

# SOME ALGORITHMS FOR CORRELATED BANDITS WITH NON-STATIONARY REWARDS : REGRET BOUNDS AND APPLICATIONS

Prathamesh Mayekar  
joint work with  
Prof N Hemachandra

Industrial Engineering and Operations Research  
Indian Institute of Technology Bombay

January 19, 2018

# Overview I

- 1 Introduction
- 2 Framework for correlated and stationary bandits
- 3 Greedy Algorithm
  - Analysis of greedy algorithm
  - Illustration of theoretical upper bounds
- 4 Bibliography
- 5

# Classical Multi-armed bandit problem

- Gambler needs to decide each arm to play at each time instant
- Each machine has different reward distribution unknown to the gambler
- Objective of the gambler is to maximize his total reward

# Our Variant of the Multi-armed bandit problem

- Set of bandit arms  $N \equiv \{1, 2 \dots n\}$  have a corresponding feature  $p_i$
- Expected reward of arm  $i$  is  $\mu_i$
- Reward of arm  $i$  at time  $t$ ,

$$X_i(t) = p_i \times D(p_i, t)$$

- $D(p_i, t)$  is a function of  $p_i$  and time  $t$  unknown to the user
- $D(p_i, t)$  leads to a non-stationary bandit problem with dependent arms

# Reward Structure of arms

$$D(p_i, t) = N(1 - F(p_i, t)) + \epsilon(t)$$

Hence,  $\mu_i(t) = p_i N(1 - F(p_i, t))$ .

- $N$  is a constant which remains same across all the arms
- $\epsilon(t)$  corresponds to a residual error term ,
  - $\epsilon(t)$  has mean 0
  - $\epsilon(t)$  is i.i.d across time periods as well as arms

# Reward Structure of arms

- $F(p, t)$  is given as follows

$$F(p, t) = 0 \quad \forall p \leq 0$$

$$F(p, t) = \frac{p}{b(t)} \quad \forall 0 \leq p \leq b(t)$$

$$F(p, t) = 1 \quad \forall p \geq b(t)$$

- Non-stationarity arises as,  $b(t)$  changes in a piece-wise constant manner at unknown time points (break points)
- Assumption  $0 \leq p_i \leq b(t) \quad \forall p_i \in \mathcal{P} \quad \forall t$

# Regression based sliding window approach

- To account for non-stationarity only the latest  $\tau$  readings are considered
- Parameters  $N, b(t)$  are same across all n-arms
- Estimate the rewards for all the arms by estimating the parameters as follows,

$$(\bar{N}_t, \bar{b}_t(t)) = \underset{N, b}{\operatorname{argmin}} \sum_{s=\max(t-\tau, 0)+1}^t (d(p_{I(s)}, s) - N(1 - \frac{p_{I(s)}}{b(t)}))^2$$

$\bar{N}_t, \bar{b}_t(t)$  denote the estimates of  $N$  and  $b(t)$  at time  $t$   
 $d(p_{I(s)}, s)$  represents the demand obtained at time  $s$   
 $I(s)$  represents the arm played at time  $s$

# Greedy Algorithm

- No padding function
- Plays the arm which has highest reward estimate
- Does not need the error term  $\epsilon(t)$  to be truncated normal



# Greedy Algorithm

**INITIALIZATION:** Play any two distinct arms in first two time periods;

**for**  $t \leq T$  **do**

**if** *all observations in the window are from one particular arm* **then** **then**

        | play any other arm randomly

**end**

**else**

$$\bar{N}_t, \bar{b}_t(t) = \operatorname{argmin}_{N,b} \sum_{s=\max(t-\tau,0)+1}^{\min(t,\tau)} (d(p_{I(s)}, s) - N(1 - \frac{p_{I(s)}}{b(t)})^2)$$

**Determine**(reward estimate  $\bar{x}_i(t)$ )

**for each arm**  $i$  **do** **do**

$$\bar{x}_i(t) = p_i \times \bar{N}(1 - \frac{p_i}{b(t)});$$

**end**

    Play the arm which has maximum  $\bar{x}_i(t)$ ;

**end**

$t = t + 1$ ;

**end**

## Theorem (Greedy algorithm worst case bounds)

*The Expected number of times a non optimal arm  $i \in \mathcal{N}$  is played, when the Greedy algorithm is applied on  $k$  bandit arms, fixed time horizon  $T$  with a sliding window of length  $\tau$  is bounded as follows :*

$$E[\tilde{N}_T(i)] \leq 1 + \min_{m_1 \in \{1, \dots, \tau-2\}} \left[ m_1 + \sum_{t=3}^{\tau} 2 \times E_{\chi^2(t-2)} \left( \frac{\alpha_{\max N}(\sigma^2 \times \chi^2(t-2), m_1)}{2} \right) \right] \\ + \min_{m_2 \in \{1, \dots, \tau-1\}} \left[ \left( \left\lceil \frac{T}{\tau} \right\rceil - 1 \right) m_2 + 2 \times (T - \tau) \times E_{\chi^2(\tau-2)} \left( \frac{\alpha_{\max N}(\sigma^2 \times \chi^2(\tau-2), m_2)}{2} \right) \right] + \gamma_T \tau$$

