

5 Concentration of Measure

Before we can start designing and analyzing algorithms we need one more tool from probability theory called **concentration of measure**. Recall that the optimal action is the one with the largest mean. Since the mean payoffs are initially unknown, they must be learned from data. We now ask how long it takes to learn about the mean reward of an action.

Suppose that X, X_1, X_2, \dots, X_n is a sequence of independent and identically distributed 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, \dots, X_n we would like to define an **estimator** of the common mean μ . The natural choice to estimate μ is to use the average of the observations, also known as the **sample mean** or **empirical mean**.

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

The question is how far from μ do we expect $\hat{\mu}$ to be? First, by the linearity of expectation (Proposition 2.1), we notice that $\mathbb{E}[\hat{\mu}] = \mu$. 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 $\mathbb{V}[\hat{\mu}] = \sigma^2/n$. From this we get

$$\mathbb{E}[(\hat{\mu} - \mu)^2] = \frac{\sigma^2}{n}, \tag{5.1}$$

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 (so larger variance means larger expected squared difference). 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 analyze 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 ?

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

The expressions above (as a function of ε) are often called the **tail probabilities** of $\hat{\mu} - \mu$, see the figure below. In particular, the first is called an upper tail probability and the second the lower tail probability. Analogously, the probability $\mathbb{P}(|\hat{\mu} - \mu| \geq \varepsilon)$ is called a two-sided tail probability.

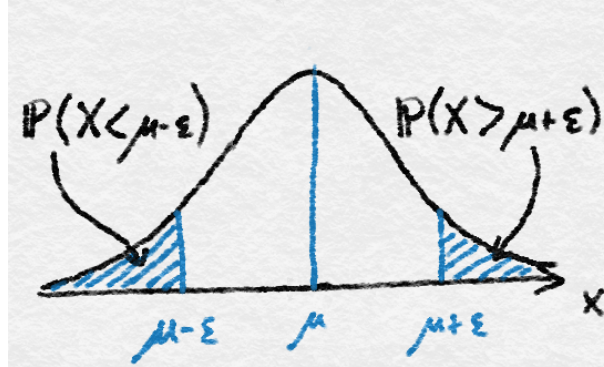


Figure 5.1 The figure shows a probability density, with the tails shaded indicating the regions where X is more than ε away from the mean μ .

5.1 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 5.1 For any random variable X with finite mean and $\varepsilon > 0$ it holds that:

$$(a) \text{ (Markov): } \mathbb{P}(|X| \geq \varepsilon) \leq \frac{\mathbb{E}[|X|]}{\varepsilon}.$$

$$(b) \text{ (Chebyshev): If } \mathbb{V}[X] < \infty, \text{ then } \mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) \leq \frac{\mathbb{V}[X]}{\varepsilon^2}.$$

We leave the proof of Lemma 5.1 as an exercise for the reader. By combining (5.1) with Chebyshev's inequality we can bound the two-sided tail directly in terms of the variance by

$$\mathbb{P}(|\hat{\mu} - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}. \quad (5.2)$$

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 in many cases the inequality is extremely loose and that huge improvement is possible if the distribution of X is well behaved. In particular, by assuming that higher moments of X exist, Chebyshev's inequality can be greatly 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 optimized. This is a bit cumbersome and thus instead we present the continuous analog of this, known as the Cramer-Chernoff method.

To calibrate our expectations on what gains to expect over Chebyshev's inequality, let us first discuss the **central limit theorem**. Let $S_n = \sum_{t=1}^n (X_t - \mu)$.

The central limit theorem (CLT) says that under no additional assumptions than the existence of the variance, the limiting distribution of $S_n/\sqrt{\sigma^2 n}$ as $n \rightarrow \infty$ is a Gaussian with mean zero and unit variance. Now, if $Z \sim \mathcal{N}(0, 1)$,

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

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

$$\begin{aligned} \int_u^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx &\leq \frac{1}{u\sqrt{2\pi}} \int_u^\infty x \exp\left(-\frac{x^2}{2}\right) dx \\ &= \sqrt{\frac{1}{2\pi u^2}} \exp\left(-\frac{u^2}{2}\right), \end{aligned} \quad (5.3)$$

which gives

$$\begin{aligned} \mathbb{P}(\hat{\mu} \geq \mu + \varepsilon) &= \mathbb{P}\left(S_n/\sqrt{\sigma^2 n} \geq \varepsilon\sqrt{n/\sigma^2}\right) \approx \mathbb{P}\left(Z \geq \varepsilon\sqrt{n/\sigma^2}\right) \\ &\leq \sqrt{\frac{\sigma^2}{2\pi n\varepsilon^2}} \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right). \end{aligned} \quad (5.4)$$

This is usually much smaller than what we obtained with Chebyshev's inequality (cf. 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 $\hat{\mu}$ rapidly concentrates around its mean. Unfortunately, since the central limit theorem is asymptotic, we cannot use it to study the regret when the number of rounds is a fixed finite number (cf. Exercise 5.5). Despite the folk-rule that $n = 30$ is sufficient for the Gaussian approximation based on the CLT to be reasonable, this is simply not true. One well known example is provided by Bernoulli variables with parameter $p \approx 1/n$, in which case the distribution of the sum is known to be much better approximated by the Poisson distribution with parameter one, which is nowhere near similar to the Gaussian distribution.

