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

May 9, 2018

Overview I

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 $N \equiv \{1, 2 ... n\}$ have a corresponding feature p_i
- Expected reward of arm i is μ_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

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

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

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

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

Regression based sliding window approach

- To account for non-stationarity only the latest τ 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,

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

$\bar{N}_t$,$\bar{b}_t(t)$ 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
        $\bar{N}_t, \bar{b}_t(t) = argmin_{N,b} \sum_{s=max(t-\tau,0)+1}^{min(t,\tau)} (d(p_{I(s)}, s) - N(1 - \frac{p_{I(s)}}{b(t)})^2$
        Determine(reward estimate $\bar{x}_i(t)$)
        for *each arm i do* **do**
            $\bar{x}_i(t) = p_i \times \tilde{N}(1 - \frac{p_i}{b(\tilde{t})});$
        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 τ is bounded as follows :

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

$$+min_{m_2 \in \{1,..,\tau-1\}} \left[\left(\left\lceil \frac{T}{\tau} \right\rceil - 1\right) m_2 + 2 \times (T-\tau) \times E_{\chi^2(\tau-2)}\left(\frac{\alpha_{maxN}(\sigma^2 \times \chi^2(\tau-2), m_2)}{2}\right)\right] + \gamma_T \tau$$

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

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

Bounds for the stationary case

With length of window τ as t

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

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
  - $b(t) = 5.5 \quad \forall t \leq 40$
  - $b(t) = 4.5 \quad \forall t \geq 40 \quad and \quad t \leq 90$
  - $b(t) = 9.0 \quad \forall t \geq 90$
- $N = 800$,$\epsilon(t) = N(0, 10^2)$
- Length of sliding window $\tau = 20$.
- Expected Reward

μ_i	$t \leq 40$	$40 \leq t \leq 90$	$t \geq 90$
μ_1	1018.18	888.88	1244.44
μ_2	1090.9090	800	1600
μ_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 τ is bounded as follows :

$$E[\tilde{N_T}(i)] \leq \tau - 1 + \gamma_T \tau$$
$$+ min_{m \in \{1,..,\tau-1\}} \Big[\big(\lceil \frac{T}{\tau} \rceil - 1 \big) m$$
$$+ 2 \times (T - \tau) \times E_{\chi^2(\tau-2)} \Big(\frac{\alpha_{maxN}(\sigma^2 \times \chi^2(\tau-2), m)}{2} \Big) \Big]$$

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

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,

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

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

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

for $t \leq \tau$,

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

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

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

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

where ,
$\mathbb{T}(\tau)$ is a set of indices such that $t \in \{\tau, ...T\}$ 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

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

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

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

- 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.

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

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

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

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,

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

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

- Now,

$$P\{\bar{x}_i(t) \geq \mu_t(i) + \frac{\Delta\mu_i(T)}{2}, n_i(t,\tau) > m\}$$

$$= \int_{x=0}^{\infty} P\{\{\bar{x}_i(t) \geq \mu_t(i) + \frac{\Delta\mu_i(T)}{2}, n_i(t,\tau) > m\} | \{\frac{sse_N}{\sigma^2} = x\}\} f_{\frac{sse_N}{\sigma^2}}(x)dx$$

(By conditioning on the value of $\frac{sse_N}{\sigma^2}$))

$$= \int_{x=0}^{\infty} P\{\bar{x}_i(t) \geq \mu_t(i) + \frac{\Delta\mu_i(T)}{2}, n_i(t,\tau) > m\} | \{\chi^2(\tau-2) = x\}\} f_{\chi^2(\tau-2)}(x)$$

(We know that when error term is $N(0,\sigma^2)$, $\frac{sse_N}{\sigma^2}$ in that case has the $\chi^2(\tau-2)$distribution.(See Klimov [?]))

$$\leq \int_{x=0}^{\infty} P(t(RV)_{\tau-2} \geq \frac{\Delta\mu_i(T)}{2 \times p_i \times \sqrt{\frac{\sigma^2 \times x}{\tau-2}} \times \sqrt{\frac{1}{\tau} + \frac{1}{m}}}) \times f_{\chi^2(\tau-2)}(x)dx$$

(Since we interpret$\Delta\mu_i(T)$ as the confidence interval *and* by using inequality

$$= E_{\chi^2(\tau-2)}((\frac{\alpha_{maxN}(\sigma^2 \times \chi^2(\tau-2), m)}{2})$$

