Thought Question Chapter 2: multi arm bandit

After I studied the multi-arm bandit problem, I have also done a lot of reading related to the combination of multi-arm slot machines, such as how to achieve good learning results in the case of limited feedback; how to solve the combination of exploration and then find the best solution; Off-line optimization based on greedy algorithm, if better combination of offline greedy algorithm and online learning; and when the expected gain of the overarm depends on the random distribution of each reference arm And not only is the average of the distribution of each reference arm.

The question “when the expected gain of the overarm depends on the random distribution of each reference arm And not only is the average of the distribution of each reference arm” is pretty interesting, so I did some research online. The stochastic combinatorial multi-armed bandit called CMAB, and the model allows a general nonlinear reward function, whose expected value may not depend only on the means of the input random variables but possibly on the entire distributions of these variables.

In Multi-arm bandit Algorithm, we have this function: “A∗ chooses the optimal super arm S ∗ = argmaxS∈F {rD(S)}”, If we are going to make it stochastic, there are some assumptions for this circumstance, and there is an Algorithm called SDCB (Stochastically dominant confidence bound) which can be used in this problem. I do not fully understand this algorithm, so I just paste the algorithm below.

Algorithm 1 SDCB (Stochastically dominant confidence bound)

Throughout the algorithm, for each arm $i \in [m]$, maintain: (i) a counter T_i which stores the number of times arm i has been played so far, and (ii) the empirical distribution $\hat{D}_i$ of the observed outcomes from arm i so far, which is represented by its CDF $\hat{F}_i$

// Initialization
for $i = 1$ **to** m **do**
  // Action in the i-th round
  Play a super arm S_i that contains arm i
  Update T_j and $\hat{F}_j$ for each $j \in S_i$
end for
for $t = m+1, m+2, \ldots$ **do**
  // Action in the t-th round
  For each $i \in [m]$, let $\underline{D}_i$ be a distribution whose CDF $\underline{F}_i$ is

$$\underline{F}_i(x) = \begin{cases} \max\{\hat{F}_i(x) - \sqrt{\frac{3\ln t}{2T_i}}, 0\}, & 0 \le x < 1 \\ 1, & x = 1 \end{cases}$$

  Play the super arm $S_t \leftarrow \mathsf{Oracle}(\underline{D})$, where $\underline{D} = \underline{D}_1 \times \underline{D}_2 \times \cdots \times \underline{D}_m$
  Update T_j and $\hat{F}_j$ for each $j \in S_t$
end for

Citation: Wei Chen, Wei Hu, Fu Li, Jian Li, Yu Liu, and Pinyan Lu. Combinatorial Multi-Armed Bandit with General Reward Functions. In Proceedings of the 29th Annual Conference on Advances in Neural Information Processing Systems (NIPS'2016), Barcelona, Spain, December 2016.