1
Chernoff–Hoeffding Bounds

1.1 What Is "Concentration of Measure"?

The basic idea of concentration of measure is well illustrated by the simplest of random experiments, and one lying at the fountainhead of probability theory: coin tossing. If we toss a fair coin once, the result is completely unpredictable – it can be "heads" or "tails" with equal probability. Now suppose that we toss the same coin a large number of times, say, a thousand times. The outcome is now *sharply predictable*! Namely, the number of heads is "very likely to be around 500". This apparent paradox, which is nevertheless familiar to everybody, is an instance of the phenomenon of the concentration of measure – although there are potentially a large number of possibilities, those that are likely to be observed are concentrated in a very narrow range, hence sharply predictable.

In more sophisticated forms, the phenomenon of the concentration of measure underlies much of our physical world. As we know now, the world is made up of microscopic particles that are governed by probabilistic laws – those of quantum and statistical physics. The reason why the macroscopic properties determined by these large ensembles of particles nevertheless appear deterministic when viewed on our larger scales is precisely the concentration of measure: the observed possibilities are concentrated into a very narrow range.

Given the obvious importance of the phenomenon, it is no surprise that large parts of treatises on probability theory are devoted to its study. The various "laws of large numbers" and the "central limit theorem" are some of the most central results of modern probability theory.

In this book we use the phenomenon of concentration of measure in the analysis of probabilistic algorithms. In analogy with the physical situation described earlier, we can argue that the observable behaviour of randomized algorithms is "almost deterministic". In this way, we can obtain the satisfaction

of deterministic results, and at the same time retain the benefits of randomized algorithms, namely their simplicity and efficiency.

In slightly more technical terms, the basic problem we want to study in this book is this: Given a random variable X with mean $\mathtt{E}[X]$, what is the probability that X deviates far from its expectation? Furthermore, we want to understand under what conditions the random variable X stays almost constant or, put in a different way, when large deviations from the mean are highly unlikely. This is the case for the familiar example of repeated coin tosses, but, as we shall see, it is a more general phenomenon.

There are several reasons why classical results from probability theory are somewhat inadequate or inappropriate for studying these questions:

- First and foremost, the results in probability theory are *asymptotic limit laws* applying in the infinite limit case. We are interested in laws that apply in finite cases.
- The probability theory results are often *qualitative*: they ensure convergence in the limit, but do not consider the *rate* of convergence. We are interested in *quantitative* laws that determine the rate of convergence, or at least good bounds on it.
- The laws of probability theory are classically stated under the assumption of *independence*. This is a very natural and reasonable assumption in probability theory, and it greatly simplifies the statement and proofs of the results. However, in the analysis of randomized algorithms, the outcome of which is the result of a complicated interaction of various processes, independence is the exception rather than the rule. Hence, we are interested in laws that are valid even without independence, or when certain known types of dependences are obtained.

We shall now embark on the development of various tools and techniques that meet these criteria.

1.2 The Binomial Distribution

Let us start with an analysis of the simple motivating example of coin tossing. The number of "heads" or successes in repeated tosses of a fair coin is a very important distribution because it models a very basic paradigm of the probabilistic method, namely to repeat experiments to boost confidence.

Let us analyse the slightly more general case of the number of "heads" in n trials with a coin of bias p, with $0 \leq p \leq 1$, i.e. $\mathtt{Pr}[\mathtt{Heads}] = p$ and $\mathtt{Pr}[\mathtt{Tails}] = 1 - p$. This is a random variable $B(n, p)$ whose distribution is

called the *binomial distribution* with parameters n and p:

$$\Pr[B(n,p)=i]=\binom{n}{i}p^i q^{n-i},\quad 0\le i\le n. \tag{1.1}$$

The general problem defined in the previous section now becomes as follows: In the binomial case the expectation is $\mathrm{E}[B(n,p)]=np$; we want to get a bound on the probability that the variable does not deviate too far from this expected value. Are such large deviations unlikely for $B(n,p)$? A direct computation of the probabilities $\Pr[B(n,p)\ge k]=\sum_{i\ge k}\binom{n}{i}p^i q^{n-i}$ is far too unwieldy. However, see Problem 1.2 for a neat trick that yields a good bound. We shall now introduce a general method that successfully solves our problem and is versatile enough to apply to many other problems we will encounter.

