CS6046: Multi-armed bandits

Course notes

Prashanth L. A.

April 1, 2019

Contents

Chapter 1

Regret minimization in K-armed bandits

1.1 The framework

Suppose we are given K arms with unknown distributions $P_k, k = 1, \ldots, K$.

The interaction of the bandit algorithm with the environment proceeds as follows:

> **Bandit interaction**
>
> **For** $t = 1, 2, \ldots, n$**, repeat**
>
> (1) Bandit algorithm selects an arm $I_t \in \{1, \ldots, K\}$.
>
> (2) The environment returns a sample X_t from the distribution P_{I_t} corresponding to the arm I_t.

Let μ_k denote the expected value of the stochastic rewards from arm k, for $k = 1, \ldots, K$. The optimal arm is one that has the highest expected value, i.e., $\mu_* = \max_{k=1,\ldots,K} \mu_k$.

The goal of the bandit algorithm is to maximize $S_n = \sum_{t=1}^{n} X_t$. Notice that S_n is a random variable (r.v.) and hence, has a distribution. So, a natural objective is to design an algorithm that maximizes $\mathbb{E}(S_n)$.

The framework outlined above captures the "exploration-exploitation dilemma". To elaborate, in any round the bandit algorithm can choose to either *explore* by pulling an arm to estimate its mean reward, or *exploit* by pulling an arm that has the highest estimated mean reward. Notice that the rewards are stochastic, i.e, each arm has a reward distribution with a mean and certain spread. Since the bandit algorithm does not know the arms' reward distributions, it has to estimate the mean rewards though sampling and the sampling has to be adaptive, i.e., in any round, based on the samples obtained so far, the bandit algorithm has to *adaptively* decide which arm to pull next. A bandit algorithm that explores too often would end with a lower expected value for the total reward S_n. On the other hand, an algorithm that does not sample the individual arms enough number of times to be confident about their mean rewards would end up pulling a sub-optimal excessively in the exploit stage and this would again lead to a low S_n in expectation. Thus, the need is for an algorithm that explores just enough to discard the bad arms (i.e., those with low mean rewards) and zeroes in on the best arm at the earliest. The notion of *regret* that we define next formalizes the exploration exploitation dilemma.

Regret

The cumulative regret R_n incurred by a bandit algorithm is defined as follows:

$$R_n = n\mu_* - \mathbb{E}\left(\sum_{t=1}^{n} X_t\right).$$

The following lemma gives a useful alternative form for the regret R_n.

Lemma 1.1.

$$R_n = \sum_{k=1}^{K} \mathbb{E}[T_k(n)]\Delta_k,$$

where $T_k(n) = \sum_{t=1}^{n} \mathbb{I}\{I_t = k\}$ *is the number of times arm* k *is pulled up to time* n *and* $\Delta_k = \mu_* - \mu_k$ *denotes the gap between the expected rewards of the optimal arm and of arm k.*

From the form for regret in the lemma above, it is apparent that the contribution to regret from pulls corresponding to optimal arm is zero, since the gap Δ_{a^*} corresponding to optimal arm $a^* = \arg\max_{i=1,\ldots,K} \mu_i$ is zero. Thus, a bandit algorithm incurs regret only by pulling suboptimal arms. However, the mean rewards for each arm has to be estimated and hence the challenge is to balance estimating the mean rewards well-enough (exploration) and minimizing regret by pulling optimal arm (exploitation) and this dilemma is captured by the notion of regret.

Proof. Notice that

$$S_n = \sum_{t=1}^{n} X_t = \sum_{t=1}^{n}\sum_{k=1}^{K} X_t \mathbb{I}\{I_t = k\},$$

where we have used the fact that $\sum_{k=1}^{K} \mathbb{I}\{I_t = k\} = 1$. So,

$$\begin{aligned}\mathbb{E}S_n &= \sum_{t=1}^{n}\sum_{k=1}^{K} \mathbb{E}\left(X_t \mathbb{I}\{I_t = k\}\right) \\ &= \sum_{t=1}^{n}\sum_{k=1}^{K} \mathbb{E}\left(\mathbb{E}\left(X_t \mathbb{I}\{I_t = k\} \mid I_t\right)\right). \qquad (1.1)\end{aligned}$$

The conditional expectation inside the summation above can be simplified as follows:

$$\mathbb{E}\left(X_t \mathbb{I}\{I_t = k\} \mid I_t\right) = \mathbb{I}\{I_t = k\}\,\mathbb{E}\left(X_t \mid I_t\right) = \mathbb{I}\{I_t = k\}\,\mu_{I_t} = \mathbb{I}\{I_t = k\}\,\mu_k,$$

where we used the fact that given I_t, $\mathbb{E}(X_t \mid I_t) = \mu_{I_t}$. Plugging the final equality above into (1.1), we obtain

$$\mathbb{E}S_n = \sum_{k=1}^{K} \sum_{t=1}^{n} \mathbb{E}\left(\mathbb{I}\{I_t = k\}\, \mu_k\right) \tag{1.2}$$

$$= \sum_{k=1}^{K} \mu_k \mathbb{E}\left(T_k(n)\right). \tag{1.3}$$

The lemma follows by substituting the above into the classic regret definition, i.e., $R_n = n\mu_* - \mathbb{E}S_n$ together with the definition of the gaps Δ_k. □

1.2 Explore and then commit

We start with a naive algorithm that clearly separates exploration and exploitation stages.

Explore-then-commit (ETC) algorithm

(1) ***Exploration phase:*** During rounds $1, \ldots, mK$, play each arm m times.

(2) ***Exploitation phase:*** During the remaining $n - mK$ rounds, play the arm with the highest empirical mean reward, i.e., $\arg\max_{k=\{1,\ldots,K\}} \hat{\mu}_k(mK)$.

In the above, the empirical mean or sample mean for arm i at any time t is defined as

$$\hat{\mu}_i(t) \triangleq \frac{1}{T_i(t)} \sum_{s=1}^{t} X_s \mathbb{I}\{I_s = i\}. \tag{1.4}$$

Notice that we have not specified the exploration parameter m in the algorithm above. The regret analysis in the following section would address this gap. In short, choosing m optimally (i.e., to minimize regret incurred) would require the knowledge of the underlying gaps and the latter information in not available in a bandit learning framework. On the other hand, a choice such as $m = \Theta(n^{2/3})$ would result in a regret of the order $\tilde{O}(n^{2/3})$[1]. As we shall see much later, when we present the UCB algorithm, a regret of $\tilde{O}(\sqrt{n})$ can be achieved on any problem instance and hence, ETC algorithm clearly exhibits suboptimal performance.

Regret analysis

Using arguments similar to that in the proof of Lemma 1.1, one can arrive at the following form for the regret R_n:

$$R_n = \sum_{t=1}^{n} \mathbb{E}\left(\Delta_{I_t}\right).$$

Observe that, in the first mK rounds, since ETC algorithm is exploring, the contribution to regret from this phase is $m \sum_{i=1}^{K} \Delta_i$. On the other hand, during the exploration phase, if arm i has

[1] $\tilde{O}(\cdot)$ is like the regular Oh-notation, except that the log factors are hidden.

the highest sample mean, then the contribution to the regret is $(n-mK)\Delta_i$. Of course, an arm, say i, getting picked for exploitation hinges on the event that $\mathbb{I}\left\{i = \operatorname{arg\,max}_{j=1,\ldots,K} \hat{\mu}_j(mK)\right\}$. Thus, the regret of ETC can be simplified as follows:

$$\begin{aligned} R_n &= m\sum_{i=1}^{K}\Delta_i + (n-mK)\sum_{i=1}^{K}\Delta_i \mathbb{E}\left(\mathbb{I}\left\{i = \operatorname*{arg\,max}_{j=1,\ldots,K} \hat{\mu}_j(mK)\right\}\right) \\ &= m\sum_{i=1}^{K}\Delta_i + (n-mK)\sum_{i=1}^{K}\Delta_i \mathbb{P}\left[i = \operatorname*{arg\,max}_{j=1,\ldots,K} \hat{\mu}_j(mK)\right]. \end{aligned}$$

The probability of the event in the second term on the RHS above can be upper-bounded by using the fact that if arm i got picked for exploitation, then its sample mean is certainly better than that of the best arm a^* and we obtain

$$\begin{aligned} \mathbb{P}\left[i = \operatorname*{arg\,max}_{j=1,\ldots,K} \hat{\mu}_j(mK)\right] &\leq \mathbb{P}\left[\hat{\mu}_i(mK) \geq \hat{\mu}_{a^*}(mK)\right] \\ &= \mathbb{P}\left[\hat{\mu}_i(mK) - \mu_i - (\hat{\mu}_{a^*}(mK) - \mu_*) \geq \Delta_i\right]. \end{aligned}$$

Thus, understanding the regret of ETC comes down to how well $\hat{\mu}_i$ and $\hat{\mu}_{a^*}$ estimate the true means μ_i and μ_*, respectively. At this point, we take a detour and understand concentration inequalities, which assist in answering the aforementioned question on estimation.

1.3 A brief tour of concentration inequalities

Suppose $X_1, \ldots, X_n$ are i.i.d. samples of a r.v. X with mean μ and variance σ^2. Using these samples, a popular estimator for μ is the sample mean $\hat{\mu}_n$, defined by

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{1.5}$$

In the following, we attempt to address the question of how quickly can $\hat{\mu}_n$ concentrate around true mean μ. First, notice that

$$\mathbb{E}\left(\hat{\mu}_n\right) = \mu, \text{Var}\left(\hat{\mu}_n\right) = \frac{\sigma^2}{n}, \text{ and } \mathbb{E}\left((\hat{\mu}_n - \mu)^2\right) = \frac{\sigma^2}{n}.$$

From the last equality above, it is apparent that the square error $(\hat{\mu}_n - \mu)^2$ converges to zero in expectation, as the number of samples n increase. However, a more useful result on the error in estimation would bound the following tail probabilities:

$$\mathbb{P}\left[\hat{\mu}_n \geq \mu + \epsilon\right] \text{ and } \mathbb{P}\left[\hat{\mu}_n \leq \mu - \epsilon\right], \text{ for any } \epsilon > 0.$$

Upper bounds on the probabilities above would ensure that $\mu \in [\hat{\mu}_n - c_n, \hat{\mu}_n + c_n]$ with high probability, for some c_n that can inferred from the bounds on the tail probabilities.

A first step is to employ Markov inequality, which states that for any positive-valued r.v. X with mean μ, $\mathbb{P}\left[X \geq \epsilon\right] \leq \frac{\mu}{\epsilon}$. A application of this inequality for a r.v. X that is not necessarily positive-valued, but with finite variance σ^2, we obtain the well-known Chebyschev's inequality:

$$\mathbb{P}\left[|X - \mu| \geq \epsilon\right] \leq \frac{\sigma^2}{\epsilon^2}.$$

Applying the inequality above to the sample mean $\hat{\mu}_n$ and using the fact that $\text{Var}\left(\hat{\mu}_n\right) = \frac{\sigma^2}{n}$, we obtain

$$\mathbb{P}\left[|\hat{\mu}_n - \mu| \geq \epsilon\right] \leq \frac{\sigma^2}{n\epsilon^2}.$$

The inequality above suggests a decay rate of the order $1/n$ for the tail probabilities. While this rate is arrived by imposing that the underlying r.v. has a finite variance, one can obtain significantly better rates for random variables whose distributions have tails that do not go above that of a Gaussian r.v. Before deriving tail bounds for such sub-Gaussian r.v.s, we look at the asymptotic tail bound that central limit theorem would provide.

Theorem 1.2. ***Central limit theorem (CLT)*** *Let $S_n = \sum_{i=1}^{n}(X_i - \mu)$, where, as before, $X_1, \ldots, X_n$ are i.i.d. samples of r.v. X with mean μ and variance σ^2. Then,*

$$\frac{S_n}{\sqrt{n}} \xrightarrow{\text{in distribution}} \mathcal{N}(0, \sigma^2) \text{ as } n \to \infty.$$

Using CLT, the tail probability concerning sample mean can be bounded as follows:

$$\begin{aligned}
\mathbb{P}\left[\hat{\mu}_n - \mu \geq \epsilon\right] &= \mathbb{P}\left[\frac{S_n}{\sqrt{n}} \geq \epsilon\sqrt{n}\right] \\
&\approx \int_{\epsilon\sqrt{n}}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-x^2}{2\sigma^2}\right) dx \\
&\leq \int_{\epsilon\sqrt{n}}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}\epsilon\sqrt{n}} x \exp\left(\frac{-x^2}{2\sigma^2}\right) dx \\
&= \sqrt{\frac{\sigma^2}{2\pi n\epsilon^2}} \times \left[-\exp\left(\frac{-x^2}{2\sigma^2}\right)\right]_{\epsilon\sqrt{n}}^{\infty} \\
&= \sqrt{\frac{\sigma^2}{2\pi n\epsilon^2}} \times \exp\left(\frac{-n\epsilon^2}{2\sigma^2}\right).
\end{aligned}$$

Observe that the bound, suggested by CLT, on the tail probability implies an exponential decay, while the bound obtained by Chebyschev's inequality was of the order $1/n$ (which is much weaker than the $\exp\left(-cn\right)$). While this is an encouraging result that suggests sample mean concentrates exponentially fast around the true mean, the CLT bound comes with a major caveat, which is that it is an asymptotic bound (or holds for large n only). One can overcome this issue and obtain non-asymptotic concentration bounds, provided the underlying distribution satisfies certain properties. As a gentle start, we investigate concentration of measure when the underlying distribution is Gaussian.

