12 The Exp3-IX Algorithm

In the last chapter we proved a sublinear bound on the expected regret of Exp3, but with a dishearteningly large variance. The objective of this chapter is to modify Exp3 so that the regret stays small in expectation and is simultaneously well concentrated about its mean. Such results are called **high probability bounds**.

One way to make Exp3 more robust is to make sure that P_{ti} is never too small. The first thing that comes to mind is to mix P_t with the uniform distribution. This is an explicit way of forcing exploration, which after further modification can be made to work. The resulting algorithm is called Exp3.P and we ask you to analyze it in Exercise 12.1. In this chapter we explore a similar idea that leads to an algorithm that is both simpler and empirically superior. The idea is to change the reward estimates to control the variance at the price of introducing some bias.

We start by summarizing what we know about the behaviour of the random regret of Exp3. Because we want to use the loss-based estimator it is more convenient to switch to losses, which we do for the remainder of the chapter. Rewriting Eq. (11.12) in terms of losses,

$$\hat{L}_n - \hat{L}_{ni} \leq \frac{\log(K)}{\eta} + \frac{\eta}{2} \sum_{j=1}^{K} \hat{L}_{nj}\,, \tag{12.1}$$

where $\hat{L}_n$ and $\hat{L}_{ni}$ are defined using the loss estimator $\hat{Y}_{tj}$ by

$$\hat{L}_n = \sum_{t=1}^{n} \sum_{j=1}^{K} P_{tj} \hat{Y}_{tj} \quad \text{and} \quad \hat{L}_{ni} = \sum_{t=1}^{n} \hat{Y}_{ti}\,.$$

⚠ Eq. (12.1) holds no matter how the loss estimators are chosen provided they satisfy $\hat{Y}_{ti} \geq 0$ for all t and i. Of course the left-hand side of Eq. (12.1) is not close to the regret unless $\hat{Y}_{ti}$ is a reasonable estimator of the loss y_{ti},

We also need to define the sum of losses observed by the learner and for each fixed action, which are

$$\tilde{L}_n = \sum_{t=1}^{n} y_{tA_t} \quad \text{and} \quad L_{ni} = \sum_{t=1}^{n} y_{ti}$$

Like in the previous chapter we need to define the (random) regret with respect to a given arm i as follows:

$$\hat{R}_{ni} = \sum_{t=1}^{n} x_{ti} - \sum_{t=1}^{n} X_t = \tilde{L}_n - L_{ni}\,. \tag{12.2}$$

By subsituting the above definitions into Eq. (12.1) and rearranging the regret with respect to any arm i is bounded by

$$\begin{aligned}\hat{R}_{ni} &= \tilde{L}_n - L_{ni} = (\tilde{L}_n - \hat{L}_n) + (\hat{L}_n - \hat{L}_{ni}) + (\hat{L}_{ni} - L_{ni}) \\ &\leq \frac{\log(K)}{\eta} + (\tilde{L}_n - \hat{L}_n) + (\hat{L}_{ni} - L_{ni}) + \frac{\eta}{2}\sum_{j=1}^{K} \hat{L}_{nj}\,. \end{aligned} \tag{12.3}$$

This means the random regret can be bounded by controlling $\tilde{L}_n - \hat{L}_n$ and $\hat{L}_{nj} - L_{nj}$ and $\hat{L}_{nj}$. As promised we now modify the loss estimate. Let $\gamma > 0$ be a small constant to be chosen later and define the biased estimator

$$\hat{Y}_{ti} = \frac{\mathbb{I}\left\{A_t = i\right\} Y_t}{P_{ti} + \gamma}\,. \tag{12.4}$$

