10 The Upper Confidence Bound Algorithm: Bernoulli Noise (†)

In previous chapters we assumed the noise of the rewards was σ-subgaussian for some known $\sigma > 0$. This has the advantage of simplicity and relative generality, but stronger assumptions are sometimes justified and often lead to stronger results. In this chapter we consider the case where the rewards are Bernoulli, which just means that $X_t \in \{0, 1\}$. This is a fundamental setting found in many applications. For example, in click-through prediction the user either clicks on the link or not. A Bernoulli bandit is characterized by the mean payoff vector $\mu \in [0, 1]^K$ and the reward observed in round t is $X_t \sim \mathcal{B}(\mu_{A_t})$.

We saw in Chapter 5 that the Bernoulli distribution is 1/2-subgaussian regardless of its mean, which means that UCB and its variants would enjoy logarithmic regret guarantees. However, the additional knowledge that the rewards are Bernoulli is not being fully exploited by these algorithms. The reason is essentially that the variance of a Bernoulli random variable depends on its mean, and when the variance is small the empirical mean concentrates faster, a fact that should be used to make the confidence intervals smaller.

10.1 Concentration for sums of Bernoulli random variables

Again we divert our attention away from bandits towards the concentration of the empirical mean towards the true value for sums of Bernoulli random variables. First we need to define a concept from information theory called the **relative entropy** or **Kullback-Leibler divergence**, which is a measure of similarity between distributions that for now we specify to the Bernoulli case. We defer the intuition for this concept until Chapter 14 where we give an introduction to information theory and specifically relative entropy.

DEFINITION 10.1 (Relative entropy between Bernoulli distributions) Let $p, q \in [0, 1]$. Then the relative entropy between Bernoulli distributions with parameters p and q respectively is defined to be

$$d(p, q) = p \log(p/q) + (1 - p) \log((1 - p)/(1 - q)) \,,$$

where the singularities are defined by taking limits so for $q \in [0, 1]$, $d(0, q) = \log(1/(1 - q))$, $d(1, q) = \log(1/q)$, and $d(p, 0) = d(p, 1) = \infty$ for $p \in (0, 1)$.

Notice that $d(p,q) = 0$ if and only if $p = q$ and $d(p,q) \geq 0$ for all p and q. So $d(\cdot,\cdot)$ is almost a metric except it is not symmetric and does not satisfy the triangle inequality. Some authors call such functions **premetrics**, but the nomenclature has not been standardized. The following lemma gives some useful properties of the relative entropy.

LEMMA 10.1 *Let $p, q, \varepsilon \in [0,1]$. The following hold:*

(a) The functions $d(\cdot, q)$ and $d(p,\cdot)$ are convex and have unique minimizers at q and p respectively.
*(b) $d(p,q) \geq 2(p-q)^2$ (**Pinsker's inequality**).*
(c) If $p \leq q - \varepsilon \leq q$, then $d(p, q-\varepsilon) \leq d(p,q) - d(q-\varepsilon, q) \leq d(p,q) - 2\varepsilon^2$.

The first inequality in (c) is a specialized version of the Pythagorean inequality for Bregman divergences. This is not important here, but see Chapter 26 for more details.

Proof For (a): $d(\cdot, q)$ is the sum of the negative binary entropy function $h(p) = p\log p + (1-p)\log(1-p)$ and a linear function. The second derivative of h is $h''(p) = 1/p + 1/(1-p)$, which is positive and hence h is convex. For fixed p the function $d(p,\cdot)$ is the sum of $h(p)$ and convex functions $p\log(1/q)$ and $(1-p)\log(1/(1-q))$. Hence $d(p,\cdot)$ is convex. The minimizer property follows because $d(p,q) > 0$ unless $p = q$ in which case $d(p,p) = d(q,q) = 0$. A more general version of (b) is given in Chapter 15. A proof of the simple version here follows by considering the function $g(x) = d(p, p+x) - 2x^2$, which obviously satisfies $g(0) = 0$. The proof is finished by showing that this is the unique minimizer of g over the interval $[-p, 1-p]$. The details are left to Exercise 10.1. For (c) notice that

$$h(p) = d(p, q-\varepsilon) - d(p,q) = p\log\frac{q}{q-\varepsilon} + (1-p)\log\frac{1-q}{1-q+\varepsilon}\,.$$

It is easy to see then that h is linear and increasing in its argument. Therefore, since $p \leq q - \varepsilon$,

$$h(p) \leq h(q-\varepsilon) = -d(q-\varepsilon, q)$$

as required for the first inequality of (c). The second inequality follows by using the result in (b). □

The next lemma controls the concentration of the sample mean of a sequence of independent and identically distributed Bernoulli random variables.

