Statistical Techniques in Robotics 16-831 - The Robotics Institute

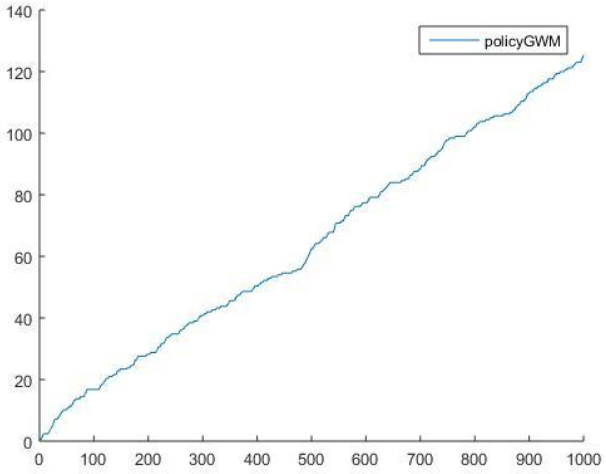
Project 2: MultiArmed Bandits

Gauri Gandhi

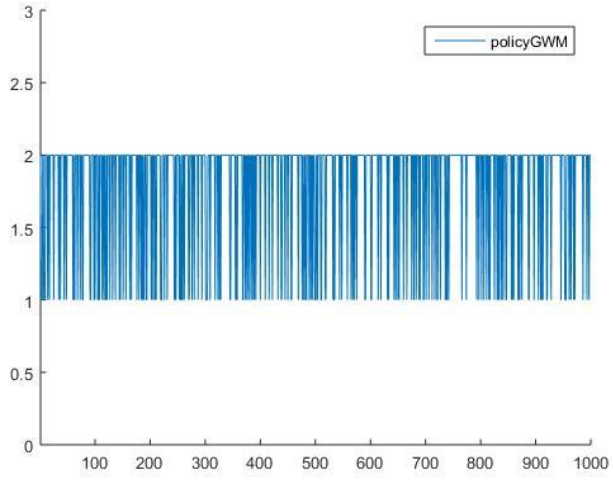
Andrew ID: gaurig

3.1.1 GWM for Bandit Setting

Regret:



Actions Chosen:



Explanation:

GWM does not work properly for the bandit setting as the losses have not been scaled with probabilities at each time step to result in unbiased estimates. Hence, we get biased estimates of true losses in partial feedback situations that results in unbounded regrets.

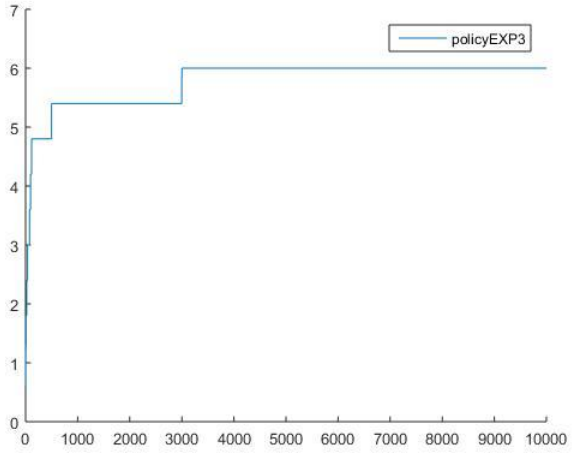
3.1.2 Regret Bound for GWM in bandit setting Theory

3.1.3 Theory

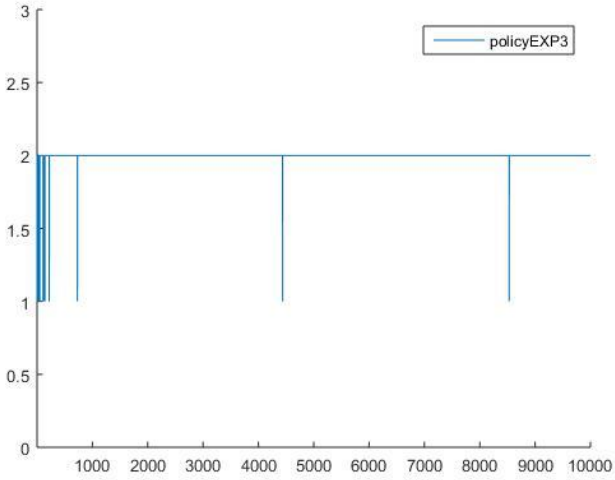
3.2.1 EXP3 on Constant Game

I plotted the constant graphs for 10000 time steps to get more intuition of the performance.

