bounded by $C\sqrt{nK\log(n)}$? The lazy way is to push part of the proof into the assumptions. For UCB this might mean replacing a subgaussian assumption with a condition that the data generating process for each arm satisfies the conclusion of the core concentration result (Corollary 5.1). A more ambitious goal is to define the subset of rewards for which the regret is bounded by some value and try to characterize this set. To our knowledge these ideas have not been explored in bandits and barely at all in machine learning more broadly.

Bibliographic remarks

The quote by George Box was used several times with different phrasings [Box, 1976, 1979]. The adversarial framework has its roots in game theory with familiar names like Hannan [1957] and Blackwell [1954] producing some of the early work. The nonstatistical approach has enjoyed enormous popularity since the 1990's and has been adopted wholeheartedly by the theoretical computer science community [Vovk, 1990, Littlestone and Warmuth, 1994, and many many others]. For bandits the earliest work that we know of is by Auer et al. [1995]. There is now a big literature on adversarial bandits, which we will cover in more depth in the chapters that follow.

This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at http://banditalgs.com. Feedback on any aspect is very welcome: banditalgs@gmail.com

11 The Exp3 Algorithm

Let $K > 1$ be the number of arms. A ***K*-armed adversarial bandit** is an arbitrary sequence of reward vectors $\nu = (x_1, \dots, x_n)$ where $x_t \in [0,1]^K$ for each $t \in [n]$. In each round the learner chooses an action $A_t \in [K]$ and observes reward $X_t = x_{tA_t}$. We do not capitalize the reward vectors (x_t) because they are not random. The learner will usually randomize their decisions so that A_t and X_t are random variables and hence capitalized.

Like in the stochastic setting, a policy can be viewed as a mapping from interaction sequences to a distribution over the actions. For stochastic bandits we did not yet make use a randomized policy, but for adversarial bandits this is crucial. Given a policy π the conditional distribution over the actions having observed $A_1, X_1, \dots, A_{t-1}, X_{t-1}$ is $P_t = \pi(\cdot \mid A_1, X_1, \dots, A_{t-1}, X_{t-1}) \in \mathcal{P}_{K-1}$.

The performance of a policy π on environment ν is measured by the expected regret, which is the expected loss in revenue of the policy relative to the best fixed action in hindsight.

$$R_n(\pi, \nu) = \max_i \sum_{t=1}^n x_{ti} - \mathbb{E}\left[\sum_{t=1}^n x_{tA_t}\right] . \tag{11.1}$$

When π and ν are clear from the context we may just write R_n in place of $R_n(\pi, \nu)$.

The only source of randomness in the regret comes from the randomness in the actions of the learner. Of course the interaction with the environment means the action chosen in round t may depend on actions $s < t$ as well as the observed rewards until round t.

Like in the stochastic setting, we are often interested in the worst-case regret over all environments, which is

$$R_n^*(\pi) = \sup_{\nu \in [0,1]^{nK}} R_n(\pi, \nu) .$$

The main question is whether or not there exist policies π for which $R_n^*(\pi)$ is sublinear in n. In Exercise 11.2 you will show that for deterministic policies $R_n^*(\pi) \geq n(1 - 1/K)$, which follows by constructing a bandit so that $x_{tA_t} = 0$ for

all t and $x_{ti} = 1$ for $i \neq A_t$. Because of this, sublinear worst-case regret is only possible by using a randomized policy.

Readers familiar with game theory will not be surprised by the need for randomization. The interaction between learner and adversarial bandit can be framed as a two-player zero-sum game between the environment and learner. The moves for the environment are the possible reward sequences and for the player they are the set of policies. The payoff for the environment/learner is the regret and its negation respectively. Since the player goes first, the only way to avoid being exploited is to choose a policy that randomizes.

While stochastic and adversarial bandits seem quite different, it turns out that the optimal worst case regret is the same up to constant factors and that lower bounds for adversarial bandits are invariably derived in the same manner as for stochastic bandits (see Part IV). In this chapter we present a simple algorithm for which the worst-case regret is suboptimal by just a logarithmic factor. First though, we explore the differences and similarities between stochastic and adversarial environments.

We already noted that deterministic strategies will have linear regret for some adversarial bandit. Since all the strategies in Part II were deterministic, they are not well suited for the adversarial setting. This immediately implies that policies that are good for stochastic bandit can be very suboptimal in the adversarial setting. What about the other direction? Will an adversarial bandit strategy have small expected regret in the stochastic setting? Let π be an adversarial bandit policy and $\nu = (\nu_1, \dots, \nu_K)$ be a stochastic bandit with ν_i supported on a subset of $[0, 1]$ for each i. Next let X_{ti} be sampled from ν_i for each $i \in [K]$ and $t \in [n]$ and assume these random variables are mutually independent. Then by Jensen's inequality and convexity of the maximum function we have

$$\begin{aligned} R_n(\nu, \pi) &= \max_i \mathbb{E}\left[\sum_{t=1}^n (X_{ti} - X_{tA_t})\right] \leq \mathbb{E}\left[\max_i \sum_{t=1}^n (X_{ti} - X_{tA_t})\right] \\ &= \mathbb{E}\left[R_n(\pi, (X_{ti}))\right] \leq R_n^*(\pi)\,, \end{aligned}$$

where the regret in the first line is the stochastic regret and in the last it is the adversarial regret. Therefore the worst-case stochastic regret is upper bounded by the worst-case adversarial regret. Going the other way, the above inequality also implies the worst-case regret for adversarial problems is lower bounded by the worst-case regret on stochastic problems with rewards bounded in $[0, 1]$. In Chapter 15 we prove the worst-case regret for stochastic bandits is at least $c\sqrt{nK}$, where $c > 0$ is a universal constant. And so for the same universal constant the minimax regret for adversarial bandits satisfies

$$R_n^* = \inf_\pi \sup_{\nu \in [0,1]^{nK}} R_n(\pi, \nu) \geq c\sqrt{nK}\,.$$

11.1 Importance-weighted estimators

A key ingredient of all adversarial bandit algorithms is a mechanism for estimating the reward of unplayed arms. Recall that P_t is the conditional distribution of the action played in round t and let P_{ti} denote the conditional probability that the policy chooses action $A_t = i$,

$$P_{ti} = \mathbb{P}\left(A_t = i \mid X_1, \dots, X_{t-1}, A_1, \dots, A_{t-1}\right) ,$$