where  $\gamma_T$  are the number of breakpoints,  $\chi^2(\tau-2)$  is chi-square distribution with  $\tau-2$  degrees of freedom and

$$\alpha_{\max}(\sigma^2 \times \chi^2(\tau-2), m) = 2 \times P(t(RV)_{\tau-2} \geq \frac{\Delta \mu_T(i)}{2 \times p_i \times \sqrt{\frac{\sigma^2 \times \chi^2(\tau-2)}{\tau-2}} \times \sqrt{\frac{1}{\tau} + \frac{1}{m}}}).$$

(here  $t(RV)_{\tau-2}$  is a  $t$  random variable with  $\tau-2$  degrees of freedom) and

$$\Delta \mu_i(T) = \min_{t \in \{1, \dots, T\}} (\mu_{i^*}(t) - \mu_i(t) : i \neq i^*).$$

# Interpretation of the upper bound

- Bound can be viewed as a sum of stationary and non-stationary part
- Non-Stationary Part
  - At most  $\gamma_T \tau$  decision points contain data from before and after the breakpoint
  - $\gamma_T \tau$  is used to bound this part

# Interpretation of the upper bound

## • Stationary Part

- 1 is used to upper bound the number of times arm  $i$  is played in first two time periods
- $\min_{m_1 \in \{1, \dots, \tau-2\}} \left[ m_1 + \sum_{t=3}^{\tau} 2 \times E_{\chi^2(t-2)} \left( \frac{\alpha_{\max N} (\sigma^2 \times \chi^2(t-2), m_1)}{2} \right) \right]$  is an upper bound from the time period  $t = 3$  to  $t = \tau$
- The third term,  $\min_{m_2 \in \{1, \dots, \tau-1\}} \left[ \left( \lceil \frac{T}{\tau} \rceil - 1 \right) m_2 + 2 \times (T - \tau) \times E_{\chi^2(\tau-2)} \left( \frac{\alpha_{\max N} (\sigma^2 \times \chi^2(\tau-2), m_2)}{2} \right) \right]$  upper bounds from the time period  $t = \tau + 1$  to  $t = T$
- second and third term monotonically increase with decrease in  $\Delta\mu_i(T)$   
 $\Delta\mu_i(T) = \min_{t \in \{1, \dots, T\}} (\mu_{i^*}(t) - \mu_i(t) : i \neq i^*)$

# Bounds for the stationary case

With length of window  $\tau$  as  $t$

$$E[\tilde{N}_T(i)] \leq 1 + \min_{m_1 \in \{1, \dots, T-2\}} \left[ m_1 + \sum_{t=3}^T 2 \times E_{\chi^2(t-2)} \left( \frac{\alpha_{\max N}(\sigma^2 \times \chi^2(t-2), m_1)}{2} \right) \right]$$

No terms of the form  $kT$

# Problem setting

- Time horizon  $T=130$
- Arm set  $\mathcal{N} \equiv \{1, 2, 3\}$ , Feature set  $\mathcal{P} \equiv \{2, 3, 4\}$
- $b(t)$  varies over time as follows
  - $b(t) = 5.5 \quad \forall t \leq 40$
  - $b(t) = 4.5 \quad \forall t \geq 40 \quad \text{and} \quad t \leq 90$
  - $b(t) = 9.0 \quad \forall t \geq 90$
- $N = 800, \epsilon(t) = N(0, 10^2)$
- Length of sliding window  $\tau = 20$ .
- Expected Reward

$\mu_i$	$t \leq 40$	$40 \leq t \leq 90$	$t \geq 90$
$\mu_1$	1018.18	888.88	1244.44
$\mu_2$	1090.9090	800	1600
$\mu_3$	1090.9090	355.56	1777.77

# Input parameters for computation of upper bound

## ① Common parameters for all arms

- ① Number of arms, 3
- ② Time horizon,  $T = 130$
- ③ Number of breakpoints,  $\gamma_T = 2$
- ④  $\sigma^2 = 10^2$