As γ increases the predictable variance decreases, but the bias increases. The optimal choice of γ depends on finding the sweet spot, which we will do once the dust has settled in the analysis. When Eq. (12.4) is used in the exponential update in Exp3, the resulting algorithm is called **Exp3-IX** (Algorithm 9). The suffix 'IX' stands for **i**mplicit e**x**ploration, a name justified by the following argument. A simple calculation shows that

$$\mathbb{E}_t[\hat{Y}_{ti}] = \frac{P_{ti} y_{ti}}{P_{ti} + \gamma} = y_{ti} - \frac{\gamma y_{ti}}{P_{ti} + \gamma} \leq y_{ti}\,.$$

Since small losses correspond to large rewards, the estimator is optimistically biased. The effect is a smoothing of P_t so that actions with large losses for which Exp3 would assign negligable probability are still chosen occasionally. As a result, Exp3-IX will explore more than the standard Exp3 algorithm (see Exercise 12.3). The reason for calling the exploration implicit is that it is a consequence of modifying the loss estimates, rather than directly altering P_t. This approach is more elegant mathematically and has nicer properties than the version that mixes P_t with the uniform distribution.

12.1 Regret analysis

We now prove the following theorem bounding the random regret of Exp3-IX with high probability.

THEOREM 12.1 *Let $\delta \in (0, 1)$ and define*

$$\eta_1 = \sqrt{\frac{2\log(K+1)}{nK}} \qquad \textit{and} \qquad \eta_2 = \sqrt{\frac{\log(K) + \log(\frac{K+1}{\delta})}{nK}}\,.$$

```
Input: $n$, $K$, $\eta$, $\gamma$
Set $\hat{L}_{0i} = 0$ for all $i$
for $t = 1, \dots, n$ do
    Calculate the sampling distribution $P_t$:
        $P_{ti} = \dfrac{\exp\left(-\eta \hat{L}_{t-1,i}\right)}{\sum_{j=1}^{K} \exp\left(-\eta \hat{L}_{t-1,j}\right)}$
    Sample $A_t \sim P_t$ and observe reward $X_t$
    Calculate $\hat{L}_{ti} = \hat{L}_{t-1,i} + \dfrac{\mathbb{I}\{A_t = i\}(1 - X_t)}{P_{t-1,i} + \gamma}$
end for
```

Algorithm 9: Exp3-IX

The following hold:

1 *If Exp3-IX is run with parameters $\eta = \eta_1$ and $\gamma = \eta/2$, then*

$$\mathbb{P}\left(\hat{R}_n \geq \sqrt{8.5nK\log(K+1)} + \left(\sqrt{\frac{nK}{2\log(K+1)}} + 1\right)\log(1/\delta)\right) \leq \delta\,. \tag{12.5}$$

2 *If Exp3-IX is run with parameters $\eta = \eta_2$ and $\gamma = \eta/2$, then*

$$\mathbb{P}\left(\hat{R}_n \geq 2\sqrt{(2\log(K+1) + \log(1/\delta))nK} + \log\left(\frac{K+1}{\delta}\right)\right) \leq \delta\,. \tag{12.6}$$

The value of η_1 is independent of δ, which means that using this choice of learning rate leads to a single algorithm with a high probability bound for all δ. On the other hand, η_2 does depend on δ so the user must choose a confidence level from the beginning. The advantage is that the bound is improved, but only for the specified confidence level. We will show in Chapter 17 that this tradeoff is unavoidable.

The proof follows by bounding each the terms in Eq. (12.3), which we do via a series of lemmas. The first of these lemmas is a new concentration bound, the statement of which requires us to introduce the notion of adapted and predictable sequences of random variables.