1.3 The Chernoff Bound

The random variable $B(n,p)$ can be written as a sum $X:=\sum_{i\in[n]}X_i$, by introducing the indicator random variables X_i, $i\in[n]$ defined by

$$X_i:=\begin{cases}1 & \text{if the } i\text{th trial is a success,}\\ 0 & \text{otherwise.}\end{cases}$$

The basic Chernoff technique that we develop now applies in many situations where such a decomposition as a sum is possible.

The trick is to consider the so-called *moment-generating function* of X, defined as $\mathrm{E}[e^{\lambda X}]$ where $\lambda>0$ is a parameter. By formal expansion of the Taylor series, we see that

$$\begin{aligned}\mathrm{E}[e^{\lambda X}]&=\mathrm{E}\left[\sum_{i\ge 0}\frac{\lambda^i}{i!}X^i\right]\\ &=\sum_{i\ge 0}\frac{\lambda^i}{i!}\mathrm{E}[X^i].\end{aligned}$$

This explains the name as the function $\mathrm{E}[e^{\lambda X}]$ is the exponential generating function of all the moments of X – it "packs" all the information about the moments of X into one function.

Now, for any $\lambda>0$, we have

$$\begin{aligned}\Pr[X>m]&=\Pr[e^{\lambda X}>e^{\lambda m}]\\ &\le\frac{\mathrm{E}[e^{\lambda X}]}{e^{\lambda m}}.\end{aligned} \tag{1.2}$$

The last step follows by *Markov's inequality*: for any non-negative random variable X, $\mathtt{Pr}[X \geq a] \leq \mathtt{E}[X]/a$.

Let us compute the moment-generating function for our example:

$$\begin{aligned}\mathtt{E}[e^{\lambda X}] &= \mathtt{E}[e^{\lambda \sum_i X_i}] \\ &= \mathtt{E}\left[\prod_i e^{\lambda X_i}\right] \\ &= \prod_i \mathtt{E}[e^{\lambda X_i}] \quad \text{by independence} \\ &= (pe^{\lambda} + q)^n. \end{aligned} \tag{1.3}$$

Substituting this back into (1.2), and using the parametrisation $m := (p+t)n$ which leads to a convenient statement of the bound, we get

$$\mathtt{Pr}[X > m] \leq \left(\frac{pe^{\lambda} + q}{e^{\lambda(p+t)}}\right)^n.$$

We can now pick $\lambda > 0$ to minimise the value within the parentheses and by a simple application of calculus, we arrive at the basic Chernoff bound:

$$\begin{aligned}\mathtt{Pr}[X > (p+t)n] &\leq \left(\left(\frac{p}{p+t}\right)^{p+t}\left(\frac{q}{q-t}\right)^{q-t}\right)^n \\ &= \exp\left(-(p+t)\ln\frac{p+t}{p} - (q-t)\ln\frac{q-t}{q}\right)^n. \end{aligned} \tag{1.4}$$

What shall we make of this mess? Certainly, this is not the most convenient form of the bound for use in applications! In Section 1.6 we derive much simpler and more intelligible formulae that can be used in applications. Now, we shall pause for a while and take a short detour to make some remarks on (1.4). This is for several reasons. First, it is the strongest form of the bound. Second, and more importantly, this same bound appears in many other situations. This is no accident for it is a very natural and insightful bound – when properly viewed! For this, we need a certain concept from information theory.

Given two (discrete) probability distributions $\boldsymbol{p} := (p_1, \ldots, p_n)$ and $\boldsymbol{q} := (q_1, \ldots, q_n)$ on a space of cardinality n, the *relative entropy distance* between them, $H(\boldsymbol{p}, \boldsymbol{q})$, is defined by:[1]

$$H(\boldsymbol{p}, \boldsymbol{q}) := \sum_i -p_i \log \frac{p_i}{q_i}.$$

[1] Note that when $\boldsymbol{q}$ is the uniform distribution, this is just the usual *entropy* of the distribution $\boldsymbol{p}$ up to an additive term of $\log n$.

The expression multiplying n in the exponent in (1.4) is exactly the relative entropy distance of the distribution $p+t, q-t$ from the distribution p, q on the two-point space $\{1, 0\}$. So (1.4) seen from the statistician's eye states: the probability of getting the "observed" distribution $\{p+t, q-t\}$ when the *a priori* or *hypothesis* distribution is $\{p, q\}$ falls exponentially in n times the relative entropy distance between the two distributions.

By considering $-X$, we get the same bound symmetrically for $\Pr[X < (p-t)n]$.