LEMMA 10.2 (Chernoff's bound) *Let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables that are Bernoulli distributed with mean μ and let $\hat{\mu} = \frac{1}{n}\sum_{t=1}^{n} X_t$ be the sample mean. Then for $\varepsilon \in [0, 1-\mu]$ it holds that*

$$\mathbb{P}\left(\hat{\mu} \geq \mu + \varepsilon\right) \leq \exp\left(-n d(\mu+\varepsilon, \mu)\right) \tag{10.1}$$

and for $\varepsilon \in [0, \mu]$ *it holds that*

$$\mathbb{P}\left(\hat{\mu} \leq \mu - \varepsilon\right) \leq \exp\left(-nd(\mu - \varepsilon, \mu)\right) . \tag{10.2}$$

Proof We will again use Chernoff's method. Let $\lambda > 0$ be some constant to be chosen later. Then

$$\begin{aligned}
\mathbb{P}\left(\hat{\mu} \geq \mu + \varepsilon\right) &= \mathbb{P}\left(\exp\left(\lambda \sum_{t=1}^{n} (X_t - \mu)\right) \geq \exp\left(\lambda n \varepsilon\right)\right) \\
&\leq \frac{\mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{n} (X_t - \mu)\right)\right]}{\exp\left(\lambda n \varepsilon\right)} \\
&= \left(\mu \exp(\lambda(1 - \mu - \varepsilon)) + (1 - \mu) \exp(-\lambda(\mu + \varepsilon))\right)^n .
\end{aligned}$$

This expression is minimized by $\lambda = \log \frac{(\mu+\varepsilon)(1-\mu)}{\mu(1-\mu-\varepsilon)}$. Therefore

$$\begin{aligned}
&\mathbb{P}\left(\hat{\mu} \geq \mu + \varepsilon\right) \\
&\leq \left(\mu \left(\frac{(\mu + \varepsilon)(1 - \mu)}{\mu(1 - \mu - \varepsilon)}\right)^{1-\mu-\varepsilon} + (1 - \mu) \left(\frac{(\mu + \varepsilon)(1 - \mu)}{\mu(1 - \mu - \varepsilon)}\right)^{-\mu-\varepsilon}\right)^n \\
&= \left(\frac{\mu}{\mu + \varepsilon} \left(\frac{(\mu + \varepsilon)(1 - \mu)}{\mu(1 - \mu - \varepsilon)}\right)^{1-\mu-\varepsilon}\right)^n \\
&= \exp\left(-nd(\mu + \varepsilon, \mu)\right) .
\end{aligned}$$

The bound on the left tail is proven identically. □

Using Pinsker's inequality, it follows that $\mathbb{P}\left(\hat{\mu} \geq \mu + \varepsilon\right), \mathbb{P}\left(\hat{\mu} \leq \mu - \varepsilon\right) \leq \exp(-2n\varepsilon^2)$, which is the same as what can be obtained from Hoeffding's lemma (see (5.8)). Solving $\exp(-2n\varepsilon^2) = \delta$ we recover the usual $1 - \delta$ confidence upper bound. In fact, this cannot be improved when $\mu \approx 1/2$, but the Chernoff bound is much stronger when μ is close to either zero or one. Can we invert the Chernoff tail bound to get confidence intervals which get tighter automatically as μ (or $\hat{\mu}$) approaches zero or one? The following corollary shows how to do this.