LEMMA 12.1 *Let $\mathbb{F} = (\mathcal{F}_t)_{0 \leq t \leq n}$ be a filtration and for $i \in [K]$ let $(\tilde{Y}_{ti})_t$ be $\mathbb{F}$-adapted such that:*

1 *For any $S \subset [K]$ with $|S| > 2$, $\mathbb{E}\left[\prod_{i \in S} \tilde{Y}_{ti} \middle| \mathcal{F}_{t-1}\right] \leq 0$.*

2 $\mathbb{E}\left[\tilde{Y}_{ti} \middle| \mathcal{F}_{t-1}\right] = y_{ti}$ *for all $t \in [n]$ and $i \in [K]$.*

Furthermore, let $(\alpha_{ti})_{ti}$ and $(\lambda_{ti})_{ti}$ be real-valued $\mathcal{F}_t$-predictable random sequences such that for all t, i it holds that $0 \le \alpha_{ti}\tilde{Y}_{ti} \le 2\lambda_{ti}$. Then for all $\delta \in (0,1)$,

$$\mathbb{P}\left(\sum_{t=1}^{n}\sum_{i=1}^{K}\alpha_{ti}\left(\frac{\tilde{Y}_{ti}}{1+\lambda_{ti}} - y_{ti}\right) \ge \log\left(\frac{1}{\delta}\right)\right) \le \delta\,.$$

The proof relies on Chernoff's method and is deferred until the end of the chapter. Equipped with this result we can easily bound the terms $\hat{L}_{ni} - L_{ni}$.

LEMMA 12.2 *Let $\delta \in (0,1)$. With probability at least $1-\delta$ the following inequalities hold simultaneously:*

$$\max_{i\in[K]}\left(\hat{L}_{ni} - L_{ni}\right) \le \frac{\log(\frac{K+1}{\delta})}{2\gamma} \qquad \textit{and} \qquad \sum_{i=1}^{K}\left(\hat{L}_{ni} - L_{ni}\right) \le \frac{\log(\frac{K+1}{\delta})}{2\gamma}\,. \tag{12.7}$$

Proof Fix $\delta' \in (0,1)$ to be chosen later. Then

$$\sum_{i=1}^{K}(\hat{L}_{ni} - L_{ni}) = \sum_{t,i}\left(\frac{A_{ti}y_{ti}}{P_{ti}+\gamma} - y_{ti}\right) = \frac{1}{2\gamma}\sum_{t,i}2\gamma\left(\frac{1}{1+\frac{\gamma}{P_{ti}}}\frac{A_{ti}y_{ti}}{P_{ti}} - y_{ti}\right)\,.$$

Introduce $\lambda_{ti} = \frac{\gamma}{P_{ti}}$, $\tilde{Y}_{ti} = \frac{A_{ti}y_{ti}}{P_{ti}}$ and $\alpha_{ti} = 2\gamma$. It is not hard to see then that the conditions of Lemma 12.1 are satisfied. In particular, for any $S \subset [K]$, $|S| > 1$, $\prod_{i\in S} A_{ti} = 0$, implying $\prod_{i\in S}\tilde{Y}_{ti} = 0$. Therefore

$$\mathbb{P}\left(\sum_{i=1}^{K}(\hat{L}_{ni} - L_{ni}) \ge \frac{\log(1/\delta')}{2\gamma}\right) \le \delta'\,. \tag{12.8}$$

Similarly, for any fixed i,

$$\mathbb{P}\left(\hat{L}_{ni} - L_{ni} \ge \frac{\log(1/\delta')}{2\gamma}\right) \le \delta'\,. \tag{12.9}$$

To see this use the previous argument with $\alpha_{tj} = \mathbb{I}\{j = i\}\,2\gamma$. The result follows by choosing $\delta' = \delta/(K+1)$ and the union bound. □

LEMMA 12.3 $\tilde{L}_n - \hat{L}_n = \gamma\sum_{j=1}^{K}\hat{L}_{nj}$.

