MFAI: Optimistic Exploration in Online Learning
Alex Ayoub

1 Concentration Inequalities

This section very closely follows Chapter 5 of [2]. Also, I highly recommend reading [1] for a more in depth survey of modern concentration inequalities. Now let's introduce some important notation and assumptions!

1.1 Tail Probabilities

Suppose that $X_1, X_2, ..., X_n$ is a sequence of independent and identically distributed (i.i.d) random variables, and assume that the mean $\mu = \mathbb{E}[][X]$ and variance $\sigma^2 = \mathbb{V}[X]$ exist. Having observed $X_1, X_2, ..., X_n$ we would like to estimate the common mean μ (this is very important in the multi-armed bandit problem as in this problem setting we are only concerned with the optimal-action which is just pulling the arm with the highest mean). The most natural estimator is

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

which is called the **sample mean** or **empirical mean**. Linearity of expectation shows that $\mathbb{E}[\hat{\mu}] = \mu$, which means that $\hat{\mu}$ is an **unbiased estimator** of μ. How far from μ do we expect $\hat{\mu}$ to be? A simple measure of the spread of the distribution of a random variable Z is its variance, $\mathbb{V}[Z] = \mathbb{E}[(Z - \mathbb{E}[Z])^2]$. A quick calculation using independence shows that

$$\mathbb{V}[\hat{\mu}] = \mathbb{E}[(\hat{\mu} - \mu)^2] = \frac{\sigma^2}{n}, \tag{1}$$

which means that we expect the squared distance between μ and $\hat{\mu}$ to shrink as n grows large at a rate of $1/n$ and scale linearly with the variance of X. While the expected squared error is important, it does not tell us very much about the distribution of the error. To do this we usually analyse the probability that $\hat{\mu}$ overestimates or underestimates μ by more than some value $\varepsilon > 0$. Precisely, how do the following quantities depend on ε?

$$\mathbb{P}(\hat{\mu} \geq \mu + \varepsilon) \text{ and } \mathbb{P}(\hat{\mu} \geq \mu - \varepsilon).$$

The expression above are called the **tail probabilities** of $\hat{\mu} - \mu$. Specifically, the first is called the upper tail probability and the second is called the lower tail probability.

1.2 The Inequalities of Markov and Chebyshev

The most straightforward way to bound the tails is by using Chebyshev's inequality, which is itself a corollary of Markov's inequality. The latter is one of the golden hammers of probability theory, and so we include it for the sake of completeness.

Lemma 1.1. *For any random variable X and $\varepsilon > 0$, the following holds:*

1. *(Markov):* $\mathbb{P}(|X| \geq \varepsilon) \leq \frac{\mathbb{E}[|X|]}{\varepsilon}$
2. *(Chevyshev):* $\mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) \leq \frac{\mathbb{V}[X]}{\varepsilon^2}$

Proof We only prove Markov's inequality, the proof for Chebyshev's inequality is left to the reader

$$\begin{aligned}\mathbb{E}[X] &= \int_0^\infty xp(x)dx = \int_0^\varepsilon xp(x)dx + \int_\varepsilon^\infty xp(x)dx \\ &\geq \int_\varepsilon^\infty xp(x)dx \geq \varepsilon \int_\varepsilon^\infty p(x)dx = \varepsilon \mathbb{P}(X > \varepsilon)\end{aligned}$$

By combining 1 with Chebyshev's inequality, we can bound the two-sided tail directly in the terms of the variance by

$$\mathbb{P}(|\hat{\mu} - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \tag{2}$$

This result is nice because it was so easily bought and relied on no assumptions other than the existence of the mean and variance. The downside is that when X is well behaved, the inequality is rather loose. By assuming that higher moments of X exist, Chebyshev's inequality can be improved by applying Markov's inequality to $|\hat{\mu} - \mu|^k$, with the positive integer k to be chosen so that the resulting bound is optimised. This can be a bit cumbersome, and thus we present the continuous analog of this, known as the Cramer-Chernoff method.