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 τ and level of significance $\alpha = \frac{1}{\tau^4}$ is bounded as follows :

$$E[\tilde{N_T}(i)] \leq \tau - 1 + 2 \times \frac{T-\tau}{\tau^4} + \gamma_T \tau$$
$$+ min_{m \in \{1,..,\tau-1\}}[(T-\tau) \times P\{\frac{\Delta\mu_i(T)^2}{\sigma^2 \times B(i,\tau)^2 \times (\frac{1}{\tau} + \frac{1}{m})} \leq \chi^2(\tau-2)\}$$
$$+ (\lceil \frac{T}{\tau} \rceil - 1)m]$$

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

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,

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

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

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

for $t \leq \tau$,

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

$$\therefore \tilde{N_T}(i) = \tau - 1 + \sum_{t=\tau+1}^{T} 1_{\{I(t)=i\neq i_t^*, n_i(t,\tau)\leq m\}} + \sum_{t=\tau+1}^{T} 1_{\{I(t)=i\neq i_t^*, n_i(t,\tau)>m\}} \tag{17}$$

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

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

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

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

($\mathbb{T}(\tau)$ is a set of indices such that $t \in \{\tau+1, ...T\}$ 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 written as a subset of following events by using lemma 5.

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

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

$$PF_i(t, \tau, p_I(t, \tau)) = p_i \times t_{\frac{\alpha}{2}, min(t-2, \tau-2)} \times mse \times \sqrt{\frac{1}{min(t, \tau)} + \frac{(p_i - \bar{p_{I(s)}})^2}{\sum_{s=t-\tau+1}^{t} (p_i - \bar{p_{I(s)}})^2}} \tag{21}$$

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

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

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 $t > \tau$

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

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

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

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

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

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

As $B(i,\tau) = (2 \times p_i \times t_{\frac{\alpha}{2},min(t-2,\tau-2)} \times \sqrt{\frac{1}{\tau-2}})$

$$\{\Delta\mu_i(T) \leq 2 \times p_i \times t_{\frac{\alpha}{2},min(t-2,\tau-2)} \times \sqrt{\frac{sse_N}{\tau-2}} \times \sqrt{\frac{1}{\tau}+\frac{1}{m}}\}$$
$$\equiv \{\Delta\mu_i(T)^2 \leq B(i,\tau)^2 \times (\frac{1}{\tau}+\frac{1}{m}) \times sse_N\} \tag{25}$$

(Since everything is positive)

$$\therefore P\{\mu_t(i^*) - \mu_t(i) \leq 2c_t(\tau,i)\}$$
$$\leq P\{\frac{\Delta\mu_i(T)^2}{B(i,\tau)^2 \times (\frac{1}{\tau}+\frac{1}{m})} \leq sse_N\}$$
$$\leq P\{\frac{\Delta\mu_i(T)^2}{B(i,\tau)^2 \times (\frac{1}{\tau}+\frac{1}{m})} \leq \sigma^2 \times \chi^2(\tau-2)\} \tag{26}$$

(See Klimov [?] et al.)

- $$P\{\bar{x_i}(t) \geq \mu_t(i) + PF_i(t, \tau, p_I(t, \tau)))\} = P\{\bar{x_i}(t) \leq \mu_t(i) - PF_i(t, \tau, p_I(t, \tau))\} = \frac{1}{\tau^4} \quad (27)$$

  (from the definition of confidence interval)
  Similarly,

  $$P\{\bar{x_{i^*}}(t) \leq \mu_t(i) - PF_{i^*}(t, \tau, p_I(t, \tau))\} = P\{\mu_t(i) \geq \bar{x_{i^*}}(t) + PF_{i^*}(t, \tau, p_I(t, \tau)\} \leq \frac{1}{\tau^4} \quad (28)$$

- From equation 20 we have

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

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

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

From equation 19 we have

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

Now from equation 30 we have

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

4. From equation 17, 18, 32 we have

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

(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

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

Lemma

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

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

$$\leq (\lceil \frac{T}{\tau} \rceil - 1) m \quad (35)$$

Proof.

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

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 $t \in (j-1)\tau+1, \ldots, (j)\tau+1$, we have $I(t) = i, n_i(t,\tau) \leq m$. Let t_j be the maximum time this happens in j^{th} interval

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

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

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

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

□

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