Regret:



Actions chosen:



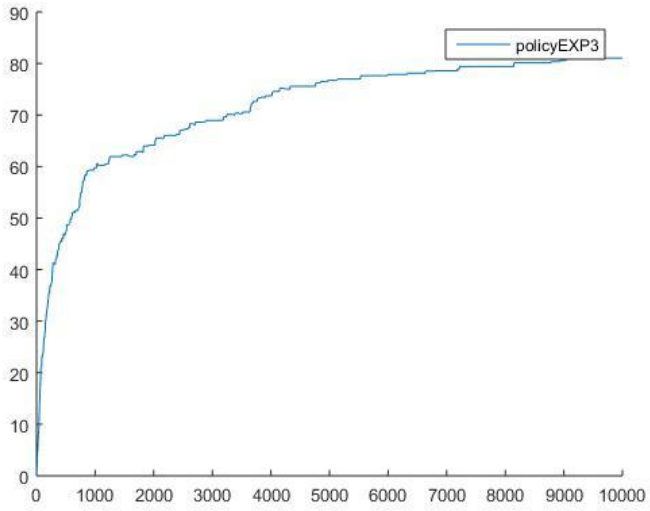
Explanation:

We see, the regret for EXP3 converges to a constant value for the constant game. Since constant functions are sublinear, average regret would converge to 0. This is because of the unbiased estimator of the loss taken in this case. The action plot shows that initially exploration happens and once the algorithm learns about the best action, exploitation happens.

3.3 Gaussian Game

3.3.2 EXP3 on Gaussian Game

Regret:



3.4 Variance Issues Theory

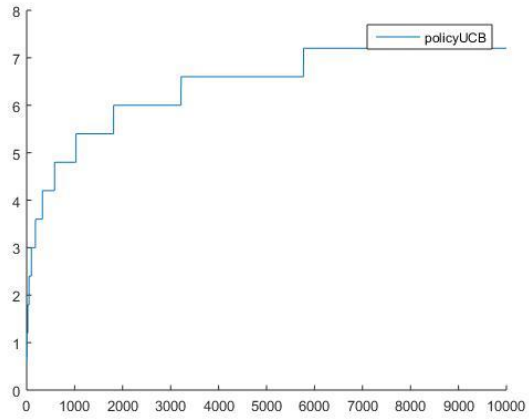
4.2 Upper Confidence Bound

4.2.1 Theory

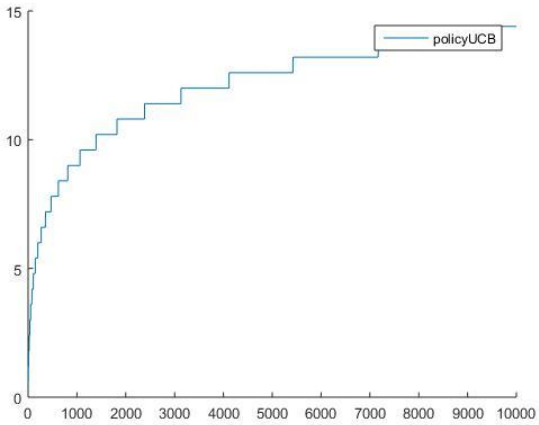
4.2.2 Theory

4.3.1 UCB on Constant Game

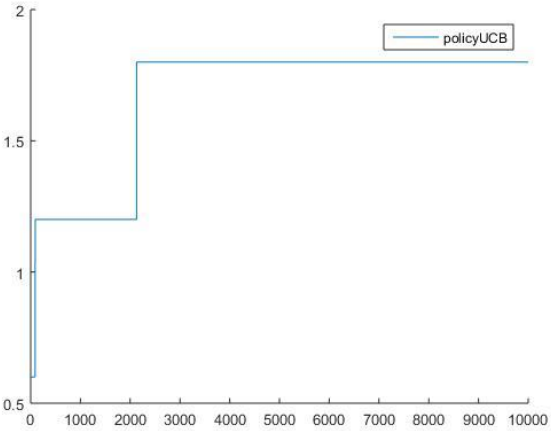
Regret: Alpha = 1



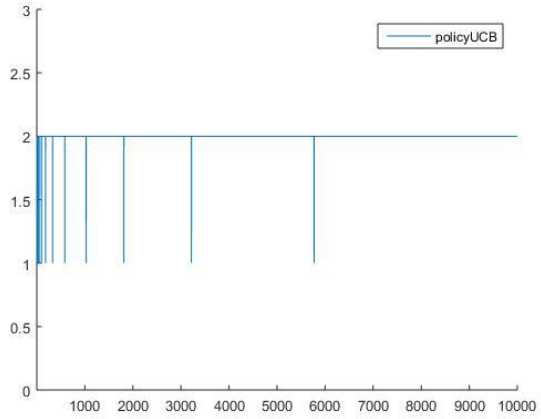
Alpha = 2



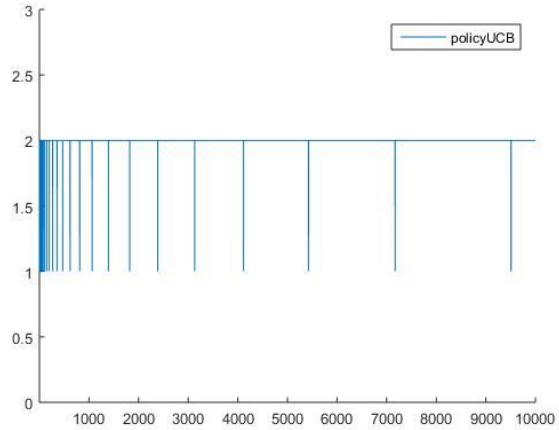
Alpha = 0.2



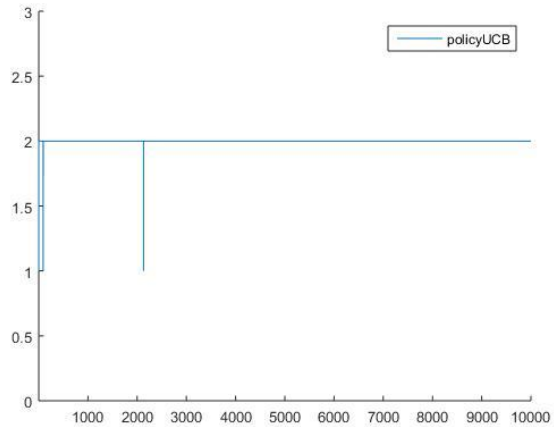
Actions: Alpha = 1



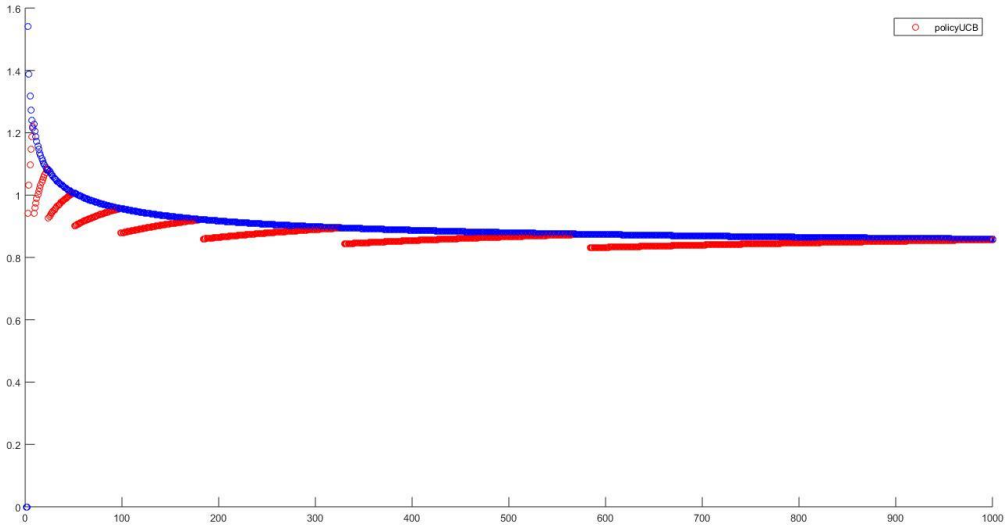
Alpha = 2



Alpha = 0.2



Bound for Actions:



Blue: Action2(Reward 0.8) Red: Action1(Reward 0.2)

Explanation:

As UCB can easily converge to the true mean for the constant game, the regret converges to a constant with t. Hence, the average converges to 0. Also