To calibrate our expectations on what improvement to expect relative to Chebyshev's inequality, let us start by recalling the central limit theorem (CLT). Let $S_n = \sum_{t=1}^{n}(X_t - \mu)$. The CLT says that under no additional assumptions than the existance of variance, the limiting distribution of $S_n/(\sqrt{n\sigma^2})$ as $n \to \infty$ is the standard normal distribution. If $Z \sim \mathcal{N}(0, 1)$, then

$$\mathbb{P}(Z \geq u) = \int_u^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$$

The integral has no closed-form solution, but is easy to bound:

$$\int_u^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \leq \frac{1}{u\sqrt{2\pi}} \int_u^\infty x \exp\left(-\frac{x^2}{2}\right) dx \tag{3}$$

$$= \sqrt{\frac{1}{2\pi u^2}} \exp\left(-\frac{u^2}{2}\right) \tag{4}$$

which gives

$$\mathbb{P}(\hat{\mu} \geq \mu + \varepsilon) = \mathbb{P}\left(S_n/\sqrt{\sigma^2 n} \geq \varepsilon\sqrt{\frac{n}{\sigma^2}}\right) \approx \mathbb{P}\left(Z \geq \varepsilon\sqrt{\frac{n}{\sigma^2}}\right) \tag{5}$$

$$\leq \sqrt{\frac{\sigma^2}{2\pi n\varepsilon^2}} \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right) \tag{6}$$

This is usually much smaller than what we obtained with Chebyshev's inequality (Exercise 5.3). In particular, the bound on the right-hand side of (5.4) decays slightly faster than the negative exponential of $n\varepsilon^2/\sigma^2$, which means that $\hat{\mu}$ rapid concentrates around its mean.

The asymptotic nature of the CLT makes it unsuitable for designing bandit algorithms. As an approximation for a finite number of observations, it provides a reasonable approximation only when close to the peak of the normal distribution; it requires a very large number of observations to stretch into the tails. In the next section, we derive finite-time analogs, which are only possible by making additional assumptions.

1.3 The Cramer Chernoff Method and Subgaussian Random Variables

For the sake of moving rapidly towards bandits, we start with a straightforward and relatively fundamental assumption on the distribution of X, known as the subgaussian assumption.

Definition 1.1. (Subgaussianity). A random variable X is σ-subgaussian if for all $\lambda \in \mathbb{R}$, it holds that $\mathbb{E}[\exp(\lambda X)] \leq \exp(\lambda^2\sigma^2/2)$.

An alternative way to express the subgaussianity condition uses the moment- generating function of X, which is a function $M_X : \mathbb{R} \to \mathbb{R}$ defined by $M_X(\lambda) = \mathbb{E}[\exp(\lambda X)]$. The condition in the definition can be written as

$$\psi_X(\lambda) = \log M_X(\lambda) \leq \frac{1}{2}\lambda^2\sigma^2$$

for all $\lambda \in \mathbb{R}$. The function ψ_X is called the cumulant-generating function. It is not hard to see that M_X need not exist for all random variables over the whole range of real numbers. For example, if X is exponentially distributed and $\lambda \geq 1$, then

$$\mathbb{E}[\exp(\lambda X)] = \int_0^\infty \exp(-x) \times \exp(\lambda x)dx = \infty$$

The moment-generating function of $X \sim \mathcal{N}(0, \sigma^2)$ satisfies $M_X(\lambda) = \exp({}^2\sigma^2/2)$, and so X is σ-subgaussian The following theorem explains the origin of the term 'subgaussian'. The tails of a σ-subgaussian random variable decay approximately as fast as that of a Gaussian with zero mean and the same variance.

Theorem 1.2. *If X is σ-subgaussian, then for any $\varepsilon \geq 0$,*

$$\mathbb{P}(X \geq \varepsilon) \leq \exp\left(\frac{\varepsilon^2}{2\sigma^2}\right) \tag{7}$$

Proof We take a generic approach called the Cramer–Chernoff method. Let $\lambda > 0$ be some constant to be tuned later. Then

$$\begin{aligned}\mathbb{P}(X \geq \varepsilon) = \mathbb{P}(\exp(\lambda X) \geq \exp(\lambda\varepsilon)) &\leq \frac{\mathbb{E}[\exp(\lambda X)]}{\exp(\lambda\varepsilon)} \text{ by Markov's Inequality} \\ &\leq \exp\left(\frac{\lambda^2\sigma^2}{2} - \lambda\varepsilon\right) \text{ Defn. of subgaussianity}\end{aligned}$$