Gaussian concentration

Consider a r.v. X with distribution $\mathcal{N}(0, \sigma^2)$. Then,

$$\begin{aligned}\mathbb{P}\left[X \geq \epsilon\right] &= \mathbb{P}\left[\exp\left(\lambda X\right) \geq \exp\left(\lambda\epsilon\right)\right], \text{ for any } \lambda > 0,\\ &\leq \exp\left(-\lambda\epsilon\right)\mathbb{E}(\exp\left(\lambda X\right)).\end{aligned} \tag{1.6}$$

The last step above follows from Markov inequality. For a Gaussian r.v. X, $\mathbb{E}(\exp\left(\lambda X\right))$ is simplified as follows:

$$\begin{aligned}\mathbb{E}(\exp\left(\lambda X\right)) &= \int_{-\infty}^{\infty} \exp\left(\lambda x\right) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-x^2}{2\sigma^2}\right) dx\\ &= \int_{-\infty}^{\infty} \exp\left(\lambda \sigma z\right) \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right) dz\\ &= \exp\left(\frac{\lambda^2\sigma^2}{2}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(z-\lambda\sigma)^2}{2}\right) dz\\ &= \exp\left(\frac{\lambda^2\sigma^2}{2}\right).\end{aligned}$$

Substituting the simplified form for $\mathbb{E}(\exp\left(\lambda X\right))$ into (1.6), we obtain

$$\mathbb{P}\left[X \geq \epsilon\right] \leq \exp\left(-\lambda\epsilon + \frac{\lambda^2\sigma^2}{2}\right).$$

A straightforward calculation yields $\frac{\epsilon}{\sigma^2}$ value for the optimum λ value that minimizes the RHS above and for the optimal λ, we have

$$\mathbb{P}\left[X \geq \epsilon\right] \leq \exp\left(\frac{-\epsilon^2}{2\sigma^2}\right). \tag{1.7}$$

The bound above is obtained through the well-known "Chernoff method".

Supposing that $X_1, \ldots, X_n$ are i.i.d. copies of a $\mathcal{N}(0, \sigma^2)$ r.v., we have

$$\mathbb{P}\left[\hat{\mu}_n \geq \mu + \epsilon\right] \leq \exp\left(\frac{-n\epsilon^2}{2\sigma^2}\right). \tag{1.8}$$

Sub-Gaussianity

We now generalize the Chernoff bound to the class of sub-Gaussian r.v.s, defined below.

Definition 1.1. *A r.v. X is σ-sub-Gaussian if there exists a $\sigma > 0$ such that*

$$\mathbb{E}(\exp\left(\lambda X\right)) \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right) \textit{ for any } \lambda \in \mathbb{R}.$$

For a σ-sub-Gaussian r.v. X, the Chernoff method gives

$$\mathbb{P}\left[X \geq \epsilon\right] \leq \exp\left(\frac{-\epsilon^2}{2\sigma^2}\right). \tag{1.9}$$

A few examples of sub-Gaussian r.v.s are given below.

Example 1.1. *A r.v. X is Rademacher if $\mathbb{P}[X=+1]=\mathbb{P}[X=-1]=\frac{1}{2}$. A Rademacher r.v. X is 1-sub-Gaussian. This can be argued as follows:*

$$\begin{aligned}\mathbb{E}(\exp(\lambda X)) &= \frac{1}{2}(\exp(\lambda)-\exp(-\lambda))\\ &= \frac{1}{2}\left(\sum_{k=0}^{\infty}\frac{(-\lambda)^k}{k!}+\sum_{k=0}^{\infty}\frac{(-\lambda)^k}{k!}\right)\\ &= \sum_{k=0}^{\infty}\frac{\lambda^{2k}}{(2k+1)!}\le \exp\left(\frac{\lambda^2}{2}\right).\end{aligned}$$

The inequality above follows by using $(2k)!\ge 2^k k!$.

Example 1.2. *A $U[-a,a]$ distributed r.v. X is a-sub-Gaussian and this can be argued as follows:*

$$\begin{aligned}\mathbb{E}(\exp(\lambda X)) &= \frac{1}{2a}\int_{-a}^{a}\exp(\lambda x) = \frac{1}{2a\lambda}(\exp(a\lambda)+\exp(-a\lambda))\\ &= \sum_{k=0}^{\infty}\frac{(a\lambda)^{2k}}{(2k)!}\le \exp\left(\frac{\lambda^2 a^2}{2}\right).\end{aligned}$$

Exercise 1.1. *A r.v. $X\in[a,b]$ is $\left(\frac{b-a}{2}\right)$-sub-Gaussian.*

A few properties satisfied by sub-Gaussian r.v.s are given below and the proofs are left as an exercise.

Property I: If X is σ-sub-Gaussian, then cX is $|c|\sigma$-sub-Gaussian.

Property II: If X_1, X_2 are σ_1 and σ_2-sub-Gaussian, respectively, then X_1+X_2 is $\sigma_1+\sigma_2$-sub-Gaussian. In addition, if X_1 and X_2 are independent, then X_1+X_2 is $\sqrt{\sigma_1^2+\sigma_2^2}$-sub-Gaussian.

Exercises

Exercise 1.2. *Prove properties I and II that are listed above.*

Exercise 1.3. *True or False? (Justify your answer)*

1. *A r.v. X distributed as $N(\mu,\sigma^2)$ for some $\mu,\sigma>0$ is sub-Gaussian.*
2. *A r.v. X distributed as Unif$[5,10]$ is sub-Gaussian.*
3. *Consider a r.v. X satisfying $\mathbb{E}(\exp(\lambda X))\le\exp\left(\frac{\lambda^2\sigma^2}{2}+\lambda\mu\right)$ for any $\lambda\in\mathbb{R}$. Then, $EX=\mu$.*
4. *For the r.v. X as in the question above, Var$(X)=\sigma^2$.*

Exercise 1.4. *Let $X_1, \ldots, X_n$ denote a sequence of σ-sub-Gaussian random variables, and $Z = \max_{i=1,\ldots,n} X_i$. Show that $\mathbb{E}[Z] \leq \sqrt{2\sigma^2 \log n}$.*

With the background on sub-Gaussian r.v.s, we are now in a position to analyze the tail probability concerning sample mean for the case when the samples $X_1, \ldots, X_n$ are i.i.d. with mean μ and in addition, $X_i - \mu$ is σ-sub-Gaussian for each i. Notice that

$$\hat{\mu}_n - \mu = \sum_{i=1}^{n} \frac{X_i - \mu}{n} \text{ is } \frac{\sigma}{\sqrt{n}}\text{-sub-Gaussian by Properties I and II}$$
$$\Rightarrow \mathbb{P}\left[\hat{\mu}_n \geq \mu + \epsilon\right] \leq \exp\left(\frac{-n\epsilon^2}{2\sigma^2}\right).$$

Exercise 1.5. ***(Hoeffding's inequality)*** *Suppose $X_1, \ldots, X_n$ are i.i.d. samples of r.v. X with mean μ. Also, $X_i \in [a, b]$ for some $a < b$ and $i = 1, \ldots, n$. Prove that*

$$\mathbb{P}\left[\hat{\mu}_n \geq \mu + \epsilon\right] \leq \exp\left(\frac{-2n\epsilon^2}{(b-a)^2}\right).$$

Hint: Use Exercise 1.1.

1.4 Back to analysis of Explore-then-commit

1.4.1 Gap-dependent bound for ETC

In the following, we derive a bound on the regret of ETC that depends on the underlying gaps Δ_i. In the subsequent section, we derive a gap-independent bound, i.e., a bound that is a function of n and holds for any problem instance. Such bounds are worst-case guarantees because they would hold for any problem instance that feeds the sample rewards (and hence, for the problem instance that leads to maximum regret for the algorithm). On the other hand, gap-dependent bounds depend on the problem instance and in particular, help understand how quickly can the algorithm learn in easier problem instance (i.e., ones with large gaps).

Recall that we had the following bound for the regret R_n of ETC:

$$R_n \leq m \sum_{i=1}^{K} \Delta_i + (n - mK) \sum_{i=1}^{K} \Delta_i \mathbb{P}\left[\hat{\mu}_i(mK) - \mu_i - (\hat{\mu}_{a^*}(mK) - \mu_*) \geq \Delta_i\right].$$

Assuming $X_t - \mathbb{E}(X_t)$ is 1-sub-Gaussian for $t = 1, \ldots, n$, we have that

$$\hat{\mu}_i(mK) - \mu_i \text{ is } \frac{1}{\sqrt{m}}\text{-sub-Gaussian for } i = 1, \ldots, K$$
$$\Rightarrow \hat{\mu}_i(mK) - \mu_i - (\hat{\mu}_{a^*}(mK) - \mu_*) \text{ is } \sqrt{\frac{2}{m}}\text{-sub-Gaussian for } i = 1, \ldots, K$$
$$\Rightarrow \mathbb{P}\left[\hat{\mu}_i(mK) - \mu_i - (\hat{\mu}_{a^*}(mK) - \mu_*) \geq \Delta_i\right] \leq \exp\left(\frac{-m\Delta_i^2}{4}\right).$$

Using the final result above in the bound for R_n, we obtain

$$R_n \leq m \sum_{i=1}^{K} \Delta_i + (n - mK) \sum_{i=1}^{K} \Delta_i \ \exp\left(\frac{-m\Delta_i^2}{4}\right).$$

Choosing m optimally is tricky and we illustrate this for the case of two-armed bandit. Clearly, we have

$$R_n \leq m\Delta + (n - mK)\Delta \ \exp\left(\frac{-m\Delta^2}{4}\right).$$

Minimizing the RHS above over m, we obtain $m^* = \left\lceil \frac{4}{\Delta^2} \log\left(\frac{n\Delta^2}{4}\right) \right\rceil$ and for this m^*, the regret R_n turns out to be

$$R_n \leq \Delta + \frac{4}{\Delta}\left(1 + \log\left(\frac{n\Delta^2}{4}\right)\right).$$

While the bound above is nearly optimal, it is obtained when the exploration parameter m is chosen optimally and this choice requires the knowledge of underlying gap Δ. In a bandit framework, the gap information is not available and hence, the requirement is for an adaptive algorithm that balances exploration and exploitation to incur lowest-possible regret, while not assuming knowledge of the underlying problem through the gaps. Before getting there, notice that the regret bound above involves the underlying gaps and we shall refer to such bounds as "gap-dependent bounds". A valid alternative is to derive gap-independent regret bound for ETC, which is the content of the exercise below.

1.4.2 Gap-independent bound for ETC

From the previous section, recall that the ETC algorithm's regret bound is minimized with the exploration parameter m is chosen using the underlying gaps. The resulting regret bound derived there was a function of the underlying gaps. An alternative is to derive a regret bound of the form $R_n \leq Cf(n)$, where C is a problem-independent constant. Such a bound is "gap independent" as it does not involve the problem instance dependent "gap" quantity. We state and prove such a bound for ETC below.