## ② Parameters specific to the arms

Parameters	Arm 1	Arm 2	Arm 3
$p_i$	2	3	4
$\Delta\mu_1$	72.729	88.85	218.182

# Bibliography I



# Thank You!!!

## Theorem (Greedy algorithm worst case bounds)

*The Expected number of times a non optimal arm  $i$  is played, when the Greedy algorithm is applied on  $k$  bandit arms, fixed time horizon  $T$  with a sliding window of length  $\tau$  is bounded as follows :*

$$\begin{aligned}
 E[\tilde{N}_T(i)] &\leq \tau - 1 + \gamma_T \tau \\
 &\quad + \min_{m \in \{1, \dots, \tau-1\}} \left[ \left( \left\lceil \frac{T}{\tau} \right\rceil - 1 \right) m \right. \\
 &\quad \left. + 2 \times (T - \tau) \times E_{\chi^2(\tau-2)} \left( \frac{\alpha_{\max N}(\sigma^2 \times \chi^2(\tau-2), m)}{2} \right) \right]
 \end{aligned}$$

where  $\gamma_T$  are the number of breakpoints and

$$\alpha_{\max}(\sigma^2 \times \chi^2(\tau-2), m) = 2 \times P(t(RV)_{\tau-2} \geq \frac{\Delta\mu_T(i)}{2 \times p_i \times \sqrt{\frac{\sigma^2 \times \chi^2(\tau-2)}{\tau-2}} \times \sqrt{\frac{1}{\tau} + \frac{1}{m}}}).$$

Also  $\Delta\mu_i(T) = \min_{t \in \{1, \dots, T\}} (\mu_{i^*}(t) - \mu_i(t) : i \neq i^*)$ .

- ① Number of times a non-optimal arm  $i$  is played can be divided as follows:

$$\tilde{N}_T(i) = \sum_{t=1}^2 1_{\{I(t)=i \neq i_t^*\}} + \sum_{t=3}^{\tau} 1_{\{I(t)=i \neq i_t^*\}} + \sum_{t=\tau+1}^T 1_{\{I(t)=i \neq i_t^*\}}$$

$$\tilde{N}_T(i) \leq 1 + (\tau - 2) + \sum_{t=\tau+1}^T 1_{\{I(t)=i \neq i_t^*\}}$$

(Since any two arms are played in the first two time periods)

Recollect that  $I(t)$  represents the arm played at time  $t$ . Also  $i_t^*$  represents the optimal arm at time  $t$ .

Now,

$$\begin{aligned} \sum_{t=\tau+1}^T 1_{\{I(t)=i \neq i_t^*\}} &= \sum_{t=\tau+1}^T 1_{\{I(t)=i \neq i_t^*, n_i(t, \tau) \leq m\}} \\ &\quad + \sum_{t=\tau+1}^T 1_{\{I(t)=i \neq i_t^*, n_i(t, \tau) > m\}} \end{aligned}$$

Here  $n_i(t, \tau)$  is the number of times the arm has been played in the last  $\tau$  periods, where for  $t > \tau$

$$n_i(t, \tau) = \sum_{s=(t-\tau+1)}^t 1_{\{I(s) = i\}}$$

for  $t \leq \tau$ ,

$$n_i(t, \tau) = \sum_{s=1}^t 1_{\{I(s) = i\}}$$

$$\begin{aligned} \therefore \tilde{N}_T(i) = \tau - 1 + \sum_{t=\tau+1}^T 1_{\{I(t)=i \neq i_t^*, n_i(t,\tau) \leq m\}} \\ + \sum_{t=\tau+1}^T 1_{\{I(t)=i \neq i_t^*, n_i(t,\tau) > m\}} \quad (1) \end{aligned}$$

- ② Using lemma 4 the first term in the right hand side of inequality 1 can be bounded as follows

$$\therefore \sum_{t=\tau+1}^T 1_{\{I(t)=i \neq i_t^*, n_i(t,\tau) \leq m\}} \leq (\lceil \frac{T}{\tau} \rceil - 1)m \quad (2)$$

- ③ The second term in the right hand side of inequality 1 can be bounded as follows

$$\sum_{t=\tau+1}^T 1_{\{I(t)=i \neq i_t^*, n_i(t, \tau) > m\}} \leq \gamma T^\tau + \sum_{t \in \mathbb{T}(\tau)} 1_{\{I(t)=i \neq i_t^*, n_i(t, \tau) > m\}} \quad (3)$$

where ,

$\mathbb{T}(\tau)$  is a set of indices such that  $t \in \{\tau, \dots, T\}$  has  $\mu_s(j) = \mu_t(j)$

$\forall t - \tau \leq s \leq t$

- Now  $\{I(t) = i \neq i_t^*, n_i(t, \tau) > m\}$  can be written as a subset of following events from lemma 5

$$\begin{aligned}
 \{I(t) = i \neq i_t^*, n_i(t, \tau) > m\} &\subset \\
 &\quad \{\bar{x}_i(t) \geq \bar{x}_i^*(t), n_i(t, \tau) > m\} \\
 &\subset \{\bar{x}_i(t) \geq \mu_t(i) + \frac{\mu_t(i^*) - \mu_t(i)}{2}, n_i(t, \tau) > m\} \\
 &\quad \cup \{\bar{x}_i^*(t) \leq \mu_t(i) - \frac{\mu_t(i^*) - \mu_t(i)}{2}, n_i(t, \tau) > m\} \quad (4)
 \end{aligned}$$