COROLLARY 10.1 *Let* $\mu, \hat{\mu}, n$ *be as above. Then, for any* $a \geq 0$,

$$\mathbb{P}\left(d(\hat{\mu}, \mu) \geq a, \hat{\mu} \leq \mu\right) \leq \exp(-na) , \tag{10.3}$$

and

$$\mathbb{P}\left(d(\hat{\mu}, \mu) \geq a, \hat{\mu} \geq \mu\right) \leq \exp(-na) . \tag{10.4}$$

Furthermore, defining

$$U(a) = \max\{u \in [0, 1] \,:\, d(\hat{\mu}, u) \leq a\} ,$$

it follows that with probability $1 - \exp(-na)$, $\mu < U(a)$. *Similarly, letting*

$$L(a) = \min\{u \in [0, 1] \,:\, d(\hat{\mu}, u) \leq a\} ,$$

with probability $1 - \exp(-na)$, $\mu > L(a)$ *holds.*

Proof First, we prove (10.3). Note that $d(\cdot, \mu)$ is decreasing on $[0, \mu]$, and thus, for $0 \leq a \leq d(0, \mu)$, $\{d(\hat{\mu}, \mu) \geq a, \hat{\mu} \leq \mu\} = \{\hat{\mu} \leq \mu - x, \hat{\mu} \leq \mu\} = \{\hat{\mu} \leq \mu - x\}$, where x is the unique solution to $d(\mu - x, \mu) = a$ on $[0, \mu]$. Hence, by Eq. (10.2) of Lemma 10.2, $\mathbb{P}\left(d(\hat{\mu}, \mu) \geq a, \hat{\mu} \leq \mu\right) \leq \exp(-na)$. When $a \geq d(0, \mu)$, the inequality trivially holds. The proof of (10.4) is entirely analogous and hence is omitted. For the second part of the corollary fix a and let $U = U(a)$. First notice that $U \geq \hat{\mu}$ and $d(\hat{\mu}, \cdot)$ is strictly increasing on $[\hat{\mu}, 1]$. Hence, $\{\mu \geq U\} = \{\mu \geq U, \mu \geq \hat{\mu}\} = \{d(\hat{\mu}, \mu) \geq d(\hat{\mu}, U), \mu \geq \hat{\mu}\} = \{d(\hat{\mu}, \mu) \geq a, \mu \geq \hat{\mu}\}$, where the last equality follows by $d(\hat{\mu}, U) = a$, which holds by the definition of U. Taking probabilities and using the first part of the corollary shows that $\mathbb{P}\left(\mu \geq U\right) \leq \exp(-na)$. The statement concerning $L = L(a)$ follows with a similar reasoning. □

Note that for $\delta \in (0, 1)$, $U = U(\log(1/\delta)/n)$ and $L = L(\log(1/\delta)/n)$ are, respectively, upper and lower confidence bounds for μ. While U and L are defined implicitly in terms of an optimization problem. Although the relative entropy has no closed form inverse, the optimization can be solved to a high degree of accuracy using Newton's method (the relative entropy d is convex in its second argument). The advantage of this confidence interval relative to the one derived from Hoeffding's bound is now clear. As $\hat{\mu}$ approaches one the width of the interval $U(a) - \hat{\mu}$ approaches zero, whereas the width of the interval provided by Hoeffding's bound stays at $\sqrt{\log(1/\delta)/(2n)}$. The same holds for $\hat{\mu} - L(a)$ as $\hat{\mu} \to 0$.

EXAMPLE 10.1 Fig. 10.1 shows a plot of $d(3/4, x)$ and the lower bound given by Pinsker's inequality. The approximation degrades as $|x - 3/4|$ grows large, especially for $x > 3/4$. As explained in Corollary 10.1, the graph of $d(\hat{\mu}, \cdot)$ can be used to derive confidence bounds by solving for $d(\hat{\mu}, x) = a = \log(1/\delta)/n$. Assuming $\hat{\mu} = 3/4$ is observed, a confidence level of 90% with $n = 10$, $a \approx 0.23$. The confidence interval be read out from the figure by finding those values where the horizontal dashed black line intersects the solid blue line. The resulting confidence interval will be highly asymmetric. Note that in this scenario the lower confidence bounds produced by both Hoeffding's inequality and Chernoff's bound are similar while the upper bound provided by Hoeffding's bound is vacuous.

10.2 The KL-UCB algorithm

The KL-UCB algorithm is nothing more than UCB, but with Chernoff's bound used to define the upper confidence bound, rather than Lemma 5.1.