In what follows we assume that $P_{ti} > 0$ almost surely, which is true for all policies considered in this chapter. Until you know how to do it, estimating the reward for all arms simultaneously using only P_t and the observed reward seems like a hopeless endeavor. The idea is to use the **importance-weighted estimator** given by

$$\hat{X}_{ti} = \frac{\mathbb{I}\{A_t = i\} X_t}{P_{ti}} . \tag{11.2}$$

One way to get a first impression about the quality of an estimator is to calculate its mean and variance. Is the mean of $\hat{X}_{ti}$ close to x_{ti}? Does $\hat{X}_{ti}$ have a small variance? Let $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid A_1, X_1, \dots, A_{t-1}, X_{t-1}]$ denote the conditional expectation given the history up to time t. Then the conditional expectation of $\hat{X}_{ti}$ satisfies

$$\mathbb{E}_t[\hat{X}_{ti}] = x_{ti} , \tag{11.3}$$

which means that $\hat{X}_{ti}$ is an unbiased estimate of x_{ti} given whatever history has been generated. To see why Eq. (11.3) holds, let $A_{ti} = \mathbb{I}\{A_t = i\}$ so that $X_t A_{ti} = x_{ti} A_{ti}$ and

$$\hat{X}_{ti} = \frac{A_{ti}}{P_{ti}} x_{ti} .$$

Now $\mathbb{E}_t[A_{ti}] = P_{ti}$ and since P_{ti} is a function of $A_1, X_1, \dots, A_{t-1}, X_{t-1}$, we get

$$\mathbb{E}_t[\hat{X}_{ti}] = \mathbb{E}_t\left[\frac{A_{ti}}{P_{ti}} x_{ti}\right] = \frac{x_{ti}}{P_{ti}} \mathbb{E}_t[A_{ti}] = \frac{x_{ti}}{P_{ti}} P_{ti} = x_{ti} .$$

By the tower rule for conditional expectation, (11.3) also implies that $\mathbb{E}[\hat{X}_{ti}] = \mathbb{E}[\mathbb{E}_t[\hat{X}_{ti}]] = x_{ti}$. For the variance we proceed in the same manner by considering the conditional variance $\mathbb{V}_t[\hat{X}_{ti}]$, which for arbitrary random variable U is

$$\mathbb{V}_t[U] = \mathbb{E}_t\left[(U - \mathbb{E}_t[U])^2\right] .$$

So $\mathbb{V}_t[\hat{X}_{ti}]$ is a random variable that measures the variance of $\hat{X}_{ti}$ conditioned on the past. Calculating the conditional variance using the definition of $\hat{X}_{ti}$ and Eq. (11.3) shows that

$$\mathbb{V}_t[\hat{X}_{ti}] = \mathbb{E}_t[\hat{X}_{ti}^2] - x_{ti}^2 = \mathbb{E}_t\left[\frac{A_{ti} x_{ti}^2}{P_{ti}^2}\right] - x_{ti}^2 = \frac{x_{ti}^2 (1 - P_{ti})}{P_{ti}} . \tag{11.4}$$

This can be extremely large when P_{ti} is small and x_{ti} is bounded away from zero. In the notes and exercises we shall see to what extent this can cause trouble. The estimator in (11.2) is the first that comes to mind, but there are alternatives. For example,

$$\hat{X}_{ti} = 1 - \frac{\mathbb{I}\{A_t = i\}}{P_{ti}}(1 - X_t)\,. \tag{11.5}$$

This estimator is still unbiased. Rewriting the formula in terms of $y_{ti} = 1 - x_{ti}$ and $Y_t = 1 - X_t$ and $\hat{Y}_{ti} = 1 - \hat{X}_{ti}$ leads to

$$\hat{Y}_{ti} = \frac{\mathbb{I}\{A_t = i\}}{P_{ti}}\, Y_t\,.$$

This is the same as (11.2) except that Y_t has replaced X_t. The terms y_{ti}, Y_t and $\hat{Y}_{ti}$ should be interpreted as **losses**. Had we started with losses to begin with then this would have been the estimator that first came to mind. For obvious reasons, the estimator in Eq. (11.5) is called the **loss-based importance-weighted estimator**. The conditional variance of this estimator is essentially the same as Eq. (11.4):

$$\mathbb{V}_t[\hat{X}_{ti}] = \mathbb{V}_t[\hat{Y}_{ti}] = y_{ti}^2 \frac{1 - P_{ti}}{P_{ti}}\,.$$

The only difference is that the variance now depends on y_{ti}^2 rather than x_{ti}^2. Which is better then depends on the rewards for arm i, with smaller rewards suggesting the superiority of the first estimator and larger rewards (or small losses) suggesting the superiority of the second estimator. At this stage, one could be suspicious about the role of zero in this argument. Can we change the estimator (either one of them) so that it is more accurate for actions whose reward is close to some specific value v? Of course! Just change the estimator so that v is subtracted from the observed reward (or loss), then use the importance sampling formula, and subsequently add back v. The problem is that the optimal value of v depends on the unknown quantity being estimated. Also note that the dependence of the variance on P_{ti} is the same for both estimators and since the rewards are bounded it is this term that usually contributes most significantly. In Exericise 11.4 we ask you to show that all unbiased estimators in this setting are importance-weighted estimators.