Proof Let $A_{ti} = \mathbb{I}\{A_t = i\}$ as before. Writing $Y_t = \sum_j A_{tj}y_{tj}$, we calculate

$$Y_t - \sum_{j=1}^{K}P_{tj}\hat{Y}_{tj} = \sum_{j=1}^{K}\left(1 - \frac{P_{tj}}{P_{tj}+\gamma}\right)A_{tj}y_{tj} = \gamma\sum_{j=1}^{K}\frac{A_{tj}}{P_{tj}+\gamma}y_{tj} = \gamma\sum_{j=1}^{K}\hat{Y}_{tj}\,.$$

Therefore $\tilde{L}_n - \hat{L}_n = \gamma\sum_{j=1}^{K}\hat{L}_{nj}$ as required. □

Proof of Theorem 12.1 By Eq. (12.3) and Lemma 12.3 we have

$$\begin{aligned}\hat{R}_n &\leq \frac{\log(K)}{\eta} + (\tilde{L}_n - \hat{L}_n) + \max_{i\in[K]}(\hat{L}_{ni} - L_{ni}) + \frac{\eta}{2}\sum_{j=1}^K \hat{L}_{nj} \\ &= \frac{\log(K)}{\eta} + \max_{i\in[K]}(\hat{L}_{ni} - L_{ni}) + \left(\frac{\eta}{2} + \gamma\right)\sum_{j=1}^K \hat{L}_{nj}\,.\end{aligned}$$

Therefore by Lemma 12.2, with probability at least $1-\delta$ it holds that

$$\begin{aligned}\hat{R}_n &\leq \frac{\log(K)}{\eta} + \frac{\log\left(\frac{K+1}{\delta}\right)}{2\gamma} + \left(\gamma + \frac{\eta}{2}\right)\left(\sum_{j=1}^K L_{nj} + \frac{\log\left(\frac{K+1}{\delta}\right)}{2\gamma}\right) \\ &\leq \frac{\log(K)}{\eta} + \left(\gamma + \frac{\eta}{2}\right)nK + \left(\gamma + \frac{\eta}{2} + 1\right)\log\left(\frac{K+1}{\delta}\right)\,,\end{aligned}$$

where the second inequality follows since $L_{nj} \leq n$ for all j. The result follows by substituting the definitions of $\eta \in \{\eta_1, \eta_2\}$ and $\gamma = \eta/2$. □

12.1.1 Proof of Lemma 12.1

We start with a technical inequality.

LEMMA 12.4 *For any $0 \leq x \leq 2\lambda$ it holds that* $\exp\left(\frac{x}{1+\lambda}\right) \leq 1 + x\,.$

Note that $1+x \leq \exp(x)$. What the lemma shows is that by slightly discounting the argument of the exponential function, in a bounded neighborhood of zero, $1+x$ can be an upper bound for the resulting function. Or, equivalently, slightly inflating the linear term in $1+x$, the linear lower bound becomes an upper bound.

Proof of Lemma 12.4 We rely on algebraic inequalities. The first is $\frac{2u}{1+u} \leq \log(1+2u)$ which holds for $u \geq 0$. The second states that $x\log(1+y) \leq \log(1+xy)$, which holds for any $x \in [0,1]$ and $y > -1$. Thanks to these inequalities,

$$\frac{x}{1+\lambda} = \frac{x}{2\lambda}\,\frac{2\lambda}{1+\lambda} \leq \frac{x}{2\lambda}\log(1+2\lambda) \leq \log\left(1 + 2\lambda\frac{x}{2\lambda}\right) = \log(1+x)\,.$$

And the proof is completed by exponentiating both sides. □

Proof of Lemma 12.1 Fix $t \in [n]$ and let $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_t]$ denote the conditional expectation with respect to $\mathcal{F}_t$. By Lemma 12.4 and the assumption that $0 \leq \alpha_{ti}\tilde{Y}_{ti} \leq 2\lambda_{ti}$ we have

$$\exp\left(\frac{\alpha_{ti}\tilde{Y}_{ti}}{1+\lambda_{ti}}\right) \leq (1 + \alpha_{ti}\tilde{Y}_{ti})\,.$$