For these reasons we need a non-asymptotic alternative to the CLT. The pathological example in the previous paragraph shows this is only possible by making additional assumptions.

5.2 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 5.1 (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 $M_X : \mathbb{R} \rightarrow \mathbb{R}$ defined by $M_X(\lambda) = \mathbb{E}[\exp(\lambda X)]$. The condition in the definition can be written as

$$\log M_X(\lambda) \leq \frac{1}{2} \lambda^2 \sigma^2 \quad \text{for all } \lambda \in \mathbb{R}.$$

Another useful function is the **cumulative generating function**, which is a function $\psi_X : \mathbb{R} \rightarrow \mathbb{R}$ defined by $\psi_X(\lambda) = \log M_X(\lambda)$. The origin of the names of M_X and ψ_X are explained in the notes and Exercise 5.9. It is not hard to see that M_X (or ψ_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

$$\mathbb{E}[\exp(\lambda X)] = \int_0^\infty \underbrace{\exp(-x)}_{\text{density of exponential}} \times \exp(\lambda x) dx = \infty.$$

Therefore the definition of a subgaussian places a nontrivial restriction on the random variables by assuming that the domain of the moment generating function is the whole real line. It is not hard to verify that the moment generating function of a zero-mean Gaussian with variance σ^2 is $\exp(\lambda^2 \sigma^2 / 2)$ from which we conclude that a centered Gaussian with standard deviation $\sigma > 0$ is σ -subgaussian.

Where does the term ‘subgaussian’ come from? The following result provides the explanation. 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 5.1 *If X is σ -subgaussian, then for any $\varepsilon \geq 0$,*

$$\mathbb{P}(X \geq \varepsilon) \leq \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right). \quad (5.5)$$

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

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

Now λ was any positive constant, and in particular may be chosen to minimize the bound above, which is achieved by $\lambda = \varepsilon / \sigma^2$. \square

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 $\mathbb{P}(|X| \geq \varepsilon) \leq 2 \exp(-\varepsilon^2 / (2\sigma^2))$. An equivalent form of these bounds is:

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

This form is often more convenient and especially the latter, which for small δ shows that with overwhelming probability the random variable 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 tail behavior of $\hat{\mu} - \mu$, we need one more lemma (the first item of the lemma is included for completeness):

LEMMA 5.2 *Suppose that X is σ -subgaussian and X_1 and X_2 are independent and σ_1 and σ_2 -subgaussian respectively, then:*

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

The proof of the lemma is left to the reader (Exercise 5.7). Note that if X_1 and X_2 are *not* independent then $X_1 + X_2$ is still guaranteed to be $(\sigma_1 + \sigma_2)$ -subgaussian. The difference between this and $\sqrt{\sigma_1^2 + \sigma_2^2}$ is the price of losing independence.

With this we are ready for our key concentration inequality. In particular, combining Lemma 5.2 and Theorem 5.1 leads to a very straightforward analysis of the tails of $\hat{\mu} - \mu$ under the assumption that $X_i - \mu$ are σ -subgaussian. Since X_i are assumed to be independent, by the lemma it holds that $\hat{\mu} - \mu = \sum_{i=1}^n (X_i - \mu)/n$ is σ/\sqrt{n} -subgaussian.

COROLLARY 5.1 *Assume that $X_i - \mu$ are independent, σ -subgaussian random variables. Then for any $\varepsilon \geq 0$*

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

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

By the inequality $\exp(-x) \leq 1/(ex)$ (which holds for all $x > 0$) we can see that except for a very small ε the above inequality is strictly stronger than what we obtained via Chebyshev's inequality and exponentially smaller (tighter) if $n\varepsilon^2$ is large relative to σ^2 .

The alternative deviation form of the above result says that under the conditions of the result, for any $\delta \in [0, 1]$, with probability at least $1 - \delta$,

$$\mu \leq \hat{\mu} + \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}. \quad (5.6)$$

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

$$\mu \geq \hat{\mu} - \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}. \quad (5.7)$$

Again, one can use a union bound to derive a two-sided inequality.

Before we finally return to bandits, one might be wondering what variables are subgaussian? We give three basic examples. First, as was already mentioned, if X is distributed like a Gaussian with zero mean and variance σ^2 , then X is

σ -subgaussian. Second, if X is bounded, zero-mean (i.e., $\mathbb{E}[X] = 0$ and $|X| \leq B$ almost surely for some $B \geq 0$) then X is B -subgaussian. A special case is when X is a shifted Bernoulli with $\mathbb{P}(X = 1 - p) = p$ and $\mathbb{P}(X = -p) = 1 - p$. In this case it also holds that X is $1/2$ -subgaussian. Finally, recall that the exponential serves the role of a classical distribution that is *not* subgaussian (instead it is sub-exponential, but we will not concern ourselves with this).



For random variables that are not **centered** ($\mathbb{E}[X] \neq 0$) we will abuse notation by saying that X is σ -subgaussian if the **noise** $X - \mathbb{E}[X]$ is σ -subgaussian. A distribution is called σ -subgaussian if a random variable drawn from that distribution is σ -subgaussian. In fact, the subgaussianity property is really a property of both a random variable and the measure on the space on which it is defined, so the nomenclature is doubly abused.