THEOREM 10.1 *If the reward in round t is $X_t \sim \mathcal{B}(\mu_{A_t})$, then the regret of*

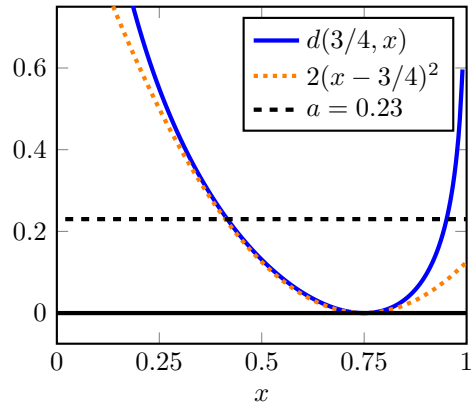

Figure 10.1 Relative entropy and Pinsker's inequality

1: **Input** K
2: Choose each arm once
3: Subsequently choose

$$A_t = \operatorname{argmax}_i \max\left\{\tilde{\mu} \in [0,1] : d(\hat{\mu}_i(t-1), \tilde{\mu}) \leq \frac{\log f(t)}{T_i(t-1)}\right\},$$
$$\text{where } f(t) = 1 + t\log^2(t)\,.$$

Algorithm 7: KL-UCB

Algorithm 7 is bounded by

$$R_n \leq \sum_{i:\Delta_i>0} \inf_{\substack{\varepsilon_1,\varepsilon_2>0 \\ \varepsilon_1+\varepsilon_2\in(0,\Delta_i)}} \Delta_i \left(\frac{f(n)}{d(\mu_i+\varepsilon_1, \mu^*-\varepsilon_2)} + \frac{1}{2\varepsilon_1^2} + 1 + \frac{1}{\varepsilon_2^2}\right).$$

Furthermore, $\limsup_{n\to\infty} \frac{R_n}{\log(n)} \leq \sum_{i:\Delta_i>0} \frac{\Delta_i}{d(\mu_i,\mu^*)}$.

Let us now compare the asymptotic result above to that given for UCB in Theorem 8.1. Specializing this result for Bernoulli rewards (which are 1/2-subgaussian), we get

$$\limsup_{n\to\infty} \frac{R_n}{\log(n)} \leq \sum_{i:\Delta_i>0} \frac{1}{2\Delta_i}.$$

By Pinsker's inequality (part (b) of Lemma 10.1) we see that $d(\mu_i,\mu^*) \geq 2(\mu^*-\mu_i)^2 = 2\Delta_i^2$, which means that the asymptotic regret of KL-UCB is never worse than that of UCB. On the other hand, a Taylor's expansion shows that when μ_i and μ^* are close (the hard case in the asymptotic regime), we have

$$d(\mu_i,\mu^*) = \frac{\Delta_i^2}{2\mu_i(1-\mu_i)} + o(\Delta_i^2)\,,$$

indicating that the regret of KL-UCB is approximately

$$\limsup_{n\to\infty} \frac{R_n}{\log(n)} \approx \sum_{i:\Delta_i>0} \frac{2\mu_i(1-\mu_i)}{\Delta_i}\,. \tag{10.5}$$

We might not be so surprised to notice that $\mu_i(1-\mu_i)$ is the variance of a Bernoulli distribution with mean μ_i. Now $\mu_i(1-\mu_i) \le 1/4$, which shows that KL-UCB is never worse than the asymptotically optimal variant of UCB presented in the last chapter. But when μ_i is close to either zero or one, then KL-UCB is a big improvement.

The proof of Theorem 10.1 relies on two lemmas. The first is used to show that the index of the optimal arm is never too far below its true value, while the second shows that the index of any other arm is not often much larger than the same value. These results mirror those given for UCB, but things are complicated by the non-symmetric and hard-to-invert divergence function.

For the next results we define $d^+(p,q) = d(p,q)\mathbb{I}\{p<q\}$, $p,q\in[0,1]$.

LEMMA 10.3 *Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli random variables with mean $\mu\in[0,1]$, $\varepsilon>0$ and*

$$\tau = \min\left\{t : \max_{1\le s\le n} d^+(\hat{\mu}_s, \mu-\varepsilon) - \frac{\log f(t)}{s} \le 0\right\}.$$

Then, $\mathbb{E}[\tau] \le \dfrac{2}{\varepsilon^2}$.