Taking the product of these inequalities over i,

$$\begin{aligned}\mathbb{E}_{t-1}\left[\exp\left(\sum_{i=1}^K \frac{\alpha_{ti}\tilde{Y}_{ti}}{1+\lambda_{ti}}\right)\right] &\le \mathbb{E}_{t-1}\left[\prod_{i=1}^K (1+\alpha_{ti}\tilde{Y}_{ti})\right] \le 1 + \mathbb{E}_{t-1}\left[\sum_{i=1}^K \alpha_{ti}\tilde{Y}_{ti}\right] \\ &= 1 + \sum_{i=1}^K \alpha_{ti} y_{ti} \le \exp\left(\sum_{i=1}^K \alpha_{ti} y_{ti}\right)\,, \qquad (12.10)\end{aligned}$$

where the second inequality follows from the assumption that for $S \subset [K]$ with $|S| > 1$, $\mathbb{E}_{t-1}\prod_{i\in S}\tilde{Y}_{ti} \le 0$, the third one follows from the assumption that $\mathbb{E}_{t-1}\tilde{Y}_{ti} = y_{ti}$, while the last one follows from $1 + x \le \exp(x)$. Define

$$Z_t = \exp\left(\sum_i \alpha_{ti}\left(\frac{\tilde{Y}_{ti}}{1+\lambda_{ti}} - y_{ti}\right)\right)$$

and let $M_t = Z_1 \dots Z_t$, $t \in [n]$ with $M_0 = 1$. By (12.10), $\mathbb{E}_{t-1}[Z_t] \le 1$. Therefore

$$\mathbb{E}[M_t] = \mathbb{E}[\mathbb{E}_{t-1}[M_t]] = \mathbb{E}[M_{t-1}\mathbb{E}_{t-1}[Z_t]] \le \mathbb{E}[M_{t-1}] \le \dots \le \mathbb{E}[M_0] = 1\,.$$

Setting $t = n$ and combining the above display with Markov's inequality leads to $\mathbb{P}\left(\log(M_n) \ge \log(1/\delta)\right) = \mathbb{P}\left(M_n\delta \ge 1\right) \le \mathbb{E}\left[M_n\right]\delta \le \delta$. □

12.2 Notes

1 An upper bound on the expected regret of Exp3-IX can be obtained by integrating the tail.

$$R_n \le \mathbb{E}[(\hat{R}_n)^+] = \int_0^\infty \mathbb{P}\left((\hat{R}_n)^+ \ge x\right) dx \le \int_0^\infty \mathbb{P}\left(\hat{R}_n \ge x\right) dx\,,$$

where the first equality follows from Proposition 2.3. The result is completed using either high probability bound in Theorem 12.1 and by straightforward integration. We leave the details to the reader in Exercise 12.4.

2 The analysis presented here uses a fixed learning rate that depends on the horizon. Replacing η and γ with $\eta_t = \sqrt{\log(K)/(Kt)}$ and $\gamma_t = \eta_t/2$ leads to an anytime algorithm with about the same regret [Kocák et al., 2014, Neu, 2015a].

3 There is another advantage of the modified importance-weighted estimators used by Exp3-IX, which leads to an improved regret in the special case that one of the arms has small losses. Specifically, it is possible to show that

$$R_n = O\left(\sqrt{K \min_{i\in[K]} L_{in} \log(K)}\right)\,.$$

In the worst case L_{in} is linear in n and the usual bound is recovered. But if the optimal arm enjoys low cumulative regret, then the above can be a big improvement over the bounds given in Theorem 12.1. Bounds of this kind are called **first order bounds**. We refer the interested reader to the papers by Allenberg et al. [2006], Abernethy et al. [2012], Neu [2015b].

4 Another situation where one might hope to have a smaller regret is when the rewards/losses for each arm do not deviate too far from their averages. Define the **quadratic variation** by

$$Q_n = \sum_{t=1}^n \sqrt{\sum_{i=1}^K (x_{ti} - \mu_i)^2}\,, \quad \text{where } \mu_i = \frac{1}{n}\sum_{t=1}^n x_{ti}\,.$$

Hazan and Kale [2011] gave an algorithm for which $R_n = O(K^2\sqrt{Q_t})$, which can be better than the worst case bound of Exp3 or Exp3-IX when the quadratic variation is very small. The factor of K^2 is suboptimal and can be removed using a careful instantiation of the mirror descent algorithm [Bubeck et al., 2018]. We do not cover this exact algorithm in this book, but the techniques based on mirror descent are presented in Chapter 28.