5.3 Notes

- 1 The Berry-Esseen Theorem (independently discovered by [Berry \[1941\]](#) and [Esseen \[1942\]](#)) quantifies the speed of convergence in the CLT. It essentially says that the distance between the Gaussian and the actual distribution decays at a rate of $1/\sqrt{n}$ under some mild assumptions (see [Exercise 5.5](#)). This is known to be tight for the class of probability distributions that appear in the Berry-Esseen result. However, this is a vacuous result when the tail probabilities themselves are much smaller than $1/\sqrt{n}$. Hence the need for concrete finite-time results.
- 2 [Theorem 5.1](#) shows that subgaussian random variables have tails that decay almost as fast as a Gaussian. A version of the converse is also possible. That is, if a centered random has tails that behave in a similar way to a Gaussian, then it is subgaussian. In particular, the following holds: Let X be a centered random variable ($\mathbb{E}[X] = 0$) with $\mathbb{P}(|X| \geq \varepsilon) \leq 2 \exp(-\varepsilon^2/2)$. Then X is

$\sqrt{5}$ -subgaussian:

$$\begin{aligned} \mathbb{E}[\exp(\lambda X)] &= \mathbb{E}\left[\sum_{i=0}^{\infty} \frac{\lambda^i X^i}{i!}\right] \leq 1 + \sum_{i=2}^{\infty} \mathbb{E}\left[\frac{\lambda^i |X|^i}{i!}\right] \\ &\leq 1 + \sum_{i=2}^{\infty} \int_0^{\infty} \mathbb{P}\left(|X| \geq \frac{i^{1/i}}{\lambda} x^{1/i}\right) dx && \text{(Exercise 2.18)} \\ &\leq 1 + 2 \sum_{i=2}^{\infty} \int_0^{\infty} \exp\left(-\frac{i^{2/i} x^{2/i}}{2\lambda^2}\right) dx && \text{(by assumption)} \\ &= 1 + \sqrt{2\pi}\lambda \left(\exp(\lambda^2/2) \left(1 + \operatorname{erf}\left(\frac{\lambda}{\sqrt{2}}\right)\right) - 1\right) && \text{(by Mathematica)} \\ &\leq \exp\left(\frac{5\lambda^2}{2}\right). \end{aligned}$$

This bound is surely loose. At the same time, there is little room for improvement: If X has density $p(x) = |x| \exp(-x^2/2)/2$, then $\mathbb{P}(|X| \geq \varepsilon) = \exp(-\varepsilon^2/2)$. And yet X is at best $\sqrt{2}$ -subgaussian, so some degree of slack is required (see Exercise 5.4).

- 3 The classical CLT only applies to sequences of independent and identically distributed random variables with finite variance. It turns out that these conditions can be relaxed significantly. One such relaxation is the removal of the condition that the sequence be identically distributed. A CLT-like result still holds under **Lindeberg's condition**, which ensures that the variance is not caused by increasingly infrequent (and catastrophic events). Formally, let $(X_t)_t$ be a sequence of independent random variables with means $(\mu_t)_t$ and variances $(\sigma_t^2)_t$ and let $s_n^2 = \sum_{t=1}^n \sigma_t^2$. Then the Lindeberg CLT says that if for all $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \frac{1}{s_n^2} \sum_{t=1}^n \mathbb{E}[(X_t - \mu_t)^2 \mathbb{I}\{|X_t - \mu_t| \geq \varepsilon s_n\}] = 0,$$

then the random variable given by $Z_n = \frac{1}{s_n} \sum_{t=1}^n (X_t - \mu_t)$ converges in distribution to a standard normal distribution. We have little use for this theorem as-is because like the CLT, it only holds asymptotically. It does, however, provide inspiration for what might be possible in finite-time and we will see similar results in subsequent chapters. The interested reader can find much more on this in the classic text by Billingsley [2008], which includes many other generalizations such as the multi-variate case.

- 4 We saw in (5.4) that if X_1, X_2, \dots, X_n are independent standard Gaussian random variables and $\hat{\mu} = \frac{1}{n} \sum_{t=1}^n X_t$, then

$$\mathbb{P}(\hat{\mu} \geq \varepsilon) \leq \sqrt{\frac{\sigma^2}{2\pi n \varepsilon^2}} \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right).$$

If $n\varepsilon^2/\sigma^2$ is relatively large, then this bound is marginally stronger than $\exp(-n\varepsilon^2/(2\sigma^2))$ that follows from the subgaussian analysis. One might ask

whether or not a similar improvement is possible more generally. And Talagrand [1995] will tell you: Yes! At least for bounded random variables (details in the paper).