1.4 Heterogeneous Variables

As a first example of the versatility of the Chernoff technique, let us consider the situation where the trials are heterogeneous: probabilities of success at different trials need not be the same. In this case, Chvatal's proof in Problem 1.2 is inapplicable, but the Chernoff method works with a simple modification. Let p_i be the probability of success at the ith trial. Then we can repeat the calculation of the moment-generating function $\mathrm{E}[e^{\lambda X}]$ exactly as in (1.3) except for the last line to get

$$\mathrm{E}[e^{\lambda X}] = \prod_i (p_i e^{\lambda} + q_i). \tag{1.5}$$

Recall that the arithmetic–geometric mean inequality states that

$$\frac{1}{n}\sum_{i=1}^{n} a_i \geq \left(\prod_{i=1}^{n} a_i\right)^{1/n}$$

for all $a_i \leq 0$. Now employing the arithmetic–geometric mean inequality, we get

$$\begin{aligned}\mathrm{E}[e^{\lambda X}] &= \prod_i (p_i e^{\lambda} + q_i) \\ &\leq \left(\frac{\sum_i (p_i e^{\lambda} + q_i)}{n}\right)^n \\ &= (pe^{\lambda} + q)^n,\end{aligned}$$

where $p := \sum_i p_i/n$ and $q := 1-p$. This is the same as (1.3) with p taken as the arithmetic mean of the p_i's. The rest of the proof is as before, and we conclude that the basic Chernoff bound (1.4) holds.

1.5 The Hoeffding Extension

A further extension by the same basic technique is possible to heterogeneous variables that need not even be discrete. Let $X := \sum_i X_i$, where each $X_i, i \in [n]$, takes values in $[0, 1]$ and has mean p_i. To calculate the moment-generating function $e^{\lambda X}$, we need, as before, to compute each individual $e^{\lambda X_i}$. This is no longer as simple as it was with the case where X_i took only two values.

However, the following convexity argument gives a simple upper bound. The graph of the function $e^{\lambda x}$ is convex and hence, in the interval $[0, 1]$, lies always below the straight line joining the endpoints $(0, 1)$ and $(1, e^{\lambda})$. This line has the equation $y = \alpha x + \beta$ where $\beta = 1$ and $\alpha = e^{\lambda} - 1$. Thus

$$\begin{aligned} \mathrm{E}[e^{\lambda X_i}] &\leq \mathrm{E}[\alpha X_i + \beta] \\ &= p_i e^{\lambda} + q_i . \end{aligned}$$

Thus we have

$$\mathrm{E}[e^{\lambda X}] \leq \prod_i \mathrm{E}[e^{\lambda X_i}] \leq \prod_i (p_i e^{\lambda} + q_i),$$

which is the same bound as in (1.5), and the rest of the proof is concluded as before.

1.6 Useful Forms of the Bound

The following forms of the Chernoff–Hoeffding bound are most useful in applications (see also Problem 1.6).

Theorem 1.1. *Let $X := \sum_{i \in [n]} X_i$ where $X_i, i \in [n]$, are independently distributed in $[0, 1]$. Then*

- *For all $t > 0$,*

$$\Pr[X > \mathrm{E}[X] + t], \Pr[X < \mathrm{E}[X] - t] \leq e^{-2t^2/n}. \tag{1.6}$$

- *For $\epsilon > 0$,*

$$\begin{aligned} \Pr[X > (1+\epsilon)\mathrm{E}[X]] &\leq \exp\left(-\frac{\epsilon^2}{3}\mathrm{E}[X]\right), \\ \Pr[X < (1-\epsilon)\mathrm{E}[X]] &\leq \exp\left(-\frac{\epsilon^2}{2}\mathrm{E}[X]\right). \end{aligned} \tag{1.7}$$

- *If* $t > 2e\mathrm{E}[X]$, *then*

$$\Pr[X > t] \leq 2^{-t}. \tag{1.8}$$

Proof. We shall manipulate the bound in (1.4). Set

$$f(t) := (p+t)\ln\frac{p+t}{p} + (q-t)\ln\frac{q-t}{q}.$$

We successively compute

$$f'(t) = \ln\frac{p+t}{p} - \ln\frac{q-t}{q}$$

and

$$f''(t) = \frac{1}{(p+t)(q-t)}.$$

Now, $f(0) = 0 = f'(0)$, and furthermore $f''(t) \geq 4$ for all $0 \leq t \leq q$ because $xy \leq 1/4$ for any two non-negative reals summing to 1. Hence by Taylor's theorem with remainder,

$$\begin{aligned} f(t) &= f(0) + f'(0)t + f''(\xi)\frac{t^2}{2!}, \quad 0 < \xi < t \\ &\geq 2t^2. \end{aligned}$$

This gives, after simple manipulations, the bound (1.6).

