

# Chapter 1

## Analyzing Randomized Search Heuristics: Tools from Probability Theory

Benjamin Doerr

*Max-Planck-Institut für Informatik,  
Campus E1 4,  
66123 Saarbrücken,  
Germany  
doerr@mpi-inf.mpg.de*

In this chapter, we collect a few probabilistic tools that are useful for analyzing randomized search heuristics. This includes elementary material like Markov, Chebyshev and Chernoff bounds, but also lesser known topics like dealing with sums of random variables that are only close to being independent or a strong lower bound for the time needed by the coupon collector process. Such results, while also of general interest, seem to be particularly useful in the analysis of randomized search heuristics.

### Contents

1.1	Introduction	2
1.2	Prerequisites	3
1.3	Linearity of Expectation	5
1.4	Deviations from the Mean	6
1.4.1	Markov Inequality	6
1.4.2	Chebyshev Inequality	7
1.4.3	Chernoff Bounds for Sums of Independent Random Variables	7
1.4.4	Chernoff Bounds for Geometric Random Variables	9
1.4.5	Chernoff-type Bounds for Independent Variables with Limited Influence	11
1.4.6	Dealing with Non-independent Random Variables	11
1.5	Coupon Collector	15
1.6	Further Tools	17
1.6.1	Talagrand's Inequality	17
1.6.2	Kim-Vu Inequality	18
1.7	Conclusion	19
	References	19

## 1.1. Introduction

Rigorous analyses of randomized search heuristics, naturally, need some understanding of probability theory. However, unlike one would expect at first, most of them are not very complicated. As the reader will see in this book, the art of analyzing randomized search heuristics consists of finding a clever way to apply one of a small collection of simple tools much more often than of having to prove deep theorems in probability theory. This, in fact, is one of the beauties of the theory of randomized search heuristics.

In this chapter, we collect many of these tools. As will be visible, some of them are very basic and could be proven in a few lines. Those which rely on deeper methods, however, often have an easy to use variant that usually is sufficient in most applications. An example for this is the famous Azuma martingale inequality. In its most common use, the bounded differences method, you will not find any martingales.

This chapter aims at serving three purposes.

- For the reader not yet experienced with this field, we aim to provide the necessary material to understand most of the proofs in the following chapters. To this aim, we start at a very elementary level in the first sections. We would also like to highlight three simple results, the linearity of expectation fact (Lemma 1.4), the waiting time argument (Theorem 1.6) and a simple version of the Chernoff bound (Corollary 1.10), that already suffice to prove many interesting results.
- The reader more familiar with the theory of randomized search heuristics might like to find an easily accessible collection of the tools he frequently uses. To this aim, we state many results in different versions. Clearly, often one easily implies the other, but we feel that saving us in future from looking for the right version of, e.g., the Chernoff bound, either by searching in different sources or by again and again deriving the version we need from the one we find in our favorite reference, is a worthwhile aim.
- Finally, we shall also present some results that are lesser known in the community, partially because they recently appeared in papers on theory of classical algorithms, partially because they adapt classical results to the needs of people working on theory of randomized search heuristics. Examples include a lower bound for the coupon collector process with weakly exponential failure probabil-

ity (Theorem 1.24), Chernoff bounds for negatively correlated random variables (Theorem 1.16) and other ways to deal with slightly dependent random variables (Lemmas 1.18 to 1.20).

## 1.2. Prerequisites

We shall assume that the reader has some basic understanding of the concepts of *probability spaces*, *events* and *random variables*. As usual in probability theory and very convenient in analysis of algorithms, we shall almost never explicitly state the probability space we are working in. Hence an intuitive understanding of the notion of a random variable should be enough to follow the remainder of this book. Having said this, here is the one statement that is easier expressed in terms of events rather than random variables.

**Lemma 1.1 (Union bound).** *Let  $E_1, \dots, E_n$  be arbitrary events in some probability space. Then*

$$\Pr\left(\bigcup_{i=1}^n E_i\right) \leq \sum_{i=1}^n \Pr(E_i).$$

While very elementary and often far from a tight estimate, the union bound is often sufficient to prove that with high probability none of a set of bad events does happen. Of course, each of the bad events has to be shown to be sufficiently rare.

Almost all random variables in this book will be *discrete*, that is, they take a finite number of values only (with non-zero probability). As a simple example, consider the random experiment of independently rolling two distinguishable dice. Let  $X_1$  denote the outcome of the first roll, that is, the number between 1 and 6 which the first die displays. Likewise, let  $X_2$  denote the outcome of the second roll. These are already two random variables. We formalize the statement that with probability  $\frac{1}{6}$  the first die shows a one by saying  $\Pr(X_1 = 1) = \frac{1}{6}$ . Also, the probability that both dice show the same number is  $\Pr(X_1 = X_2) = \frac{1}{6}$ . The *complementary event* that they show different numbers, naturally has a probability of  $\Pr(X_1 \neq X_2) = 1 - \Pr(X_1 = X_2) = \frac{5}{6}$ .

We can add random variables (defined over the same probability space), e.g.,  $X := X_1 + X_2$  is the sum of the numbers shown by the two dice, and we can multiply a random variable by a number, e.g.,  $X := 2X_1$  is twice the number shown by the first die.

The most common type of random variable we shall encounter in this book is an extremely simple one called *binary random variable* or *Bernoulli random variable*. It takes the values 0 and 1 only. In consequence, the probability distribution of a binary random variable  $X$  is fully described by its probability  $\Pr(X = 1)$  of being one, since  $\Pr(X = 0) = 1 - \Pr(X = 1)$ .

Binary random variable usually show up as indicator variables for some random event. For example, if the random experiment is a simple roll of a fair (six-sided) die, we may define a random variable  $X$  by setting  $X = 1$ , if the die shows a ‘six’, and  $X = 0$  otherwise. We say that  $X$  is the indicator random variable for the event “die shows a ‘six’ ”.

