EE194/290S: ML for Sequential Decision under Uncertainty **Fall 2018**

Lecture 22: Adversarial Bandits

Lecturer: Anant Sahai, Vidya Muthukumar *Scribes: Brian Hung, Jihan Yin*

In Contextual Bandits, our regret is relative to a class Π of policies. A policy is a function that takes in the context x and maps it to an action a:[1] $\pi(x) \mapsto a$. We would like to do as well in hindsight as we would have done if we had known which policy to use and had seen everything.

22.1 Approach

Recall that in past lectures we developed algorithms for the Experts problem, where at each round we had full information about how each expert performed regardless of whether we had selected them or not. The leading algorithm in this setting is Multiplicative Weights, which we showed had $\mathcal{O}(\sqrt{T})$ regret.

We are now concerned with the Bandits problem in which feedback is limited: we only get to know the losses or rewards of the expert that we had selected. A central question of this lecture is can we can still do as well only given partial information? In a previous lecture, we saw that we could reduce the Bandits to the Experts problem by using a blocking trick, which segmented time into separate intervals of exploration and exploitation. When we randomized when these phases occurred and ran Multiplicative Weights, we found it was possible to achieve a regret bound of $\mathcal{O}(\sqrt[3]{T^2})$.

In this lecture, we want to recover the $\mathcal{O}(\sqrt{T})$ bound for the Bandits problem. The key idea is that we don't have allocate separate intervals to exploration and exploitation, nor do we have to mix in some amount of uniform exploration. Instead, it turns out that Multiplicative Weights, when given a small modification in the partial feedback setting, will have sufficient inherent exploration to give us the desired $\mathcal{O}(\sqrt{T})$ bound.

22.1.1 EXP3

Let's first consider the non-stochastic, adversarial multi-arm bandit environment *without* policies and contexts. How do we choose the learning rate and associated parameters (if there are any)? And how does our regret scale with respect to

- T, the time horizon, and
- $|\mathcal{A}| = N$, the number of different actions?

Recall the scenario where we had semi-sparse random explore times, and played the multiplicative weights algorithms to exploit outside of those times. What if we don't look at what happens during

[1] A policy can also map to a distribution over all arms, as we will see in EXP4. In that case, $\pi(x) \mapsto \boldsymbol{a}$.

exploit times? This was multiplicative weights with ϵ-exploration, and we got a regret that scales sub-linearly, $\mathcal{O}(T^{2/3})$. So what if we actually use the information we get during exploit times?

Instead of using the actual cumulative loss vectors $L_t(a) = \sum_{i=1}^{t} l_i(a)$, which we don't have, we will use the estimated cumulative loss $\hat{L}_t(a) = \sum_{i=1}^{t} \hat{l}_i(a)$. What is $\hat{l}_i(a)$? We would like that in expectation for $\hat{l}_i$ to be an unbiased estimator of l_t, that is $\mathbb{E}[\hat{l}_i(a)] = l_i(a)$. As a consequence of using $\hat{l}_i$, by linearity of expectation, $\hat{L}_t(a)$ will also be an unbiased estimator of $L_t(a)$.

This can be achieved by the following strategy of "the grass is greener on the other side",

$$\hat{l}_i(a) = \begin{cases} \frac{1}{p} l_i(a) & \text{if } A_i = a \text{ (observed the actual loss)} \\ 0 & \text{otherwise} \end{cases}$$

where p is the probability of choosing a at time step i. Let's verify that $\hat{l}_i$ is indeed an unbiased estimator.

$$\mathbb{E}[\hat{l}_i(a)] = \mathbb{E}\Big[\Pr(A_i = a) \frac{l_i(a)}{\Pr(A_i = a)} + (1 - \Pr(A_i = a)) \cdot 0 \Big] = l_i(a)$$

Intuitively, without $1/p$ scalar, our estimated loss would only be estimate for $p \times l_i$, the probability of seeing the actual loss times the actual loss itself. We divide $l_i(a)$ by the probability of choosing that action so that we can maintain our expectation property.

Beyond being unbiased, why is $\hat{l}_i$ a good estimator to use? $\hat{l}_i$ has the additional property that it is highly optimistic for arms that have not been pulled yet, and hence it is likely to explore those arms. Note that by setting the loss equal to 0, we are effectively maintaining the weight between rounds for that arm:

$$w_{t+1}(a) = w_t(a) \underbrace{\exp(\eta \cdot 0)}_{1} = w_t(a)$$

When we combine the ideas laid out above, the result is an "Exponential-weight algorithm for Exploration and Exploitation" (EXP3)[2].