Now consider $g(x) := f(px)$. Then $g'(x) = pf'(px)$ and $g''(x) = p^2 f''(px)$. Thus, $g(0) = 0 = g'(0)$ and $g''(x) = p^2/(p+px)(q-px) \geq p/(1+x) \geq 2p/3x$. Now by Taylor's theorem, $g(x) \geq px^2/3$. This gives the upper tail in (1.7).

For the lower tail in (1.7), set $h(x) := g(-x)$. Then $h'(x) = -g'(-x)$ and $h''(x) = g''(-x)$. Thus $h(0) = 0 = h'(0)$ and $h''(x) = p^2/(p-px)(q+px) \geq p$. Thus by Taylor's theorem, $h(x) \geq px^2/2$, and this gives the result.

For (1.8), see Problem 1.6. ∎

Often, we would apply the bounds given earlier to a sum $\sum_i X_i$ where we do not know the exact values of the expectations $\mathrm{E}[X_i]$ but only upper or lower bounds on it. In such situations, one can nevertheless apply the Chernoff–Hoeffding bounds with the known bounds instead, as you should verify in the following exercise.

Exercise 1.1. *In this exercise, we explore a very useful extension of the Chernoff–Hoeffding bounds. Suppose* $X := \sum_{i=1}^{n} X_i$ *as in Theorem 1.1, and suppose* $\mu_L \leq \mu \leq \mu_H$. *Show that*

(a) For any $t > 0$,

$$\Pr[X > \mu_H + t], \Pr[X < \mu_L - t] \leq e^{-2t^2/n}. \tag{1.9}$$

(b) For $0 < \epsilon < 1$,

$$\Pr[X > (1+\epsilon)\mu_H] \leq \exp\left(-\frac{\epsilon^2}{3}\mu_H\right),$$

$$\Pr[X < (1-\epsilon)\mu_L] \leq \exp\left(-\frac{\epsilon^2}{2}\mu_L\right).$$

These bounds are useful because often one only has a bound on the expectation. You may need to use the following useful and intuitively obvious fact that we prove in Section 7.4. Let $X_1, \ldots, X_n$ *be independent random variables distributed in* $[0,1]$ *with* $\mathrm{E}[X_i] = p_i$ *for each* $i \in [n]$. *Let* $Y_1, \ldots, Y_n$ *and* $Z_1, \ldots, Z_n$ *be independent random variables with* $\mathrm{E}[Y_i] = q_i$ *and* $\mathrm{E}[Z_i] = r_i$ *for each* $i \in [n]$. *Now suppose* $q_i \leq p_i \leq r_i$ *for each* $i \in [n]$. *Then, if* $X := \sum_i X_i$, $Y := \sum_i Y_i$ *and* $Z := \sum_i Z_i$, *for any* t,

$$\Pr[X > t] \leq \Pr[Z > t], \quad \text{and} \quad \Pr[X < t] \leq \Pr[Y < t].$$

1.7 A Variance Bound

Finally, we give an application of the basic Chernoff technique to develop a form of the bound in terms of the variances of the individual summands – a form that can be considerably sharper than those derived earlier, and one which will be especially useful for applications we encounter in later chapters.

Let us return to the basic Chernoff technique with $X := X_1 + \cdots + X_n$ and $X_i \in [0,1]$ for each $i \in [n]$. Set $\mu_i := \mathrm{E}[X_i]$ and $\mu := \mathrm{E}[X] = \sum_i \mu_i$. Then

$$\begin{aligned}
\Pr[X > \mu + t] &= \Pr\left[\sum_i (X_i - \mu_i) > t\right] \\
&= \Pr[e^{\lambda \sum_i (X_i - \mu_i)} > e^{\lambda t}] \\
&\leq \mathrm{E}[e^{\lambda \sum_i (X_i - \mu_i)}]/e^{\lambda t},
\end{aligned}$$

for each $\lambda > 0$. The last line follows again from Markov's inequality.