Indicator random variables are useful for counting. If we roll a die  $n$  times and  $X_1, \dots, X_n$  are the indicator random variables for the events that the corresponding roll showed a ‘six’, then  $\sum_{i=1}^n X_i$  is a random variable describing the number of times we saw a ‘six’ in these  $n$  rolls. In general, a random variable  $X$  that is the sum of  $n$  binary random variables being one all with equal probability  $p$ , is called a *binomial random variable* (with success probability  $p$ ). We have  $\Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$  for all  $k \in \{0, \dots, n\}$ .

A different question is how long we have to wait until we roll a ‘six’. Assume that we have an infinite sequence of die rolls and  $X_1, X_2, \dots$  are the indicator variables for the event that the corresponding roll showed a ‘six’. Then we are interested in the random variable  $Y = \min\{k \in \mathbb{N} \mid X_k = 1\}$ . Again for the general case of all  $X_i$  being one with probability  $p$ , this random variable  $Y$  is called *geometric random variable* (with success probability  $p$ ). We have  $\Pr(Y = k) = (1 - p)^{k-1} p$  for all  $k \in \mathbb{N}$ .

Before continuing with probabilistic tools, let us mention two simple, but highly useful estimates.

**Lemma 1.2.** *For all  $x \in \mathbb{R}$ ,*

$$1 + x \leq e^x. \quad (1.1)$$

While valid for all  $x \in \mathbb{R}$ , naturally this estimate is strongest for  $x$  close to zero. This is demonstrated by the following estimate.

**Lemma 1.3.** *For all  $n \in \mathbb{N}$ ,*

$$\left(1 - \frac{1}{n}\right)^n \leq \frac{1}{e} \leq \left(1 - \frac{1}{n}\right)^{n-1}. \quad (1.2)$$

Note that the left inequality is just a special case of the previous lemma.

### 1.3. Linearity of Expectation

The *expectation* (or mean) of a random variable  $X$  taking values in some set  $\Omega \subseteq \mathbb{R}$  is defined by  $E(X) = \sum_{\omega \in \Omega} \omega \Pr(X = \omega)$ , where we shall always assume that the sum exists and is finite. As a trivial example, we immediately see that if  $X$  is a binary random variable, then  $E(X) = \Pr(X = 1)$ .

An elementary, but very useful property is that expectation is linear.

**Lemma 1.4.** *Let  $X_1, \dots, X_n$  be arbitrary random variables and  $a_1, \dots, a_n \in \mathbb{R}$ . Then*

$$E\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i E(X_i).$$

This fact is usually exploited by writing a complicated random variable as sum of simpler ones and then deriving its expectation from the expectation of the simple random variables. For example, let  $X$  be a binomial random variable with success probability  $p$ , that is, we have  $\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ . Then, as seen above,  $X$  is the sum of independent binary random variables  $X_1, \dots, X_n$ , each satisfying  $\Pr(X_i = 1) = p$ . In consequence,  $E(X) = E(X_1 + \dots + X_n) = E(X_1) + \dots + E(X_n) = np$ . Note that we did not need that the  $X_i$  are independent. We just proved the following.

**Lemma 1.5.** *Let  $X$  be a binomial random variable describing  $n$  trials with success probability  $p$ . Then  $E(X) = np$ .*

Computing the expectation of a geometric random variable is slightly more difficult. Assume that  $X$  is a geometric random variable with success probability  $p$ . Intuitively, we feel that the time needed to see a success is  $1/p$ . This intuition is guided by the fact that after  $1/p$  repetitions of the underlying binary random experiment, the expected number of successes is exactly one.

Note that this does *not* prove our claim. Fortunately the claim is still true, and follows either from standard results in Markov chain theory, or some elementary, though non-trivial, calculations. It is one of the results most often used in the analysis of randomized search heuristics.

**Theorem 1.6 (Waiting time argument).** *Let  $X$  be a geometric random variable with success probability  $p$ . Then  $E(X) = 1/p$ .*

**Proof.** By definition,  $E(X) = \sum_{i=1}^{\infty} i(1-p)^{i-1}p$ . Note that  $\sum_{i=1}^{\infty} (1-p)^{i-1}p = \sum_{i=1}^{\infty} \Pr(X = i) = \Pr(X \in \mathbb{N}) = 1$  by definition of the geometric random variable. We thus compute

$$\begin{aligned} E(X) &= \sum_{i=1}^{\infty} i(1-p)^{i-1}p \\ &= \sum_{i=1}^{\infty} (1-p)^{i-1}p + (1-p) \sum_{i=1}^{\infty} (i-1)(1-p)^{i-2}p \\ &= 1 + (1-p) \sum_{i=2}^{\infty} (i-1)(1-p)^{i-2}p \\ &= 1 + (1-p) \sum_{i=1}^{\infty} i(1-p)^{i-1}p = 1 + (1-p)E(X). \end{aligned}$$

Solving this equation shows that  $E(X) = 1/p$ .  $\square$

## 1.4. Deviations from the Mean

Often, we are interested not so much in the expectation of some random variable, e.g., the run-time of an algorithm, but we would like to have a bound that holds with high probability. Since computing the expectation is so easy, a successful approach is to first compute the expectation and then bound the probability that the random variable exceeds this expectation by a certain amount. Inequalities serving this purpose are called *tail inequalities* or *large deviation inequalities*.

### 1.4.1. Markov Inequality

A simple one-line proof shows the most general deviation bound, valid for *all* non-negative random variables.

**Lemma 1.7 (Markov inequality).** *Let  $X$  be a non-negative random variable. Then for all  $\lambda \geq 1$ ,*

$$\Pr(X \geq \lambda E(X)) \leq \frac{1}{\lambda}.$$

**Proof.** We have  $E(X) = \sum_{\omega} \omega \Pr(X = \omega) \geq \sum_{\omega \geq \mu} \mu \Pr(\omega) = \mu \Pr(X \geq \mu)$  for all  $\mu \geq 0$ . Setting  $\mu = \lambda E(X)$  proves the claim.  $\square$

### 1.4.2. Chebyshev Inequality

For completeness, we quickly state a second elementary inequality, although it seems that is seldom used in the theory of randomized search heuristics. The *variance* of a discrete random variable  $X$  is  $\text{Var}(X) = E((X - E(X))^2) = E(X^2) - E(X)^2$ . Just by definition, it is a measure of how well  $X$  is concentrated around its mean. Applying Markov's inequality to the random variable  $(X - E(X))^2$  easily yields the following.