- 5 The name ‘moment generating function’ comes from the following fact. Suppose that $M_X(\lambda)$ exists in a neighborhood of zero, then all the moments of the underlying random variable can be read out from its derivatives at zero (see Exercise 5.9). Furthermore, the moment generating function in any small neighborhood of zero uniquely determines the distribution of the underlying random variable. In particular, the following holds: Let X and Y be random variables and $a > 0$ and assume that $\text{dom}(M_X) \supseteq [-a, a]$ and $\text{dom}(M_Y) \supseteq [-a, a]$ and $M_X(\lambda) = M_Y(\lambda)$ on $[-a, a]$. Then
- (e) X and Y have the same distribution: $\mathbb{P}(X \geq x) = \mathbb{P}(Y \geq x)$ for all $x \in \mathbb{R}$.
 - (e) Suppose additionally that X_1, X_2, \dots are a sequence of random variables such that $\text{dom}(M_{X_t}) \supseteq [-a, a]$ and $\lim_{t \rightarrow \infty} M_{X_t}(\lambda) = M_X(\lambda)$ for all $\lambda \in [-a, a]$. Then $\lim_{t \rightarrow \infty} \mathbb{P}(X_t \geq x) = \mathbb{P}(X \geq x)$ for all x , which is equivalent to saying that $(X_t)_t$ converges in distribution to X .

The proof of this result does not belong here (but could serve as a challenging exercise). Most probability texts prove the analogous result for the characteristic function (see the next note), which is known as **Lévy’s continuity theorem**, but the above is sometimes also given [Billingsley, 2008, §30]. The significance of these results is that the moment generating function is often more convenient to work with than the distribution. For example if X and Y are independent, then the distribution of $X + Y$ is the convolution of the distributions of X and Y , which, in a way, is a complicated object. The moment generating function, on the other hand, satisfies $M_{X+Y}(\lambda) = M_X(\lambda)M_Y(\lambda)$. To illustrate the usefulness of this, let X, X_1, X_2, \dots, X_n be a sequence of independent random variables with zero mean, unit variance and $M_X(\lambda)$ defined for all $\lambda \in [-a, a] \subset \mathbb{R}$ with $a > 0$ and let $Z_n = \sum_{t=1}^n X_t/\sqrt{n}$. By the multiplicative property above, for any $\lambda \in [-a, a]$, we have

$$\begin{aligned} \lim_{n \rightarrow \infty} \log M_{Z_n}(\lambda) &= \lim_{n \rightarrow \infty} n \log M_X(\lambda/\sqrt{n}) \\ &= \lim_{n \rightarrow \infty} n \log \mathbb{E}[\exp(\lambda X/\sqrt{n})] \\ &= \lim_{n \rightarrow \infty} n \log \left(1 + \frac{\lambda^2}{2n} + \sum_{i=3}^{\infty} \frac{\lambda^i \mathbb{E}[X^i]}{i!n^{i/2}} \right) \\ &= \frac{\lambda^2}{2} \\ &= \log M_Z(\lambda), \end{aligned}$$

where Z is distributed like a standard Gaussian. Therefore the sequence Z_n converges in distribution to Z , which is exactly the statement of the central limit theorem. Of course here we required the additional assumption that $M_X(\lambda)$ was defined over $[-a, a]$ with $a > 0$, which does not normally appear in the statement of the CLT.

- 6 The non-existence of the moment generating function is one of the motivations to introduce the **characteristic function**, which is defined as $\phi_X(\lambda) = \mathbb{E}[\exp(\lambda i X)]$ with $i = \sqrt{-1}$ being the imaginary unit and which always exists and shares many properties with the moment generating function. An example application of characteristic functions is the classical proof of the central limit theorem. Mathematically inclined readers will notice that the moment generating function of random variable X is the Laplace transform of $-X$ (and the characteristic function is the Fourier transform). There are many books on these topics, so we'll just mention that they are essential tools for a probabilist and leave it at that.
- 7 Hoeffding's lemma states that for a zero-mean random variable X such that $X \in [a, b]$ almost surely for real values $a < b$, then $M_X(\lambda) \leq \exp(\lambda^2(b-a)^2/8)$. Applying Chernoff's method shows that if X_1, X_2, \dots, X_n are independent and $X_t \in [a_t, b_t]$ almost surely with $a_t < b_t$ for all t , then

$$\mathbb{P}\left(\sum_{t=1}^n (X_t - \mathbb{E}[X_t]) \geq \varepsilon\right) \leq \exp\left(\frac{2n^2\varepsilon^2}{\sum_{t=1}^n (b_t - a_t)^2}\right). \quad (5.8)$$

For details see Exercise 5.13. There are many variants of this result that provide tighter bounds when X satisfies certain additional distributional properties. For example, if X has small variance, then **Bernstein's inequality** supplies a useful improvement. For details see the texts mentioned below.

- 8 The Cramer-Chernoff method is applicable beyond the subgaussian case, even when the moment generating function is not defined globally. One example where this occurs is when X_1, X_2, \dots, X_n are independent standard Gaussian and $Y = \sum_{i=1}^n X_i^2$. Then Y has a χ^2 -distribution with n degrees of freedom. An easy calculation shows that $M_Y(\lambda) = (1 - 2\lambda)^{-n/2}$ for $\lambda \in [0, 1/2)$ and $M_Y(\lambda)$ is undefined for $\lambda \geq 1/2$. By the Cramer-Chernoff method we have

$$\begin{aligned} \mathbb{P}(Y \geq n + \varepsilon) &\leq \inf_{\lambda \in [0, 1/2)} M_\lambda(Y) \exp(-\lambda(n + \varepsilon)) \\ &\leq \inf_{\lambda \in [0, 1/2)} \left(\frac{1}{1 - 2\lambda}\right)^{\frac{n}{2}} \exp(-\lambda(n + \varepsilon)) \end{aligned}$$

Choosing $\lambda = \frac{1}{2} - \frac{n}{2(n+\varepsilon)}$ leads to $\mathbb{P}(Y \geq n + \varepsilon) \leq \left(1 + \frac{\varepsilon}{n}\right)^{\frac{n}{2}} \exp\left(-\frac{\varepsilon}{2}\right)$, which turns out to be about the best you can do [Laurent and Massart, 2000].

