IE613: Online Machine Learning **Jan-Apr 2018**

Lecture 11: High Probability Regret of EXP3.P Algorithm

Lecturer: M. K. Hanawal *Scribes: Richa Dhingra*

11.1 Notations

We shall stay consistent with the following notations throughout the document:

With respect to the loss setting , we had :

I_t : Arm chosen in round t randomly

K : Total number of arms available to choose from

$l_{t,i}$: Loss incurred by choosing arm i in round t

A : Set of arms available to choose from where A = $\{1, 2, \ldots, K\}$

$\tilde{R}_T$: $\sum_{t=1}^{T} l_{I_t,t}$- $\min_{k \in [K]} \sum_{t=1}^{T} l_{k,t}$

R_T : $\mathbf{E}\tilde{R}_T$

R_T^- : $\mathbf{E}[\sum_{t=1}^{T} l_{I_t,t}] - min_{k\in[K]}\mathbf{E}[\sum_{t=1}^{T} l_{k,t}]$

With respect to the gains setting, we have :

I_t : Arm chosen in round t randomly

K : Total number of arms available to choose from

$g_{t,i}$: Gain by choosing arm i in round t

A : Set of arms available to choose from where A = $\{1, 2, \ldots, K\}$

$\tilde{R}_T$: + $\max_{k \in [K]} \sum_{t=1}^{T} g_{t,k}$-$\sum_{t=1}^{T} g_{t,I_t}$

R_T : $\mathbf{E}\tilde{R}_T$

R_T^- : $\max_{k \in [K]}$ $\mathbf{E}$ $\sum_{t=1}^{T} g_{t,k} - \sum_{t=1}^{T}$ $\mathbf{E}$ g_{t,I_t}

11.2 Recall

In the last few lectures ,adversarial bandit setting was considered which was a variant of multi-armed bandit scenario where the loss incurred in every round t with arm chosen to be i, $l_{t,i}$ is non-stochastic and bounded i.e. $l_{t,i} \in [0,1]$. We introduced EXP3 algorithm where the pseudo regret was bounded by $O(\sqrt{KT \log K})$ but the actual regret i.e. $\tilde{R}_T$ cannot be bounded in high probability owing to the high variance in the estimates of the losses(usually of the order $O(\frac{1}{p_{t,i}})$) which can blow up if the algorithm is executed for large T. Therefore, we discussed a new algorithm EXP3.P which is given by algorithm 2.

Algorithm 1 Exp.3P in Loss setting

Parameters : $\eta \in \mathbb{R}^+$ and $\gamma, \beta \in [0,1]$
Let P_1 be the uniform distribution over $\{1, 2, .., K\}$
For each round t =1,2,...,T :
1. Draw an arm I_t from the probability distribution P_t.
2. Compute the estimated loss for each arm as :
$$\tilde{l}_{t,i} = \frac{l_{i,t}\mathbb{1}_{\{I_t=i\}}+\beta}{p_{t,i}}$$
3. Compute the estimated cumulative loss for each arm as :
$$\tilde{L}_{t,i} = \sum_{s=1}^{t} \tilde{l}_{s,i}$$
4. Compute the new probability distribution over the arms $P_{t+1} = (p_{t+1,1}, p_{t+1,2}, ..., p_{t+1,K})$ where :
$$p_{t+1,i} = (1-\gamma)\frac{exp(-\eta \tilde{L}_{t,i})}{\sum_{k=1}^{K} exp(-\eta \tilde{L}_{t,k})} + \frac{\gamma}{K}$$

It differs from EXP3 in the following ways

1. **Definition of the estimator :**
While in the case EXP3 , we had an unbiased estimator for losses in each round and for every arm,the estimator given by EXP3.P is positively biased and hence gives an overestimation.

2. **Definition of Probabilities**
In the case of EXP3, we just had the probabilities given by exponential weights which directly depended on the cumulative losses incurred up to that round,in the case of EXP3.P , its a convex combination of distribution defined on the basis of exponential weights as well uniform distribution on the set of arms.