**Lemma 1.8 (Chebyshev inequality).** *For all  $\lambda > 0$ ,*

$$\Pr(|X - E(X)| \geq \lambda \sqrt{\text{Var}(X)}) \leq \frac{1}{\lambda^2}.$$

Note that Chebyshev's inequality automatically yields two-sided bounds (that is, for both cases that the random variable is larger and smaller than its expectation), as opposed to Markov's inequality (giving just a bound for exceeding the expectation) and the Chernoff bounds in the following section (giving different bounds for both sides). There is a one-sided version of the Chebyshev inequality attributed to Cantelli, replacing the  $\frac{1}{\lambda^2}$  by  $\frac{1}{\lambda^2+1}$ , a gain not adding a lot in most applications.

### 1.4.3. Chernoff Bounds for Sums of Independent Random Variables

Above, we computed that the expectation of a binomial random variable is  $np$ . For example, if we flip a fair coin  $n$  times, we expect to see 'tails'  $n/2$  times. However, we also feel that the actual outcome should be relatively close to this expectation if  $n$  is sufficiently large. That this is true in a very strong sense, can be computed from the probability distribution of this random variable. However, deeper methods show such results for much more general settings. We shall omit most of the proofs and refer the interested reader to, e.g., the standard text book by Alon and Spencer (2008). A condensed, but self-contained proof was given by Hagerup and Rüb (1990).

In the general setting, we assume that our random variable of interest is the sum of independent random variables, not necessarily having the same distribution. The bounds presented below are all known under names like *Chernoff* or *Hoeffding* bounds, referring to the seminal papers by Chernoff (1952) and Hoeffding (1963). Since the first bounds of this type were actually proven by Bernstein (1924), the name Bernstein inequality would be

more appropriate. We shall not be that precise and instead use the most common name *Chernoff inequalities* for all such bounds.

For the readers' convenience, we shall give several versions of these bounds.

**Theorem 1.9.** *Let  $X_1, \dots, X_n$  be independent random variables taking values in  $[0, 1]$ . Let  $X = \sum_{i=1}^n X_i$ .*

- (a) *Let  $\delta \in [0, 1]$ . Then  $\Pr(X \leq (1 - \delta)E(X)) \leq \left(\left(\frac{1}{1-\delta}\right)^{1-\delta} e^{-\delta}\right)^{E(X)}$ .*  
 (b) *Let  $\delta \geq 0$ . Then  $\Pr(X \geq (1 + \delta)E(X)) \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^{E(X)}$ .*

**Corollary 1.10.** *Let  $X_1, \dots, X_n$  be independent random variables taking values in  $[0, 1]$ . Let  $X = \sum_{i=1}^n X_i$ .*

- (a) *Let  $\delta \in [0, 1]$ . Then  $\Pr(X \leq (1 - \delta)E(X)) \leq \exp(-\delta^2 E(X)/2)$ .*  
 (b) *Let  $\delta \geq 0$ . Then  $\Pr(X \geq (1 + \delta)E(X)) \leq \left(\frac{e}{1+\delta}\right)^{(1+\delta)E(X)}$ .*  
 (c) *Let  $\delta \in [0, 1]$ . Then  $\Pr(X \geq (1 + \delta)E(X)) \leq \exp(-\delta^2 E(X)/3)$ .*  
 (d) *For all  $d \geq 6E(X)$ ,  $\Pr(X \geq d) \leq 2^{-d}$ .*

Part (a) and (b) follow from estimating the corresponding expressions in (a) and (b) of Theorem 1.9. Part (c) and (d) are derived from bound (b) in Theorem 1.9, and are often more convenient to work with. The different constants  $1/2$  and  $1/3$  in (a) and (c) cannot be avoided. Ignoring the constants, (a) and (c) tell us that the probability that such a random variable deviates from its expectation by a small constant factor, is exponentially small in the expectation. In consequence, we get inverse polynomial failure probabilities if the expectation is at least logarithmic, and exponentially small ones, if the expectation is  $\Theta(n)$ , e.g., if the  $X_i$  are one with constant probability.

We also see that we obtain useful bounds for  $\delta = o(1)$ . More precisely, from  $\delta \gg 1/\sqrt{E(X)}$  on, the bounds become non-trivial, and for  $\delta \gg \log(n)/\sqrt{E(X)}$  inverse-polynomial.

If  $E(X) = \Theta(n)$ , it is often more convenient to use a Chernoff bound for additive deviations. The following version (which is Theorem 2 in the seminal paper by Hoeffding (1963)) in addition does not require the random variables not be non-negative.

**Theorem 1.11.** *Let  $X_1, \dots, X_n$  be independent random variables. Assume that each  $X_i$  takes values in a real interval of length  $c_i$  only. Let  $X =$*

$\sum_{i=1}^n X_i$ . Then for all  $\lambda > 0$ ,

$$\Pr(X \geq E(X) + \lambda) \leq \exp\left(-2\lambda^2 / \sum_i c_i^2\right),$$

$$\Pr(X \leq E(X) - \lambda) \leq \exp\left(-2\lambda^2 / \sum_i c_i^2\right).$$

Using similar arguments as those that prove Theorem 1.11, Hoeffding (1963) also proves a bound that takes into account the variance of the random variables.