- 9 Distributions for which the moment generating function is infinite for all $\lambda > 0$ are called **heavy tailed**. Distributions that are not heavy tailed are **light tailed**.

5.4 Bibliographical remarks

We will be returning to concentration of measure many times, but note here that it is an interesting (and still active) topic of research. What we have seen is only the tip of the iceberg. Readers that are interested to dive into this exciting

field might enjoy the book by [Boucheron et al. \[2013\]](#). For matrix versions of many standard results there is a recent book by [Tropp \[2015\]](#). The survey of [McDiarmid \[1998\]](#) has many of the classic results. We'll often see concentration of the empirical mean for random variables that are not quite independent and very far from identically distributed. There is a useful type of concentration bound that are 'self-normalized' by the variance. A nice book on this is by [Peña et al. \[2008\]](#). Another tool that is occasionally useful for deriving concentration bounds in more unusual setups is called **empirical process theory**. There are several references for this, including those by [van de Geer \[2000\]](#) or [Dudley \[2014\]](#). Of course these are just a few of the many textbooks (not to mention the papers). We will return to concentration many times throughout the book, so more details will follow, especially on martingales.

5.5 Exercises

There are too many candidate exercises to list. We heartily recommend *all* the exercises in Chapter 2 of the book by [Boucheron et al. \[2013\]](#).

5.1 Let X_1, X_2, \dots, X_n be a sequence of independent and identically distributed random variables with mean μ and variance $\sigma^2 < \infty$. Let $\hat{\mu} = \frac{1}{n} \sum_{t=1}^n X_t$ and show that $\mathbb{V}[\hat{\mu}] = \mathbb{E}[(\hat{\mu} - \mu)^2] = \sigma^2/n$.

5.2 Prove Markov's inequality (Lemma 5.1).

5.3 Compare the Gaussian tail probability bound on the right-hand side of (5.4) and the one on (5.2). What values of ε make one smaller than the other? Discuss your findings.

5.4 Let X be a random variable on \mathbb{R} with density with respect to the Lebesgue measure of $p(x) = |x| \exp(-x^2/2)/2$. Show that:

- (a) $\mathbb{P}(|X| \geq \varepsilon) = \exp(-\varepsilon^2/2)$.
- (b) That X is not $\sqrt{(2-\varepsilon)}$ -subgaussian for any $\varepsilon > 0$.

5.5 Let X_1, X_2, \dots, X_n be a sequence of independent and identically distributed random variables with mean μ , variance σ^2 and bounded third absolute moment

$$\rho = \mathbb{E}[|X_1 - \mu|^3] < \infty.$$

Let $S_n = \sum_{t=1}^n (X_t - \mu)/\sigma$. The Berry-Esseen theorem shows that

$$\left| \mathbb{P}\left(\frac{S_n}{\sqrt{n}} \geq x\right) - \underbrace{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp(-y^2/2) dy}_{\Phi(x)} \right| \leq \frac{C\rho}{\sqrt{n}},$$

where $C < 1/2$ is a universal constant.

- (a) Let $\hat{\mu}_n = \frac{1}{n} \sum_{t=1}^n X_t$ and derive a tail bound from the Berry-Esseen theorem. That is, give a bound of the form $\mathbb{P}(\hat{\mu}_n \geq \mu + \varepsilon)$ for positive values of ε .
- (b) Compare your bound with the one that can be obtained from the Cramer-Chernoff method. Argue pro- and contra- for the superiority of one over the other.

5.6 We mentioned that invoking the central limit theorem to approximate the distribution of sums of independent Bernoulli random variables using a Gaussian can be a bad idea. Let $X_1, \dots, X_n \sim \mathcal{B}(p)$ be independent Bernoulli random variables with common mean $p = \lambda/n$ where $\lambda \in (0, 1)$. For $x \in \mathbb{N}$ natural number, let $P_n(x) = \mathbb{P}(X_1 + \dots + X_n = x)$.

- (a) Show that $\lim_{n \rightarrow \infty} P_n(x) = e^{-\lambda} \lambda^x / (x!)$, which is a Poisson distribution with parameter λ .
- (b) Explain why this does not contradict the CLT and discuss the implications of the Berry-Esseen.
- (c) In what way does this show that the CLT is indeed a poor approximation in some cases?
- (d) Based on Monte-Carlo simulations, plot the distribution of $X_1 + \dots + X_n$ for $n = 30$ and some well-chosen values of λ . Compare the distribution to what you would get from the CLT. What can you conclude?

5.7 Prove Lemma 5.2.



Use Taylor series.

