CME323 Report: Distributed Multi-Armed Bandits

Milind Rao
milind@stanford.edu

1 Introduction

Consider the multi-armed bandit (MAB) problem. In this sequential optimization problem, a player gets to pull one of K arms at every time instant. Arm i when pulled yields a reward $Y_{i,k} \overset{\text{iid}}{\sim} P_i$. The distribution and means of the rewards for each arm are unknown. The goal of the player is the maximize rewards or minimize regret (difference between an oracle who knows the distribution of rewards and the player). MAB problems are used in online ad placement, ranking of search results, accelerating model selection among other applications. There is a balance between exploring arms which have not been seen before and exploiting the arm that is seen till now to offer the best rewards.

The analysis of the multiarmed bandit problem was given by Gittens [1]. Upper Confidence Bound and a variant of online gradient approach to solving the single player case was proposed by Auer et al [2]. The analysis of Posterior Sampling was done in [3].

In this project, we consider the distributed version of the problem where there are n players distributed on a network. Each node plays the game in parallel and can communicate with their neighbour. We investigate how much cooperation can help players converge to the optimal arm. A version of this problem was studied in Liu [4] where players play with common arms and there are collisions if multiple players pick the same arm. There is no communication between players. In [5], the distributed bandit problem in considered on a P2P network where any node can communicate with any other node. A distributed ϵ-greedy approach was mooted and an analysis for a simplified case was presented. Finally, in [6], the impact of infrequent all-to-all communication on the convergence of distributed multiarmed bandits is analysed. In this work, we extend the online gradient approach to the distributed network case.

2 Single Player Solutions

We further make assumptions that the rewards on each arm are subgaussian with parameter σ^2. An example could be that the rewards on the K arms arise from bernoulli distributions of different means.I.e. $y_i \overset{\text{iid}}{\sim} \mathcal{B}(p_i)$ is the distribution of the rewards on the i^{th} arm. Suppose that the i^{*th} arm is optimal and the mean of the rewards from this arm is p^* and that the reward from the arm i the expected regret of a policy u can be defined as,

$$R_T(u) = \mathbf{E}\left[\sum_{t=1}^{T} p^* - y_{u(t,y_1^t)}\right] \tag{1}$$

A variety of approaches have been proposed to solve the problem when there is a single player [7]. We go over a few briefly here:

1. ϵ-greedy approach: The player maintains empirical estimates of the mean for each arm as they see rewards. With probability -1ϵ, they play the best arm seen till now, or otherwise randomly pick an arm to drive exploration.

2. Upper Confidence Bound (UCB) approaches: In this approach, the player maintains a range for the mean of the rewards on each arm. The mean would fall in this range with high probability. At each time instant, the player picks the arm with the highest upper confidence bound and this drives exploration. As the player sees more instances of the reward from a particular arm, they can tighten the high probability bounds they have for the mean. With no assumptions on the gap between means of optimal and suboptimal arms, the regret scales as $\mathcal{O}(\sqrt{K\sigma^2 T \log T})$.

3. Posterior Sampling: This approach, also known as Thompson sampling is a Bayesian approach. A prior is placed on the parameters (in this case the means) of the rewards on each arm. At each time instant, the parameters are sampled from the prior and based on this, the best arm is chosen. The posterior is now updated based on the reward from the arm chosen. Posterior sampling achieves regret bounds close to UCB of $\mathcal{O}(\sqrt{KT \log T} + K)$.

4. Online Gradient Descent Approach: We describe this approach in greater detail as we will expand to the case with collaborative distributed multi-armed bandits. In this algorithm, the player maintains a distribution from which the best arm is sampled at each time instant. The reward seen is used to generate an unbiased estimate of the mean rewards and this is appended to a running estimate. The distribution maximizing the expected reward with the running estimate is now chosen. This is further highlighted in Alg. 1

   As seen, g is the running estimate of the rewards of each arm, w is the distribution that is maintained. The regret of this algorithm is $\mathcal{O}(\sqrt{KT \log T})$. The algorithm is obtained from online mirror descent approaches with entropic regularization. Given running estimate of rewards, the distribution that maximizes expected reward is the solution to the following optimization problem

$$\min_{w \in \Delta_K} \quad g_t^\intercal w + \frac{1}{\eta} \sum_{k \in [K]} w_k \log w_k .$$

