Bandit Algorithm学习笔记1

6 Explore-Then-Commit 算法

算法流程

Input m.

In round t choose action

$$A_t = \begin{cases} (t \bmod k) + 1\,, & \text{if } t \leq mk\,; \\ \operatorname{argmax}_i \hat{\mu}_i(mk)\,, & t > mk\,. \end{cases}$$

(ties in the argmax are broken arbitrarily)

Algorithm 1: Explore-then-commit.

这个公式的物理意义是，前mk轮根据顺序选择动作，成为探索阶段。m*k轮以后选择平均奖励最大的那个动作。

平均奖励由这个公式给出：

$$\hat{\mu}_i(t) = \frac{1}{T_i(t)} \sum_{s=1}^{t} \mathbb{I}\{A_s = i\} X_s$$

定理6.1 ETC遗憾上界

Theorem 6.1. *When ETC is interacting with any 1-subgaussian bandit and* $1 \leq m \leq n/k$,

$$R_n \leq m \sum_{i=1}^{k} \Delta_i + (n - mk) \sum_{i=1}^{k} \Delta_i \exp\left(-\frac{m\Delta_i^2}{4}\right).$$

其中m表示1到k个动作至少执行的次数。

证明过程

不失一般性地，我们假设第一个动作为最优动作，根据遗憾分解引理，有：

$$R_n = \sum_{i=1}^{k} \Delta_i \mathbb{E}\left[T_i(n)\right] .$$

在前m*k轮，ETC策略是确定的，每个动作都被选择m次。因此每个动作被选择的次数期望是，他一定小于等于每次都选择最优动作的概率：

$$\begin{aligned}\mathbb{E}\left[T_i(n)\right] &= m + (n - mk)\mathbb{P}\left(A_{mk+1} = i\right) \\ &\leq m + (n - mk)\mathbb{P}\left(\hat{\mu}_i(mk) \geq \max_{j \neq i} \hat{\mu}_j(mk)\right) .\end{aligned}$$

物理意义是前m*k轮确定被选择了m次，mk+1到n轮的次数则由轮数乘以该动作被选择的概率决定。

因为我们不失一般性地将第一个动作假设为最优动作，结合第i个动作的次优间隙的定义可以得到如下不等式：

$$\begin{aligned}\mathbb{P}\left(\hat{\mu}_i(mk) \geq \max_{j \neq i} \hat{\mu}_j(mk)\right) &\leq \mathbb{P}\left(\hat{\mu}_i(mk) \geq \hat{\mu}_1(mk)\right) \\ &= \mathbb{P}\left(\hat{\mu}_i(mk) - \mu_i - (\hat{\mu}_1(mk) - \mu_1) \geq \Delta_i\right) .\end{aligned}$$

之后为了下面用霍夫定界放缩，我们将上式变形：

代入*式

$$\begin{aligned}\Rightarrow (*) &= \mathbb{P}\Big(\frac{1}{m}\sum_{s=1}^{m} X_{k(s-1)+i} - \frac{1}{m}\sum_{s=1}^{m} X_{k(s-1)+1} - (\mu_i - \mu_1) \geq \Delta_i\Big) \\ &= \mathbb{P}\Big(\frac{1}{m}\sum_{s=1}^{m}(X_{k(s-1)+i} - X_{k(s-1)+1}) - \frac{1}{m}\sum_{s=1}^{m}(\mu_i - \mu_1) \geq \Delta_i\Big) \\ &= \mathbb{P}\Big(\sum_{s=1}^{m}(X_{k(s-1)+i} - X_{k(s-1)+1}) - \sum_{s=1}^{m}(\mu_i - \mu_1) \geq m\Delta_i\Big) \\ &= \mathbb{P}\Big(\sum_{s=1}^{m}\underbrace{(X_{k(s-1)+i} - X_{k(s-1)+1} - (\mathbb{E}[X_{k(s-1)+i}] - \mathbb{E}[X_{k(s-1)+1}])}_{1-次高斯} \geq m\Delta_i\Big)\end{aligned}$$

根据假设我们知道奖励是1-次高斯的，因此用霍夫丁引理放缩得到：

$$\mathbb{P}\left(\hat{\mu}_i(mk) - \mu_i - \hat{\mu}_1(mk) + \mu_1 \geq \Delta_i\right) \leq \exp\left(-\frac{m\Delta_i^2}{4}\right) .$$

最后结合Rn的定义式能够得到：

$$
\begin{aligned}
R_n &= \sum_{i=1}^{k} \Delta_i \, \mathbb{E}(T_i(n)) \\
&\leq \sum_{i=1}^{k} \Delta_i \left[m + (n-km) \exp\left(-\frac{m\Delta_i^2}{4}\right) \right] \\
&= m \sum_{i=1}^{k} \Delta_i + (n-mk) \sum_{i=1}^{k} \exp\left(-\frac{m\Delta_i^2}{4}\right)
\end{aligned}
$$

证明完毕

总结

ETC算法是在线学习Bandit问题的一个相对简单的算法，本文小结介绍explore-then-commit算法，并对其遗憾上界进行推导。