Although the two estimators seem quite similar, it should be noted that the first estimator takes values in $[0, \infty)$ while the second takes values in $(-\infty, 1]$. Soon we will see that this difference has a big impact on the usefulness of these estimators when used in the Exp3 algorithm.

11.2 The Exp3 algorithm

The importance weighted estimator provides us with the means to estimate the reward. The next step is to choose the distribution over actions $P_t = (P_{ti})_i$. The simplest algorithm for adversarial bandits is called Exp3, which stands for "**Exp**onential-weight algorithm for **Exp**loration and **Exp**loitation". The reason for this name will become clear after the explanation of the algorithm. Let $\hat{S}_{ti} = \sum_{s=1}^{t} \hat{X}_{si}$ be the total estimated reward by the end of round t. It seems natural to choose the action-selection probabilities so that actions with larger estimated reward receive more weight. While there are many ways to map $\hat{S}_{ti}$ into probabilities, a simple and popular choice is called **exponential weighting**, which for tuning parameter $\eta > 0$ sets

$$P_{ti} = \frac{\exp(\eta \hat{S}_{t-1,i})}{\sum_j \exp(\eta \hat{S}_{t-1,j})} . \tag{11.6}$$

The parameter η is called the **learning rate** and its role is to control how aggressively P_{ti} is pushed towards arms for which the estimated cumulative reward is highest. As $\eta \to \infty$, the probability mass in P_t quickly concentrates on $\operatorname{argmax}_i \hat{S}_{t-1,i}$. It is here that the exploration/exploitation dilemma raises its head. If η is large, then the resulting policy will explore with low probability. But when P_{ti} is small, then the variance of the importance-weighted estimator is large and the estimates for these arms could be very poor. The consequence is the usual tension. Large η leads to overconfident policies and small η leads to excessive exploration.

There are many ways to set η, including allowing it to vary with time. In this chapter we restrict our attention to the simplest case by choosing η to depend only on the number of actions K and the horizon n. Since the algorithm depends on η this means that the horizon must be known in advance. This is relaxed in subsequent chapters.

For practical implementations it is useful to note that P_t can be calculated incrementally by

$$P_{t+1,i} = \frac{P_{ti} \exp(\eta \hat{X}_{ti})}{\sum_{j=1}^{K} P_{tj} \exp(\eta \hat{X}_{tj})} . \tag{11.7}$$

Computing the summation in the denominator can be numerically unstable because its terms can vary by several orders of magnitude. There are a variety of approaches for summing floats in a numerically stable way. One of the simplest is Kahan's algorithm [Kahan, 1965]. An even better approach is to note that Eq. (11.7) does not change if all $\hat{S}_{ti}$ are translated by some fixed

Input: n, K, η
Set $\hat{S}_{0i} = 0$ for all i
for $t = 1, \ldots, n$ **do**
    Calculate the sampling distribution P_t:

$$P_{ti} = \frac{\exp\left(\eta \hat{S}_{t-1,i}\right)}{\sum_{j=1}^{K} \exp\left(\eta \hat{S}_{t-1,j}\right)}$$

    Sample $A_t \sim P_t$ and observe reward X_t
    Calculate $\hat{S}_{ti}$:

$$\hat{S}_{ti} = \hat{S}_{t-1,i} + 1 - \frac{\mathbb{I}\{A_t = i\}(1 - X_t)}{P_{ti}}$$

end for

Algorithm 8: Exp3

amount. Let $\tilde{S}_{ti} = \hat{S}_{ti} - \min_j \hat{S}_{tj}$ so that

$$P_{t+1,i} = \frac{\exp(\eta \tilde{S}_{ti})}{\sum_{j=1}^{K} \exp(\eta \tilde{S}_{tj})}\,.$$

Care is still required for the summation in the denominator, but the number of floating point multiplications has been reduced significantly.

11.3 Regret analysis

We are now ready to bound the expected regret of Exp3 (Algorithm 8).

THEOREM 11.1 *Let $\nu = (x_{ti}) \in [0,1]^{nK}$ be an arbitrary adversarial bandit and π be the policy of Exp3 (Algorithm 8) with learning rate $\eta = \sqrt{\log(K)/(nK)}$. Then*

$$R_n(\pi, \nu) \leq 2\sqrt{nK \log(K)}\,.$$

Proof For any arm i define

$$R_{ni} = \sum_{t=1}^{n} x_{ti} - \mathbb{E}\left[\sum_{t=1}^{n} X_t\right],$$

which is the expected regret relative to using action i in all the rounds. The result will follow by bounding R_{ni} for all i, including the optimal arm. For the remainder of the proof, let i be some fixed arm. By the unbiasedness property of

$\hat{X}_{ti}$,

$$\mathbb{E}[\hat{S}_{ni}] = \sum_{t=1}^{n} x_{ti} \qquad \text{and also} \qquad \mathbb{E}_t[X_t] = \sum_{i=1}^{K} P_{ti} x_{ti} = \sum_{i=1}^{K} P_{ti} \mathbb{E}_t[\hat{X}_{ti}] \,.$$