But

$$\{\bar{x}_i(t) \geq \mu_t(i) + \frac{\mu_t(i^*) - \mu_t(i)}{2}\} \subset \{\bar{x}_i(t) \geq \mu_t(i) + \frac{\Delta\mu_T(i)}{2}\}$$

and

$$\{\bar{x}_i^*(t) \leq \mu_t(i) - \frac{\mu_t(i^*) - \mu_t(i)}{2}\} \subset \{\bar{x}_i^*(t) \leq \mu_t(i) - \frac{\Delta\mu_T(i)}{2}\}$$

(recollect that  $\Delta\mu_i(T) = \min_{t \in \{1, \dots, T\}} (\mu_{i^*}(t) - \mu_i(t) : i \neq i^*)$ ).

Hence,

$$\begin{aligned} \{I(t) = i \neq i_t^*, n_i(t, \tau) > m\} &\subset \\ &\quad \{\bar{x}_i(t) \geq \bar{x}_i^*(t), n_i(t, \tau) > m\} \\ &\subset \{\bar{x}_i(t) \geq \mu_t(i) + \frac{\Delta\mu_T(i)}{2}, n_i(t, \tau) > m\} \\ &\quad \cup \{\bar{x}_i^*(t) \leq \mu_t(i) - \frac{\Delta\mu_T(i)}{2}, n_i(t, \tau) > m\} \quad (5) \end{aligned}$$

- We know that  $\epsilon(t)$  is normal random variable. The sum of squared error ( $sse_N$ ) for normal random variable is defined as follows,

$$sse_N := \sum_{t=\tau+1}^t (D(p_{I(s)}, s) - \bar{N}(1 - \frac{p_{I(s)}}{\bar{b}(t)}))^2 \quad (6)$$



- The mean square error mse is defined as follows,

$$mse := \sqrt{\frac{sse_N}{\tau - 2}}$$

- Now we interpret  $\frac{\Delta\mu_T(i)}{2}$  as the confidence interval for the normal distribution.

$$\frac{\Delta\mu_i(T)}{2} = p_i \times t_{\frac{\alpha}{2}, \min(t-2, \tau-2)} \times mse \times \sqrt{\frac{1}{\min(t, \tau)} + \frac{(p_i - p_{I(s)})^2}{\sum_{s=t-\tau+1}^t (p_i - p_{I(s)})^2}}$$

(Recollect that  $p_I(t, \tau)$  is the vector of all possible prices played in last  $\tau$  periods. Hence

$$p_I(t, \tau) = (p_{I(t-\tau+1)}, \dots, p_{I(t)})$$

and  $\bar{p}_I(t)$  is the mean of all the prices in  $p_I(t)$ .)

Since  $n_i(t, \tau) \geq m$ ,

we have

$$\frac{\Delta\mu_i(T)}{2} \leq p_i \times t_{\frac{\alpha}{2}, \min(t-2, \tau-2)} \times mse \times \sqrt{\frac{1}{\min(t, \tau)} + \frac{1}{m}} \quad (7)$$

- Since  $t \geq \tau$  we have

$$\therefore t_{\frac{\alpha}{2}, \tau-2} \geq \frac{\Delta\mu_i(T)}{2 \times p_i \times mse \times \sqrt{\frac{1}{\tau} + \frac{1}{m}}}$$

$$\begin{aligned} \frac{\alpha}{2} &= P(t(RV)_{\tau-2} \geq t_{\frac{\alpha}{2}, \tau-2}) \\ &\leq P(t(RV)_{\tau-2} \geq \frac{\Delta\mu_i(T)}{2 \times p_i \times \sqrt{\frac{sse_N}{\tau-2}} \times \sqrt{\frac{1}{\tau} + \frac{1}{m}}}) \quad (8) \end{aligned}$$

Let us denote the right hand side as,

$$\alpha_{\max}(sse_N, m) = 2 \times P(t(RV)_{\tau-2} \geq \frac{\Delta\mu_i(T)}{2 \times p_i \times \sqrt{\frac{sse_N}{\tau-2}} \times \sqrt{\frac{1}{\tau} + \frac{1}{m}}}) \quad (9)$$

where  $t(RV)_{\tau-2}$  is a random variable following  $t$  distribution with  $\tau - 2$  degrees of freedom.