Algorithm 1 Online gradient descent method to solve the bandit problem.

```
**Input:** stepsize $\eta$, initial vector $w_1 = \frac{1}{K}\mathbf{1}$, initial loss estimate $g_1 = \mathbf{0}$.
**repeat**Choose arm $i$ with probability $w_{t,i}$, receive loss $y_i$
    $$g_{t+1,j} \leftarrow g_{t,j} + \begin{cases} y_j/w_{t,j} & \text{if } j = i \\ 0 & \text{else.} \end{cases}$$
    $$w_{t+1,j} \leftarrow \frac{\exp(-\eta g_{t+1,i})}{\sum_k \exp(-\eta g_{t+1,k})} \forall j \in [K]$$
**until** $t = T$
```

The solution is exponentiated gradient descent. This method is similar to the follow-the-regularized leader approach.

3 Distributed Solution

We follow the approach in Duchi et al [8] for generalizing the online gradient method for networks.

Let us consider that there are n nodes on an undirected network $G = (V, E)$ each playing the bandit game independently but the arms for each player offer rewards from the same distribution. At the end of each time instant, a node can send a message of size $\mathcal{O}(K)$ to all its neighbours. This is similar to the Pregel framework seen in class. Our goal is minimize the average regret seen by all players.

In [8], distributed dual averaging is an approach to generalizing dual averaging or follow-the-regularized-leader problems. Suppose that A is the adjacency matrix. Let D be a matrix whose diagonal elements are the degrees of the various nodes. We can obtain a double stochastic matrix P where $P_{i,j}$ is non-zero only if nodes i and j are neighbours. P can be obtained as,

$$P = I - \frac{1}{\max(\text{diag}(D))}(D - A)$$

The nodes now mix the running estimates of the rewards from neighbours. This enables rapid exploration initially as neighbours may play different arms. The algorithm is described in Alg. 2

As was shown in [8], the average $\bar{g}_T^\alpha = \frac{1}{T}\sum_{t=1}^{T} g_t^\alpha$ converges to the optimal value for each node. The convergence analysis of this algorithm has not yet been performed and is future work. It is believed that the spectral gap $1 - \sigma_2(P)$ or the gap between the first and second eigenvalues of P governs the convergence rate of the process. For a well connected graph, the spectral gap is larger and the random walk mixing time is small. When there are no edges between nodes, the spectral gap is 0 and the random walk mixing time in infinity.

We now empirically verify the performance of the scheme. In Fig. 1, we contrast the performance of the distributed online gradient descent algorithm with

Algorithm 2 Distributed online gradient descent method to solve the bandit problem.

```
Input: stepsize $\eta$, initial vector $w_1^\alpha = \frac{1}{K}\mathbf{1}$, initial loss estimate $g_1^\alpha = \mathbf{0}$ for all nodes $\alpha$.
repeat
    for All nodes $\alpha$ in parallel do choose arm $i$ with probability $w_{t,i}^\alpha$, receive loss $y_i^\alpha$
        $$g_{t+1,j}^\alpha \leftarrow \sum_\beta P_{\alpha,\beta} g_{t,j}^\beta + \begin{cases} y_j^\alpha / w_{t,j}^\alpha & \text{if } j = i \\ 0 & \text{else.} \end{cases}$$
        $$w_{t+1,j}^\alpha \leftarrow \frac{\exp(-\eta g_{t+1,i}^\alpha)}{\sum_k \exp(-\eta g_{t+1,k}^\alpha)} \forall j \in [K]$$
    end for
until $t = T$
```

10 nodes and 30 arms when the nodes are not connected, connected via a chain with 2 neighbours and are completely connected. It is seen that cooperation reduces the regret.

In Fig. 2, we see the performance of grid configuration and a random graph with the same number of edges. The random graph is seen to perform marginally better.

This framework can be extended to the case where the links stochastically vary mimicking conditions in wireless sensor networks.

4 Conclusion and Further Work

In this work, we have proposed an extension of the multiarmed bandit problem to the distributed scenario where nodes can only communicate with neighbours. We have seen that cooperation reduces the expected regret. Also, the graph configuration which enables rapid mixing or quick communication of values from one to the other is better. This informs the design of wireless sensor networks.