The tower rule says that $\mathbb{E}[\mathbb{E}_t[X_t]] = \mathbb{E}[X_t]$, which together with the linearity of expectation and the above display means that

$$R_{ni} = \mathbb{E}\left[\hat{S}_{ni}\right] - \mathbb{E}\left[\sum_{t=1}^{n}\sum_{i=1}^{K} P_{ti}\hat{X}_{ti}\right] = \mathbb{E}\left[\hat{S}_{ni} - \hat{S}_n\right] , \tag{11.8}$$

where the last equality serves as the definition of $\hat{S}_n = \sum_{t,i} P_{ti}\hat{X}_{ti}$. To bound the right-hand side of Eq. (11.8) let

$$W_t = \sum_{j=1}^{K} \exp\left(\eta \hat{S}_{tj}\right) .$$

By convention an empty sum is zero, which means that $S_{0j} = 0$ and $W_0 = K$. Then

$$\exp(\eta \hat{S}_{ni}) \le \sum_{j=1}^{K} \exp(\eta \hat{S}_{nj}) = W_n = W_0 \frac{W_1}{W_0} \cdots \frac{W_n}{W_{n-1}} = K \prod_{t=1}^{n} \frac{W_t}{W_{t-1}} \,.$$

The ratio in the product can be rewritten in terms of P_t by

$$\frac{W_t}{W_{t-1}} = \sum_{j=1}^{K} \frac{\exp(\eta \hat{S}_{t-1,j})}{W_{t-1}} \exp(\eta \hat{X}_{tj}) = \sum_{j=1}^{K} P_{tj} \exp(\eta \hat{X}_{tj}) \,. \tag{11.9}$$

We need the following facts:

$$\exp(x) \le 1 + x + x^2 \text{ for all } x \le 1 \quad \text{and} \quad 1 + x \le \exp(x) \text{ for all } x \in \mathbb{R} \,.$$

Using these two inequalities leads to

$$\frac{W_t}{W_{t-1}} \le 1 + \eta \sum_{j=1}^{K} P_{tj}\hat{X}_{tj} + \eta^2 \sum_{j=1}^{K} P_{tj}\hat{X}_{tj}^2 \le \exp\left(\eta \sum_{j=1}^{K} P_{tj}\hat{X}_{tj} + \eta^2 \sum_{j=1}^{K} P_{tj}\hat{X}_{tj}^2\right) .$$

Notice that this was only possible because $\hat{X}_{tj}$ is defined by Eq. (11.5), which ensures that $\hat{X}_{tj} \le 1$ and would not have been true had we used Eq. (11.2). Putting the inequalities together we get

$$\exp\left(\eta \hat{S}_{ni}\right) \le K \exp\left(\eta \hat{S}_n + \eta^2 \sum_{t=1}^{n}\sum_{j=1}^{K} P_{tj}\hat{X}_{tj}^2\right) .$$

Taking the logarithm of both sides, dividing by $\eta > 0$ and reordering gives

$$\hat{S}_{ni} - \hat{S}_n \le \frac{\log(K)}{\eta} + \eta \sum_{t=1}^{n}\sum_{j=1}^{K} P_{tj}\hat{X}_{tj}^2 \,. \tag{11.10}$$

As noted earlier, the expectation of the left-hand side is R_{ni}. The first term on

the right-hand side is a constant, which leaves us to bound the expectation of the second term. Letting $y_{tj} = 1 - x_{tj}$ and $Y_t = 1 - X_t$, then expanding the definition of $\hat{X}_{tj}^2$ leads to

$$\begin{aligned}
\mathbb{E}\left[\sum_{t,j} P_{tj}\hat{X}_{tj}^2\right] &= \mathbb{E}\left[\sum_{t=1}^{n}\sum_{j=1}^{K} P_{tj}\left(1 - \frac{\mathbb{I}\{A_t = j\}\, y_{tj}}{P_{tj}}\right)^2\right] \\
&= \sum_{t=1}^{n}\mathbb{E}\left[\sum_{j=1}^{K} P_{tj}\left(1 - 2\,\frac{\mathbb{I}\{A_t = j\}\, y_{tj}}{P_{tj}} + \frac{\mathbb{I}\{A_t = j\}\, y_{tj}^2}{P_{tj}^2}\right)\right] \\
&= \sum_{t=1}^{n}\mathbb{E}\left[1 - 2Y_t + \mathbb{E}_t\left[\sum_{j=1}^{K}\frac{\mathbb{I}\{A_t = j\}\, y_{tj}^2}{P_{tj}}\right]\right] \\
&= \sum_{t=1}^{n}\mathbb{E}\left[1 - 2Y_t + \sum_{j=1}^{K} y_{tj}^2\right] \\
&= \sum_{t=1}^{n}\mathbb{E}\left[(1 - Y_t)^2 + \sum_{j\neq A_t} y_{tj}^2\right] \\
&\leq nK\,.
\end{aligned}$$

By substituting this into Eq. (11.10), we get

$$R_{ni} \leq \frac{\log(K)}{\eta} + \eta nK = 2\sqrt{nK\log(K)}\,,$$

where the equality follows by substituting $\eta = \sqrt{\log(K)/(nK)}$, which was chosen to optimize this bound. □

At the heart of the proof are the inequalities:

$$1 + x \leq \exp(x) \text{ for all } x \in \mathbb{R} \qquad \text{and} \qquad \exp(x) \leq 1 + x + x^2 \text{ for } x \leq 1\,.$$

Attentive readers will notice that the former of these inequalities is an ansatz derived from the first order Taylor expansion of $\exp(x)$ about $x = 0$. The latter, however, is not the second order Taylor expansion, which would be $1 + x + x^2/2$. The problem is that the second order Taylor series is not an upper bound on $\exp(x)$ for $x \leq 1$, but only for $x \leq 0$:

$$\exp(x) \leq 1 + x + \frac{1}{2}x^2 \text{ for all } x \leq 0\,. \tag{11.11}$$

But it is nearly an upper bound, and this can be exploited to improve the bound in Theorem 11.1. The mentioned upper and lower bounds on $\exp(x)$ are shown in Fig. 11.1, from which it is quite obvious that the new bound is significantly tighter when $x \leq 0$.