**Theorem 1.12.** *Let  $X_1, \dots, X_n$  be independent random variables. Let  $b$  be such that  $X_i \leq E(X_i) + b$  for all  $i = 1, \dots, n$ . Let  $X = \sum_{i=1}^n X_i$ . Let  $\sigma^2 = \frac{1}{n} \sum_{i=1}^n \text{Var}(X_i)$ . Then*

$$\Pr(X \geq E(X) + \lambda) \leq \left( \left(1 + \frac{b\lambda}{n\sigma^2}\right)^{-\left(1 + \frac{b\lambda}{n\sigma^2}\right) \frac{\sigma^2}{b^2 + \sigma^2}} \left(1 - \frac{\lambda}{nb}\right)^{-\left(1 - \frac{\lambda}{nb}\right) \frac{b^2}{b^2 + \sigma^2}} \right)^n$$

$$\leq \exp\left(-\frac{\lambda}{b} \left( \left(1 + \frac{n\sigma^2}{b\lambda}\right) \ln\left(1 + \frac{b\lambda}{n\sigma^2}\right) - 1 \right)\right).$$

As discussed in Hoeffding (1963), the second bound is the same as inequality (8b) in Bennett (1962). Using the abbreviations  $r = \lambda/b$  and  $s = b\lambda/(n\sigma^2)$ , the bound of Bennett (1962) becomes  $\exp(-r((1 + 1/s) \ln(1 + s) - 1))$ . This is stronger than the bound of  $\exp(-rs/(2 + \frac{2}{3}s)) = \exp(-\lambda^2/(2n\sigma^2(1 + b\lambda/(3n\sigma^2))))$  due to Bernstein (1924) and the bound of  $\exp(-r \arcsin(s/2)/2)$  due to Prokhorov (1956).

We end this subsection with a real gem from Hoeffding’s paper, which, in spite of its usefulness, seems rarely known. It states that the last two theorems also hold, simultaneously, for all partial sums of the sequence  $X_1, \dots, X_n$ .

**Theorem 1.13.** *In the settings of Theorem 1.11 or 1.12, let  $Y_k := \sum_{i=1}^k X_i$ . Let  $X = \max\{Y_k \mid 1 \leq k \leq n\}$ . Then the bounds of Theorem 1.11 and 1.12 also hold for  $X$  defined like this.*

#### 1.4.4. Chernoff Bounds for Geometric Random Variables

If  $X$  is a sum of independent geometric random variables  $X_1, \dots, X_n$ , we would still think that  $X$  displays some concentration behavior similar to that guaranteed by Chernoff bounds. It is true that the  $X_i$  are unbounded, but the probability  $\Pr(X_i = k)$  decays exponentially with  $k$ . Therefore, we

would feel that an  $X_i$  taking a large valued should be a sufficiently rare event. In fact, the following is true.

**Theorem 1.14.** *Let  $p \in ]0, 1[$ . Let  $X_1, \dots, X_n$  be independent geometric random variables with  $\Pr(X_i = j) = (1 - p)^{j-1}p$  for all  $j \in \mathbb{N}$  and let  $X := \sum_{i=1}^n X_i$ .*

*Then for all  $\delta > 0$ ,*

$$\Pr(X \geq (1 + \delta)E(X)) \leq \exp\left(-\frac{\delta^2(n-1)}{2(1+\delta)}\right).$$

**Proof.** Let  $Y_1, Y_2, \dots$  be an infinite sequence of independent, identically distributed binary random variables such that  $Y_i$  is one with probability  $\Pr(Y_i = 1) = p$ . Note that the random variable “smallest  $j$  such that  $Y_j = 1$ ” has the same distribution as each  $X_i$ . In consequence,  $X$  has the same distribution as “smallest  $j$  such that exactly  $n$  of the variables  $Y_1, \dots, Y_j$  are one”. In particular,  $\Pr(X \geq j) = \Pr(\sum_{i=1}^{j-1} Y_i \leq n - 1)$  for all  $j \in \mathbb{N}$ . This manipulation reduces our problem to the analysis of independent binary variables and enables us to use the classical Chernoff bounds.

By Theorem 1.6, the expected value of each  $X_i$  is  $E(X_i) = \frac{1}{p}$ . Hence  $E(X) = \frac{n}{p}$ . Let  $Y := \sum_{i=1}^{\lfloor (1+\delta)E(X) - 1 \rfloor} Y_i$ . By the above,

$$\Pr(X \geq (1 + \delta)E(X)) = \Pr(Y \leq n - 1).$$

The expected value of  $Y$  is bounded by

$$E(Y) = \lfloor (1 + \delta)E(X) - 1 \rfloor p \geq (1 + \delta)n - p > (1 + \delta)(n - 1).$$

Now let  $\delta' := 1 - \frac{n-1}{E(Y)}$ . Then  $0 < \delta' \leq 1$  and  $\Pr(Y \leq n - 1) = \Pr(Y \leq (1 - \delta')E(Y))$ . Hence we can apply Corollary 1.10 to get

$$\begin{aligned} \Pr(X \geq (1 + \delta)E(X)) &= \Pr(Y \leq (1 - \delta')E(Y)) \\ &\leq \exp\left(-\frac{1}{2}E(Y) \left(1 - \frac{n-1}{E(Y)}\right)^2\right) \\ &\leq \exp\left(-\frac{1}{2}E(Y) \left(1 - \frac{1}{1+\delta}\right)^2\right) \\ &\leq \exp\left(-\frac{1}{2}(n-1)(1+\delta) \left(\frac{\delta}{1+\delta}\right)^2\right). \end{aligned}$$

□

### 1.4.5. Chernoff-type Bounds for Independent Variables with Limited Influence

Often, the random variable of interest can be expressed as sum of independent random variables. Then, as seen above, Chernoff bounds give excellent estimates on the deviation from the mean. Sometimes, the random variable we are interested in is also determined by the outcomes of many independent random variables, however, not as simply as in the case of a sum. Nevertheless, if each of the independent random variables only has a limited influence on the outcome, then bounds similar to those of Theorem 1.11 can be proven.