Theorem 1.3. *For the two-armed bandit problem, with stochastic rewards from arms' distribution bounded within* $[0, 1]$*, the regret* R_n *of ETC with* $m = n^{2/3}(\log n)^{1/3}$ *satisfies*

$$R_n \leq cn^{2/3}(\log n)^{1/3},$$

for some universal constant c*.*

Remark 1.1. *The regret upper bound of the order* $O(n^{2/3})$ *for ETC is far from being optimal and in the next section, we shall present the well-known UCB algorithm whose regret is bounded above by* $O(\sqrt{n})$*. Further, we shall establish even later that the UCB upper bound is the best-achievable in the minimax sense – a topic handled in detail under "lower bounds".*

Proof. From the discussion about sub-Gaussianity earlier, recall that the sample mean $\hat{\mu}_m$ formed out of m samples of a bounded (in $[0,1]$) r.v. with mean μ satisfies:

$$\mathbb{P}\left[\hat{\mu}_m \geq \mu + \epsilon\right] \leq \exp\left(\frac{-n\epsilon^2}{2}\right).$$

A straightforward transformation of the bound above yields

$$\mathbb{P}\left[\hat{\mu}_m \geq \mu + \sqrt{\frac{2\log(\frac{1}{\delta})}{m}}\right] \leq \delta.$$

Along similar lines, it is easy to obtain

$$\mathbb{P}\left[\hat{\mu}_m \leq \mu - \sqrt{\frac{2\log(\frac{1}{\delta})}{m}}\right] \leq \delta.$$

We need to pick a δ that is small enough to guarantee that the two tail events above do not affect the regret bound for a horizon n. To simplify the presentation, we shall pick $\delta = \frac{1}{n^2}$. One could choose a better δ and optimize the constants in the regret bound, but the order of n, would not be affected. For $\delta = \frac{1}{n^2}$, the sample means $\hat{\mu}_1(2m)$ and $\hat{\mu}_2(2m)$ corresponding to arms 1 and 2, respectively, satisfies the following:

$$\mathbb{P}\left[|\hat{\mu}_j(2m) - \mu_j| \leq \sqrt{\frac{4\log(n)}{m}}\right] \geq 1 - \frac{2}{n^2}, \text{ for } j = 1, 2.$$

Let E denote the event that the condition inside the probability above holds for both arm 1 and 2. We shall refer to E as the good event, since on E the sample mean is close to the corresponding true mean with high probability, for both arms.

Without loss of generality, assume 1 is the optimal arm. Let $\hat{R}_n = n\mu_1 - \sum_{t=1}^{n} X_t$. Then, we have

$$\begin{aligned} R_n &= \mathbb{E}(\hat{R}_n) \\ &= \mathbb{E}\left(\hat{R}_n \mid E\right)\mathbb{P}\left[E\right] + \mathbb{E}\left(\hat{R}_n \mid E^c\right)\mathbb{P}\left[E^c\right] \\ &\leq \mathbb{E}\left(\hat{R}_n \mid E\right) + n \times \frac{2}{n^2} \end{aligned} \tag{1.10}$$

In the last inequality, we have used the fact that the gap $\Delta_2 \leq 1$ since the rewards are bounded within $[0,1]$ and that $\mathbb{P}\left[E^c\right] \leq \frac{1}{n^2}$.

Regret of ETC, conditioned on the good event E, is simplified as follows: After exploration stage, suppose that ETC chooses arm 2 and not 1, for exploitation. This happens only if $\hat{\mu}_2(2m) > \hat{\mu}_1(2m)$. Since we are conditioning on the good event E, the sample means are not too far from the true means and hence, we have

$$\mu_2 + \sqrt{\frac{4\log(n)}{m}} \geq \hat{\mu}_2(2m) > \hat{\mu}_1(2m) \geq \mu_1 - \sqrt{\frac{4\log(n)}{m}},$$

which implies

$$\mu_1 - \mu_2 \le 2\sqrt{\frac{4\log(n)}{m}}.$$

In other words, if a sub-optimal arm is pulled during exploitation, then there is a good chance that its mean is close to the mean of the optimal arm. Alternatively, the regret incurred in each round during exploitation is bounded above by $2\sqrt{\frac{4\log(n)}{m}}$. So,

$$\mathbb{E}\left(\hat{R}_n \mid E\right) \le m + (n-2m)2\sqrt{\frac{4\log(n)}{m}} \le m + 2n\sqrt{\frac{4\log(n)}{m}}.$$

The first inequality above follows from the fact that the per-round regret during exploration is bounded above by 1 and arm 2 is pulled m times. Since the first term on the RHS of the final inequality above is increasing with m and the other is decreasing with m, a simple way to optimize m is to equate the two terms roughly. This simplification leads to the value $n^{2/3}(\log n)^{1/3}$ for m and for this value of m, the regret on event E turns out to be

$$\mathbb{E}\left(\hat{R}_n \mid E\right) \le 5n^{2/3}(\log n)^{1/3},$$

and the overall regret bound for ETC, from (1.10), simplifies to

$$R_n \le n^{2/3}(\log n)^{1/3} + \frac{2}{n} = cn^{2/3}(\log n)^{1/3},$$

for some problem independent constant c. □

Exercises

Exercise 1.6. *For a K-armed stochastic bandit problem, with $m = \left(\frac{n}{K}\right)^{2/3}(\log n)^{1/3}$, show that the regret R_n of the explore-then-commit (ETC) algorithm satisfies*

$$R_n \le cn^{2/3}(K\log n)^{1/3},$$

for some universal constant c.

Exercise 1.7. *Consider the following bandit algorithm:*

ϵ-greedy algorithm

For $t = 1, 2, \dots, n$, repeat

(1) Let i_t be the arm with the highest sample mean so far, i.e.,
$i_t = \underset{k=1,\dots,K}{\arg\max}\, \hat{\mu}_k(t-1)$, where $\hat{\mu}_k(t-1)$ is the average of rewards obtained from arm k upto time t.

(2) With probability $1-\epsilon_t$, play arm i_t and with probability ϵ_t, play a random arm.

For a two-armed bandit problem, show that the regret R_n incurred by the ϵ-greedy algorithm, with $\epsilon_t = 1/t^{1/3}$, satisfies

$$R_n \leq c n^{2/3} (\log n)^{1/3},$$

for some universal constant c.

Exercise 1.8. *Consider the following game that proceeds over n rounds: In each round $t \in \{1, \ldots, n\}$, you choose either to play or do nothing. If you do nothing, then your reward is $X_t = 0$. If you play, then your reward is $X_t = 1$ with probability p and $X_t = -1$ otherwise. You do not know p and we will assume it could take any value in $[0, 1]$.*

Answer the following:

(i) Formulate the game above as a stochastic bandit problem with horizon n.

(ii) Write down the expression for the regret incurred by any algorithm $\mathcal{A}$.

(iii) Describe an optimal way of choosing actions, i.e., the best algorithm, when p is known.

(iv) For the unknown p case, apply ETC algorithm to the bandit problem formulated above and derive a bound on its regret.

(v) Does exploiting the fact that the reward is zero for "doing nothing" lead to an improved regret bound for ETC?

Exercise 1.9. *Consider a two-armed bandit problem with Bernoulli arms. Let Δ denote the gap. Consider a variant of the ETC algorithm, where the exploration parameter m is chosen based on the samples observed. More precisely, this ETC algorithm plays arms 1 and 2 alternately until some condition is satisfied, and then plays the arm with the highest sample mean, i.e., the arm I_t is chosen as follows:*

$$I_t = \begin{cases} 1\,, & \textit{if } t \textit{ is odd and } t \leq 2M; \\ 2\,, & \textit{if } t \textit{ is even and } t \leq 2M; \\ \textit{arg max}\, \hat{\mu}_i(2M)\,, & \textit{if } t > 2M\,, \end{cases}$$

If the decision at the end of round t is to stop exploring, we set $M = t/2$: M is thus the (random) number of observations that the algorithm will use when committing to one arm for the rest of the rounds.

Design the part of the algorithm that decides about when to stop, and show that the regret of the resulting ETC algorithm satisfies

$$R_n \leq c_1 \left(\Delta + \frac{\log(n)}{\Delta} \right), \textit{ and } R_n(v) \leq c_2 n^{2/3}$$

where c_1, c_2 are universal constants. Note that the decision to stop at any time instant t can be based on the sample rewards seen so far, and should not assume knowledge of the underlying means.

Exercise 1.10. ***(Simulation experiment)***
Consider a two-armed bandit problem, where each arm's distribution is Bernoulli. Consider the following three problem variants, with respective Bernoulli distribution parameters specified for each arm:

Problem	*Arm 1*	*Arm 2*
P1	0.9	0.6
P2	0.9	0.8
P3	0.55	0.45

Write a program (in your favorite language) to simulate each of the above bandit problems. In particular, do the following for each problem instance:

(i) *Choose the horizon n as 10000.*

(ii) *For each algorithm, repeat the experiment 100 times.*

(iii) *Store the number of times an algorithm plays the optimal arm, for each round $t = 1, \ldots, n$.*

(iv) *Store the regret in each round $m = 1, \ldots, n$.*

(v) *Plot the percentage of optimal arm played and regret against the rounds $t = 1, \ldots, n$.*

(vi) *For each plot, add standard error bars.*

Do the above for the following bandit algorithms:

- *The explore-then-commit (ETC) algorithm with exploration parameter m chosen optimally so that the gap-dependent regret is minimum (this choice for m would require information about underlying gap).*
- *The ETC algorithm with a heuristic choice for exploration parameter m. Try different values for m and summarize your findings, say by tabulating regret for different m.*

Interpret the numerical results and submit your conclusions. In particular, discuss the following:

(a) *Explain the results obtained for ETC with optimal m and correlate the results to the theoretical findings.*

(b) *Explain the results obtained for ETC with a heuristic choice for m. In particular, how does ETC with a m that is far from the optimal, perform?*

1.5 Upper confidence bound (UCB) algorithm

1.5.1 Basic algorithm

Given i.i.d. samples $X_1, \ldots, X_m$ of a r.v X with mean μ and assuming 1-sub-Gaussianity of $X_i - \mu$, for all i, we have that

$$\mathbb{P}\left[\hat{\mu}_m \geq \mu + \epsilon\right] \leq \exp\left(\frac{-m\epsilon^2}{2}\right) \text{ and } \mathbb{P}\left[\hat{\mu}_m \leq \mu - \epsilon\right] \leq \exp\left(\frac{-m\epsilon^2}{2}\right).$$

Or equivalently, for any $\delta \in (0,1)$,

$$\mathbb{P}\left[\mu \in [\hat{\mu}_m - \sqrt{\frac{2\log(\frac{1}{\delta})}{m}}, \hat{\mu}_m + \sqrt{\frac{2\log(\frac{1}{\delta})}{m}}]\right] \geq 1 - 2\delta.$$

In this section, we describe the UCB algorithm that balances exploration and exploitation in each round $t = 1, \ldots, n$. A vital ingredient in this balancing act is the concentration inequality given above. An important question here is, what should be the δ value, so that the UCB algorithm can ignore the errors in estimation (i.e., the true means falling outside the confidence intervals) and not suffer linear regret. The finite sample analysis provided by Auer et al. [2002] chose to set $\delta = \frac{1}{t^4}$ and this is good enough to guarantee a sub-linear regret. For this value of δ, at any round t of UCB, we have the following high-confidence guarantee for any arm $k \in \{1, \ldots, K\}$:

$$\mathbb{P}\left[\mu_k \in [\hat{\mu}_k(t-1) - \sqrt{\frac{8\log t}{T_k(t-1)}}, \hat{\mu}_k(t-1) + \sqrt{\frac{8\log t}{T_k(t-1)}}]\right] \geq 1 - \frac{2}{t^4}.$$

In the above, the quantity $\hat{\mu}_k(t-1)$ denotes the sample mean of rewards seen from arm k so far and $T_k(t-1)$ samples from arm k's distribution are used to form this sample mean.

We are now ready to present the UCB algorithm.

UCB algorithm

Initialization: Play each arm once,

For $t = K+1, \ldots, n$**, repeat**

(1) Play arm $I_t = \arg\max_{k=1,\ldots,K} \mathrm{UCB}_t(k)$, where
$\mathrm{UCB}_t(k) \triangleq \hat{\mu}_k(t-1) + \sqrt{\frac{8\log t}{T_k(t-1)}}$.

(2) Observe sample X_t from the distribution P_{I_t} corresponding to the arm I_t.

Notice that the confidence estimates are applicable only if there is at least one sample for any arm and hence, in the initialization phase, UCB pulls each arm once. Further, UCB is an *anytime* algorithm, since the UCB index for any arm depends only on the round index t and does not require the horizon n.

Intuitively, the first term in the UCB index for any arm is geared towards exploitation (i.e., if the sample mean of an arm is high, then the UCB index is high and hence the algorithm is likely to play this arm), while the second term is to do with exploration, since an arm that has not been played often would get a higher UCB index through the second term. For a rigorous justification for the explicit form used in the UCB index's second term, we now bound the number of times the UCB algorithm pulls a suboptimal arm.