3. **Variance in Estimation**
In the case of EXP3, we had a very high variance due to the involvement of the $\frac{1}{p_{t,i}}$ factor where $p_{t,i}$ can take as small value as possible while in the case of EXP3.P,we have the value of $\frac{1}{p_{t,i}}$ always bounded below by $\frac{\gamma}{K}$, hence the variance of the estimator cannot blow up arbitrarily.

11.3 Remark

In order to prove the bounds on the regret for the above algorithm in high probability, we shall consider the gains setting as its easier to develop to give a tight bound in this setting in comparison to the loss setting.

Algorithm 2 Exp.3P in Gain setting

Parameters : $\eta \in \mathbb{R}^+$ and $\gamma, \beta \in [0,1]$
Let P_1 be the uniform distribution over $\{1, 2, .., K\}$
For each round t =1,2,...,T :
1. Draw an arm I_t from the probability distribution P_t.
2. Compute the estimated loss for each arm as :
$$\tilde{g}_{t,i} = \frac{g_{t,i}\mathbb{1}_{\{I_t=i\}}+\beta}{p_{t,i}}$$
3. Compute the estimated cumulative loss for each arm as :
$$\tilde{G}_{t,i} = \sum_{s=1}^{t} \tilde{l}_{s,i}$$
4. Compute the new probability distribution over the arms $P_{t+1} = (p_{t+1,1}, p_{t+1,2}, ..., p_{t+1,K})$ where :
$$p_{t+1,i} = (1-\gamma)\frac{exp(\eta\tilde{G}_{t,i})}{\sum_{k=1}^{K} exp(\eta\tilde{G}_{t,k})} + \frac{\gamma}{K}$$

11.4 Introduction

11.4.1 Exploration Vs Exploitation

In this section, we shall explore the concept of exploration and exploitation which mainly differentiate the fundamental logics of algorithms EXP3 and $EXP3.P$. In both of the algorithms,the P_t ,which is the probability distribution of choosing arms in round t is convex combination of Exploration distribution and Exploitation distribution. the two algorithms vary on the basis of the scalars used to compute these convex combinations.

Exploitation :
This factor defines the weights that should be assigned to each of the arms or actions based on the cumulative losses that have been incurred up to the round t for that particular action and get updated with every round. So, an action that accumulates a large loss in the beginning gets a very low weight and hence is chosen a very few number of times until it recovers in terms of its cumulative loss and hence comes back in to picture.

Exploration:
This factor contributes to the weights assigned to all the arms in every round and is independent of the losses accumulated by each arm in any given round and just depends on the fixed original distribution that we begin with in order to choose an arm in any given round. Hence, even if the cumulative loss of any arm is low would not let the probability factor of choosing that arm to diminish as this quantity will always be bounded below by the exploration factor.

Remark:
We observe that in the case of :

EXP3:
This algorithm is completely based on the exploitation factor and does not involve any exploration factor so we can see it as a special case of EXP3.P with γ to be 0 and β also to be zero.
EXP3.P:
This algorithm has an explicit addition of an exploration factor to the exponentially weighted EXP3 algorithm and so we have $\gamma \in (0,1)$.In such a case,the scenario of latching onto one specific arm with gives lesser

losses and ignoring the other arms just on the basis of cumulative loss gets avoided as the exploration factor always brings that arm in to picture with certain assured lower bound on the probability of choosing that arm. Hence,an action that performs poorly in the beginning and having a potential of improvising later does not get ruled out outrightly on the basis of its initial performance and vice versa.

11.4.2 Anytime Algorithms

Anytime Algorithms are algorithms that do not demand the requirement to know the number of rounds for which the algorithm will be executed apriori.So, the parameters that are used to define these algorithms are independent of T. Hence, the bounds on the regret for such algorithms are much more practical and desirable in comparison to the other algorithms which have parameters depending on T ans it is not always possible to know the number of rounds for which the algorithm will be executed.

