Advanced Topics in Machine Learning - Assignment 5

Christoffer Thrysøe - dfv107

October 11, 2017

1. Introduction of New Products

Comment: For this assignment, unfortunately I discovered very late that we could not use the time horizon T for the algorithm. Therefore I have left my old assignment in section 1.1, as per discussion with TA, where the algorithm depends on knowing the time horizon T and written a new approach for unknown T in section 1.2. Unfortunately I did not have time to finish the pseudo regret analysis of the algorithm, independent of time horizon T. If possible to be graded on the algorithm for known T then please disregard section 1.2.

1.1 Known T

We have an esablished product on the market, which we know sells with probability $\mu = 0.5$. We also have a new product, which sells with an unknown probability μ. At every sales round, we can only offer one product. We wish to propose a strategy, which maximizes the sales and analyse it's pseudo regret. For this assignment, I assume that we know the time horizon T, beforehand. I also assume that we are in bandit setting, i.e if we should try to sell the established product, where $\mu = 0.5$, we do not know if the new product would sell or not.

Since we know the time horizon T, we can use the exploration/exploitation approach, where the exploration phase determines the optimal product and the exploitation phase sells the optimal product. Since the sale probability of the established product is known, we only need to explore the new product. If we denote the established product by a_{old} and the new product a_{new}. The algorithm for employing the above strategy can be described as followed:

Algorithm 1 Maximize Sale

```
Input: Time Horizon: $T$
error = 0
for $t = 1, 2, \dots, \epsilon T$ do
    Sell $A_t = a_{new}$
    if not sale($A_t$) then
        error ± 1
sale_prob = $1 -$ (error/$\epsilon T$)
$\Delta = 0.5 -$ sale_prob
if $\Delta < 0$ then
    $a_{opt} = a_{new}$
else
    $a_{opt} = a_{old}$
for $t = \epsilon T + 1, \dots, T$ do
    Sell $A_t = a_{opt}$
```

Where we do ϵT iterations of exploration (i.e. try to sell the unknown product) and $(1 - \epsilon)T$ iterations of exploitation. The gap Δ is negative if the estimated sale probability is greater than 0.5, in which the optimal product is the new product. We wish to maximise the sale of the above algorithm, thus we want to pick ϵ in such a way that it minimises the regret of the above algorithm. Here I will use the same approach as for the lecture notes, where we define $\delta(\epsilon)$ to be the probability of picking the wrong product to be the one after ϵT rounds of exploration. Then we bind the regret as followed:

$$\bar{R}_T \leq \frac{1}{2}\Delta\epsilon T + \delta(\epsilon)\Delta(1-\epsilon)T \leq \frac{1}{2}\Delta\epsilon T + \delta(\epsilon)\Delta T = \left(\frac{1}{2}\epsilon + \delta(\epsilon)\right)\Delta T \tag{1}$$

Now we wish to express the probability $\delta(\epsilon)$, which we will do by binding it. We measure the algorithms performance of the exploration phase by $\hat{\mu}_{\epsilon T}(a)$, which is the empirical loss/reward at round ϵT. We want to pick the product, which resulted in the most optimal empirical loss/reward. Below binds the probability of picking a suboptimal arm after the exploration phase.

$$\begin{align}
\delta(\epsilon) &= P\{\hat{\mu}_{\epsilon T}(a) \geq \hat{\mu}_{\epsilon T}(a^*)\} \tag{2}\\
&\leq P\{\hat{\mu}_{\epsilon T}(a) \geq \mu(a) + \frac{1}{2}\} + P\{\hat{\mu}_{\epsilon T}(a^*) \leq \mu^* - \frac{1}{2}\Delta\} \tag{3}\\
&\leq 2e^{-2\epsilon T(\frac{1}{2}\Delta)^2} = 2e^{-\epsilon T\Delta^2/2} \tag{4}
\end{align}$$

where the last line follows from Hoeffding's inequality. Note that since we are exploring the unknown product at each sales round, we get full information as we know the sales probability of the existing product. That is why we don't divide by two in (4). Now that we have a bound for $\delta(\epsilon)$ we can use this in (1), resulting in the following:

$$\bar{R}_T \leq \left(\frac{1}{2}\epsilon + 2e^{-\epsilon T\Delta^2/2}\right)\Delta T \tag{5}$$