1.5.2 Regret analysis

Bounding the number of pulls of a suboptimal arm

Let 1 denote the optimal arm, without loss of generality. Fix a round $t \in \{1, \ldots, n\}$ and suppose that a sub-optimal arm k is pulled in this round. Then, we have

$$\hat{\mu}_k(t-1) + \sqrt{\frac{8 \log t}{T_k(t-1)}} \geq \hat{\mu}_1(t-1) + \sqrt{\frac{8 \log t}{T_1(t-1)}}.$$

The UCB-value of arm k can be larger than that of 1 *only if* one of the following three conditions holds:

(1) μ_1 is outside the confidence interval

$$\hat{\mu}_{1,T_1(t-1)} \leq \mu_1 - \sqrt{\frac{8 \log t}{T_1(t-1)}}, \tag{1.11}$$

(2) μ_k is outside the confidence interval

$$\hat{\mu}_{k,T_k(t-1)} \geq \mu_k + \sqrt{\frac{8 \log t}{T_k(t-1)}}, \tag{1.12}$$

(3) Gap Δ_k is small If we negate the two conditions above and use the fact $UCB_t(k) \geq UCB_t(1)$, then we obtain

$$\begin{aligned}
&\mu_k + 2\sqrt{\frac{8 \log t}{T_k(t-1)}} \geq \hat{\mu}_{k,T_k(t-1)} + \sqrt{\frac{8 \log t}{T_k(t-1)}} \geq \hat{\mu}_{1,T_1(t-1)} + \sqrt{\frac{8 \log t}{T_1(t-1)}} > \mu_1 \\
\Rightarrow \quad &\Delta_k < 2\sqrt{\frac{8 \log t}{T_k(t-1)}} \text{ or } T_k(t-1) \leq \frac{32 \log t}{\Delta_k^2} \qquad (1.13)
\end{aligned}$$

Let $u = \dfrac{32 \log n}{\Delta_k^2} + 1$. When $T_k(t-1) \geq u$, i.e., when the condition in (1.13) does not hold, then either (i) arm k is not pulled at time m, or (ii) (1.11) or (1.12) occurs. Thus, we have

$$
\begin{aligned}
T_k(n) &= 1 + \sum_{t=K+1}^{n} \mathbb{I}\{I_t = k\} \\
&\leq u + \sum_{t=u+1}^{n} \mathbb{I}\{I_t = k; T_k(t) \geq u\} \\
&\leq u + \sum_{t=u+1}^{n} \mathbb{I}\left\{\hat{\mu}_{k,T_k(t-1)} + \sqrt{\frac{8\log t}{T_k(t-1)}} \geq \hat{\mu}_{1,T_1(t-1)} + \sqrt{\frac{8\log t}{T_1(t-1)}}; T_k(t-1) \geq u\right\} \\
&\leq u + \sum_{t=1}^{\infty}\sum_{s=1}^{t-1}\sum_{s_k=u}^{t-1} \mathbb{I}\left\{\hat{\mu}_{k,s_k} + \sqrt{\frac{8\log t}{s_k}} \geq \hat{\mu}_{1,s} + \sqrt{\frac{8\log t}{s}}\right\} \\
&\leq u + \sum_{t=1}^{\infty}\sum_{s=1}^{t-1}\sum_{s_k=u}^{t-1} \mathbb{I}\left\{\left(\hat{\mu}_{1,s} \leq \mu_1 - \sqrt{\frac{8\log t}{s}}\right)\right. \\
&\qquad\qquad \left. \text{or } \left(\hat{\mu}_{k,s_k} \geq \mu_k + \sqrt{\frac{8\log t}{s_k}}\right) \text{ occurs}\right\}.
\end{aligned}
$$

From the discussion earlier on concentration inequalities, we can upper bound the probability of occurence of each of the two events inside the indicator on the RHS of the final display above as follows:

$$
\mathbb{P}\left[\hat{\mu}_{1,s} \leq \mu_1 - \sqrt{\frac{8\log t}{s}}\right] \leq \frac{1}{t^4} \text{ and } \mathbb{P}\left[\hat{\mu}_{k,s_k} \geq \mu_k + \sqrt{\frac{8\log t}{s_k}}\right] \leq \frac{1}{t^4}.
$$

Plugging the bounds on the events above and taking expectations on $T_k(n)$ related inequality above, we obtain

$$
\begin{aligned}
\mathbb{E}[T_k(n)] &\leq u + \sum_{t=1}^{\infty}\sum_{s=1}^{t-1}\sum_{s_k=u}^{t-1} \frac{2}{t^4} \\
&\leq u + 2\sum_{t=1}^{\infty} \frac{1}{t^2} \leq \frac{32\log n}{\Delta_k^2} + \left(1 + \frac{\pi^2}{3}\right).
\end{aligned}
$$

The preceding analysis together with the fact that $R_n = \sum_{k=1}^{K} \Delta_k \mathbb{E}[T_k(n)]$ leads to the following regret bound:

Theorem 1.4. *For a K-armed stochastic bandit problem where the stochastic rewards from the arms' distributions are bounded within $[0,1]$, the regret R_n of UCB is satisfies*

$$
R_n \leq \sum_{\{k:\Delta_k > 0\}} \frac{32\log n}{\Delta_k} + K\left(1 + \frac{\pi^2}{3}\right)
$$

.

Gap-independent regret bound

We arrive at a gap-independent bound on UCB algorithm's regret as follows:

$$\begin{aligned}
R_n &= \sum_k \Delta_k \, \mathbb{E}[T_k(n)] \\
&= \sum_k \left(\Delta_k \sqrt{\mathbb{E}[T_k(n)]}\right)\left(\sqrt{\mathbb{E}[T_k(n)]}\right) \\
&\leq \left(\sum_k \Delta_k^2 \, \mathbb{E}[T_k(n)]\right)^{\frac{1}{2}} \left(\sum_k \mathbb{E}[T_k(n)]\right)^{\frac{1}{2}} && \text{(Cauchy-Schwarz inequality)} \\
&\leq \left(K\left(32\log n + \frac{\pi^2}{3} + 1\right)\right)^{\frac{1}{2}} \sqrt{n}, \\
&= \sqrt{Kn\left(32\log n + \frac{\pi^2}{3} + 1\right)}.
\end{aligned}$$

where the final inequality follows from the fact that $\sum_k \mathbb{E}[T_k(n)] = n$, together with the following bound on the number of pulls of a a suboptimal arms, i.e., k with $\Delta_k > 0$:

$$\mathbb{E}[T_k(n)] \leq \frac{32\log n}{\Delta_k^2} + \left(1 + \frac{\pi^2}{3}\right).$$

Thus, we have obtained a $\tilde{O}(\sqrt{n})$ regret bound for UCB and this is clearly better than the corresponding $\tilde{O}(n^{2/3})$ regret bound for ETC algorithm. A natural question that arises is if $\tilde{O}(\sqrt{n})$ is the best achievable regret bound and we shall arrive at a positive answer, when we discuss lower bounds on regret in the next section.

Exercises

Exercise 1.11. *Consider a stochastic K-armed bandit problem where the rewards of arms' distributions are bounded within $[\frac{1}{2}, \frac{1+\epsilon}{2}]$ for some $\epsilon \in (0,1)$. Construct a variant of UCB algorithm that uses the knowledge of ϵ. Derive gap-dependent and gap-independent regret bounds for this UCB variant and discuss their dependence on ϵ.*

Hint: Use Hoeffding's inequality.

Exercise 1.12. *For each of the two-armed bandit algorithms listed below, answer if they achieve sub-linear regret. An intuitive justification will suffice. Notation: For $i = 1, 2$, let $\hat{\mu}_i(t)$ denote the sample mean of arm i from the rewards seen up to time t.*

(a) Play arm $I_t =$ arg max$_{i=1,2}\, \hat{\mu}_i(t-1)$.

(b) Fix two sequences $A_1 = \{1, 2, 4, 8, 16, \ldots\}$ and $A_2 = \{3, 9, 27, 81, \ldots\}$. If $t \in A_i$, then play arm i, else play $I_t =$ arg max$_{i=1,2}\, \hat{\mu}_i(t-1)$.

(c) Suppose the means are in the set $S = \{\mu, \mu - \epsilon\}$ and the bandit algorithm is aware of the set S. Consider the following algorithm: At time t, if $\max \hat{\mu}_i(t-1) > \mu - \epsilon/2$, then pull the arm that has the maximum sample mean. Otherwise pull both arms once.

Exercise 1.13. *Consider a two-armed bandit problem with Bernoulli reward distributions. Let* 1 *be the best arm without loss of generality, and let* $\Delta = \mu_1 - \mu_2$ *denote the gap. Show that the UCB algorithm satisfies the following:*

$$\mathbb{P}\left[\hat{R}_n > \Delta\left(1 + \frac{32}{\Delta^2}\log n\right)\right] \leq \frac{a}{(\log n)^b},$$

where $b > 0$ *is a problem independent constant and* $\hat{R}_n = \Delta T_2(n)$, *with* $T_2(n)$ *denoting the number of times arm* 2 *was pulled upto time* n.

Hint: For any $\tau \in \mathbb{R}$, *any integer* $u > 1$ *and any sub-optimal arm* k, *we have*

$$\mathbb{P}\left[T_2(n) > u\right] \leq \sum_{t=u+1}^{n} \mathbb{P}\left[\hat{\mu}_{2,u} + \sqrt{\frac{8\log t}{u}} > \tau\right] + \sum_{s=1}^{n-u} \mathbb{P}\left[\hat{\mu}_{1,s} + \sqrt{\frac{8\log(u+s)}{u}} \leq \tau\right].$$

In the above, $\hat{\mu}_{2,u}$ *is the sample mean of* u *i.i.d. samples from the second arm's distribution.*

Exercise 1.14. *Consider a two-armed Bernoulli bandit problem. Suppose that the underlying means are in the set* $\{\theta, 1-\theta\}$ *and the bandit algorithm is aware of* θ. *Does there exist an algorithm* $\mathcal{A}$ *that satisfies*

$$R_n(\mathcal{A}) \leq \frac{c}{2\theta - 1},$$

where $R_n(\mathcal{A})$ *is the expected regret with horizon* n *and* c *is a problem-independent constant. If yes, describe the algorithm and derive the regret bound.*

Hint: Try the algorithm in part (c) of Exercise (1.12) *above, or the following variant that uses upper confidence bounds: If the UCB of an arm is better than the optimal mean, play that arm, else alternate between the arms.*

1.6 A brief tour of information theory

Before presenting regret lower bounds, we briefly cover the necessary information theory concepts. In the following, we assume that the underlying random variables are discrete and leave it to the reader to fill in the necessary details for the continuous extension.

1.6.1 Entropy

Definition 1.2. *Consider a discrete r.v.* X *taking values in the set* $\mathcal{X}$ *with p.m.f.* p. *Then, the entropy* $H(X)$ *is defined as*

$$H(X) = -\sum_{x \in \mathcal{X}} p(x)\log p(x),$$

where the log *is to base* 2.

It is easy to see that $H(X) \geq 0$ for any X, since $\log p(x) \leq 0$ for $p(x) \in [0,1]$.

The entropy of a Bernoulli r.v. X with parameter p is $H(X) = -p\log p - (1-p)\log(1-p)$. Plotting $H(X)$ as a function of p, it is easy to infer that $H(X)$ is maximized at $p = 1/2$, $H(X) = 0$ at $p = 0$ and $p = 1$.

The notion of entropy came has roots in information theory, as it gives the expected number of bits necessary to encode a random signal (= a random variable). We illustrate this interpretation through the following r.v.:

$$X = \begin{cases} a & \text{w.p. } 1/2, \\ b & \text{w.p. } 1/4, \\ c & \text{w.p. } 1/8, \\ d & \text{w.p. } 1/8. \end{cases}$$

If one were to design a sequence of binary questions to infer the value of the r.v. X and ask the minimum number of questions in expectation, then it would serve him/her to start with "Is $X = a$?" rather than start with "Is $X = d$?". Now using the pmf of X given above, the expected number of questions asked is $1 \times \frac{1}{2} + 2 \times \frac{1}{4} + 31 \times \frac{1}{8} + 3 \times \frac{1}{8} = \frac{7}{4}$. It is not a coincidence that $H(X)$ turns out to be $\frac{7}{4}$ for this r.v.

An equivalent interpretation is the following: Suppose that the value a is represented by the code "1", b by "01", c by "001" and d by "000". Then, the average code length, assuming that the values a, b, c, d occur with probabilities given above, then the average code length turns out to be the same as $H(X)$.