11.5 Prerequisites

11.5.1 Gains Setting

Before laying down the foundation of the proof for the bounds to regret for exp3p , lets acquaint ourselves with the analogous gains settings that is used in online learning and how the regrets are defined accordingly.

Gain
We say that in round t, the learner had a gain of $g_{t,i}$ in round t by using arm i if he manages to give a correct prediction in terms of the label to the corresponding data instance and hence the loss $l_{t,i}$ is zero. So ,in general ,we have :
$g_{t,i}$=1-$l_{t,i}$

Regret The regret in setting can be seen as the difference in the gains one could earn if a learner used a single arm that maximized the total gains through out the T rounds in comparison to the actual gains earned through the algorithm.
Mathematically :
$R_T = \max_{k \text{ in } [K]} \sum_{t=1}^{T} g_{t,k} - \sum_{t=1}^{T} g_{t,I_t}$

11.5.2 Important results

In this section , we shall write down the following set of inequalities and lemma which shall be used in writing the proof of the theorem stated below.
We shall use the following inequalities extensively through out the proofs on the bound with high probability on the regret of EXP3P.
1. **Log Inequality**
$\log(x) \leq x - 1$ where $x \leq 1$

2. **Exponential Inequality**
$exp(-x) + x - 1 \leq \frac{x^2}{2}$ where $x \leq 1$

3. **Lemma**

Lemma 11.1. *For $\beta \leq 1$,let $\tilde{g_{t,i}} = \frac{\beta + g_{t,i}\mathbb{1}_{\{I_t=i\}}}{p_{t,i}}$ and a given arm i in [K], we have :*
$\sum_{t=1}^{T} g_{t,i} - \sum_{t=1}^{T} \tilde{g}_{t,i} \leq \frac{\log \frac{1}{\delta}}{\beta}$, *with probability at least $1-\delta$,for all $\delta \in (0,1)$.*

Lemma 11.2. *For $\beta \leq 1$,let $\tilde{g_{t,i}} = \frac{\beta + g_{t,i}\mathbb{1}_{\{I_t=i\}}}{p_{t,i}}$, we have :*
$\sum_{t=1}^{T} g_{t,i} - \sum_{t=1}^{T} \tilde{g_{t,i}} \leq \frac{\log \frac{K}{\delta}}{\beta}$, *with probability at least $1-\delta$,for all $\delta \in (0,1)$ and for all $i \in$ [K].*

11.6 Developing a bound on the regret in High Probability for EXP3.P algorithm

In this section, we shall prove the following theorem which tries to bound the regret associated with EXP3.P algorithm in high probability.

Theorem 11.3. *EXP3.P Theorem*
1. For any δ in$(0,1)$, if we execute EXP3.P algorithm with the following parametric values , $\beta = \sqrt{\frac{\log \frac{K}{\delta}}{TK}}$,
$\eta = 0.95\sqrt{\frac{\log K}{TK}}$ and $\gamma = 1.05\sqrt{\frac{K \log K}{T}}$, then with probability at least 1-δ, $\tilde{R_T} \leq 5.15\sqrt{TK \ln \frac{K}{\delta}}$;

2. Furthermore, if we set the values of : $\beta = \sqrt{\frac{\log K}{TK}}$,$\eta = 0.95\sqrt{\frac{\log K}{TK}}$ and $\gamma = 1.05\sqrt{\frac{K \log K}{T}}$, then with probability at least 1-δ, $\tilde{R_T} \leq 5.15\sqrt{TK \log K} + \sqrt{\frac{TK}{\log K}} \log \frac{1}{\delta}$ for any $\delta \in (0,1)$.