• Now,

$$\begin{aligned}
 & P\{\bar{x}_i(t) \geq \mu_t(i) + \frac{\Delta\mu_i(T)}{2}, n_i(t, \tau) > m\} \\
 &= \int_{x=0}^{\infty} P\{\{\bar{x}_i(t) \geq \mu_t(i) + \frac{\Delta\mu_i(T)}{2}, n_i(t, \tau) > m\} | \{\frac{sse_N}{\sigma^2} = x\}\} f_{\frac{sse_N}{\sigma^2}}(x) dx \\
 &\quad \text{(By conditioning on the value of } \frac{sse_N}{\sigma^2} \text{)} \\
 &= \int_{x=0}^{\infty} P\{\bar{x}_i(t) \geq \mu_t(i) + \frac{\Delta\mu_i(T)}{2}, n_i(t, \tau) > m\} | \{\chi^2(\tau-2) = x\}\} f_{\chi^2(\tau-2)}(x) dx \\
 &\quad \text{(We know that when error term is } N(0, \sigma^2), \frac{sse_N}{\sigma^2} \text{ in that case has the } \\
 &\quad \chi^2(\tau-2) \text{ distribution. (See Klimov [? ]))} \\
 &\leq \int_{x=0}^{\infty} P(t(RV)_{\tau-2} \geq \frac{\Delta\mu_i(T)}{2 \times p_i \times \sqrt{\frac{\sigma^2 \times x}{\tau-2}} \times \sqrt{\frac{1}{\tau} + \frac{1}{m}}}) \times f_{\chi^2(\tau-2)}(x) dx \\
 &\quad \text{(Since we interpret } \Delta\mu_i(T) \text{ as the confidence interval and by using inequality)} \\
 &= E_{\chi^2(\tau-2)}\left(\left(\frac{\alpha_{\max N}(\sigma^2 \times \chi^2(\tau-2), m)}{2}\right)\right)
 \end{aligned}$$

## Theorem (Worst case bounds for Regression Based UCB algorithm when the error term is normal)

*The Expected number of times a non optimal arm  $i$  is played, when the Regression Based UCB algorithm is applied on  $k$  bandit arms, fixed time horizon  $T$  with a sliding window of length  $\tau$  and level of significance  $\alpha = \frac{1}{\tau^4}$  is bounded as follows :*

$$E[\tilde{N}_T(i)] \leq \tau - 1 + 2 \times \frac{T - \tau}{\tau^4} + \gamma_T \tau$$

$$+ \min_{m \in \{1, \dots, \tau-1\}} [(T - \tau) \times P\left\{ \frac{\Delta\mu_i(T)^2}{\sigma^2 \times B(i, \tau)^2 \times (\frac{1}{\tau} + \frac{1}{m})} \leq \chi^2(\tau - 2) \right\} + (\lceil \frac{T}{\tau} \rceil - 1)m]$$

Here  $\gamma_T$  are the number of breakpoints,  $B(i, \tau) = (2 \times p_i \times t_{\frac{\alpha}{2}, \min(\tau-2, \tau-2)} \times \sqrt{\frac{1}{\tau-2}})$  and  $\Delta\mu_i(T) = \min_{t \in \{1, \dots, T\}} (\mu_{i^*}(t) - \mu_i(t) : i \neq i^*)$

- ① Number of times a non-optimal arm  $i$  is played can be divided as follows:

$$\tilde{N}_T(i) = \sum_{t=1}^2 1_{\{I(t)=i \neq i_t^*\}} + \sum_{t=3}^{\tau} 1_{\{I(t)=i \neq i_t^*\}} + \sum_{t=\tau+1}^T 1_{\{I(t)=i \neq i_t^*\}}$$

$$\tilde{N}_T(i) \leq 1 + (\tau - 2) + \sum_{t=\tau+1}^T 1_{\{I(t)=i \neq i_t^*\}}$$

(Since any two arms are played in the first two time periods)

Recollect that  $I(t)$  represents the arm played at time  $t$ . Also  $i_t^*$  represents the optimal arm at time  $t$ .

Now,

$$\begin{aligned} \sum_{t=\tau+1}^T \mathbf{1}_{\{I(t)=i \neq i_t^*\}} &= \sum_{t=\tau+1}^T \mathbf{1}_{\{I(t)=i \neq i_t^*, n_i(t, \tau) \leq m\}} \\ &\quad + \sum_{t=\tau+1}^T \mathbf{1}_{\{I(t)=i \neq i_t^*, n_i(t, \tau) > m\}} \end{aligned}$$

Here  $n_i(t, \tau)$  is the number of times the arm has been played in the last  $\tau$  periods, where for  $t > \tau$

$$n_i(t, \tau) = \sum_{s=(t-\tau+1)}^t \mathbf{1}_{\{I(s) = i\}}$$

for  $t \leq \tau$ ,

$$n_i(t, \tau) = \sum_{s=1}^t \mathbf{1}_{\{I(s) = i\}}$$

$$\begin{aligned} \therefore \tilde{N}_T(i) = \tau - 1 + \sum_{t=\tau+1}^T 1_{\{I(t)=i \neq i_t^*, n_i(t,\tau) \leq m\}} \\ + \sum_{t=\tau+1}^T 1_{\{I(t)=i \neq i_t^*, n_i(t,\tau) > m\}} \quad (17) \end{aligned}$$