Note that a similar algorithm can be derived for rewards by advantaging arms that have yielded a better return r, via

$$w_{t+1}(a) = w_t(a) \exp(+\eta \hat{r}_t(a)/N)$$

22.1.2 EXP4

Let's now turn to the non-stochastic, adversarial multi-arm bandit environment *with* policies and contexts. How do we choose the learning rate and associated parameters (if there are any)? And how does our regret scale with respect to

- T, the time horizon,

[2]The original paper for the algorithm by Auer et al. is linked in the references section.

Algorithm 1: EXP3 for losses

function EXP3 (η)

initialize $w_0(a) = 1$ for $a = 1, \ldots, N$

for $t = 1$ to T **do**

set $W_t = \sum_{a=1}^{N} w_t(a)$, and set for $a = 1, \ldots, N$

$$\Pr(A_t = a) = (1 - \eta)\frac{w_t(a)}{W_t} + \eta\frac{1}{N}$$

draw A_t randomly accordingly to the probabilities $\Pr(A_t = a)$

receive loss $l_t(a) \in [0, 1]$

set for $a = 1, \ldots, N$

$$\hat{l}_t(a) = \begin{cases} l_t(a)/\Pr(A_t = a) & \text{if } A_t = a \\ 0 & \text{otherwise} \end{cases}$$

$$w_{t+1}(a) = w_t(a) \exp(-\eta \hat{l}_t(a)/N)$$

end

- $|\mathcal{A}| = N$, the number of different actions, and
- $|\Pi|$, the number of policies?

We could approach the problem using the multi-armed bandits framework that we constructed above, treating each policy like an arm. The problem here is, regret will scale with respect to the number of policies. Because policies are functions of context, there can be a huge number of them — and this will cause regret to scale very badly!

To avoid this combinatorial explosion, we will instead focus on the space of possible actions when it comes selecting arms. In general, we can assume $N << |\Pi|$ and that in each round, multiple policies when given contexts will map to a same action. However, we still maintain a weight for each policy $w_t(\pi) \propto \exp(-\eta \hat{L}_{t-1}(\pi))$ where π includes the contexts we have seen and η is the learning rate. Note that for updating the weights, we now use losses associated with policies, not actions. These losses are given by

$$L_t(\pi) = \sum_{i=1}^{t} l_i(\pi(x_i))$$
$$l_i(\pi(x_i)) = l_i(a)$$

Again, as with EXP3, we will be substituting in estimated loss vectors $\hat{l}_t(\pi)$ for $l_t(\pi)$.

One unresolved question is, what is the probability of pulling an arm given policies and contexts? The probability of any arm a at time t is equal to sum of probabilities of all policies which would have chosen a as its action at time t.

The result is an "Exponential-weight algorithm for Exploration and Exploitation using Expert[3] advice" (EXP4). In this algorithm, during each round, each policy generates an advice vector which is a distribution over all arms and indicates that policy's recommended probability of playing an action a at time t. Note that the context x_t does not appear in the algorithm since it is only used by the policy to generate advice.

Note that a similar algorithm can be derived for rewards by advantaging arms that have yielded a better return r, via

$$w_{t+1}(a) = w_t(a) \exp(+\eta \hat{r}_t(a)/N)$$

[3]In the literature, the term 'experts' are often used interchangably with 'policies'

Algorithm 2: EXP4 for losses

function EXP4 (η)

initialize $w_0(\pi) = 1$ for $\pi = 1, \ldots, K$

for $t = 1$ to T **do**

get advice vectors $\boldsymbol{\xi}_t^1, \ldots, \boldsymbol{\xi}_t^K$

set $W_t = \sum_{\pi=1}^{K} w_t(\pi)$, and set for $a = 1, \ldots, N$

$$\Pr(A_t = a) = (1 - \eta) \sum_{\pi=1}^{K} \frac{w_t(\pi)\xi_t^\pi(a)}{W_t} + \eta \frac{1}{N}$$

draw A_t randomly accordingly to the probabilities $\Pr(A_t = a)$

receive loss $l_t(a) \in [0, 1]$

set for $a = 1, \ldots, N$

$$\hat{l}_t(a) = \begin{cases} l_t(a) / \Pr(A_t = a) & \text{if } A_t = a \\ 0 & \text{otherwise} \end{cases}$$

set for $\pi = 1, \ldots, K$

$$\hat{y}_t(\pi) = \boldsymbol{\xi}_t^\pi \cdot \hat{\boldsymbol{l}}_t$$

$$w_{t+1}(\pi) = w_t(\pi) \exp(-\eta \hat{y}_t(\pi)/N)$$

end

22.2 Regret Analysis

We know from previous lectures that using Multiplicative Weights as a blackbox algorithm achieves sublinear regret relative to the loss vectors it is given. In the case of adversarial bandits, we would just be feeding in the estimated loss vectors $\hat{l}_i$. However, does the analysis from before still hold true when given these new loss vectors?