5 An alternative to the algorithm presented here is to mix the probability distribution computed using exponential weights with the uniform distribution, while biasing the estimates. This leads to the Exp3.P algorithm due to Auer et al. [2002b] who considered the case where δ is given and derived a bound that is similar to Eq. (12.6) of Theorem 12.1. With an appropriate modification of their proof it is possible to derive a weaker bound similar to Eq. (12.5) where the knowledge of δ is not needed by the algorithm. This has been explored by Beygelzimer et al. [2010] in the context of a related algorithm, which will be considered in Chapter 18. One advantage of this approach is that it generalizes to the case where the loss estimators are sometimes negative, a situation that can arise in more complicated settings. For technical details we advise the reader to work through Exercise 12.1.

12.3 Bibliographic remarks

The Exp3-IX algorithm is due to Kocák et al. [2014], who also introduced the biased loss estimators. The focus of that paper was to improve algorithms for more complex models with potentially large action-sets and side information, though their analysis can still be applied to the model studied in this chapter. The observation that this algorithm also leads to high probability bounds appeared in a followup paper by Neu [2015a]. High probability bounds for adversarial bandits were first provided by Auer et al. [2002b] and explored in a more generic way by Abernethy and Rakhlin [2009]. The idea to reduce the variance of importance-weighted estimators is not new and seems to have been applied in various forms [Uchibe and Doya, 2004, Wawrzynski and Pacut, 2007, Ionides, 2008, Bottou et al., 2013]. All of these papers are based on truncating the estimators, which makes the resulting estimator less smooth. Surprisingly, the variance reduction technique used in this chapter seems to be recent [Kocák et al., 2014].

12.4 Exercises

12.1 In this exercise we ask you to analyze the Exp3.P algorithm, which as we mentioned in the notes is another way to obtain high probability bounds. The idea is to modify Exp3 by biasing the estimators and introducing some forced exploration. Let $\hat{Y}_{ti} = A_{ti}y_{ti}/P_{ti} - \eta/P_{ti}$ be a biased version of the loss-based importance-weighted estimator that was used in the previous chapter. Define $\hat{L}_{ti} = \sum_{s=1}^t \hat{Y}_{si}$ and consider the policy that samples $A_t \sim P_t$ where

$$P_{ti} = (1-\gamma)\tilde{P}_{ti} + \frac{\gamma}{K} \qquad \text{with} \qquad \tilde{P}_{ti} = \frac{\exp\left(-\eta \hat{L}_{t-1,i}\right)}{\sum_{j=1}^K \exp\left(-\eta \hat{L}_{t-1,j}\right)}\,.$$

(a) Let $\delta \in (0,1)$ and $i \in [K]$. Show that with probability $1-\delta$, the random regret $\hat{R}_{ni}$ against i (cf. (12.2)) satisfies

$$\hat{R}_{ni} < n\gamma + (1-\gamma)\sum_{t=1}^n \sum_{j=1}^K \tilde{P}_{tj}(\hat{Y}_{tj} - y_{ti}) + \sum_{t=1}^n \frac{\beta}{P_{tA_t}} + \sqrt{\frac{n\log(1/\delta)}{2}}\,.$$

(b) Show that

$$\sum_{t=1}^n \sum_{j=1}^K \tilde{P}_{tj}(\hat{Y}_{tj} - y_{ti}) = \sum_{t=1}^n \sum_{j=1}^K \tilde{P}_{tj}(\hat{Y}_{tj} - \hat{Y}_{ti}) + \sum_{t=1}^n (\hat{Y}_{ti} - y_{ti})\,.$$

