Chapter 1

Introduction to Probability Theory

In this chapter, we review the basic concepts in the area of probability theory and present probabilistic techniques that allow the solution of a wide range of coordination problems. A special emphasis will be given on modeling dependence and presenting different tools to deal with dependent events or dependent random variables. Our intention will not be to give a mathematically rigorous introduction with all proofs, but rather to concentrate on the main ideas and techniques that lead to the results presented in this section.

1.1 Basic Notation and Concepts

In the following, small letters usually represent scalars, capital letters represent sets or events, and letters in bold face represent vectors (like $\boldsymbol{a}, \boldsymbol{b}, \ldots$) or matrices (like $\boldsymbol{A}, \boldsymbol{B}, \ldots$). A family of sets is denoted by $\mathcal{A}, \mathcal{B}, \ldots$.

1.1.1 Set Theory

Let M be an arbitrary set. Given any subset $A \subseteq M$, $|A|$ denotes its *cardinality*, that is, the number of elements in A, and $\bar{A}$ denotes its *complement* $M \setminus A$. $\mathcal{P}(M)$ is called the *power set* of M and consists of all subsets $A \subseteq M$. We summarize a few standard sets.

- $\emptyset$ denotes the empty set.
- $\mathbb{N}$ means the set of integers $\{1, 2, 3, 4, \ldots\}$, and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$.
- For any $c \in \mathbb{N}$, the set $[c]$ represents the set $\{1, \ldots, c\}$.
- $\mathbb{R}_+$ denotes the set of all non-negative reals.
- For any two reals $x, y \in \mathbb{R}$ with $x < y$, $[x, y]$ denotes the set of all real numbers z with $x \leq z \leq y$, and (x, y) denotes the set of all real numbers z with $x < z < y$. (Also mixed forms $[x, y)$ are possible.)
- For any $n \in \mathbb{N}$, S_n denotes the set of all permutations on $[n]$.

Given two sets A and B,

- $A \cap B = \{x \mid x \in A \text{ and } x \in B\}$ is called the *intersection* of A and B,
- $A \cup B = \{x \mid x \in A \text{ or } x \in B\}$ is called the *union* of A and B, and
- $A \setminus B = \{x \mid x \in A \text{ and } x \notin B\}$ is called the *difference* of A and B.

Two sets A and B are called *disjoint* if $A \cap B = \emptyset$.

A very important principle in set theory is the *inclusion-exclusion principle*: Consider any collection of n sets $A_1, \dots, A_n$. Then,

$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{k=1}^{n} (-1)^{k+1} \sum_{i_1 < i_2 < \dots < i_k} \left|\bigcap_{j=1}^{k} A_{i_j}\right| .$$

1.1.2 Combinatorics

Basic Notation

We frequently use the following notation.

- If we use "e" in formulas, we always mean the Euler number (2.71828..).
- By "log" we mean the logarithm to the base of 2. "ln" denotes the logarithm to the base of e. If we use any other base, say b, we write "$\log_b$".
- $\log^k n$ means $(\log n)^k$.
- $n! = n \cdot (n-1) \cdot \ldots \cdot 2 \cdot 1$ is called *n factorial*. By definition, $0! = 1$. For any $x \in \mathbb{R}$, the *rth descending factorial* of x is $x^{(r)} = x \cdot (x-1) \cdot \ldots \cdot (x-r+1)$, and the *$r$th ascending factorial* of x is $x^{[r]} = x \cdot (x+1) \cdot \ldots \cdot (x+r-1)$. By definition, $x^{[0]} = x^{(0)} = 1$ for every $x \in \mathbb{R}$.
- For all $0 \le k \le n$,
$$\binom{n}{k} = \frac{n!}{(n-k)! \cdot k!} = \frac{n^{(k)}}{k!} .$$

Identities

We will require sometimes to switch between different representations. For this, the following list of identities will be helpful. For every $x \in \mathbb{R}$,

$$\mathrm{e}^x = \sum_{i=0}^{\infty} \frac{x^i}{i!} \quad \text{and} \quad \ln(1+x) = \sum_{i=1}^{\infty} \frac{(-1)^{i-1}}{i} \cdot x^i .$$

For every $n \ge 0$ and $x \in \mathbb{R}$ it holds that

$$x^{(n)} = \sum_{k=0}^{n} s_{n,k} \cdot x^k$$

where the $s_{n,k}$'s are called the *Stirling numbers of the first kind*. They are defined as $s_{n,k} = (-1)^{n-k}c_{n,k}$ for all $k, n \in \mathbb{N}_0$, where $c_{n,k}$ is called a *signless* Stirling number of the first kind. The $c_{n,k}$'s are recursively defined by

$$
\begin{aligned}
c_{0,0} &= 1 \\
c_{n,0} &= 0 \quad \text{for all } n \in \mathbb{N} \\
c_{n,k} &= (n-1) \cdot c_{n-1,k} + c_{n-1,k-1} \quad \text{for all } k, n \in \mathbb{N}.
\end{aligned}
$$

$c_{n,k}$ represents the number of permutations $\pi \in S_n$ with exactly k cycles. Some values of the $c_{n,k}$'s are given in Figure 1.1.

n/k	0	1	2	3	4	5	6	7	8
0	1								
1		1							
2		1	1						
3		2	3	1					
4		6	11	6	1				
5		24	50	35	10	1			
6		120	274	225	85	15	1		
7		720	1764	1624	735	175	21	1	
8		5040	1368	13132	6769	1960	322	28	1

Figure 1.1: Table of the signless Stirling numbers of the first kind

Furthermore, for every $n \geq 0$ and $x \in \mathbb{R}$ it holds that

$$x^n = \sum_{k=0}^{n} S_{n,k} \cdot x^{(k)}$$

where the $S_{n,k}$'s are called *Stirling numbers of the second kind*. They are recursively defined by

$$
\begin{aligned}
S_{n,k} &= 0 \quad \text{for } k = 0 \text{ or } k > n \\
S_{n,1} &= S_{n,n} = 1 \\
S_{n,k} &= k \cdot S_{n-1,k} + S_{n-1,k-1} \quad \text{for all } k, n \in \mathbb{N}
\end{aligned}
$$

$S_{n,k}$ represents the number of ways of partitioning $[n]$ into k non-empty sets. Some values of the $S_{n,k}$'s are given in Figure 1.2.

Furthermore, for every $n \geq 0$ and $x \in \mathbb{R}$ we have

$$x^{[n]} = \sum_{k=0}^{n} c_{n,k} \cdot x^k$$

and

$$x^n = \sum_{k=0}^{n} (-1)^{n-k} S_{n,k} \cdot x^{[k]} \ .$$

n/k	0	1	2	3	4	5	6	7	8
0	1								
1		1							
2		1	1						
3		1	3	1					
4		1	7	6	1				
5		1	15	25	10	1			
6		1	31	90	65	15	1		
7		1	63	301	350	140	21	1	
8		1	127	966	1701	1050	266	28	1

Figure 1.2: Table of the Stirling numbers of the second kind

Useful inequalities

Using the identity for e^x above, one can show that

$$1 + x \leq \mathrm{e}^x \quad \text{for all } x \in \mathbb{R}$$

and

$$1 - x \leq \mathrm{e}^{-x} \leq 1 - x + x^2/2 \quad \text{for all } x \in [0, 1].$$

Furthermore,

$$\left(1 + \frac{1}{n}\right)^n \leq \mathrm{e} \leq \left(1 + \frac{1}{n}\right)^{n+1}$$

for all $n \in \mathbb{N}$, implying that $\lim_{n\to\infty}(1 + 1/n)^n = \mathrm{e}$. Also,

$$\left(1 - \frac{1}{n}\right)^n \leq \frac{1}{\mathrm{e}} \leq \left(1 - \frac{1}{n}\right)^{n-1}$$

for all $n \in \mathbb{N}$, implying that $\lim_{n\to\infty}(1 - 1/n)^n = 1/\mathrm{e}$.

Frequently, one wishes to express $n!$ or $\binom{n}{k}$ by terms based on standard operations. For all $n \in \mathbb{N}$, Stirling's formula says that

$$n! = \alpha(n)\sqrt{2\pi n}\left(\frac{n}{\mathrm{e}}\right)^n \quad \text{where } \alpha(n) \in [1, \mathrm{e}^{1/12n}] \ .$$

Often, the simpler inequality $(n/\mathrm{e})^n \leq n! \leq n^n$ is used.

Stirling's formula implies that for all $k, n \in \mathbb{N}$ with $k < n$,

$$\binom{n}{k} = \beta(n,k)\left(\frac{n}{k}\right)^k\left(\frac{n}{n-k}\right)^{n-k}\sqrt{\frac{n}{2\pi k(n-k)}} \quad \text{where } \beta(n,k) \in [\mathrm{e}^{-1/6k}, 1] \ .$$

Often, the simpler inequality $(n/k)^k \leq \binom{n}{k} \leq (\mathrm{e}n/k)^k$ is used.

For any $0 < \alpha \leq 1/2$,

$$\sum_{0\leq k\leq \alpha n}\binom{n}{k} \leq 2^{H(\alpha)\cdot n} \ ,$$

where $H(\alpha) = -\alpha \log \alpha - (1-\alpha)\log(1-\alpha)$ is the binary entropy function.

We list some other important inequalities taken from [Hof87]. For any $n \in \mathbb{N}$, let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be any two sequences of non-negative real numbers.

Fact 1.1 (Chebychev) *If $a_1 \geq a_2 \geq \ldots \geq a_n$ and $b_1 \geq b_2 \geq \ldots \geq b_n$ then*

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

Fact 1.2 (Hölder) *If $p > 1$, $q > 1$ and $p^{-1} + q^{-1} = 1$, then*

$$\sum_{i=1}^{n} a_i b_i \leq \left(\sum_{i=1}^{n} a_i^p\right)^{1/p} \left(\sum_{i=1}^{n} b_i^q\right)^{1/q} .$$

For the special case of $p = q = 2$, this inequality is called the Cauchy-Schwartz inequality.

Fact 1.3 (Minkowski) *If $p > 1$ then*

$$\left(\sum_{i=1}^{n} (a_i + b_i)^p\right)^{1/p} \leq \left(\sum_{i=1}^{n} a_i^p\right)^{1/p} + \left(\sum_{i=1}^{n} b_i^p\right)^{1/p} .$$

Fact 1.4 (The Three Means) *Given a sequence of real numbers $x_1, \ldots, x_n$,*

- *its* arithmetic mean *is defined as $A = \sum_i x_i/n$,*
- *its* geometric mean *is defined as $G = (\prod_i x_i)^{1/n}$, and*
- *its* harmonic mean *is defined as $H = (\sum_i x_i^{-1}/n)^{-1}$.*

If $x_1, \ldots, x_n$ are non-negative, then

$$A \geq G \geq H .$$

Basic Counting

Many combinatorial problems can be viewed as drawing balls. Consider the problem of determining the number of ways of drawing k balls one after the other out of n balls, numbered from 1 to n. The most important cases are summarized in Figure 1.3.

	ordered	unordered
with replacement	n^k	$\binom{n+k-1}{k}$
without replacement	$n^{(k)}$	$\binom{n}{k}$

Figure 1.3: Number of outcomes of drawing k out of n balls

1.2 Basic Concepts in Probability Theory

In this section we will introduce some basic concepts in probability theory.

Let us consider a random experiment of which all possible results are included in a non-empty set Ω, usually called the *sample space*. An element $\omega \in \Omega$ is called a *sample point* or *outcome* of the experiment. An *event* of a random experiment is specified as a subset of Ω. Event A is called *true* if an outcome $\omega \in \Omega$ has been chosen with $\omega \in A$. Otherwise A is called *false*. A system $\mathcal{A}$ of events (or, in general, subsets of Ω) is called an *algebra* if

- $\Omega \in \mathcal{A}$,
- if $A, B \in \mathcal{A}$, then $A \cup B \in \mathcal{A}$ and $A \cap B \in \mathcal{A}$, and
- if $A \in \mathcal{A}$, then $\bar{A} \in \mathcal{A}$.

Given an algebra $\mathcal{A}$, a function $\mu : \mathcal{A} \to \mathsf{IR}_+$ is called a *measure* on $\mathcal{A}$ if for every pair of disjoint sets $A, B \in \mathcal{A}$ we have

$$\mu(A \cup B) = \mu(A) + \mu(B) \ .$$

This definition clearly implies that $\mu(\emptyset) = 0$ and that for any set of pairwise disjoint events $A_1, \ldots, A_k \in \mathcal{A}$ we have

$$\mu\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{n} \mu(A_i) \ .$$

Furthermore, it implies that for any pair of sets $A, B \in \mathcal{A}$ we have

$$\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B) \ .$$

We say that a function $p : \mathcal{A} \to [0, 1]$ is a *probability measure* if

- p is a measure on $\mathcal{A}$ and
- $p(\Omega) = 1$.

Given a probability measure p, the *probability* of an event A to be true is defined as

$$\Pr[A] = p(A) \ .$$

We say that a triple $(\Omega, \mathcal{A}, p)$ is a *probability space* if $\mathcal{A}$ is an algebra over Ω and p is a probability measure on $\mathcal{A}$.

1.2.1 Events

Starting from a given collection of sets that represent events, we can form new events by means of statements containing the logical connectives "or," "and," and "not," which correspond in the language of set theory to the operations "union," "intersection," and "complement."

If A and B are events, their *union*, denoted by $A \cup B$, is the event consisting of all outcomes realizing either A or B. The *intersection* of A and B, denoted by $A \cap B$, consists of all outcomes realizing both A and B. The *difference* of B and A, denoted by $B \setminus A$, consists of all outcomes that belong to B but not to A. If A is a subset of Ω, its *complement*, denoted by $\bar{A}$, is the set of outcomes in Ω that do not belong to A. That is, $\bar{A} = \Omega \setminus A$.

Two events A and B are called *disjoint* if $A \cap B$ is empty. In probability theory, $\emptyset$ is called the *impossible* event. The set Ω is naturally called the *certain* event.

If Ω is a countable sample space (i.e., its elements can be arranged in a sequence so that the rth element is identifiable for any $r \in \mathbb{N}$), we define the *size* of an event A, denoted by $|A|$, to be the number of outcomes it contains.

1.2.2 The Inclusion-Exclusion Principle

Let $A_1, \ldots, A_n$ be any collection of events. The *inclusion-exclusion principle* stated in Section 1.1.1 implies that

$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] = \sum_{k=1}^{n} (-1)^{k+1} \sum_{i_1<i_2<\ldots<i_k} \Pr\left[\bigcap_{j=1}^{k} A_{i_j}\right] . \tag{1.1}$$

For the special case of $n = 2$ we obtain

$$\Pr[A_1 \cup A_2] = \Pr[A_1] + \Pr[A_2] - \Pr[A_1 \cap A_2] .$$

In cases where it is too difficult to evaluate (1.1) exactly, *Bonferroni's inequalities* may be used to find suitable approximations:

- For every odd m,

$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] \leq \sum_{k=1}^{m} (-1)^{k+1} \sum_{i_1<i_2<\ldots<i_k} \Pr\left[\bigcap_{j=1}^{k} A_{i_j}\right] .$$

- For every even m,

$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] \geq \sum_{k=1}^{m} (-1)^{k+1} \sum_{i_1<i_2<\ldots<i_k} \Pr\left[\bigcap_{j=1}^{k} A_{i_j}\right] .$$

Special cases of these inequalities are *Boole's inequalities*:

$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] \leq \sum_{i=1}^{n} \Pr[A_i]$$

and

$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] \geq \sum_{i=1}^{n} \Pr[A_i] - \sum_{1\leq i<j\leq n} \Pr[A_i \cap A_j] .$$

We demonstrate the usefulness of the inclusion-exclusion principle by some simple examples.

Example 1: Consider the problem of coloring the nodes of a hypergraph G with the help of 2 colors such that no hyperedge is monochromatic (i.e., no hyperedge only contains nodes of a single color). We would like to identify a class of hypergraphs for which this is always possible. Such a class is given in the following claim.

Claim 1.5 *For every hypergraph G with n hyperedges in which every hyperedge is of size at least $\log n + 2$, there is a 2-coloring of the nodes such that no hyperedge is monochromatic.*

Proof. Consider the random experiment of choosing for each node independently and uniformly at random one of the two possible colors. In this case, the probability that a hyperedge of size k is monochromatic is equal to $2 \cdot 2^{-k} = 2^{-k+1}$. Assume the hyperedges to be numbered from 1 to n and let A_i be the event that hyperedge i is monochromatic. Then, by the inclusion-exclusion principle,

$$\Pr[A_1 \cup \ldots \cup A_n] \leq \sum_{i=1}^{n} \Pr[A_i] \leq \sum_{i=1}^{n} 2^{-(\log n+1)} = \frac{1}{2} \ .$$

Hence,

$$\Pr[\bar{A}_1 \cap \ldots \cap \bar{A}_n] = 1 - \Pr[A_1 \cup \ldots \cup A_n] \geq \frac{1}{2} \ ,$$

and therefore there must exist a 2-coloring such that no edge is monochromatic. □

Example 2: Assume that we have n balls and n bins. Each ball is placed in a bin that is chosen independently and uniformly at random. The goal is to provide an upper bound for the maximum number of balls in a bin. This will be given in the following claim.

Claim 1.6 *With a probability of at least $1/2$, the maximum number of balls in a bin is at most $(1 + o(1)) \log n / \log \log n$.*

Proof. Let $k = \alpha \log n / \log \log n$ for some non-negative α that will be specified later. For every $i \in [n]$, let the event A_i be true if and only if bin i has at least k balls. Furthermore, let E_1 denote the event that there exists a bin with at least k balls. Then, by Boole's inequality,

$$\Pr[E_1] = \Pr[A_1 \cup \ldots \cup A_n] \leq \sum_{i=1}^{n} \Pr[A_i] \ .$$

For any subset $S \subseteq [n]$, let the event $A_{i,S}$ be true if and only if all balls in S are placed in bin i. Then, again by Boole's inequality,

$$\Pr[A_i] = \Pr\left[\bigcup_{S \subseteq [n],\ |S|=k} A_{i,S}\right] \leq \sum_{S \subseteq [n],\ |S|=k} \Pr[A_{i,S}]$$

for all i. Since $|\{S \subseteq [n],\ |S| = k\}| = \binom{n}{k}$ and $\Pr[A_{i,S}] = 1/n^{|S|}$ for all $S \subseteq [n]$, we obtain

$$\begin{aligned}
\Pr[E_1] &\leq n \cdot \binom{n}{k} \left(\frac{1}{n}\right)^k \\
&\leq n \cdot \left(\frac{\mathrm{e} \cdot n}{k}\right)^k \left(\frac{1}{n}\right)^k = n \cdot \left(\frac{\mathrm{e}}{k}\right)^k \\
&\leq n \cdot 2^{-\alpha \frac{\log n}{\log \log n} \cdot \log\left(\frac{\alpha \log n}{\mathrm{e} \log \log n}\right)} \\
&\leq n \cdot 2^{-\log n - 1} \leq \frac{1}{2}
\end{aligned}$$

if $\alpha = 1 + o(1)$ and n are sufficiently large. □

We note that the bound for the maximum number of balls in Claim 1.6 is essentially best possible, since the probability that the maximum number of balls is at least $(1 - o(1)) \log n / \log \log n$ can be shown to be also at least $1/2$.

Example 3: Again, consider the situation that we have n balls and n bins, and each ball is placed in a bin chosen independently and uniformly at random.

Claim 1.7 *The probability that bin 1 has at least one ball is at least* $1/2$.

Proof. For any $i \in [n]$, let A_i be the event that ball i is placed in bin 1. Then, by Boole's inequality,

$$\begin{aligned}
\Pr[\text{bin 1 has at least one ball}] &= \Pr\left[\bigcup_{i \in [n]} A_i\right] \\
&\geq \sum_{1 \leq i \leq n} \Pr[A_i] - \sum_{1 \leq i < j \leq n} \Pr[A_i \cap A_j] \\
&= \sum_{1 \leq i \leq n} \frac{1}{n} - \sum_{1 \leq i < j \leq n} \frac{1}{n^2} \\
&= 1 - \binom{n}{2} \frac{1}{n^2} \geq 1 - \frac{1}{2} = \frac{1}{2} .
\end{aligned}$$

□

Observe that the probability bound in Claim 1.7 is not far away from the exact bound:

$$\Pr[\text{bin 1 has at least one ball}] = 1 - \left(1 - \frac{1}{n}\right)^n \overset{n \to \infty}{=} 1 - \frac{1}{\mathrm{e}} .$$

1.2.3 Conditional Probability

The *conditional probability* of event B assuming an event A with $\Pr[A] > 0$ is denoted by $\Pr[B \mid A]$ and defined as

$$\Pr[B \mid A] = \frac{\Pr[A \cap B]}{\Pr[A]} .$$

From this definition it follows that for all events A and B with $\Pr[A] > 0$,

$$\Pr[A \cap B] = \Pr[A] \cdot \Pr[B \mid A] . \tag{1.2}$$

This can be generalized as follows: If $A_1, \ldots, A_n$ are events with $\Pr[A_1 \cap \ldots \cap A_{n-1}] > 0$, then

$$\Pr[A_1 \cap \ldots \cap A_n] = \prod_{i=1}^{n} \Pr[A_i \mid A_1 \cap \ldots \cap A_{i-1}] .$$

Suppose that A and B are events with $\Pr[A] > 0$ and $\Pr[B] > 0$. Then, in addition to the equality (1.2), we have

$$\Pr[A \cap B] = \Pr[B] \cdot \Pr[A \mid B] . \tag{1.3}$$

From (1.2) and (1.3) we obtain *Bayes's formula*

$$\Pr[A \mid B] = \frac{\Pr[A] \cdot \Pr[B \mid A]}{\Pr[B]} .$$

Two events A and B are called *independent* if and only if

$$\Pr[B \mid A] = \Pr[B] .$$

Note that, due to Bayes's formula, in this case also $\Pr[A \mid B] = \Pr[A]$, that is, the independence property is *symmetric*.

If $\Pr[B \mid A] \neq \Pr[B]$, then A and B are said to be *correlated*. A and B are called

- *negatively correlated* if $\Pr[B \mid A] < \Pr[B]$ and
- *positively correlated* if $\Pr[B \mid A] > \Pr[B]$.

By Bayes's formula, all of these correlation properties are also symmetric.

As an example, any two disjoint events A and B with positive probabilities cannot be independent, since $\Pr[B \mid A] = 0$. However, they are always negatively correlated. Furthermore, they have the property that

$$\Pr[A \cup B] = \Pr[A] + \Pr[B] .$$

To illustrate negative and positive correlation, let us give a very simple example.

Example: Consider the problem of throwing n balls into n bins, where each ball is placed in a bin chosen independently and uniformly at random. Let A be the event that bin 1 has no ball, and let B be the event that some fixed bin $i > 1$ has no ball. For $n = 2$ it holds

$$\Pr[B \mid A] = 0 \leq \frac{1}{4} = \Pr[B] ,$$

and for $n = 3$ it holds

$$\Pr[B \mid A] = \frac{1}{8} \leq \frac{8}{27} = \Pr[B] .$$

Thus, A and B are negatively correlated for these n. (One can observe that this also holds for higher n.)

Now, let A be the event that bin 1 has exactly one ball, and let B be the event that some fixed bin $i > 1$ has exactly one ball. Since the expected number of balls in a bin is 1, one may assume that A does not influence B. However, for $n = 2$ it holds

$$\Pr[B \mid A] = 1 \geq \frac{1}{2} = \Pr[B] ,$$

and for $n = 3$ it holds

$$\Pr[B \mid A] = \frac{1}{2} \geq \frac{4}{9} = \Pr[B] .$$

Thus, A and B are positively correlated for these n. However, one can observe that $\Pr[B \mid A]$ tends to $\Pr[B]$ for $n \to \infty$.

If we define A and B to be events for which a bin has exactly $k > 1$ balls, one can show that A and B are, as in the first case, negatively correlated.

1.2.4 Random Variables

Any numerical function $X = X(\omega)$ defined on a sample space Ω may be called a *random variable*. In this thesis we will only consider real-valued random variables, i.e., functions of the form $X : \Omega \to \mathbb{R}$. It has been found convenient to separate the random variables into three categories: discrete, continuous, and mixed.

A *discrete* random variable is supposed to take only isolated values with nonzero probabilities. The number of values it is allowed to take may be infinite, but it is essential that they are *countable*. That is, it must be possible to arrange its values in a sequence so that the rth number is identifiable for any $r \in \mathbb{N}$. An *(absolutely) continuous* random variable X is one for which $\Pr[X \leq x]$ can be expressed as the integral

$$F_X(x) = \int_{-\infty}^{x} p_X(x)dx$$

of a function $p_X(x)$ commonly called the *probability density* of X. F_X is called the *distribution function* of X. A random variable is called *mixed* if it has both discrete and continuous parts.

In the following, we will only consider discrete random variables. A random variable X is called *non-negative* if $X(\omega) \geq 0$ for all $\omega \in \Omega$. For the special case that X maps elements in Ω to $\{0, 1\}$, X is called a *binary* or *Bernoulli* random variable. A binary random variable X is called an *indicator* of event A (denoted by I_A) if $X(\omega) = 1$ if and only if $\omega \in A$ for all $\omega \in \Omega$.

For any random variable X and any number $x \in \mathbb{R}$, we define $[X = x] = \{\omega \in \Omega : \; X(\omega) = x\}$. Instead of using set operations to express combinations of events associated with random variables, we will use logical expressions in the following, that is,

- instead of $\Pr[[X = x] \cap [Y = y]]$ we write $\Pr[X = x \wedge Y = y]$, and
- instead of $\Pr[[X = x] \cup [Y = y]]$ we write $\Pr[X = x \vee Y = y]$.

For any discrete random variable X and any $k \in \mathbb{R}$, we define

$$\Pr[X \leq k] = \sum_{\ell \leq k} \Pr[X = \ell] \quad \text{and} \quad \Pr[X \geq k] = \sum_{\ell \geq k} \Pr[X = \ell] \; .$$

The function $p_X(k) = \Pr[X = k]$ is called the *probability distribution* of X, and the function $F_X(k) = \Pr[X \leq k]$ is called the *(cumulative) distribution function* of X. Furthermore, the function $G_X(k) = \Pr[X \geq k]$ is called the *survival distribution function* of X. For any two random variables X and Y, X is said to *(stochastically) dominate* Y (denoted by $X \succ Y$) if $G_X(k) \geq G_Y(k)$ for all $k \in \mathbb{R}$.

1.2.5 Expectation

The most important measure used in combination with random variables is the expectation.

Definition 1.8 *Let $(\Omega, \mathcal{A}, p)$ denote an arbitrary probability space and $X : \Omega \to \mathbb{R}$ be an arbitrary discrete function. Then the* expectation *of X is defined as*

$$\mathrm{E}[X] = \sum_{x \in \mathbb{R}} x \cdot \Pr[X = x] \; . \tag{1.4}$$

Assume that the set $S = \{x_1, x_2, \ldots\}$ contains all values X can take, where $x_1 < x_2 < \ldots$. Let the function $\Delta_X : \mathbb{R} \to \mathbb{R}_+$ be defined as

$$\Delta_X(x) = \begin{cases} 0 & : \quad x \notin S \\ x & : \quad x = x_1 \\ x - x_{i-1} & : \quad x = x_i \text{ for some } i \geq 2 \end{cases}$$

That is, $\Delta_X(x)$ provides the distance to the predecessor of x in S. If X is non-negative, it follows from equation (1.4) and the definition of Δ_X that

$$\mathrm{E}[X] = \sum_{x \geq 0} \Delta_X(x) \cdot \Pr[X \geq x] = \sum_{x \geq 0} \Delta_X(x) \cdot G_X(x) \; . \tag{1.5}$$

In the case that X is integer-valued, we can simplify this expression to

$$\mathrm{E}[X] = \sum_{x \in \mathbb{N}} \Pr[X \geq x] = \sum_{x \in \mathbb{N}} G_X(x) \; . \tag{1.6}$$

For binary random variables X it holds that

$$\mathrm{E}[X] = \Pr[X = 1] \; .$$

Basic properties

The following fact lists some basic properties of the expectation.