Let us now put Eq. (11.11) to use in proving the following improved version of Theorem 11.1 for which the regret is smaller by a factor of $\sqrt{2}$. This looks quite insignificant, but in relative terms shaves off approximately thirty percent

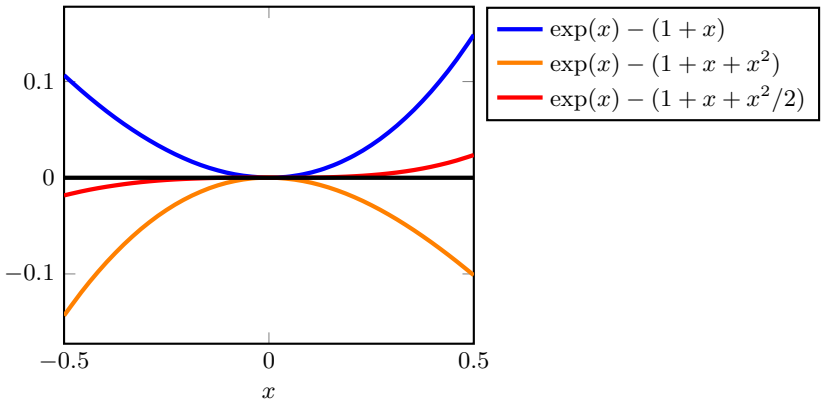

Figure 11.1 Approximations for $\exp(x)$ on $[-1/2, 1/2]$.

of the previous bound. The algorithm is unchanged except for a slightly increased learning rate.

THEOREM 11.2 *Let $\nu = (x_{ti}) \in [0,1]^{nK}$ be an adversarial bandit and π be the policy of Exp3 with learning rate $\eta = \sqrt{2\log(K)/(nK)}$. Then*

$$R_n(\pi, \nu) \le \sqrt{2nK \log(K)}\,.$$

Proof By construction $\hat{X}_{tj} \le 1$. Therefore

$$\begin{aligned}\exp\left(\eta \hat{X}_{tj}\right) &= \exp(\eta)\exp\left(\eta(\hat{X}_{tj} - 1)\right) \\ &\le \exp(\eta)\left\{1 + \eta(\hat{X}_{tj} - 1) + \frac{\eta^2}{2}(\hat{X}_{tj} - 1)^2\right\}.\end{aligned}$$

Using the fact that $\sum_j P_{tj} = 1$ and the inequality $1 + x \le \exp(x)$ we get

$$\frac{W_t}{W_{t-1}} = \sum_{j=1}^{K} P_{tj}\exp(\eta \hat{X}_{tj}) \le \exp\left(\eta \sum_{j=1}^{K} P_{tj}\hat{X}_{tj} + \frac{\eta^2}{2}\sum_{j=1}^{K} P_{tj}(\hat{X}_{tj} - 1)^2\right),$$

where the equality is from Eq. (11.9). We see that here we need to bound $\sum_j P_{tj}(\hat{X}_{tj} - 1)^2$. Let $\hat{Y}_{tj} = 1 - \hat{X}_{tj}$. Then

$$P_{tj}(\hat{X}_{tj} - 1)^2 = P_{tj}\hat{Y}_{tj}\hat{Y}_{tj} = \mathbb{I}\{A_t = j\}\, y_{tj}\hat{Y}_{tj} \le \hat{Y}_{tj}\,,$$

where the last inequality used $\hat{Y}_{tj} \ge 0$ and $y_{tj} \le 1$. Thus,

$$\sum_{j=1}^{K} P_{tj}(\hat{X}_{tj} - 1)^2 \le \sum_{j=1}^{K} \hat{Y}_{tj}\,.$$

With the same calculations as before, we get

$$\hat{S}_{ni} - \hat{S}_n \le \frac{\log(K)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{n}\sum_{j=1}^{K} \hat{Y}_{tj}\,. \tag{11.12}$$

The result is completed by taking expectations of both sides, using $\mathbb{E}\sum_{t,j}\hat{Y}_{tj} = \mathbb{E}\sum_{t,j}\mathbb{E}_t\hat{Y}_{tj} = \mathbb{E}\sum_{t,j}y_{tj} \le nK$, and then substituting the learning rate. □

11.4 Notes

1 The expected regret of Exp3 cannot be improved significantly, but the distribution of its regret is poorly behaved. Define the **random regret** to be the random variable measuring the actual deficit of the learner relative to the best arm in hindsight:

$$\hat{R}_n = \underbrace{\max_{i\in[K]}\sum_{t=1}^{n}x_{ti} - \sum_{t=1}^{n}X_t}_{\text{in terms of rewards}} = \underbrace{\sum_{t=1}^{n}Y_t - \min_{i\in[K]}\sum_{t=1}^{n}y_{ti}}_{\text{in terms of losses}},$$

In Exercise 11.5 you will show that for all large enough n and reasonable choices of η there exists a bandit such that the random regret of Exp3 satisfies $\mathbb{P}(\hat{R}_n \ge n/4) > 1/131$. This is quite a troubling result and motivates the introduction of algorithms in the next chapter for which the distribution of $\hat{R}_n$ is well behaved.