5.8 Let X_i be σ_i -subgaussian for $i \in \{1, 2\}$ with $\sigma_i \geq 0$. Prove that $X_1 + X_2$ is $(\sigma_1 + \sigma_2)$ -subgaussian. Do *not* assume independence of X_1 and X_2 .

5.9 [Properties of moment/cumulant generating functions] Let X be a real-valued random variable and let $M_X(\lambda) = \mathbb{E}[\exp(\lambda X)]$ be its moment-generating function defined over $\text{dom}(M_X) \subset \mathbb{R}$ where the expectation takes on finite values. Show that the following properties hold:

- (a) M_X is convex and in particular $\text{dom}(M_X)$ is an interval containing zero.
- (b) $M_X(\lambda) \geq e^{\lambda \mathbb{E}[X]}$ for all $\lambda \in \text{dom}(M_X)$.
- (c) For any λ in the interior of $\text{dom}(M_X)$, M_X is infinitely many times differentiable.
- (d) Let $M_X^{(k)}(\lambda) = \frac{d^k}{d\lambda^k} M_X(\lambda)$. Then, for λ in the interior of $\text{dom}(M_X)$, $M^{(k)}(\lambda) = \mathbb{E}[X^k \exp(\lambda X)]$.
- (e) Assuming 0 is in the interior of $\text{dom}(M_X)$, $M_X^{(k)}(0) = \mathbb{E}[X^k]$ (hence the name of M_X).
- (f) ψ_X is convex (that is, M_X is log-convex).



For part (a) use the convexity of $x \mapsto e^x$.

5.10 Let X, X_1, X_2, \dots, X_n be a sequence of independent and identically distributed random variables with zero mean and moment generating function M_X with $\text{dom}(M_X) = \mathbb{R}$. Let $\hat{\mu}_n = \frac{1}{n} \sum_{t=1}^n X_t$.

(a) Show that for any $\varepsilon > 0$,

$$\frac{1}{n} \log \mathbb{P}(\hat{\mu}_n \geq \varepsilon) \leq -\psi_X^*(\varepsilon) = -\sup_{\lambda} (\lambda \varepsilon - \log M_X(\lambda)). \quad (5.9)$$

(b) Let $\sigma^2 = \mathbb{V}[X]$. The central limit theorem says that for any $x \in \mathbb{R}$,

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(\hat{\mu}_n \sqrt{\frac{n}{\sigma^2}} \geq x\right) = \Phi(x),$$

where $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp(-y^2/2) dy$ is the cumulative distribution of the standard Gaussian. Let Z be a random variable distributed like a standard Gaussian. A careless application of this result might suggest that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}(\hat{\mu}_n \geq \varepsilon) \stackrel{?}{=} \lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}\left(Z \geq \varepsilon \sqrt{\frac{n}{\sigma^2}}\right).$$

Evaluate the right-hand side and explain why the question-marked equality does *not* hold.



As it happens, the inequality in (5.9) may be replaced by an equality as $n \rightarrow \infty$. The assumption that the moment-generating function exists everywhere may be relaxed significantly. We refer the interested reader to the classic text by [Dembo and Zeitouni \[2009\]](#). The function ψ_X^* is called the **Legendre transform**, **convex conjugate** or **Fenchel dual** of the convex function ψ_X . Convexity will play a role in some of the later chapters and will be discussed in more detail then. The standard reference is by [Rockafellar \[2015\]](#).

5.11 A **Rademacher** random variable X satisfies $X \in \{-1, 1\}$ with $\mathbb{P}(X = 1) = 1/2$ and $\mathbb{P}(X = -1) = 1/2$. Prove that $M_X(\lambda) = \cosh(\lambda) \leq \exp(\lambda^2/2)$.

5.12 Prove that if X is zero-mean and $a \leq X \leq b$ almost surely for some $a < 0 < b$, then X is $(b - a)$ -subgaussian.



Let X' be an independent copy of X and consider $M_{X-X'}$. Use $M_X \geq 1$, which follows from Exercise 5.9 and then use $M_{X-X'} = M_{\varepsilon(X-X')}$ where $\varepsilon \in \{-1, +1\}$ is a Rademacher random variable independent of (X, X') . Then use the result of the previous exercise.



Since $Y = X - X'$ has the same distribution as $-Y$ (that is, Y has a symmetric distribution), the trick of considering Y instead of X is called the **symmetrization** trick or symmetrization device. The symmetrization argument is useful in a variety of contexts, but may not give the best possible constants. In fact, in the next exercise, you are asked to sharpen the result of Exercise 5.12.

5.13 [Hoeffding's lemma] Suppose that X is zero-mean and $X \in [a, b]$ almost surely for constants $a < b$.

- Show that X is $(b - a)/2$ -subgaussian.
- Prove Hoeffding's inequality (5.8).



For part (a) it suffices to prove that $\psi_X(\lambda) \leq \lambda^2(b - a)^2/4$. By Taylor's theorem, for some λ' between 0 and λ , $\psi_X(\lambda) = \psi_X(0) + \psi_X'(0)\lambda + \psi_X''(\lambda')\lambda^2/2$. To bound the last term, introduce the distribution P_λ for $\lambda \in \mathbb{R}$ arbitrary: $P_\lambda(dz) = e^{-\psi_X(\lambda)} e^{\lambda z} P(dz)$. Show that $\psi_X''(\lambda) = \mathbb{V}[Z]$ where $Z \sim P_\lambda$. Now, since $Z \in [a, b]$ with probability one, argue (without relying on $\mathbb{E}[Z]$) that $\mathbb{V}[Z] \leq (b - a)^2/4$.