Fact 1.9 *For arbitrary random variables X and Y and any $c \in \mathbb{R}$ it holds:*

- *If X is non-negative, then* $\mathrm{E}[X] \geq 0$.
- $|\mathrm{E}[X]| \leq \mathrm{E}[|X|]$.
- $\mathrm{E}[c \cdot X] = c \cdot \mathrm{E}[X]$.
- $\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$.

Two random variables X and Y are called *independent* if, for all $x, y \in \mathbb{R}$,

$$\Pr[X = x \mid Y = y] = \Pr[X = x] \; .$$

Independent random variables have the following important property.

Claim 1.10 *If X and Y are independent random variables, then* $\mathrm{E}[X \cdot Y] = \mathrm{E}[X] \cdot \mathrm{E}[Y]$.

Proof. For all $x, y \in \mathbb{R}$, let A_x and B_y be the events with

$$A_x = \{\omega \in \Omega : \; X(\omega) = x\} \quad \text{and} \quad B_y = \{\omega \in \Omega : \; Y(\omega) = y\} \; .$$

Since X and Y are independent, $\Pr[A_x \cap B_y] = \Pr[A_x] \cdot \Pr[B_y]$ for all $x, y \in \mathbb{R}$. Recall that, for any event E, the indicator variable I_E is equal to 1 if and only if $\omega \in E$. As $X = \sum_x x \cdot I_{A_x}$ and $Y = \sum_y y \cdot I_{B_y}$, we obtain

$$\begin{aligned}
\mathrm{E}[X \cdot Y] &= \mathrm{E}\left[\left(\sum_{x \in \mathbb{R}} x \cdot I_{A_x}\right) \cdot \left(\sum_{y \in \mathbb{R}} y \cdot I_{B_y}\right)\right] \\
&= \mathrm{E}\left[\sum_{x,y \in \mathbb{R}} x \cdot y \cdot I_{A_x \cap B_y}\right] = \sum_{x,y \in \mathbb{R}} x \cdot y \cdot \Pr[A_x \cap B_y] \\
&= \sum_{x,y \in \mathbb{R}} x \cdot y \cdot \Pr[A_x] \cdot \Pr[B_y] \\
&= \left(\sum_{x \in \mathbb{R}} x \cdot \Pr[A_x]\right) \cdot \left(\sum_{y \in \mathbb{R}} y \cdot \Pr[B_y]\right) = \mathrm{E}[X] \cdot \mathrm{E}[Y] \ .
\end{aligned}$$

□

A set $X_1, \ldots, X_n$ of random variables is called *(mutually) independent* if, for all $i \in [n]$ and $S \subseteq [n] \setminus \{i\}$,

$$\Pr\left[X_i = x_i \mid \bigwedge_{j \in S} X_j = x_j\right] = \Pr[X_i = x_i] \tag{1.7}$$

for all $x_i \in \mathbb{R}$ and $x_j \in \mathbb{R}$ with $j \in S$. $X_1, \ldots, X_n$ are called *k-wise independent* if (1.7) holds for all $S \subseteq [n] \setminus \{i\}$ with $|S| \leq k$.

Modeling Dependence

The *conditional expectation* of a random variable Y with respect to an event A is defined by

$$\mathrm{E}[Y \mid A] = \sum_{y \in \mathbb{R}} y \cdot \Pr[Y = y \mid A] \ . \tag{1.8}$$

If the event A is $X = a$ for some random variable X, this equation defines a function f with

$$f(a) = \mathrm{E}[Y \mid X = a] \ .$$

Thus, $\mathrm{E}[Y \mid X]$ is a random variable, namely the random variable $f(X)$. (Observe that it does *not* hold in general that $\mathrm{E}[X \cdot Y] = \mathrm{E}[X] \cdot \mathrm{E}[Y \mid X]$. A way to study the relationship between $\mathrm{E}[X \cdot Y]$ and $\mathrm{E}[X]$ and $\mathrm{E}[Y]$ will be given in Section 1.2.6.)

A fundamental property of the conditional expectation is that

$$\mathrm{E}[Y] = \mathrm{E}[\mathrm{E}[Y \mid X]]$$

for any two random variables X and Y (which is easy to check). If Y is independent of X, then (1.8) implies that $\mathrm{E}[Y \mid X] = \mathrm{E}[Y]$. An important case for which $\mathrm{E}[Y]$ depends on X is given in the following definition.

Definition 1.11 *A sequence of random variables $X_0, X_1, \ldots$ is called a* martingale *if for all $i \geq 1$,*

$$\mathrm{E}[X_i \mid X_0, X_1, \ldots, X_{i-1}] = X_{i-1} \ .$$

Martingales will be considered in a later section.

Tail estimates

We start with a fact that contains some straightforward probability bounds.

Fact 1.12 *Let X be an arbitrary random variable. Then*

$$\Pr[X < \mathrm{E}[X]] < 1 \quad \textit{and} \quad \Pr[X > \mathrm{E}[X]] < 1 .$$

The next result provides a first, simple probability bound that depends on the deviation from the expected value. It has apparently first been used by Chebychev (which is why some authors call it Chebychev inequality [Shi96]), but it is commonly called *Markov inequality*.

Theorem 1.13 (Markov Inequality) *Let X be an arbitrary non-negative random variable. Then, for any $k > 0$,*

$$\Pr[X \geq k] \leq \frac{\mathrm{E}[X]}{k} .$$

Proof. Obviously,

$$\mathrm{E}[X] = \sum_{x \geq 0} x \cdot \Pr[X = x] \geq k \cdot \Pr[X \geq k] .$$

□

As we will show later, this inequality has a tremendous number of consequences. We will start with some simple examples.

Example 1: A directed graph G is called a *tournament* if for every pair of nodes u, v with $u \neq v$, either (u, v) or (v, u) is an edge in G. A *Hamiltonian path* in a directed graph G is a directed path that visits every node of G exactly once. We will show the following claim.

Claim 1.14 *There exists a tournament of size n with at least $n!/2^{n-1}$ Hamiltonian paths.*

Proof. Let K_n be the complete undirected graph of size n. We generate a random tournament T out of K_n by choosing a direction for each edge independently and uniformly at random. Let π be a permutation of the nodes in T. We define the binary random variable X_π to be 1 if and only if π defines a directed path in T. Clearly, for all $\pi \in S_n$, $\Pr[X_\pi = 1] = 2^{-(n-1)}$. Let X be the random variable counting the number of Hamiltonian paths in T, that is,

$$X = \sum_{\pi \in S_n} X_\pi .$$

Since $|S_n| = n!$, we obtain

$$\mathrm{E}[X] = \sum_{\pi \in S_n} \Pr[X_\pi = 1] = n!/2^{n-1} .$$

According to Fact 1.12, $\Pr[X < \mathrm{E}[X]] < 1$. Thus, there must exist a tournament with at least $n!/2^{n-1}$ Hamiltonian paths. □

Example 2 [Sch98]: Consider an arbitrary network G of size n, and let S_n denote the set of all permutations on $[n]$. For a permutation $\pi \in S_n$, let $R(G,\pi)$ be the minimum possible number of steps required to route packets in G according to π. Then the *routing number* R of G is defined as

$$R = \max_{\pi \in S_n} R(G,\pi) .$$

Let $\bar{R} = (1/|S_n|) \sum_{\pi \in S_n} R(G,\pi)$ denote the average number of steps to route a permutation in G using a best possible routing strategy. We intend to show the following claim.

Claim 1.15 *For any graph G, $\bar{R} = \Theta(R)$.*

Proof. It is obvious that $\bar{R} \leq R$ and therefore $\bar{R} = O(R)$. Hence, it remains to be shown that $\bar{R} = \Omega(R)$.

Consider any fixed permutation π. In order to bound the minimum number of steps to route π, we will use a probabilistic argument based on Valiant's trick (see [Val82]) by first sending the packets to random intermediate destinations before sending them to the destinations prescribed by π. Let X be a random variable denoting the minimum number of steps necessary to route a randomly chosen permutation. In this case, $\mathrm{E}[X] = \bar{R}$. According to the Markov inequality it holds:

$$\Pr[X \geq 3\bar{R}] \leq \tfrac{1}{3} .$$

Therefore, for a randomly chosen permutation φ for the intermediate destinations, it holds that the minimum number of steps to route the packets first to their intermediate destinations (prescribed by φ) and then to their final destinations (prescribed by π) is at least $6\bar{R}$ with probability at most $\frac{1}{3} + \frac{1}{3} < 1$. Hence, there exists an offline protocol that routes π in at most $6\bar{R}$ steps. Thus, $R \leq 6\bar{R}$, which completes the proof. □

1.2.6 Correlation

Two random variables X and Y are said to be *correlated* if and only if

$$\mathrm{E}[X \cdot Y] \neq \mathrm{E}[X] \cdot \mathrm{E}[Y] .$$

Otherwise, X and Y are called *uncorrelated*. Note that random variables may be uncorrelated, but not independent. Here is a simple example. Let the random variable X take the values 0, $\pi/2$ and π with probability $1/3$. Then $Y = \sin X$ and $Z = \cos X$ are uncorrelated. However, they are not independent, since

$$\Pr[Y = 1 \ \wedge \ Z = 1] = 0 \neq \frac{1}{9} = \Pr[Y = 1] \cdot \Pr[Z = 1] .$$

For the special case that X and Y are binary random variables, X and Y are uncorrelated if and only if X and Y are independent. Two random variables X and Y are called

- *negatively correlated* if $\mathrm{E}[X \cdot Y] \leq \mathrm{E}[X] \cdot \mathrm{E}[Y]$ and
- *positively correlated* if $\mathrm{E}[X \cdot Y] \geq \mathrm{E}[X] \cdot \mathrm{E}[Y]$.

In general, a collection of random variables $X_1, \ldots, X_n$ is called *(mutually) uncorrelated* if

$$\mathrm{E}\left[\prod_{i \in S} X_i\right] = \prod_{i \in S} \mathrm{E}[X_i] \tag{1.9}$$

for every $S \subseteq [n]$. $X_1, \ldots, X_n$ are called *(mutually) negatively correlated* or *(mutually) positively correlated* if instead of "=" we have "$\leq$" or "$\geq$", respectively. Furthermore, $X_1, \ldots, X_n$ are called k*-wise uncorrelated* if (1.9) holds for every $S \subseteq [n]$ with $|S| \leq k$. k-wise negative and positive correlation is defined accordingly.

Panconesi and Srinivasan defined in [PS92] the notion of self-weakening random variables, which can be seen as a generalization to the notion of negatively correlated random variables. A collection of random variables $X_1, \ldots, X_n$ is called *self-weakening with parameter* λ if

$$\mathrm{E}\left[\prod_{i \in S} X_i\right] \leq \lambda \prod_{i \in S} \mathrm{E}[X_i] \tag{1.10}$$

for every $S \subseteq [n]$.

In the following, we present two methods of proving correlation properties. The first uses stochastic domination (recall the definition in Section 1.2.4).

Claim 1.16 *Consider any two non-negative random variables X and Y.*

(1) If $Y \succ (Y \mid X \geq x)$ for all $x \in \mathrm{IR}$, then X and Y are negatively correlated.

(2) If $Y \prec (Y \mid X \geq x)$ for all $x \in \mathrm{IR}$, then X and Y are positively correlated.

Proof. We only prove (1). The proof for (2) is analogous. For all $x, y \in \mathrm{IR}$, let A_x and B_y be the events with

$$A_x = \{\omega \in \Omega : \; X(\omega) \geq x\} \quad \text{and} \quad B_y = \{\omega \in \Omega : \; Y(\omega) \geq y\} \; .$$

Since $Y \succ (Y \mid X \geq x)$ for all $x \in \mathrm{IR}$, $\Pr[Y \geq y] \geq \Pr[Y \geq y \mid X \geq x]$ for all $x, y \in \mathrm{IR}$. Hence,

$$\Pr[A_x \cap B_y] = \Pr[A_x] \cdot \Pr[B_y \mid A_x] \leq \Pr[A_x] \cdot \Pr[B_y]$$

for all $x, y \in \mathrm{IR}$. It is easy to verify that X is equal to $\sum_{x \geq 0} \Delta_X(x) I_{A_x}$ (see equation (1.5)). Thus,

$$\begin{aligned}
\mathrm{E}[X \cdot Y] &= \mathrm{E}\left[\left(\sum_{x \geq 0} \Delta_X(x) \cdot I_{A_x}\right) \cdot \left(\sum_{y \geq 0} \Delta_Y(y) \cdot I_{B_y}\right)\right] \\
&= \mathrm{E}\left[\sum_{x,y \geq 0} \Delta_X(x) \cdot \Delta_Y(y) \cdot I_{A_x \cap B_y}\right] = \sum_{x,y \geq 0} \Delta_X(x) \cdot \Delta_Y(y) \cdot \Pr[A_x \cap B_y] \\
&\leq \sum_{x,y \geq 0} \Delta_X(x) \cdot \Delta_Y(y) \cdot \Pr[A_x] \cdot \Pr[B_y] \\
&= \left(\sum_{x \geq 0} \Delta_X(x) \cdot \Pr[A_x]\right) \cdot \left(\sum_{y \geq 0} \Delta_Y(y) \cdot \Pr[B_y]\right) = \mathrm{E}[X] \cdot \mathrm{E}[Y] \; ,
\end{aligned}$$

which implies that X and Y are negatively correlated. $\square$

Another important tool for the study of the relationship between $\mathrm{E}[X \cdot Y]$ and $\mathrm{E}[X] \cdot \mathrm{E}[Y]$ is the so-called FKG inequality. Before we present it, we introduce some notation.

A *lattice* is a partially ordered set in which every two elements x and y have a unique minimal upper bound, denoted by $x \vee y$ and called the *join* of x and y, and a unique maximal lower bound, denoted by $x \wedge y$ and called the *meet* of x and y. A lattice is *distributive* if for all $x, y, z \in L$,

$$x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z)$$

or, equivalently, if for all $x, y, z \in L$,

$$x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z) \ .$$

To give an example, any subset L of Ω which is closed under the union and intersection operations is a distributive lattice. Here, the join of two members $A, B \in L$ is simply their union $A \cup B$ and their meet is the intersection $A \cap B$.

A function $\mu : L \to \mathrm{I\!R}_+$, where L is a distributive lattice, is called *log-supermodular* if

$$\mu(x)\mu(y) \le \mu(x \vee y)\mu(x \wedge y)$$

for all $x, y \in L$. A function $f : L \to \mathrm{I\!R}$ is *non-decreasing* if $f(x) \le f(y)$ whenever $x \le y$ and is *non-increasing* if $f(x) \ge f(y)$ whenever $x \le y$.

Motivated by a problem from statistical mechanics, Fortuin, Kasteleyn, and Ginibre [FGK71] proved the following useful inequality.

Theorem 1.17 (FKG Inequality) *Let L be a finite distributive lattice and let $\mu : L \to \mathrm{I\!R}_+$ be a log-supermodular function. Then, if $f, g : L \to \mathrm{I\!R}_+$ are both non-decreasing or non-increasing, we have*

$$\left(\sum_{x\in L} \mu(x) f(x)\right) \cdot \left(\sum_{x\in L} \mu(x) g(x)\right) \le \left(\sum_{x\in L} \mu(x) f(x) g(x)\right) \cdot \left(\sum_{x\in L} \mu(x)\right) \ .$$

If one of the functions is non-decreasing and the other is non-increasing, then the reverse inequality holds.

It is helpful to view μ as a measure on L. Assuming μ is not identically 0, we can define for any $f : L \to \mathrm{I\!R}$ its expectation by

$$\mathrm{E}_\mu[f] = \frac{\sum_{x\in L} \mu(x) f(x)}{\sum_{x\in L} \mu(x)} \ .$$

In this notation, the FKG inequality asserts that for any log-supermodular μ and any two functions $f, g : L \to \mathrm{I\!R}_+$,

$$\mathrm{E}_\mu[f] \cdot \mathrm{E}_\mu[g] \le \mathrm{E}_\mu[f \cdot g]$$

if f, g are both non-decreasing or non-increasing, and

$$\mathrm{E}_\mu[f] \cdot \mathrm{E}_\mu[g] \ge \mathrm{E}_\mu[f \cdot g]$$

if one of the functions is non-decreasing and the other is non-increasing.

To demonstrate the usefulness of Claim 1.16 and the FKG inequality, we present a simple example.

Example [DR94]: Suppose we throw m balls into n bins uniformly and independently at random, for positive integers m and n. Let X_i be the random variable denoting the number of balls in the ith bin.

A possible *configuration* of the experiment can be represented by a vector $\boldsymbol{a} = (a_1, \ldots, a_m)$, with $a_i \in [n]$ for every $i \in [m]$. This is the configuration where ball i goes into bin a_i for every $i \in [m]$. Define the lattice L to be the set of all such configurations ordered component-wise:

$$\boldsymbol{a} \leq_L \boldsymbol{b} \quad \Leftrightarrow \quad a_i \leq b_i \text{ for every } i \in [m].$$

It turns out that this in fact defines a distributive lattice, with join and meet given by the following equation on the components:

$$(\boldsymbol{a} \wedge \boldsymbol{b})_i = \max[a_i, b_i] \quad \text{and} \quad (\boldsymbol{a} \vee \boldsymbol{b})_i = \min[a_i, b_i] \ .$$

Distributivity follows because of the following property of the integers:

$$\begin{aligned} \min[a, \max[b, c]] &= \max[\min[a, b], \min[a, c]] \ , \\ \max[a, \min[b, c]] &= \min[\max[a, b], \max[a, c]] \ . \end{aligned}$$

Now define $\mu : L \to \mathrm{IR}_+$ by $\mu(\boldsymbol{a}) = 1/n^m$ for every $\boldsymbol{a} \in L$. So defined, μ is trivially log-supermodular, and crucially, it makes each configuration equally likely, representing the fact that the balls are thrown uniformly and independently at random into the bins.

For any configuration $\boldsymbol{a}$ and any $i \in [n]$, we introduce the random variable

$$X_i(\boldsymbol{a}) = |\{j : \ a_j = i\}| \ .$$

This gives the number of balls in the ith bin in configuration $\boldsymbol{a}$. Let $I, J \subseteq [n]$ be two index sets such that either $I \cap J = \emptyset$ or $I \cup J = [n]$. With no loss of generality, we can arrange it by renumbering, so that $J = \{1, \ldots, |J|\}$ and $I = \{n - |I| + 1, \ldots, n\}$. Let t_I and t_J be arbitrary non-negative integers. Define $f, g : L \to \mathrm{IR}_+$ as indicator functions by

$$f(\boldsymbol{a}) = \begin{cases} 1 & : \ \sum_{i \in I} X_i(\boldsymbol{a}) \geq t_I \\ 0 & : \ \text{otherwise} \end{cases}$$

and

$$g(\boldsymbol{a}) = \begin{cases} 1 & : \ \sum_{j \in J} X_j(\boldsymbol{a}) \geq t_J \\ 0 & : \ \text{otherwise} \end{cases}$$

The definition of the lattice order ensures that f is non-decreasing while g is non-increasing. Also, note that

$$\begin{aligned} \sum_{\boldsymbol{a} \in L} \mu(\boldsymbol{a}) f(\boldsymbol{a}) &= \frac{1}{n^m} \cdot |\{\boldsymbol{a} \in L : \ \sum_{i \in I} X_i(\boldsymbol{a}) \geq t_I\}| \\ &= \Pr[\sum_{i \in I} X_i \geq t_I] \ . \end{aligned}$$

Similarly,

$$\begin{aligned} \sum_{\boldsymbol{a} \in L} \mu(\boldsymbol{a}) g(\boldsymbol{a}) &= \Pr[\sum_{j \in J} X_j \geq t_J] \quad \text{and} \\ \sum_{\boldsymbol{a} \in L} \mu(\boldsymbol{a}) f(\boldsymbol{a}) g(\boldsymbol{a}) &= \Pr[\sum_{i \in I} X_i \geq t_I \ \wedge \ \sum_{j \in J} X_j \geq t_J] \ . \end{aligned}$$

Applying the FKG inequality, we obtain the following correlation inequality on the random variables X_i:

Claim 1.18 *Let $I, J \subseteq [n]$ be index sets such that $I \cap J = \emptyset$ or $I \cup J = [n]$, and let t_I, t_J be arbitrary non-negative integers. Then*

$$\Pr[\sum_{i \in I} X_i \geq t_I \mid \sum_{j \in J} X_j \geq t_J] \leq \Pr[\sum_{i \in I} X_i \geq t_I] \ .$$

This claim and Claim 1.16 together imply that the X_i's are pairwise negatively correlated.

1.2.7 Variance

The *variance* (or *dispersion*) of a random variable X, denoted by $\mathrm{V}[X]$, is defined as

$$\mathrm{V}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2] \ .$$

The number $\sigma = \sqrt{\mathrm{V}[X]}$ is called the *standard deviation* of X.

Basic properties

Since

$$\mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{E}[X^2 - 2X \cdot \mathrm{E}[X] + (\mathrm{E}[X])^2] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2 \ ,$$

we have

$$\mathrm{V}[X] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2 \ .$$

As $\mathrm{E}[X^2] \geq \mathrm{E}[X]^2$ for any random variable X, $\mathrm{V}[X] \geq 0$. Furthermore, it follows from the definition that

$$\mathrm{V}[a + bX] = b^2 \mathrm{V}[X]$$

for any two constants a and b.

Let X and Y be random variables. Then

$$\begin{aligned} \mathrm{V}[X + Y] &= \mathrm{E}[((X - \mathrm{E}[X]) + (Y - \mathrm{E}[Y]))^2] \\ &= \mathrm{V}[X] + \mathrm{V}[Y] + 2\mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])] \ . \end{aligned}$$

$\mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]$ is called the *covariance* of X and Y and denoted by $\mathrm{cov}(X, Y)$. We observe immediately that if X and Y are independent, so are $X - \mathrm{E}[X]$ and $Y - \mathrm{E}[Y]$. Consequently, in this case,

$$\mathrm{cov}(X, Y) = \mathrm{E}[X - \mathrm{E}[X]] \cdot \mathrm{E}[Y - \mathrm{E}[Y]] = 0 \ .$$

Thus, if X and Y are independent, then

$$\mathrm{V}[X + Y] = \mathrm{V}[X] + \mathrm{V}[Y] \ .$$

In general, if $X_1, \ldots, X_n$ are independent, then

$$\mathrm{V}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{V}[X_i] + 2\sum_{i<j} \mathrm{cov}(X_i, X_j) = \sum_{i=1}^{n} \mathrm{V}[X_i] \ .$$

Since $\mathrm{cov}(X_i, X_j) = 0$ if and only if X_i and X_j are uncorrelated, the equality already holds for pairwise uncorrelated $X_1, \ldots, X_n$. Furthermore, the following fact holds.

Fact 1.19 *For any two random variables X and Y,*

- $\mathrm{cov}(X,Y) = 0$ *if and only if X and Y are uncorrelated,*
- $\mathrm{cov}(X,Y) \leq 0$ *if and only if X and Y are negatively correlated, and*
- $\mathrm{cov}(X,Y) \geq 0$ *if and only if X and Y are positively correlated.*

Because of this fact, the correlation property is often defined by the use of cov.

Modeling Dependence

The *conditional variance* of a random variable Y with respect to an event A is defined by

$$\mathrm{V}[Y \mid A] = \mathrm{E}[(Y - \mathrm{E}[Y \mid A])^2 \mid A] \ . \tag{1.11}$$

If the event A is $X = a$ for some random variable X, this equation defines a function f with

$$f(a) = \mathrm{V}[Y \mid X = a] \ .$$

Thus, $\mathrm{V}[Y \mid X]$ is a random variable, namely the random variable $f(X)$. If Y is independent of X then (1.11) implies that $\mathrm{V}[Y \mid X] = \mathrm{V}[Y]$.

A fundamental property of the conditional variance is that

$$\mathrm{V}[Y] = \mathrm{E}[\mathrm{V}[Y \mid A]] + \mathrm{V}[\mathrm{E}[Y \mid A]]$$

for any random variable Y and any event A (which is easy to check). Thus, observe that (in contrast to $\mathrm{E}[\mathrm{E}[Y \mid X]] = \mathrm{E}[Y]$) it does *not* hold in general that $\mathrm{E}[\mathrm{V}[Y \mid X]] = \mathrm{V}[Y]$.

Conditional variances have rarely been considered, and so far no important tail estimates exist that use them. Hence, we will not consider them any further in this thesis.

Tail estimates

The most well-known tail estimate that uses the variance of a random variable is due to Chebychev.

Theorem 1.20 (Chebychev Inequality) *Let X be an arbitrary random variable. Then, for every $k > 0$,*

$$\Pr[|X - \mathrm{E}[X]| \geq k] \leq \frac{\mathrm{V}[X]}{k^2} \ .$$

Proof. From the Markov inequality it follows for all random variables X that

$$\Pr[|X| \geq k] = \Pr[X^2 \geq k^2] \leq \mathrm{E}[X^2]/k^2 \ .$$

Replacing $|X|$ by $|X - \mathrm{E}[X]|$ yields the theorem. □

Let us apply this theorem to some simple examples.

Example 1: Consider again the example in Section 1.2.6, that is, we throw m balls into n bins uniformly and independently at random. Let the binary random variable X_i be 1 if and only if at least one ball falls into bin i, and let $X = \sum_{i=1}^{n} X_i$ count the number of non-empty bins. Since we know from the example above that the X_i's are pairwise negatively correlated, we have

$$\mathrm{V}[X] \leq \sum_{i=1}^{n} \mathrm{V}[X_i] \ .$$

Thus, we only have to calculate the variance of each X_i. It is not difficult to see that

$$\begin{aligned} \mathrm{E}[X_i^2] &= \mathrm{E}[X_i] = \sum_{j=1}^{m} \binom{m}{j} \left(\frac{1}{n}\right)^j \left(1 - \frac{1}{n}\right)^{m-j} \\ &= 1 - F(n,m) \ . \end{aligned}$$

where $F(n,m) = (1 - 1/n)^m$. Hence,

$$\begin{aligned} \mathrm{V}[X_i] &= \mathrm{E}[X_i^2] - \mathrm{E}[X_i]^2 = (1 - F(n,m)) - (1 - F(n,m))^2 \\ &= F(n,m)(1 - F(n,m)) \ . \end{aligned}$$

Therefore, $\mathrm{V}[X] \leq n \cdot F(n,m)(1 - F(n,m))$. Moreover,

$$\mathrm{E}[X] = \sum_{i=1}^{n} \mathrm{E}[X_i] = n \cdot (1 - F(n,m)) \ .$$

Suppose that $m = n$ and n is infinitely large. Then $F(n,m) = 1/\mathrm{e}$ according to Section 1.1.2. In this case, $\mathrm{E}[X] = n(1 - 1/\mathrm{e})$ and $\mathrm{V}[X] \leq n/\mathrm{e}$. From the Chebychev Inequality it follows for all $k > 0$ that

$$\Pr\left[|X - n(1 - 1/\mathrm{e})| \geq k\sqrt{n/\mathrm{e}}\right] \leq \frac{1}{k^2}$$

Example 2: (The Coupon Collector Problem)
Suppose that we have n different bins. In each iteration we throw one ball into a bin chosen uniformly at random. The problem is to bound the number of time steps until every bin has at least one ball.

Let T be the random variable that counts the number of time steps until every bin has at least one ball. Our aim is to compute the expected value and the variance of T, so that we can use Chebychev's inequality.