For future work, the primary problem is to analyse the convergence of the regret of the algorithm. In the scheme, each node transmits a message of size $\mathcal{O}(K)$ which can be prohibitively large if the number of arms is large. Schemes which transmit a sample of the message and the convergence analysis of these schemes will be useful to know. Finally, the optimal choice of the step-size in the online algorithm depends on the number of nodes, the number of time-steps and the configuration of the network. Investigating the dependence on these parameters will be required for practical implementations.

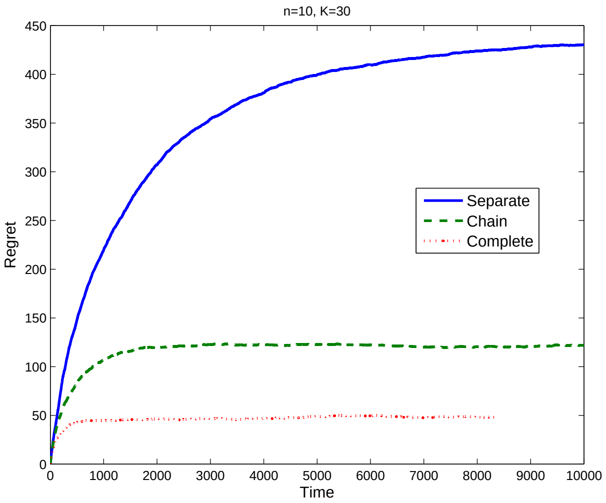

Figure 1: Regret varying for 10 nodes and 30 arms when they do not coordinate, communicate through a chain graph and on a complete graph.

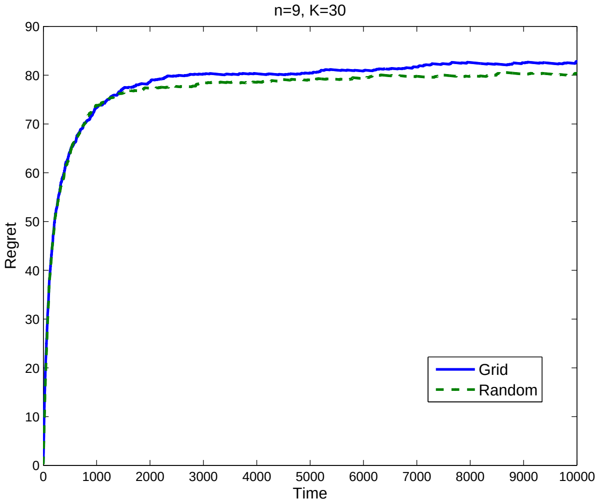

Figure 2: Regret varying for 9 nodes and 30 arms with a grid and random graph with same number of nodes.

References

[1] J. Gittins, K. Glazebrook, and R. Weber, *Multi-armed bandit allocation indices.* John Wiley & Sons, 2011.

[2] P. Auer, N. Cesa-Bianchi, and P. Fischer, “Finite-time analysis of the multiarmed bandit problem,” *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[3] D. Russo and B. Van Roy, “An information-theoretic analysis of thompson sampling,” *arXiv preprint arXiv:1403.5341*, 2014.

[4] K. Liu and Q. Zhao, “Distributed learning in multi-armed bandit with multiple players,” *Signal Processing, IEEE Transactions on*, vol. 58, no. 11, pp. 5667–5681, 2010.

[5] B. Szorenyi, R. Busa-Fekete, I. Hegedüs, R. Ormándi, M. Jelasity, and B. Kégl, “Gossip-based distributed stochastic bandit algorithms,” in *30th International Conference on Machine Learning (ICML 2013)*, vol. 28. Acm Press, 2013, pp. 19–27.

[6] E. Hillel, Z. S. Karnin, T. Koren, R. Lempel, and O. Somekh, “Distributed exploration in multi-armed bandits,” in *Advances in Neural Information Processing Systems*, 2013, pp. 854–862.

[7] J. C. Duchi, “Notes for ee377/stats311 information theory and statistics,” 2016.

[8] J. C. Duchi, A. Agarwal, and M. J. Wainwright, “Dual averaging for distributed optimization: convergence analysis and network scaling,” *Automatic control, IEEE Transactions on*, vol. 57, no. 3, pp. 592–606, 2012.