To minimize $\frac{1}{2}\epsilon + 2e^{-\epsilon T\Delta^2/2}$ we take the derivative and set it to zero, which leads to $\epsilon = \frac{\ln(T\Delta^2)}{T\Delta^2/2}$. Plugging this back this value of ϵ into (5) we get the following:

$$\begin{align}
\bar{R}_T &\leq \left(\frac{(\ln(T\Delta^2)}{T\Delta^2} + 2e^{-\ln(T\Delta^2)}\right)\Delta T \tag{6}\\
&= \left(\frac{(\ln(T\Delta^2)}{T\Delta^2} + \frac{2}{T\Delta^2}\Delta T\right) \tag{7}\\
&= \frac{\ln(T\Delta^2)}{\Delta} + \frac{2}{\Delta} \tag{8}
\end{align}$$

Therefore the regret of the approach can be bounded by:

$$\bar{R}_T \leq \frac{\ln(T\Delta^2)}{\Delta} + \frac{2}{\Delta} \tag{9}$$

The optimal number of explorations rounds, which minimizes the pseudo regret is given by:

$$\epsilon T = \frac{\ln(T\Delta^2)}{\Delta^2/2} \tag{10}$$

Thus the previous proposed algorithm, from this task, can be altered to use the number of exploration rounds ϵT from (10), which minimizes the pseudo regret.

1.2 Unknown T

If the time horizon T is unknown, we cannot do an exploration phase, as we will not know the optimal length of this time period. If we instead do an approach similar to "Follow the leader", we can choose the most successful product at each t and since we know the $\mu(a_{old})$ of the existing product, we only need to check if $\hat{\mu}(a_{new})$ is greater than 0.5. The issue is, however that since we are in bandit setting, if we pick the existing product, we will have no way of estimating the sale probability of the new product. For example if the unknown product has a bad start, we will only pick the existing product, leaving an uncertainty of the new product. Thus we wish to add a reward/upside to picking a product if we are uncertain about the true sales probability of this product. This is done by the `UCB1` algorithm, which takes an upper confidence bound $U_t(a)$ on $\mu(a)$:

$$U_t(a) = \hat{\mu}_{t-1}(a) + \sqrt{\frac{3 \ln t}{2N_{t-1}(a)}} \tag{11}$$

where $N_{t-1}(a)$ is the number of times the arm a has been pulled. Thus the added term increases when the arm is unexplored and decreases if the arm is explored. This added term will help us explore the unknown product, until we are cetain about it's performance. Thus the algorithm for maximizing the sale could be the following:

Algorithm 2 Maximize Sale 2

```
**Initialization:** Play the unknown action $a_{new}$ once
for $t = 2, 3, \ldots$ do
    Play $A_t = \max\left(a_{old} = 0.5, \hat{\mu}_{t-1}(a_{new}) + \sqrt{\frac{3 \ln t}{2N_{t-1}(a_{new})}}\right)$
```

To bound the pseudo regret we must take the following:

$$\bar{R}_T = \sum_a \Delta(a)\mathbb{E}[N_T(a)] \tag{12}$$

if we are given the optimality gap, when picking a_{new}:

$$\Delta = 0.5 - \mu \tag{13}$$

and when picking a_{old} we have the optimality gap:

$$\Delta = \mu - 0.5 \tag{14}$$

then the gap will be negative if the sold product is the optimal product. I believe if we can quantify the expected number of times each action will be played, we can multiply it by the optimality gap and derive a pseudo regret. I believe the expected number of times an action is played can be estimated by taking the probability of the upper confidence bound being greater than 0.5, thus being chosen by the UCB1 algorithm, and summing the probability over T, as we did in the previous assignment for the bound on the FTL approach. When we know the gap and the expected number of times each product will be sold, we can get the pseudo regret from (12).

2. Tighter analysis of the Hedge algorithm

We wish to bind the following, using Hoeffding's lemma:

$$\sum_a e^{-\eta X_t^a} p_t(a) \tag{15}$$

Hoeffding's lemma is defined as:

$$\mathbb{E}[e^{\lambda X}] \leq e^{\lambda \mathbb{E}[X] + \frac{\lambda^2 (b-a)^2}{8}} \tag{16}$$

where b is the largest value X_t can take, and a is the smallest value X_t can take. First we note that (15) is the same as the expected value of $e^{-\eta X_t^a}$ with respect to the distribution p_t, thus we can write:

$$\sum_a e^{-\eta X_t^a} p_t(a) = \mathbb{E}_{p_t}[e^{-\eta X_t^a}] \tag{17}$$

Applying hoeffding's lemma to the right hand side of (17), we get the following:

$$\mathbb{E}_{p_t}[e^{-\eta X_t^a}] \leq e^{-\eta \mathbb{E}_{p_t}[X_t^a] + \frac{\eta^2 (\max(X_t) - \min(X_t))^2}{8}} \tag{18}$$

$$= e^{-\eta \sum\limits_a X_t^a p_t(a) + \frac{\eta^2 (\max(X_t) - \min(X_t))^2}{8}} \tag{19}$$

where (19) follows from writing out the expected value.
We now continue the proof from the lecture notes, where we defined $W_t = \sum_a e^{-\eta L_t(a)}$, where $L_t(a)$ is the accumulated loss for arm a at time t.
We now have the ratio, which we must sum for each t:

$$\frac{W_T}{W_0} \leq e^{-\eta \sum_{t=1}^{T} \sum_a X_t^a p_t(a) + \frac{\eta^2}{8} \sum_{t=1}^{T} (\max(X_t) - \min(X_t))^2} \tag{20}$$

If we assume that the losses X_a^t can take values in the interval $[0, 1]$ we have the following:

$$\frac{W_T}{W_0} \leq e^{-\eta \sum_{t=1}^{T} \sum_a X_t^a p_t(a) + \frac{\eta^2}{8} \sum_{t=1}^{T} 1} \tag{21}$$

From the lecture notes we have that:

$$\frac{W_T}{W_0} = \frac{\sum\limits_a e^{-\eta L_T(a)}}{K} \geq \frac{\max_a e^{-\eta L_T(a)}}{K} = \frac{e^{-\eta \min_a L_T(a)}}{K} \tag{22}$$

Combining the inequalities in (21) and (22) and taking the logarithm on both sides, we get:

$$-\eta \min_a L_T(a) - \ \ln K \leq -\eta \sum_{t=1}^{T} \sum_a X_t^a p_t(a) + \frac{\eta^2}{8} \sum_{t=1}^{T} 1 \tag{23}$$

Rearranging (23) and dividing by η we get the following:

$$\sum_{t=1}^{T} \sum_a X_t^a p_t(a) - \min_a L_T(a) \leq \frac{\ln k}{\eta} + \frac{\eta}{8} \sum_{t=1}^{T} 1 \tag{24}$$

We note that the term $\sum_{t=1}^{T}\sum_{a=1}^{K} p_t(a)X_t^a$ is the expected cumulative loss of the hedge after T rounds. Subtracting this with the minimum loss, we get the expected regret of the Hedge. Thus we can write:

$$\mathbb{E}[R_T] \leq \frac{\ln K}{\eta} + \frac{\eta}{8}T \tag{25}$$

where T follows from summing 1 T times. Now we wish to determine what η minimizes the expected regret. To do so, we take the derivative of (25), with respect to η and set it to zero:

$$\frac{\partial}{\partial \eta}\left[\frac{\ln k}{\eta} + \frac{\eta}{8}T\right] = -\frac{\ln K}{\eta^2} + \frac{T}{8} \tag{26}$$

setting it to zero and solving for η:

$$0 = -\frac{\ln K}{\eta^2} + \frac{T}{8} \Rightarrow \tag{27}$$

$$\eta = \sqrt{\frac{8 \ln K}{T}} \tag{28}$$

And if we take the second derivative we get

$$\frac{\partial^2}{\partial \eta^2}\left[\frac{\ln k}{\eta} + \frac{\eta}{8}T\right] = \frac{\ln K}{\eta^3} > 0 \tag{29}$$

Thus we have an extrema point. Now we can plug this value back into the bound on the expected regret in (25) and get the following:

$$\mathbb{E}[R_T] \leq \frac{\ln K}{\eta} + \frac{\eta}{8}T \tag{30}$$

$$= \frac{\ln K}{\sqrt{\frac{8 \ln K}{T}}} + \frac{\sqrt{\frac{8 \ln K}{T}}}{8}T \tag{31}$$

$$= \frac{\sqrt{T}\sqrt{\ln K}}{\sqrt{2}} \tag{32}$$

$$= \sqrt{\frac{1}{2}T \ln K} \tag{33}$$

Which is what we wanted to prove. Thus, using Hoeffding's lemma, we get

$$\mathbb{E}[R_T] \leq \sqrt{\frac{1}{2}T \ln K} \tag{34}$$

instead of:

$$\mathbb{E}[R_T] \leq \sqrt{2T \ln K} \tag{35}$$

3. Empirical comparison of FTL and Hedge

1

We wish to compare the empirical performance of the "Follow the Leader" algorithm with the hedge algorithm, using various learning rates. Their performance will be tested on predicting a binary sequence $X_1, X_2, ...$, that is, at each round the algorithm can predict a 0 or 1. The algorithms do not know the bias of the sequence μ, which the sequences are generated with respect to.