Let the random variable Y_i count the number of time steps that have to pass by until a ball lands in an empty bin, assuming i bins are already filled with at least one ball. Then $T = \sum_{i=0}^{n-1} Y_i$. Clearly, for every $i \in \{0, \ldots, n-1\}$ we have

$$\Pr[Y_i = k] = p_i^{k-1}(1 - p_i)$$

for every $k \in \mathbb{N}$, where $p_i = i/n$. Thus,

$$\begin{aligned} \mathrm{E}[Y_i] &= \sum_{k=1}^{\infty} k \cdot p_i^{k-1}(1 - p_i) = (1 - p_i) \left(\sum_{k=0}^{\infty} p_i^k\right)^2 \\ &= (1 - p_i) \cdot \frac{1}{(1 - p_i)^2} = \frac{n}{n - i} \ . \end{aligned}$$

and

$$\mathrm{E}[T] = \sum_{i=0}^{n-1} \mathrm{E}[Y_i] = \sum_{i=0}^{n-1} \frac{n}{n-i} = n \sum_{j=1}^{n} \frac{1}{j} = n \cdot H_n \; .$$

Next we calculate the variance of each Y_i. It holds that

$$\begin{aligned}
\mathrm{V}[Y_i] &= \mathrm{E}[Y_i^2] - \mathrm{E}[Y_i]^2 = \mathrm{E}[Y_i(Y_i-1)] + \mathrm{E}[Y_i] - \mathrm{E}[Y_i]^2 \\
&= \frac{2p_i}{(1-p_i)^2} + \frac{1}{1-p_i} - \frac{1}{(1-p_i)^2} \\
&= \frac{p_i}{(1-p_i)^2} \; .
\end{aligned}$$

Note that the random variables Y_i are independent. Therefore,

$$\begin{aligned}
\mathrm{V}[T] &= \mathrm{V}\left[\sum_{i=0}^{n-1} Y_i\right] = \sum_{i=0}^{n-1} \mathrm{V}[Y_i] \\
&= \sum_{i=0}^{n-1} \frac{i}{n} \cdot \frac{n^2}{(n-i)^2} = n \sum_{j=1}^{n} \frac{n-j}{j^2} \\
&= n^2 \sum_{j=1}^{n} \frac{1}{j^2} - n \sum_{j=1}^{n} \frac{1}{j} = n^2 \cdot \frac{\pi^2}{6} - n \cdot H_n \\
&\leq n^2 \cdot \frac{\pi^2}{6} \; .
\end{aligned}$$

It follows from the Chebychev inequality that for all $k > 0$,

$$\Pr[|T - n \cdot H_n| \geq k \cdot \pi n/\sqrt{6}] \leq \frac{1}{k^2} \; .$$

This bound is quite weak. In the following we will introduce techniques that will allow to obtain much better tail estimates.

1.2.8 Higher Order Moments

Expectations of powers of random variables are called *moments* and play an essential role in the investigations of probability theory. For us they will be of particular interest in the study of sums of random variables. Given a random variable X, the following expected values of functions of X are often studied:

- $\mathrm{E}[X^k]$, $k \in \mathbb{N}$: kth *moment* of X
- $\mathrm{E}[|X|^k]$, $k \in \mathbb{N}$: kth *absolute moment* of X
- $\mathrm{E}[|X - \mathrm{E}[X]|^k]$, $k \in \mathbb{N}$: kth *absolute central moment* of X
- $\mathrm{E}[X^{(k)}]$, $k \in \mathbb{N}$ where $X^{(k)} = X(X-1)\cdots(X-k+1)$: kth *(descending) factorial moment* of X
- $\mathrm{E}[X^{[k]}]$, $k \in \mathbb{N}$ where $X^{[k]} = X(X+1)\cdots(X+k-1)$: kth *ascending factorial moment* of X
- $\mathrm{E}[\mathrm{e}^{h \cdot X}]$, $h > 0$: *exponential moment* of X

According to Section 1.1.2, these moments obey the following identities:

- For every $h \in \mathbb{R}$,
$$\mathrm{E}[\mathrm{e}^{h \cdot X}] = \sum_{i=0}^{\infty} \frac{h^i \cdot \mathrm{E}[X^i]}{i!} .$$

- For every $r \in \mathbb{N}$,
$$\mathrm{E}[X^{(r)}] = \sum_{i=0}^{r} s_{r,i} \cdot \mathrm{E}[X^i] \quad \text{and} \quad \mathrm{E}[X^r] = \sum_{i=0}^{r} S_{r,i} \cdot \mathrm{E}[X^{(i)}] .$$

- For every $r \in \mathbb{N}$,
$$\mathrm{E}[X^{[r]}] = \sum_{i=0}^{r} |s_{r,i}| \cdot \mathrm{E}[X^i] \quad \text{and} \quad \mathrm{E}[X^r] = \sum_{i=0}^{r} (-1)^{r-i} S_{r,i} \cdot \mathrm{E}[X^{[i]}] .$$

These identities allow us to express moments in terms of other moments.

Basic properties

In the following we list a few properties of moments stated in [Loé77]. Let k be any positive integer. While a kth moment may not exist, absolute moments always exist but may be infinite. On the other hand, if the kth absolute moment of X is finite, then its kth moment exists and is finite; and conversely. Furthermore, since $|X|^r \leq 1 + |X|^s$ for all $0 < r < s$, finiteness of a moment of X implies existence and finiteness of all moments of X of lower order.

When applying the elementary inequality
$$|a + b|^r \leq 2^{r-1}(|a|^r + |b|^r)$$
for all $r \in \mathbb{N}$, replacing a by X and b by Y and taking expectations on both sides, we obtain the following inequality:
$$\mathrm{E}[|X + Y|^r] \leq 2^{r-1}(\mathrm{E}[|X|^r] + \mathrm{E}[|Y|^r]) \quad \text{for all } r \in \mathbb{N}.$$

This inequality shows that if the rth absolute moments of X and Y exist and are finite, so is the rth absolute moment of $X + Y$. The *Hölder inequality* gives a relation for products of random variables:
$$\mathrm{E}[|X \cdot Y|] \leq \mathrm{E}[|X|^r]^{1/r} \cdot \mathrm{E}[|Y|^s]^{1/s} \quad \text{for all } r, s \text{ with } r > 1 \text{ and } \tfrac{1}{r} + \tfrac{1}{s} = 1.$$

Hölder's inequality with $r = s = 2$ is called the *Schwartz inequality*:
$$\mathrm{E}[|X \cdot Y|]^2 \leq \mathrm{E}[X^2] \cdot \mathrm{E}[Y^2] .$$

Furthermore, the *Minkowski inequality* states that
$$\mathrm{E}[|X + Y|^r]^{1/r} \leq \mathrm{E}[|X|^r]^{1/r} + \mathrm{E}[|Y|^r]^{1/r} \quad \text{for all } r \in \mathbb{N}.$$

A function f is said to be *convex* on $I \in \mathbb{R}$ if, for every pair of points $x, x' \in I$,

$$f\left(\frac{x+x'}{2}\right) \leq \frac{1}{2}f(x) + \frac{1}{2}f(x') \ .$$

If f is twice differentiable on I, then the convexity property is equivalent to $f'' \geq 0$ on I. *Jensen's inequality* states that if f is convex and $\mathrm{E}[X]$ is finite, then

$$f(\mathrm{E}[X]) \leq \mathrm{E}[f(X)] \ .$$

Since $f(x) = x^r$ is convex for all $r \in \mathbb{N}$, this implies that $\mathrm{E}[|X|] \leq \mathrm{E}[|X|^r]^{1/r}$, which is a special case of the Hölder inequality.

The following fact contains one more Chebychev inequality (see, e.g., [Pet95]).

Fact 1.21 (Chebychev) *Let (a,b) be any (finite or infinite) interval on $\mathbb{R}$. Let u and v be both non-decreasing or both non-increasing functions defined on (a,b). Let X be a random variable taking values only from the interval (a,b). Then*

$$\mathrm{E}[u(X)] \cdot \mathrm{E}[v(X)] \leq \mathrm{E}[u(X) \cdot v(X)]$$

if these means exist. If one of u and v is non-decreasing and the other is non-increasing, then the reverse inequality holds.

Modeling Dependence

Next we use higher moments to define the notion of *association*, which is in some situations more useful than the correlation property. We present here a definition that is not as strict as the definitions in [JP83, DR94], but it suffices for all situations studied in this thesis.

Two variables X and Y are called *associated* if there are $k, \ell \in \mathbb{N}$ with

$$\mathrm{E}[X^k \cdot Y^\ell] \neq \mathrm{E}[X^k] \cdot \mathrm{E}[Y^\ell]$$

or, equivalently, $\mathrm{cov}(X^k, Y^\ell) \neq 0$. Two variables X and Y are called *negatively associated* if, for all $k, \ell \in \mathbb{N}$, $\mathrm{cov}(X^k, Y^\ell) \leq 0$ and are called *positively associated* if, for all $k, \ell \in \mathbb{N}$, $\mathrm{cov}(X^k, Y^\ell) \geq 0$. Clearly, all association properties are symmetric.

In general, a collection of random variables $X_1, \ldots, X_n$ is called *(mutually) non-associated* if

$$\mathrm{E}\left[\prod_{j \in S} X_j^{k_j}\right] = \prod_{j \in S} \mathrm{E}[X_j^{k_j}]$$

for any $S \subseteq [n]$ and any $k_j \in \mathbb{N}$ with $j \in S$. Furthermore, $X_1, \ldots, X_n$ are called *k-wise non-associated* if the property above holds for every S with $|S| \leq k$.

In the following, we list some basic properties of associated random variables. For two binary random variables X and Y it is necessary and sufficient that $\mathrm{cov}(X,Y) \leq 0$ (resp. $\mathrm{cov}(X,Y) \geq 0$) for X and Y to be negatively (resp. positively) associated.

Claim 1.22 *Consider any two non-negative random variables X and Y.*

(1) If $Y \succ (Y|X \geq x)$ for all $x \in \mathbb{R}_+$, then X and Y are negatively associated.

(2) If $(Y|X \geq x) \succ Y$ for all $x \in \mathbb{R}_+$, then X and Y are positively associated.

Proof. We only prove (1). The proof for (2) is analogous. Consider any fixed $k, \ell \in \mathbb{N}$. For all $x, y \in \mathbb{R}_+$, let A_x and B_y be events with

$$A_x = \{\omega \in \Omega : \ X(\omega)^k \geq x\} \quad \text{and} \quad B_y = \{\omega \in \Omega : \ Y(\omega)^\ell \geq y\} .$$

Since $Y \succ (Y|X \geq x)$ for all $x \in \mathbb{R}_+$, $\Pr[Y \geq y] \geq \Pr[Y \geq y \mid X \geq x]$ for all $x, y \in \mathbb{R}_+$, and therefore, $\Pr[Y^\ell \geq y] \geq \Pr[Y^\ell \geq y \mid X^k \geq x]$ for all $x, y \in \mathbb{R}_+$. Hence,

$$\Pr[A_x \cap B_y] = \Pr[A_x] \cdot \Pr[B_y \mid A_x] \leq \Pr[A_x] \cdot \Pr[B_y]$$

for all $x, y \in \mathbb{R}$. The rest of the proof is identical to the proof of Claim 1.16. □

As we will see, our notion of association is strong enough to replace in the following all requirements that a collection of random variables be mutually independent by the requirement that this collection be mutually non-associated. However, note that the fact that a collection of random variables is non-associated does not imply that they are independent. If, for instance, two random variables X and Y are independent, then $f(X)$ and $g(Y)$ are independent for any two functions f and g. Thus, $\mathrm{cov}(f(X), g(Y)) = 0$ for all functions f and g, which may not be the case if X and Y are just known to be non-associated. However, we will show that for certain random variables and certain functions this is true. For this we need the following well-known theorem.

Theorem 1.23 (Weierstrass) *For every continuous function $f : [0,1] \to \mathbb{R}$ and every $\epsilon > 0$ there is a polynomial p such that $|f(x) - p(x)| < \epsilon$ for every $x \in [0,1]$.*

Theorem 1.23 allows to prove the following result.

Theorem 1.24 *For any two discrete, non-associated random variables X and Y and any two (in the ranges of X and Y) non-negative, continuous functions f and g with finite $\mathrm{E}[f(X)]$, $\mathrm{E}[g(Y)]$, and $\mathrm{E}[f(X)g(Y)]$,*

$$\mathrm{E}[f(X)g(Y)] = \mathrm{E}[f(X)] \cdot \mathrm{E}[g(Y)] .$$

Proof. For any discrete random variable X and any (finite or infinite) set $S \subseteq \mathbb{R}$, we define $\mathrm{E}_S[X]$ as

$$\mathrm{E}_S[X] = \sum_{x \in S} x \cdot \Pr[X = x] .$$

Now let X, Y, f, and g be chosen as in Theorem 1.24. Consider any fixed $\epsilon > 0$. Since f and g are non-negative and $\mathrm{E}[f(X)]$, $\mathrm{E}[g(Y)]$, and $\mathrm{E}[f(X)g(Y)]$ are finite, there is is a finite interval $I \subseteq \mathbb{R}$ such that

$$\mathrm{E}_{\mathbb{R}\setminus I}[f(X)] < \epsilon , \quad \mathrm{E}_{\mathbb{R}\setminus I}[g(Y)] < \epsilon , \quad \text{and} \quad \mathrm{E}_{\mathbb{R}\setminus I}[f(X)g(Y)] < \epsilon . \tag{1.12}$$

As f and g are continuous in I, there must be two polynomials p and q according to Theorem 1.23 with $|f(z) - p(z)| < \epsilon$ and $|g(z) - q(z)| < \epsilon$ for every $z \in I$. Since f and g are non-negative,

$$f(x)g(y) \leq (p(x) + \epsilon)(q(y) + \epsilon)$$

for all $x, y \in I$. Hence,

$$\begin{aligned} \mathrm{E}_I[f(X)g(Y)] &\leq \mathrm{E}_I[(p(X)+\epsilon)(q(Y)+\epsilon)] \\ &= (\mathrm{E}_I[p(X)]+\epsilon)(\mathrm{E}_I[q(Y)]+\epsilon)\ , \end{aligned} \tag{1.13}$$

where the equality follows from the fact that X and Y are non-associated. Furthermore, since $f(z) \geq p(z) - \epsilon$ and $g(z) \geq q(z) - \epsilon$ for every $z \in I$,

$$\mathrm{E}_I[f(X)] \geq \mathrm{E}_I[p(X)] - \epsilon \quad \text{and} \quad \mathrm{E}_I[g(Y)] \geq \mathrm{E}_I[q(Y)] - \epsilon\ . \tag{1.14}$$

Combining (1.13) and (1.14) yields

$$\mathrm{E}_I[f(X)g(Y)] \leq (\mathrm{E}_I[f(X)] + 2\epsilon)(\mathrm{E}_I[g(X)] + 2\epsilon)\ . \tag{1.15}$$

From (1.12) it follows that

$$\begin{aligned} \mathrm{E}[f(X)g(Y)] &\leq \mathrm{E}_I[f(X)g(Y)] + \epsilon \\ &\leq (\mathrm{E}_I[f(X)] + 2\epsilon)(\mathrm{E}_I[g(X)] + 2\epsilon) + \epsilon \\ &\leq (\mathrm{E}[f(X)] + 2\epsilon)(\mathrm{E}[g(X)] + 2\epsilon) + \epsilon\ . \end{aligned} \tag{1.16}$$

On the other hand,

$$f(x)g(y) \geq (p(x) - \epsilon)(q(y) - \epsilon) - 4\epsilon^2$$

for all $x, y \in I$. (The $4\epsilon^2$ is due to the fact that, in the worst case, $f(x) = g(y) = 0$ and $p(x) = q(y) = -\epsilon$, resulting in $f(x)g(y) = (p(x) - \epsilon)(q(y) - \epsilon) - 4\epsilon^2$.) Thus, since X and Y are non-associated,

$$\begin{aligned} \mathrm{E}_I[f(X)g(Y)] &\geq (\mathrm{E}_I[p(X)] - \epsilon)(\mathrm{E}_I[q(Y)] - \epsilon) - 4\epsilon^2 \\ &\geq (\mathrm{E}_I[f(X)] - 2\epsilon)(\mathrm{E}_I[g(Y)] - 2\epsilon) - 4\epsilon^2\ . \end{aligned} \tag{1.17}$$

From (1.12) it follows that

$$\begin{aligned} \mathrm{E}[f(X)g(Y)] &\geq \mathrm{E}_I[f(X)g(Y)] \\ &\geq (\mathrm{E}_I[f(X)] - 2\epsilon)(\mathrm{E}_I[g(Y)] - 2\epsilon) - 4\epsilon^2 \\ &\geq (\mathrm{E}[f(X)] - 3\epsilon)(\mathrm{E}[g(Y)] - 3\epsilon) - 4\epsilon^2\ . \end{aligned} \tag{1.18}$$

Combining (1.16) and (1.18) yields

$$\mathrm{E}[f(X)g(Y)] \stackrel{\epsilon \rightrightarrows 0}{=} \mathrm{E}[f(X)] \cdot \mathrm{E}[g(Y)]\ .$$

□

If we assume to have random variables that only take a finite number of values in $\mathbb{R}$, Theorem 1.24 can be improved to the following result.

Theorem 1.25 *For any two non-associated random variables X and Y taking only a finite number of values in $\mathbb{R}$ and any two functions f and g with finite $f(x)$ for any $x \in \mathbb{R}$ with $\Pr[X = x] > 0$ and finite $g(y)$ for any $y \in \mathbb{R}$ with $\Pr[Y = y] > 0$,*

$$\mathrm{E}[f(X)g(Y)] = \mathrm{E}[f(X)] \cdot \mathrm{E}[g(Y)]\ .$$

Proof. Since X and Y only take a finite number of values in $\mathbb{R}$, there are polynomials p and q such that $f(x) = p(x)$ for any $x \in \mathbb{R}$ with $\Pr[X = x] > 0$ and $g(y) = q(y)$ for any $y \in \mathbb{R}$ with $\Pr[Y = y] > 0$. Thus, we have

$$\mathrm{E}[f(X)g(Y)] = \mathrm{E}[p(X)q(Y)] = \mathrm{E}[p(X)] \cdot \mathrm{E}[q(Y)] = \mathrm{E}[f(X)] \cdot \mathrm{E}[g(Y)] \ .$$

□

Theorems 1.24 and 1.25 demonstrate that the expectation of products of functions of non-associated random variables can be treated in many cases as if the random variables were independent.

Tail estimates

Loéve stated the following result in [Loé77], which easily follows from the Markov inequality.

Theorem 1.26 (Loéve Inequality) *Let X be an arbitrary non-negative random variable and g be an arbitrary function that is non-negative and non-decreasing on $\mathbb{R}_+$. Then, for any $k \geq 0$ with $g(k) > 0$,*

$$\Pr[X \geq k] \leq \frac{\mathrm{E}[g(X)]}{g(k)} \ .$$

Proof. Since g is non-negative and non-decreasing on $\mathbb{R}_+$, it holds for every $k \geq 0$ that

$$\mathrm{E}[g(X)] = \sum_{x \in \mathbb{R}_+} g(x) \cdot \Pr[X = x] \geq g(k) \Pr[X \geq k] \ .$$

□

1.2.9 Basic Probability Distributions

In this section we introduce the most important probability distributions used in combination with discrete random variables. Many of the results presented in this section have been taken from [JK77]. Besides giving the expected value and the variance of these distributions, we will also state the factorial moments. Using these moments in Theorem 1.26 already yields very good tail estimates.

For instance, consider a random variable X with finite descending factorial moment $\mathrm{E}[X^{(r)}]$ for some $r \in \mathbb{N}$. Then, for any $k \in \mathbb{N}$ with $k \geq r$,

$$\Pr[X \geq k] \leq \frac{\mathrm{E}[X^{(r)}]}{k^{(r)}} \ .$$

Uniform distribution

A random variable X is called *uniformly distributed* over a finite set M of values in $\mathbb{R}$ if

$$\Pr[X = x] = \frac{1}{|M|}$$

for all $x \in M$. If $M = \{a, a+1, \ldots, b\}$ for some integers a, b with $a < b$, then

$$\mathrm{E}[X] = \frac{a+b}{2} \quad \text{and} \quad \mathrm{V}[X] = \frac{(b-a)^2}{12} + \frac{b-a}{6}$$

Since it is very easy to obtain sharp tail estimates in this case, we omit the definition of factorial moments.

Binomial distribution

A random variable X is called *binomially distributed* if there are parameters $p \in [0,1]$ and $n \in \mathbb{N}$ such that
$$\Pr[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$$
for all $k \in \{0, \ldots, n\}$. In this case,
$$\mathrm{E}[X] = n \cdot p \quad \text{and} \quad \mathrm{V}[X] = n \cdot p(1-p) \; .$$
Furthermore,
$$\mathrm{E}[X^{(r)}] = n^{(r)} p^r$$
for all $r \in \mathbb{N}$.

As an example, consider the situation that we have m balls in an urn, of which m_1 are red and m_2 are blue. We draw n balls at random out of this urn, one after the other, replacing each drawn ball. Let the random variable X denote the number of red balls drawn. Then X is binomially distributed with $p = m_1/m$.

Poisson distribution

A random variable X is called *Poisson distributed* if there is a parameter $\lambda > 0$ such that
$$\Pr[X = k] = \frac{\lambda^k}{k!} \cdot \mathrm{e}^{-\lambda}$$
for all $k \in \mathbb{N}_0$. In this case,
$$\mathrm{E}[X] = \lambda \quad \text{and} \quad \mathrm{V}[X] = \lambda \; .$$
Furthermore,
$$\mathrm{E}[X^{(r)}] = \lambda^r$$
for all $r \in \mathbb{N}$.

The binomial distribution converges towards the Poisson distribution if, in the example above, the number of balls in the urn, m, and the number of draws, n, is increased while keeping the number of red balls constant.

Geometric distribution

A random variable X is called *geometrically distributed* if there is a parameter $p \in [0,1]$ such that
$$\Pr[X = k] = (1-p)^{k-1} p$$
for all $k \in \mathbb{N}$. In this case,
$$\mathrm{E}[X] = \frac{1}{p} \quad \text{and} \quad \mathrm{V}[X] = \frac{1-p}{p^2} \; .$$
Furthermore,
$$\mathrm{E}[X^{[r]}] = \frac{r!}{p^r}$$

for all $r \in \mathbb{N}$.

As an example, suppose that we again have an urn with m balls, of which m_1 balls are red and m_2 balls are blue. We draw a ball at random from this urn and replace it until we draw a red ball. Let the random variable X denote the number of balls drawn. Then X is geometrically distributed with $p = m_1/m$.

Hypergeometric distribution

A random variable X is called *hypergeometrically distributed* if there are parameters N, M, and n such that

$$\Pr[X = k] = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

for all $k \in \{0, \ldots, n\}$. In this case, we obtain for $p = M/N$

$$\mathrm{E}[X] = n \cdot p \quad \text{and} \quad \mathrm{V}[X] = \frac{N-n}{N-1} \cdot n \cdot p(1-p) \ .$$

Furthermore,

$$\mathrm{E}[X^{(r)}] = \frac{n^{(r)} \cdot M^{(r)}}{N^{(r)}}$$

for all $r \in \mathbb{N}$.

As an example, consider the situation that we have N balls in an urn, of which M are red and $N-M$ are blue. We draw n balls at random out of this urn, one after the other, *without* replacing each drawn ball. Let the random variable X denote the number of red balls drawn. Then X is hypergeometrically distributed.

Negative binomial distribution

A random variable X is called *negative binomially distributed* if there are parameters $p \in (0, 1]$ and $j \in \mathbb{N}$ such that

$$\Pr[X = k] = \binom{k-1}{j-1} p^j (1-p)^{k-j}$$

for all $k \in \mathbb{N}$ with $k \geq j$. Instead of using p and $1 - p$, the negative binomial distribution is also sometimes expressed by the use of $Q = 1/p$ and $P = (1-p)/p$:

$$\Pr[X = k] = \binom{k-1}{j-1} P^{-j} \left(\frac{P}{Q}\right)^k \ .$$

For this distribution we obtain

$$\mathrm{E}[X] = Q \quad \text{and} \quad \mathrm{V}[X] = j \cdot Q(1-Q) \ .$$

Furthermore,

$$\mathrm{E}[X^{[r]}] = j^{[r]} \cdot Q^r$$

for all $r \in \mathbb{N}$. For $j = 1$ the negative binomial distribution is equal to the geometric distribution with parameter p.

As an example, consider the situation that we have an urn with m balls, of which m_1 balls are red and m_2 balls are blue. We draw one ball after the other at random from this urn and replace it until we have drawn altogether j red balls. Let the random variable X denote the number of balls drawn. Then X is negative binomially distributed with $p = m_1/m$.

Negative hypergeometric distribution

A random variable X is called *negative hypergeometrically distributed* if there are parameters k, j, N, and M such that

$$\Pr[X = k] = \binom{k-1}{j-1} \frac{M^{(j)}(N-M)^{(k-j)}}{N^{(k)}} .$$

for all $k \in \mathbb{N}$ with $k \geq j$. For this distribution we obtain

$$\mathrm{E}[X] = \frac{j \cdot (N+1)}{M+1} \quad \text{and} \quad \mathrm{V}[X] = \frac{j(N+1)(N-M)(M+1-j)}{(M+1)^2(M+2)} .$$

Furthermore,

$$\mathrm{E}[X^{[r]}] = \frac{j^{[r]}(N+1)^{[r]}}{(M+1)^{[r]}} .$$

for all $r \in \mathbb{N}$.

As an example, consider the situation that we have an urn with N balls, of which M balls are red and $N - M$ balls are blue. We draw one ball after the other at random from this urn *without* replacing them until we have drawn altogether j red balls. Let the random variable X denote the number of balls drawn. Then X is negative hypergeometrically distributed.

1.2.10 Approximations of Probability Distributions

In this section we demonstrate that in certain cases we can approximate probability distributions via standard distributions such as the Poisson or Gaussian probability distribution.

Approximating the binomial distribution via the Poisson distribution

Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with probabilities $p_i = \Pr[X_i = 1]$ for all $1 \leq i \leq n$. We define $S_n = \sum_{i=1}^n X_i$. Let $B = (b_0, b_1, \ldots, b_n)$ be the probability distribution of the sum S_n, where $b_k = \Pr[S_n = k]$. Furthermore, let $\Pi = (\pi_0, \pi_1, \ldots)$ be the Poisson distribution with parameter λ, where

$$\pi_k = \frac{\lambda^k}{k!} \cdot \mathrm{e}^{-\lambda} .$$

Let $||B - \Pi||$ be defined as $\sum_{k=0}^{\infty} |b_k - \pi_k|$. Then the following theorem holds.

Theorem 1.27 (Presman [Pre85]) *For $\lambda = \mathrm{E}[S_n]$ it holds*

$$||B - \Pi|| \leq 2 \sum_{i=1}^{n} p_i^2 .$$

Since $\sum_{i=1}^n p_i^2 \leq \lambda \max_{i \in [n]} p_i$, this also implies that $||B - \Pi|| \leq 2\lambda \max_{i \in [n]} p_i$.

Approximations using the normal distribution

The *normal* or *Gaussian* distribution is defined by the probability density function

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \cdot \mathrm{e}^{-x^2/2} .$$

The corresponding distribution function is defined by

$$\Phi(x) = \int_{-\infty}^{x} \phi(u)du = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-u^2/2}du .$$

X is called a *Gaussian random variable* if $\Pr[X \leq x] = \Phi(x)$ for all $x \in \mathrm{IR}$.