Choosing $\lambda = \varepsilon/\sigma^2$ completes the proof. □

A similar inequality holds for the left tail. By using the union bound $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$, we also find that $\mathbb{P}(|X| \leq \varepsilon) \leq 2\exp(-\varepsilon^2/(2\sigma^2)))$. An equivalent form of these bounds is

$$\mathbb{P}\left(X \geq \sqrt{2\sigma^2\log(1/\delta)}\right) \leq \delta \text{ and } \mathbb{P}\left(|X| \geq \sqrt{2\sigma^2\log(2/\delta)}\right) \leq \delta$$

This form is often more convenient and especially the latter, which for small δ shows that with overwhelming probability X takes values in the interval

$$\left(-\sqrt{2\sigma^2\log(2/\delta)}, \sqrt{2\sigma^2\log(2/\delta)}\right)$$

To study the behavior of $\hat{\mu} - \mu$, we need one more lemma.

Lemma 1.3. *Suppose that X is σ-subgaussian with X_1 and X_2 are independent and σ_1 and σ_2-subgaussian, respectively, then:*

1. *$\mathbb{E}[X] = 0$ and $\mathbb{V}[X] \leq \sigma^2$.*
2. *cX is $|c|\sigma$-subgaussian for all $c \in \mathbb{R}$.*
3. *$X_1 + X_2$ is $\sqrt{\sigma_1^2 + \sigma_2^2}$-subgassian*

The proof of this lemma is left to the reader, however, both 2 and 3 follow from the Taylor expansion of e^x. Combining Lemma 1.3 and Theorem 1.2 leads to a straightforward bound on the tails of $\hat{\mu} - \mu$.

Corollary 1.4. *Assume that $X_i - \mu$ are independent, σ-subgaussian random variables. Then for any $\varepsilon \geq 0$,*

$$\mathbb{P}(\hat{\mu} \geq \mu + \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right) \text{ and } \mathbb{P}(\hat{\mu} \leq \mu - \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right),$$

where $\hat{\mu} = \frac{1}{n}\sum_{t=1}^{n} X_t$.

Proof By Lemma 1.3, it holds that $\hat{\mu} - \mu = \sum_{i=1}^{n}(X_i - \mu)/n$ is $\sigma/\sqrt{n}$-subgaussian. Then apply Theorem 1.2

For $x > 0$, it holds that $\exp(-x) \leq 1/(ex)$, which shows that the above inequality is stronger than what we obtained via Chevyshev's inequality except when ε is very small. It is exponentially smaller if $n\varepsilon^2$ is large relative to σ^2. The deviation form of the above results says that under the conditions of the result, for any $\delta \in [0, 1]$, with probability $1 - \delta$,

$$\mu \leq \hat{\mu} + \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}} \tag{8}$$

Symmetrically, it also follows that with probability at least $1 - \delta$,

$$\mu \geq \hat{\mu} - \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}} \tag{9}$$

Again, one can use a union bound to derive a two-sided inequality. Now we are ready to construct an online learning algorithm!

2 UCB for Multi-Armed Bandits

This section closely follows the first part of Chapter 7 in [2]. Using the bounds we constructed in Equations 8 and 9, we can construct a good algorithm for solving the multi-armed bandit problem. Now let's informally define the multi-armed bandit problem, more a more formal definition of the multi-armed bandit problem read Chapters 4 and 6 of [2] or chapter 2 of [4]. Let the multi-armed bandit problem be informally defined as follows:

1. The number of actions, also called arms, is denoted by a natural number k.
2. For simplicity, we assume all multi-armed bandit instances are 1-subgaussian in this talk.
3. Each arm has a mean, μ_i for $i \in [k]$, and when the arm is pulled, a reward $X_t \sim \mathcal{N}(\mu_i, 1)$ is observed, for $t \in [n]$ where n is the horizon or the length an agent interacts with the multi-armed bandit environment.
4. The objective is to minimize the regret when interacting with a bandit environment, the regret is defined as follows, $R_n = \sum_{t=1}^{n} \Delta_i \mathbb{E}[T_i(n)]$ where $\Delta_i = \mu^* - \mu_i$, $\mu^* = \max_i(\mu_i)$, and $T_i(n)$ is the number of times each arm was pulled.