(c) Show that

$$\sum_{t=1}^n \sum_{j=1}^K \tilde{P}_{tj}(\hat{Y}_{tj} - \hat{Y}_{ti}) \le \frac{\log(K)}{\eta} + \eta \sum_{t=1}^n \sum_{j=1}^K \tilde{P}_{tj}\hat{Y}_{tj}^2\,.$$

(d) Show that

$$\sum_{t=1}^n \sum_{j=1}^K \tilde{P}_{tj}\hat{Y}_{tj}^2 \le \frac{nK^2\beta^2}{\gamma} + \sum_{t=1}^n \frac{1}{P_{tA_t}}\,.$$

(e) Apply the result of Exercise 5.17 to show that for any $\delta \in (0,1)$, the following hold:

$$\mathbb{P}\left(\sum_{t=1}^n \frac{1}{P_{tA_t}} \ge 2nK + \frac{K}{\gamma}\log\left(\frac{1}{\delta}\right)\right) \le \delta\,.$$

$$\mathbb{P}\left(\sum_{t=1}^n \hat{Y}_{ti} - y_{ti} \ge \frac{1}{\beta}\log\left(\frac{1}{\delta}\right)\right) \le \delta\,.$$

(f) Combining the previous steps, show that there exists a universal constant $C > 0$ such that for any $\delta \in (0,1)$, for an appropriate choice of η, γ and β, with probability at least $1-\delta$ it holds that the random regret $\hat{R}_n$ of Exp3.P satisfies

$$\hat{R}_n \le C\sqrt{nK\log(K/\delta)}$$

(g) In which step did you use the modified estimators?

(h) Show a bound where the algorithm parameters η, γ, β can only depend on n, K, but not on δ.

(i) Compare the bounds with the analogous bounds for Exp3-IX in Theorem 12.1.

12.2 This exercise is concerned with a generalization of the core idea underlying Exp3.P of the previous exercise in that rather than giving explicit expressions for the biased loss estimates, we focus on the key properties of these that makes Exp3.P "tick". To reduce clutter we assume for the remainder that t ranges in $[n]$ and $a \in [K]$. Let $(\Omega, \mathcal{F}, \mathcal{G} \doteq (\mathcal{G}_t)_{t=0}^n, \mathbb{P})$ be a filtered probability space. Let (Z_t), $(\hat{Z}_t)$, $(\tilde{Z}_t)$, (β_t) be sequences of random elements in $\mathbb{R}^K$, where $\tilde{Z}_t = \hat{Z}_t - \beta_t$ and $(Z_t), (\beta_t)$ are $\mathcal{G}$-predictable, whereas $(\hat{Z}_t)$ and therefore also $(\tilde{Z}_t)$ are $\mathcal{G}$-adapted (think of $\hat{Z}_t$ as the estimate of Z_t that uses randomization, and β_t is the bias as in the previous exercise). Given positive constant η define the probability vector $P_t \in \mathcal{P}_{K-1}$ by

$$P_{ta} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \tilde{Z}_{sa}\right)}{\sum_{b=1}^K \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{Z}_{sb}\right)} .$$

Let $\mathbb{E}_{t-1}[\cdot] = \mathbb{E}[\cdot|\mathcal{G}_{t-1}]$. Assume the following hold for all $a \in [K]$:

(a) $\eta|\hat{Z}_{ta}| \le 1$, *(b)* $\eta\beta_{ta} \le 1$,

(c) $\eta\mathbb{E}_{t-1}[\hat{Z}_{ta}^2] \le \beta_{ta}$ almost surely, *(d)* $\mathbb{E}_{t-1}[\hat{Z}_{ta}] = Z_{ta}$ almost surely.

Let $A^* = \operatorname{argmin}_{a\in[K]} \sum_{t=1}^n Z_{ta}$ and $R_n = \sum_{t=1}^n \sum_{a=1}^K P_{ta}(Z_{ta} - Z_{tA^*})$.