Now, consider the function

$$\Theta(x) = \mathrm{e}^{x^2/2} \int_{x}^{\infty} \phi(u)du .$$

Thus $\Theta(0) = 1/2$, and it is not difficult to see that Θ decreases for $x \geq 0$. Furthermore, it has been shown in [IM74] p. 17 and [JK82] p. 505 that

$$\frac{2}{\sqrt{2\pi}(x + \sqrt{x^2+4})} \leq \Theta(x) \leq \frac{4}{\sqrt{2\pi}(3x + \sqrt{x^2+8})} .$$

This implies the following result.

Theorem 1.28 *If X is a Gaussian random variable, then*

$$\frac{2}{\sqrt{2\pi}(x + \sqrt{x^2+4})} \cdot \mathrm{e}^{-x^2/2} \leq \Pr[X \geq x] \leq \frac{4}{\sqrt{2\pi}(3x + \sqrt{x^2+8})} \cdot \mathrm{e}^{-x^2/2}$$

for any $x \geq 0$. Or, in a more simple way,

$$\frac{1}{\sqrt{2\pi}(1+x)} \cdot \mathrm{e}^{-x^2/2} \leq \Pr[X \geq x] \leq \frac{1}{\sqrt{\pi}(1+x)} \cdot \mathrm{e}^{-x^2/2} .$$

As stated in the following well-known result, in the infinite case, this inequality holds also for sums of random variables with any other common distribution.

Theorem 1.29 (Central Limit Theorem) *Let $S_n = \sum_{i=1}^{n} X_i$ be the sum of n discrete independent random variables with common distribution having finite expected value μ and variance ν. Then, for any $a < b$,*

$$\lim_{n\to\infty} \Pr\left[a < \frac{S_n - n\mu}{\sqrt{n \cdot \nu}} < b\right] = \Phi(b) - \Phi(a) .$$

Combining Theorem 1.29 with Theorem 1.28, we obtain the following result.

Corollary 1.30 *Let $S_n = \sum_{i=1}^{n} X_i$ be the sum of n discrete independent random variables with common distribution having finite expected value μ and variance ν. Then, for $n \to \infty$, it holds for any $x \geq 0$ that*

$$\frac{1}{\sqrt{2\pi}(1+x)} \cdot \mathrm{e}^{-x^2/2} \leq \Pr\left[\frac{S_n - n\mu}{\sqrt{n \cdot \nu}} \geq x\right] \leq \frac{1}{\sqrt{\pi}(1+x)} \cdot \mathrm{e}^{-x^2/2}$$

The drawback of the Central Limit Theorem is that it does not say anything about the rate of convergence towards the Gaussian distribution. We will quantify this rate of convergence for the case that S_n is normalized to

$$S_n = \frac{\sum_{i=1}^n X_i}{\sqrt{n \cdot \nu}}$$

where $X_1, \ldots, X_n$ is a sequence of independent, identically distributed random variables with $\mathrm{E}[X_i] = 0$, $\mathrm{V}[X_i] = \nu$, and $\mathrm{E}[|X_1|^3] = \gamma < \infty$. For all $x \in \mathbb{R}$, let $F_n(x) = \Pr[S_n \leq x]$. Then the following theorem holds (for a proof, see e.g., [GK68]).

Theorem 1.31 (Berry and Esseen) *For every $n \in \mathbb{N}$ we have*

$$\sup_x |F_n(x) - \Phi(x)| \leq \frac{\gamma}{\sqrt{n \cdot \nu^3}} \cdot \frac{1}{(1+|x|)^3} \, .$$

The bound in this theorem is sharp up to some constant $(2\pi)^{-1/2}$ [Shi96].

Convergence bounds have also been shown for non-identically distributed random variables. Let $X_1, \ldots, X_n$ be independent binary random variables and let

$$S = \sum_{i=1}^n X_i, \quad \mu = \mathrm{E}[S] = \sum_{i=1}^n \mathrm{E}[X_i], \quad \sigma^2 = \mathrm{V}[S], \quad \text{and } \gamma = \mathrm{E}[(S-\mu)^3] \, .$$

Furthermore, let

$$\Gamma(x) = \Phi(x) + \frac{\gamma}{6\sigma^3}(1-x^2)\phi(x) \, .$$

Then the following theorem holds.

Theorem 1.32 (Mikhailov [Mik93]) *Suppose that $\sigma \geq 10$. Then, for every $0 \leq m \leq n$,*

$$\left| \Pr[S \leq m] - \Gamma\left(\frac{m - \mu + 1/2}{\sigma}\right) \right| \leq \frac{\sigma + 3}{4\sigma^3} \, .$$

As noted by Mikhailov, particular examples demonstrate that the theorem also holds in the range $0 < \sigma \leq 10$. The restriction $\sigma \geq 10$ is only a technical problem. For the range $5 \leq \sigma \leq 10$, the same proof yields a rougher bound of $(2\sigma + 5)/(5\sigma^3)$.

1.3 Tail Estimates

In the following we present tail estimates for sums of random variables that are in many cases much better than the tail estimates provided by the theorems in Section 1.2.10, and that apply to much more cases than studied in Section 1.2.9. We will always assume that $X_1, \ldots, X_n$ are discrete random variables with finite first and second moments. We define

$$S_n = \sum_{i=1}^n X_i$$

to be the sum of these variables and

$$A_n = \frac{S_n}{n}$$

to be the average. If $\mathrm{E}[X_i] = \mu_i$ and $\mathrm{V}[X_i] = \nu_i$ for all $i \in [n]$, then

$$\mathrm{E}[S_n] = \sum_{i=1}^{n} \mu_i \quad \text{and} \quad \mathrm{V}[S_n] = \sum_{i=1}^{n} \nu_i \ .$$

Furthermore,

$$\mathrm{E}[A_n] = \mathrm{E}[S_n]/n \quad \text{and} \quad \mathrm{V}[A_n] = \mathrm{V}[S_n]/n^2 \ .$$

We often use $\mu = \mathrm{E}[S_n]$, $\nu = \mathrm{V}[S_n]$, $\bar{\mu} = (\mu_1 + \ldots + \mu_n)/n$, and $\bar{\nu} = (\nu_1 + \ldots + \nu_n)/n$.

Let us first review the concepts of majorization and Schur-convexity. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be vectors in $\mathbb{R}^n$, and let $x_{[i]}$ and $y_{[i]}$ denote the ith largest component of $\boldsymbol{x}$ and $\boldsymbol{y}$ respectively. Then we say that $\boldsymbol{x} \prec \boldsymbol{y}$ ($\boldsymbol{x}$ is *majorized* by $\boldsymbol{y}$) if

$$\sum_{i=1}^{k} x_{[i]} \leq \sum_{i=1}^{k} y_{[i]} \quad \text{for all } 1 \leq k \leq n \text{ and } \quad \sum_{i=1}^{n} x_{[i]} = \sum_{i=1}^{n} y_{[i]} \ .$$

To give an intuition for this definition, $\boldsymbol{x}$ is majorized by $\boldsymbol{y}$ if its entries are, in a certain sense, more balanced than the entries in $\boldsymbol{y}$. (Note that the symbol $\prec$ has already been used for stochastic domination. Since for both notions $\prec$ is common in the literature, we avoided using a different symbol for one of them. To ensure a unique meaning of $\prec$, it will always mean majorization in the context of vectors and dominance in the context of random variables.)

A function f of n arguments is *Schur-convex* on a set $A \subseteq \mathbb{R}^n$ if, for all $\boldsymbol{x}, \boldsymbol{y} \in A$,

$$\boldsymbol{x} \prec \boldsymbol{y} \quad \Rightarrow f(\boldsymbol{x}) \leq f(\boldsymbol{y}) \ .$$

A function f is *Schur-concave* if

$$\boldsymbol{x} \prec \boldsymbol{y} \quad \Rightarrow f(\boldsymbol{x}) \geq f(\boldsymbol{y}) \ .$$

Since $\prec$ is transitive (i.e., if $\boldsymbol{x} \prec \boldsymbol{y}$ and $\boldsymbol{y} \prec \boldsymbol{z}$, then $\boldsymbol{x} \prec \boldsymbol{z}$), it is often quite easy to prove that a function f is Schur-convex or Schur-concave: all that has to be shown is that for every vector, all of its direct successors or predecessors concerning $\prec$ fulfill the respective Schur condition.

The following fact follows immediately from the definition of Schur-convex and Schur-concave functions.

Fact 1.33 *Assume that we have a set A of vectors with the property that for some $\boldsymbol{x} \in A$, $\boldsymbol{x} \prec \boldsymbol{y}$ for all $\boldsymbol{y} \in A$. If f is Schur-convex on A, then $f(\boldsymbol{x}) \leq f(\boldsymbol{y})$ for all $\boldsymbol{y} \in A$. If f is Schur-concave, the inequality holds with $\geq$ in place of $\leq$.*

Furthermore, assume that we have a set B of vectors with the property that for some $\boldsymbol{x} \in B$, $\boldsymbol{x} \succ \boldsymbol{y}$ for all $\boldsymbol{y} \in B$. If f is Schur-convex on A, then $f(\boldsymbol{x}) \geq f(\boldsymbol{y})$ for all $\boldsymbol{y} \in B$. If f is Schur-concave, the inequality holds with $\leq$ in place of $\geq$.

The fact immediately implies the following important result. (We call a set $A \subseteq \mathbb{R}^n$ *convex* w.r.t. $\succ$ if all $\boldsymbol{x} \in A$ fulfill $\sum_i x_i = m$ for some fixed $m \in \mathbb{R}$ and for every $\boldsymbol{x}$ in A it holds that all $\boldsymbol{y} \in \mathbb{R}^n$ with $\boldsymbol{y} \prec \boldsymbol{x}$ are also in A.)

Corollary 1.34 *Let f be a Schur-convex function on a convex set $A \subseteq \mathbb{R}^n$. Then for any $\boldsymbol{x} = (x_1, \ldots, x_n) \in A$,*

$$f(x_1, \ldots, x_n) \geq f(m, \ldots, m) \quad \textit{where } m = \frac{x_1 + \ldots + x_n}{n} \; .$$

For a Schur-concave function f, the inequality holds with $\leq$ in place of $\geq$.

Corollary 1.34 will be useful when dealing with sums of non-identically distributed random variables.

1.3.1 Tail Estimates Based on the Expectation

First, we study sums of independent binary random variables. Tail estimates for these sums have among many others been investigated by Chernoff [Che52] and are often called *Chernoff bounds*. It has been found convenient to use exponential moments of random variables to derive these tail estimates. The idea of using exponential moments was apparently first used by S. N. Bernstein. His method had a significant influence on deriving tail estimates also for many other problems. We present here a proof given by Hagerup and Rüb [HR90]. Bound (1.20) was taken from [McD98].

Theorem 1.35 (Chernoff Bound) *If $X_1, \ldots, X_n$ are independent binary random variables, then it holds for all $\delta \geq 0$ that*

$$\begin{aligned} \Pr[S_n \geq (1+\delta)\mu] &\leq \left(\frac{\mathrm{e}^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu} && (1.19) \\ &\leq \mathrm{e}^{-\frac{\delta^2\mu}{2(1+\delta/3)}} && (1.20) \\ &\leq \mathrm{e}^{-\min[\delta^2,\,\delta]\cdot\mu/3} \; . \end{aligned}$$

Furthermore, it holds for all $0 \leq \delta \leq 1$ that

$$\begin{aligned} \Pr[S_n \leq (1-\delta)\mu] &\leq \left(\frac{\mathrm{e}^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu} && (1.21) \\ &\leq \mathrm{e}^{-\delta^2\mu/2} \; . \end{aligned}$$

Proof. We will only show (1.19) and (1.21). First we present a proof for inequality (1.19). For all $i \in [n]$, let $p_i = \mathrm{E}[X_i]$. According to the Loéve inequality (see Theorem 1.26) we obtain that, for any function $g(x) = e^{h \cdot x}$ with $h > 0$ and any $\delta \geq 0$,

$$\Pr[S_n \geq (1+\delta)\mu] \leq \mathrm{e}^{-h(1+\delta)\mu} \cdot \mathrm{E}[\mathrm{e}^{h \cdot S_n}] \; . \quad (1.22)$$

Since $X_1, \ldots, X_n$ are independent, it follows from Claim 1.10 that

$$\begin{aligned} \mathrm{E}[\mathrm{e}^{h \cdot S_n}] &= \mathrm{E}[\mathrm{e}^{h(X_1+\ldots+X_n)}] = \mathrm{E}[\mathrm{e}^{h \cdot X_1} \cdots \mathrm{e}^{h \cdot X_n}] = \textstyle\prod_{i=1}^n \mathrm{E}[\mathrm{e}^{h \cdot X_i}] \\ &= \prod_{i=1}^n (p_i \mathrm{e}^h + (1-p_i)) = \textstyle\prod_{i=1}^n (1 + p_i(\mathrm{e}^h - 1)) \\ &\leq \prod_{i=1}^n \mathrm{e}^{p_i(\mathrm{e}^h-1)} \quad \text{since } 1 + x \leq \mathrm{e}^x \text{ for all } x \\ &= \mathrm{e}^{\mu(\mathrm{e}^h-1)} \; . \end{aligned}$$

This yields together with inequality (1.22) that

$$\Pr[S_n \geq (1+\delta)\mu] \leq \mathrm{e}^{-h(1+\delta)\mu} \cdot \mathrm{e}^{\mu(\mathrm{e}^h - 1)} = \mathrm{e}^{-(1+h(1+\delta)-\mathrm{e}^h)\mu} \ . \tag{1.23}$$

The right-hand side of (1.23) attains its minimum at $h = h_0$, where $h_0 = \ln(1+\delta)$. Inserting this in (1.23) yields

$$\Pr[S_n \geq (1+\delta)\mu] \leq (1+\delta)^{-(1+\delta)\mu} \cdot \mathrm{e}^{\delta \cdot \mu} = \left(\frac{\mathrm{e}^\delta}{(1+\delta)^{1+\delta}}\right)^\mu \ .$$

This completes the proof for inequality (1.19). The proof of (1.21) is similar and can be found in Section 1.3.4. □

If $X_1, \ldots, X_n$ are identically distributed binary random variables, the bounds in Theorem 1.35 can be easily improved:

Suppose that $\mathrm{E}[X_i] = p$ for all $i \in [n]$. Let $0 < t < 1$, $t \geq p$, and $h \geq 0$. Choosing δ to make $(1+\delta)\mu = t \cdot n$, we get from the calculations in the proof of Theorem 1.35 that

$$\Pr[S_n \geq t \cdot n] \leq \mathrm{e}^{-h \cdot t \cdot n}(p \cdot \mathrm{e}^h + (1-p))^n \ .$$

For $h = \ln(t(1-p)/[p(1-t)])$ this becomes

$$\Pr[S_n \geq t \cdot n] \leq \left(\frac{p(1-t)}{t(1-p)}\right)^{t \cdot n} \left(\frac{t(1-p)}{1-t} + (1-p)\right)^n = \left(\frac{p(1-t)}{t(1-p)}\right)^{t \cdot n} \left(\frac{1-p}{1-t}\right)^n \ ,$$

from which we obtain

$$\Pr[S_n \geq t \cdot n] \leq \left[\left(\frac{p}{t}\right)^t \left(\frac{1-p}{1-t}\right)^{1-t}\right]^n \ . \tag{1.24}$$

It is not difficult to show that this bound is better than the bounds in Theorem 1.35. Therefore, one may ask, whether this bound also holds for non-identically distributed binary random variables. In fact, Gleser [Gle75] showed that if $X_1, \ldots, X_n$ are independent binary random variables with $\Pr[X_i = 1] = p_i$ for all $i \in [n]$, then the function

$$F(p_1, \ldots, p_n, t) = \Pr[X_1 + \ldots + X_n \leq t]$$

is Schur-concave in $(p_1, \ldots, p_n)$ for $0 \leq t \leq n\bar{p} - 2$ and Schur-convex for $n\bar{p} + 1 \leq t \leq n$, where $\bar{p} = (p_1 + \ldots + p_n)/n$. Using this property together with Corollary 1.34 implies that the tail estimate (1.24) also holds for non-identically distributed binary random variables. This will be shown explicitly in the proof of the next theorem.

Bounded random variables

The next theorem is due to Hoeffding [Hoe63]. Therefore, it is usually called *Chernoff-Hoeffding bound* or *Hoeffding bound*.

Theorem 1.36 (Hoeffding Bound) *If $X_1, \dots, X_n$ are independent and $0 \le X_i \le 1$ for all $i \in [n]$, then for $\bar{\mu} < t < 1$*

$$\begin{aligned} \Pr[A_n \ge t] &\le \left[\left(\frac{\bar{\mu}}{t}\right)^t \left(\frac{1-\bar{\mu}}{1-t}\right)^{1-t} \right]^n && (1.25) \\ &\le \mathrm{e}^{-nd^2 g(\bar{\mu})} && (1.26) \\ &\le \mathrm{e}^{-2nd^2} , && (1.27) \end{aligned}$$

where $d = t - \bar{\mu}$ and

$$g(x) = \begin{cases} \frac{1}{1-2x} \ln \frac{1-x}{x} & : \quad 0 < x < \frac{1}{2} \\ \frac{1}{2x(1-x)} & : \quad \frac{1}{2} \le x < 1 \end{cases} .$$

The proof of Theorem 1.36 can be found in Section 1.3.4. It is easy to extend the theorem to arbitrary sequences of random variables with $a \le X_i \le b$ by simply shifting the X_i's to a range of $[0, b-a]$ and then scaling them to $[0,1]$. In this case, the values $\bar{\mu}$, t, and d in the three upper bounds of the theorem have to be replaced by $(\bar{\mu}-a)/(b-a)$, $(t-a)/(b-a)$, and $d/(b-a)$, respectively. As noted by Hoeffding [Hoe63], the bound in (1.25) is the best that can be obtained from the exponential moment method (i.e., using the Loéve Inequality with $g(x) = \mathrm{e}^{hx}$). Naveau [Nav97] and Philips and Nelson [PN95] showed that there are functions that yield better tail estimates (we will come back to this point in the proof of Theorem 1.46).

If the ranges of the random variables differ significantly, then it might be more useful to apply the following theorem, which is also due to Hoeffding. Its proof can be found, e.g., in [Hoe63, McD98].

Theorem 1.37 ([Hoe63]) *If $X_1, \dots, X_n$ are independent and $a_i \le X_i \le b_i$ for all $i \in [n]$, then for $d > 0$*

$$\Pr[A_n - \bar{\mu} \ge d] \le \mathrm{e}^{-2n^2d^2/(\sum_{i=1}^n (b_i - a_i)^2)} .$$

Unbounded random variables

If some of the X_i's are unbounded, then the previous theorems cannot be applied. We will show that for the case that the X_i's are independent geometrically distributed random variables, good tail estimates can nevertheless be shown. Recall that a random variable is geometrically distributed if there is a $0 < p < 1$ with $\Pr[X = k] = (1-p)^{k-1}p$ for all $k \ge 1$. See Section 1.3.4 for the proof of the following result.

Theorem 1.38 *Let $X_1, \dots, X_n$ are independent, geometrically distributed random variables with parameters $p_1, \dots, p_n \in (0,1)$, and let $p = \min_i p_i$. Then*

- *for all $\delta > 0$ it holds that*

$$\Pr[S_n \ge (1+\delta)\mu] \le \left(1 + \frac{\delta\mu p}{n}\right)^n \mathrm{e}^{-\delta\mu p} , \tag{1.28}$$

- *and for all $0 \le \delta \le 1$, it holds that*

$$\Pr[S_n \le (1-\delta)\mu] \le \mathrm{e}^{-\delta^2\mu p/2} . \tag{1.29}$$

This theorem can be used, for instance, to prove much better tail estimates for the Coupon Collector Problem than presented in Example 2 of Section 1.2.7.

The following theorem due to Merkle and Petrovic [MP97] offers alternative ways of proving tail estimates for geometrical — and, in general, negative binomial — random variables.

Theorem 1.39 ([MP97]) *Let $X_1, \ldots, X_n$ be independent geometric random variables with parameters $p_1, \ldots, p_n$. That is,*

$$\Pr[X_i = k] = (1 - p_i)^{k-1} p_i$$

for all $k \geq 1$. Then the function

$$F(p_1, \ldots, p_n, t) = \Pr[X_1 + \ldots + X_n \leq t]$$

is Schur-concave with respect to $(p_1, \ldots, p_n)$ for every real t.

Theorem 1.39 and Fact 1.33 together imply that $\Pr[S_n \geq t]$ is maximal if as many of the X_i as possible have a parameter p_i that is equal to $p = \min_i p_i$. Furthermore, $\Pr[S_n \leq t]$ is maximal if all of the X_i have the same probability parameter. This allows to bound $\Pr[S_n \leq t]$ in terms of $\bar{p} = (p_1 + \ldots + p_n)/n$ instead of using p as in Theorem 1.38. However, note that changing the parameters $p_1, \ldots, p_n$ changes $\mathrm{E}[S_n]$, even if their sum $p_1 + \ldots + p_n$ remains fixed.

Using bounds for the expectation

Sometimes it may not be possible to calculate $\mathrm{E}[S_n]$ exactly. In this case, it may be more convenient to use upper or lower bounds for $\mathrm{E}[S_n]$. We will show that in all bounds presented above, $\mathrm{E}[S_n]$ can be replaced by upper or lower bounds on $\mathrm{E}[S_n]$ (that do not violate the restrictions on the X_i in S_n), depending on whether we need estimates for $\Pr[S_n \geq t]$ or for $\Pr[S_n \leq t]$. For this we need the following claim. (Recall that here "$\succ$" means stochastic domination.)

Claim 1.40 *If X_1, X_2, Y_1, and Y_2 are independent random variables with $X_1 \succ Y_1$ and $X_2 \succ Y_2$, then $(X_1 + X_2) \succ (Y_1 + Y_2)$.*

Proof. For all $z \in \mathbb{R}$ it holds:

$$\begin{aligned}
\Pr[X_1 + X_2 \geq z] &= \sum_{z'} \Pr[X_1 = z'] \cdot \Pr[X_2 \geq z - z'] \\
&\geq \sum_{z'} \Pr[X_1 = z'] \cdot \Pr[Y_2 \geq z - z'] \\
&= \Pr[X_1 + Y_2 \geq z] = \sum_{z'} \Pr[Y_2 = z'] \cdot \Pr[X_1 \geq z - z'] \\
&\geq \sum_{z'} \Pr[Y_2 = z'] \cdot \Pr[Y_1 \geq z - z'] = \Pr[Y_1 + Y_2 \geq z] \ .
\end{aligned}$$

□

This claim immediately yields the following result.

Lemma 1.41 *For any two sets of independent random variables $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ it holds:*

- *if* $X_i \succ Y_i$ *for all* $i \in [n]$, *then*

$$\Pr\left[\sum_{i=1}^{n} X_i \leq t\right] \leq \Pr\left[\sum_{i=1}^{n} Y_i \leq t\right]$$

for every $t \in \mathrm{IR}$, *and*

- *if* $X_i \prec Y_i$ *for all* $i \in [n]$, *then*

$$\Pr\left[\sum_{i=1}^{n} X_i \geq t\right] \leq \Pr\left[\sum_{i=1}^{n} Y_i \geq t\right]$$

for every $t \in \mathrm{IR}$.

Clearly, for any of the cases studied in Section 1.3.1 it holds that for any sum S_n of random variables with $\mathrm{E}[S_n] \leq \mu$, where μ does not violate the restrictions on the random variables in S_n, a sum S'_n of random variables with the same restrictions that dominate the variables in S_n can be defined with $\mathrm{E}[S'_n] = \mu$. Furthermore, for any sum S_n of random variables with $\mathrm{E}[S_n] \geq \mu$, where μ does not violate the restrictions on the random variables in S_n, a sum S'_n of random variables with the same restrictions that are dominated by the variables in S_n can be defined with $\mathrm{E}[S'_n] = \mu$. Thus, Lemma 1.41 implies that we can replace μ in the bounds for $\Pr[S_n \geq (1+\delta)\mu]$ above by any $\mu \geq \mathrm{E}[X]$, and we can replace μ in the bounds for $\Pr[S_n \leq (1-\delta)\mu]$ by any $\mu \leq \mathrm{E}[X]$, subject to the condition that μ does not violate the restrictions on the random variables in S_n.

Tail estimates for products of random variables

Suppose that we want to bound the probability that the product of independent, non-negative random variables $X_1, \ldots, X_n$ is at most or at least t. In this case we simply set $Y_i = \ln X_i$ for all $i \in [n]$. Then $\Pr[\prod_i X_i \geq t]$ is equivalent to $\Pr[\sum_i Y_i \geq \ln t]$, and $\Pr[\prod_i X_i \leq t]$ is equivalent to $\Pr[\sum_i Y_i \leq \ln t]$. Thus we can apply the tail estimates derived above for sums of independent random variables.

1.3.2 Tail Estimates Based on the Variance

Now we present tail estimates that use the variances of the involved random variables. We start with a result (not far away from the "ideal" bound in Corollary 1.30) that has a simple proof.

Theorem 1.42 ([DP98]) *If* $X_1, \ldots, X_n$ *are independent random variables with* $X_i - \mathrm{E}[X_i] \leq 1$ *and* $\mathrm{V}[\sum_i X_i] = \nu$, *then for any* $0 \leq d \leq 2\nu$,

$$\Pr[S_n \geq \mu + d] \leq \mathrm{e}^{-d^2/4\nu} .$$

Proof. For all $i \in [n]$, let $\mu_i = \mathrm{E}[X_i]$ and $\nu_i = \mathrm{V}[X_i]$. It holds for every $h > 0$ that

$$\begin{aligned}\Pr[S_n \geq \mu + d] &= \Pr[\sum_i (X_i - \mu_i) \geq d] = \Pr[\mathrm{e}^{h\sum_i (X_i - \mu_i)} \geq \mathrm{e}^{h \cdot d}] \\ &\leq \mathrm{E}[\mathrm{e}^{h\sum_i (X_i - \mu_i)}]/\mathrm{e}^{h \cdot d} .\end{aligned}$$

It is easy to check that $e^x \leq 1 + x + x^2$ for every $x \leq 1$ and $e^x \geq 1 + x$ for every $x \in \mathbb{R}$. If $h(X_i - \mu_i) \leq 1$ for every $i \in [n]$, we have

$$\begin{aligned} \mathrm{E}[e^{h\sum_i (X_i-\mu_i)}] &= \prod_i \mathrm{E}[e^{h(X_i-\mu_i)}] \\ &\leq \prod_i \mathrm{E}[1 + h(X_i - \mu_i) + h^2(X_i - \mu_i)^2] \\ &= \prod_i (1 + h^2\nu_i) \leq \prod_i e^{h^2\nu_i} = e^{h^2\nu} . \end{aligned}$$

Thus,

$$\Pr[S_n \geq \mu + d] \leq e^{h^2\nu}/e^{h\cdot d} . \tag{1.30}$$

This term is minimized for $h = h_0$ with $h_0 = d/(2\nu)$. Using this in inequality (1.30) yields the theorem. □

Theorem 1.42 yields better tail estimates than the above theorems if ν is much smaller than μ. (This is, for instance, the case for Example 1 in Section 1.2.7 if $m \gg n$.) Even better bounds are given in the following theorem (recall the definition of $\bar{\nu}$ at the beginning of Section 1.3). Its proof can be found in Section 1.3.4.

Theorem 1.43 ([Hoe63]) *If $X_1, \ldots, X_n$ are independent, and $\mathrm{E}[X_i] = 0$ and $X_i \leq b$ for all $i \in [n]$, then for any $0 < t < b$,*

$$\begin{aligned} \Pr[A_n \geq t] &\leq \left[\left(1 + \frac{bt}{\bar{\nu}}\right)^{-(1+\frac{bt}{\bar{\nu}})\frac{\bar{\nu}}{b^2+\bar{\nu}}} \left(1 - \frac{t}{b}\right)^{-(1-\frac{t}{b})\frac{b^2}{b^2+\bar{\nu}}} \right]^n && (1.31) \\ &\leq e^{-\frac{nt}{b}\left[(1+\frac{\bar{\nu}}{bt})\ln(1+\frac{bt}{\bar{\nu}})-1\right]} . && (1.32) \end{aligned}$$

Note that this theorem can be applied to any collection of upper-bounded random variables $X_1, \ldots, X_n$ by using shifts and scaling. As noticed by Hoeffding [Hoe63], inequality (1.31) is the best that can be obtained from using the exponential moment method.