Remark : If we closely study the implications of the above algorithm, we see that in the first part of the theorem,the parameter β has a dependence on the value of δ which gives us our confidence level , so in order to achieve such a bound with high probability, we need to feed in the value of δ to β apriori while in the result 2 of the theorem , its just the bound that depends on the confidence level while the other parameters are completely independent of this value. Also the above the results hold only if we know the value of T apriori ,hence its not an anytime algorithm bound.

Proof : We shall use the following claim to prove the above result:
Claim 1 :
If $\gamma \leq 1/2$ and $(1+\beta)\eta KT \leq \gamma$,then :
$\mathrm{R_T} \leq \beta TK + \gamma T + (1+\beta)\eta KT + \frac{\log(K/\delta)}{\beta} + \frac{\log K}{\eta}$
Proof for the claim :
STEP 0 : Positively biased Estimator of gain

$$\begin{aligned}\mathbf{E}_{i\ P_t}\tilde{g}_{t,i} &= \sum_{i=1}^{k}\tilde{g}_{t,i}p_{t,i}\\ &= \sum_{i=1}^{k}\frac{g_{t,i}\mathbb{1}_{\{I_t=i\}}+\beta}{p_{t,i}}p_{t,i}\\ &= g_{t,I_t}+\beta K\end{aligned}$$

STEP 1 :Defining the regret in terms of Expectation

$$\begin{aligned}R &= \sum_{t=1}^{T}g_{t,k}-\sum_{t=1}^{T}g_{t,I_t}\\ &= \sum_{t=1}^{T}g_{t,k}-\mathbf{E}_{i\ P_t}[\tilde{g}_{t,i}-\beta K]\\ &= \sum_{t=1}^{T}g_{t,k}-\mathbf{E}_{i\ P_t}[\tilde{g}_{t,i}]+\beta KT\end{aligned}$$

Let us define :
u=(1/K,1/K,...../1/K) to be uniform distribution on A and
$w_t = \frac{p_t-\gamma u}{1-\gamma}$
Then , we get that that by adding and subtracting the cumulant generating function of $\tilde{g}_{t,i}$:

$$\begin{aligned}-\mathbf{E}_{i\ P_t}[\tilde{g}_{t,i}] &= -(1-\gamma)\mathbf{E}_{i\ w_t}[\tilde{g}_{t,i}]-\gamma\mathbf{E}_{i\ u}[\tilde{g}_{t,i}]\\ &= (1-\gamma)(\log\mathbf{E}_{i\ w_t}exp(\eta(\tilde{g}_{t,i}-\mathbf{E}_{k\ w_t}[\tilde{g}_{t,k}]))-\log\mathbf{E}_{i\ w_t}exp(\eta(\tilde{g}_{t,i})/\eta-\gamma\mathbf{E}_{i\ u}[\tilde{g}_{t,i}]\\ &= \sum_{t=1}^{T}g_{t,k}-\mathbf{E}_{i\ P_t}[\tilde{g}_{t,i}]+\beta KT\end{aligned}$$

STEP 2 :Bounding the expectation

$$\begin{aligned}\log\mathbf{E}_{i\ w_t}exp(\eta(\tilde{g}_{t,i}-\mathbf{E}_{k\ w_t}[\tilde{g}_{t,k}])) &= \log\mathbf{E}_{i\ w_t}exp(\eta(\tilde{g}_{t,i})exp(-\eta\mathbf{E}_{k\ w_t}[\tilde{g}_{t,k}]))\\ &= -\eta\mathbf{E}_{k\ w_t}[\tilde{g}_{t,k}]+\log\mathbf{E}_{i\ w_t}exp(\eta(\tilde{g}_{t,i})\\ &\le -\eta\mathbf{E}_{k\ w_t}[\tilde{g}_{t,k}]+\mathbf{E}_{i\ w_t}exp(\eta(\tilde{g}_{t,i})-1 \text{ (using the log inequality)}\\ &= \mathbf{E}_{k\ w_t}[\eta\tilde{g}_{t,k}+exp(\eta(\tilde{g}_{t,k})-1] \text{ (due to linearity property of expectation)}\\ &\le -\eta\mathbf{E}_{k\ w_t}[\eta^2\tilde{g}_{t,k}^2]/2 \text{ (using the exponential inequality)}\\ &\le -\eta\mathbf{E}_{k\ w_t}[\eta^2\tilde{g}_{t,k}^2](1)\end{aligned}$$