The goal here is to minimize the regret which is equivalent to maximizing the number of times the optimal arm is pulled. So for each arm, i, we have a stream of data $X_1^i, X_2^i, ..., X_t^i$ and using this observed data we would like learn μ_i in order to determine with high probability which arm is the best arm? Before we do this let us talk about **optimism**.

2.1 Optimism in the Face of Uncertainty

The UCB algorithm is based on the principle of optimism in the face of uncertainty, which states that one should act as if the environment is as nice as plausibly possible. As we shall see in later chapters, the principle is applicable beyond the finite-armed stochastic bandit problem.

Imagine visiting a new country and making a choice between sampling the local cuisine or visiting a well-known multinational chain. Taking an optimistic view of the unknown local cuisine leads to exploration because without data, it could be amazing. After trying the new option a few times, you can update your statistics and make a more informed decision. On the other hand, taking a pessimistic view of the new option discourages exploration, and you may suffer significant regret if the local options are delicious. Just how optimistic you should be is a difficult decision, which we explore for the rest of the chapter in the context of finite-armed bandits.

For bandits, the optimism principle means using the data observed so far to assign to each arm a value, called the upper confidence bound that with high probability is an overestimate of the unknown mean. The intuitive reason why this leads to sublinear regret is simple. Assuming the upper confidence bound assigned to the optimal arm is indeed an overestimate, then another arm can only be played if its upper confidence bound is larger than that of the optimal arm, which in turn is larger than the mean of the optimal arm. And yet this cannot happen too often because the additional data provided by playing a suboptimal arm means that the upper confidence bound for this arm will eventually fall below that of the optimal arm.

In order to make this argument more precise, we need to define the upper confidence bound. Let $X_1, X_2, ..., X_n$ be a sequence of independent 1-subgaussian random variables with mean μ and $\hat{\mu} = \frac{1}{n}\sum_{t=1}^{n} X_t$ by Equation 8,

$$\mathbb{P}\left(\mu \geq \hat{\mu} + \sqrt{\frac{2\log(1/\delta)}{n}}\right) \leq \delta \text{ for all } \delta \in (0,1) \tag{10}$$

When considering its options in round t, the agent has observed $T_i(t-1)$ samples of arm i and received rewards from that arm with an empirical mean of $\hat{\mu}_i(t-1)$. Then a reasonable candidate for "as large as plausibly possible" for the unknown mean of the ith arm is

$$\text{UCB}_i(t-1,\delta) = \begin{cases} \infty, & \text{if } T_i(t-1) = 0 \quad (11) \\ \hat{\mu}_i(t-1) + \sqrt{\dfrac{2\log(1/\delta}{T_i(t-1)}}, & \text{otherwise} \quad (12) \end{cases}$$

Great care is required when comparing 10 and 11 because in the former the number of samples is the constant n, but in the latter it is a random variable $T_i(t-1)$. By and large, however, this is merely an annoying technicality, and the intuition remains that δ is approximately an upper bound on the probability of the event that the above quantity is an underestimate of the true mean.

The value inside the argmax is called the index of arm i. Generally speaking, an index algorithm chooses the arm in each round that maximises some value (the index), which usually only depends on the current time step and the samples from that arm. In the case of UCB, the index is the sum of the empirical mean of rewards experienced so far and the exploration bonus, which is also known as the confidence width.

Besides the slightly vague 'optimism guarantees optimality or learning' intuition we gave before, it is worth exploring other intuitions for the choice of index. At a very basic level, an algorithm should explore arms more often if they are (a) promising because $\hat{\mu}_i(t-1)$ is large or (b) not well explored because $T_i(t-1)$ is small. As one can plainly see, the definition in Eq. 11 exhibits this behaviour. This explanation is not completely satisfying, however, because it does not explain why the form of the functions is just so.

References

[1] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

[2] Tor Lattimore and Csaba Szepavári. *Bandit Algorithms*. 2019.

[3] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. *CoRR*, abs/1003.0146, 2010.

[4] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.