Definition 1.3. *The joint entropy $H(X, Y)$ of r.v. pair (X, Y) with joint pmf $p(x, y)$ is defined as*

$$H(X, Y) = -\sum_x \sum_y p(x, y) \log p(x, y).$$

Definition 1.4. *The conditional entropy $H(Y \mid X)$, assuming the r.v. pair (X, Y) has joint pmf $p(x, y)$, is defined as*

$$\begin{aligned} H(Y \mid X) &= \sum_x p(x) H(Y \mid X = x) \\ &= -\sum_x p(x) \sum_y p(y \mid x) \log p(y \mid x) \\ &= -\sum_x \sum_y p(x, y) \log p(y \mid x). \end{aligned}$$

Theorem 1.5. $H(X, Y) = H(X) + H(Y \mid X)$.

Proof. Follows by using the definition of $H(X, Y)$ followed by a separation of terms using $p(x, y) = p(x)p(y \mid x)$ to obtain $H(X)$ and $H(Y \mid X)$. □

We now are ready to define the concept of KL-divergence between two probability distributions, a notion that serves us well in obtaining regret lower bounds in a bandit framework.

1.6.2 KL-divergence (aka relative entropy)

Definition 1.5. *The KL-divergence $D(p, q)$ between two pmfs p and q is defined as*

$$D(p, q) = \sum_x p(x) \log \left(\frac{p(x)}{q(x)} \right),$$

where $0 \log \frac{0}{q} = 0$ *and* $p \log \frac{p}{0} = \infty$.

Example 1.3. *Let* p *and* q *be pmfs of Bernoulli r.v.s with parameters* α *and* β*, respectively. Then,*

$$D(p, q) = \alpha \log \frac{\alpha}{\beta} + (1 - \alpha) \log \frac{1 - \alpha}{1 - \beta}.$$

Plugging in values $1/4$ *and* $1/2$ *for* α *and* β*, it is easy to see that* $D(p, q)$ *is not equal to* $D(q, p)$.

KL-divergence is not a metric because it is not symmetric, as shown in the above example. Moreover, KL-divergence does not satisfy the triangle inequality. However, KL-divergence is non-negative and zero if and only if the probability distributions are the same – a claim made precise below.

Lemma 1.6. *The KL-divergence* $D(p, q)$ *between two pmfs* p *and* q *is non-negative and equals zero if and only if* $p(x) = q(x), \forall x$.

Proof. Let $A = \{x \mid p(x) > 0\}$ be the support of p. Then, using Jensen's inequality for the concave $\log$ function, we have

$$\begin{aligned}
-D(p, q) &= -\sum_{x \in A} p(x) \log \left(\frac{p(x)}{q(x)} \right) \\
&= \sum_{x \in A} p(x) \log \left(\frac{q(x)}{p(x)} \right) \\
&\leq \log \left(\sum_{x \in A} p(x) \frac{p(x)}{q(x)} \right) \\
&= \log \left(\sum_{x \in A} q(x) \right) \leq \log \left(\sum_{x} q(x) \right) \\
&= \log 1 = 0,
\end{aligned}$$

which proves the first part of the claim. For the second part, observe that $\log$ is stricly concave and hence, equality holds in Jensen's if and only if $\frac{p(x)}{q(x)} = 1, \forall x$. □

Definition 1.6. *The conditional KL-divergence between two pmfs* p *and* q *is defined as*

$$D(p(y \mid x), q(y \mid x)) = \sum_{x} p(x) \sum_{y} p(y \mid x) \log \frac{p(y \mid x)}{q(y \mid x)}.$$

Lemma 1.7. *(Chain rule)*

$$D(p(x, y), q(x, y)) = D(p(x), q(x)) + D(p(y \mid x), q(y \mid x)).$$

In addition, if x *and* y *are independent, then*

$$D(p(x, y), q(x, y)) = D(p(x), q(x)) + D(p(y), q(y)).$$

Proof.

$$\begin{aligned}
D(p(x,y), q(x,y)) &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{q(x,y)} \\
&= \sum_x \sum_y p(x,y) \log \frac{p(x)}{q(x)} + \sum_x \sum_y p(x,y) \log \frac{p(y \mid x)}{q(y \mid x)} \\
&= D(p(x), q(x)) + D(p(y \mid x), q(y \mid x)).
\end{aligned}$$

□

1.6.3 Pinsker's inequality

Lemma 1.8. *(Pinsker's inequality) Given two pmfs p and q, for any event A, we have*

$$2(p(A) - q(A))^2 \leq D(p,q).$$

Proof. Fix an event A. Then, we have

$$\sum_x p(x) \log \frac{p(x)}{q(x)} \geq p(A) \log \frac{p(A)}{q(A)}. \tag{1.14}$$

The proof of the claim above is as follows: Letting $p_A(x) = \frac{p(x)}{p(A)}$ and $q_A(x) = \frac{q(x)}{q(A)}$, we have

$$\begin{aligned}
\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} &= p(A) \sum_{x \in A} p_A(x) \log \frac{p(A) p_A(x)}{q(A) q_A(x)} \\
&= p(A) \log \frac{p(A)}{q(A)} \sum_{x \in A} p_A(x) + p(A) \sum_{x \in A} p_A(x) \log \frac{p_A(x)}{q_A(x)} \\
&\geq p(A) \log \frac{p(A)}{q(A)},
\end{aligned}$$

where the last inequality follows from the fact that $\sum_x p_A(x) \log \frac{p_A(x)}{q_A(x)} = D(p_A, q_A) \geq 0$ and $\sum_x p_A(x) = 1$.

Letting $\alpha = p(A)$ and $\beta = q(A)$ and using (1.14), we have

$$\begin{aligned}
D(p,q) &\geq \alpha \log \frac{\alpha}{\beta} + (1-\alpha) \log \frac{1-\alpha}{1-\beta} \\
&= \int_\alpha^\beta \left(\frac{-\alpha}{x} + \frac{1-\alpha}{1-x} \right) dx \\
&= \int_\alpha^\beta \left(\frac{x-\alpha}{x(1-x)} \right) dx \geq \int_\alpha^\beta \frac{x-\alpha}{1/4} dx \qquad \text{since } x(1-x) \leq 1/4, \\
&= 2(\alpha - \beta)^2.
\end{aligned}$$

Hence proved. □

Lemma 1.9. *(Pinsker's inequality: a variant)*

Given two pmfs p and q, for any event A, we have

$$P(A) + Q(A^c) \geq \frac{1}{2}\exp(-D(p,q)),$$

where $P(A)$ (resp. $Q(A^c)$) is shorthand for $\sum_{x\in A} p(x)$ (resp. $\sum_{x\in A^c} q(x)$).

Proof. Notice that

$$\begin{aligned}\sum_x \min(p(x),q(x)) &= \sum_{x\in A}\min(p(x),q(x)) + \sum_{x\in A^c}\min(p(x),q(x))\\ &\leq \sum_{x\in A} p(x) + \sum_{x\in A^c} q(x) = P(A) + Q(A^c).\end{aligned}$$

So, it is enough to prove a lower bound on $\sum_{x\in A}\min(p(x),q(x))$. We claim that

$$\sum_x \min(p(x),q(x)) \geq \frac{1}{2}\left(\sum_x \sqrt{p(x)q(x)}\right)^2.$$

The inequality above holds because

$$\begin{aligned}\left(\sum_x \sqrt{p(x)q(x)}\right)^2 &= \left(\sum_x \sqrt{\min(p(x),q(x))\max(p(x),q(x))}\right)^2\\ &\leq \left(\sum_x \min(p(x),q(x))\right)\left(\sum_x \max(p(x),q(x))\right)\\ &\leq 2\sum_x \min(p(x),q(x)),\end{aligned}$$

where the last inequality holds because

$$\sum_x \max(p(x),q(x)) = \sum_x (p(x)+q(x)-\min(p(x),q(x))) \leq 2 - \sum_x \min(p(x),q(x)) \leq 2.$$

Now, we have

$$\begin{aligned}\left(\sum_x \sqrt{p(x)q(x)}\right)^2 &= \exp\left(2\log\left(\sum_x \sqrt{p(x)q(x)}\right)\right)\\ &= \exp\left(2\log\left(\sum_x p(x)\sqrt{\frac{q(x)}{p(x)}}\right)\right)\\ &\geq \exp\left(2\left(\sum_x p(x)\log\sqrt{\frac{q(x)}{p(x)}}\right)\right) && \text{(Jensen's inequality)}\\ &= \exp\left(\sum_x p(x)\log\frac{q(x)}{p(x)}\right)\\ &= \exp\left(-D(p,q)\right).\end{aligned}$$

□

For establishing lower bounds for regret minimization setting of this chapter as well as best arm identification setting treated in the next chapter, it would be handy to have the KL-divergence between two Bernoulli r.v.s bounded above and the following claim makes this bound explicit.

Lemma 1.10. *For some $0 < \Delta < 1/2$, let p, q and r correspond to the pmfs of Bernoulli r.v.s with parameters $\frac{1}{2}$, $\frac{1+\Delta}{2}$ and $\frac{1-\Delta}{2}$, respectively. Then,*

$$D(p,q) \leq \Delta^2, D(q,p) \leq 2\Delta^2, D(p,r)) \leq \Delta^2 \text{ and } D(r,q)) \leq 4\Delta^2.$$

Proof.

$$\begin{aligned} D(p,q) &= \frac{1}{2}\log\left(\frac{1}{1+\Delta}\right) + \frac{1}{2}\log\left(\frac{1}{1-\Delta}\right) \\ &= -\frac{1}{2}\log\left(1-\Delta^2\right) \\ &\leq -\frac{1}{2}(-2\Delta^2) = \Delta^2, \qquad (\log(1-\Delta^2) \geq -2\Delta^2 \text{ for } \Delta^2 \leq \frac{1}{2}). \end{aligned}$$

Along similar lines, it is easy to see that $D(p,r) \leq \Delta^2$.

$$\begin{aligned} D(q,p) &= \frac{1+\Delta}{2}\log\left(1+\Delta\right) + \frac{1-\Delta}{2}\log\left(1-\Delta\right) \\ &= \frac{1}{2}\log\left(1-\Delta^2\right) + \frac{\Delta}{2}\log\left(\frac{1+\Delta}{1-\Delta}\right) \\ &\leq \frac{\Delta}{2}\log\left(1+\frac{2\Delta}{1-\Delta}\right), \qquad (\log(1-\Delta^2) < 0) \\ &\leq \frac{\Delta}{2}\frac{2\Delta}{1-\Delta} \leq 2\Delta^2. \end{aligned}$$

Finally,

$$\begin{aligned} D(r,q) &= \frac{1-\Delta}{2}\log\left(\frac{1-\Delta}{1+\Delta}\right) + \frac{1+\Delta}{2}\log\left(\frac{1+\Delta}{1-\Delta}\right) \\ &= \Delta\log\left(\frac{1-\Delta}{1+\Delta}\right) \\ &\leq \Delta\log\left(1+\frac{2\Delta}{1-\Delta}\right) \\ &\leq \Delta\frac{2\Delta}{1-\Delta} \leq 4\Delta^2 \qquad \text{(Since } \Delta \leq 1/2\text{)}. \end{aligned}$$

□

Exercises

Exercise 1.15. *For distributions P and Q of a continuous random variable, the KL-divergence is defined to be the integral:*

$$D(P,Q) = \int p(x)\log\left(\frac{p(x)}{q(x)}\right)dx,$$

where p and q denote the densities of P and Q, respectively.

Answer the following:

(a) Suppose that P and Q correspond to univariate Gaussian distributions with means μ_1, μ_2, and variances σ_1^2, σ_2^2, respectively. Show that

$$D(P, Q) = \frac{1}{2}\left(\log \frac{\sigma_2^2}{\sigma_1^2} + \frac{\sigma_1^2}{\sigma_2^2} - 1\right) + \frac{(\mu_1 - \mu_2)^2}{2\sigma_2^2}.$$

(b) Suppose that P and Q correspond to bivariate Gaussian distributions with zero mean and covariance matrices $\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ and $\begin{bmatrix} 1 & \rho^2 \\ \rho^2 & 1 \end{bmatrix}$, where $\rho \in (0, 1)$. Calculate $D(P, Q)$, upper bound it using the simplest possible function of ρ.

Exercise 1.16.

Suppose there are two coins. The first is a fair coin, while the second one is biased (i.e., it falls heads with probability $\frac{3}{4}$). Suppose n sample outcomes $X_1, \ldots, X_n$ are generated using one of the two coins and an algorithm, say $\mathcal{A}$, uses these samples to identify the source coin. Let $\hat{I}_n$ denote the index that the algorithm $\mathcal{A}$ returns as its estimate of the source coin. Let P_v (resp. $P_{v'}$) denote the law of the observed samples $(X_1, \ldots, X_n)$, when the underlying source is the fair (resp. biased) coin.