Now look at :
$w_{t,i} = (p_{t,i}-\gamma/k)/(1-\gamma)$

$\leq p_{t,i}/(1-\gamma)$
Hence : $w_{t,i}/(p_{t,i} \leq 1/(1-\gamma)$ (2)

From (1) and (2) :
$\log \mathbf{E}_{i\ w_t} exp(\eta(\tilde{g}_{t,i} - \mathbf{E}_{k\ w_t}[\tilde{g}_{t,k}])) \leq \eta^2 \sum_{k=1}^{K} \tilde{g}_{t,k}^2 w_{t,k}$

$$= \eta^2 \sum_{k=1}^{K} \left(\frac{g_{t,i}\mathbb{1}_{\{I_t=i\}} + \beta}{p_{t,k}}\right)^2 w_{t,k}$$

$$\leq \frac{\eta^2(1+\beta)}{(1-\gamma)}\left(\sum_{k=1}^{K} \tilde{g}_{t,k}\right)$$

STEP 3 :Substitution in one of the terms of the regret expression
$-\sum_{t=1}^{T}[\mathbf{E}_{i\ P_t}[\tilde{g}_{t,i}]] =$
$\sum_{t=1}^{T}[(1-\gamma)(\log \mathbf{E}_{i\ w_t} exp(\eta(\tilde{g}_{t,i} - \mathbf{E}_{k\ w_t}[\tilde{g}_{t,k}])) - \log \mathbf{E}_{i\ w_t} exp(\eta(\tilde{g}_{t,i})/\eta - \gamma \mathbf{E}_{i\ u}[\tilde{g}_{t,i}]]$

$$\leq \sum_{t=1}^{T}[(1-\gamma)(\log \mathbf{E}_{i\ w_t} exp(\eta(\tilde{g}_{t,i} - \mathbf{E}_{k\ w_t}[\tilde{g}_{t,k}])) - \log \mathbf{E}_{i\ w_t} exp(\eta(\tilde{g}_{t,i}))]/\eta$$

$$\leq \frac{(1+\beta)(1-\gamma)\eta^2}{\eta(1-\gamma)} \sum_{t=1}^{T}\sum_{k=1}^{K} \tilde{g}_{t,k} - \frac{(1-\gamma)}{\eta} \sum_{t=1}^{T} \log \mathbf{E}_{i\ w_t} exp(\eta(\tilde{g}_{t,i}))$$

(from step 2 bound)

$$\leq (1+\beta)\eta \sum_{t=1}^{T}\sum_{k=1}^{K} \tilde{g}_{t,k} - \frac{(1-\gamma)}{\eta} \log \prod_{t=1}^{T} \mathbf{E}_{i\ w_t} exp(\eta(\tilde{g}_{t,i}))$$

$$\leq (1+\beta)\eta \sum_{t=1}^{T}\sum_{k=1}^{K} \tilde{g}_{t,k} - \frac{(1-\gamma)}{\eta} \log \prod_{t=1}^{T}\sum_{k=1}^{K} \frac{exp(\eta \tilde{G}_{t,i})}{\sum_{k=1}^{K} exp(\eta \tilde{G}_{t-1,k})}$$

$$\leq (1+\beta)\eta \sum_{k=1}^{K}\sum_{t=1}^{T} \tilde{g}_{t,k} - \frac{(1-\gamma)}{\eta} \log \sum_{k=1}^{K} \exp(\eta \tilde{G}_{T,k}) + \frac{\log K}{\eta}$$