we can see the exploration exploitation trade-offs with changing alpha. At high alpha, there is more exploration while at low alpha, there is more exploitation.

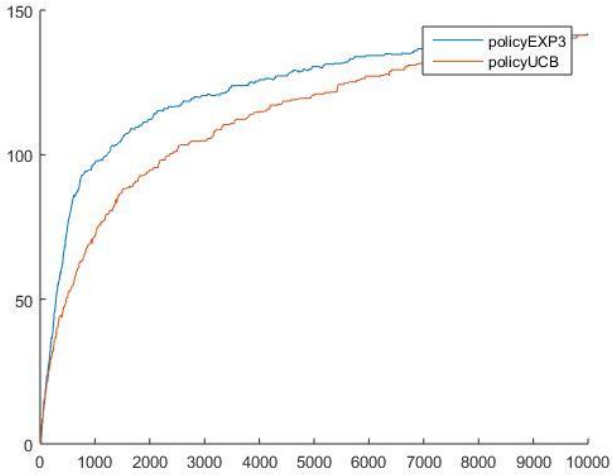
Confidence bound explanation:

$$a^t \leftarrow \underset{i=1,\ldots,N}{\arg\max} \frac{S_i}{C_i} + \sqrt{\alpha \frac{\log t}{2C_i}};$$

The first term of the bound is more significant when t is small. Hence action2 with higher S and C will always have a higher confidence bound. While, when t increases to a very large value, second term gets significant and the action1 with small C starts getting higher confidence bound.

4.4.1 UCB and EXP3 on Gaussian Game

Regret:

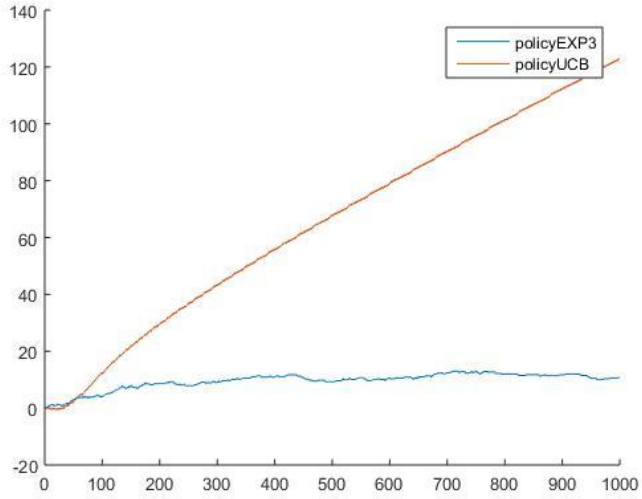


Explanation:

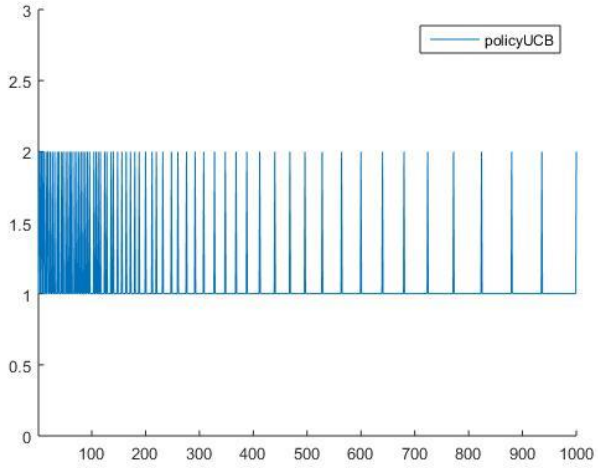
UCB performs better than EXP3 in Gaussian(stochastic) as it tries to converge to the true mean of the distributions and efficiently manages between exploration and exploitation. EXP3 does not adopt such a policy.

4.5.1 UCB and EXP3 on Adversarial Game

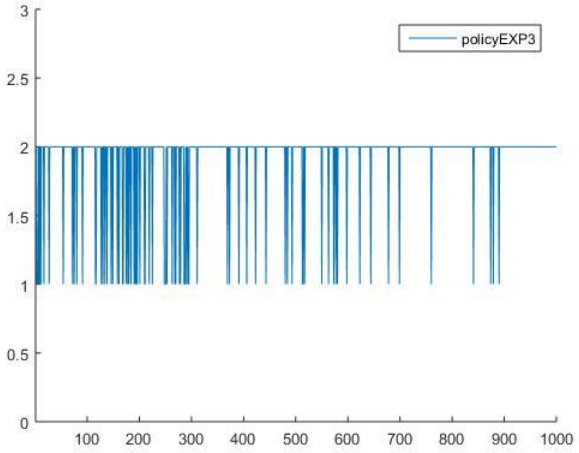
Regret:



Actions UCB:



Actions EXP3:

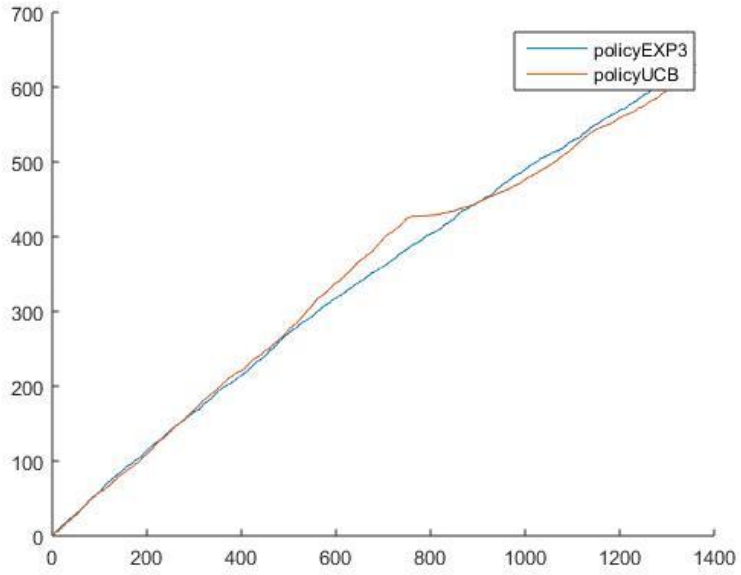


Explanation:

Here, EXP3 performs better than UCB as UCB tries to estimate the mean of the underlying distribution assuming there exists some stochastic model for the actions. Since there is no model and there is a deterministic sequence for each action, its regret is unbounded.