2 What happens when the range of the rewards is unbounded? This has been studied by Allenberg et al. [2006], where some (necessarily much weaker) positive results are presented.

3 A more basic problem than the one considered here is when the learner receives all $(x_{ti})_i$ at the end of round t, but the reward is still x_{tA_t}. This setting is called the **full-information** setting or **prediction with expert advice**. Exponential weighting is still a good idea, but the estimated rewards can now be replaced by the actual rewards. The resulting algorithm is sometimes called Hedge or the Exponential Weights Algorithm (EWA). The proof as written goes through in almost the same way, but one should replace the polynomial upper bound on $\exp(x)$ with Hoeffding's lemma. This analysis gives a regret of $\sqrt{n\log(K)/2}$, which is optimal in an asymptotic sense [Cesa-Bianchi and Lugosi, 2006].

4 A more sophisticated algorithm and analysis shaves a factor of $\sqrt{\log(K)}$ from the regret upper bound [Audibert and Bubeck, 2009, 2010a, Bubeck and Cesa-Bianchi, 2012]. The algorithm is an instantiation of the mirror descent algorithm from convex optimization, which we present in Chapter 28 for the more general adversarial linear bandit problem. Exercise 28.10 in that chapter explains the steps needed to solve this problem.

5 The initial distribution (the 'prior') P_1 does not have to be uniform. By biasing the prior towards a specific action the regret can be reduced when the favored action turns out to be optimal. There is a price for this, however, if the optimal arm is not favored [Lattimore, 2015a].

6 It was assumed in this chapter that the environment chose the rewards

at the start of the game. Such environments are called **oblivious** because the choices of the environment do not depend on those of the learner. A **reactive environment** is one where x_t is allowed to depend on the history $a_1, x_1, \ldots, a_{t-1}, x_{t-1}$. Despite the fact that this is clearly a harder problem the result we obtained can be generalized to this setting without changes to the analysis. It is another question whether the definition of regret makes sense for such reactive environments.

7 Building on the previous note, suppose the reward vector in round t is $X_t = f_t(A_1, \ldots, A_t)$ and $f_1, \ldots, f_n$ are a sequence of functions chosen in advance by the adversary with $f_t : [K]^t \to [0, 1]$. Let $\Pi \subset [K]^n$ be a set of action-sequences. Then the expected **policy regret** with respect to Π is

$$\max_{a_1,\ldots,a_n \in \Pi} \sum_{t=1}^{n} f_t(a_1, \ldots, a_t) - \mathbb{E}\left[\sum_{t=1}^{n} f_t(A_1, \ldots, A_t)\right] .$$

Even if Π only consists of constant sequences, there still does not exist a policy guaranteeing sublinear regret. The reason is simple. Consider the two candidate choices of $f_1, \ldots, f_n$. In the first choice $f_t(a_1, \ldots, a_t) = \mathbb{I}\{a_1 = 1\}$ and in the second we have $f_t(a_1, \ldots, a_t) = \mathbb{I}\{a_1 = 2\}$. Clearly the learner must suffer linear regret in at least one of these two reactive bandit environments. The problem is that the learner's decision in the first round determines the rewards available in all subsequent rounds and there is no time for learning. By making additional assumptions sublinear regret is possible, however. For example, by assuming the adversary has limited memory [Arora et al., 2012].

8 There is a common misconception that the adversarial framework is a good fit for nonstationary environments. While the framework does not assume the rewards are stationary, the regret concept used in this chapter has stationarity built in. A policy designed for minimizing the regret relative to the best action in hindsight is seldom suitable for nonstationary bandits where the whole point is to adapt to changes in the optimal arm. In such cases a better benchmark is to compete with a sequence of actions. For more on nonstationary bandits see Chapter 31.

9 The estimators in Eq. (11.2) and Eq. (11.5) both have conditional variance $\mathbb{V}_t[\hat{X}_{ti}] \approx 1/P_{ti}$, which blows up for small P_{ti}. It is instructive to think about whether and how P_{ti} can take on very small values. Consider the loss-based estimator given by (11.5). For this estimator, when P_{tA_t} and X_t are both small, $\hat{X}_{tA_t}$ can take on a large negative value. Through the update formula (11.6) this then translates into P_{t+1,A_t} being squashed aggressively towards zero. A similar issue arises with the reward-based estimator given by (11.2). The difference is that now it will be a 'positive surprise' (P_{tA_t} small, X_t large) that pushes the probabilities towards zero. But note that in this case $P_{t+1,i}$ is pushed towards zero for all $i \neq A_t$. This means that dangerously small probabilities are expected to be more frequent for the gains estimator Eq. (11.2).

10 We argued at the beginning of the chapter that deterministic policies are

no good for adversarial bandit problems, which rules out all of the policies analyzed in Part II. We also showed the regret of Exp3 grows with at most the square root of the horizon on both stochastic and nonstochastic bandits. One might wonder if there exists a policy with (near-)optimal regret for adversarial bandits and logarithmic regret for stochastic bandits. There is a line of work addressing this question, which shows that such algorithm do exist [Bubeck and Slivkins, 2012, Seldin and Slivkins, 2014, Auer and Chiang, 2016, Seldin and Lugosi, 2017]. There are some complications, however, depending on whether or not the adversary is oblivious or not. The situation is best summarized by Auer and Chiang [2016], where the authors present upper and lower bounds on what is possible in various scenarios.

11 Exp3 requires advance knowledge of the horizon. The doubling trick can be used to overcome this issue, but perhaps a more elegant solution is to use a decreasing learning rate. The analysis in this chapter can be adapted to this case. More discussion is provided in the notes and exercises of Chapter 28 where we give a more generic solution to this problem.