(a) Show that

$$\begin{aligned}
&\sum_{t=1}^n \sum_{a=1}^K P_{ta}(Z_{ta} - Z_{tA^*}) \\
&\quad = \underbrace{\sum_{t=1}^n \sum_{a=1}^K P_{ta}(\tilde{Z}_{ta} - \tilde{Z}_{tA^*})}_{\text{(A)}} + \underbrace{\sum_{t=1}^n \sum_{a=1}^K P_{ta}(Z_{ta} - \tilde{Z}_{ta})}_{\text{(B)}} + \underbrace{\sum_{t=1}^n (\tilde{Z}_{tA^*} - Z_{tA^*})}_{\text{(C)}} .
\end{aligned}$$

(b) Show that

$$\text{(A)} \le \frac{\log(K)}{\eta} + \eta \sum_{t=1}^n \sum_{a=1}^K P_{ta}\hat{Z}_{ta}^2 + 3\sum_{t=1}^n \sum_{a=1}^K P_{ta}\beta_{ta} .$$

(c) Show that with probability at least $1-\delta$,

$$\text{(B)} \le 2\sum_{t=1}^n \sum_{a=1}^K P_{ta}\beta_{ta} + \frac{\log(1/\delta)}{\eta} .$$

(d) Show that with probability at least $1 - K\delta$,

$$(\text{C}) \le \frac{\log(1/\delta)}{\eta}\,.$$

(e) Conclude that for any $\delta \le 1/(K+1)$, with probability at least $1-(K+1)\delta$,

$$R_n \le \frac{3\log(1/\delta)}{\eta} + \eta \sum_{t=1}^{n}\sum_{a=1}^{K} P_{ta}\hat{Z}_{ta}^2 + 5\sum_{t=1}^{n}\sum_{a=1}^{K} P_{ta}\beta_{ta}\,.$$

This is a long and challenging exercise. You may find it helpful to use the result in Exercise 5.17. The solution is also available.

12.3 Consider the Bernoulli bandit with $K = 5$ arms and $n = 10^4$ with means $\mu_1 = 1/2$ and $\mu_i = 1/2 - \Delta$ for $i > 1$. Plot the regret of Exp3 and Exp3-IX for $\Delta \in [0, 1/2]$. You should get something similar to the graph in Fig. 12.1. Does the result surprise you? Repeat the experiment in Part (k) of Exercise 11.5 with Exp3-IX and convince yourself that this algorithm is more robust than Exp3.

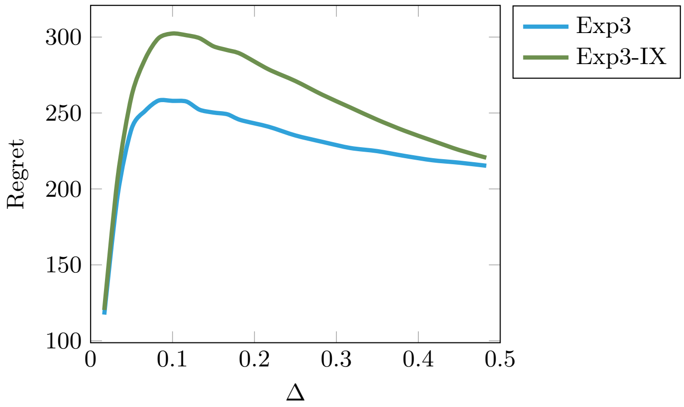

Figure 12.1 Comparison between Exp3 and Exp3-IX on Bernoulli bandit

12.4 In this exercise you will complete the steps explained in Note 1 to prove a bound on the expected regret of Exp3-IX.

(a) Find a choice of η and universal constant $C > 0$ such that

$$R_n \le C\sqrt{Kn\log(K)}\,.$$

(b) What happens as η grows? Write a bound on the expected regret of Exp3-IX in terms of η and K and n.