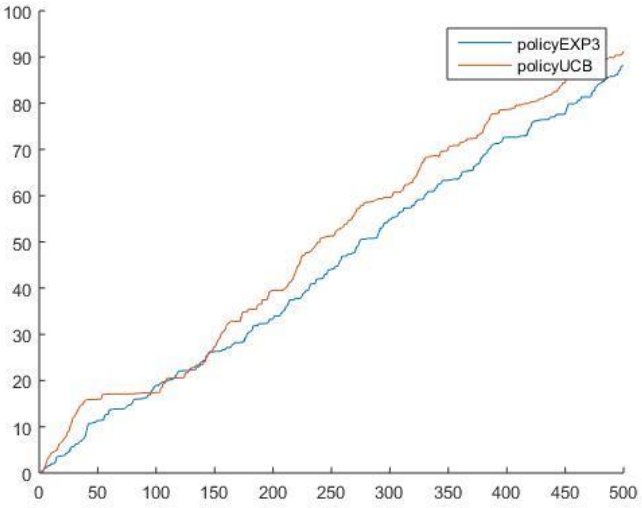
5.2.1 EXP3 and UCB on University Latency Dataset:

Regret:



5.3.1 EXP3 and UCB on Planner Dataset:

Regret:



3.1.2

Doing regret analysis

Let the potential function be

$$\phi^{(t)} = \sum_{n=1}^{N} w_n^{(t)}$$

N = total actions.
w = associated weights

$$\therefore \phi^{(t+1)} = \sum_{n=1}^{N} w_n^{(t+1)}$$

$$= \sum_{n=1}^{N} w_n^{(t)} e^{-\eta^{(t)} l_n^{(t)}}$$

$$= \phi^{(t)} \sum_{n=1}^{N} p_n^{(t)} e^{-\eta^{(t)} l_n^{(t)}} \quad \left[\because p_n^{(t)} = \frac{w_n^{(t)}}{\sum_{n=1}^{N} w_n^{(t)}} = \frac{w_n^{(t)}}{\phi^{(t)}}\right]$$

$$= \phi^{(t)} \sum_{n=1}^{N} p_n^{(t)} b^{l_n^{(t)}} \quad \text{where } b = e^{-\eta^{(t)}}$$

Upper bound potential :

$$\phi^{(t+1)} \leq \phi^{(t)} \sum_{n=1}^{N} p_n^{(t)} \left(1 - (1-b) l_n^{(t)}\right)$$

$$\left[\because b^x \leq 1 - (1-b)x \quad b \in (0,1) \quad x \in [0,1]\right]$$

taking log on both sides

$$\ln \phi^{(t+1)} \leq \ln \phi^{(t)} + \ln\left(\sum_{n=1}^{N} p_n^{(t)} \left(1 - (1-b) l_n^{(t)}\right)\right)$$

$$\leq \ln \phi^{(t)} + \ln\left(\underbrace{\sum_{n=1}^{N} p_n^{(t)}}_{1} - \sum_{n=1}^{N} p_n^{(t)} (1-b) l_n^{(t)}\right)$$

$$\leq \ln \phi^{(t)} + \ln\left(1 - \sum_{n=1}^{N} p_n^{(t)} (1-b) l_n^{(t)}\right)$$

$$\leq \ln \phi^{(t)} - \sum_{n=1}^{N} p_n^{(t)} (1-b) l_n^{(t)}$$

$$\left[\because \ln(1+x) \leq x \quad x = -\sum_{n=1}^{N} p_n^{(t)} (1-b) l_n^{(t)}\right]$$

$$\ln \phi^{(t+1)} \leq \ln \phi^{(t)} - \sum_{n=1}^{N} p_n^{(t)} (1-b) l_n^{(t)}$$

$$\sum_{t=1}^{T} \ln \phi^{(t+1)} \leq \sum_{t=1}^{T} \ln \phi^{(t)} - \sum_{t=1}^{T} \sum_{n=1}^{N} p_n^{(t)} (1-b) l_n^{(t)}$$

$$\cancel{\sum_{t=2}^{T} \ln \phi^{(t)}} + \ln \phi^{(T+1)} \leq \ln \phi^{(1)} + \cancel{\sum_{t=2}^{T} \ln \phi^{(t)}} - \sum_{t=1}^{T} \sum_{n=1}^{N} p_n^{(t)} (1-b) l_n^{(t)}$$