- ② Using lemma 4 the first term in the right hand side of inequality 17 can be bounded as follows

$$\therefore \sum_{t=\tau+1}^T 1_{\{I(t)=i \neq i_t^*, n_i(t,\tau) \leq m\}} \leq (\lceil \frac{T}{\tau} \rceil - 1)m \quad (18)$$



- ③ The second term in the right hand side of equation 17 can be bounded as follows.

$$\sum_{t=\tau+1}^T \mathbf{1}_{\{I(t)=i \neq i_t^*, n_i(t, \tau) > m\}} \leq \gamma T \tau + \sum_{t \in \mathbb{T}(\tau)} \mathbf{1}_{\{I(t)=i \neq i_t^*, n_i(t, \tau) > m\}} \quad (19)$$

$(\mathbb{T}(\tau))$  is a set of indices such that  $t \in \{\tau + 1, \dots, T\}$  has  $\mu_s(j) = \mu_t(j)$   
 $\forall t - \tau \leq s \leq t$

- Now  $\{I(t) = i \neq i_t^*, n_i(t, \tau) > m\}$  can be written as a subset of following events by using lemma 5.

$$\begin{aligned} \{I(t) = i \neq i_t^*, n_i(t, \tau) > m\} \subset \\ \{\bar{x}_i(t) + PF_i(t, \tau, p_I(t, \tau)) \geq \bar{x}_{i^*}(\tau, i^*) + PF_{i^*}(\tau, \tau, p_I(t, \tau)), n_i(t, \tau) > m\} \\ \subset \{\bar{x}_i(t) \geq \mu_t(i) + PF_i(t, \tau, p_I(t, \tau))\} \cup \{\bar{x}_{i^*}(t) \leq \mu_t(i) - PF_{i^*}(\tau, i^*)\} \\ \cup \{\mu_t(i^*) - \mu_t(i) \leq 2PF_i(\tau, i), n_i(t, \tau) > m\} \quad (20) \end{aligned}$$

- Padding function for arm  $i=PF_i(t, \tau, p_I(t, \tau))$

$$\begin{aligned}
 & PF_i(t, \tau, p_I(t, \tau)) \\
 &= p_i \times t_{\frac{\alpha}{2}, \min(t-2, \tau-2)} \times mse \times \sqrt{\frac{1}{\min(t, \tau)} + \frac{(p_i - \bar{p}_{I(s)})^2}{\sum_{s=t-\tau+1}^t (p_i - \bar{p}_{I(s)})^2}} \\
 & \hspace{15em} (21)
 \end{aligned}$$

(Recollect that  $p_I(t, \tau)$  is the vector of all possible prices played in last  $\tau$  periods. Hence

$$p_I(t, \tau) = (p_{I(t-\tau+1)}, \dots, p_{I(t)})$$

and  $\bar{p}_I(t)$  is the mean of all the prices in  $p_I(t)$ .)

Since  $n_i(t, \tau) \geq m$

$$\sum_{s=t-\tau+1}^t (p_i - \bar{p}_{I(s)})^2 \geq m(p_i - \bar{p}_{I(s)})^2$$

$$\therefore PF_i(t, \tau, p_I(t, \tau)) \leq p_i \times t_{\frac{\alpha}{2}, \min(t-2, \tau-2)} \times mse \times \sqrt{\frac{1}{\min(t, \tau)} + \frac{1}{m}}$$

when  $t > \tau$

$$PF_i(t, \tau, p_I(t, \tau)) \leq p_i \times t_{\frac{\alpha}{2}, \tau-2} \times mse \times \sqrt{\frac{1}{\tau} + \frac{1}{m}} \quad (22)$$

- We know that  $\epsilon(t)$  is normal random variable. The sum of squared error is defined as

$$sse_N := \sum_{t=\tau+1}^t (D(p_{I(s)}, s) - \bar{N}(1 - \frac{p_{I(s)}}{b(t)}))^2 \quad (23)$$

We know that when error term is  $N(0, \sigma^2)$ , the sum of squared error in that case ( $sse_N$ ) has the  $\sigma^2 \times \chi^2(\tau - 2)$  distribution. See Klimov [?] .  
The mean square error  $mse$  is defined as

$$mse := \sqrt{\frac{sse_N}{\tau - 2}}$$

Now when  $t \in \mathbb{T}(\tau)$

$$\begin{aligned}
 & \{\mu_t(i^*) - \mu_t(i) \leq 2PF_i(t, \tau, p_I(t, \tau)), n_i(t, \tau) > m\} \\
 & \quad \subset \{\Delta\mu_i(T) \leq 2PF_i(t, \tau, p_I(t, \tau)), n_i(t, \tau) > m\} \\
 & \quad \subset \{\Delta\mu_i(T) \leq 2 \times p_i \times t_{\frac{\alpha}{2}, \min(t-2, \tau-2)} \times mse \times \sqrt{\frac{1}{\tau} + \frac{1}{m}}\} \\
 & \quad \quad \quad (\text{Because of equation 22}) \\
 & \equiv \{\Delta\mu_i(T) \leq 2 \times p_i \times t_{\frac{\alpha}{2}, \min(t-2, \tau-2)} \times \sqrt{\frac{sse_N}{\tau-2}} \times \sqrt{\frac{1}{\tau} + \frac{1}{m}}\} \quad (24)
 \end{aligned}$$