The answer unfortunately turns out to be no. Recall that in those proofs, the original losses l_i were bounded between 0 and 1. In the limited feedback setting, the same cannot be said for $\hat{l}_i$ because we are scaling the observed losses by $1/p$ to recover the expectation property mentioned earlier. Here, l_t/p is a random variable with an increased range and potentially high variance. As a result of $\hat{l}_i$ being unbounded, we will need to redo the proof to show that our algorithm is still sublinear.

As we will see, instead of relying on a bound on the original losses, we will invoke a second moment-type bound[4] on l_t/p.

22.2.1 EXP3

Theorem 22.1

$$\mathbb{E}\Big[\text{Regret}_T(\text{EXP3})\Big] \leq \frac{1}{\eta}\ln N + \eta T N$$

Here, the regret and expected regret are defined as

$$R_T = \sum_{t=1}^{T} \hat{l}_t(a_t) - \min_{a\in\mathcal{A}} \sum_{t=1}^{T} l_t(a) \quad \text{and} \quad \mathbb{E}[R_T] = \sum_{t=1}^{T} \boldsymbol{w}_t \cdot \hat{\boldsymbol{l}}_t - \min_{a\in\mathcal{A}} \sum_{t=1}^{T} l_t(a).$$

Proof. See below. ■

As with our regret analysis of Hedge, we want to express it in terms of mixed loss. For time step t, this is defined as

$$\Phi_t = \frac{1}{\eta}\ln(\sum_{a=1}^{N} \exp(-\eta L_t(a))$$

The difference then across all rounds is

[4] In the original paper for EXP3 and EXP4 by Auer et al., the authors added a uniform-exploration component to lower-bound p. In effect, this also upper-bounds l_t/p. While this helps with the actualized performance of the algorithms, it was found later that this uniform part was unnecessary in the analysis of expected regret.

$$\begin{aligned}\Phi_T - \Phi_0 &= \sum_{t=1}^{T} \Phi_t - \Phi_{t-1} \\ &= \sum_{t=1}^{T} \frac{1}{\eta} \ln\left(\frac{\sum_{a=1}^{N} \exp(-\eta L_{t-1}(a) - \eta l_t(a))}{\sum_{a=1}^{N} \exp(-\eta L_{t-1}(a)}\right) \\ &= \sum_{t=1}^{T} \frac{1}{\eta} \ln\left(\sum_{a=1}^{N} w_t(a) \exp(-\eta l_t(a))\right)\end{aligned}$$

How can we simplify this?