$$\ln \phi^{(T+1)} \leq \ln N - \sum_{t=1}^{T} \sum_{n=1}^{N} p_n^{(t)} (1-b) l_n^{(t)} \quad \left[\because \phi^{(1)} = \sum_{n=1}^{N} w_n^{(1)} = N \right]$$

$$(\because \textstyle\sum_i w_i^{(1)} = 1 \; \forall n)$$

$$\ln \phi^{(T+1)} \leq \ln N - (1-b) E(L)$$

$$\left(\because E(L) = p_n^{(t)} l_n^{(t)} \right)$$

$$(1-b) E(L) \leq \ln N - \ln \phi^{(T+1)}$$

Lower Bound →

$$E(L) \leq \frac{\ln N}{1-b} - \frac{\ln w_i^{(T+1)}}{1-b} \quad \left[\because \phi^{(T+1)} \geq w_i^{(T+1)} \; \forall i \right]$$

$$= \frac{\ln(N)}{1-b} - \frac{\ln b^{(L_h^{(T)})}}{1-b}$$

$$= \frac{\ln N}{1-b} - \frac{L_h^{(T)} \ln b}{1-b}$$

$$\left[\because w_i^{(T+1)} = \prod^{T} e^{-\eta^{(t)}} \underbrace{w_i^{(1)}}_{1} \right.$$

$$= e^{-\eta \left(\sum_{t=1}^{T} l_i^{(t)} \right)} L_i^{(T)}$$

$$\left. = b^{L_i} \right]$$

Let $\varepsilon = 1 - b \in (0, 0.5)$

$$\therefore E(L) \leq \frac{\ln N}{\varepsilon} - L_h^{(T)} \frac{\ln(1-\varepsilon)}{\varepsilon}$$

$$E(L) \leq \frac{\ln(N)}{\varepsilon} + L_n^{(T)} \frac{\varepsilon^2+\varepsilon}{\varepsilon}$$

$$\left[\because -\ln(1-\varepsilon) \leq \varepsilon^2+\varepsilon \text{ for } \varepsilon \in (0, \tfrac{1}{2})\right]$$

$$\therefore \text{Regret } E(L) \leq \frac{\ln(N)}{\varepsilon} + L_n^{(T)}(\varepsilon+1)$$

$$\therefore \text{Regret } E(L) - L_n^{(T)} \leq \frac{\ln N}{\varepsilon} + L_n^{(T)} \varepsilon$$

$$R \leq \frac{\ln N}{\varepsilon} + L_n^{(T)} \varepsilon$$

$$\boxed{R \leq \frac{\ln N}{1-e^{-\eta(T)}} + L_n^{(T)}\left(1-e^{-\eta(T)}\right)}$$

3.13 $$\sum_{t=1}^{T} \tilde{l}_n^t = \frac{\sum_{t=1}^{T} l_n^t}{p_n^t} \mathbb{1}_{a^t=n}$$ where l_n^t = true loss

In GLOM, the action is chosen over each time step by sampling (multinomial) from a prob distribution p_n^t based on the true weights.

∴ If the sampled value = s_n^t (true)

~~with p_n^t, estimate = $s_n^t \times p_n^t$ (biased)~~

~~Dividing this by p_n^t~~

~~The estimate of sampled value~~

When the complete loss vector is not known (only the value for the sampled one is known),

~~$s_n^t \times p_n^t$ is a biased estimate~~

we assign probabilities p_n^t to the sample

∴ Estimate of chosen sample = True value × prob

$= s_n^t \times p_n^t$

= biased

for an unbiased estimate

$$\text{unbiased} \leftarrow \frac{s_n^t \times p_n^t}{p_n^t}$$

∴ $\sum_{t=1}^{T} \tilde{l}_n^t = \sum_{t=1}^{T} \frac{l_n^t}{p_n^t} \mathbb{1}_{a^t=n}$ is an unbiased estimator.

