional subspace of RKHS space. Under the assumptions that the NTK matrix is **positive definite** and that the context has **unit norm** and component j is the same as component $j + d/2$, the regret can be shown to have $\mathcal{O}(\sqrt{T})$ bound using a set of parameters and restrictions. The experiments they performed hinted that the NTK theory had some limitations. When T is unknown, m can be set **adaptively** by dividing time using powers of 2.

For the proof, they assume that the **network width** m has some bounds to control the approximation error. The authors define **events** for mean and variance under which the estimated mean reward is similar to the expected reward. They define **saturated arms** as those whose standard deviation of the estimate is smaller than that of the optimal arm, thus obtaining a bound on the regret of unsaturated arms. The expectation of regret is also **bounded** conditioned on the mean event. They use **Azuma-Hoeffding** inequality to bound the difference between the true reward of the optimal arm and another arm conditioned on the mean event. Combining all the intermediate results, they obtain a theoretical bound on the cumulative regret.

The empirical evaluation is performed on several **benchmarks** including adult, covertype, magic telescope, mushroom, shuttle, as well as MNIST, with comparisons to several **algorithms** including linear and kernelized TS, linear and kernelized UCB, BootstrapNN and ϵ-greedy for NNs. According to the authors, NeuralTS performed **among the best** in 6 datasets and was significantly better than other baselines in 2 of them. NeuralTS also degraded "more gracefully" than NeuralUCB upon increasing the **reward delay**.