If $n < 4 \log 2$, then show that no algorithm can ensure

$$\max(P_v(\hat{I}_n = 2), P_{v'}(\hat{I}_n = 1)) \leq 0.22.$$

Hint: Use Pinsker's inequality.

1.7 Regret lower bounds

1.7.1 Worst-case lower bounds

Consider two bandit problems with Bernoulli distributions for the arms, with means given in the following table: For some $\Delta > 0$ to be specified later,

	Arm 1	Arm 2
Problem 1	$\frac{1}{2}$	$\frac{1-\Delta}{2}$
Problem 2	$\frac{1}{2}$	$\frac{1+\Delta}{2}$

Let p_1, p_2 and p_2' denote the pmfs of Bernoulli r.v.s with means $\frac{1}{2}$, $\frac{1-\Delta}{2}$ and $\frac{1+\Delta}{2}$, respectively. Let v_t (resp. v_t') denote the product distribution $(p_1, p_2)^{\otimes t}$ (resp. $(p_1, p_2')^{\otimes t}$), for $t = 1, \ldots, n$. The distributions v_t and v_t' govern the sample rewards up to time t for any bandit algorithm.

Let $\mathbb{P}_{v_t}$ (resp. $\mathbb{P}_{v_t'}$) denote the probability law with underlying distribution v_t (resp. v_t') and with the arms chosen by the algorithm $\mathcal{A}$. Further, let $R_n(v)$ (resp. $R_n(v')$) denote the regret incurred by the algorithm when the underlying problem instance is 1 (resp. 2), i.e., when the sample rewards are generated from v_n (resp. v_n'). For the sake of notational convenience, we have suppressed the dependence of regret $R_n(v)$ and $\mathbb{P}_v$ on the algorithm $\mathcal{A}$. Then, we have the following claim:

Theorem 1.11. *For any bandit algorithm $\mathcal{A}$, we have*

$$\max(R_n(v), R_n(v')) \geq \frac{1}{16\Delta} \log\left(n\Delta^2\right).$$

Proof. Notice that, on problem 1, the bandit algorithm incurs a regret of $\Delta/2$ if it pulls arm 2 and the number of times it pulls arm 2 is $T_2(n)$ leading to

$$\begin{aligned} R_n(v) &\geq \frac{\Delta}{2}\mathbb{E}_{v_n} T_2(n) \\ \Rightarrow \max\left(R_n(v), R_n(v')\right) &\geq R_n(v) \geq \frac{\Delta}{2}\mathbb{E}_v T_2(n). \end{aligned} \tag{1.15}$$

Since the max is greater than the average, we obtain

$$\begin{aligned} \max\left(R_n(v), R_n(v')\right) &\geq \frac{1}{2}\left(R_n(v) + R_n(v')\right) \\ &= \frac{\Delta}{4}\sum_{t=1}^{n}\Big(\mathbb{P}_{v_t}(I_t = 2) + \mathbb{P}_{v_t'}(I_t = 1)\Big) \\ &\geq \frac{\Delta}{8}\sum_{t=1}^{n}\exp(-D(P_{v_t}, \mathbb{P}_{v_t'})) \qquad \text{(Pinsker's inequality)} \\ &= \frac{\Delta}{8}\sum_{t=1}^{n}\exp(-4\mathbb{E}_{v_t}T_2(t)\Delta^2) \qquad (1.16) \\ &\geq \frac{n\Delta}{8}\exp(-4\mathbb{E}_{v_n}T_2(n)\Delta^2). \end{aligned}$$

In the above, we have used the following fact in inequality (1.16):

$$\begin{aligned} D(P_{v_t}, \mathbb{P}_{v_t'}) &= \sum_{s=1}^{\mathbb{E}_v T_2(n)} D((p_1, p_2), (p_1, p_2')) \\ &= \sum_{s=1}^{\mathbb{E}_v T_2(n)} D(p_2, p_2') \leq \mathbb{E}_v T_2(n) 4\Delta^2, \qquad \text{(Lemma 1.10)} \end{aligned}$$

where we used the fact that the KL-divergence, between the arms distributions under problem 1 and 2, is not zero only when the bandit algorithm pulls arm 2 in a certain round and also that the underlying distribution is i.i.d. in time.

Combining (1.15) and (1.16), we have

$$\begin{aligned} \max\left(R_n(v), R_n(v')\right) &\geq \frac{\Delta}{4}\left(\mathbb{E}_{v_n}T_2(n) + \frac{n}{4}\exp(-4\Delta^2\mathbb{E}_\nu T_2(n))\right) \\ &\geq \min_{x\in[0,n]}\frac{\Delta}{4}\left(x + \frac{n}{4}\exp(-4\Delta^2 x)\right) \\ &\geq \frac{\log(n\Delta^2)}{16\Delta}. \end{aligned}$$

Hence proved. $\square$

Plugging in a value of $\frac{2}{\sqrt{n}}$ for Δ, we have the following gap-independent regret lower bound:

Corollary 1.1. *For any bandit algorithm,*

$$\max\left(R_n(v), R_n(v')\right) \geq \frac{1}{32}\sqrt{n}.$$

We now generalize the result in Corollary 1.1 to a setting involving more than two arms.

Theorem 1.12. *For any bandit algorithm $\mathcal{A}$, there exists a problem instance v such that*

$$R_n(v) \geq c\sqrt{Kn},$$

where $R_n(v)$ is the regret[2] *incurred by algorithm $\mathcal{A}$ on problem v and c is a universal constant.*

Proof. As in the proof for the case of two-armed bandit, we consider two bandit problem instances v and v' such that a bandit algorithm $\mathcal{A}$ that does well on v would end up suffering high regret on v' and vice-versa. Let v correspond to a Bernoulli-armed bandit problem instance, with the underlying means given by $p_1 \sim \text{Ber}(\frac{1}{2})$ and $p_i \sim \text{Ber}(\frac{1-\Delta}{2})$, for $i \neq 1$ and for some $\Delta > 0$ to be specified later. Let $\mathbb{P}_v$ denote the probability law of the rewards when algorithm $\mathcal{A}$ is run on problem v (for the sake of notational convenience, we have suppressed the dependence on $\mathbb{P}_v$ on $\mathcal{A}$. Let i be the arm that is pulled least (in expectation) by $\mathcal{A}$ on v, i.e.,

$$i = \underset{j=2,\ldots,K}{\arg\min}\, \mathbb{E}_v(T_j(n)).$$

In the above, $\mathbb{E}_v$ is the expectation under $\mathbb{P}_v$ and $T_j(n)$ is the number of times arm j is pulled up to time n. Define the problem instance v' as follows: For $j \neq i$, $p'_j = p_j$, i.e., the arms' distributions are unchanged, while $p_j \sim \text{Ber}(\frac{1+\Delta}{2})$. Let v' refer to the problem where arm i's distribution is modified as described before and let $\mathbb{P}_{v'}$ denote the probability law of the sample rewards when algorithm $\mathcal{A}$ is run on problem v'. Then, it is easy to see that

$$R_n(v) \geq \mathbb{P}_v(T_1(n) \leq \frac{n}{2})\frac{n\Delta}{4}.$$

The inequality above holds because an algorithm that pulls the optimal arm 1 on problem v less than $n/2$ times would suffer at least a regret of $n/2 \times \Delta/2$. Along similar lines, it can be argued that

$$R_n(v') \geq \mathbb{P}_{v'}(T_1(n) > \frac{n}{2})\frac{n\Delta}{4}.$$

Invoking Pinsker's inequality with event A defined as $\{T_1(n) \leq \frac{n}{2}\}$, we obtain

$$\begin{aligned}
R_n(v) + R_n(v') &\geq \frac{n\Delta}{4}\left(P_v(A) + P_{v'}(A^c)\right) \\
&\geq \frac{n\Delta}{8}\exp(-D(P_v, P_{v'})).
\end{aligned}$$

[2]For notational convenience, we have suppressed the dependence of regret on the algorithm.

As in the two-armed bandit case, it can be shown that $D(P_v, P_{v'})) \leq E_v(T_i(n))4\Delta^2$, which leads to the following bound:

$$R_n(v) + R_n(v') \geq \frac{n\Delta}{4} \exp\left(-E_v(T_i(n))4\Delta^2\right),$$

Notice that $E_v(T_i(n)) \leq \frac{n}{K-1}$. Suppose not. Then, $n = \sum_j E_v(T_i(n)) > \sum_j \frac{n}{K-1} > n$, a contradiction. Hence, we have

$$R_n(v) + R_n(v') \geq \frac{n\Delta}{8} \exp\left(-\frac{n4\Delta^2}{K-1}\right),$$

Choosing $\Delta = \sqrt{\frac{K-1}{8n}}$, we have that

$$\begin{aligned}
\max(R_n(v), R_n(v')) &\geq \frac{1}{2}\left(R_n(v) + R_n(v')\right) \\
&\geq \frac{n\Delta}{16} \exp\left(-\frac{n4\Delta^2}{K-1}\right) \\
&\geq c\sqrt{Kn},
\end{aligned}$$

for some problem-independent constant c. The claim follows. □

1.7.2 Instance dependent lower bounds

to be done

Exercises

Exercise 1.17. *A regret upper bound of $O(\log n)$ was shown for UCB algorithm, while a lower bound of $O(\sqrt{n})$ (ignoring the dependence on number of arms K) was also derived. How does one resolve the apparent contradiction between these two bounds?*

Exercise 1.18. *Consider a two-armed bandit problem. Recall that the ETC algorithm chooses each arm m number of times and then plays the arm with the highest sample mean $(n-2m)$ number of times. For any horizon n and exploration parameter m (chosen non-adaptively, i.e., before sampling any arm), there exists a problem instance with underlying arms' distribution $v = \mathcal{N}(\mu_1, 1) \times \mathcal{N}(\mu_2, 1)$, such that the regret $R_n(v)$ of ETC on v satisfies*

$$R_n(v) \geq cn^{2/3},$$

where c is a problem-independent constant.

Exercise 1.19. *Consider a two-armed bandit problem with underlying joint distribution $\nu = p_1 \times p_2$, where p_1 and p_2 are Bernoulli distributions with parameters θ and $1-\theta$, respectively, for some $\theta \in (\frac{1}{2}, 1)$. Let $v' = p_2 \times p_1$ denote the underlying distribution for a permuted bandit problem. Then, for any bandit algorithm $\mathcal{A}$,*

$$\max(R_n(v), R_n(v') \geq \frac{c}{2\theta - 1},$$

where $R_n(v)$ (resp. $R_n(v')$) is the expected regret with horizon n on problem v (resp. v') and c is a problem-independent constant.

1.8 A tour of Bayesian inference

So far, we have been working in a setting that statisticians would classify as frequentist - a worldview where probability refers to a long-term frequency (think of tossing a coin a large number of times. The frequency of heads would coincide with the intuitive notion of the probability that the coin turns up heads). Further, a *good* algorithm here (in the frequentist setting) would come with long run frequency guarantees (for e.g., sample mean and its associated convergence guarantees).

An alternative view is *Bayesian*, where probability refers to a (subjective) belief and the algorithms are built around updating the beliefs, based on observations from the real-world. We illustrate the difference between the frequentist and Bayesian view in the following example.

Consider the problem of mean estimation of a normally distributed r.v. with unit variance. To be more precise, let $X_1, \ldots, X_n$ be i.i.d. samples from $\mathcal{N}(\theta, 1)$, i.e., the normal distribution with mean θ and variance 1. The well-known sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ satisfies

$$\mathbb{P}_\theta(\theta \in C) = 0.95, \text{ where } C = \left(\bar{X}_n - \frac{1.96}{\sqrt{n}}, \bar{X}_n + \frac{1.96}{\sqrt{n}}\right].$$

The statement above implies that the random quantity C would contain the constant parameter θ with high probability.

The Bayesian approach to the mean estimation would proceed as follows:

1. Choose a prior density $\pi(\cdot)$ that reflects your belief about the parameter θ;
2. Collect data $D_n = \{X_1, \ldots, X_n\}$, which is sampled iid from $p(X \mid \theta)$. The latter quantity denotes the density of X conditioned on θ;
3. Update your belief, i.e., the posterior density $p(\theta \mid D_n)$, using the Bayes theorem as follows:

$$\begin{aligned} p(\theta \mid D_n) &= \frac{p(D_n \mid \theta)\pi(\theta)}{p(D_n)} \\ &= \frac{L_n(\theta)\pi(\theta)}{c_n}, \end{aligned}$$

   where $L_n(\theta) = \prod_{i=1}^n p(X_i \mid \theta)$ is the likelihood function and $c_n = \int_\theta L_n(\theta)\pi(\theta)d\theta$ is a normalization constant. Note that we have assumed independence for the samples from the conditional density $p(X \mid \theta)$ and hence, $L_n(\theta)$ splits into a nice product.

