OPNS530 Sequential Learning Lecture 5 - April 12

Gaussian UCB Regret Bound Proof

Lecturer: Daniel Russo *Scribe: Daniel Russo*

1 Gaussian independent arm bandit problem

Consider a bandit problem with k actions $[k] = \{1, ..., k\}$. Each action i is associated with an unknown mean θ_i that is drawn from a prior $\theta \sim N(\mu_{1,i}, \sigma_{1,i}^2)$. At each time t the algorithm selects an action $I_t \in [k]$ and observes a reward

$$R_t = \theta_{I_t} + W_t$$

where $W_1, W_2, ... \overset{i.i.d}{\sim} N(0, \sigma^2)$. Let

$$H_t = (I_1, R_1, ...I_t, R_t)$$

denote the history of observations up to time t. The posterior distribution θ_i at based on observations before to time t is still Gaussian, with

$$\theta_i | H_{t-1} \sim N(\mu_{t,i}, \sigma_{t,i}^2).$$

The posterior parameters $(\mu_{t,i}, \sigma_{t,i})$ have a simple closed form. $N_{t,i} = \sum_{\ell=1}^{t-1} \mathbf{1}_{\{I_\ell = i\}}$ and $\hat{\mu}_{t,i} = N_{t,i}^{-1} \sum_{\ell=1}^{t-1} \mathbf{1}_{\{I_\ell = i\}} R_\ell$ denote respectively the number of times i has been sampled prior to time t and the empirical mean of these samples. Then

$$\sigma_{t,i}^2 = \left(\frac{1}{\sigma_{1,i}^2} + \frac{N_{t,i}}{\sigma^2} \right)^{-1}$$

and

$$\mu_{t,i} = \sigma_{t,i}^2 \left(\frac{\mu_{1,i}}{\sigma_{1,i}^2} + \frac{N_{t,i} \cdot \hat{\mu}_{t,i}}{\sigma^2} \right).$$

Notice that when $\sigma_{t,i}^2 \leq \sigma^2 / N_{t,i}$.

1.1 Gaussian UCB

In class, we introduced a Gaussian UCB algorithm. At each time step n, this method chooses the action

$$I_t \in \arg\max_{i \in [k]} \mu_{t,i} + \beta \sigma_{t,i}$$

where $\beta \in \mathbb{R}_+$ is a tuning parameter.

1.2 Regret bound

In class, we discussed the following result.

Theorem 1. *Suppose there is a common prior standard deviation $\sigma_{1,i} = \sigma_1$ for all $i \in [k]$. If Gaussian UCB is applied with parameter $\beta = \sqrt{2 \log\left(\frac{T\sigma_1}{\sqrt{2\pi}}\right)}$, then*

$$\mathbb{E} \sum_{t=1}^{T} \left(\max_{i \in [k]} \theta_i - \theta_{I_t} \right) \leq k(1 + \sigma_1 \beta) + 2\beta\sigma\sqrt{kT}$$

2 Proof of Regret Bound

2.1 Two Basic Facts

Fact 2. *If $X \sim N(\mu, \sigma^2)$ with $\mu < 0$ then*

$$\mathbb{E}[X\mathbf{1}(X \geq 0)] = \frac{\sigma}{\sqrt{2\pi}} e^{-\mu/2\sigma^2}.$$

This follows directly from integrating the Guassian density.

Fact 3.

$$\sum_{i=1}^{L} \frac{1}{\sqrt{i}} \leq \int_0^L \frac{1}{\sqrt{x}} dx = 2\sqrt{L}$$

2.2 Notation

- $U_{t,i} := \mu_{t,i} + \beta\sigma_{t,i}$
- $H_t := (I_1, R_1, ..., I_t, R_t)$
- $I^* := \arg\max_i \theta_i$

It is worth emphasizing that I_t and $(U_{t,i})_{i\in[k]}$ are known given H_{t-1} (i.e. they can be written as functions of H_{t-1}). The index I^* of the optimal action is itself a random variable as it is a function of $\theta = (\theta_1, ..., \theta_k)$.