First I will discuss how the sequence is generated, then I will comment on the algorithms and last compare

their performance.
For each experiment, we are given a bias μ. At each iteration, a random number $r \in [0,1]$ is drawn and the i'th element of the sequence is sampled as followed:

$$X_i = \begin{cases} 0 & \text{if } \mu < r \\ 1 & \text{if } \mu \geq r \end{cases} \tag{36}$$

Thus at each element in the sequence, we have the probability:

$$P\{X_i = 1\} = \mu \tag{37}$$
$$P\{X_i = 0\} = 1 - \mu \tag{38}$$

The "Follow the Leader" algorithm picks at round t, the arm with the best empirical performance over the previous $t-1$ pulls. The algorithm is defined as followed:

Algorithm 3 Follow The Leader

```
procedure FTL
    for $t = 1, 2, ...do$ do
        Play $A_t = \arg\max_a \hat{\mu}_{t-1}(a)$
```

The Hedge algorithm is defined as followed:

Algorithm 4 Hedge

```
Input: Learning rates $\eta_i \geq \eta_2 .... > 0$
$\forall a : L_0(a) = 0$
for $t = 1, 2, ...do$ do
    $\forall a : p_t(a) = \frac{e^{-\eta L_{t-1}(a)}}{\sum_a e^{-\eta L_{t-1}(a)}}$
    Sample $A_t$ according to $p_t$ and play it
    Observe $l_t^1, ..., l_t^K$ and suffer $l_t^{A_t}$
    $\forall a : L_t(a) = L_{t-1}(a) + l_t^a$
```

Where $L_{t-1}(a)$ is the accumulated loss, until round t. The prediction of the Hedge algorithm is done as followed, $r \in [0,1]$ is picked randomly:

$$X_i = \begin{cases} 0 & \text{if } r < p_t(a_0) \\ 1 & \text{else} \end{cases} \tag{39}$$

The experiment is performed with the following learning rates for the hedge algorithm:

$$\eta = \sqrt{\frac{2 \ln K}{T}}, \quad \eta = \sqrt{\frac{8 \ln K}{T}}, \quad \eta_t = \sqrt{\frac{\ln K}{t}}, \quad \eta_t = 2\sqrt{\frac{\ln K}{t}} \tag{40}$$

where T is the time horizon, and t is the given time index. Thus the two first weights are constant. The tested values of μ are the following:

$$\mu = \frac{1}{2} - \frac{1}{4}, \quad \mu = \frac{1}{2} - \frac{1}{8}, \quad \mu = \frac{1}{2} - \frac{1}{16} \tag{41}$$

The performance of the algorithms is given by the pseudo regret, which is defined as:

$$\bar{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} l_t^{A_t}\right] - T \min_a \mu(a) \tag{42}$$

We now run the five algorithms (FTL and four variations of the Hedge algorithm), in which we will get a series of predictions. At each time t, we know μ of the optimal and suboptimal arm. To get the expected value of (42) we simply sum up the μ for each prediction of either 0 or 1, that is either μ or $1-\mu$. The experiment is performed for each μ in (41), the average pseudo regret over 10 runs and the average pseudo regret + the standard deviation is plotted.

Results:

The following plots of the pseudo regret and the pseudo regret + standard deviations are shown for the three different settings of μ. The scribbled lines shows the average pseudo regret + the standard deviation for each t. The color of the scribbled lines identify what algorithm they are plotted for. The results will be commented and discussed in the next section.