**Theorem 1.15 (Azuma’s inequality).** *Let  $X_1, \dots, X_n$  be independent random variables taking values in the sets  $\Omega_1, \dots, \Omega_n$ , respectively. Let  $\Omega := \Omega_1 \times \dots \times \Omega_n$ . Let  $f : \Omega \rightarrow \mathbb{R}$  and  $c_1, \dots, c_n > 0$  be such that for all  $i \in \{1, \dots, n\}$ ,  $\omega, \bar{\omega} \in \Omega$  we have that if for all  $j \neq i$ ,  $\omega_j = \bar{\omega}_j$ , then  $|f(\omega) - f(\bar{\omega})| \leq c_i$ . Let  $X = f(X_1, \dots, X_n)$ . Then for all  $\lambda > 0$ ,*

$$\Pr(X \geq E(X) + \lambda) \leq \exp\left(-2\lambda^2 / \sum_i c_i^2\right),$$

$$\Pr(X \leq E(X) - \lambda) \leq \exp\left(-2\lambda^2 / \sum_i c_i^2\right).$$

Theorem 1.15 is usually known under the name Azuma’s inequality or *method of bounded differences*. It is a special case of a similar bound for martingales due to Azuma (1967), which, however, already appeared in Hoeffding (1963). Again, we will stick to the most common name and not care whether it is the most appropriate one.

The version of Azuma’s inequality given above is due to McDiarmid (1998), while most authors state a slightly weaker bound of  $\exp(-\lambda^2/2 \sum_i c_i^2)$ .

### 1.4.6. Dealing with Non-independent Random Variables

All bounds of Chernoff type presented so far build on a large number of *independent* random variables. Often, in particular, in the analysis of randomized search heuristics, this is too much to ask for. In this subsection, we present two ways to still obtain useful bounds.

The first approach uses negative correlation. Let  $X_1, \dots, X_n$  binary random variables. We call them *negatively correlated*, if for all  $I \subseteq \{1, \dots, n\}$

the following holds.

$$\Pr(\forall i \in I : X_i = 0) \leq \prod_{i \in I} \Pr(X_i = 0),$$

$$\Pr(\forall i \in I : X_i = 1) \leq \prod_{i \in I} \Pr(X_i = 1).$$

In simple words, we require that the event that a set of variables is all zero or all one, is at most as likely as in the case of independent  $X_i$ . It feels that this condition should make life rather easier than in the independent case, and exactly this is what Panconesi and Srinivasan (1997) were able to prove.

**Theorem 1.16 (Chernoff bounds, negative correlation).**

Let  $X_1, \dots, X_n$  be negatively correlated binary random variables. Let  $a_1, \dots, a_n \in [0, 1]$  and  $X = \sum_{i=1}^n a_i X_i$ . Then  $X$  satisfies the Chernoff bounds given in Theorem 1.9 and Corollary 1.10.

Here is an example of how to apply Theorem 1.16, namely to derive Chernoff bounds for *hypergeometric* distributions. Say we choose randomly  $n$  elements from a given  $N$ -element set *without replacement*. For a given  $m$ -element subset of full set, we wonder how many of its elements we have chosen. This random variable is called hypergeometrically distributed with parameters  $N$ ,  $n$  and  $m$ .

More formally, let  $S$  be any  $N$ -element set. Let  $T \subseteq S$  have exactly  $m$  elements. Let  $U$  be a subset of  $S$  chosen uniformly among all  $n$ -element subsets of  $S$ . Then  $X = |U \cap T|$  is a random variable with hypergeometric distribution (with parameters  $N$ ,  $n$  and  $m$ ).

It is easy to see that  $E(X) = |U||T|/|S| = mn/N$ : Enumerate  $T = \{t_1, \dots, t_m\}$  in an arbitrary manner (before choosing  $U$ ). For  $i = 1, \dots, m$ , let  $X_i$  be the binary random variable that is one if and only if  $t_i \in U$ . Clearly,  $\Pr(X_i = 1) = |U|/|S| = n/N$ . Since  $X = \sum_{i=1}^m X_i$ , we have  $E(X) = mn/N$  by linearity of expectation (Lemma 1.4).

It is also obvious that the  $X_i$  are not independent. If  $n < m$  and  $X_1 = \dots = X_n = 1$ , then necessarily we have  $X_i = 0$  for  $i > n$ . Fortunately, however, these dependencies are of the negative correlation type. This is intuitively clear, but also straight-forward to prove. Let  $I \subseteq \{1, \dots, m\}$ . Let  $W = \{t_i \mid i \in I\}$  and  $w = |W| = |I|$ . Then  $\Pr(\forall i \in I : X_i = 1) = \Pr(W \subseteq U)$ . Since  $U$  is uniformly chosen, it suffices to count the number of  $U$  that contain  $W$  (these are  $\binom{|S \setminus W|}{|U \setminus W|}$ ) and compare them with the total

number of possible  $U$ . Hence  $\Pr(W \subseteq U) = \binom{N-w}{n-w} / \binom{N}{n} = \frac{n \cdots (n-w+1)}{N \cdots (N-w+1)} < (n/N)^w = \prod_{i \in I} \Pr(X_i = 1)$ . A similar argument shows that the  $X_i$  also satisfy the first part of the definition of negative correlation.

**Theorem 1.17.** *If  $X$  is a random variable with hypergeometric distribution, then it satisfies the Chernoff bounds given in Theorem 1.9 and Corollary 1.10.*

A second situation often encountered is that a sequence of random variables is not independent, but each member of the sequence has a good chance of having a desired property conditional on each possible outcome of its predecessors. More formally, we have the following.

**Lemma 1.18 (Chernoff bound, lower tail, moderate independence).** *Let  $X_1, \dots, X_n$  be arbitrary binary random variables. Let  $X_1^*, \dots, X_n^*$  be binary random variables that are mutually independent and such that for all  $i$ ,  $X_i^*$  is independent of  $X_1, \dots, X_{i-1}$ . Assume that for all  $i$  and all  $x_1, \dots, x_{i-1} \in \{0, 1\}$ ,*

$$\Pr(X_i = 1 | X_1 = x_1, \dots, X_{i-1} = x_{i-1}) \geq \Pr(X_i^* = 1).$$

Then for all  $k \geq 0$ , we have

$$\Pr\left(\sum_{i=1}^n X_i < k\right) \leq \Pr\left(\sum_{i=1}^n X_i^* < k\right),$$

and the latter term can be bounded by Chernoff bounds for independent random variables.