2.3 Regret Decomposition

Let us isolate regret under a single period t. One has

$$\begin{aligned} \theta_{I^*} - \theta_{I_t} &= \theta_{I^*} - U_{t,I_t} + U_{t,I_t} - \theta_{I_t} \\ &\leq \underbrace{\theta_{I^*} - U_{t,I^*}}_{\text{typically } \leq 0} + \underbrace{U_{t,I_t} - \theta_{I_t}}_{\text{optimism at } I_t}. \end{aligned}$$

where the inequality follows because $U_{t,I_t} \geq U_{t,I^*}$ (by definition I_t maximizes the upper confidence bound). Overall then, we find

$$\begin{aligned} \mathbb{E}\sum_{t=1}^{T}(\theta_{I^*} - \theta_{I_t}) &\leq \mathbb{E}\sum_{t=1}^{T}(\theta_{I^*} - U_{t,I^*}) + \mathbb{E}\sum_{t=1}^{T}(U_{t,I_t} - \theta_{I_t}) \\ &= \mathbb{E}\sum_{t=1}^{T}(\theta_{I^*} - U_{t,I^*}) + \mathbb{E}\sum_{t=1}^{T}\mathbb{E}[U_{t,I_t} - \theta_{I_t}|H_{t-1}] \\ &= \underbrace{\mathbb{E}\sum_{t=1}^{T}(\theta_{I^*} - U_{t,I^*})}_{A} + \underbrace{\mathbb{E}\sum_{t=1}^{T}(U_{t,I_t} - \mu_{t,I_t})}_{B}. \end{aligned}$$

2.4 Bounding term B

Since $U_{t,I_t} - \mu_{t,I_t} = \beta\sigma_{t,I_t}$, we have that

$$B \leq \beta \sum_{t=1}^{T} \sigma_{t,I_t}.$$

Let $T_i = \{t \leq T | I_t = i\}$ denote the set of periods in which action i was chosen. Since $\sigma^2_{t,I_t} \leq \sigma^2/N_{t,i}$ we have

$$\begin{aligned} B \leq \beta \sum_{i=1}^{k} \sum_{t \in T_i} \sigma_{t,i} \leq \beta \sum_{i=1}^{k} \left(\sigma_1 + \sum_{\ell=1}^{|T_i|} \frac{1}{\sqrt{\ell}} \right) \leq \beta k \sigma_1 + 2\beta \sum_{i=1}^{k} \sqrt{|T_i|} &\leq \beta k \sigma_1 + 2\beta \sqrt{k \sum_{i=1}^{k} |T_i|} \\ &= \beta k \sigma_1 + 2\beta \sqrt{kT} \end{aligned}$$

where the second to last step uses Cauchy-Schwartz.

2.5 Bounding term A

Now we show term A is less than the number of arms k. Since $\theta_i - U_{t,i} | H_{t-1} \sim N(-\beta \sigma_{t,i}, \sigma^2_{t,i})$, by Fact 2,

$$\mathbb{E}[(\theta_i - U_{t,i})_+ | H_{t-1}] = \frac{\sigma_{t,i}}{\sqrt{2\pi}} e^{-\beta^2/2} \leq \frac{\sigma_1}{\sqrt{2\pi}} e^{-\beta^2/2} = \frac{1}{T}.$$

Therefore

$$\begin{aligned} \mathbb{E} \sum_{t=1}^{T} (\theta_{I_t} - U_{t,I_t}) \leq \mathbb{E} \sum_{t=1}^{T} (\theta_{I_t} - U_{t,I_t})_+ &\leq \mathbb{E} \sum_{t=1}^{T} \sum_{i=1}^{k} (\theta_i - U_{t,i})_+ \\ &\leq \mathbb{E} \sum_{t=1}^{T} \sum_{i=1}^{k} \mathbb{E}[(\theta_i - U_{t,i})_+ | H_{t-1}] \\ &\leq k \sum_{t=1}^{T} \frac{1}{T} = k. \end{aligned}$$