$\mu = \mathbf{\frac{1}{2} - \frac{1}{4}}$:

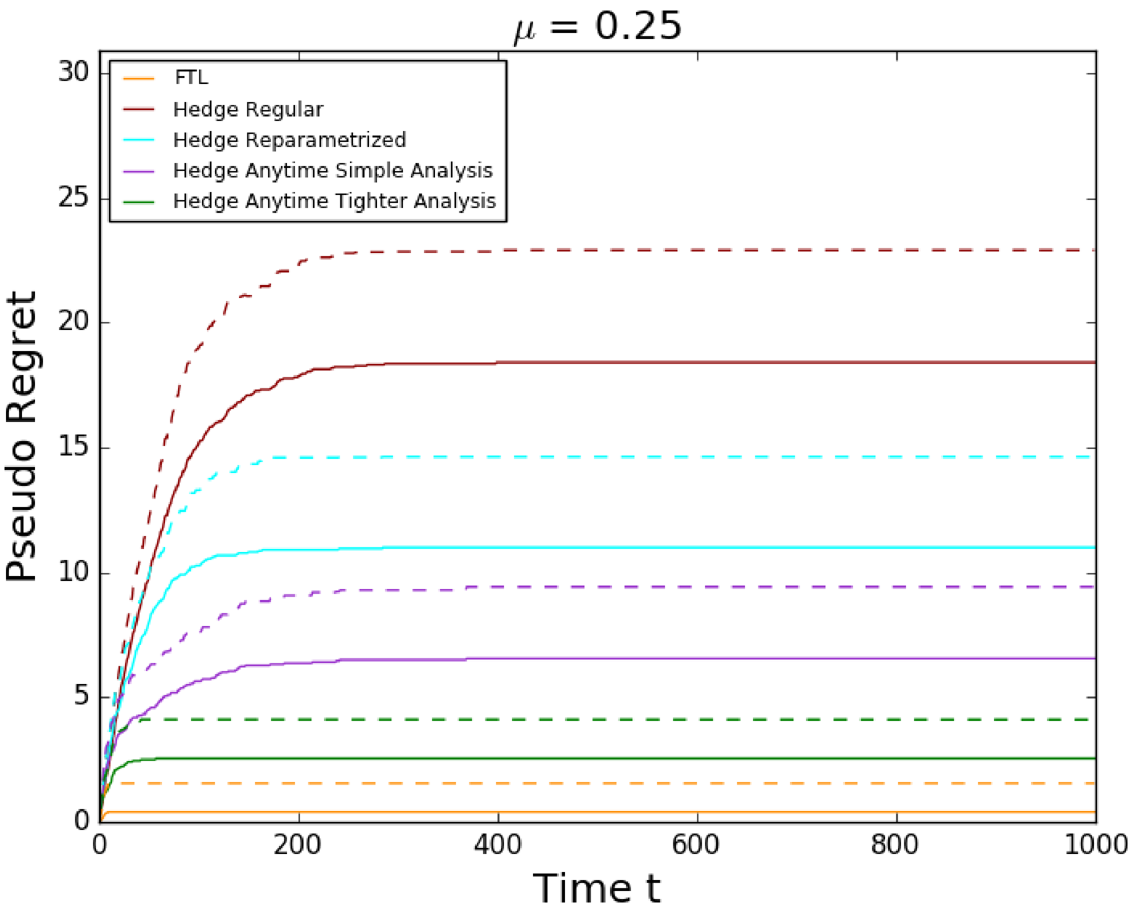

Figure 1: Average pseudo regret and average pseudo regret + standard deviation (scribbled line) over 10 runs with $\mu = 0.25$

$\mu = \mathbf{\frac{1}{2} - \frac{1}{8}}$:

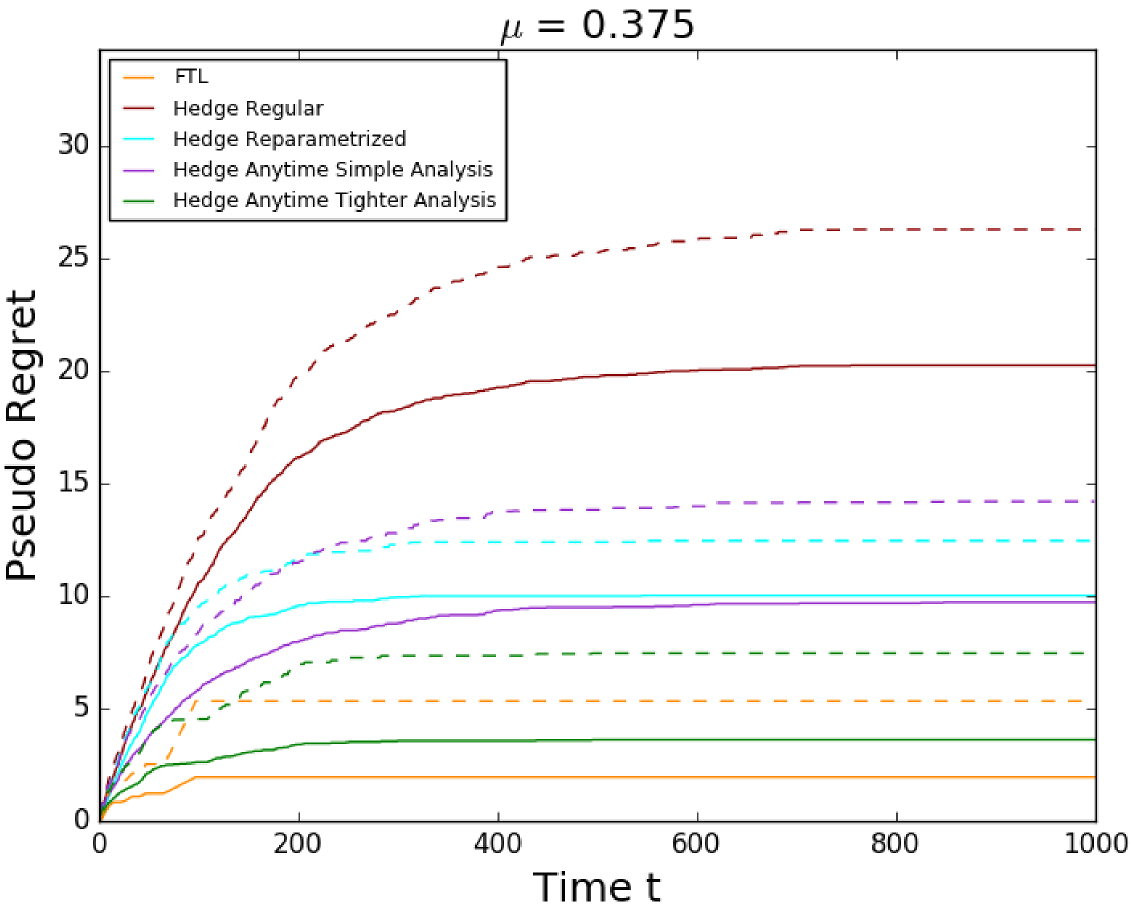

Figure 2: Average pseudo regret and average pseudo regret + standard deviation (scribbled line) over 10 runs with $\mu = 0.375$

$\mu = \mathbf{\frac{1}{2} - \frac{1}{16}}$:

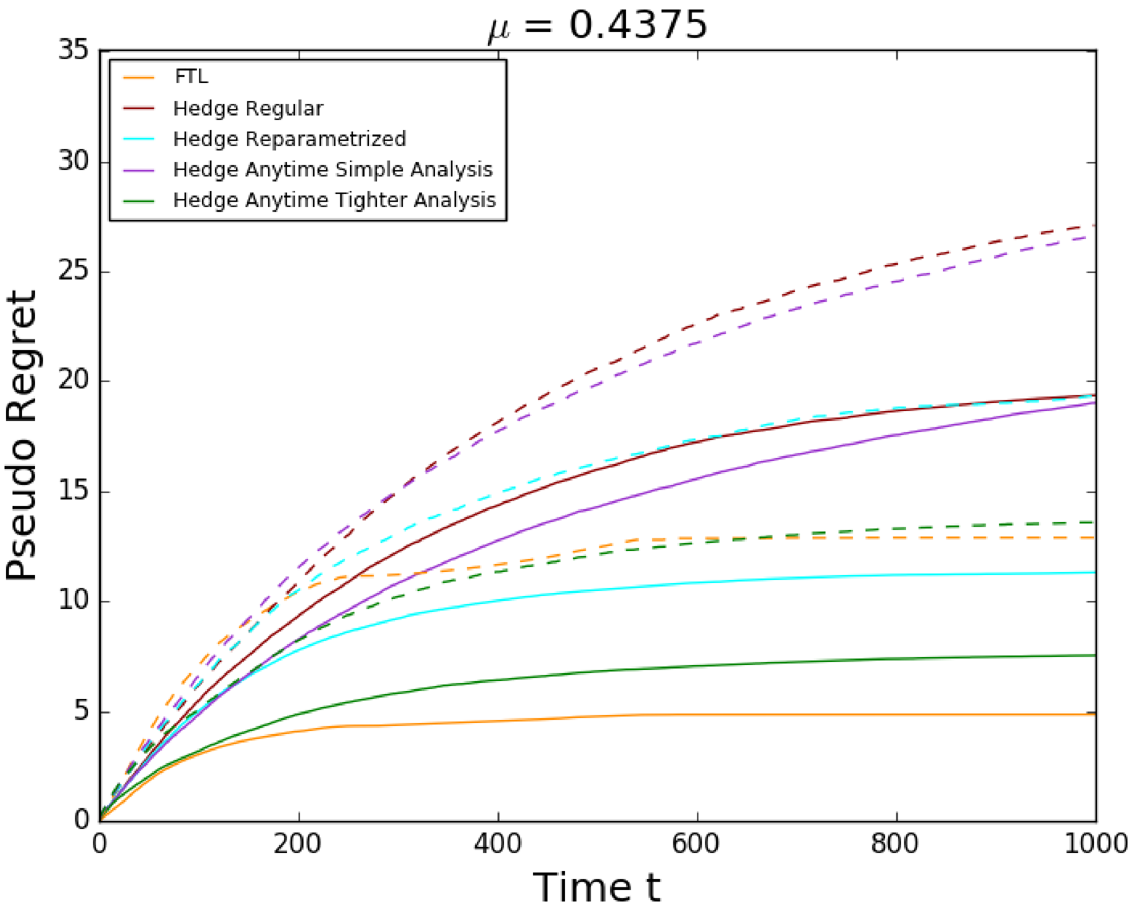

Figure 3: Average pseudo regret and average pseudo regret + standard deviation (scribbled line) over 10 runs with $\mu = 0.4375$

2

The "Follow the leader" algorithm has a low pseudo regret compared to the other algorithms. This is because in the pseudo regret, the series of pulled arms is compared to pulling the most optimal arm throughout the experiment. When the bias is low (i.e. the probability is great for one arm and small for the other), the optimal arm will quickly be identified and the "Follow the Leader" algorithm will pull the same arm throughout the experiment. Because it will pull the same arm as the optimal one, the pseudo regret will be low. As evident from figure 1, the pseudo regret is very low for the FTL approach. There is a small increase in the pseudo regret for the first couple of iterations and then the pseudo regret flattens out. This is because the approach will change between the two arms before identifying the most optimal one. As t increases, the output of the experiment will go towards the true μ, which is why the pseudo regret flattens out. As μ increases, the optimal arm will be harder to identify, because the gap between the two arms is smaller and the leader will be more inclined to switch between the arms. This is evident from figure 2, where the approach takes a longer time identifying the most optimal arm, and 3 where it takes even longer time before identifying the optimal arm. Interestingly the hedge algorithm appears to be more invariant to the increase in μ. Notably the Hedge Anytime with tighter analysis has the best performance, where the Hedge Anytime with simple analysis gets worse as μ increases. The Hedge with fixed weights η performs invariant to the change in μ.

3

For this task, we wish to create a sequence, such that the "Follow the Leader" approach has a bad performance on the sequence. The performance will then be compared to the various versions of the Hedge algorithm. The performance of each algorithm will be measured in regret, which is defined as:

$$R_T = \sum_{t=1}^{T} l_t^{A_t} - \min_a \sum_{t=1}^{T} l_t^a \tag{43}$$

thus the performance is measured by the accumulated loss of each pulled arm by the given algorithm, minus the accumulated loss of the arm which resulted in the lowest error in hindsight. The error is the zero-one loss, which is one if the wrong arm is predicted and zero otherwise. The adversarial sequence to the FTL algorithm is simply a sequence, which interchanging picks 0 and 1. This sequence has been chosen because the leader will shift at each iteration. For example if we consider the following sequence, and we take that ties are chosen to be 1:

$$0, 1, 0, 1, 0, 1, 0, 1, 0$$

then the FTL approach will predict the following sequence:

$$1, 0, 1, 0, 1, 0, 1, 0, 1$$

Thus the FTL algorithm will predict wrong at each iteration, whereas the optimal arm only will be wrong half of the iterations. The performance on the adversarial sequence is shown in figure 4 for the five algorithms.