(using the telescopic sum of the sequence)

$$\leq (1+\beta)\eta K \max_{k in [K]} \tilde{G}_{t,k} - \frac{(1-\gamma)}{\eta} \log \sum_{k=1}^{K} \exp(\eta \tilde{G}_{T,k}) + \frac{\log K}{\eta}$$

$$\leq -[-(1+\beta)\eta K + (1-\gamma)] \max_{k in [K]} \tilde{G}_{t,k} + \frac{\log K}{\eta}$$

(since we have positive sums being subtracted)

$$\leq -[-(1+\beta)\eta K + (1-\gamma)] \max_{k in [K]} \sum_{t=1}^{T} \tilde{g}_{t,k} + \frac{\log K}{\eta}$$

$$\leq -[-(1+\beta)\eta K + (1-\gamma)] \max_{k in [K]} \sum_{t=1}^{T} g_{t,k} + \frac{\log K}{\eta} + [-(1+\beta)\eta K + (1-\gamma)] \frac{\log K\delta^{-1}}{\beta}$$

(using lemma 2)

$$\leq -[-(1+\beta)\eta K + (1-\gamma)] \max_{k in [K]} \sum_{t=1}^{T} g_{t,k} + \frac{\log K}{\eta} + \frac{\log K\delta^{-1}}{\beta}$$

(using the fact that $[-(1+\beta)\eta K + (1-\gamma)] < 1$ by the assumption in the claim)

Step 4 : Substituting in the regret expression

We get that : $\mathrm{R} = \sum_{t=1}^{T} g_{t,k} - \mathbf{E}_{i\ P_t}[\tilde{g}_{t,i}] + \beta KT$

$$\leq -[-(1+\beta)\eta K + (1-\gamma)] \max_{k in [K]} \sum_{t=1}^{T} g_{t,k} + \frac{\log K}{\eta} + \frac{\log K\delta^{-1}}{\beta} + \beta KT$$

$$\leq -[-(1+\beta)\eta K + (1-\gamma)]T + \frac{\log K}{\eta} + \frac{\log K\delta^{-1}}{\beta} + \beta KT$$

$$\leq (1+\beta)\eta KT + \gamma T + \frac{\log K}{\eta} + \frac{\log K\delta^{-1}}{\beta} + \beta KT$$

with probability atleast 1 - δ from the claim, we see that $\mathrm{R_T}$

$$\leq (1+\beta)\eta KT + \gamma T + \frac{\log K}{\eta} + \frac{\log K\delta^{-1}}{\beta} + \beta KT$$

with probability atleast 1 - δ (since the above result holds for any general arm chosen, so will hold in particular for the reward maximizing arm) with probability 1-δ

Hence this proves our claim.
Now we shall use this claim to prove the required result.
By plugging in the values of the parameters as per the statement of the theorem, we see that :
With the following values : $\beta = \sqrt{\frac{\log \frac{K}{\delta}}{TK}}$,
$\eta = 0.95\sqrt{\frac{\log K}{TK}}$ and $\gamma = 1.05\sqrt{\frac{K \log K}{T}}$ $\mathrm{R_T}$

$$\sqrt{\frac{\log \frac{K}{\delta}}{TK}} TK + 1.05\sqrt{\frac{K \log K}{T}} T + (1+\sqrt{\frac{\log \frac{K}{\delta}}{TK}}) 0.95\sqrt{\frac{\log K}{TK}} TK + \log(K\delta^{-1}) / \sqrt{\frac{\log \frac{K}{\delta}}{TK}} + \log(K) / 0.95\sqrt{\frac{\log K}{TK}}$$

$$\leq 5.05\sqrt{\log k\delta^{-1}TK} + 0.95\sqrt{\log K \log K\delta^{-1}}$$

Case 1: If we consider the case when T $\geq 5.15\sqrt{TK\ln\frac{K}{\delta}}$, then using this condition, we can bound the last term in the above expression of R_T by $\frac{0.45}{5.15}\sqrt{\log K \log K\delta^{-1}}$ and hence achieve the bound of $5.15\sqrt{TK\ln\frac{K}{\delta}}$