Using the posterior density, the posterior mean can be calculated as

$$\bar{\theta}_n = \int \theta p(\theta \mid D_n)d\theta = \frac{\int \theta L_n(\theta)\pi(\theta)d\theta}{\int \theta L_n(\theta)\pi(\theta)d\theta}.$$

Further, one can find c and d such that

$$\int_{-\infty}^{c} \partial(\theta \mid D_n)d\theta = \int_{d}^{\infty} \partial(\theta \mid D_n)d\theta = \frac{\alpha}{2},$$

leading to the Bayesian counterpart $C = (c, d)$ of the confidence interval. In other words, we have

$$P(\theta \in C \mid D_n) = 1 - \alpha.$$

Notice that θ above is a random variable, unlike the frequentist setting.

Example 1.4. *Suppose we start with a uniform prior, i.e., $\pi(\theta) = 1, \forall\theta$ and the conditional density $p(X \mid \theta)$ is Bernoulli with parameter θ. Then, the posterior density can be calculated as follows:*

$$\begin{aligned} p(\theta \mid D_n) &= c\pi(\theta)L_n(\theta) \\ &= c\prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i} \\ &= c\theta^{S_n}(1-\theta)^{n-S_n}, \text{ where } S_n = \sum_{i=1}^{n} X_i \\ &= c\theta^{S_n+1-1}(1-\theta)^{n-S_n+1-1}. \end{aligned} \tag{1.17}$$

Recall that a r.v. is Beta-distributed with parameters α and β, if the underlying density is

$$\pi_{\alpha,\beta}(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1},$$

where Γ is the gamma function. Further, the mean of a beta r.v. is $\frac{\alpha}{\alpha+\beta}$. Comparing (1.17) *with beta density, we have*

$$\theta \mid D_n \sim \textit{Beta}(S_n + 1, n - S_n + 1),$$

where $\sim$ is short hand for "distributed according to". The posterior mean is given by

$$\bar{\theta}_n = \frac{S_n + 1}{n + 2} = \left(\frac{n}{n+2}\right)\frac{S_n}{n} + \left(1 - \frac{n}{n+2}\right)\frac{1}{2}.$$

Thus, the posterior mean is a convex combination of the prior mean and the sample average, with the latter factor dominating for large n.

Repeating the calculation in the example above, starting with a non-uniform prior $\pi \sim \text{Beta}(\alpha, \beta)$, it is easy to see that

$$\theta \mid D_n \sim \text{Beta}(\alpha + S_n, \beta + n - S_n).$$

Notice that the special case of $\alpha = \beta = 1$ corresponds to the uniform prior. The posterior mean is given by

$$\bar{\theta}_n = \frac{\alpha + S_n}{\alpha + \beta + n} = \left(\frac{n}{\alpha+\beta+n}\right)\frac{S_n}{n} + \left(1 - \frac{n}{\alpha+\beta+n}\right)\frac{\alpha}{\alpha+\beta}.$$

Exercise 1.20. *Repeat the calculations in the example above for the case when $p(X \mid \theta)$ is $\mathcal{N}(\theta, \sigma^2)$ and the prior π is $\mathcal{N}(\theta_0, \sigma_0^2)$. Conclude that the posterior density $p(\theta \mid D_n)$ corresponds to $\mathcal{N}(\theta_n, \sigma_n^2)$, where*

$$\theta_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\right)\frac{S_n}{n} + \left(\frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\right)\theta_0.$$

Discuss the case when $\sigma_0 = 0$ and $\sigma_0 \neq 0$ and n is large.

Exercise 1.21. *Let θ denote a univariate parameter and $X_1, \ldots, X_n$ denote i.i.d. samples from a density $p(X \mid \theta)$ (referred to as likelihood below). Consider the following likelihoods for the samples:*

1. *Poisson: $p(X = m \mid \theta) = \frac{\theta^m e^{-\theta}}{m!}, \ m \geq 0.$*

2. *Normal with mean 0 and variance θ^2: $p(X = x \mid \theta) = \frac{1}{\sqrt{2\pi}\theta} e^{-\frac{x^2}{2\theta^2}}.$*

For each of the likelihood choices above, answer the following:

(a) Work out the posterior update and highlight the form of the posterior density (ignoring the normalization constant).

(b) Under what choice for the prior is conjugacy guaranteed?

(c) Derive the expression for posterior mean and variance and discuss the asymptotics (i.e., when the number of samples n become large).

1.9 Bayesian bandits

A Bayes approach to bandits would involve a prior over problem instances and the notion of Bayes regret would be the regret suffered for each problem instance, combined with an averaging over the problem instances through the prior. In contrast, the frequentist regret notion that we have discussed earlier fixes the problem instance and considers the regret incurred for any algorithm on the chosen problem instance. The crucial difference in the Bayesian view is that the problem instance is chosen randomly through the prior. We make the Bayesian view of bandits precise now.

1.9.1 Setting and Bayes regret

Consider a K-armed stochastic bandit problem, where the underlying arms' distributions are parameterized. In particular, for the sake of simplicity, we assume 1-parameter distributions for the arms. Bernoulli and normal distribution with known variance are popular examples in this class. Let $\theta_1, \ldots, \theta_K$ denote the parameters for the arms $i = 1, \ldots, K$. The overall flow is as follows:

1. The prior density π is used to pick a $\theta = (\theta_1, \ldots, \theta_K)$, which corresponds to a problem instance. We assume independence in the prior, i.e., $(\theta_i)_{i=1,\ldots,K}$ is drawn independently from $(\pi_i)_{i=1,\ldots,K}$.

2. Conditioned on the problem instance corresponding to θ, the samples $(Y_{t,1})_{t\geq 1}, \ldots, Y_{t,K})_{t\geq 1}$ are jointly independent and iid with marginals denoted by $\nu_{\theta_1}, \ldots, \nu_{\theta_K}$. The expectation of the marginal distribution ν_{θ_j} is denoted by μ_j, for $j = 1, \ldots, K$.

Let $\mu^*(\theta) = \max_{i=1,\ldots,K} \mu_i(\theta_i)$ denote the mean of the best arm, under the parameter θ. The expected regret $R_n(\theta)$, conditioned on the parameter θ, for an algorithm that chooses arm I_t in round $t = 1, \ldots, n$, is defined as

$$R_n(\theta) = \sum_{t=1}^{n} \mathbb{E}\left(\mu^* - \mu_{I_t} \mid \theta\right).$$

The Bayes regret R_n^B would average the regret defined above, over problem instances (or, equivalently over the parameter θ), with the averaging governed by the prior density, i.e.,

$$R_n^B = \mathbb{E}_\pi(R_n(\theta)) = \sum_{t=1}^{n} \mathbb{E}_\pi\left(\mathbb{E}\left(\mu^* - \mu_{I_t} \mid \theta\right)\right).$$

1.9.2 Thompson sampling for Bernoulli bandits

We illustrate the main ideas behind the Thompson sampling algorithm, for the case when the arms' distributions are Bernoulli. Let $\theta = (\theta_1, \ldots, \theta_K)$ denote the vector of Bernoulli parameters. From the discussion in the earlier section on Bayesian inference, it is apparent that a Beta prior is a good choice for this problem, due to its conjugacy property. So, suppose that the prior $\pi = (\pi_1, \ldots, \pi_K)$, with $\pi_i \sim \text{Beta}(\alpha_i, \beta_i)$.

Under the Beta prior, the posterior update is straightforward. Letting I_t denote the arm pulled by the bandit algorithm in round t and r_t the sample reward obtained, the posterior update is

$$(\alpha_{I_t}, \beta_{I_t}) \leftarrow (\alpha_{I_t}, \beta_{I_t}) + (r_t, 1 - r_t).$$

What remains to be specified the decision rule for choosing arms in each round, based on the beliefs specified by the posterior density $p(\theta_i \mid H_t)$ for each arm. Here $H_t = (a_{I_1}, r_1, a_{I_2}, r_2, \ldots, a_{I_{t-1}}, r_{t-1}$ is the history of actions chosen and sample rewards observed up to time t.

A greedy playing choice is to select the arm with the highest posterior mean. In the Bernoulli bandit setting considered here, this is equivalent to playing the arm that achieves $\max_{i=1,\ldots,K} \frac{\alpha_i}{\alpha_i+\beta_i}$. Such a decision rule is prone to high regret because of reasons similar to those arguing against an algorithm that picks the arm with the highest sample mean (in the frequentist setting). To illustrate, consider a contrived scenario as discussed in Russo et al. [2017], where there are three arms, which have been pulled 1000, 1000 and 5 times and their posterior means are 0.6, 0.6 and 0.4. The greedy choice would end up pulling arm 1, without resolving the uncertainty around the mean of arm 3 and given the small number of samples used for arm 3, there is a positive probability that its mean is higher than 0.6. The UCB algorithm has an exploration factor that ensures arm 3 ends up being pulled well enough to resolve the uncertainty around its mean. The Thompson sampling (TS) algorithm achieves the same effect by sampling from the posterior densities. With the Beta prior and posterior update as mentioned above, TS choice for the arm I_t to be played in each round is given by

$$I_t = \mathop{\arg\max}_{i=1,\ldots,K} \hat{\theta}_i, \text{ where } \hat{\theta}_i \sim \text{Beta}(\alpha_i, \beta_i).$$

Thus, TS obtains a random estimate of each arm's mean, by sampling from its posterior density and then plays the arm with the highest sample. An equivalent description for TS choice of arm played in each round is given below:

$$I_t \sim p(a^* \mid H_t).$$

The statement above is equivalent to saying that TS plays the arm that has the highest posterior probability of being the best arm, given the history of sample rewards so far. Finally, if we view a^* as a r.v. with density $\pi(\cdot)$ (i.e., the prior), then we have that

$$p(a^* = i \mid H_t) = p(I_t = i \mid H_t), \quad \text{for } i = 1, \ldots, K.$$

The statement above follows by definition of I_t and in simple terms means that the optimal arm a^* and the arm I_t played by TS have the same distribution, when conditioned on the history H_t. Let us return to the three armed example used while discussing the shortcomings of the greedy algorithm. In that example setting, TS would do better exploration as there would be a small probability that arm 3's posterior sample is better than that of arm 1. Moreover, the posterior sampling ensures that arm 1, which has the highest sample mean would be played with much higher probability than arm 3, while arm 2 would have zero chances, given it has lower sample mean and enough samples to rule out any uncertainty.

Kinship of TS to optimistic exploration ala UCB

Let $I_t = \arg\max_{i=1,\ldots,K} \text{UCB}_t(i)$ denote the UCB index for arm i in round t, where $\text{UCB}_t(i)$ is as defined earlier (in Section 1.5). Let θ denote the underlying parameter vector and $\mu^*(\theta) = \max_{i=1,\ldots,K} \mu_i$ the optimal mean value, achieved by an arm $a^* \in \{1, \ldots, K\}$. As mentioned in the setting description above, we have assumed 1-parameter distributions for the arms and $\mu_i(\theta)$ denotes the mean value for arm i under parameter θ_i. Notice that

$$\begin{aligned} \mu^*(\theta) - \mu_{I_t}(\theta) &= \mu^*(\theta) - \text{UCB}_t(I_t) + \text{UCB}_t(I_t) - \mu_{I_t}(\theta) \\ &\le \underbrace{\mu^*(\theta) - \text{UCB}_t(a^*)}_{\text{(A)}} + \underbrace{\text{UCB}_t(I_t) - \mu_{I_t}(\theta)}_{\text{(B)}}. \end{aligned} \tag{1.18}$$

The last inequality above holds since $\text{UCB}_t(I_t) \ge \text{UCB}_t(i)$, for all i. If $\text{UCB}_t(a^*)$ is an upper-confidence bound, i.e., $\text{UCB}_t(a^*) > \mu^*$ with high probability, then the term (A) is negative and can be ignored. On the other hand, the term (B) in (1.18) related to how well the algorithm has estimated the mean of a sub-optimal arm and the confidence width that is of the order $O(\sqrt{\frac{\log n}{T_{I_t}(t)}})$ would play a role in bounding (B).