(recollect that  $\Delta\mu_i(T) = \min_{t \in \{1, \dots, T\}} (\mu_{i^*}(t) - \mu_i(t) : i \neq i^*)$ )

$$\begin{aligned}
\text{As } B(i, \tau) &= (2 \times p_i \times t_{\frac{\alpha}{2}, \min(t-2, \tau-2)} \times \sqrt{\frac{1}{\tau-2}}) \\
\{\Delta\mu_i(T) &\leq 2 \times p_i \times t_{\frac{\alpha}{2}, \min(t-2, \tau-2)} \times \sqrt{\frac{sse_N}{\tau-2}} \times \sqrt{\frac{1}{\tau} + \frac{1}{m}}\} \\
&\equiv \{\Delta\mu_i(T)^2 \leq B(i, \tau)^2 \times (\frac{1}{\tau} + \frac{1}{m}) \times sse_N\} \quad (25)
\end{aligned}$$

(Since everything is positive)

$$\begin{aligned}
\therefore P\{\mu_t(i^*) - \mu_t(i) \leq 2c_t(\tau, i)\} \\
&\leq P\left\{\frac{\Delta\mu_i(T)^2}{B(i, \tau)^2 \times (\frac{1}{\tau} + \frac{1}{m})} \leq sse_N\right\} \\
&\leq P\left\{\frac{\Delta\mu_i(T)^2}{B(i, \tau)^2 \times (\frac{1}{\tau} + \frac{1}{m})} \leq \sigma^2 \times \chi^2(\tau-2)\right\}
\end{aligned} \quad (26)$$

(See Klimov [?] et al. )



$$\begin{aligned}
 &P\{\bar{x}_i(t) \geq \mu_t(i) + PF_i(t, \tau, p_I(t, \tau))\} \\
 &= P\{\bar{x}_i(t) \leq \mu_t(i) - PF_i(t, \tau, p_I(t, \tau))\} = \frac{1}{\tau^4} \quad (27)
 \end{aligned}$$

(from the definition of confidence interval)

Similarly,

$$\begin{aligned}
 &P\{\bar{x}_{i^*}(t) \leq \mu_t(i) - PF_{i^*}(t, \tau, p_I(t, \tau))\} \\
 &= P\{\mu_t(i) \geq \bar{x}_{i^*}(t) + PF_{i^*}(t, \tau, p_I(t, \tau))\} \leq \frac{1}{\tau^4} \quad (28)
 \end{aligned}$$

- From equation 20 we have

$$\begin{aligned}
 P\{I(t) = i \neq i_t^*, n_i(t, \tau) > m\} &\leq P\{\bar{x}_i(t) \geq \mu_t(i) + PF_i(t, \tau, p_I(t, \tau))\} \\
 &\quad + P\{\bar{x}_{i^*}(t) \leq \mu_t(i) - PF_{i^*}(t, \tau, p_I(t, \tau))\} \\
 &\quad + P\{\mu_t(i^*) - \mu_t(i) \leq 2PF_i(t, \tau, p_I(t, \tau), n_i(t, \tau) > m\} \quad (29)
 \end{aligned}$$

From equations 26,27,28, when  $t \in \mathbb{T}(\tau)$

$$\begin{aligned}
 \therefore P\{I(t) = i \neq i_t^*, n_i(t, \tau) > m\} &\leq \frac{2}{\tau^4} \\
 &\quad + P\left\{\frac{\Delta\mu_i(T)^2}{\sigma^2 \times B(i, \tau)^2 \times \left(\frac{1}{\tau} + \frac{1}{m}\right)} \leq \chi^2(\tau - 2)\right\} \quad (30)
 \end{aligned}$$

From equation 19 we have

$$\mathbb{E}\left(\sum_{t=\tau+1}^T 1_{\{I(t)=i \neq i_t^*, n_i(t,\tau) > m\}}\right) \leq \gamma_T \tau + \mathbb{E}\left(\sum_{t \in \mathbb{T}(\tau)} 1_{\{I(t)=i \neq i_t^*, n_i(t,\tau) > m\}}\right) \quad (31)$$

Now from equation 30 we have

$$\begin{aligned} \mathbb{E}\left(\sum_{t=\tau+1}^T 1_{\{I(t)=i \neq i_t^*, n_i(t,\tau) > m\}}\right) &\leq \gamma_T \tau + \sum_{t \in \mathbb{T}} \left(\frac{2}{\tau^4}\right. \\ &\quad \left.+ P\left\{\frac{\Delta\mu_i(T)^2}{\sigma^2 \times B(i, \tau)^2 \times \left(\frac{1}{\tau} + \frac{1}{m}\right)} \leq \chi^2(\tau - 2)\right\}\right) \end{aligned} \quad (32)$$