Lemma 22.2 For all $x \geq 0$,

$$e^{-x} \leq 1 - x + \frac{1}{2}x^2$$

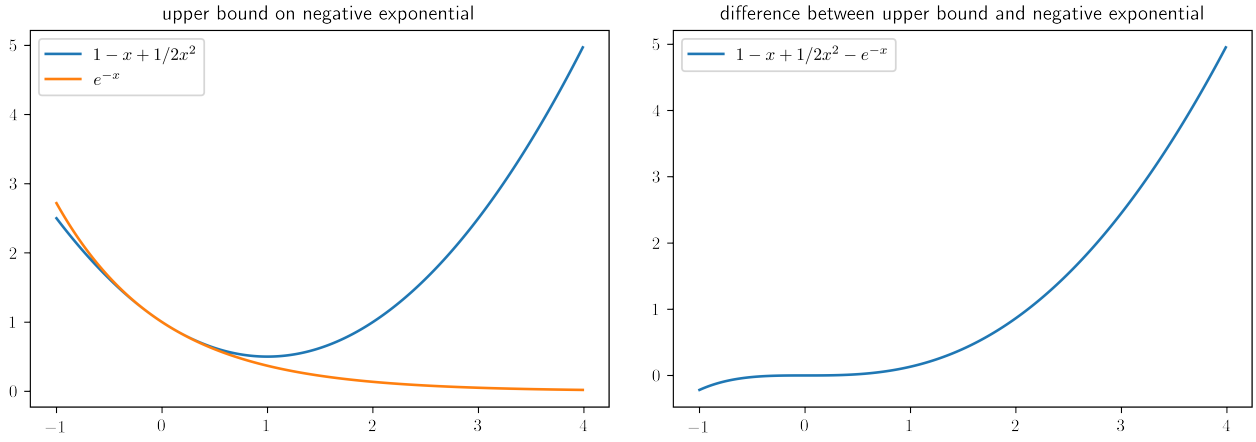

Figure 22.1: visual proof of Lemma 22.1

Proof. Just show that $e^{-x} - 1 + x - \frac{1}{2}x^2$ has a maximum at $x = 0$ and is decreasing using the first and second derivative tests. ■

Lemma 22.3 For all x,

$$\ln(1+x) \leq x$$

Proof. Same method as for proving the previous lemma. ■

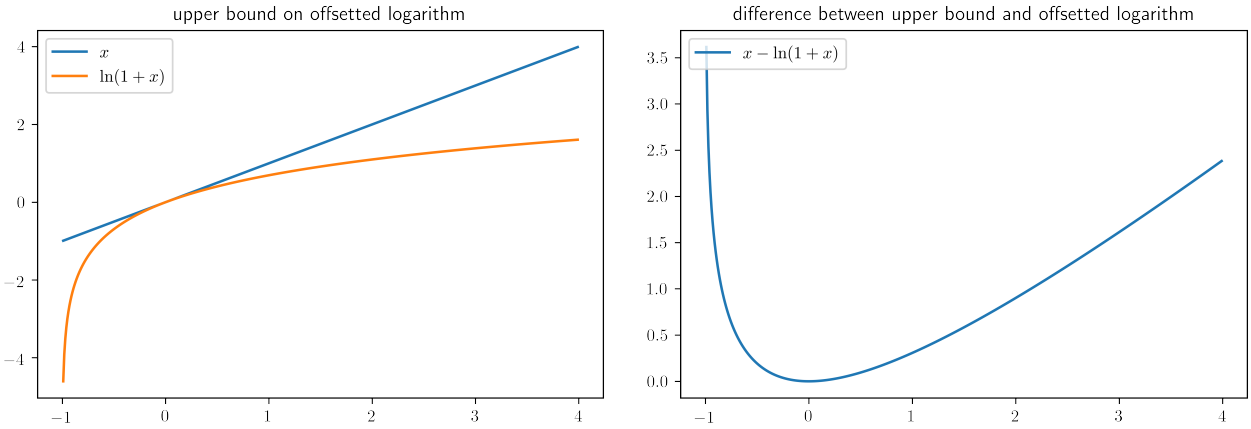

Figure 22.2: visual proof of Lemma 22.2

Using these two lemmas, we can derive an upper bound on the cumulative mixed loss.

$$\begin{aligned}
\Phi_T - \Phi_0 &= \sum_{t=1}^{T} \frac{1}{\eta} \ln\Big(\sum_{a=1}^{N} w_t(a) \exp(-\eta l_t(a))\Big) \\
&\leq \sum_{t=1}^{T} \frac{1}{\eta} \ln\Big(\sum_{a=1}^{N} w_t(a)\Big[1 - \eta l_t(a) + \frac{1}{2}\eta^2 l_t(a)^2\Big]\Big) && \text{Lemma 22.1} \\
&\leq \sum_{t=1}^{T} \frac{1}{\eta} \ln\Big(\Big[1 - \eta \sum_{a=1}^{N} w_t(a) l_t(a) + \frac{1}{2}\eta^2 \sum_{a=1}^{N} w_t(a) l_t(a)^2\Big]\Big) \\
&\leq \sum_{t=1}^{T} \frac{1}{\eta}\Big[- \eta \sum_{a=1}^{N} w_t(a) l_t(a) + \frac{1}{2}\eta^2 \sum_{a=1}^{N} w_t(a) l_t(a)^2\Big] && \text{Lemma 22.2} \\
&\leq \sum_{t=1}^{T} \Big[- \boldsymbol{w}_t \cdot \boldsymbol{l}_t + \eta \sum_{a=1}^{N} w_t(a) l_t(a)^2\Big]
\end{aligned}$$