12 There is a connection between adversarial learning and simultaneous-action zero sum games. This is discussed in a little more detail in the notes and exercises of Chapter 28.

11.5 Bibliographic remarks

Exponential weighting has been a standard tool in online learning since the papers by Vovk [1990] and Littlestone and Warmuth [1994]. Exp3 and several variations were introduced by Auer et al. [1995], which was also the first paper to study bandits in the adversarial framework. The algorithm and analysis presented here differs slightly because we do not add any additional exploration, while the version of Exp3 in that paper explores uniformly with low probability. The fact that additional exploration is not required was observed by Stoltz [2005].

11.6 Exercises

11.1 In order to implement Exp3 you need a way to sample from the exponential weights distribution. Many programming language provide a standard way to do this. For example in Python you can use the Numpy library and `numpy.random.multinomial`. In more basic languages, however, you only have access to a function `rand()` that returns a floating point number 'uniformly' distributed in $[0, 1]$. Describe an algorithm that takes as input a probability vector $p \in \mathcal{P}_{d-1}$ and uses a single call to `rand()` to return $X \in [d]$ with $\mathbb{P}(X = i) = p_i$.

Of course, on most computers `rand()` will return a pseudo-random number and since there are only finitely many floating point numbers the resulting distribution will not really be uniform on $[0, 1]$. Thinking about these issues is a worthy endeavour, and sometimes it really matters. For this exercise you may ignore these issues, however.

11.2 Show that for any deterministic policy π there exists an environment ν such that $R_n(\pi, \nu) \geq n(1 - 1/K)$. What does your result say about the policies designed in Part II?

11.3 Suppose we had defined the regret by

$$R_n^{\text{track}}(\pi, \nu) = \mathbb{E}\left[\sum_{t=1}^{n} \max_{i \in [K]} x_{ti} - \sum_{t=1}^{n} x_{tA_t}\right].$$

At first sight this definition seems like the right thing because it measures what you actually care about. Unfortunately, however, it gives the adversary too much power. Show that for any policy π (randomised or not) there exists a $\nu \in [0, 1]^{Kn}$ such that

$$R_n^{\text{track}}(\pi, \nu) \geq n\left(1 - \frac{1}{K}\right).$$

11.4 Let $P \in \mathcal{P}_{K-1}$ be a probability vector and suppose $\hat{X} : [K] \times \mathbb{R} \to \mathbb{R}$ is a function such that for all $x \in \mathbb{R}^K$,

$$\mathbb{E}[\hat{X}(i, x_i)] = \sum_{i=1}^{K} P_i \hat{X}(i, x_i) = x_1 .$$

Show there exists an $a \in \mathbb{R}$ such that $\hat{X}(i, x) = a + \dfrac{\mathbb{I}\{i = 1\} x_1 - a}{P_1}$.

11.5 In this exercise you will show that if $\eta \in [n^{-p}, 1]$ for some $p \in (0, 1)$, then for sufficiently large n there exist bandits on which Exp3 has a constant probability of suffering linear regret and hence the variance of its regret is $\Omega(n^2)$. Let $x \in [1/4, 1/2]$ be a constant to be tuned subsequently and define two-armed adversarial bandit in terms of its losses by

$$y_{t1} = \begin{cases} 0 & \text{if } t \leq n/2 \\ 1 & \text{otherwise} \end{cases} \quad \text{and} \quad y_{t2} = \begin{cases} x & \text{if } t \leq n/2 \\ 0 & \text{otherwise} . \end{cases}$$

We are interested in analyzing the algorithm that samples $A_t \sim P_t$ where $P_{ti} \propto \exp(-\eta \sum_{s=1}^{t-1} \hat{Y}_{si})$ with $\hat{Y}_{si} = y_{si} A_{si} / P_{si}$.

(a) Define sequence of real-valued functions $q_1, \ldots, q_n$ on domain $[1/4, 1/2]$ inductively by $q_0(x) = 1/2$ and

$$q_{s+1}(x) = \frac{q_s(x) \exp(-\eta x / q_s(x))}{1 - q_s(x) + q_s(x) \exp(-\eta x / q_s(x))} .$$

Show for $t \le 1 + n/2$ that $P_{t2} = q_{T_2(t-1)}(x)$.

(b) Show that q_s is continuous on its domain and that $\frac{d}{dx} q_s(x) \le 0$ for all $s \ge 0$.

(c) Let $s = \min\{u : q_u(1/2) < 1/(8n)\}$. Show that for $s \ge 4$ there exists an $x \in [1/4, 1/2]$ such that $q_s(x) = 1/(8n)$.

(d) Let s and x be as in the previous part. Show for large enough n it holds that $\sum_{u=1}^{s-1} 1/q_u(x) \le n/8$.

(e) Let $N(t)$ be a discrete counting process with $N(1) = 0$ and $N(t+1) - N(t) \in \{0, 1\}$ almost surely and $\mathbb{P}\left(N(t+1) - N(t) = 1 \mid N(t)\right) = q_{N(t)}(x)$. Prove that

$$\mathbb{P}\left(X\left(2\sum_{u=1}^{s} 1/q_u(x)\right) \ge s\right) \ge \frac{1}{2}\,.$$

(f) Prove that $\mathbb{P}\left(T_2(n/4) \ge s\right) \ge \frac{1}{2}$.