**Lemma 1.19 (Chernoff bound, upper tail, moderate independence).** *Let  $X_1, \dots, X_n$  be arbitrary binary random variables. Let  $X_1^*, \dots, X_n^*$  be binary random variables that are mutually independent and such that for all  $i$ ,  $X_i^*$  is independent of  $X_1, \dots, X_{i-1}$ . Assume that for all  $i$  and all  $x_1, \dots, x_{i-1} \in \{0, 1\}$ ,*

$$\Pr(X_i = 1 | X_1 = x_1, \dots, X_{i-1} = x_{i-1}) \leq \Pr(X_i^* = 1).$$

Then for all  $k \geq 0$ , we have

$$\Pr\left(\sum_{i=1}^n X_i > k\right) \leq \Pr\left(\sum_{i=1}^n X_i^* > k\right),$$

and the latter term can be bounded by Chernoff bounds for independent random variables.

Both Lemmas are simple corollaries from the following general result, which itself might be useful in the analysis of randomized search heuristics when encountering random variables taking more values than just zero and one.

**Lemma 1.20 (Dealing with moderate dependencies).** *Let  $X_1, \dots, X_n$  be arbitrary integral random variables. Let  $X_1^*, \dots, X_n^*$  be random variables that are mutually independent and such that for all  $i$ ,  $X_i^*$  is independent of  $X_1, \dots, X_{i-1}$ . Assume that for all  $i$  and all  $x_1, \dots, x_{i-1} \in \mathbb{Z}$ , we have  $\Pr(X_i \geq m | X_1 = x_1, \dots, X_{i-1} = x_{i-1}) \leq \Pr(X_i^* \geq m)$  for all  $m \in \mathbb{Z}$ , that is,  $X_i^*$  dominates  $(X_i | X_1 = x_1, \dots, X_{i-1} = x_{i-1})$ . Then for all  $k \in \mathbb{Z}$ , we have*

$$\Pr\left(\sum_{i=1}^n X_i \geq k\right) \leq \Pr\left(\sum_{i=1}^n X_i^* \geq k\right).$$

**Proof.** Define  $P_j := \Pr(\sum_{i=1}^j X_i + \sum_{i=j+1}^n X_i^* \geq k)$  for  $j \in \{0, \dots, n\}$  and  $\mathcal{X}_k^n := \{(x_1, \dots, x_n) \in \mathbb{Z}^n \mid \sum_{i=1}^n x_i = k\}$ . Then

$$\begin{aligned} P_{j+1} &= \Pr\left(\sum_{i=1}^{j+1} X_i + \sum_{i=j+2}^n X_i^* \geq k\right) \\ &= \sum_{m \in \mathbb{Z}} \Pr\left(\sum_{i=1}^j X_i + \sum_{i=j+2}^n X_i^* = k - m \wedge X_{j+1} \geq m\right) \\ &= \sum_{m \in \mathbb{Z}} \sum_{(x_1, \dots, x_j, x_{j+2}, \dots, x_n) \in \mathcal{X}_{k-m}^{n-1}} \Pr(X_1 = x_1, \dots, X_j = x_j) \cdot \\ &\quad \Pr(X_{j+1} \geq m | X_1 = x_1, \dots, X_j = x_j) \cdot \prod_{i=j+2}^n \Pr(X_i^* = x_i) \\ &\leq \sum_{m \in \mathbb{Z}} \Pr\left(\sum_{i=1}^j X_i + \sum_{i=j+2}^n X_i^* = k - m\right) \cdot \Pr(X_{j+1}^* \geq m) \\ &= \Pr\left(\sum_{i=1}^j X_i + \sum_{i=j+1}^n X_i^* \geq k\right) \\ &= P_j. \end{aligned}$$

Thus, we have

$$\Pr\left(\sum_{i=1}^n X_i \geq k\right) = P_n \leq P_{n-1} \leq \dots \leq P_1 \leq P_0 = \Pr\left(\sum_{i=1}^n X_i^* \geq k\right). \quad \square$$

Lemma 1.19 follows directly by noting that its assumptions imply the domination assumption in Lemma 1.20. To prove Lemma 1.18, note that the assumptions there imply that  $-X_i^*$  dominates  $-(X_i|X_1 = x_1, \dots, X_{i-1} = x_{i-1})$  for all  $x_1, \dots, x_{i-1}$ . Hence applying Lemma 1.20 to the negated variables proves the claim.

### 1.5. Coupon Collector

The coupon collector problem is the following simple question. Assume that there are  $n$  types of coupons available. Whenever you buy a certain product, you get one coupon, with its type chosen uniformly at random from the  $n$  types. How long does it take until you have a coupon of each type?

Besides an interesting problem to think about, understanding its solution also pays off when analyzing randomized search heuristics, because very similar situations are encountered here. Since often the setting is similar, but not exactly as in the coupon collector problem, it helps a lot not just to know the coupon collector theorem, but also its proof. We start with a simple version.

**Theorem 1.21 (Coupon collector, expectation).** *The expected time to collect all  $n$  coupons is  $nH_n$ , where  $H_n := \sum_{i=1}^n \frac{1}{i}$  is the  $n$ th harmonic number. Since  $\ln(n) < H_n < 1 + \ln(n)$ , the coupon collector needs an expected time of  $(1 + o(1))n \ln(n)$ .*

**Proof.** Given that we already have  $i$  different coupons for some  $i \in \{0, \dots, n - 1\}$ , the probability that the next coupon is one that we do not already have, is  $(n - i)/n$ . By the waiting time argument (Theorem 1.6), we see that the time  $T_i$  needed to obtain a new coupon given that we have exactly  $i$  different ones, satisfies  $E(T_i) = n/(n - i)$ . Clearly, the total time  $T$  needed to obtain all coupons is  $\sum_{i=0}^{n-1} T_i$ . Hence, by linearity of expectation (Lemma 1.4),  $E(T) = \sum_{i=1}^{n-1} E(T_i) = nH_n$ .