Proof We start with a high probability bound and then integrate to control the expectation.

$$\begin{aligned}
\mathbb{P}(\tau > t) &\le \mathbb{P}\left(\exists 1\le s\le n : d^+(\hat{\mu}_s, \mu-\varepsilon) > \frac{\log f(t)}{s}\right) \\
&\le \sum_{s=1}^{n} \mathbb{P}\left(d^+(\hat{\mu}_s, \mu-\varepsilon) > \frac{\log f(t)}{s}\right) \\
&= \sum_{s=1}^{n} \mathbb{P}\left(d(\hat{\mu}_s, \mu-\varepsilon) > \frac{\log f(t)}{s}, \hat{\mu}_s < \mu-\varepsilon\right) \\
&\le \sum_{s=1}^{n} \mathbb{P}\left(d(\hat{\mu}_s, \mu) > \frac{\log f(t)}{s} + 2\varepsilon^2, \hat{\mu}_s < \mu\right) && \text{((c) of Lemma 10.1)} \\
&\le \sum_{s=1}^{n} \exp\left(-s\left(2\varepsilon^2 + \frac{\log f(t)}{s}\right)\right) && \text{(Eq. (10.3) of Corollary 10.1)} \\
&\le \frac{1}{f(t)} \sum_{s=1}^{n} \exp\left(-2s\varepsilon^2\right) \\
&\le \frac{1}{2f(t)\varepsilon^2}\,.
\end{aligned}$$

To finish, we integrate the tail,

$$\mathbb{E}[\tau] \le \int_0^\infty \mathbb{P}(\tau \ge t)\, dt \le \frac{1}{2\varepsilon^2} \int_0^\infty \frac{dt}{f(t)} \le \frac{2}{\varepsilon^2}\,. \qquad \square$$

LEMMA 10.4 *Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli random variables with mean μ. Further, let $\Delta > 0$, $a > 0$ and define*

$$\kappa = \sum_{s=1}^{n} \mathbb{I}\left\{ d(\hat{\mu}_s, \mu + \Delta) \le \frac{a}{s} \right\}.$$

Then, $\mathbb{E}[\kappa] \le \inf_{\varepsilon \in (0,\Delta)} \left(1 + \dfrac{a}{d(\mu+\varepsilon, \mu+\Delta)} + \dfrac{1}{2\varepsilon^2} \right)$.

Proof Let $\varepsilon \in (0, \Delta)$ and $u = a/d(\mu+\varepsilon, \mu+\Delta)$. Then

$$\begin{aligned}
\mathbb{E}[\kappa] &= \sum_{s=1}^{n} \mathbb{P}\left(d(\hat{\mu}_s, \mu+\Delta) \le \frac{a}{s} \right) \\
&\le \sum_{s=1}^{n} \mathbb{P}\left(\hat{\mu}_s \ge \mu + \varepsilon \text{ or } d(\mu+\varepsilon, \mu+\Delta) \le \frac{a}{s} \right) \\
&\qquad\qquad (d(\cdot, \mu+\Delta) \text{ is decreasing on } [0, \mu+\Delta]) \\
&\le u + \sum_{s=\lceil u \rceil}^{n} \mathbb{P}(\hat{\mu}_s \ge \mu + \varepsilon) \\
&\le u + \sum_{s=1}^{\infty} \exp(-s d(\mu+\varepsilon, \mu)) \qquad \text{(Lemma 10.2)} \\
&\le 1 + \frac{a}{d(\mu+\varepsilon, \mu+\Delta)} + \frac{1}{d(\mu+\varepsilon, \mu)} \\
&\le 1 + \frac{a}{d(\mu+\varepsilon, \mu+\Delta)} + \frac{1}{2\varepsilon^2} \qquad \text{(Pinsker's inequality/Lemma 10.1(b))}
\end{aligned}$$

as required. $\square$

Proof of Theorem 10.1 As in other proofs we assume without loss of generality that $\mu_1 = \mu^*$ and bound $\mathbb{E}[T_i(n)]$ for suboptimal arms i. To this end, fix a suboptimal arm i and let $\varepsilon_1 + \varepsilon_2 \in (0, \Delta_i)$ with both ε_1 and ε_2 positive. Define