(g) Let E be the event that $\sum_{t=1}^{n/2} \hat{Y}_{t2} \ge 2n$. Prove that

$$\mathbb{P}\left(\sum_{t=n/2+1}^{n} \mathbb{I}\left\{A_t = 1\right\} \ge n/2 \;\middle|\; E\right) \ge 1 - n\exp(-\eta n)\,.$$

(h) Prove that $\mathbb{P}\left(E\right) \ge \frac{1}{2}(1 - \exp(-1/32))$.

(i) Prove that $\mathbb{P}(\hat{R}_n \ge \frac{n}{4}) \ge \frac{1}{2}(1 - \exp(-1/32))(1 - n\exp(-\eta n))$.

(j) You have shown that for large enough n, $\mathbb{P}(\hat{R}_n \ge cn) \ge c$ for some universal constant c. Explain why does this not contradict the proof that $R_n = \mathbb{E}[\hat{R}_n] = O(\sqrt{n})$.

(k) Let $n = 10^5$ and $\eta = \sqrt{2\log(2)/(2n)}$. Find the value of x satisfying the conditions in Part (c) and simulate Exp3 to demonstrate linear regret with constant probability.

11.6 Show that Theorem 11.1 stays valid for an adversarially stopped Exp3. That is, imagine that an adversary is given the power to stop Exp3 at some random time $\tau \in [n]$. The adversary is restricted in this decision in that while it can use $A_1, \ldots, A_t$ when deciding about whether Exp3 should be stopped in t, it cannot use $A_{t+1}, \ldots, A_n$. That is, $\{\tau = t\}$ must be $\mathcal{F}_t = \sigma(A_1, \ldots, A_t)$-measurable. Show that $\mathbb{E}\left[\hat{R}_\tau\right] \le 2\sqrt{nK\log(K)}$, where $\hat{R}_n = \sum_{t=1}^{n} x_{ti} - \sum_{t=1}^{n} X_t$ is the random regret of Exp3.

Use the identity $\sum_{t=1}^{\tau} U_t = \sum_{t=1}^{n} \mathbb{I}\left\{t \le \tau\right\} U_t$, the tower rule and argue that $\{t \le \tau\}$ is $\mathcal{F}_{t-1}$-measurable.

11.7 Let $a_1, \ldots, a_K$ be positive real values and $U_1, \ldots, U_K$ be a sequence of independent and identically distributed uniform random variables. Then let $G_i = -\log(-\log(U_i))$, which follows a **standard Gumbel distribution**. Prove

that

$$\mathbb{P}\left(\log(a_i) + G_i = \max_{k\in[K]}(\log(a_k) + G_k)\right) = \frac{a_i}{\sum_{k=1}^{K} a_k}\,.$$

Let $(Z_{ti})_{ti}$ be a collection of independent and identically distributed random variables. The follow the perturbed leader algorithm chooses

$$A_t = \operatorname{argmax}_{i\in[K]}\left(Z_{ti} - \eta\sum_{s=1}^{t-1}\hat{\ell}_{ti}\right)\,.$$

The previous exercise shows that choosing the distribution of Z_{ti} to be a standard Gumbel distribution makes follow the perturbed leader the same as exponential weights. This viewpoint will prove extremely useful when we tackle combinatorial bandits in Chapter 30.

11.8 In this exercise we compare UCB and Exp3 on stochastic data. Suppose we have a two-armed stochastic Bernoulli bandit with $\mu_1 = 0.5$ and $\mu_2 = \mu_1 + \Delta$ with $\Delta = 0.05$.

(a) Plot the regret of UCB and Exp3 on the same plot as a function of the horizon n using the learning rate from Theorem 11.2.
(b) Now fix the horizon to $n = 10^5$ and plot the regret as a function of the learning rate. Your plot should look like Fig. 11.2.
(c) Investigate how the shape of this graph changes as you change Δ.
(d) Find empirically the choice of η that minimizes the worst-case regret over all reasonable choices of Δ and compare to the value proposed by the theory.
(e) What can you conclude from all this? Tell an interesting story.

The performance of UCB depends greatly on which version you use. For best results remember that Bernoulli distributions are 1/2-subgaussian or use the KL-UCB algorithm from Chapter 10.

11.9 Stress test your implementation of Exp3 from the previous exercise. What happens when $K = 2$ and the sequence of rewards is $x_{t1} = \mathbb{I}\{t \le n/4\}$ and $x_{t2} = \mathbb{I}\{t > n/4\}$?

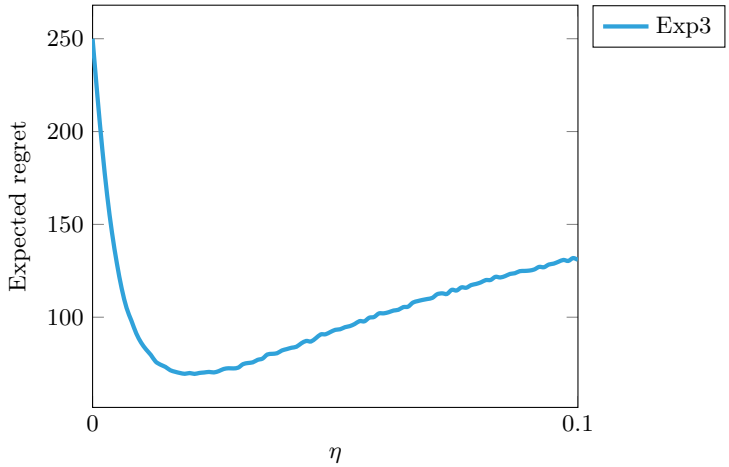

Figure 11.2 Expected regret for Exp3 for different learning rates over $n = 10^5$ rounds on a Bernoulli bandit with means $\mu_1 = 0.5$ and $\mu_2 = 0.55$.