5.14 Let X_p be a Bernoulli distribution with mean p , which means that $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$.

- Show that X_p is $1/2$ -subgaussian for all p .
- Let $Q : [0, 1] \rightarrow [0, 1/2]$ be the function given by $Q(p) = \sqrt{\frac{1-2p}{2 \ln((1-p)/p)}}$ where undefined points are defined in terms of their limits. Show that X_p is $Q(p)$ -subgaussian.
- Plot $Q(p)$ as a function of p . How does it compare to $\sqrt{\mathbb{V}[X_p]} = \sqrt{p(1-p)}$?



Readers looking for a hint to part (b) in the previous exercise might like to look at the short paper by [Ostrovsky and Sirota \[2014\]](#).

5.15 In this question we try to understand the concentration of the empirical mean for Bernoulli random variables. Let X_1, X_2, \dots, X_n be independent Bernoulli random variables with mean $p \in [0, 1]$ and $\hat{p}_n = \sum_{t=1}^n X_t/n$. Let Z_n be normally distributed random variable with mean p and variance $p(1-p)/n$.

- Write down expressions for $\mathbb{E}[\hat{p}_n]$ and $\mathbb{V}[\hat{p}_n]$.
- What does the central limit theorem say about the relationship between \hat{p}_n and Z_n as n gets large?
- For each $p \in \{1/10, 1/2\}$ and $\delta = 1/100$ and $\Delta = 1/10$ find the minimum n such that $\mathbb{P}(\hat{p}_n \geq p + \Delta) \leq \delta$.

(d) Let $p = 1/10$ and $\Delta = 1/10$ and

$$\begin{aligned} n_1(\delta, p, \Delta) &= \min \{n : \mathbb{P}(\hat{p}_n \geq p + \Delta) \leq \delta\} \\ n_2(\delta, p, \Delta) &= \min \{n : \mathbb{P}(Z_n \geq p + \Delta) \leq \delta\}. \end{aligned}$$

(i) Evaluate empirically or analytically the value of

$$\lim_{\delta \rightarrow 0} \frac{n_1(\delta, 1/10, 1/10)}{n_2(\delta, 1/10, 1/10)}$$

(ii) In light of the central limit theorem, explain why the answer you got in (i) was not 1.

5.16 Let X_1, \dots, X_n be a sequence of independent random variables with $X_t - \mathbb{E}[X_t] \leq b$ almost surely and $S_n = \sum_{t=1}^n (X_t - \mathbb{E}[X_t])$ and $V_n = \sum_{t=1}^n \mathbb{V}[X_t]$.

- (a) Show that $g(x) = \frac{1}{2} + \frac{x}{3!} + \frac{x^2}{4!} + \dots = (\exp(x) - 1 - x)/x^2$ is monotone increasing.
- (b) Let X be a random variable with $\mathbb{E}[X] = 0$ and $X \leq b$ almost surely. Show that $\mathbb{E}[\exp(X)] \leq 1 + g(b)\mathbb{V}[X]$.
- (c) Prove that $(1 + \alpha) \log(1 + \alpha) - \alpha \geq \frac{3\alpha^2}{6+2\alpha}$ for all $\alpha \geq 0$.
- (d) Let $\varepsilon > 0$ and $\alpha = b\varepsilon/V$ and prove that

$$\begin{aligned} \mathbb{P}\left(\sum_{t=1}^n (X_t - \mathbb{E}[X_t]) \geq \varepsilon\right) &\leq \exp\left(-\frac{V_n}{b^2} ((1 + \alpha) \log(1 + \alpha) - \alpha)\right) \\ &\leq \exp\left(-\frac{\varepsilon^2}{2V(1 + \frac{b\varepsilon}{3V})}\right). \end{aligned} \quad (5.10)$$

(e) Use the previous result to show that

$$\mathbb{P}\left(S_n \geq \sqrt{2 \sum_{t=1}^n \mathbb{V}[X_t] \log\left(\frac{1}{\delta}\right)} + \frac{2b}{3n} \log\left(\frac{1}{\delta}\right)\right) \leq \delta.$$


(f) What can be said if X_1, \dots, X_n are Gaussian? Discuss empirically or theoretically whether or not a dependence on b is avoidable or not.



The bound in Eq. (5.10) is called Bernstein's inequality. There are several generalizations, the most notable of which is the martingale version that slightly relaxes the independence assumption. We will see martingale techniques in Chapter 20. Another useful variant (under slightly different conditions) replaces the actual variance with the empirical variance. This is useful in the common case that the variance is unknown. For more on this see papers by [Audibert et al. \[2007\]](#), [Mnih et al. \[2008\]](#), [Maurer and Pontil \[2009\]](#) or skip ahead to Exercise 7.7.