$$\begin{aligned}
\tau &= \min\left\{ t : \max_{1 \le s \le n} d^+(\hat{\mu}_{1s}, \mu_1 - \varepsilon_2) - \frac{\log f(t)}{s} \le 0 \right\}, \text{ and} \\
\kappa &= \sum_{s=1}^{n} \mathbb{I}\left\{ d(\hat{\mu}_{is}, \mu_i + \Delta_i - \varepsilon_2) \le \frac{\log f(n)}{s} \right\}.
\end{aligned}$$

By a similar reasoning as in the proof of Theorem 8.1,

$$\begin{aligned}
\mathbb{E}[T_i(n)] &= \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{I}\{A_t = i\}\right] \\
&\leq \mathbb{E}[\tau] + \mathbb{E}\left[\sum_{t=\tau+1}^{n} \mathbb{I}\{A_t = i\}\right] \\
&\leq \mathbb{E}[\tau] + \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{I}\left\{A_t = i \text{ and } d(\hat{\mu}_{i,T_i(t-1)}, \mu_1 - \varepsilon_2) \leq \frac{\log f(t)}{T_i(t-1)}\right\}\right] \\
&\leq \mathbb{E}[\tau] + \mathbb{E}[\kappa] \\
&\leq 1 + \frac{2}{\varepsilon_2^2} + \frac{f(n)}{d(\mu_i + \varepsilon_1, \mu^* - \varepsilon_2)} + \frac{1}{2\varepsilon_1^2},
\end{aligned}$$

where the second inequality follows since by the definition of τ, if $t > \tau$, then the index of the optimal arm is at least as large as $\mu_1 - \varepsilon_2$. The third inequality follows from the definition of κ as in the proof of Theorem 8.1. The final inequality follows from Lemmas 10.3 and 10.4. The first claim of the theorem is completed by substituting the above into the standard regret decomposition

$$R_n = \sum_{i=1}^{k} \Delta_i \mathbb{E}[T_i(n)].$$

The asymptotic claim is left as an exercise. □

10.3 Notes

1 The new concentration inequality (Lemma 10.2) actually holds more generally for any sequence of independent and identically distributed random variables $X_1, X_2, \ldots, X_n$ provided only that $X_t \in [0,1]$ almost surely. Therefore all results in this section also hold if the assumption that the noise is Bernoulli is relaxed to the case where it is simply supported in $[0,1]$ (or other bounded sets by shifting/scaling).

2 Expanding on the previous note, all that is required is a bound on the moment generating function for random variables X where $X \in [0,1]$ almost surely. Garivier and Cappé [2011, Lemma 9] noted that $f(x) = \exp(\lambda x) - x(\exp(\lambda) - 1) - 1$ is negative on $[0,1]$ and so

$$\mathbb{E}[\exp(\lambda X)] \leq \mathbb{E}[X(\exp(\lambda) - 1) + 1] = \mu \exp(\lambda) + 1 - \mu,$$

which is precisely the moment generating function of the Bernoulli distribution with mean μ. Then the remainder of the proof of Lemma 10.2 goes through unchanged. This shows that for any bandit $\nu = (P_i)_i$ with $\text{Supp}(P_i) \in [0,1]$ for

all i the regret of the policy in Algorithm 7 satisfies

$$\limsup_{n\to\infty} \frac{R_n}{\log(n)} \leq \sum_{i:\Delta_i>0} \frac{\Delta_i}{d(\mu_i, \mu^*)} .$$

3 The bounds obtained using the argument in the previous note are not quite tight. Specifically one can show there exists an algorithm such that for all bandits $\nu = (P_i)_i$ with P_i the reward distribution of the ith arm supported on $[0, 1]$, then

$$\limsup_{n\to\infty} \frac{R_n}{\log(n)} = \sum_{i:\Delta_i>0} \frac{\Delta_i}{d_{[0,1]}(P_i, \mu^*)} ,$$

where

$$d_{[0,1]}(P_i, \mu^*) = \inf\{\mathrm{D}(P_i, P) : \mu(P) > \mu^* \text{ and } \mathrm{Supp}(P) \subset [0, 1]\} ,$$

where $\mathrm{D}(P, Q)$ is the relative entropy between measures P_i and P, which we define in Chapter 14. This last quantity is never smaller than $d(\mu_i, \mu^*)$. For details on this we refer the reader to the paper by Honda and Takemura [2010].

