Problem Formulation

Shirley Shu, Tianfang Chang

1 Multi-Armed Bandit

- We have k-armed bandit with winning probabilities, $p_1, ..., p_k$.
- At each time step t, we take an action (pull an arm) on the machine and observe a reward X_t.
- Objective: maximize the total rewards.
- A is a set of actions, each referring to the interaction with one arm. $A_t \in \{1, 2, ..., k\}$
- Reward $X_t \sim P_{A_t}$, where $P_1, P_2, ..., P_k$ are unknown distributions.
- Measuring performance with Regret $R_n = n\mu^* - \mathbb{E}[\sum_{t=1}^{n} X_t]$, where $\mu^* = \max_i \mu_i$, μ_i is the mean reward of P_i.

2 Problem Formulation

Formulate the problem with Upper Confidence Bound (UCB). Let $(X_t)_{t-1}^{n}$ be a sequence of independent 1-subgaussian random variables with mean μ and $\hat{\mu} = \frac{1}{n}\sum_{t=1}^{n} X_t$, for all $\delta \in (0, 1)$,

$$P\left(\mu > \hat{\mu} + \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}}\right) \leq \delta \tag{1}$$

- Arms: k subflows.
- Reward (X): estimated throughput $X_t = \frac{MSS_t}{RTT_t} * \frac{1}{\sqrt{p_t}}$, where MSS is maximum segment size and p_t is number of lost packets in t-th round.
- Action Set (A): $A_t = \arg\max_i \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}}$, where $\hat{\mu}_i$ is the expected value of rewards with i-th subflow, and T_i is number of samples choosing subflow i in t-th round. δ is the error probability.
- Exploration: During the previous rounds, if the use of subflow i smaller, $\sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}}$ larger. So, it is more possible to choose i if $\hat{\mu}_i$ is same.

3 Current Algorithm

Algorithm 1 MPQUIC Scheduler with UCB

Require: k, δ

1: Choose each action once;
2: **for** $t \in 1, 2, ..., n$ **do**
3: Choose action $A_t = \arg\max_i \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}}$;
4: Observe reward $X_t = \frac{MSS_t}{RTT_t} * \frac{1}{\sqrt{p_t}}$;
5: Update $\hat{\mu}_i$ = expected value of X_t;
6: **end for**