McDiarmid [McD98] simplified the bounds in Theorem 1.43 to obtain bounds that look very similar to the Chernoff bounds (see Theorem 1.35).

Theorem 1.44 ([McD98]) *If $X_1, \ldots, X_n$ are independent, and $X_i - \mathrm{E}[X_i] \leq b$ for all $i \in [n]$, then for any $t \geq 0$,*

$$\begin{aligned} \Pr[S_n - \mu \geq d] &\leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^{\nu/b^2} \\ &\leq e^{-\frac{d^2}{2\nu(1+\delta/3)}} && (1.33) \end{aligned}$$

where $\delta = b \cdot d/\nu$.

Comparing (1.33) with Corollary 1.30, it becomes clear why Hoeffding's bounds are essentially best possible. What is left is essentially the factor in front of the e-Formula in Corollary 1.30. Talagrand presents bounds that also include an approximation for this factor [Tal95b]. His bounds are the best obtained so far.

Using bounds for the variance

As for the expected value $\mathrm{E}[S_n]$, it may be sometimes not possible to calculate $\mathrm{V}[S_n]$ exactly. In this case it may be more convenient to use an upper bound for $\mathrm{V}[S_n]$. Clearly, for the theorems presented above it holds that the probability bounds are monotonically decreasing with an increasing $\mathrm{V}[S_n]$. Hence, upper bounds cause no problems.

1.3.3 Tail Estimates for Dependent Random Variables

Clearly, all bounds presented above also hold for collections of mutually non-associated random variables, since all of them are based on Bernstein's exponential moment method and $\mathrm{E}[\mathrm{e}^{hX}]$ can be expressed as $\sum_{i=0}^{\infty} h^i \mathrm{E}[X^i]/i!$. In the following we will study the case of k-wise independent or k-wise non-associated random variables and of k-wise uncorrelated random variables. We will also consider some more general types of dependent random variables.

k-wise independent random variables

First, we consider the case that the random variables are k-wise independent. Recall that $\cosh(x) = (\mathrm{e}^x + \mathrm{e}^{-x})/2$. The following theorem is due to Schmitt, Siegel, and Srinivasan.

Theorem 1.45 ([SSS95]) *If $X_1, \ldots, X_n$ are k-wise independent random variables that satisfy $|X_i - \mathrm{E}[X_i]| \leq 1$, then for any even k and any $d > 0$,*

$$\Pr[|S_n - \mu| \geq d] \leq \sqrt{2} \cosh\left(\sqrt{\frac{k^3}{36\nu}}\right) \cdot \left(\frac{k \cdot \nu}{\mathrm{e}\, d^2}\right)^{k/2} .$$

For non-negative random variables, the following theorem was shown by the same authors. Its proof can be found in Section 1.3.4.

Theorem 1.46 ([SSS95]) *If S is the sum of k-wise independent random variables, each of which is confined to the interval $[0, 1]$, then we have:*

(1) For $\delta \leq 1$,

(a) if $k \leq \lfloor \delta^2 \mu \cdot \mathrm{e}^{-1/3} \rfloor$, then $\Pr[|S - \mu| \geq \delta\mu] \leq \mathrm{e}^{-\lfloor k/2 \rfloor}$, and

(b) if $k = \lfloor \delta^2 \mu \cdot \mathrm{e}^{-1/3} \rfloor$, then $\Pr[|S - \mu| \geq \delta\mu] \leq \mathrm{e}^{-\lfloor \delta^2 \mu/3 \rfloor}$.

(2) For $\delta \geq 1$,

(a) if $k \leq \lfloor \delta\mu \cdot \mathrm{e}^{-1/3} \rfloor$, then $\Pr[|S - \mu| \geq \delta\mu] \leq \mathrm{e}^{-\lfloor k/2 \rfloor}$, and

(b) if $k = \lfloor \delta\mu \cdot \mathrm{e}^{-1/3} \rfloor$, then $\Pr[|S - \mu| \geq \delta\mu] \leq \mathrm{e}^{-\lfloor \delta\mu/3 \rfloor}$.

(3) If $\delta \geq 1$ and $k = \lceil \delta\mu \rceil$,

$$\Pr[|S - \mu| \geq \delta\mu] \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu} .$$

k-wise non-associated and k-wise uncorrelated random variables

In the proof of Theorem 1.46 in Section 1.3.4 solely multilinear polynomials of the form

$$P_j(X_1, \dots, X_n) = \sum_{i_1 < i_2 < \dots < i_j} X_{i_1} X_{i_2} \dots X_{i_j}$$

are used. Since their expected values are the same for independent, non-associated, and also uncorrelated random variables, it suffices to have a collection of k-wise uncorrelated random variables to obtain the following result.

Theorem 1.47 *If S is the sum of k-wise uncorrelated random variables, each of which is confined to the interval $[0, 1]$, then we have:*

(1) For $\delta \le 1$,

(a) if $k \le \lfloor \delta^2 \mu \cdot \mathrm{e}^{-1/3} \rfloor$, then $\Pr[|S - \mu| \ge \delta\mu] \le \mathrm{e}^{-\lfloor k/2 \rfloor}$, and

(b) if $k = \lfloor \delta^2 \mu \cdot \mathrm{e}^{-1/3} \rfloor$, then $\Pr[|S - \mu| \ge \delta\mu] \le \mathrm{e}^{-\lfloor \delta^2 \mu/3 \rfloor}$.

(2) For $\delta \ge 1$,

(a) if $k \le \lfloor \delta \mu \cdot \mathrm{e}^{-1/3} \rfloor$, then $\Pr[|S - \mu| \ge \delta\mu] \le \mathrm{e}^{-\lfloor k/2 \rfloor}$, and

(b) if $k = \lfloor \delta \mu \cdot \mathrm{e}^{-1/3} \rfloor$, then $\Pr[|S - \mu| \ge \delta\mu] \le \mathrm{e}^{-\lfloor \delta \mu/3 \rfloor}$.

(3) If $\delta \ge 1$ and $k = \lceil \delta\mu \rceil$,

$$\Pr[|S - \mu| \ge \delta\mu] \le \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}} \right)^{\mu} .$$

In the case of k-wise negatively correlated random variables, the same bounds hold if $\Pr[|S - \mu| \ge \delta\mu]$ is replaced by $\Pr[S - \mu \ge \delta\mu]$. This is due to the fact that the expected values of the multilinear polynomials used in the proof are at most their expected values for the case that the random variables are k-wise uncorrelated.

For the case of k-wise self-weakening random variables with parameter λ (recall the definition in Section 1.2.6), all bounds for $\Pr[S - \mu \ge \delta\mu]$ given in Theorem 1.47 still hold, but with a factor of λ in front of each bound. The reason for this is that the expected values of the multilinear polynomials used in the proof have to be increased by a factor of λ, compared to the case of k-wise uncorrelated random variables.

Other types of dependent random variables

Suppose that T is a random variable that can be written in the form

$$T = p_1 T_1 + p_2 T_2 + \dots + p_N T_N \ ,$$

where each T_i is a sum of independent random variables and $p_1, p_2, \dots, p_N$ are non-negative numbers with $p_1 + p_2 + \dots + p_N = 1$. The random variables T_i need not to be (mutually or even pairwise) independent. Then the following theorem holds

Theorem 1.48 ([Hoe63]) *If T is a random variable with properties defined above, then for every $t \geq 0$ and $h > 0$,*

$$\Pr[T \geq t] \leq \sum_{i=1}^{N} p_i \mathrm{E}[\mathrm{e}^{hT_i}]/\mathrm{e}^{ht} .$$

$\mathrm{E}[\mathrm{e}^{hT_i}]$ may be obtained by using one of the bounds derived in the proofs of the previous theorems.

Proof of Theorem 1.48: For any $h > 0$,

$$\Pr[T \geq t] \leq \mathrm{e}^{-ht}\mathrm{E}[\mathrm{e}^{hT}] . \tag{1.34}$$

Since the exponential function is convex, we have by Jensen's inequality

$$\exp(hT) = \exp\left(h \sum_{i=1}^{N} p_i T_i\right) \leq \sum_{i=1}^{N} p_i \exp(hT_i) .$$

Using this in inequality (1.34) yields the theorem. □

Another technique to model dependent random variables is to use so-called dependency graphs. Consider any collection $X_1, \ldots, X_n$ of random variables. A directed graph $G = ([n], E)$ is called a *dependency graph* of $X_1, \ldots, X_n$, if edges are selected in this graph such that, for all $i \in [n]$, X_i is independent of any set of X_j's with $\{i, j\} \notin E$. Thus, any independent set of nodes in G represents a set of mutually independent random variables. In the following, we present techniques to find good tail estimates in such a situation.

Theorem 1.49 *Let $X_1, \ldots, X_n$ are binary random variables with a dependency graph of degree at most d. Then, for any $\delta > 0$,*

$$\Pr[S \geq (1+\delta)(d+1)\mu] \leq \left(\frac{\mathrm{e}^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu} .$$

Proof. Let G be a dependency graph of $X_1, \ldots, X_n$ of degree at most d. Furthermore, let the random variable S' denote the size of the index set $I \subseteq [n]$ of maximum cardinality with the property that

- I is an independent set in G and
- all X_i with $i \in I$ are equal to 1.

According to the Loéve inequality, for any $\delta > 0$ and $h > 0$,

$$\Pr[S' \geq (1+\delta)\mu] \leq \mathrm{e}^{-h(1+\delta)\mu} \cdot \mathrm{E}[\mathrm{e}^{hS'}] .$$

Now, let $Y_1, \ldots, Y_n$ be independent binary random variables with $\mathrm{E}[Y_i] = \mathrm{E}[X_i]$ for all i and let $T = \sum_{i=1}^{n} Y_i$. Using the identity $\mathrm{e}^x = \sum_{i=0}^{\infty} x^i/i!$, it follows that $\mathrm{E}[\mathrm{e}^{hS'}] \leq \mathrm{E}[\mathrm{e}^{hT}]$, since all products of X_i's in $\mathrm{E}[\mathrm{e}^{hS'}]$ that are not independent have an expected value of 0, whereas in $\mathrm{E}[\mathrm{e}^{hT}]$ the corresponding products have an expected value of at least 0. Hence, using the same technique as in the proof of Theorem 1.35, we obtain for every $\delta > 0$

$$\Pr[S' \geq (1+\delta)\mu] \leq \mathrm{e}^{-h(1+\delta)\mu} \cdot \mathrm{E}[\mathrm{e}^{hT}] \leq \left(\frac{\mathrm{e}^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}$$

if h is set to $\ln(1+\delta)$.

For any $k \in \mathbb{N}_0$ it clearly holds that if $S' = k$ then S can be at most $k(d+1)$, since otherwise there must be at least $k+1$ independent X_i's that are 1. Thus,

$$\Pr[S \geq (1+\delta)(d+1)\mu] \leq \Pr[S' \geq (1+\delta)\mu] ,$$

which concludes the proof. □

If the X_i's are identically distributed, Rabani and Tardos gave a better tail estimate [RT96]. We improve their technique to non-identically distributed random variables.

Theorem 1.50 *If $X_1, \ldots, X_n$ are binary random variables with a dependency graph of degree at most d and $\mathrm{E}[X_i] \leq p$ for all i, then for any $\delta > 0$,*

$$\Pr[S \geq (1+2\delta)\mu + 4\delta p d] \leq 4d \cdot \mathrm{e}^{-\frac{\delta^2\mu}{4d(1+\delta/3)}} .$$

Proof. We will use the following vital fact.

Fact 1.51 *For any sequence of non-negative reals $a_1, \ldots, a_n$ with $A = \sum_{i=1}^n a_i$ and $a_i \leq m$ for all $i \in [n]$ and any $k \in \mathbb{N}$, there are integers $1 \leq i_1 < i_2 < \ldots < i_{k-1} \leq n$ such that, with $i_0 = 1$ and $i_k = n+1$, $\sum_{j=i_\ell}^{i_{\ell+1}-1} a_j \leq A/k + m$ for all $0 \leq \ell \leq k-1$.*

It is not difficult to see that any graph of degree d can be partitioned into at most $d+1$ independent sets. Let $I_1, \ldots, I_{d+1}$ be the corresponding sets in the dependency graph of $X_1, \ldots, X_n$. Fact 1.51 implies that for any $k \in \mathbb{N}$, $I_1, \ldots, I_{d+1}$ can be split into $m \leq k \cdot d + (d+1)$ independent sets $I'_1, \ldots, I'_m$ with the property that

$$\mu_j = \mathrm{E}[\sum_{i \in I'_j} X_i] \leq \frac{\mu}{k \cdot d} + p$$

for all $1 \leq j \leq m$. Let us choose $k = 2$, and let $S_j = \sum_{i \in I'_j} X_i$ for all j. Since I'_j is an independent set, its random variables are independent and therefore, according to the Chernoff bounds,

$$\begin{aligned}
\Pr[S_j \geq \mu_j + \delta(\mu/(2d)+p)] &= \Pr[S_j \geq (1+\delta')\mu_j] \quad \text{with } \delta' = \delta(\mu/(2d)+p)/\mu_j \\
&\leq \mathrm{e}^{-\frac{(\delta')^2\mu_j}{2(1+\delta'/3)}} \\
&\leq \mathrm{e}^{-\frac{\delta^2\mu/(2d)}{2(1+\delta/3)}} .
\end{aligned}$$

The last equality follows from the fact that $\frac{(\delta x/y)^2 y}{2(1+(\delta x/y)/3)} \geq \frac{\delta^2 x}{2(1+\delta/3)}$ for any $\delta, x, y > 0$ with $x \geq y$. Thus we obtain

$$\Pr[S \geq (1+2\delta)\mu + 4\delta p d] \leq \Pr[\bigvee_{j \in [m]} (S_j \geq \mu_j + \delta(\mu/(2d)+p))] \leq 4d \cdot \mathrm{e}^{-\frac{\delta^2\mu/(2d)}{2(1+\delta/3)}} .$$

□

Dependency graphs are very useful as we will see later, but in the case that every random variable weakly depends on every other, Theorems 1.48 and 1.50 might result in far too bad bounds. In such a situation, the following theorem may still help.

Theorem 1.52 *For any k and any set of random variables $X_1, \ldots, X_n$ with $X_i \in [0,1]$ and the property that*

$$P_j(X_1, \ldots, X_n) \leq P_j(p, \ldots, p)$$

for all $j \leq k$, where P_j is defined as in the proof of Theorem 1.46, the same probability bounds hold for $\Pr[S_n \geq (1+\delta)\mu]$ as given in Theorem 1.46 for the case that $X_1, \ldots, X_n$ are k-wise independent random variables with $\mathrm{E}[S_n] = \mu = p \cdot n$.

Despite of its innocent-looking nature, this theorem is extremely powerful. All that has to be shown in order for $P_j(X_1, \ldots, X_n) \leq P_j(p, \ldots, p)$ to hold for all $j \leq k$ is that for all $j \leq k$ we have

$$\mathrm{E}[\textstyle\prod_{i \in S} X_i] \leq p^j$$

for a set S chosen uniformly at random out of the set of all subsets of $[n]$ of size j. In order to demonstrate that Theorem 1.52 is useful, we give the following example.

Example: Consider the problem of embedding a graph $G = (V, E)$ of maximum degree d into a graph $H = (W, F)$ with $|V| = |W| = n$. In particular, we are searching for a way to map the nodes of G one-to-one to the nodes of H and to select paths in H simulating the edges of G such that the congestion and dilation of the path collection is as small as possible.

Let R be the routing number of H (for the definition of the routing number, see Example 2 in Section 1.2.5). To facilitate the selection of the paths, we assume that we are given a system of paths $P = (p_{u,v})_{u,v \in W}$ with dilation at most R and congestion at most $(n-1)R$. That such a system of paths indeed exists can be shown quite easily: take any best possible path collection for the permutation routing problem defined by π_i with $\pi_i(x) = (x+i) \bmod n$ for all $i \in \{1, \ldots, n-1\}$ and $x \in \{0, \ldots, n-1\}$, and define the union of these path collections as the system of paths we intend to use. To find a mapping of G to H, we perform the following random experiment:

Choose a permutation π uniformly at random from the set of all possible permutations on $[n]$. Given π, every node $v \in V$ is mapped to node $\pi(v)$ in W, and every edge $(v, w) \in E$ is represented by the path $p_{\pi(v), \pi(w)}$ in H.

Clearly, the dilation of the paths used for the embedding is at most R. Hence, it remains to bound the congestion of the selected paths. Let e be any fixed edge in H. We start with bounding the expected congestion.

For every pair of nodes $u, v \in W$ with $u \neq v$, let the binary random variable $X_{u,v}$ be 1 if and only if e is contained in $p_{u,v}$ and $(\pi^{-1}(u), \pi^{-1}(v)) \in E$. Furthermore, let $X = \sum_{u,v \in W} X_{u,v}$. Thus X denotes the congestion caused at e by the selected paths. Since π is chosen uniformly at random and the degree of G is d, the probability that for some given path $p_{u,v}$ we have $(\pi^{-1}(u), \pi^{-1}(v)) \in E$ is at most

$$\frac{d \cdot n}{n(n-1)} = \frac{d}{n-1} \, .$$

Since at most $(n-1)R$ paths are crossing e,

$$\mathrm{E}[X] \leq (n-1)R \cdot \frac{d}{n-1} = d \cdot R \, .$$

Next we show that w.h.p. the congestion at e is not far away from $\mathrm{E}[X]$. This is trivial for $d > n/2$. Hence, in the following we assume that $d \leq n/2$. Since there may be high dependencies among the $X_{u,v}$, this cannot be done by simply using Chernoff bounds. We first need the following claim.

Claim 1.53 *Consider any fixed $(u,v) \in W^2$ and any set of pairs $(u_1,v_2),\ldots,(u_k,v_k) \in W^2$ with $k \geq 1$. If it holds that either u or v is a node that has not been used in any pair (u_i,v_i), then*

$$\Pr[X_{u,v} = 1 \mid X_{u_1,v_1} = \ldots = X_{u_k,v_k} = 1] \leq \frac{d}{n-2k-1} .$$

Proof. Let m be the number of distinct nodes in the pairs (u_i,v_i), $i \in [k]$. First we consider the case that u is (and therefore v is not) identical to some node u_i or v_i, $i \in [k]$ (or vice versa). In this case, there are $n-m$ nodes of G that can be embedded in v, and at most d of them are edges in G. Since a mapping of G to H is chosen uniformly at random, we therefore have

$$\Pr[X_{u,v} = 1 \mid X_{u_1,v_1} = \ldots = X_{u_k,v_k} = 1] \leq \frac{d}{n-m} \leq \frac{d}{n-2k} .$$

For the case that u and v are not equal to any node u_i or v_i, $i \in [k]$, the number of pairs of nodes in G that can be embedded in (u,v) is equal to $(n-m)(n-m-1)$. Moreover, the number of remaining edges in G that can be represented by $p_{u,v}$ is at most $d(n-m)$. Thus we obtain

$$\begin{aligned} \Pr[X_{u,v} = 1 \mid X_{u_1,v_1} = \ldots = X_{u_k,v_k} = 1] &\leq \frac{d(n-m)}{(n-m)(n-m-1)} \\ &= \frac{d}{n-m-1} \leq \frac{d}{n-2k-1} , \end{aligned}$$

which concludes the proof. □

Now, consider the random experiment of choosing a set $\{(u_1,v_1),\ldots,(u_k,v_k)\}$ of pairs of distinct nodes uniformly at random among all sets of pairs of distinct nodes of size k. Since the probability that both u_i and v_i are already used in the pairs $(u_1,v_1),\ldots,(u_{i-1},v_{i-1})$ is at most

$$\frac{(2(i-1))(2(i-1)-1)-(i-1)}{n(n-1)-(i-1)} \leq \frac{4k^2}{n^2}$$

if k is sufficiently small compared to n, it follows from Claim 1.53 that

$$\Pr[X_{u_i,v_i} = 1 \mid X_{u_1,v_1} = \ldots = X_{u_{i-1},v_{i-1}} = 1] \leq \frac{4k^2}{n^2} + \frac{d}{n-2k-1} ,$$

where the probability is taken according to the random choice of (u_i,v_i) and the random mapping of G to H. Thus,

$$\Pr[X_{u_1,v_1} = \ldots = X_{u_k,v_k} = 1] \leq \left(\frac{4k^2}{n^2} + \frac{d}{n-2k-1}\right)^k$$

In the notation of Theorem 1.52, this implies that

$$P_k(\{X_{u,v} \ : \ u \neq v\}) \leq P_k(p_k,\ldots,p_k) \quad \text{where} \quad p_k = \frac{4k^2}{n^2} + \frac{d}{n-2k-1} .$$

For $k = (\sqrt{n}-1)/2$ we have $p_j \leq (d+1)/(n-\sqrt{n})$ for all $j \leq k$. Hence,

$$P_j(\{X_{u,v} \ : \ u \neq v\}) \leq P_j(p,\ldots,p) \quad \text{with} \quad p = \frac{d+1}{n-\sqrt{n}} \tag{1.35}$$

for all $j \leq k$. Let $\mu = p \cdot (n-1)R$. Using (1.35) in Theorem 1.52 yields that, for any $\delta > 0$,

$$\Pr[S \geq (1+\delta)\mu] \leq \mathrm{e}^{-\min[\delta^2\mu/3,\ \delta\mu/3,\ k/2]} .$$

Hence, the probability can be made as low as $1/n^3$ that the congestion at edge e exceeds $(d+1)R + O(\max[\sqrt{(d+1)R\log n},\ \log n])$. Since H can have at most n^2 edges, the probability that the congestion of the embedding exceeds the same bound can be made less than $1/n$.

We note that Valiant's trick (see Example 2 in Section 1.2.5) would have enabled us to use a much simpler proof to obtain a congestion of $O(d \cdot R + \log n)$, w.h.p, for the embedding. However, using this trick, it is at best possible to show a bound of $2d \cdot R$, whereas the proof above shows that a bound of approximately $(d+1)R$ can be reached.

1.3.4 Proofs of the Theorems in Section 1.3

Proof of Theorem 1.35: It remains to prove inequality (1.21). For $0 \leq \delta \leq 1$ and $h > 0$,

$$\begin{aligned}\Pr[S_n \leq (1-\delta)\mu] &= \Pr[-S_n \geq -(1-\delta)\mu] = \Pr[\mathrm{e}^{-hS_n} \geq \mathrm{e}^{-h(1-\delta)\mu}] \\ &\overset{\text{Markov}}{\leq} \mathrm{e}^{h(1-\delta)\mu} \cdot \mathrm{E}[\mathrm{e}^{-h\cdot S_n}] \end{aligned} \tag{1.36}$$

and

$$\begin{aligned}\mathrm{E}[\mathrm{e}^{-h\cdot S_n}] &= \mathrm{E}[\mathrm{e}^{-h(X_1+\ldots+X_n)}] = \mathrm{E}[\mathrm{e}^{-h\cdot X_1}\cdots \mathrm{e}^{-h\cdot X_n}] = \textstyle\prod_{i=1}^n \mathrm{E}[\mathrm{e}^{-h\cdot X_i}] \\ &= \prod_{i=1}^n (p_i\mathrm{e}^{-h} + (1-p_i)) = \prod_{i=1}^n (1 - p_i(1-\mathrm{e}^{-h})) \leq \prod_{i=1}^n \mathrm{e}^{-p_i(1-\mathrm{e}^{-h})} \\ &= \mathrm{e}^{-\mu(1-\mathrm{e}^{-h})} .\end{aligned}$$

This yields together with inequality (1.36) that

$$\Pr[S_n \leq (1-\delta)\mu] \leq \mathrm{e}^{h(1-\delta)\mu}\mathrm{e}^{-\mu(1-\mathrm{e}^{-h})} = \mathrm{e}^{-(1-h(1-\delta)-\mathrm{e}^{-h})\mu} . \tag{1.37}$$

The right-hand side of (1.37) attains its minimum at $h = h_1$, where $h_1 = -\ln(1-\delta)$. Inserting this in (1.37) yields

$$\Pr[S_n \leq (1-\delta)\mu] \leq (1-\delta)^{-(1-\delta)\mu} \cdot \mathrm{e}^{-\delta\cdot\mu} = \left(\frac{\mathrm{e}^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu} ,$$

which completes the proof of Theorem 1.35. □

Proof of Theorem 1.36: We will only prove inequality (1.25). The proofs of the other inequalities can be found in [Hoe63].