5.17 Let X_1, X_2, \dots, X_n be a sequence of random variables adapted to filtration $\mathbb{F} = (\mathcal{F}_t)_t$. Abbreviate $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$ and $\mu_t = \mathbb{E}_{t-1}[X_t]$. Suppose that $\eta > 0$ satisfies $\eta X_t \leq 1$ almost surely. Prove that

$$\mathbb{P} \left(\sum_{t=1}^n (X_t - \mu_t) \geq \eta \sum_{t=1}^n \mathbb{E}_{t-1}[X_t^2] + \frac{1}{\eta} \log \left(\frac{1}{\delta} \right) \right) \leq \delta.$$


 Use Chernoff's method and the fact that $\exp(x) \leq 1 + x + x^2$ for all $x \leq 1$ and $\exp(x) \geq 1 + x$ for all x .

5.18 Let X_1, \dots, X_n be independent random variables with $\mathbb{P}(X_t \leq x) \leq x$ for each $x \in [0, 1]$ and $t \in [n]$. Prove for any $\varepsilon > 0$ that

$$\mathbb{P} \left(\sum_{t=1}^n \log(1/X_t) \geq \varepsilon \right) \leq \left(\frac{\varepsilon}{n} \right)^n \exp(n - \varepsilon).$$

5.19 Let X_1, \dots, X_n be an independent and identically distributed sequence taking values in $[m]$. For $i \in [m]$ let $p(i) = \mathbb{P}(X_1 = i)$ and $\hat{p}(i) = \frac{1}{n} \sum_{t=1}^n \mathbb{I}\{X_t = i\}$. Show that for any $\delta \in (0, 1)$,

$$\mathbb{P} \left(\|p - \hat{p}\|_1 \geq \sqrt{\frac{2m \log(2/\delta)}{n}} \right) \leq \delta. \quad (5.11)$$

 This is quite a tricky exercise. The result is due to [Weissman et al. \[2003\]](#). It is worth comparing this to what can be obtained from Hoeffding's inequality, which implies for any $i \in [m]$ and $\delta \in (0, 1)$ that with probability $1 - \delta$,

$$|\hat{p}(i) - p(i)| < \sqrt{\frac{2 \log(2/\delta)}{n}}.$$

By a union bound this ensures that with probability $1 - \delta$,

$$\sum_i |\hat{p}(i) - p(i)| < m \sqrt{\frac{2 \log(2m/\delta)}{n}},$$

which is significantly weaker than the upper bound in (5.11). A standard approach to deriving a stronger inequality is to use the fact that $\|p - \hat{p}\|_1 = \sup_{x: \|x\|_\infty \leq 1} \langle p - \hat{p}, x \rangle$. Choose finite subset $S \subset B = [-1, 1]^m$ such that for any point in $x \in B$ there exists a $y \in S$ such that $\|x - y\|_\infty \leq \varepsilon/3$. Let $x^* = \operatorname{argmax}_{x \in B} \langle p - \hat{p}, x \rangle$ and $s = \operatorname{argmin}_{u \in S} \|s - x^*\|_\infty$. Then $\langle p - \hat{p}, x^* \rangle = \langle p - \hat{p}, s \rangle + \langle p - \hat{p}, s - x^* \rangle \leq \langle p - \hat{p}, s \rangle + 2 \sup_{p: \|p\|_1 \leq 1} \langle p, s - x^* \rangle = \langle p - \hat{p}, s \rangle + 2\|s - x^*\|_\infty = \langle p - \hat{p}, s \rangle + 2\varepsilon/3$. By applying Hoeffding's inequality to $\langle p - \hat{p}, s \rangle$ and a union bound we see that if $n \geq 2 \log(|S|/\delta)/\varepsilon^2$, then with probability $1 - \delta$ it holds that $\|p - \hat{p}\|_1 \leq \varepsilon$. Choosing S to have the fewest elements gives a bound similar to that of Lemma 37.2. In particular, S can be

chosen to be the regular grid with stride $\varepsilon/3$, giving $|S| = (6/\varepsilon)^m$. The quantity $\sup_{x \in X} \langle p - \hat{p}, x \rangle$ is called an **empirical process**. Such empirical processes are the subject of extensive study in the field of empirical process theory, which has many applications within statistics, machine learning and also beyond these field in almost all areas of mathematics [Vaart and Wellner, 1996, Dudley, 2014, van de Geer, 2000].

5.20 Let X_1, X_2, \dots, X_n be a sequence of nonnegative random variables adapted to filtration $(\mathcal{F}_t)_{t=0}^n$ such that $\sum_{t=1}^n X_t \leq 1$ almost surely. Prove that for all $x > 1$,

$$\mathbb{P} \left(\sum_{t=1}^n \mathbb{E}[X_t | \mathcal{F}_{t-1}] \geq x \right) \leq f_n(x) = \begin{cases} \left(\frac{n-x}{n-1} \right)^{n-1}, & \text{if } x < n; \\ 0, & \text{if } x \geq n, \end{cases}$$

where the equality serves as the definition of $f_n(x)$.



This problem does not use the techniques introduced in the chapter. Prove that Bernoulli random variables are the worst case and use backwards induction. Although this result is new to our knowledge, a weaker version was derived by Kirschner and Krause [2018] for the analysis of information directed sampling. The bound is tight in the sense that there exists a sequence of random variables and filtration for which equality holds.