How can we use mixed loss to upper bound the regret we defined earlier? By a small slight of hand,

$$\Phi_0 = \frac{1}{\eta} \ln(\sum_{a=1}^{N} 1) = \frac{1}{\eta} \ln N$$

we can rewrite the formula above:

$$\begin{aligned}
\Phi_T - \Phi_0 &\leq -\sum_{t=1}^{T} \boldsymbol{w}_t \cdot \boldsymbol{l}_t + \eta \sum_{t=1}^{T} \sum_{a=1}^{N} w_t(a) l_t(a)^2 \\
\Phi_T - \frac{1}{\eta} \ln N &\leq \\
\Phi_T + \sum_{t=1}^{T} \boldsymbol{w}_t \cdot \boldsymbol{l}_t &\leq \frac{1}{\eta} \ln N + \eta \sum_{t=1}^{T} \sum_{a=1}^{N} w_t(a) l_t(a)^2
\end{aligned}$$

The additional step is to substitute Φ_T for $-L_T(a)$. The inequality trivially holds true for all values of a since Φ_T and $L_T(a)$ are both ≥ 0.

$$\begin{aligned}
\sum_{t=1}^{T} \boldsymbol{w}_t \cdot \boldsymbol{l}_t + \Phi_T &\leq \frac{\ln N}{\eta} + \eta \sum_{t=1}^{t} \sum_{a=1}^{N} w_t(a) l_t(a)^2 \\
\sum_{t=1}^{T} \boldsymbol{w}_t \cdot \boldsymbol{l}_t - L_T(a) &\leq
\end{aligned}$$

Using this bound on the original losses, we can obtain a bound on the estimated losses and on the expected regret

$$\begin{aligned}
\mathbb{E}\Big[\sum_{t=1}^{T} \boldsymbol{w}_t \cdot \hat{\boldsymbol{l}}_t - \min_a L_T(a)\Big] &\leq \frac{\ln N}{\eta} + \eta \sum_{t=1}^{T} \sum_{a=1}^{N} \mathbb{E}\Big[w_t(a) \cdot \hat{l}_t(a)^2\Big] \\
&\leq \frac{\ln N}{\eta} + \eta \sum_{t=1}^{T} \sum_{a=1}^{N} \mathbb{E}\Big[w_t(a)\Big] \cdot \frac{1}{\Pr(a)} l_i(a)^2 \\
&\leq \frac{\ln N}{\eta} + \eta \sum_{t=1}^{T} \sum_{a=1}^{N} l_i(a)^2 \\
&\leq \frac{\ln N}{\eta} + \eta \sum_{t=1}^{T} \sum_{a=1}^{N} 1 \\
&\leq \frac{\ln N}{\eta} + \eta T N
\end{aligned}$$

To jump from the first to the fourth equation, we used the facts that

- $\mathbb{E}\Big[\hat{l}_t(a)^2\Big] = \frac{1}{\Pr(A_t = a)} l_t(a)^2$, which got us the 2nd equation,
- $\mathbb{E}\Big[w_t(a)\Big] = \Pr(a)$, which got us the 3rd equation, and
- $l_i(a) \in [0,1] \implies l_i(a)^2 \in [0,1]$, which got us the 4th equation.

And that concludes our proof.

Theorem 22.4

$$\mathbb{E}\Big[\text{Regret}_T(\text{EXP3})\Big] \leq \sqrt{2TN\ln N}$$

Proof. Set

$$\eta = \sqrt{\frac{\ln N}{TN}}$$

and evaluate the bound above. ■

This is a better regret scaling than when we had used multiplicative weights and ignored the losses incurred during exploration, which had regret $\mathcal{O}(\sqrt[3]{T^2})$.

The key difference between the EXP3 algorithm and multiplicative weights with ϵ-exploration is that EXP3 does not drop observations in any round, as opposed to multiplicative weights which drops observations with probability $1-\epsilon$.

22.2.2 EXP4

What if we take policies and contexts into account?

Theorem 22.5

$$\mathbb{E}\Big[\text{Regret}_T(\text{EXP4})\Big] \leq \frac{1}{\eta}\ln|\Pi| + \eta TN$$

Here, the regret and expected regret are defined as

$$R_T = \sum_{t=1}^{T}\hat{l}_t(a_t) - \min_{\pi\in\Pi}\sum_{t=1}^{T}\boldsymbol{\xi}_t^{\pi}\cdot\hat{\boldsymbol{l}}_t \quad \text{and} \quad \mathbb{E}[R_T] = \sum_{t=1}^{T}\boldsymbol{w}_t\cdot\hat{\boldsymbol{l}}_t - \min_{\pi\in\Pi}\sum_{t=1}^{T}\boldsymbol{\xi}_t^{\pi}\cdot\hat{\boldsymbol{l}}_t$$