Let X be a random variable such that $a \leq X \leq b$. Since the exponential function $\exp(hX)$ is convex, its graph is bounded above on the interval $[a, b]$ by the straight line which connects its ordinates at $X = a$ and $X = b$. Thus,

$$\mathrm{e}^{hX} \leq \frac{b-X}{b-a} \cdot \mathrm{e}^{h\cdot a} + \frac{X-a}{b-a} \cdot \mathrm{e}^{h\cdot b} .$$

Hence, we obtain

$$\mathrm{E}[\mathrm{e}^{hX}] \le \frac{b - \mathrm{E}[X]}{b-a} \cdot \mathrm{e}^{h\cdot a} + \frac{\mathrm{E}[X] - a}{b-a} \cdot \mathrm{e}^{h\cdot b} \ . \tag{1.38}$$

Similar to the proof of Theorem 1.35 it holds for any $h > 0$ that

$$\Pr[A_n \ge t] \le \mathrm{e}^{-h\cdot n\cdot t} \cdot \prod_{i=1}^{n} \mathrm{E}[\mathrm{e}^{hX_i}] \ . \tag{1.39}$$

By assumption, $0 \le X_i \le 1$. Let $\mu_i = \mathrm{E}[X_i]$. Using inequality (1.38), we have

$$\prod_{i=1}^{n} \mathrm{E}[\mathrm{e}^{hX_i}] \le \prod_{i=1}^{n} (1 - \mu_i + \mu_i \mathrm{e}^h) \ .$$

Since the geometric mean does not exceed the arithmetic mean (see Fact 1.4),

$$\left[\prod_{i=1}^{n} (1 - \mu_i + \mu_i \mathrm{e}^h)\right]^{1/n} \le \frac{1}{n} \sum_{i=1}^{n} (1 - \mu_i + \mu_i \mathrm{e}^h) = 1 - \bar{\mu} + \bar{\mu}\mathrm{e}^h \ . \tag{1.40}$$

It follows from (1.39) that

$$\Pr[A_n \ge t] \le \left[\mathrm{e}^{-h\cdot t}(1 - \bar{\mu} + \bar{\mu}\mathrm{e}^h)\right]^n \ . \tag{1.41}$$

The right hand side of (1.41) attains its minimum at $h = h_0$, where

$$h_0 = \ln \frac{(1-\bar{\mu})t}{(1-t)\bar{\mu}} \ .$$

Since $\bar{\mu} < t < 1$, h_0 is positive. Inserting h_0 in (1.41) yields inequality (1.25). □

Proof of Theorem 1.38: We start with proving inequality (1.28). According to the Loéve inequality it holds that

$$\Pr[S_n \ge (1+\delta)\mu] \le \mathrm{e}^{-h(1+\delta)\mu} \cdot \mathrm{E}[\mathrm{e}^{h\cdot S}] \ . \tag{1.42}$$

Since $X_1, \ldots, X_n$ are independent, it follows that

$$\begin{aligned} \mathrm{E}[\mathrm{e}^{h\cdot S}] &= \mathrm{E}[\mathrm{e}^{h\cdot X_1} \cdots \mathrm{e}^{h\cdot X_n}] = \prod_{i=1}^{n} \mathrm{E}[\mathrm{e}^{h\cdot X_i}] = \prod_{i=1}^{n} \left(\sum_{j\ge 1} (1-p_i)^{j-1} p_i \cdot \mathrm{e}^{h\cdot j} \right) \\ &= \prod_{i=1}^{n} \left(p_i \cdot \mathrm{e}^h \sum_{j \ge 0} ((1-p_i)\cdot \mathrm{e}^h)^j \right) \ . \end{aligned}$$

In order for the right-hand side to be finite we require that $\mathrm{e}^h < 1/(1-p_i)$ for all $i \in [n]$, or equivalently, $\mathrm{e}^h < 1/(1-p)$. In this case, we obtain for $\mu_i = \mathrm{E}[X_i]$ for all $i \in [n]$

$$\mathrm{E}[\mathrm{e}^{h\cdot S}] = \prod_{i=1}^{n} \left(p_i \mathrm{e}^h \cdot \frac{1}{1-(1-p_i)\mathrm{e}^h} \right) = \prod_{i=1}^{n} \frac{1}{1 - \mu_i + \mu_i \mathrm{e}^{-h}} \ .$$

Thus together with (1.42) we have

$$\Pr[S \geq (1+\delta)\mu] \leq \prod_{i=1}^{n} \left(\frac{\mathrm{e}^{-h(1+\delta)\mu_i}}{1-\mu_i+\mu_i \mathrm{e}^{-h}} \right) .$$

For $\mu_i = 1/p$, the fraction $\mathrm{e}^{-h(1+\delta)\mu_i}/(1-\mu_i+\mu_i\mathrm{e}^{-h})$ attains its minimum for $h = h_0 = \ln(1 - \delta p/(1+\delta))^{-1}$, which fulfills the condition that $\mathrm{e}^h < 1/(1-p)$ for all $\delta > 0$. For $h = h_0$ we have

$$\begin{aligned} \frac{\mathrm{e}^{-h(1+\delta)\mu_i}}{1-\mu_i+\mu_i\mathrm{e}^{-h}} &= \left(1 - \frac{\delta p}{1+\delta}\right)^{(1+\delta)\mu_i} \cdot \frac{1}{1-\mu_i \cdot \frac{\delta p}{1+\delta}} \\ &\leq \mathrm{e}^{-\delta p \mu_i} \cdot \left(1 + \frac{\delta p \mu_i}{1+\delta(1-p\mu_i)}\right) \leq (1+\delta p \mu_i) \cdot \mathrm{e}^{-\delta p \mu_i} . \end{aligned}$$

For any vector $\boldsymbol{x} = (x_1, \ldots, x_2) \in \mathrm{IR}_+^n$, let the function $f : \mathrm{IR}_+^n \to \mathrm{IR}_+$ be defined as $f(\boldsymbol{x}) = \prod_{i=1}^n (1+x_i)$. It is not difficult to show that f is Schur-concave. Hence,

$$\begin{aligned} \Pr[S \geq (1+\delta)\mu] &\leq \prod_{i=1}^{n} \left((1+\delta p \mu_i)\mathrm{e}^{-\delta p \mu_i}\right) \\ &\leq \left(\frac{\sum_i (1+\delta p \mu_i)}{n}\right)^n \mathrm{e}^{-\sum_i \delta p \mu_i} = \left(1 + \frac{\delta \mu p}{n}\right)^n \cdot \mathrm{e}^{-\delta \mu p} . \end{aligned}$$

Next we prove inequality (1.29). Using the Markov inequality (see the proof of Theorem 1.35), it holds that

$$\Pr[S \leq (1-\delta)\mu] \leq \mathrm{e}^{h(1-\delta)\mu} \cdot \mathrm{E}[\mathrm{e}^{-h \cdot S}] . \tag{1.43}$$

Since $X_1, \ldots, X_n$ are independent, it follows that

$$\begin{aligned} \mathrm{E}[\mathrm{e}^{-h\cdot S}] &= \prod_{i=1}^{n} \mathrm{E}[\mathrm{e}^{-h \cdot X_i}] = \textstyle\prod_{i=1}^n \left[\sum_{j\geq 1} (1-p_i)^{j-1} p_i \mathrm{e}^{-hj}\right] \\ &= \prod_{i=1}^{n} \left[\frac{p_i}{\mathrm{e}^h} \sum_{j \geq 0} \left(\frac{1-p_i}{\mathrm{e}^h}\right)^j\right] = \prod_{i=1}^{n} \frac{p_i}{\mathrm{e}^h} \cdot \frac{1}{1-(1-p_i)/\mathrm{e}^h} \end{aligned}$$

for all $h \geq 0$. In this case,

$$\mathrm{E}[\mathrm{e}^{-h \cdot S}] = \prod_{i=1}^{n} \frac{1}{\mathrm{e}^h/p_i + 1 - 1/p_i} = \prod_{i=1}^{n} \frac{1}{1-\mu_i+\mu_i\mathrm{e}^h} . \tag{1.44}$$

Inserting this in (1.43) yields

$$\Pr[S \leq (1-\delta)\mu] \leq \prod_{i=1}^{n} \left(\frac{\mathrm{e}^{h(1-\delta)\mu_i}}{1-\mu_i+\mu_i\mathrm{e}^h}\right) \tag{1.45}$$

The fraction on the right-hand side of (1.45) is minimal for $h = h_1$, where

$$h_1 = \ln\left(1 + \frac{\delta}{\mu_i(1-\delta)}\right) .$$

We choose $h = \ln(1 + \delta p/(1-\delta))$. For this we have

$$\Pr[S \le (1-\delta)\mu] \le \prod_{i=1}^{n} \left(\frac{(1+\delta p/(1-\delta))^{(1-\delta)\mu_i}}{1+\delta p \mu_i/(1-\delta)} \right) .$$

Since $p\mu_i \le 1$ it holds that

$$\left(1 + \frac{\delta p}{1-\delta}\right)^{(1-\delta)\mu_i} \le \mathrm{e}^{\delta p \mu_i} \le 1 + \delta p \mu_i \sum_{i \ge 0} \tfrac{\delta^i}{(i+1)!} .$$

Thus,

$$\begin{aligned} \frac{(1+\delta p/(1-\delta))^{(1-\delta)\mu_i}}{1+\delta p \mu_i/(1-\delta)} &\le \frac{1 + \delta p \mu_i \sum_{i\ge 0} \frac{\delta^i}{(i+1)!}}{1 + \delta p \mu_i \sum_{i \ge 0} \delta^i} \\ &\le 1 - (1-\delta)\delta p \mu_i \Big[\Big(\textstyle\sum_{i\ge 0} \delta^i\Big) - \Big(\sum_{i \ge 0} \frac{\delta^i}{(i+1)!}\Big)\Big] , \end{aligned}$$

since

$$\begin{aligned} & 1 + \delta p \mu_i \sum_{i \ge 0} \frac{\delta^i}{(i+1)!} \\ \le\ & 1 + \delta p \mu_i \sum_{i\ge 0} \delta^i - (1-\delta) \left[\left(1 + \delta p \mu_i \sum_{i \ge 0} \delta^i \right) - \left(1 + \delta p \mu_i \sum_{i \ge 0} \frac{\delta^i}{(i+1)!} \right) \right] \left(1 + \delta \sum_{i \ge 0} \delta^i \right) \\ \le\ & 1 + \delta p \mu_i \sum_{i\ge 0} \delta^i - (1-\delta) \left[\delta p \mu_i \sum_{i \ge 0} \delta^i - \delta p \mu_i \sum_{i \ge 0} \frac{\delta^i}{(i+1)!} \right] \left(1 + \delta p \mu_i \sum_{i \ge 0} \delta^i \right) \\ =\ & \left(1 - (1-\delta)\delta p \mu_i \left[\sum_{i \ge 0} \delta^i - \sum_{i\ge 0} \frac{\delta^i}{(i+1)!} \right] \right) \left(1 + \delta p \mu_i \sum_{i \ge 0} \delta^i \right) . \end{aligned}$$

Hence,

$$\begin{aligned} \frac{(1+\delta p/(1-\delta))^{(1-\delta)\mu_i}}{1+\delta p \mu_i/(1-\delta)} &\le \mathrm{e}^{-(1-\delta)\delta p \mu_i \left[\left(\sum_{i\ge 0} \delta^i\right) - \left(\sum_{i\ge 0} \frac{\delta^i}{(i+1)!}\right)\right]} \\ &\le \mathrm{e}^{-(1-\delta)\delta p \mu_i \cdot \frac{1}{2} \sum_{i \ge 1} \delta^i} = \mathrm{e}^{-\delta^2 p \cdot \mu_i / 2} . \end{aligned}$$

Inserting this in (1.45) yields

$$\Pr[S \le (1-\delta)\mu] \le \mathrm{e}^{-\delta^2 p \mu/2} .$$

□

Proof of Theorem 1.43: We only show inequality (1.31). For this we need two lemmas.

Lemma 1.54 *If X is a random variable such that $\mathrm{E}[X] = 0$, $\mathrm{V}[X] = \nu$ and $X \le b$, then for any positive number h*

$$\mathrm{E}[\mathrm{e}^{hX}] \le \frac{b^2}{b^2 + \nu} \cdot \mathrm{e}^{-(\nu/b)h} + \frac{\nu}{b^2+\nu} \cdot \mathrm{e}^{bh} .$$

A proof of this lemma can be found in Bennett [Ben62].

Lemma 1.55 *If $c > 0$, the function*

$$f(x) = \ln\left(\frac{1}{1+x}\mathrm{e}^{-cx} + \frac{x}{1+x}\mathrm{e}^{c}\right)$$

has a negative second derivative for $x \geq 0$.

For a proof of this lemma see [Hoe63]. By assumption, $\mathrm{E}[X_i] = 0$ and $X_i \leq b$. Let $\nu_i = \mathrm{V}[X_i]$. From Lemma 1.54 it follows that

$$\begin{aligned}\Pr[A_n \geq t] &\leq \mathrm{e}^{-h\cdot n\cdot t}\prod_{i=1}^{n}\left(\frac{b^2}{b^2+\nu_i}\cdot\mathrm{e}^{-(\nu_i/b)h} + \frac{\nu_i}{b^2+\nu_i}\cdot\mathrm{e}^{bh}\right)\\ &= \mathrm{e}^{-h\cdot n\cdot t+\sum_{i=1}^{n} f(\nu_i/b^2)}\ ,\end{aligned}$$

where f is the function defined in Lemma 1.55 with $c = bh$. Since, by Lemma 1.55, $f(x)$ has a negative second derivative, $-f(x)$ is convex for $x \geq 0$. Therefore by Jensen's inequality

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} f(\nu_i/b^2) &\leq f\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\nu_i}{b^2}\right) = f(\bar{\nu}/b^2)\\ &= \ln\left(\frac{b^2}{b^2+\bar{\nu}}\mathrm{e}^{-(\bar{\nu}/b)h} + \frac{\bar{\nu}}{b^2+\bar{\nu}}\mathrm{e}^{bh}\right)\ .\end{aligned}$$

It follows that

$$\Pr[A_n \geq t] \leq \left(\frac{b^2}{b^2+\bar{\nu}}\mathrm{e}^{-(t+\bar{\nu}/b)h} + \frac{\bar{\nu}}{b^2+\bar{\nu}}\mathrm{e}^{(b-t)h}\right)^{n}\ . \tag{1.46}$$

The right-hand side of this inequality attains it minimum at $h = h_0$, where

$$h_0 = \frac{b}{b^2+\bar{\nu}}\ln\frac{1+tb/\bar{\nu}}{1-t/b}\ .$$

Inserting this in (1.46), we obtain inequality (1.31) of Theorem 1.43. □

Proof of Theorem 1.46: We will only prove the case when $\delta \geq 1$ and $k = \lceil\delta\mu\rceil$. First we assume that we have n binary random variables $X_1, \ldots, X_n$. Instead of using functions of the form e^{hS} to achieve good probability bounds, we intend to use a family of symmetric multilinear polynomials $P_j(\boldsymbol{z})$ with $j \in [n]$ and $\boldsymbol{z} = (z_1, \ldots, z_n) \in \mathrm{I\!R}^n$, where

$$P_j(\boldsymbol{z}) = \sum_{i_1<i_2<\ldots<i_j} z_{i_1}z_{i_2}\cdots z_{i_j}\ .$$

If $X_1, \ldots, X_n$ are independent, then $\mathrm{E}[P_j(X_1, \ldots, X_n)]$ is explicitly available:

$$\mathrm{E}[P_j(X_1, \ldots, X_n)] = P_j(p_1, \ldots, p_n)\ ,$$

where $p_i = \Pr[X_i = 1]$.

First, we will demonstrate that using multilinear polynomials is a very general approach that includes the exponential moment method. Clearly, for any sequence of variables $z_1, \ldots, z_n$ in $\{0, 1\}$ and any positive integer j,

$$(z_1 + z_2 + \ldots + z_n)^j = \sum_{i=1}^{\min[j,n]} a_i P_i(z_1, \ldots, z_n) \tag{1.47}$$

for some non-negative reals $a_1, \ldots, a_{\min[j,n]}$ with $\sum_i a_i \binom{n}{i} = n^j$. Let $S = \sum_{i=1}^n X_i$. Due to the fact that the X_i's only take values in $\{0, 1\}$ and the a_i's are limited by n^j, the expansion $e^{hS} = \sum_{j=1}^{\infty} (h^j/j!) \cdot S^j$ is finite. Since all of the coefficients $(h^j/j!)$ are positive if $h > 0$, we obtain the following lemma.

Lemma 1.56 *Let $z_1, \ldots, z_n \in \{0, 1\}$. Then, for any $h > 0$, there are non-negative reals $a_1, \ldots, a_n$ such that*

$$\mathrm{e}^{h(z_1+z_2+\ldots+z_n)} = \sum_{i=0}^{n} a_i P_i(z_1, \ldots, z_n) \ .$$

Hence, the family of functions $\{\mathrm{e}^{hX} : \ h > 0\}$ is just a special case of the family of polynomials we use here.

Now we show how to obtain tail estimates using multilinear polynomials. For any $\boldsymbol{y} = (y_0, y_1, \ldots, y_n) \in \mathbb{R}_+^{n+1}$ and $\boldsymbol{z} = (z_1, \ldots, z_n) \in \mathbb{R}^n$, define $f_{\boldsymbol{y}}(\boldsymbol{z}) = \sum_{i=0}^n y_i P_i(\boldsymbol{z})$. With this notation, Lemma 1.56 can be restated as $\forall h \ \exists \boldsymbol{y}$: $f_{\boldsymbol{y}}(X_1, \ldots, X_n) = \mathrm{e}^{h \cdot S}$. Clearly, for any non-negative integer m, $S = m$ if and only if

$$f_{\boldsymbol{y}}(X_1, \ldots, X_n) = \sum_{i=0}^{m} y_i \binom{m}{i} \tag{1.48}$$

and hence,

$$\Pr[S \geq t] \ = \ \Pr\left[f_{\boldsymbol{y}}(X_1, \ldots, X_n) \geq \sum_{i=0}^{t} y_i \binom{t}{i} \right] \leq \frac{\mathrm{E}[f_{\boldsymbol{y}}(X_1, \ldots, X_n)]}{\sum_{i=0}^t y_i \binom{t}{i}} \ . \tag{1.49}$$

Our goal is to find a $\boldsymbol{y}$ for which the right-hand side of this formula is minimal. We will restrict our attention to those $\boldsymbol{y}$ with $y_{t+1} = \ldots = y_n = 0$. To find a best possible assignment for the rest, we state the following lemma. For a proof see [SSS95]. (It essentially has to be shown that $P_i(.)$ is Schur-concave.)

Lemma 1.57 *For any $i > 0$ and $s \geq 0$, $P_i(z_1, \ldots, z_n)$ is maximized by setting $z_1 = \ldots = z_n = s/n$, when subject to the constraints that $z_1, \ldots, z_n \in \mathbb{R}_+$ and $\sum_{i=1}^n z_i = s$.*

Inequality (1.49) and Lemma 1.57 imply that if $p = (\sum_{i=1}^n p_i)/n = \mu/n$, then for any $y \in \mathbb{R}_+^{n+1}$,

$$\Pr[X \geq t] \leq \frac{\sum_{i=0}^t y_i \binom{n}{i} p^i}{\sum_{i=0}^t y_i \binom{t}{i}} \ .$$

This is minimized by setting $y_i = 1$ if $i = i^*$ and 0 otherwise, where i^* is the integer at which $\binom{n}{i}p^i/\binom{t}{i}$ is minimized. Since

$$\frac{\binom{n}{i+1}p^{i+1}/\binom{t}{i+1}}{\binom{n}{i}p^i/\binom{t}{i}} = \frac{(n-i)p}{t-i} \ ,$$

which is less than, equal to, or greater than 1 depending on if i is less than, equal to, or greater than $(t-np)/(1-\mu/n)$, $\binom{n}{i}p^i/\binom{t}{i}$ is minimized at

$$i^* = \left\lceil \frac{t-np}{1-\mu/n} \right\rceil = \left\lceil \frac{\delta\mu}{1-\mu/n} \right\rceil$$

for $t = (1+\delta)\mu$. Since $i^* > \lceil \delta\mu \rceil$ and $X_1, \ldots, X_n$ are only $\lceil \delta\mu \rceil$-wise independent, we choose $i^* = \lceil \delta\mu \rceil$. In this case,

$$\begin{aligned}
\Pr[S \geq t] &\leq \binom{n}{i^*}p^{i^*}/\binom{t}{i^*} \\
&\leq \left(\frac{n}{t}\right)^{i^*}\left(\frac{\mu}{n}\right)^{i^*}\left(\frac{n}{n-i^*}\right)^{n-i^*}\left(\frac{t-i^*}{t}\right)^{t-i^*} \\
&\leq \left(\frac{n}{(1+\delta)\mu}\right)^{\delta\mu}\left(\frac{\mu}{n}\right)^{\delta\mu}\left(\frac{n}{n-\delta\mu}\right)^{n-\delta\mu}\left(\frac{\mu}{(1+\delta)\mu}\right)^{\mu} \\
&= \frac{(n/n-\delta\mu)^{n-\delta\mu}}{(1+\delta)^{(1+\delta)\mu}} = \left(\frac{1}{1+\delta}\right)^{(1+\delta)\mu}\left(1+\frac{\delta\mu}{n-\delta\mu}\right)^{n-\delta\mu} \\
&\leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu} \ .
\end{aligned}$$

So far, we have assumed that the X_i are binary random variables. For the case that X_i can take any real value in $[0,1]$, we need the following lemma.

Lemma 1.58 *Let z_i be real numbers with $0 \leq z_i \leq 1$ for all $i \in [n]$ and suppose that $m \geq 0$, $j \leq m$ and $\sum_{i=1}^n z_i \geq m$. Then,*

$$P_j(z_1, \ldots, z_n) \geq \binom{m}{j} \ .$$

Using this lemma in (1.48) yields the theorem. □

1.4 The Method of Bounded Differences

In this section, we present methods of finding tail estimates for more general functions than just sums of random variables.

1.4.1 Independent Bounded Differences Inequalities

First, we present tail estimates that are based on bounds on the difference of functions over independent random variables. The following result is due to McDiarmid [McD89].

Theorem 1.59 ([McD89]) *Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a vector of independent random variables with X_i taking values in a set A_i for each i. Suppose that the real-valued function f defined on $A_1 \times \ldots \times A_n$ satisfies*

$$|f(\boldsymbol{x}) - f(\boldsymbol{x}')| \le c_i$$

whenever the vectors $\boldsymbol{x}$ and $\boldsymbol{x}'$ differ only in the ith coordinate. Let μ be the expected value of $f(\boldsymbol{X})$. Then for any $d \ge 0$,

$$\Pr[f(\boldsymbol{X}) - \mu \ge d] \le \mathrm{e}^{-2d^2/\sum_{i=1}^n c_i^2}$$

and

$$\Pr[f(\boldsymbol{X}) - \mu \le -d] \le \mathrm{e}^{-2d^2/\sum_{i=1}^n c_i^2} .$$

If we assume the X_i's to be binary random variables and $f(\boldsymbol{X}) = \sum_{i=1}^n X_i$, then we obtain

$$\Pr[f(\boldsymbol{X}) - \mu \ge d] \le \mathrm{e}^{-2d^2/n} . \tag{1.50}$$

The problem with this result is that it only provides reasonable probability bounds for $d \ge \sqrt{n}$. To overcome this problem, we introduce the following notation.

Given a random variable X, its *median* $\mathrm{M}[X]$ is defined as a number with the property that

$$\Pr[X \le \mathrm{M}[X]] \ge \frac{1}{2} \quad \text{and} \quad \Pr[X \ge \mathrm{M}[X]] \ge \frac{1}{2} .$$

Using the median of a random variable, Talagrand proved the following theorem (see [Tal95a, Mol98]).

Theorem 1.60 ([Tal95a]) *Let X be a random variable determined by n independent trials $T_1, \ldots, T_n$ and satisfying*

(1) changing the outcome of any one trial can affect X by at most c, and

(2) for any s, if $X \ge s$ then there are s trials $T_{i_1}, \ldots, T_{i_s}$ whose outcomes certify that $X \ge s$,

then for any $0 < d \le \mathrm{M}[X]$,

$$\Pr[|X - \mathrm{M}[X]| > d] \le 2\mathrm{e}^{-\frac{d^2}{16c^2\mathrm{M}[X]}} .$$

The fact that Talagrand's inequality proves concentration around the median rather than the expected value is not a serious problem, as in the situation where Talagrand's inequality applies, those two values are very close together, and so concentration around one implies concentration around the other:

Fact 1.61 *Under the conditions of Talagrand's inequality,*

$$|\mathrm{E}[X] - \mathrm{M}[X]| \le 3c\sqrt{\mathrm{E}[X]} .$$

Using this fact and Talagrand's inequality in Theorem 1.59, we obtain the following result (see also [Mol98]).

Theorem 1.62 *Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a vector of independent random variables with X_i taking values in a set A_i for each i. Suppose that the real-valued function f defined on $A_1 \times \ldots \times A_n$ satisfies*

(1) changing the outcome of any one trial can affect f by at most c, and

(2) for any s, if $f(\boldsymbol{X}) \geq s$ then there are s variables $X_{i_1}, \ldots, X_{i_s}$ whose outcomes certify this,

then for any $0 < d \leq \mu$ with $\mu = \mathrm{E}[f(\boldsymbol{X})]$,

$$\Pr[|f(\boldsymbol{X}) - \mu| > d + 3c\sqrt{\mu}] \leq 2\mathrm{e}^{-\frac{d^2}{16c^2\mu}} .$$

Example: We will give a simple example of how to use McDiarmid's bounded differences method.

Claim 1.63 *Let n balls be thrown randomly into n bins. With high probability, at most $n/2^c$ balls are in bins with load at least c, for $8 \leq c \leq \log n/4$.*

Proof. We define the binary random variables X_i, $1 \leq i \leq n$, to be 1 if and only if ball i resides in a bin with a load of at least c. Obviously, it holds

$$\begin{aligned} \Pr[X_i = 1] &\leq \binom{n-1}{c-1}\left(\frac{1}{n}\right)^{c-1} \\ &\leq \left(\frac{\mathrm{e}}{c-1}\right)^{c-1} \leq \left(\frac{1}{2}\right)^{c+1} \quad \text{for } c \geq 8. \end{aligned}$$

Thus, for $X = \sum_{i=1}^n X_i$ we obtain $\mathrm{E}[X] \leq n/2^{c+1}$. Now we use Theorem 1.59 to bound the deviation from the expected value. Let the random variable Y_j, $1 \leq j \leq n$, denote the random destination chosen by ball j. Then we can view X, the number of balls that reside in a bin of load at least c, as a function f in the Y_j's, i.e.,

$$X = f(\{Y_j,\ 1 \leq j \leq n\}) .$$

If the value of Y_j changes for any j, then X changes by at most c. Therefore, by Theorem 1.59, we obtain for $c \leq \log n/4$,

$$\Pr[X \geq \mathrm{E}[X] + \tfrac{n}{2^{c+1}}] \leq \mathrm{e}^{-\frac{2(n/2^c)^2}{nc^2}} \leq \mathrm{e}^{-\frac{2n^2}{\sqrt{n}\cdot n\cdot(\log n/4)^2}} = \mathrm{e}^{-32\sqrt{n}/\log^2 n} .$$

□

1.4.2 Martingales and the Method of Bounded Differences

Martingales are a well-studied concept in classical probability theory. Here we will define them in a discrete setting in the simplest form, which is sufficient for our purposes.

Definition 1.64 *A* martingale *is a sequence of random variables $X_0, X_1, \ldots$ such that*

$$\mathrm{E}[X_i \mid X_0, X_1, \ldots, X_{i-1}] = X_{i-1}$$

for all $i \geq 1$.

Let us give a simple example of a martingale. Suppose we are given a sum of independent random variables $Y = Y_1 + \ldots + Y_n$. Let $X_0 = 0$ and the random variables X_i be equal to $\sum_{j=1}^{i}(Y_i - \mathrm{E}[Y_i])$ for all $1 \le i \le n$. We will show that the variables X_i, $i \ge 0$, define a martingale. First, observe that they satisfy the so-called Markov property, $\mathrm{E}[X_n \mid X_0, X_1, \ldots, X_{n-1}] = \mathrm{E}[X_n \mid X_{n-1}]$, which intuitively says that the future outcome depends only on the current state. Hence,

$$\begin{aligned}
\mathrm{E}[X_n \mid X_0, X_1, \ldots, X_{n-1}] &= \mathrm{E}[X_n \mid X_{n-1}] \\
&= \mathrm{E}[X_{n-1} + (Y_n - \mathrm{E}[Y_n]) \mid X_{n-1}] \\
&= X_{n-1} + \mathrm{E}[Y_n - \mathrm{E}[Y_n] \mid X_{n-1}] \\
&= X_{n-1} + \mathrm{E}[Y_n - \mathrm{E}[Y_n]] \quad \text{by independence of the } Y_i \\
&= X_{n-1}\,, \quad \text{since } \mathrm{E}[Y_n - \mathrm{E}[Y_n]] = 0.
\end{aligned}$$

A general way of constructing martingales is given in the following definition.

Definition 1.65 *The* Doob sequence *of a function f with respect to a sequence of random variables $X_1, \ldots, X_n$ is defined by*

$$Y_i = \mathrm{E}[f \mid X_1, \ldots, X_i]$$

for all $0 \le i \le n$. In particular, $Y_0 = \mathrm{E}[f]$ and $Y_n = f(X_1, \ldots, X_n)$.

It is shown, for instance, in [DP98] that this sequence always defines a martingale (choose $f = \sum_{i=1}^{n} X_i$ to see why).

The following concentration result for martingales is known as Azuma's Inequality [Azu67] although it was first shown (implicitly) in a paper by Hoeffding (see [Hoe63]).

Theorem 1.66 (Azuma Inequality) *Let $X_0, X_1, \ldots$ be a martingale satisfying the property that $|X_i - X_{i-1}| \le c_i$ for all $i \ge 1$. Then for any $t \ge 0$,*

$$\Pr[X_n > X_0 + t] \le \mathrm{e}^{-t^2/(2\sum_{i=1}^{n} c_i^2)}$$

and

$$\Pr[X_n < X_0 - t] \le \mathrm{e}^{-t^2/(2\sum_{i=1}^{n} c_i^2)}\,.$$

Using the Doob martingale, this theorem can be generalized to obtain a very powerful and useful generalization of the Chernoff bounds (see, e.g., [DP98]).