The estimates on  $H_n$  follow from the fact that  $\int_1^{n+1} \frac{1}{x} dx$  is a lower bound for  $H_n$ , and likewise,  $1 + \int_1^n \frac{1}{x} dx$  is an upper bound.  $\square$

As discussed before, the expectation is only one aspect of a random variable, and often we would be happy to also know that the random variable is close to its expectation with good probability. The always applicable Markov inequality tells us that the probability to need more than  $cnH_n$  rounds, is bounded by  $1/c$ .

Chebyshev's inequality (Lemma 1.8) can be used to prove the following two-sided bound.

**Lemma 1.22.** *Let  $T$  denote the time needed to complete the coupon collector process with  $n$  types of coupons. Then  $\Pr(|T - nH_n| \geq cn) \leq 2/c^2$ .*

Stronger bounds for the upper tail can be derived very elementary. Note that the probability that a particular coupon is missed for  $t$  rounds, is  $(1 - \frac{1}{n})^t$ . By a union bound argument (see Lemma 1.1), the probability that there is a coupon not obtained within  $t$  rounds, and hence equivalently, that  $T > t$ , satisfies  $\Pr(T > t) \leq n(1 - \frac{1}{n})^t$ . Using the simple estimate of Lemma 1.2 (or 1.3), we compute  $\Pr(T \geq cn \ln(n)) \leq n \exp(-c \ln n) = n^{-c+1}$ .

**Theorem 1.23.** *Let  $T$  denote the time needed to collect all  $n$  coupons. Then  $\Pr(T \geq (1 + \varepsilon)n \ln(n)) \leq n^{-\varepsilon}$ .*

Note that we used the variable  $c$  to parameterize the deviations in the estimates above, because often a constant  $c$  is a useful choice. However, since there are no asymptotics involved,  $c$  may as well depend on  $n$ . The strongest results in terms of the asymptotics, to be found, e.g., in the book by Motwani and Raghavan (1995), is that for any real constant  $c$ , we have

$$\lim_{n \rightarrow \infty} \Pr(n \ln(n) - cn \leq T \leq n \ln(n) + cn) = e^{-e^{-c}} - e^{-e^c}.$$

For the analysis of randomized search heuristics, however, the non-asymptotic results given previously might be more useful.

Surprisingly, there seem to be no good lower bounds for the coupon collector time published. In the following, we present a simple solution to this problem.

**Theorem 1.24.** *Let  $T$  denote the time needed to collect all  $n$  coupons. Then for all  $\varepsilon > 0$ ,  $\Pr(T < (1 - \varepsilon)(n - 1) \ln(n)) \leq \exp(-n^\varepsilon)$ .*

**Proof.** Let  $t = (1 - \varepsilon)(n - 1) \ln(n)$ . For  $i = 1, \dots, n$ , let  $X_i$  be the indicator random variable for the event that a coupon of type  $i$  is obtained within  $t$  rounds. Note first that  $\Pr(X_i = 1) = 1 - (1 - \frac{1}{n})^t \leq 1 - \exp(-(1 - \varepsilon) \ln(n)) = 1 - n^{-1+\varepsilon}$ , where the estimate follows from Lemma 1.3.

Let  $I \subset \{1, \dots, n\}$ ,  $j \in \{1, \dots, n\} \setminus I$ . By the law of total probability, we have

$$\begin{aligned} \Pr(\forall i \in I : X_i = 1) &= \Pr(\forall i \in I : X_i = 1 | X_j = 1) \cdot \Pr(X_j = 1) + \\ &\Pr(\forall i \in I : X_i = 1 | X_j = 0) \cdot \Pr(X_j = 0). \end{aligned} \quad (1.3)$$

We have  $\Pr(\forall i \in I : X_i = 1 | X_j = 0) \geq \Pr(\forall i \in I : X_i = 1)$ —the left part is the probability to obtain all coupons with types in  $I$  if in each round a random coupon out of  $n - 1$  different ones (including  $I$ ) is drawn. This is clearly easier than achieving the same with  $n$  coupons, which is the right hand side.

From (1.3) we conclude that  $\Pr(\forall i \in I : X_i = 1 | X_j = 1) \leq \Pr(\forall i \in I : X_i = 1)$ . Equivalently, we have  $\Pr(\forall i \in I \cup \{j\} : X_i = 1) \leq \Pr(\forall i \in I : X_i = 1) \cdot \Pr(X_j = 1)$ . By induction, we conclude

$$\begin{aligned} \Pr(\forall i \in \{1, \dots, n\} : X_i = 1) &\leq \prod_{i=1}^n \Pr(X_i = 1) \\ &\leq (1 - n^{-1+\epsilon})^n \\ &\leq \exp(-n^\epsilon) \end{aligned}$$

by Lemma 1.2. □

### 1.6. Further Tools

There are a number of results that, to the best of our knowledge, have not been used in the analysis of randomized search heuristics, but where we see a good chance that they might be useful. Therefore, we briefly describe them in this section.

#### 1.6.1. Talagrand’s Inequality

Consider the setting of Azuma’s inequality (Theorem 1.15), that is, we have a function  $f$  defined over a product of  $n$  probability spaces such that changing a single component of the input to  $f$  has only a limited influence on the function value. Assume for simplicity that all these influences  $c_i$  are at most one.

One weakness of Azuma’s inequality is that does give useful bounds only for deviations  $\lambda$  of size  $\Omega(\sqrt{n})$ . This is different to most Chernoff type bounds, where it often suffices that  $\lambda = \Omega(\sqrt{E(X)})$ . The bound  $\Pr(X \leq (1 - \delta)E(X) \leq \exp(-\delta^2 E(X)/2)$  for example can be rewritten as  $\Pr(X \leq E(X) - \lambda) \leq \exp(-\lambda^2/(2E(X)))$ .