We shall now use the simple inequalities $e^x \leq 1 + x + x^2$, for $0 < |x| < 1$, and $e^x \geq 1 + x$. Now, if $\lambda \max(\mu_i, 1 - \mu_i) < 1$ for each $i \in [n]$, we have

$$\begin{aligned}
\mathrm{E}[e^{\lambda \sum_i (X_i - \mu_i)}] &= \prod_i \mathrm{E}[e^{\lambda(X_i - \mu_i)}] \\
&\leq \prod_i \mathrm{E}[1 + \lambda(X_i - \mu_i) + \lambda^2 (X_i - \mu_i)^2] \\
&= \prod_i (1 + \lambda^2 \sigma_i^2) \\
&\leq \prod_i e^{\lambda^2 \sigma_i^2} \\
&= e^{\lambda^2 \sigma^2},
\end{aligned}$$

where σ_i^2 is the variance of X_i for each $i \in [n]$ and σ^2 is the variance of X. Thus,

$$\Pr[X > \mu + t] \leq e^{\lambda^2 \sigma^2} / e^{\lambda t},$$

for λ satisfying $\lambda \max(\mu_i, 1 - \mu_i) < 1$ for each $i \in [n]$. By calculus, take $\lambda := t/2\sigma^2$ and we get the bound

$$\Pr[X > \mu + t] \leq \exp\left(\frac{-t^2}{4\sigma^2}\right),$$

for $t < 2\sigma^2 / \max_i \max(\mu_i, 1 - \mu_i)$.

Exercise 1.2. *Check that for random variables distributed in* $[0, 1]$*, this is of the same form as the Chernoff–Hoeffding bound derived in the previous section up to constant factors in the exponent. You may need to use the fact that for a random variable distributed in the interval* $[a, b]$*, the variance is bounded by* $(b - a)^2/4$.

The following bound is often referred to as Bernstein's inequality:

Theorem 1.2 (Bernstein's inequality)**.** *Let the random variables* $X_1, \ldots, X_n$ *be independent with* $X_i - \mathrm{E}[X_i] \leq b$ *for each* $i \in [n]$*. Let* $X := \sum_i X_i$ *and let* $\sigma^2 := \sum_i \sigma_i^2$ *be the variance of* X*. Then, for any* $t > 0$*,*

$$\Pr[X > \mathrm{E}[X] + t] \leq \exp\left(-\frac{t^2}{2\sigma^2(1 + bt/3\sigma^2)}\right).$$

Exercise 1.3. *Check that for random variables in* $[0, 1]$ *and* $t < 2\sigma^2/b$*, this is roughly the same order bound as we derived earlier.*

In typical applications, the 'error' term $bt/3\sigma^2$ will be negligible. Suppose that the random variables $X_1, \ldots, X_n$ have the same bounded distribution with positive variance c^2, so $\sigma^2 = nc^2$. Then for $t = o(n)$, this bound is $\exp\left(-(1+o(1))t^2/2\sigma^2\right)$, which is consistent with the central limit theorem assertion that in the asymptotic limit, $X - \mathtt{E}[X]$ is normal with mean 0 and variance σ^2.

Exercise 1.4. *Let $X := \sum_i X_i$ where the $X_i, i \in [n]$, are i.i.d. with $\mathtt{Pr}[X_i = 1] = p$ for each $i \in [n]$ for some $p \in [0,1]$. Compute the variance of X and apply and compare the two bounds as well as the basic Chernoff–Hoeffding bound. Check that when $p = 1/2$, all these bounds are roughly the same.*

1.8 Pointers to the Literature

The original technique is from Chernoff [19] although the idea of using the moment-generating function to derive tail bounds is attributed to Bernstein. The extension to continuous variables is due to Hoeffding [43]. Our presentation was much influenced by [65]. The quick derivation in Problems 1.2 and 1.3 is due to Chvátal [22].

1.9 Problems

Problem 1.1. A set of n balls is drawn by sampling with replacement from an urn containing N balls, M of which are red. Give a sharp concentration result for the number of red balls in the sample drawn. ▽

Problem 1.2. In this problem, we outline a simple proof of the Chernoff bound due to Chvátal.

(a) Argue that for all $x \geq 1$, we have

$$\mathtt{Pr}[B(n,p) \geq k] \leq \sum_{i \geq 0} \binom{n}{i} p^i q^{n-i} x^{i-k}.$$

(b) Now use the binomial theorem and thereafter calculus to optimise the value of x. ▽

Problem 1.3 (hypergeometric distribution)**.** A set of n balls is drawn by sampling **without** replacement from an urn containing N balls, M of which are