4 The approximation in Eq. (10.5) was used to show that the regret for KL-UCB is closely related to the variance of the Bernoulli distribution. It is natural to ask whether or not this result could be derived, at least asymptotically, by appealing to the central limit theorem. The answer is no! First, the quality of the approximation in Eq. (10.5) does not depend on n, so asymptotically it is not true that the Bernoulli bandit behaves like a Gaussian bandit with variances tuned to match. The reason is that as n tends to infinity, the confidence level should be chosen so that the risk of failure also tends to zero. But the central limit theorem does not provide information about the tails with probability mass less than $O(n^{-1/2})$. See Note 1 in Chapter 10.

5 This style of analysis is easily generalized to a wide range of alternative noise models, with the easiest being single parameter exponential families (Exercise 10.4).

6 Chernoff credits Lemma 10.2 to his friend Herman Rubin [Chernoff, 2014], but the name seems to have stuck.

10.4 Bibliographic remarks

Several authors have worked on Bernoulli bandits and the asymptotics have been well-understood since the article by Lai and Robbins [1985]. The earliest version of the algorithm presented in this chapter is due to Lai [1987] who provided asymptotic analysis. The finite-time analysis of KL-UCB was given by two groups simultaneously (and published in the same conference!) by Garivier and Cappé [2011] and Maillard et al. [2011] (see also the combined journal article: Cappé et al. 2013). Two alternatives are the DMED Honda and Takemura [2010] and IMED [Honda and Takemura, 2015] algorithms. These works go

after the problem of understanding the asymptotic regret for the more general situation where the rewards lie in a bounded interval (see Note 3). The latter work covers even the semi-bounded case where the rewards are almost surely upper bounded. Both algorithms are asymptotically optimal. Ménard and Garivier [2017] combined MOSS and KL-UCB to derive an algorithm that is minimax optimal and asymptotically optimal for single parameter exponential families. While the subgaussian and Bernoulli examples are very fundamental, there has also been work on more generic setups where the unknown reward distribution for each arm is known to lie in some class $\mathcal{F}$. The article by Burnetas and Katehakis [1996] gives the most generic (albeit, asymptotic) results. These generic setups remain wide open for further work.

10.5 Exercises

10.1 [Pinsker's inequality] Prove Lemma 10.1(b).

Consider the function $g(x) = d(p, p+x) - 2x^2$ over the $[-p, 1-p]$ interval. By taking derivatives, show that $g \geq 0$.

10.2 Let $\mathbb{F} = (\mathcal{F}_t)_t$ be a filtration, $(X_t)_t$ be $[0,1]$-valued, $\mathbb{F}$-adapted sequence, such that $\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] = \mu_t$ for some $\mu_1, \ldots, \mu_n \in [0,1]$ non-random numbers. Define $\mu = \frac{1}{n}\sum_{t=1}^n \mu_t$, $\hat{\mu} = \frac{1}{n}\sum_{t=1}^n X_t$. Prove that the conclusion of Lemma 10.2 still holds.

Read note 2 at the end of this chapter. Let $g(\cdot, \mu)$ be the cumulant generating function of the μ-parameter Bernoulli distribution. For $X \sim \mathcal{B}(\mu)$, $\lambda \in \mathbb{R}$, $g(\lambda, \mu) = \log \mathbb{E}[\exp(\lambda X)]$. Show that $g(\lambda, \cdot)$ is concave. Next, use this and the tower rule to show that $\mathbb{E}[\exp(\lambda n(\hat{\mu} - \mu))] \leq g(\lambda, \mu)^n$.

The bound of the previous exercise is most useful when all μ_t are either all close to 0 or they are all close to 1. In particular, if half of the $\{\mu_t\}$ is close to zero, half of them is close to one, the bound will degrade to Hoeffding's bound. Irrespective of this, it is useful to notice that the claims made in Corollary 10.1 continue to hold for $\hat{\mu}$, μ as defined in the exercise.