④ From equation 17, 18, 32 we have

$$E[\tilde{N}_T(i)] \leq 1 + \max(\tau - k, 0) + (\lceil \frac{T}{\tau} \rceil - 1)m + \gamma_T \tau + 2 \times \frac{T - \tau}{\tau^4} \\ + (T - \tau) \times P\left\{ \frac{\Delta \mu_i(T)^2}{\sigma^2 \times B(i, \tau)^2 \times (\frac{1}{\tau} + \frac{1}{m})} \leq \chi^2(\tau - 2) \right\} \quad (33)$$

(Since  $|\mathbb{T}(\tau)| \leq T - \tau$ )

But this holds for every positive integer  $m$ . Note that the bound becomes trivial if  $m \geq \tau$ . Hence we have

$$E[\tilde{N}_T(i)] \leq \min_{m \in \{1, \dots, \tau-1\}} (1 + \max(\tau - k, 0) + (\lceil \frac{T}{\tau} \rceil - 1)m + \gamma_T \tau + 2 \times \frac{T - \tau}{\tau^4} \\ + (T - \tau) \times P\left\{ \frac{\Delta \mu_i(T)^2}{\sigma^2 \times B(i, \tau)^2 \times (\frac{1}{\tau} + \frac{1}{m})} \leq \chi^2(\tau - 2) \right\}) \quad (34)$$

## Lemma

*For an arm  $i \in \mathbb{N}$  and for a sliding window length  $\tau$ ,*

$$\begin{aligned} \sum_{t=\tau+1}^T 1_{\{I(t)=i, n_i(t,\tau) \leq m\}} &\leq \left(\lceil \frac{T}{\tau} \rceil - \lfloor \frac{\tau}{\tau} \rfloor\right)m \\ &\leq \left(\lceil \frac{T}{\tau} \rceil - 1\right)m \quad (35) \end{aligned}$$

## Proof.

$$\sum_{t=\tau+1}^T 1_{\{I(t)=i, n_i(t, \tau) \leq m\}} \leq \sum_{j=\lfloor \frac{T}{\tau} \rfloor + 1}^{\lceil \frac{T}{\tau} \rceil} \sum_{t=(j-1)\tau+1}^{j\tau} 1_{\{I(t)=i, n_i(t, \tau) \leq m\}}$$

Now,  $\sum_{t=(j-1)\tau+1}^{j\tau} 1_{\{I(t)=i, n_i(t, \tau) \leq m\}}$  is non-zero if for some  $t \in (j-1)\tau+1, \dots, (j)\tau+1$ , we have  $I(t) = i, n_i(t, \tau) \leq m$ .

Let  $t_j$  be the maximum time this happens in  $j^{\text{th}}$  interval

$$\therefore t_j = \max(t \in (j-1)\tau+1, \dots, (j)\tau+1 : I(t) = i, n_i(t, \tau) \leq m)$$

$$\begin{aligned} \therefore \sum_{t=(j-1)\tau+1}^{j\tau} 1_{\{I(t)=i, n_i(t, \tau) \leq m\}} &= \sum_{t=(j-1)\tau+1}^{t_j} 1_{\{I(t)=i, n_i(t, \tau) \leq m\}} \\ &\leq \sum_{t=t_j-\tau+1}^{t_j} 1_{\{I(t)=i, n_i(t, \tau) \leq m\}} \leq \sum_{t=t_j-\tau+1}^{t_j} 1_{\{I(t)=i\}} = n_i(t, \tau) \leq m \end{aligned}$$

□

## Lemma

*If*

$$D \cap A^c \cap B^c \subset C$$

*than*

$$D \subset A \cup B \cup C$$

Proof.

$$D \cap A^c \cap B^c \subset C$$

$$\therefore D \cap (A \cup B)^c \subset C$$

$$\therefore (D \cap (A \cup B)^c) \cup (A \cup B) \subset C \cup A \cup B$$

$$\therefore (D \cup (A \cup B)) \cap ((A \cup B)^c \cup (A \cup B)) \subset C \cup A \cup B$$

$$\therefore D \cup (A \cup B) \subset C \cup A \cup B$$

But

$$D \subset D \cup (A \cup B)$$

$$\therefore D \subset C \cup A \cup B$$



# Computation of expected regret

- The computation of expected regret is done by averaging the number of times a each arm is played at each time period over 100 replications
- This value of fraction of times a each arm is played at each time period is multiplied by the regret associated with that arm at that time period
- Summing over the value of regret of all arms at that time period we get the instantaneous value of regret at that time period
- Summing over all such values upto that time period, we get the value of expected regret