3.4 Variance Issues

3.4.1 We know $Var(X) = E(X^2) - [E(X)]^2$ — (1)

here $X = \sum_{t=1}^{T} \tilde{l}_n^t$

$$E(X^2) = E\left(\left(\sum_{t=1}^{T} \tilde{l}_n^t\right)^2\right)$$

$$\geq E\left(\sum_{t=1}^{T} \tilde{l}_{n,t}^2\right)$$

$$\geq \sum_{t=1}^{T} \left(E\left(l_{n,t}^2\right)\right)$$

$$\geq \sum_{t=1}^{T} \frac{l_{n,t}^2}{P_{n,t}} \quad \text{— (2)}$$

$$[E(X)]^2 = \left[E\left(\sum_{t=1}^{T} \tilde{l}_n^t\right)\right]^2$$

$$= \left(\sum_{t=1}^{T} E\left(\tilde{l}_n^t\right)\right)^2$$

$$= \left(\sum_{t=1}^{T} l_{n,t}\right)^2 \quad \text{— (3)}$$

∴ Using (1), (2) and (3),

$$V = E(X^2) + [E(X)]^2$$

$$\geq \sum_{t=1}^{T} \frac{l_{n,t}^2}{P_{n,t}} + \left(\sum_{t=1}^{T} l_{n,t}\right)^2$$

∴ Both terms of V are of the order $O(T^2)$

∴ The lower bound of V is unbounded.

∴ V is unbounded.

4.2 Upper Confidence Bound

4.2.1 Using Hoeffding's inequality,

$$P(\mu - \hat{\mu} \geq \varepsilon) \leq e^{-2m\varepsilon^2} \quad -(1)$$

where μ = true mean

$\hat{\mu}$ = estimated mean $= \frac{1}{m}\sum_{i=1}^{m} X_i$

m = number of sampled values

(1) can be written as

$$P(\mu \geq \hat{\mu} + \varepsilon) \leq e^{-2m\varepsilon^2} \quad -(2)$$

(2) is the upper bound on probability of

$$\mu \geq \hat{\mu} + \varepsilon$$

$$\therefore 1 - P(\mu \geq \hat{\mu} + \varepsilon) \geq 1 - e^{-2m\varepsilon^2}$$

or $$P(\mu \leq \hat{\mu} + \varepsilon) \geq 1 - e^{-2m\varepsilon^2} \quad -(3)$$

let $e^{-2m\varepsilon^2} = \delta$

taking ln on both sides

$$-2m\varepsilon^2 = \ln \delta$$

$$\varepsilon^2 = \frac{-\ln \delta}{2m}$$

$$\varepsilon = \sqrt{\frac{\ln \delta^{-1}}{2m}} \quad -(4)$$

Substituting (4) in (3)

$$P\left(\mu \leq \hat{\mu} + \sqrt{\frac{\ln \delta^{-1}}{2m}}\right) \geq 1 - \delta \quad (5)$$

where $\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} X_i$

(5) shows that probability of $\mu \leq \hat{\mu} + \sqrt{\frac{\ln \delta^{-1}}{2m}}$ is lower bounded by $1-\delta$

4.2.2 from (5)

$$P\left(\mu \leq \frac{1}{m}\sum_{i=1}^{m} x_i + \sqrt{\frac{\log \delta^{-1}}{2m}}\right) \geq 1 - \delta$$

In case of a stochastic bandit problem, $n \in N$ actions,

C_n^t = number of times n is selected till t time steps

μ_n^t = true mean of distribution associated with n

$\hat{\mu}_n^t$ = sample mean of rewards for action n

∴ substituting in (5)

$$\mu_n^t \leq \hat{\mu}_n^t + \sqrt{\frac{\log \delta^{-1}}{2 C_n^t}}$$

let $\delta = \frac{1}{t}$

$$\therefore \mu_n^t \leq \hat{\mu}_n^t + \sqrt{\frac{\log t}{2 C_n^t}} \quad -(6)$$

4.2.3

Collaborators:

Richa

Keerthana

Rohan

Jimit