10.3 Prove the asymptotic claim in Theorem 10.1.

Choose $\varepsilon_1, \varepsilon_2$ to decrease slowly with n and use the first part of the theorem.

10.4 Let h be a measure on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ and $T : \mathbb{R} \to \mathbb{R}$. The function T is called the **sufficient statistic**. Define

$$\frac{dP_\theta}{dh}(x) = \exp(\theta T(x) - A(\theta))\,,$$

where $A(\theta)$ is the **log partition function** given by

$$A(\theta) = \log \int_{\mathbb{R}} \exp(\theta T(x)) dh(x)\,.$$

Let $\Theta = \{\theta \in \mathbb{R} : A(\theta) \text{ exists}\}$. The set $\{P_\theta : \theta \in \Theta\}$ is called an **exponential family**. For more details see the note after the exercise.

(a) Prove that for $\theta \in \Theta$ the function $P_\theta : \mathfrak{B}(\mathbb{R}) \to [0, 1]$ given by

$$P_\theta(A) = \int_A \frac{dP_\theta}{dh}(x) dh(x)$$

is a probability measure on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$.

(b) Let $\mathbb{E}_\theta$ denote expectations with respect to P_θ and show that $A'(\theta) = \mathbb{E}_\theta[T(x)]$.

(c) Find a choice of h and T such that $\{P_\theta : \theta \in \Theta\}$ is the family of Bernoulli distributions.

(d) Find a choice of h and T such that $\{P_\theta : \theta \in \Theta\}$ is the family of Gaussian distributions with unit variance means in $\mathbb{R}$.

(e) Let $\theta \in \Theta$ and $X \sim P_\theta$. Show that

$$\mathbb{E}_\theta[\exp(\lambda T(X))] = \exp(A(\lambda + \theta) - A(\theta))\,.$$

(f) Given $\theta, \theta' \in \Theta$, show that

$$d(\theta, \theta') = \mathbb{E}_\theta\left[\log\left(\frac{p_\theta(X)}{p_{\theta'}(X)}\right)\right] = A(\theta') - A(\theta) - (\theta' - \theta)A'(\theta)\,.$$

(g) Let $\theta, \theta' \in \Theta$ be such that $A'(\theta') \geq A'(\theta)$ and $X_1, \ldots, X_n$ be independent and identically distributed and $\hat{T} = \frac{1}{n}\sum_{t=1}^n T(X_t)$. Show that

$$\mathbb{P}\left(\hat{T} \geq A'(\theta')\right) \leq \exp\left(-nd(\theta, \theta')\right)\,,$$

(h) Let $\mathcal{E}$ be the set of all bandits with reward distributions in family $\{P_\theta : \theta \in \Theta\}$. Design a policy π such that for all $\nu \in \mathcal{E}$ it holds that

$$\lim_{n\to\infty} \frac{R_n(\pi, \nu)}{\log(n)} \leq \sum_{i:\Delta_i > 0} \frac{\Delta_i}{d(\theta_i, \theta_{i^*})}\,,$$

where θ_i is such that P_{θ_i} is the distribution of the rewards for arm i and i^* is the optimal arm.

Exponential families represent a wide range of statistical models. We discuss them in more detail in Chapter 34. The function $d(\theta, \theta')$ is called the **relative entropy** between P_θ and $P_{\theta'}$. We discuss this concept more in Chapter 14. The bound in the last part of the exercise cannot be improved as we explain in Chapter 16.

10.5 In this exercise you compare KL-UCB and UCB empirically.

(a) Implement Algorithm 7 and Algorithm 5 where the latter algorithm should be tuned for 1/2-subgaussian bandits so that

$$A_t = \operatorname{argmax}_{i \in [K]} \hat{\mu}_i(t-1) + \sqrt{\frac{\log(f(t))}{2T_i(t-1)}}\,.$$

(b) Let $n = 10000$ and $K = 2$. Plot the expected regret of each algorithm as a function of Δ when $\mu_1 = 1/2$ and $\mu_2 = 1/2 + \Delta$.

(c) Repeat the above experiment with $\mu_1 = 1/10$ and $\mu_1 = 9/10$.

(d) Discuss your results.