Theorem 1.67 (Method of Bounded Martingale Differences) *Let $X_1, \ldots, X_n$ be an arbitrary set of random variables and let f be a function satisfying the property that for each $i \in [n]$, there is a non-negative c_i such that*

$$|\mathrm{E}[f \mid X_1, \ldots, X_i] - \mathrm{E}[f \mid X_1, \ldots, X_{i-1}]| \le c_i\,.$$

Then

$$\Pr[f \ge \mathrm{E}[f] + t] \le \mathrm{e}^{-t^2/(2\sum_{i=1}^{n} c_i^2)}$$

and

$$\Pr[f \le \mathrm{E}[f] - t] \le \mathrm{e}^{-t^2/(2\sum_{i=1}^{n} c_i^2)}\,.$$

Note that the X_i's in this theorem are *not* required to be independent.

1.5 Witness Structures

There are certain applications where all previous theorems may not be good enough to obtain reasonable bounds. In this case it is sometimes possible to establish a connection between the performance of the algorithm and some combinatorial structure in the following way:

> *If algorithm A has a performance of P, then there exists a combinatorial structure of type $\mathcal{C}(P)$ that witnesses this.*

If it is possible to show that for certain values of P the probability of $\mathcal{C}(P)$ to exist is very low, then this immediately implies that only with a very low probability the algorithm will have a performance that matches one of these values.

There are no fixed rules according to which to construct combinatorial witness structures. We will demonstrate the approach by an example. Assume that we have n balls and n bins. Each ball is placed in a bin that is chosen independently and uniformly at random. We noticed at the end of Example 2 in Section 1.2.2 that with a probability of at least $1/2$ the maximum number of balls in a bin is at least $(1 - o(1)) \log n / \log\log n$. In the following we demonstrate that a much better distribution of balls among the bins can be achieved if the balls have more than just one option.

Suppose that n balls are thrown sequentially into n bins in the following way: When it is the turn of ball i, 2 bins are chosen uniformly and independently at random for it. The ball is placed in the bin with the minimum number of already placed balls. Ties are broken arbitrarily. Let this placement strategy be called *minimum rule*. Then the following theorem holds. Its proof is based on a proof in [CFM$^+$98], which itself builds on proofs in [CMM$^+$98]. Both papers use so-called witness trees for the analysis. The idea of using witness structures for balls-into-bins strategies goes back to [KLM92]. Witness trees for balls-into-bins strategies have first been used in [Sch93].

Theorem 1.68 *If n is sufficiently large, the maximum load of a bin caused by the minimum rule is at most $4 \log\log n$, w.h.p.*

Proof. We prove the theorem in two parts. First, we show that if there is a bin at the end with $4d$ balls, where $d = \log\log n$, there exists a tree structure that witnesses this situation. Afterwards, we show that, with high probability, no such witness tree exists.

Constructing a witness tree

We will consider witness trees that are rooted trees of directed edges that are labeled with numbers in $\{1, \ldots, n\}$. Each node in a witness tree represents a bin and each edge (v, w) with label i represents ball i with first choice bin v and second choice bin w. The edges can be either directed towards or away from the root.

Suppose that some bin r has a load of at least $4d$ at the end. We construct a witness tree for this situation as follows. The root of the tree corresponds to bin r. Let $b_{i_1}, \ldots, b_{i_{4d}}$ be (any $4d$ of) the balls placed in r, $i_1 < i_2 < \ldots < i_{4d}$. Let r_j be the other bin choice associated with ball b_{i_j}. For each $1 \le j \le 4d$, the root r is connected via a directed edge with label i_j to a node representing bin r_j. The edge direction depends on whether r or r_j is the first choice of ball b_{i_j}. Since $i_1 < i_2 < \ldots < i_{4d}$, the load of bin r before ball b_{i_j} was inserted was at least $j-1$. Therefore, the load of bin r_j, the other choice of b_{i_j}, was at least $j-1$ at that time. We use this fact to recursively grow a tree rooted at each r_j.

The witness tree we constructed in this manner is irregular. However, it contains as a subgraph a d-ary tree of depth d (that is, of $d+1$ levels) such that

- The root in level 0 has d children that are internal nodes.
- Each internal node on levels 1 to $d-2$ has two children that are internal nodes and $d-2$ children that are leaves.
- Each internal node on level $d-1$ has d children that are leaves.

For convenience we refer to this subtree as the actual witness tree henceforth.

Constructing a pruned witness tree

If the nodes of the witness tree are guaranteed to represent distinct bins, proving our probabilistic bound is a relatively easy matter. However, this is not the case: a ball may reappear several times in a witness tree, leading to dependencies that are difficult to resolve. This makes it necessary to *prune* the tree. We will show that it suffices to prune it in such a way that each *internal* node of the resulting tree represents a distinct bin. This is done as follows:

We visit the nodes of the witness tree iteratively in a breadth-first search order starting at the root and ending with the last node in level $d-1$. Assume that we reached node v and let $B(v)$ denote the set of all nodes of the possibly already pruned tree that were visited before v. If v represents a bin that is different from the bins represented by the nodes in $B(v)$, we do nothing. Otherwise, we prune all nodes in the subtree rooted at v, not including v. The procedure continues until all nodes have been visited. The tree resulting from this process is called a *pruned witness tree*.

We note two vital properties of pruned witness trees.

Lemma 1.69 *The pruned witness tree has the following properties:*

(1) All endpoints of the edges pointing to an internal node of the pruned witness tree represent distinct choices of the balls.

(2) Each ball is represented by at most two edges of the pruned witness tree, and if it has two edges, then both of them must be incident to a leaf.

Proof. We start with a proof for item (1). From the construction of a pruned witness tree it immediately follows that all internal nodes of the pruned witness tree must be distinct. Thus, there can be no choice of a ball that is represented by more than one endpoint of an edge pointing to an internal node of the pruned witness tree.

Next we prove item (2). Suppose that a ball is represented by more than two edges. Then its edges are incident to at least 3 different internal nodes. Since each ball only has two optional bins, 2 of these internal nodes must represent the same bin, which is a contradiction to the construction of a witness tree. Further, suppose that a ball is represented by two edges, where one of them connects only internal nodes. Then, again, its edges must be incident to at least 3 different internal nodes. Again, this is not possible, which concludes the proof. □

Bounding the expected number of pruned witness trees

We showed above that if there is a bin with at least $4d$ balls, then there is a pruned witness tree that fulfills the conditions of Lemma 1.73. In order to bound the probability that such a pruned witness tree exists, we intend to bound the expected number of ways of constructing a pruned witness tree and assigning balls to its edges such that

- all choices of the balls pointing to a common node in the tree represent the same bin and
- Lemma 1.73 is true.

Such an assignment is called *legal*.

To bound the expected number of pruned witness trees with a legal assignment, we distinguish between two cases:

(1) At least one tree rooted at a child of root r has not been pruned.

(2) Every tree rooted at a child of the root r has been pruned at least once.

In the first case, we restrict the witness tree to any one of the non-pruned subtrees under the root and throw away all other subtrees. This results in a d-ary witness tree of depth of $d-1$, called *of type 1*, that fulfills the conditions of Lemma 1.73. We start with counting the number of possibilities of assigning bins to the nodes and balls to the edges of a witness tree of type 1, called T_1 henceforth. Clearly, T_1 consists of $\Delta = \sum_{i=0}^{d-2} 2^i = 2^{d-1}-1$ internal nodes and a total number of $m = d \cdot \Delta + 1$ nodes. There are n possibilities to select a bin for the root of T_1. Given any internal node v that has already been assigned to a bin, there are $\binom{n}{d}$ possibilities of choosing d balls for the edges leading to the children of v. If v does not have only leaves as children, then there are $\binom{d}{2}$ further possibilities of determining those 2 of the d balls whose edges lead to the 2 children that are internal nodes. Moreover, each edge can have 2 orientations, depending on whether the bin represented by v denotes its first or second choice. Once a ball and the number of its choice is fixed, the child of v with which its edge is connected has a uniquely determined bin (the other choice of the ball, whatever this might be). Hence, there are at most

$$\begin{aligned} n \cdot \left(\binom{n}{d} \binom{d}{2} 2^d \right)^{\Delta} &\leq n \cdot \left(\left(\frac{2\mathrm{e}n}{d} \right)^d \cdot \frac{d^2}{2} \right)^{\Delta} \\ &\leq n \cdot \left(\frac{4\mathrm{e}n}{d} \right)^{d \cdot \Delta} = n \cdot \left(\frac{4\mathrm{e}n}{d} \right)^{m-1} \end{aligned}$$

ways of assigning balls and bins to the edges and nodes of T_1. Due to Lemma 1.73 (1), all ball-choice pairs represented by the endpoints of the edges towards the root must be distinct. Since for each ball-choice pair a bin is selected uniformly and independently at random, the probability that the assignment of balls and bins to T_1 is legal is at most $n^{-(m-1)}$. Thus, the expected number of legal assignments to T_1 is at most

$$\begin{aligned} n \cdot \left(\frac{4\mathrm{e}n}{d} \right)^{m-1} \cdot \left(\frac{1}{n} \right)^{m-1} &= n \cdot \left(\frac{4\mathrm{e}}{d} \right)^{m-1} \\ &\leq n \cdot \left(\frac{1}{2} \right)^{d(2^{d-1}-1)} \leq n^{-\log\log n/2+2} \end{aligned}$$

if n is sufficiently large.

Now we consider the second case, that is, every tree rooted at a child of the root r has been pruned at least once. In this case, for each tree rooted at a child of r we only keep its levels up to the first level that has a pruned node. Thus all nodes of this level become leaves. This results in a pruned witness tree of type 2 with root r that also fulfills Lemma 1.73. Let us omit the fact for a while that we have pruned nodes. In this case, the expected number of ways of having a legal assignment of balls to any possible witness tree of type 2 is at most

$$n \cdot \underbrace{\binom{n}{d} 2^d \left(\frac{1}{n}\right)^d}_{(*)} \underbrace{\left(\sum_{\ell=0}^{d-1} \left(\binom{n}{d}\binom{d}{2} 2^d \left(\frac{1}{n}\right)^d\right)^{2^\ell - 1}\right)^d}_{(**)} \tag{1.51}$$

where $(*)$ counts the expected number of legal assignments for the edges incident to r, and $(**)$ counts the expected number of legal assignments for the d trees rooted at the children of r, using the formula in case 1. The problem with this bound is that it is by far too weak. To improve this bound, we exploit the fact that in each subtree we have a node that caused a pruning.

Consider some fixed witness tree T_2 of type 2 with root r. Let the tree rooted at the ith child of r, $1 \leq i \leq d$, be called T_i'. For each T_i', let r_i be its root and ℓ_i be its depth. If $\ell_i = 0$, then the bin assigned to r_i must be equal to some bin assigned to r_j, $1 \leq j < i$. Since distinct balls have to be assigned to the edges incident to r and each ball chooses its bins independently at random, we obtain for each i with $\ell_i = 0$ a probability of d/n that the choice of the ball is legal. Now suppose that $\ell_i > 0$. In this case there are at most $d \cdot 2^{\ell_i}$ ways of selecting a node v_i in T_i' that caused a pruning. Let e_i be the edge incident to v_i. Every pruned node v_i has two possibilities:

(1) The ball assigned to e_i has not been assigned to some edge e_j visited earlier in the breadth-first-search order. In this case, we use the fact that the probability that the ball assigned to e_i points to a bin that has already been used somewhere else in T_2 (which has to be the case for a pruned node) is at most $d \cdot 2^{d-1}/n$.

(2) The ball assigned to e_i has already been assigned to some edge e_j visited earlier in the breadth-first-search order. According to Lemma 1.73 (2) we know that e_j must be incident to a leaf. Since the highest level in which the leaf of e_i can be is level $d-1$, the leaf of e_j must be a pruned node. If e_j is also in subtree T_i', we can treat e_i as in case (1). Otherwise, we assume w.l.o.g. that the leaf of e_j is the pruned node in the other subtree. (We can always reselect the pruned nodes accordingly.) In this case, we use the fact that there are at most d ways of choosing a ball for e_i, since only balls assigned to edges with pruned nodes can be possible candidates. We do not have to count 2 ways of orienting e_i, since only one orientation can be valid: the opposite orientation of e_j. Otherwise, Lemma 1.73 (1) is violated.

Suppose that one of the children of node v has been selected as a pruned node. Then the bound $\binom{n}{d}\binom{d}{2}2^d n^{-d}$ for v changes to

$$\binom{n}{d} 2^d \left(\frac{1}{n}\right)^d \cdot \frac{d \cdot 2^{d-1}}{n}$$

for case (1) and

$$\binom{n}{d-1} 2^{d-1} \left(\frac{1}{n}\right)^{d-1} \cdot d$$

for case (2). Obviously, at least $d/2$ of the pruned nodes must fulfill case (1). (Otherwise Lemma 1.73 (2) is violated.)

Since a) there are at most 2^d ways of selecting one of the two cases for each of the d pruned nodes and b) the number of ways to select a pruned node in a subtree of depth ℓ_i is at most $d \cdot 2^{\ell_i}$, the expected number of legal assignments of balls and bins to any possible T_2 is at most

$$n \cdot \binom{n}{d} 2^d \left(\frac{1}{n}\right)^d \cdot \underbrace{2^d}_{a)} \cdot$$
$$\left(\underbrace{\frac{d}{n}}_{\ell_i=0} + \sum_{\ell=1}^{d-1} \underbrace{d \cdot 2^{\ell_i}}_{b)} \cdot \left(\binom{n}{d} \binom{d}{2} 2^d \left(\frac{1}{n}\right)^d \right)^{2^\ell - 2} \cdot \underbrace{\binom{n}{d} 2^d \left(\frac{1}{n}\right)^d \cdot \frac{d \cdot 2^{d-1}}{n}}_{(1)} \right)^{d/2} \cdot$$
$$\left(\underbrace{\frac{d}{n}}_{\ell_i=0} + \sum_{\ell=1}^{d-1} \underbrace{d \cdot 2^{\ell_i}}_{b)} \cdot \left(\binom{n}{d} \binom{d}{2} 2^d \left(\frac{1}{n}\right)^d \right)^{2^\ell - 2} \cdot \underbrace{\binom{n}{d-1} 2^{d-1} \left(\frac{1}{n}\right)^{d-1} \cdot d}_{(2)} \right)^{d/2} .$$

If n is large enough, then $\binom{n}{d} 4^d n^{-d} \leq 1$ and $\binom{n}{d-1} 2^{d-1} n^{d-1} d \leq 1$. This simplifies the bound above to at most

$$\begin{aligned} & n \cdot \left(\frac{d}{n} + d(2^d - 1) \cdot \frac{d \cdot 2^{d-1}}{n} \right)^{d/2} \left(\frac{d}{n} + d(2^d - 1) \right)^{d/2} \\ \leq \;& n \cdot \left(\frac{(d2^d)^2}{n} \right)^{d/2} \cdot \left(d2^d \right)^{d/2} = n \cdot \left(\frac{(d2^d)^3}{n} \right)^{d/2} \\ = \;& n^{-\Omega(\log \log n)} . \end{aligned}$$

From our construction above it follows that if there is a bin with at least $4d$ balls, then there either exists a type 1 witness tree, or a type 2 witness tree. Let the random variable X denote the number of witness trees of type 1 and 2 that can be constructed from the given choices of the balls. According to our calculations above it follows that

$$\mathrm{E}[X] = n^{-\Omega(\log \log n)} .$$

If there exists a bin with at least $4d$ balls, then we must have $X \geq 1$. However, according to the Markov inequality,

$$\Pr[X \geq 1] \leq \mathrm{E}[X] = n^{-\Omega(\log \log n)} .$$

This proves the theorem. □

1.6 The Lovász Local Lemma

Probabilistic methods have been used with great success to prove the existence of objects with desirable properties by showing that a randomly chosen object from an appropriate probability distribution has the desired properties with positive probability. In most applications this probability is not only positive, but is actually high and frequently tends to 1 as the parameters of the problem tend to infinity. In such cases, the proof usually supplies an efficient randomized algorithm for producing a structure of the desired type.

There are, however, certain examples, where one can prove the existence of the required combinatorial structure by probabilistic arguments that deal with rare events; events that hold with positive probability which is exponentially small in the size of the input. This happens often when using the Lovász Local Lemma. We present this lemma in its general form.

A directed graph $G = (V, E)$ on the set of nodes $V = \{1, 2, \ldots, n\}$ is called a *dependency graph* for the events $A_1, \ldots, A_n$ if for every i, $1 \le i \le n$, the event A_i is independent of any set of events A_j with $(i, j) \notin E$.

Lemma 1.70 (Lovász Local Lemma [EL75]) *Let $A_1, \ldots, A_n$ be "bad" events in an arbitrary probability space. Suppose that $G = (V, E)$ is a dependency graph of these events and suppose there are real numbers $x_i \in [0, 1)$ for all $1 \le i \le n$ with*

$$\Pr[A_i] \le x_i \prod_{(i,j)\in E} (1 - x_j)$$

for all $1 \le i \le n$. Then,

$$\Pr[\bigcap_{i=1}^{n} \bar{A}_i] \ge \prod_{i=1}^{n} (1 - x_i) \ .$$

In particular, with positive probability no bad event A_i holds.

Proof. We first prove, by induction on s, that for any set $S \subset \{1, \ldots, n\}$, $|S| = s$, and $i \notin S$,

$$\Pr[A_i \mid \bigcap_{j\in S} \bar{A}_j] \le x_i \ . \tag{1.52}$$

This is certainly true for $s = 0$. Assuming it holds for all $s' < s$, we prove it for S. Set $S_1 = \{j \in S : (i, j) \in E\}$ and $S_2 = S \setminus S_1$. Then

$$\Pr[A_i \mid \bigcap_{j\in S} \bar{A}_j] = \frac{\Pr[A_i \cap (\bigcap_{j\in S_1} \bar{A}_j) \mid \bigcap_{\ell\in S_2} \bar{A}_\ell]}{\Pr[\bigcap_{j\in S_1} \bar{A}_j \mid \bigcap_{\ell\in S_2} \bar{A}_\ell]} \ . \tag{1.53}$$

To bound the numerator, observe that since A_i is mutually independent of the events $\{A_\ell : \ \ell \in S_2\}$,

$$\begin{aligned} \Pr[A_i \cap (\bigcap_{j\in S_1} \bar{A}_j) \mid \bigcap_{\ell\in S_2} \bar{A}_\ell] &\le \Pr[A_i \mid \bigcap_{\ell\in S_2} \bar{A}_\ell] \\ &= \Pr[A_i] \le x_i \prod_{(i,j)\in E} (1 - x_j) \ . \end{aligned} \tag{1.54}$$

The denominator, on the other hand, can be bounded by the induction hypothesis. Indeed, suppose $S_1 = \{j_1, j_2, \ldots, j_r\}$. If $r = 0$, then the denominator is 1, and (1.53) follows. Otherwise

$$\begin{aligned} \Pr[\bigcap_{j\in S_1} \bar{A}_j \mid \bigcap_{\ell\in S_2} \bar{A}_\ell] &= \prod_{k=1}^{r} \left(1 - \Pr[A_{j_k} \mid \bar{A}_{j_1} \cap \ldots \cap \bar{A}_{j_{k-1}} \cap (\bigcap_{\ell\in S_2} \bar{A}_\ell)]\right) \\ &\ge \prod_{k=1}^{r} (1 - x_{j_k}) \ge \prod_{(i,j)\in E} (1 - x_j) \ . \end{aligned} \tag{1.55}$$

Substituting (1.55) and (1.56) into (1.54), we conclude that $\Pr[A_i \mid \bigcap_{j\in S} \bar{A}_j] \le x_i$, completing the proof of the induction.

The assertion of Lemma 1.74 now follows easily, as

$$\begin{aligned}\Pr[\bigcap_{i=1}^{n} \bar{A}_i] &= \prod_{i=1}^{n}\left(1 - \Pr[A_i \mid \bigcap_{j=1}^{i-1} \bar{A}_j]\right) \\ &\ge \prod_{i=1}^{n}(1-x_i)\ .\end{aligned}$$

□

Lemma 1.71 (The Local Lemma, Symmetric Case) *Let $A_1, \ldots, A_n$ be events in an arbitrary probability space. Suppose that each event A_i is mutually independent of a set of all the other events A_j but at most d, and that $\Pr[A_i] \le p$ for all $1 \le i \le n$. If*

$$\mathrm{e}p(d+1) \le 1 \tag{1.56}$$

then $\Pr[\bigcap_{i=1}^{n} \bar{A}_i] > 0$.

Proof. If $d = 0$, the result is trivial. Otherwise, by the assumption, there is a dependency graph $G = (V, E)$ for the events $A_1, \ldots, A_n$ in which for each i, $|\{j : (i,j) \in E\}| \le d$. The result now follows from Lemma 1.74 by taking $x_i = 1/(d+1) < 1$ for all i and using the fact that for any $d \ge 1$, $[1 - 1/(d+1)]^d > 1/\mathrm{e}$. □

It is worth noting that, as was shown by Shearer [She85], the constant "e" is the best possible constant in inequality (1.57). In the following, we present various applications of the Local Lemma. There is no known proof of any of these results that does not use the Local Lemma.

1.6.1 Existential Proofs using the Local Lemma

First, we present several existential results using the Local Lemma. We start with two important, elementary problems.

Hypergraph coloring

Consider the problem of coloring the nodes of a hypergraph $H = (V, E)$. H is called *c-colorable* if c colors are sufficient to color its nodes in such a way that no hyperedge is *monochromatic*, i.e., only has nodes of a single color. We prove a sufficient condition for the 2-colorability of hypergraphs.

Theorem 1.72 ([EL75]) *Let $H = (V, E)$ be a hypergraph in which every edge has at least k elements, and suppose that every edge of H intersects at most d other edges. If $\mathrm{e}(d+1) \le 2^{k-1}$, then H is 2-colorable.*

Proof. Color each node v of H randomly and independently, either blue or red with equal probability. For each edge $e \in E$, let A_e be the event that e is monochromatic. Obviously, $\Pr[A_e] = 2/2^{|e|} \le 1/2^{k-1}$. Moreover, each event A_e is clearly mutually independent of all the events $A_{e'}$ for which e' does not intersect e. The result now follows from Lemma 1.75. □

Set partitioning

The *set partitioning problem* [FK86] is defined as follows:

Let $H = (V, E)$ be a set system with $V = [n]$ and $E = \{S_1, \dots, S_M\} \subseteq \mathcal{P}(V)$. Given a positive integer ℓ, the problem is to partition V into ℓ parts, so that the sets are split well: we want a $\chi : V \to [\ell]$ that minimizes $Y = \max_{j \in [M], k \in [\ell]} |\{i \in S_j : \ \chi(i) = k\}|$.

We prove the following result.

Theorem 1.73 *Let $H = (V, E)$ be a system of sets of size k, and suppose that every set in H intersects at most d other sets. For any $\ell \in \mathbb{N}$ it holds that if*

$$\mathrm{e} \cdot \ell(d+1) \leq \mathrm{e}^{\frac{\epsilon^2 k}{2\ell(1+\epsilon/3)}} ,$$

then there is a set partitioning for H with $Y \leq (1+\epsilon)k/\ell$.

Proof. Consider the random experiment of choosing independently and uniformly at random one out of ℓ colors for each node. For each set $S \in E$, let A_S denote the event that S has a set of more than $(1+\epsilon)k/\ell$ nodes of the same color. For some fixed $S \in E$, let X_i denote the number of nodes in S with color i, $1 \leq i \leq \ell$. Obviously, $\mathrm{E}[X_i] = k/\ell$. Since the nodes choose their colors independently of the other nodes, we can use the Chernoff bounds to obtain

$$\Pr[X_i \geq (1+\epsilon)\mathrm{E}[X_i]] \leq \mathrm{e}^{-\frac{\epsilon^2 k/\ell}{2(1+\epsilon/3)}} .$$

From this it follows that

$$\Pr[A_S] \leq \ell \cdot \mathrm{e}^{-\frac{\epsilon^2 k/\ell}{2(1+\epsilon/3)}} .$$

The theorem now follows directly from Lemma 1.75. □

Routing

Next we will present a more advanced application of the Local Lemma that requires to use the set partitioning problem several times. Consider the problem of routing packets in an arbitrary collection of simple paths with congestion C and dilation D, one packet along each path. The proof of the following result has been taken from [Sch98].

Theorem 1.74 ([LMR94]) *For any collection of paths with congestion C and dilation D, there is an offline schedule that needs at most $O(C+D)$ time steps to send a packet along each of these paths.*

Proof. The global framework of our strategy for constructing an efficient schedule is the same as in [LMR94]. It consists of a succession of refinements, starting with an initial schedule S_0, in which each packet is forwarded at every time step until it reaches its final destination. This initial schedule is as short as possible. Its length is exactly D. Unfortunately, up to C packets may have to use an edge at a single time step in S_0, whereas in the final schedule at most one packet is allowed to use an edge at each step. Each refinement will bring us closer to meeting this requirement by bounding the congestion within smaller and smaller time intervals. The proof uses the Chernoff bounds and the Lovász Local Lemma at each refinement step. In the following, a *t-interval* denotes a time interval consisting of t consecutive time steps. A *t-frame* is defined as a t-interval that starts at an integer multiple of t.

Let $I_0 = \max[C, D]$ and $I_j = \max[\log I_{j-1},\ 6]$ for all $j \geq 1$. The first step is to assign an initial delay to each packet, chosen independently and uniformly at random from the range $\Delta_1 = [C]$. In the resulting schedule, S_1, a packet that is assigned a delay of δ waits in its source node for δ steps, then moves on without waiting again until it reaches its destination. The length of S_1 is at most $D + C$. We use the Lovász Local Lemma to show that if the delays are chosen independently and uniformly at random and I_1 is sufficiently large, then with nonzero probability the congestion at any edge in any I_1^3-interval is at most $(1 + \frac{4}{I_1})I_1^3$. Thus, such a set of delays must exist.

To apply the Lovász Local Lemma, we associate a bad event with each edge. The bad event for edge e is that more than $C_1 = (1 + \frac{4}{I_1})I_1^3$ packets use e in some I_1^3-interval. For this we have to bound the dependence b among the bad events and the probability p that a bad event occurs.

We first bound the dependence. Whether or not a bad event occurs depends solely on the delays assigned to the packets that pass through the corresponding edge. Thus, two bad events are independent unless some packet passes through both of the corresponding edges. Since at most C packets pass through an edge and each of these packets passes through at most D other edges, the dependence b of the bad events is at most $C \cdot D$.

Next we bound the probability that a bad event occurs. Let us consider some fixed edge e and I_1^3-interval J. Let the packets traversing e be numbered from 1 to $m \leq C$. For every $i \in \{1, \ldots, m\}$, let the binary random variable X_i be one if and only if packet i traverses e during J. Let $X = \sum_{i=1}^m X_i$. Since the packets choose their delays uniformly at random from a range of size C, we get $\Pr[X_i = 1] \leq I_1^3/C$ for all $i \in \{1, \ldots, m\}$. Hence,

$$\mathrm{E}[X] = \sum_{i=1}^{m} \mathrm{E}[X_i] \leq I_1^3 \ .$$

Together with the Chernoff bounds we therefore get with $\epsilon = \frac{4}{I_1}$ (which is at most 1 because of $I_1 \geq 6$) that

$$\Pr[X \geq C_1] = \Pr[X \geq (1+\epsilon)I_1^3] \leq \mathrm{e}^{-\left(\frac{4}{I_1}\right)^2 I_1^3/3} \leq 2^{-6I_1} \ .$$

Since for each edge there are at most $D + C$ many I_1^3-intervals to consider, the probability that a bad event occurs for edge e is bounded by

$$p \leq (D+C)2^{-6I_1} < \frac{1}{\mathrm{e}(CD+1)}$$

for all $I_1 \geq 6$ (assuming that $C, D \geq 6$). Thus, the product $\mathrm{e}p(b+1)$ is less than 1 and therefore, by the Lovász Local Lemma, there is some assignment of delays such that the congestion in each I_1^3-interval is bounded by $(1 + \frac{4}{I_1})I_1^3$.