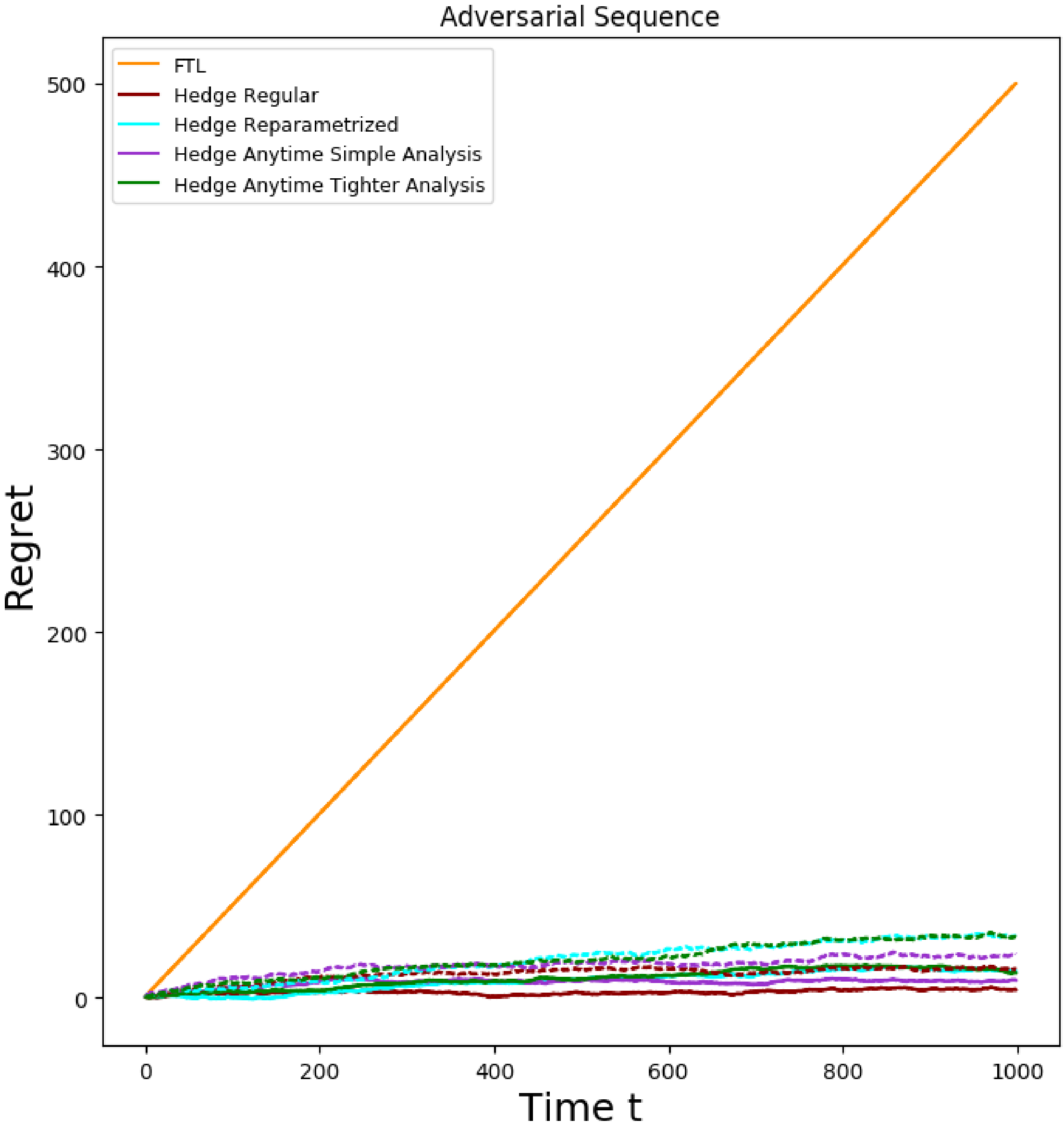

Figure 4: Average pseudo regret and average pseudo regret + standard deviation (scribbled line) over 10 runs with the adversarial sequence plotted

As evident from figure 4, the regret of the FTL algorithm is high, whereas the regret for the different variations of the hedge algorithm are low. The regret of the Hedge algorithms are magnitudes lower than the FTL, although fluctuating. It is difficult to identify which algorithm did the best job of predicting the adversarial sequence, but it appears that the Hedge algorithms with fixed weights performed slightly better. But in conclusion the Hedge algorithm performed much better than FTL on the adversarial sequence.