Case 2: If we have when T $\leq 5.15\sqrt{TK\ln\frac{K}{\delta}}$, then we clearly see that the $\gamma \leq 0.21$ (by using this bound on the value of the parameter) and $\beta \leq 0.1$ so by using the condition that $(1+\beta)\eta K \leq \gamma \leq 0.5$, we can obtain the above bound by simply substituting the upper bounds of the respective parameters i.e., we can show that this forms an upper bound on the regret with probability at least $1-\delta$. Similarly, in order to prove part 2 , of the theorem, we follow exactly the same line of proof upto the bound we got on the regret in terms of the parameter.
Here, now we have :
$\beta = \sqrt{\frac{\log K}{TK}}$,$\eta = 0.95\sqrt{\frac{\log K}{TK}}$ and $\gamma = 1.05\sqrt{\frac{K\log K}{T}}$, so we substitute these values in the bound claimed through the claim 1:
We get that :

$$\sqrt{TK\log K} + 0.95\sqrt{\frac{\log KT}{K}} + 0.95\sqrt{\log KTK} + 0.95TK\log K + \frac{\log K\delta^{-1}}{\sqrt{TK\log K}} + \sqrt{\frac{TK}{\log K}}/0.95$$

$$\leq 3.952\sqrt{TK\log K} + 0.95TK\log K + \frac{\log(k\delta^{-1})}{\sqrt{TK\log(K)}}$$

$$\leq 5.15\sqrt{TK\log K} + \sqrt{\frac{TK}{\log K}}\log\frac{1}{\delta}$$

for any $\delta \in$ (0,1)

Theorem 11.4. *EXP3.P Theorem for expected regret*
For any δ in$(0,1)$, if we execute EXP3.P algorithm with the following parametric values , $\beta = \sqrt{\frac{\log\frac{K}{\delta}}{TK}}$,
$\eta = 0.95\sqrt{\frac{\log K}{TK}}$ *and* $\gamma = 1.05\sqrt{\frac{K\log K}{T}}$, *then* $\mathbf{E}R_T \leq 5.15\sqrt{TKlogK} + \frac{TK}{logK}$

Proof :
We can use the above theorem and the following result to prove this : $\mathbf{E}R_T \leq \int_0^1 \delta^{-1}\mathbb{P}(W > \log 1/\delta)d\delta$ where W is any random variable which is real valued.
If we look at the RHS of the above statement , then :
From the result 2 of the theorem, we see that : $\tilde{R_T} \leq 5.15\sqrt{TK\log K} + \sqrt{\frac{TK}{\log K}}\log\frac{1}{\delta}$ for any

$\delta \in (0,1)$
We define W = $\sqrt{\frac{\log K}{TK}}(R_T - 5.15\sqrt{nK\log K})$
So, we get that $\mathbf{P}(W > \log 1/\delta)$ is atmost δ
Now, $\mathbf{E}W \leq \int_0^1 \delta * \frac{1}{\delta} d\delta$
$\leq \int_0^1 1 d\delta$
≤ 1
Hence we get that $\mathbf{E}W \leq 1$
Hence we have $\mathbf{E}\sqrt{\frac{\log K}{TK}}(R_T - 5.15\sqrt{nK\log K}) \leq 1$
Hence , we get :

$\mathbf{E}R_T \leq 5.15\sqrt{TK\log K} + \frac{TK}{\log K}$
Hence Proved

11.6.1 Conclusion

We saw a new variant of the existing EXP3 algorithm which involved and explicit exploration factor and through that, we could control the variance and also develop a bound on the regret in high probability. We note that in terms of expected regret, both EXP3 and EXP3P are of the same order O($TK\log K$) . Later , in the coming lecture, we shall look at an implicit way of incorporating an exploration factor in the existing EXP3 algorithm and study the bounds on its regrets in high probability.