Talagrand’s inequality is a way to overcome this short-coming of Azuma’s inequality. It comes, however, with a certain prize. Still in the notation of Azuma’s theorem, let  $m$  denote a median of  $X = f(X_1, \dots, X_n)$ . Hence  $m$  is a number such that  $\Pr(X \leq m) \geq 1/2$  and  $\Pr(X \geq m) \geq 1/2$ . Assume further that  $f$  has certificates of size  $s$  for exceeding  $m$ . That

means, that for all  $\omega$  with  $f(\omega) \geq m$ , there are  $s$  components of  $\omega$  such that any  $\omega'$  which agrees with  $\omega$  on these components, also satisfies  $f(\omega') \geq m$ . This sounds more obscure than it actually is. If each  $\Omega_i = \{0, 1\}$  and  $f$  is simply the sum of the components, then a certificate for having sum greater than  $m$  are just  $m$  components that are one.

Given that  $f$  is as described above, then a simple version of Talagrand's bound ensures

$$\Pr(X \leq m - \lambda) \leq 2 \exp(-t^2/(4s)).$$

Hence if the certificate size can be chosen of order  $E(X)$ , we obtain bounds of order comparable to Chernoff bounds, apart from the issue that we also need to show that the median is close to the expectation. This usually is true if the random variable is sufficiently concentrated (which we hope for anyway).

Nevertheless, it is clear that we pay with a significant technical overhead for the hope for a stronger bound. Talagrand's bound first appeared in Talagrand (1995). It is also covered in the textbooks by Alon and Spencer (2008) and Janson *et al.* (2000). Our presentation is greatly inspired by the unpublished lecture notes by Jirka Matoušek and Jan Vondrák, accessible from the first author's homepage.

### 1.6.2. *Kim-Vu Inequality*

Both Azuma's and Talagrand's inequality require that each variable has only a bounded influence on the function value. One might imagine situations where this is not fulfilled, but still a strong concentration behavior is observed. Such settings might in particular occur if some variables may have a large influence on the function value, but only with very small probability. An example met in this text (though solved with different methods) are sums of independent, geometrically distributed random variables.

In such situations, we might hope that a suitable condition on the typical influence of each variable does suffice. Indeed, this is what the recent inequality due to Kim and Vu (2000) does. The result, however, is quite technical. Therefore, we omit further details and point the interested reader to the original paper by Kim and Vu (2000) or the textbook by Alon and Spencer (2008).

## 1.7. Conclusion

In this chapter, we collected a number of tools that found application in the analysis of randomized search heuristics. The collection shows a slight dichotomy. On the one hand, a number of standard arguments from the classical theory of randomized algorithms can be used just as simple as there. On the other hand, the particular nature of randomized search heuristics asks for special tools. This can be because the objective of the analysis is different, e.g., we rarely care for constant factors, but do care a lot about guarantees that hold with high probability. This motivates the use of a different coupon collector theorem. Another reason is the type of random experiments that show up when running a randomized search heuristic. Here we often observe an abundance of dependencies. Hence methods that allow to use classical estimates also in such dependent environments are highly needed. This is different from the classical theory of randomized algorithms. Here, the researcher would design the algorithms in a way that both the algorithm is efficient and that avoids, as far as possible, such difficulties in the analysis.

While the theory community in the randomized search heuristics field made great progress in the last ten year to find mathematical tools suitable for the analysis of such heuristics, it remains a challenge to go further. When analyzing a particular problem, how often do we stumble upon the situation that we ‘know’ what is the answer, but it seems very hard to prove that, simply because the setting is minimally different from a classical setting? This is when not only solving the particular problem is asked for, but in addition to develop tools that are robust against such slight modifications, which cause all the trouble (and much of the fun as well).

## References

- Alon, N. and Spencer, J. H. (2008). *The Probabilistic Method*, 3rd edn. (Wiley-Interscience, New York).
- Azuma, K. (1967). Weighted sums of certain dependent variables, *Tohoku Math. Journal* **3**, pp. 357–367.
- Bennett, G. (1962). Probability inequalities for the sum of independent random variables, *Journal of the American Statistical Association* **57**, pp. 33–45.
- Bernstein, S. (1924). On a modification of Chebyshevs inequality and of the error formula of Laplace, *Ann. Sci. Inst. Sav. Ukraine, Sect. Math.* **1** **4**, pp. 38–49.

- Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, *Ann. Math. Statistics* **23**, pp. 493–507.
- Hagerup, T. and Rüb, C. (1990). A guided tour of Chernoff bounds, *Inf. Process. Lett.* **33**, pp. 305–308.
- Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables, *J. Amer. Statist. Assoc.* **58**, pp. 13–30.
- Janson, S., Luczak, T. and Ruciński, A. (2000). *Random Graphs* (Wiley-Interscience, New York).
- Kim, J. H. and Vu, V. H. (2000). Concentration of multivariate polynomials and its applications, *Combinatorica* **20**, pp. 417–434.
- McDiarmid, C. (1998). Concentration, in *Probabilistic Methods for Algorithmic Discrete Mathematics, Algorithms Combin.*, Vol. 16 (Springer, Berlin), pp. 195–248.
- Motwani, R. and Raghavan, P. (1995). *Randomized Algorithms*. (Cambridge University Press).
- Panconesi, A. and Srinivasan, A. (1997). Randomized distributed edge coloring via an extension of the Chernoff–Hoeffding bounds, *SIAM J. Comput.* **26**, pp. 350–368.
- Prokhorov, Y. V. (1956). Convergence of random processes and limit theorems in probability theory, *Theory of Probability and Its Applications* **1**, pp. 157–214.
- Talagrand, M. (1995). Concentration of measure and isoperimetric inequalities in product spaces, *Inst. Hautes Études Sci. Publications Mathématiques* **81**, pp. 73–205.