UCB алгоритм

2021

Гунаев Руслан, 774 группа

1 Верхняя оценка ожидаемого регрета UCB алгоритма

Пусть $R(T)$ – регрет UCB алгоритма для некоторого многорукого бандита, тогда для любого T верна верхняя оценка

$$\boldsymbol{E}[R(T,\Theta)] \leq \sum_{i:\mu_i<\mu^*} \frac{4\ln T}{\Delta_i} + 8\Delta_i,\ \Delta_i = \mu^* - \mu_i.$$

1.1 Предисловие

Положим $n_{i,t}$ – количество раз, когда была сыграна ручка i до момента времени t. r_t – награда, которую мы получаем в момент времени t. $I_t \in \{1, 2, \dots, N\}$ – выбранная ручка в момент времени t. Эмпирическая оценка награды $\mathrm{arm}\, i$ в момент t:

$$\hat{\mu}_{i,t} = \frac{\sum_{s=1:I_s=i}^{l} r_s}{n_{i,t}}.$$

Регрет задается следующим образом

$$R(T) = T\mu^\star - \sum_{t=0}^{T-1} \mathbb{E}\left[r^t\right]$$

UCB присваивает каждой ручке в каждый момент времени следующее значение:

$$\mathrm{UCB}_{i,t} := \hat{\mu}_{i,t} + \sqrt{\frac{\ln t}{n_{i,t}}}$$

Алгоритм задается следующим образом:

Algorithm 1: UCB algorithm

Data: N arms, number of rounds $T \geq N$
for $t = 1 \dots N$ **do**
  play arm t
end
for $t = N+1 \dots T$ **do**
  play arm

$$I_t = \underset{i\in\{1\dots N\}}{\arg\max} \mathrm{UCB}_{i,t-1}$$

end

1.2 Доказательство

Есть более фундаментальная причина выбора $\sqrt{\frac{\ln t}{n_{i,t}}}$. Эта верхняя оценка вытекает из неравенства Чернова-Хоффдинга. Для каждой ручки верно

$$|\hat{\mu}_{i,t} - \mu_i| < \sqrt{\frac{\ln t}{n_{i,t}}}$$

с вероятностью не меньше $1 - 2/t^2$. Их этого получаем два важных неравенства:

1. Нижняя граница для $\mathbf{UCB}_{i,t}$. С вероятностью не меньше $1 - 2/t^2$,

$$\mathrm{UCB}_{i,t} > \mu_i$$

2. Верхняя граница для $\hat{\mu}_{i,t}$ с большим числом семплов. При $n_{i,t} \geq \frac{4\ln t}{\Delta_i^2}$, с вероятностью не меньшей $1 - 2/t^2$ верно,

$$\hat{\mu}_{i,t} < \mu_i + \frac{\Delta_i}{2}$$

1 показывает, что значение UCB, вероятно, равно истинному вознаграждению: в этом смысле алгоритм UCB оптимистичен. 2 – что при наличии достаточного количества (а именно, по крайней мере, $\frac{4\ln t}{\Delta_i^2}$) семплов оценка вознаграждения, вероятно, не превышает истинное вознаграждение более чем на $\Delta_i/2$. Эти ограничения показывают, что алгоритм быстро находит субоптимальную ручку.

Лемма 1.1. В любой момент времени t, если субоптимальная ручка i ($\mu_i < \mu^*$) была сыграна $n_{i,l} \geq \frac{4\ln t}{\Delta_i^2}$ раз, тогда $UCB_{i,t} < UCB_{I^*,t}$ с вероятностью $1 - 4/t^2$. Это значит, что любого t,

$$P\left(I_{t+1} = i \mid n_{i,t} \geq \frac{4\ln t}{\Delta_i^2}\right) \leq \frac{4}{t^2}$$

Доказательство,

$$\begin{aligned}
\mathrm{UCB}_{i,t} = \hat{\mu}_{i,t} + \sqrt{\frac{\ln t}{n_{i,t}}} \leq \hat{\mu}_{i,t} + \frac{\Delta_i}{2} & \qquad \text{since } n_{i,t} \geq \frac{4\ln L}{\Delta_i^2} \\
< \left(\mu_i + \frac{\Delta_i}{2}\right) + \frac{\Delta_i}{2} & \\
= \mu^* \qquad \text{since } \Delta_i := \mu^* - \mu_i & \\
< \mathrm{UCB}_{i^*,t} &
\end{aligned}$$

Лемма 1.2. Пусть $n_{i,T}$ – количество раз, когда ручка i была выбрана алгоритмом . Тогда для любой ручки с $\mu_i < \mu^*$,

$$\mathbb{E}\left[n_{i,T}\right] \leq \frac{4\ln T}{\Delta_i} + 8$$

Доказательство. Для любой ручки i ожидаемое число раз, когда она была сыграна

$$
\begin{aligned}
\mathbf{E}\left[n_{i,T}\right] &= 1 + \mathbb{E}\left[\sum_{t=N}^{T} \mathbb{I}\left(I_{t+1} = i\right)\right] \\
&= 1 + \mathbb{E}\left[\sum_{t=N}^{T} \nVdash \left(I_{t+1} = i, n_{i,t} < \frac{4\ln t}{\Delta_i^2}\right)\right] + \mathbb{E}\left[\sum_{t=N}^{T} \mathbb{I}\left(I_{t+1} = i, n_{i,t} \geq \frac{4\ln t}{\Delta_i^2}\right)\right. \\
&\leq \frac{4\ln T}{\Delta_i^2} + \mathbb{E}\left[\sum_{t=N}^{T} \nVdash \left(I_{l+1} = i, n_{i,t} \geq \frac{4\ln t}{\Delta_i^2}\right)\right] \\
&= \frac{4\ln T}{\Delta_i^2} + \sum_{t=N}^{T} P\left(I_{l+1} = i, n_{i,l} \geq \frac{4\ln t}{\Delta_i^2}\right) \\
&= \frac{4\ln T}{\Delta_i^2} + \sum_{t=N}^{T} P\left(I_{t+1} = i \mid n_{i,t} \geq \frac{4\ln T}{\Delta_i^2}\right) P\left(n_{i,t} \geq \frac{4\ln t}{\Delta_i^2}\right) \\
&\leq \frac{4\ln T}{\Delta_i^2} + \sum_{t=N}^{T} \frac{4}{t^2} \\
&\leq \frac{4\ln T}{\Delta_i^2} + 8
\end{aligned}
$$

Тогда пользуясь леммами, итоговый ожидаемый регрет до времени T:

$$
\mathbf{E}[R(T,\Theta)] = \sum_{i:\mu_i<\mu^*} \mathbf{E}\left[n_{i,T}\right] \Delta_i \leq \sum_{i:\mu_i<\mu^*} \frac{4\ln T}{\Delta_i} + 8\Delta_i
$$