Summing over t in (1.18), we obtain

$$R_n^B(\text{UCB}) \le \mathbb{E} \sum_{t=1}^{n} (\mu^*(\theta) - \text{UCB}_t(a^*)) + \mathbb{E} \sum_{t=1}^{n} (\text{UCB}_t(I_t) - \mu_{I_t}(\theta)). \tag{1.19}$$

We now analyze the Bayes regret of TS. Letting I_t denote the arm played by TS in round t, using samples from the posterior distributions for each arm and H_t denote the history of sample rewards and actions chosen by TS upto time t, we have

$$\begin{aligned}
\mathbb{E}\left(\mu^*(\theta) - \mu_{I_t}(\theta)\right) &= \mathbb{E}\left(\mathbb{E}\left(\mu^*(\theta) - \mu_{I_t}(\theta) \mid H_t\right)\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(\mu^*(\theta) - \mathrm{UCB}_t(I_t) + \mathrm{UCB}_t(I_t) - \mu_{I_t}(\theta) \mid H_t\right)\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(\mu^*(\theta) - \mathrm{UCB}_t(a^*) + \mathrm{UCB}_t(I_t) - \mu_{I_t}(\theta) \mid H_t\right)\right) \qquad (1.20) \\
&= \underbrace{\mathbb{E}\left(\mu^*(\theta) - \mathrm{UCB}_t(a^*)\right)}_{\text{(A')}} + \underbrace{\mathbb{E}\left(\mathrm{UCB}_t(I_t) - \mu_{I_t}(\theta)\right)}_{\text{(B')}}. \qquad (1.21)
\end{aligned}$$

In the above, for arriving at the equality in (1.20), we have used the fact that $\mathbb{E}\left(\mathrm{UCB}_t(a^*) \mid H_t\right) = \mathbb{E}\left(\mathrm{UCB}_t(I_t) \mid H_t\right)$, which holds because (i) a^* and I_t (chosen by TS) have the same distribution conditioned on the history H_t and (ii) $\mathrm{UCB}_t(\cdot)$ is a deterministic function given H_t.

Summing over t in (1.21), we obtain

$$R_n^B(\mathrm{TS}) \leq \mathbb{E}\sum_{t=1}^{n}\left(\mu^*(\theta) - \mathrm{UCB}_t(a^*)\right) + \mathbb{E}\sum_{t=1}^{n}\left(\mathrm{UCB}_t(I_t) - \mu_{I_t}(\theta)\right). \qquad (1.22)$$

The RHS in the equality above bears a striking resemblance to that in (1.19), even though the TS algorithm did not incorporate confidence widths explicitly. Thus, posterior sampling in TS is performing the task of optimistic exploration, while explicit confidence bounds did a similar job in UCB algorithm. Further, any upper confidence bound can be used to arrive at (1.21) for TS, while UCB algorithm requires clear specification of the confidence widths to handle the exploration-exploitation dilemma. This fact about implicitness of optimistic exploration in TS makes it advantageous to use TS in complicated settings, where the confidence widths are not apparent, for instance, due to dependencies between arms. The flip side to TS is the computational expense involved in updating the posterior, a problem that can be eased with conjugacy under well-known arms' distribution choices.

The Bayes regret bound for TS in (1.22) holds in general, since no distributional assumptions were made. We now specialize the bound for the case of Bernoulli bandit, which in turn leads to a $\tilde{O}(\sqrt{KT})$ bound on $R_n^B(\mathrm{TS})$. For notational convenience, we shall drop the dependence on θ in μ_i.

Let us define upper and lower confidence bounds as follows: For any arm i,

$$\mathrm{UCB}_t(i) = \hat{\mu}_i(t-1) + \sqrt{\frac{2\log n}{T_i(t-1)}}, \text{ and } \mathrm{LCB}_t(i) = \hat{\mu}_i(t-1) - \sqrt{\frac{2\log n}{T_i(t-1)}},$$

where $\hat{\mu}_i(\cdot)$ and $T_i(\cdot)$ denote the sample mean and number of times arm i is pulled up to time t. From the discussion in Section 1.3, we know that

$$\mathbb{P}\left[\mu_i(\theta) > \mathrm{UCB}_t(i)\right] \leq \frac{1}{n}, \text{ for any arm } i.$$

Hence,

$$\begin{aligned}
&\mathbb{E}(\mu_i(\theta) - \mathrm{UCB}_t(i)) \leq \frac{1}{n}, \forall i, \\
\Rightarrow &\mathbb{E}(\mu^*(\theta) - \mathrm{UCB}_t(a^*)) \leq \max_i \mathbb{E}(\mu_i(\theta) - \mathrm{UCB}_t(i)) \leq \frac{1}{n},
\end{aligned}$$

where we used the fact that, for any r.v. X with $|X| \leq c$, $\mathbb{E}X \leq cP(X > 0)$.

Thus, term (A) in (1.22) can be bounded as follows:

$$\mathbb{E}\sum_{t=1}^{n}(\mu^*(\theta) - \text{UCB}_t(a^*)) \leq \sum_{t=1}^{n}\frac{1}{n} = 1.$$

For the term (B), observe that $\text{UCB}_t(i) - \mu_i = \text{LCB}_t(i) - \mu_i + 2\sqrt{\frac{2\log n}{T_i(t-1)}}$. From Hoeffding's inequality, it is clear that

$$\mathbb{P}\left[\mu_i(\theta) < \text{LCB}_t(i)\right] \leq \frac{1}{n}, \text{ for any arm } i.$$

Hence, we have

$$\begin{aligned}
\mathbb{E}\sum_{t=1}^{n}(\text{UCB}_t(I_t) - \mu_{I_t}) &= \mathbb{E}\sum_{t=1}^{n}(\text{LCB}_t(I_t) - \mu_{I_t}) + \mathbb{E}\sum_{t=1}^{n}2\sqrt{\frac{2\log n}{T_{I_t}(t-1)}} \\
&\leq 1 + 2\,\mathbb{E}\left(\sum_{i=1}^{K}\sum_{t:I_t=i}\sqrt{\frac{2\log n}{T_i(t-1)}}\right) \\
&\leq 1 + 2\sqrt{2\log n}\,\mathbb{E}\left(\sum_{i=1}^{K}\sum_{j=1}^{T_i(t-1)}\sqrt{\frac{1}{j}}\right) \\
&\leq 1 + 2\sqrt{2\log n}\,\mathbb{E}\left(\sum_{i=1}^{K}2\sqrt{T_i(t-1)}\right) \\
&\leq 1 + 4\sqrt{2\log n}\,\mathbb{E}\left(\sqrt{K\sum_{i=1}^{K}T_i(t-1)}\right) \\
&= 1 + 4\sqrt{2\log n}\sqrt{Kn} = O(\sqrt{Kn\log n}).
\end{aligned}$$

To arrive at the penultimate inequality above, we have compared a sum with an integral, while the last inequality follows from Cauchy-Schwarz.

The above analysis allows us to conclude the following:

Theorem 1.13. *For the K-armed bandit problem with Bernoulli arms, the Bayes regret of Thompson sampling satisfies*

$$R_n^B(TS) = O(\sqrt{Kn\log n}).$$

Exercise 1.22. *Solve Exercise 1.14 using an algorithm based on Thompson sampling.*

Exercise 1.23. ***(Simulation experiment)***
Consider three multi-armed bandit problem instances, where each arm's distribution is Bernoulli with parameters as follows:

Arms $\rightarrow$	1	2	3	4	5	6	7	8	9	10
P1	0.5	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4
P2	0.5	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48
P3	0.5	0.2	0.1	*No other arms*						

Write a program (in your favorite language) to simulate each of the above bandit problems and implement the following bandit algorithms:

- *Thompson sampling (TS) with a Beta*$(1,1)$ *prior.*
- *A variant of TS where the prior has mean* 0.2 *instead of* 0.5.
- *The UCB algorithm.*

Do the following for each problem instance:

(i) *Choose the horizon* n *as* 10000.

(ii) *For each algorithm, repeat the experiment* 100 *times.*

(iii) *Store the regret in each round* $m = 1, \ldots, n$.

(iv) *For TS and its variant, store the (posterior) probability of playing each arm.*

(v) *Plot regret against the rounds* $t = 1, \ldots, n$. *For TS variants, plot the arm playing probabilities as well.*

(vi) *For each plot, add standard error bars.*

(vii) *In the figures that report regret performance, plot the gap-dependent lower bound as well as worst case lower bound.*

Interpret the numerical results and submit your conclusions. In particular, discuss the following:

(a) *Comparison of the regret performance of TS with Beta*$(1,1)$ *prior against that of UCB. How do both algorithm fare when compared to the lower bounds (esp. the gap-dependent one).*

(b) *For the TS variant with a prior mean* 0.2, *discuss the results, while including comparison to TS with Beta*$(1,1)$ *prior.*

1.10 Bibliographic remarks

A bandit framework for learning dates back to Thompson [1933], where the motivation was clinical trials. The stochastic K-armed bandit problem was formulated by Robbins [1952]. The presentation of the ETC algorithm and concentration inequalities is based on blog posts by Szepesvári and Lattimore [2017]. Some of the material on sub-Gaussian r.v.s is based on Chapter 2 of [Wainwright, 2019].

The seminal upper confidence type index strategy was first proposed by Lai and Robbins [1985], refined to use a sample mean-based index by Agrawal [1995] and Burnetas and Katehakis [1996]. The finite-time analysis of UCB using Hoeffding-type bounds was done by Auer et al. [2002] and the UCB regret bound presented here closely follows this approach. To know further into UCB-type approaches, the reader is referred to KL-UCB [Cappé et al., 2013], UCB-V Audibert et al. [2009] and UCB-Improved [Auer and Ortner, 2010]. For further reading, the reader is referred to [Salomon and Audibert, 2011] and the survey article [Munos, 2014].

The presentation of lower bounds is based on blog posts by Szepesvári and Lattimore [2017] and Chapter 3 of [Slivkins, 2017]. The information theory background is based on the classic text book by Cover and Thomas [2012].

The presentation of Bayesian inference related topics is based on Chapter 11 of [Wasserman, 2013]. The Bayes regret analysis of Thompson sampling is from [Russo and Van Roy, 2014], while a high-level introduction to Thompson sampling is available in [Russo et al., 2017]. A frequentist regret analysis, which is skipped in this chapter, is available in [Korda et al., 2013, Agrawal and Goyal, 2012, 2013, Kaufmann et al., 2012b]. For further reading on TS, the reader is referred to [Liu and Li, 2016, Kaufmann et al., 2012a, Gopalan et al., 2014, Bubeck and Liu, 2013, Russo and Van Roy, 2016, 2013].

Bibliography

R. Agrawal. Sample mean based index policies by o (log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995.

S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, volume 23, pages 39.1–39.26, 2012.

S. Agrawal and N. Goyal. Further optimal regret bounds for thompson sampling. In *Artificial Intelligence and Statistics*, pages 99–107, 2013.

J. Y. Audibert, R. Munos, and C. Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

P. Auer and R. Ortner. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

S. Bubeck and C. Y. Liu. Prior-free and prior-dependent regret bounds for thompson sampling. In *Advances in Neural Information Processing Systems*, pages 638–646, 2013.

A. N. Burnetas and M. N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

O. Cappé, A. Garivier, O. Maillard, R. Munos, and G. Stoltz. Kullback–Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

A. Gopalan, S. Mannor, and Y. Mansour. Thompson sampling for complex online problems. In *International Conference on Machine Learning*, pages 100–108, 2014.

E. Kaufmann, O. Cappé, and A. Garivier. On Bayesian upper confidence bounds for bandit problems. In *Artificial Intelligence and Statistics*, pages 592–600, 2012a.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory*, pages 199–213. Springer, 2012b.

Nathaniel Korda, Emilie Kaufmann, and Remi Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems*, pages 1448–1456, 2013.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

C. Y. Liu and L. Li. On the prior sensitivity of thompson sampling. In *International Conference on Algorithmic Learning Theory*, pages 321–336. Springer, 2016.

R. Munos. From bandits to monte-carlo tree search: The optimistic principle applied to optimization and planning. *Foundations and Trends® in Machine Learning*, 7(1):1–129, 2014.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematics Society*, 58:527–535, 1952.

D. Russo and B. Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, pages 2256–2264, 2013.

D. Russo and B. Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

D. Russo and B. Van Roy. An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, and Ian Osband. A tutorial on thompson sampling. *arXiv preprint arXiv:1707.02038*, 2017.

A. Salomon and J. Y. Audibert. Deviations of stochastic bandit regret. In *International Conference on Algorithmic Learning Theory*, pages 159–173. Springer, 2011.

Alex Slivkins. Multi-armed bandits. `http://slivkins.com/work/MAB-book.pdf`, 2017.

C. Szepesvári and T. Lattimore. Bandit Algorithms. `http://banditalgs.com/`, 2017.

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

M. J. Wainwright. High-dimensional statistics: A non-asymptotic viewpoint, 2019.

L. Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.