Proof. From the regret bound for EXP3, we have

$$\begin{aligned}
\mathbb{E}\Big[\sum_{t=1}^{T} \boldsymbol{w}_t \cdot \hat{\boldsymbol{l}}_t - \min_{\pi\in\Pi}\sum_{t=1}^{T} \boldsymbol{\xi}_t^{\pi} \cdot \hat{\boldsymbol{l}}_t\Big] &\leq \frac{1}{\eta}\ln|\Pi| + \eta\sum_{t=1}^{T}\mathbb{E}\Big[\sum_{\pi=1}^{K} w_t(\pi)\cdot \hat{l}_t(\pi(x_t))^2\Big] \\
&\leq \frac{1}{\eta}\ln|\Pi| + \eta\sum_{t=1}^{T}\sum_{a=1}^{N}\sum_{\pi:\pi(x_t)=a}\mathbb{E}\Big[w_t(\pi)\cdot \hat{l}_t(\pi(x_t))^2\Big] \\
&\leq \frac{1}{\eta}\ln|\Pi| + \eta\sum_{t=1}^{T}\sum_{a=1}^{N}\sum_{\pi:\pi(x_t)=a}\mathbb{E}\Big[w_t(\pi)\cdot \hat{l}_t(a)^2\Big] \\
&\leq \frac{1}{\eta}\ln|\Pi| + \eta\sum_{t=1}^{T}\sum_{a=1}^{N}\mathbb{E}\Big[w_t(a)\cdot \hat{l}_t(a)^2\Big] \\
&\leq \frac{1}{\eta}\ln|\Pi| + \eta\sum_{t=1}^{T}\sum_{a=1}^{N}\mathbb{E}\Big[w_t(a)\Big]\cdot\frac{1}{\Pr(a)} l_t(a)^2 \\
&\leq \frac{1}{\eta}\ln|\Pi| + \eta\sum_{t=1}^{T}\sum_{a=1}^{N} l_t(a)^2 \\
&\leq \frac{1}{\eta}\ln|\Pi| + \eta\sum_{t=1}^{T}\sum_{a=1}^{N} 1 \\
&\leq \frac{1}{\eta}\ln|\Pi| + \eta TN
\end{aligned}$$

To jump from the first to the third equation, we used the facts that

- $\mathbb{E}\Big[\sum_{\pi=1}^{K} w_t(\pi)\cdot \hat{l}_t(\pi(x_t))^2\Big] = \sum_{\pi:\pi(x_t)=a}\mathbb{E}\Big[w_t(\pi)\cdot \hat{l}_t(\pi(x_t))^2\Big]$, which is just rewriting loss in terms of actions instead of policies, and
- $\mathbb{E}\Big[w_t(\pi)\cdot \hat{l}_t(\pi(x_t))^2\Big] = \mathbb{E}\Big[w_t(\pi)\cdot \hat{l}_t(a)^2\Big]$, which is equivalent since $\pi(x_t) = a$.

And that concludes our proof. ■

Theorem 22.6

$$\mathbb{E}\Big[\text{Regret}_T(\text{EXP4})\Big] \leq \sqrt{2TN\ln|\Pi|}$$

Proof. Set

$$\eta = \sqrt{\frac{\ln|\Pi|}{TN}}$$

and evaluate the bound above. ■

22.3 Recap

algorithm	**setting**	**regret bound**
multiplicative weights	experts with full feedback	$\mathcal{O}(\sqrt[2]{T})$
ϵ-multiplicative weights	experts with limited feedback	$\mathcal{O}(\sqrt[3]{T^2})$
EXP3	non-stochastic, adversarial bandits	$\mathcal{O}(\sqrt[2]{TN \log N})$
EXP4	non-stochastic, adversarial bandits with contexts	$\mathcal{O}(\sqrt[2]{TN \log K})$

22.4 References

1. Auer et al. The non-stochastic multi-armed bandit problem. 18 November 2001.
2. Abernethy, Jacob. Lecture 22: Adversarial Multi-Armed Bandits. EECS 589-005: Theoretical Foundations of Machine Learning. Fall 2015. University of Michigan.
3. Dekel, Ofer. Lecture 9: EXP3 Regret Analysis. CSE599s: Online Learning. 29 April 2014. University of Washington.
4. Zhou, Li. A Survey on Contextual Multi-armed Bandits. 13 August 2015. Carnegie Mellon University.