We now break schedule S_1 into I_1^4-frames and continue to schedule each frame independently. So each frame can be viewed as a separate scheduling problem where the origin of a packet is its location at the beginning of the frame, and its destination is its location at the end of the frame. Our next refinement step will be to choose, for each frame, a random initial delay for each packet that visits at least one edge within this frame. In the resulting schedule S_2, the frames (enlarged by their delay ranges) are executed one after the other in a way that a packet that is assigned a delay δ in some frame F waits at its first edge in F for (additional) δ steps, and then moves on without waiting until it traverses its last edge in F.

Let us concentrate in the following on some fixed frame F. Let each packet choose an additional delay out of the range $\Delta_2 = [I_1^3 - I_2^3]$. Hence the length of the resulting schedule for this frame is at most $I_1^4 + I_1^3$. We use the Lovász Local Lemma to show that if the delays are chosen independently and uniformly at random and I_2 is sufficiently large, then with nonzero probability the congestion in any I_2^3-interval is at most

$$C_2 = \left(1 + \frac{4}{I_1}\right)\left(1 + \frac{6}{I_2}\right)\left(\frac{1}{1 - (I_2/I_1)^3}\right) I_2^3 \ .$$

We again associate a bad event with each edge. The bad event for edge e is that more than C_2 packets use e in some I_2^3-interval.

We first bound the dependence of these events. Whether or not a bad event occurs solely depends on the delays assigned to the packets that pass through the corresponding edge. Since at most $C_1 \cdot I_1$ packets pass through an edge in any frame and each of these packets passes through at most I_1^4 other edges, the dependence of the bad events is at most $C_1 \cdot I_1^5$.

Next we bound the probability that a bad event occurs. Let us consider some fixed edge e and I_2^3-interval J. We know according to the first refinement step that every I_1^3-interval has congestion at most C_1. Since the packets can only choose delays up to $I_1^3 - I_2^3$ and J covers I_2^3 time steps, at most C_1 packets are able to traverse this edge during J. Let these packets be numbered from 1 to $m \le C_1$. For every $i \in \{1, \ldots, m\}$, let the binary random variable X_i be one if and only if packet i traverses e during J. Let $X = \sum_{i=1}^m X_i$. Since the packets choose their delays uniformly at random from a range of size $I_1^3 - I_2^3$, we get $\Pr[X_i = 1] \le I_2^3/(I_1^3 - I_2^3)$ for all $i \in \{1, \ldots, m\}$. Hence,

$$\mathrm{E}[X] = \sum_{i=1}^{m} \mathrm{E}[X_i] \le \left(1 + \frac{4}{I_1}\right)\left(\frac{1}{1 - (I_2/I_1)^3}\right) I_2^3 \ .$$

Together with the Chernoff bounds we therefore get with $\epsilon = \frac{6}{I_2}$ and $\alpha_2 = (1 + \frac{4}{I_1})(\frac{1}{1-(I_2/I_1)^3})$ that

$$\Pr[X \ge C_2] = \Pr[X \ge (1+\epsilon)\alpha_2 I_2^3] \le \mathrm{e}^{-\left(\frac{6}{I_2}\right)^2 \alpha_2 I_2^3/3} \le \mathrm{e}^{-12\alpha_2 I_2} \ .$$

Since for each edge there are at most $I_1^4 + I_1^3$ many I_2^3-intervals to consider, the probability that a bad event occurs for edge e is bounded by

$$p \le (I_1^4 + I_1^3)\mathrm{e}^{-12\alpha_2 I_2} < \frac{1}{\mathrm{e}(C_1 I_1^5 + 1)}$$

for all $I_2 \ge 6$. Thus, the product $\mathrm{e}p(b+1)$ is less than 1 and therefore, by the Lovász Local Lemma, there is some assignment of delays such that the congestion in each I_2^3-interval is bounded by C_2.

We continue to refine each I_2^4-frame in F recursively until we reach a round k, in which $I_k = 6$. We end up with a schedule S_k with total length at most

$$(D + C)\prod_{j=1}^{k-1}\left(1 + \frac{1}{I_j}\right) = O(D + C)$$

and with a congestion C_k in each I_k^3-interval of at most

$$I_k^3\left(1 + \frac{4}{I_1}\right)\prod_{j=1}^{k-1}\left(1 + \frac{6}{I_{j+1}}\right)\left(\frac{1}{1 - (I_{j+1}/I_j)^3}\right) = O(I_k^3) \ .$$

In order to obtain a feasible schedule, we simply schedule each I_k^3-frame in schedule S_k in a greedy way. Since each I_k^3-frame has a constant congestion and dilation, we altogether end up with a routing schedule of runtime $O(C + D)$. □

1.6.2 Constructive Proofs using the Local Lemma

Many applications of the Local Lemma can be found in the literature (see, e.g., [ABLP91, Alo94, AMR91, ASE92, BFU97, EL75, ES91, FS98, HMR97, LMR94, Mol98, Ree98, RS98, Spe95, Sri96]). To turn proofs using the Lovász Local Lemma into efficient algorithms, even random ones, proved to be difficult for many of these applications. In a breakthrough paper [Bec91], Beck presented a method of converting some applications of the Lovász Local Lemma into polynomial time algorithms (with some sacrifices made with regards to the constants in the original application). Alon [Alo91] provided a parallel variant of the algorithm and simplified the arguments used. Leighton, Maggs and Richa [LMR99] extended Beck's arguments in 1996 so that they can also be applied to scheduling problems. Similar techniques were later used by Molloy and Reed [MR98] to find polynomial time algorithms for a general class of applications of the Lovász Local Lemma.

In the following, we will concentrate on the presentation and extension of the methods given by Alon and Molloy and Reed. The first result constructivizes Theorem 1.76.

An algorithm for 2-coloring hypergraphs

Theorem 1.75 ([Alo91]) *Let $H = (V, E)$ be a hypergraph in which every edge has at least k elements, and every edge intersects less than d other edges, where k and d are fixed. Suppose that for some $\alpha > 0$*

$$4\mathrm{e}d^3 \leq 2^{(1-H(\alpha))k} \tag{1.57}$$

and

$$2\mathrm{e}d \leq 2^{\alpha k} \ . \tag{1.58}$$

Then there is a randomized algorithm that finds a 2-coloring for H in polynomial time.

Proof. Suppose that we want to find a 2-coloring for a hypergraph H consisting of n nodes and m edges, labeled e_1 to e_m. The algorithm for finding a proper 2-coloring of H consists of several passes. In the first pass, all nodes of H are colored randomly and independently by one out of two colors, where each point is colored either red or blue with equal probability. We call an edge *bad* if at most αn of its points are red or at most αn of its points are blue. According to Section 1.1.2, the probability of a fixed edge to be bad is at most

$$2 \sum_{i \leq \alpha k} \binom{k}{i} 2^{-k} \leq 2 \cdot 2^{(H(\alpha)-1)k} \ ,$$

where $H(\alpha)$ is the binary entropy function. Put $p = 2 \cdot 2^{(H(\alpha)-1)k}$ and let B denote the set of all bad edges.

Let G be the dependency graph for the problem, whose nodes are the hyperedges of H and in which two nodes are connected if and only if their corresponding hyperedges intersect. Observe that if S is an independent set in G, then the probability that $S \subset B$ is at most $p^{|S|}$, since these $|S|$ events are mutually independent. Let us call a set of nodes C of G a 1,2-tree if C is the set of nodes of a connected subgraph in the square of G, i.e., C is a 1,2-tree if the nodes in C are such that drawing an edge between two nodes in C if their distance in G is either 1 or 2 the resulting graph is connected.

Lemma 1.76 *The probability that every 1,2-tree in G all of whose nodes belong to B has a size of at most $d \log(2m)$ is at least $1/2$.*

Proof. We call $T \subseteq G$ a 2,3-tree if the nodes in T are such that all their mutual distances in G are at least two and so that, drawing an edge between two nodes in T if their distance in G is either 2 or 3 the resulting graph is connected. We first bound the number of 2,3-trees of size s in G. Consider the graph on the set of nodes of G in which two nodes are adjacent if their distance in G is either 2 or 3. Every 2,3-tree on a set T of nodes of G must contain a tree on T in this new graph. The new graph has a maximum degree of at most $D = (d-1)^3 + (d-1)^2 < d^3$. It is well known (see, e.g., [Knu69], p. 369 (Exercise 11)) that an infinite D-regular rooted tree contains precisely $\frac{1}{(D-1)s+1}\binom{Ds}{s}$ rooted subtrees of size s, and this easily implies that the number of trees of size s containing one specific given node in any graph with maximum degree at most D does not exceed this number, which is smaller than $(\mathrm{e}D)^s$.

For any particular 2,3-tree T of size s we know that $\Pr[T \subseteq B] \le p^s$. Hence, the expected number of 2,3-trees $T \subseteq B$ of size s is at most $m(\mathrm{e}Dp)^s$. As $\mathrm{e}Dp \le 1/2$ by (1.58), if $s = \log(2m)$ this term is at most $1/2$. Thus with probability at least $1/2$ there is no 2,3-tree of size bigger than $\log(2m)$ all of whose nodes are in B. We actually want to bound the size of any 1,2-tree C of G all of whose nodes lie in B. A maximal 2,3-tree T in such a C must have the property that every node in C is a neighbor (in G) of a node in T. This is because otherwise there is a node in C which is not a neighbor of any node in T, and yet it is of distance 2 or 3 from some node in T. Such a node can be added to T, contradicting its maximality. Since there are less than d neighbors of any node in G, $\log(2m) \ge |T| \ge |C|/d$ and so

$$|C| \le d \log(2m)$$

completing the proof of the lemma. □

Let us call the first pass successful if there is no 1,2-tree of size greater than $d\log(2m)$ all of whose nodes lie in B. By the last lemma, the probability that the first pass is successful is at least $1/2$. In case it is not, we simply repeat the entire procedure. In expected linear time the first pass is successful.

We can now fix the coloring by recoloring, in the second pass, the nodes of H that belong to the bad edges. Let us call an edge *dangerous* if it contains at least αk nodes that belong to bad edges. (Thus, in particular, bad edges are also dangerous.) Observe that if an edge is not dangerous then it will not become monochromatic after the recoloring. This is because less than αk of its points will change color, and it has at least αk points of each color before the recoloring. Thus we only have to worry about the dangerous edges. However, if we recolor all the nodes in the bad edges randomly and independently, then we recolor at least αk nodes in each dangerous edge, and hence the probability that it becomes monochromatic does not exceed $2^{-\alpha k+1}$. Since each dangerous edge intersects less than d others, it follows from (1.59) and Theorem 1.76 that there is a recoloring in which no edge is monochromatic.

The crucial point is that the recoloring of the points in the edges of each maximal 1,2-tree C of bad edges can be done separately. This is because there is no dangerous edge that intersects edges from two distinct such maximal 1,2-trees, and hence it suffices to recolor the points in the edges of each such C in a way that only makes sure that no dangerous edge intersecting an edge in C becomes monochromatic. Since we assume d and k to be fixed, each of the 1,2-trees C as above has only $O(\log m)$ nodes that have to be recolored. Thus we can find the required recoloring by exhaustive enumeration, as examining all possible two-colorings in each such C only takes time polynomial in m. □

Extending the 2-coloring algorithm to non-constant d and k

If d and k are not fixed, in order to obtain a polynomial time algorithm, one can make another pass similar to the first one in each 1,2-tree separately, which results in new 1,2-trees of size $O(d\log(d\log m))$. However, since this may reduce the size of the hyperedges up to some α^2 fraction of their original size, we have to change the conditions for α in Theorem 1.79 to

$$4\mathrm{e}d^3 \le 2^{(1-H(\alpha))\alpha k} \quad \text{and} \quad 2\mathrm{e}d \le 2^{\alpha^2 k} \,.$$

to ensure that a second reduction is possible. At this point, we distinguish between two cases. Let c be a sufficiently large constant. If $k \le c\log\log\log n$ then $d = (\log\log n)^{O(1)}$. Thus, we can simply use exhaustive search in each 1,2-tree to obtain a polynomial time algorithm. Otherwise, we perform another random experiment of choosing a color independently and uniformly for each node covered by a 1,2-tree. Since there are at most $O(d\log(d\log m))$ hyperedges in a 1,2-tree we have: If c is large enough, with probability at least $1/2$ a coloring will be chosen in a 1,2-tree such that all edges are monochromatic. Thus, we also end up with a polynomial time algorithm in this case.

A general method for the symmetric LLL

In this section, we present a technique developed by Molloy and Reed [MR98] that seems to capture all applications to which Beck's technique will apply.

In the following, let $\mathcal{T} = \{t_1,\ldots,t_n\}$ be a set of independent random trials, and let $\mathcal{A} = \{A_1,\ldots,A_m\}$ be a set of events such that each A_i is determined by the outcome of the trials in $T_i \subseteq \mathcal{T}$. We say that T_i *intersects* T_j and A_i *intersects* A_j ($A_i \sim A_j$) if $T_i \cap T_j \ne \emptyset$.

For any $t_{j_1},\ldots,t_{j_k} \in T_i$ and any $w_{j_1},\ldots,w_{j_k}$ in the domains of $t_{j_1},\ldots,t_{j_k}$, we define $\Pr^*[A_i \mid t_{j_1} = w_{j_1},\ldots,t_{j_k} = w_{j_k}]$ to be the probability of A_i conditional on the event that the outcomes of $t_{j_1},\ldots,t_{j_k}$ are $w_{j_1},\ldots,w_{j_k}$. We sometimes just say $\Pr^*[A_i]$ if it causes no ambiguity, always meaning $t_{j_1},\ldots,t_{j_k}$ to be the set of trials already carried out and $w_{j_1},\ldots,w_{j_k}$ to be their outcomes. For $k=0$, we have $\Pr^*[A_i] = \Pr[A_i]$.

Theorem 1.77 ([MR98]) *If we have the following:*

(1) for each $1 \le i \le m$, $\Pr[A_i] \le p$;

(2) each T_i *intersects less than* d *other* T_j*'s;*

(3) $p \cdot d^9 \le (1/2\mathrm{e})^3$;

(4) for each $1 \le i \le m$, $|T_i| \le w$;

(5) for each $1 \le j \le n$*, the size of the domain of* t_i *is at most* γ*, and we can carry out the random trial in time* τ_1*;*

(6) for each $1 \le i \le m$, $t_{j_1}, \dots t_{j_k} \in T_i$ and $w_{j_1}, \dots, w_{j_k}$ in the domains of $t_{j_1}, \dots t_{j_k}$, we can compute $\Pr^*[A_i]$ in time τ_2;

then we have a randomized $O(n \cdot d \cdot (\tau_1 + \tau_2) + n \cdot \gamma^{wd(\log(d \log m)+2)})$-time algorithm which will find outcomes for $t_1, \dots, t_n$ such that none of the events in $\mathcal{A}$ holds.

The algorithm and the proof are essentially the same as those in [Alo91] and [Bec91]. The new idea here is to consider the conditional probabilities $\Pr^*[A_i]$ and to occasionally undo a trial.

In the first pass, we carry out the trials $t_1, \dots, t_n$ in sequential order. After each trial t_j, we compute $\Pr^*[A_i]$ for every i with $t_j \in T_i$. If $\Pr^*[A_i] > p^{2/3}$, then we say that A_i is *dangerous*, and we

- undo t_j, i.e., we cancel its outcome and carry out the trial again at a later time, and
- freeze t_j and all other remaining trials in T_i, i.e., we will not carry out those trials during this pass and so when it comes to their turn we will skip them.

Note that at the end of the first pass, $\Pr^*[A_i] \le p^{2/3}$ for all i. Therefore, by the (symmetric form of the) Local Lemma, upon carrying out the remaining trials,

$$\Pr[\bar{A}_1 \cap \dots \cap \bar{A}_m] > 0 \ .$$

That is, there is a feasible solution extending the partial solution given by the trials that have already been carried out. We will see that so few of the events become dangerous that it is now nearly feasible to find a good set of outcomes for the remaining trials using exhaustive search.

Lemma 1.78 *For each $1 \le i \le m$, the probability that A_i becomes dangerous is at most $p^{1/3}$.*

Proof. If the probability that $\Pr^*[A_i]$ will ever exceed $p^{2/3}$ is greater than $p^{1/3}$, then $\Pr[A_i] > p$. □

We denote by H the hypergraph with node set $V = \mathcal{T}$ and edge set $E = \{T_1, \dots, T_m\}$ representing the events. Let a 1,2-tree and a 2,3-tree of hyperedges in H be defined as in Section 1.6.2. We call an a, b-tree *dangerous* if all of its hyperedges correspond to dangerous events.

The key observation is this: No A_i intersects two events which belong to different maximal dangerous 1,2-trees. Thus, we can deal with the frozen trials contained in each maximal dangerous 1,2-tree independently, and so as long as these trees is sufficiently small, an exhaustive search for each tree of all the combinations of possible outcomes of the corresponding trials is feasible.

Lemma 1.79 *With probability at least $1/2$, all dangerous 1,2-trees have a size of at most $d \log(2m)$.*

Proof. The proof is similar to Lemma 1.80. We require that $\mathrm{e}d^3 p^{1/3} \le 1/2$ for the lemma to hold. This is fulfilled due to condition (3) above. □

If after the first pass we have dangerous 1,2-trees of size greater than $d \log(2m)$, then we repeat the first pass. The expected number of repetitions is constant, and each repetition takes time $O(n \cdot d \cdot (\tau_1 + \tau_2))$. At this point, an exhaustive search for the satisfactory outcomes to the frozen trials corresponding to each dangerous 1,2-tree takes time $O(\gamma^{wd \log(2m)})$. Thus, if d, γ, and w are constant, this can be done in polynomial time. To improve the runtime, we run a second pass in the same manner as the first pass, where we carry out the frozen trials in sequence, and an event becomes dangerous if its conditional probability exceeds $p^{1/3}$. Within an expected linear number of repetitions of the second pass, there will be no dangerous 1,2-trees of size greater than $d\log(2m')$ where $m' = d\log(2m)$, and so we can complete our exhaustive search of each one in time $O(\gamma^{wd(\log(d\log m)+2)})$. Thus the total expected running time is

$$O\left(n \cdot d \cdot (\tau_1 + \tau_2) + n \cdot \gamma^{wd(\log(d\log m)+2)}\right) .$$

An improved general method for the symmetric LLL

Now we show how to improve the algorithm by Molloy and Reed to obtain polynomial time algorithms for (almost) all applications that match their conditions.

Theorem 1.80 *If we have the following:*

(1) for each $1 \le i \le m$, $\Pr[A_i] \le p$*;*

(2) each T_i *intersects less than* d *other* T_j*'s;*

(3) for each $1 \le i \le m$, $|T_i| \le w$*;*

(4a) $p \cdot d^9 \le (1/2\mathrm{e})^3$*;*

(4b) $p \cdot w \le 1$*; (this is the only additional condition!)*

(5) for each $1 \le j \le n$*, we can carry out the random trial in time* τ_1*;*

(6) for each $1 \le i \le m$, $t_{j_1}, \dots t_{j_k} \in T_i$ *and* $w_{j_1}, \dots, w_{j_k}$ *in the domains of* $t_{j_1}, \dots t_{j_k}$*, we can compute* $\Pr^*[A_i]$ *in time* τ_2*;*

then we have a randomized $O(n \cdot d \cdot (\tau_1 + \tau_2) + n(\tau_1 \cdot \log^{O(1)} m + 2^{(\log\log m)^{O(1)}}))$*-time algorithm which will find outcomes for* $t_1, \dots, t_n$ *such that none of the events in* $\mathcal{A}$ *holds.*

Proof. In the following, we always assume that m is sufficiently large. The first two passes work in the same way as in the algorithm by Molloy and Reed. However, instead of using exhaustive search at the end, we distinguish between the following two cases:

If $p^{1/3} \le 1/(2d\log(2d\log(2m)))$, then we simply carry out all the random trials in each 1,2-tree T. Since the probability of a dangerous event to be true is ensured to be at most $p^{1/3}$ and every 1,2-tree consists of at most $d\log(2d\log(2m))$ dangerous events, the probability that there is an event left in T that is true is at most $1/2$. Hence, in an expected number of $O(\tau_1 \cdot n)$ steps all events are made false.

Suppose otherwise that $p^{1/3} > 1/(2d\log(2d\log(2m)))$. According to condition (4a) it holds that $d^3 \le 1/(2\mathrm{e}p^{1/3})$. Thus, if m is sufficiently large, this restricts d to be at most $\log\log m$. From this we obtain together with condition (4b) that

$$w \le 1/p \le (2d\log(2d\log(2m)))^3 \le (3\log\log m)^6 .$$

Thus, altogether $d \cdot w \le 3^6(\log\log m)^7$. If $\gamma = \log^{O(1)} m$, then we use exhaustive search to find outcomes for the trials so that all events are false. This takes at most $2^{(\log\log m)^{O(1)}}$ time for T. If $\gamma = \log^{\Omega(1)} m$, we will first use a strategy presented below that reduces γ to a value in $\log^{O(1)} m$. Afterwards, we continue as for $\gamma = \log^{O(1)} m$.

In order to reduce γ from $\log^{\Omega(1)} m$ to $\log^{O(1)} m$, we use the following simple sampling algorithm for each 1,2-tree T:

For each random trial t_j in T, perform s random experiments (s will be in the order of $\log^{O(1)} m$). The outcomes of these random experiments are stored in vector $\boldsymbol{v}_j = (v_{j,1}, \ldots, v_{j,s})$.

We show that if we change now the random experiment in a way that each t_j is determined by choosing an entry in $\boldsymbol{v}_j$ independently and uniformly at random, then we still have a probability of at most $2p^{1/3}$ for a dangerous event to become true, w.h.p. In this case, condition (4a) implies that according to the symmetric LLL there must be outcomes for the trials from the sample sets such that all events are false. (If this is not true, because the sampling failed to guarantee a probability of at most $2p^{1/3}$ for a dangerous event to become true, we will notice it after the exhaustive search and simply restart the sampling procedure above.)

Lemma 1.81 *Consider some fixed 1,2-tree T. If $s = \log^{O(1)} m$, then with probability at least $1 - 1/m$ outcome vectors are chosen such that every dangerous event in T has a probability of at most $2p^{1/3}$ to be true.*

Proof. We will use Theorem 1.52 to prove the lemma. Consider any fixed event A in T, and let T_A denote the set of its trials, $|T_A| = q$. W.l.o.g. we assume that $T_A = \{t_1, \ldots, t_q\}$. Let $M = [s^q]$ represent the set of all possibilities of choosing an outcome vector $O = (v_{1,j_1}, \ldots, v_{1,j_q})$ out of the sample vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_q$. Suppose that we choose uniformly at random a subset $U \subset M$ of size k, representing the outcome vectors $O_\ell = (v_{1,j_{1,\ell}}, \ldots, v_{q,j_{q,\ell}})$, $1 \leq \ell \leq k$. Let the binary random variable X_ℓ be 1 if and only O_ℓ implies that A is true. Our aim is to bound

$$z_\ell = \Pr[X_\ell = 1 \mid X_1 = \ldots = X_{\ell-1} = 1]$$

for all $1 \leq \ell \leq k$. Let the event E_ℓ be true if and only if outcome O_ℓ has at least one entry in common with an outcome $O_{\ell'}$ with $1 \leq \ell' < \ell$, that is, there is an $\ell' \in [\ell - 1]$ and an $i \in [q]$ such that $j_{i,\ell} = j_{i,\ell'}$. Clearly,

$$\Pr[E_\ell] \leq \binom{q(\ell-1)}{1} \frac{s^{q-1}}{s^q} = \frac{q \cdot (\ell - 1)}{s} .$$

On the other hand, we know that

$$\Pr[X_\ell = 1 \mid X_1 = \ldots = X_{\ell-1} = 1 \;\wedge\; \bar{E}_\ell] \leq p^{1/3} .$$

Thus, altogether we have

$$\begin{aligned} z_\ell &= \Pr[E_\ell] \cdot \Pr[X_\ell = 1 \mid X_1 = \ldots = X_{\ell-1} = 1 \;\wedge\; E_\ell] + \\ &\quad\; \Pr[\bar{E}_\ell] \cdot \Pr[X_\ell = 1 \mid X_1 = \ldots = X_{\ell-1} = 1 \;\wedge\; \bar{E}_\ell] \\ &\leq \Pr[E_\ell] + \Pr[X_\ell = 1 \mid X_1 = \ldots = X_{\ell-1} = 1 \;\wedge\; \bar{E}_\ell] \\ &\leq \frac{q \cdot k}{s} + p^{1/3} . \end{aligned}$$

Let us choose $s = q \cdot k \cdot (\log m)/p^{1/3}$. From above we know that $q \leq (3 \log\log m)^6$. Furthermore, we intend to choose $k = \alpha \log m$ for some constant $\alpha > 0$. Thus, $s = (\log m)^{O(1)}$. Moreover, we obtain

$$z_\ell \leq \left(1 + \tfrac{1}{\log m}\right) p^{1/3}$$

for all $1 \leq \ell \leq k$.

For any $i \in M$, let the binary random variable Y_i be 1 if and only if A is true for outcome vector i, and let $Y = \sum_{i \in M} Y_i$. Obviously, it holds that $\Pr[Y_i = 1] \leq p^{1/3}$. Thus,

$$\mathrm{E}[Y] \leq s^q \cdot p^{1/3} \ .$$

From above we know that, for the multilinear polynomial $P_j(.)$ used in the proof of Theorem 1.52, it holds that

$$P_j(Y_1, \ldots, Y_{s^q}) \leq P_j(p', \ldots, p') \quad \text{with } p' = (1 + \tfrac{1}{\log m})p^{1/3}$$

for all $1 \leq j \leq k$. Thus, according to Theorem 1.52, we have

$$\Pr[Y \geq 2p^{1/3}] \leq \mathrm{e}^{-\min[(1/2)s^q p^{1/3}/3,\ \lfloor k/2 \rfloor]} \leq \frac{1}{m^2}$$

if $k = \alpha \log m$ and α is a sufficiently large constant. Hence, the probability of A to become true (if each t_j in A is determined by choosing an entry in $\boldsymbol{v}_j$ independently and uniformly at random) is at most $2p^{1/3}$ with probability at least $1 - 1/m^2$. Since T has less than m dangerous events, the lemma follows. $\square$

Lemma 1.85 concludes the proof of Theorem 1.84. $\square$

Bibliography

[ABLP91] N. Alon, A. Bar-Noy, N. Linial, and D. Peleg. A lower bound for radio broadcast. *Journal of Computer and System Sciences*, 43(2):290–298, 1991.

[Alo91] N. Alon. A parallel algorithmic version of the local lemma. *Random Structures and Algorithms*, 2(4):367–387, 1991.

[Alo94] N. Alon. Probabilistic methods in coloring and decomposition problems. *Discrete Mathematics*, 127:31–46, 1994.

[AMR91] N. Alon, C. McDiarmid, and B. Reed. Acyclic coloring of graphs. *Random Structures and Algorithms*, 2(3):277–288, 1991.

[ASE92] N. Alon, J. Spencer, and P. Erdős. *The Probabilistic Method*. John Wiley & Sons, New York, 1992.

[Azu67] K. Azuma. Weighted sums of certain dependent random variables. *Tôkuku Math. J.*, 19:357–367, 1967.

[Bec91] J. Beck. An algorithmic approach to the Lovász local lemma. *Random Structures and Algorithms*, 2(4):343–365, 1991.

[Ben62] G. Bennet. Probability inequalities for sums of independent random variables. *Journal of the American Statistical Association*, 57:33–45, 1962.

[BFU97] A. Z. Broder, A. M. Frieze, and E. Upfal. Static and dynamic path selection on expander graphs: a random walk approach. In *Proc. of the 29th ACM Symp. on Theory of Computing (STOC)*, pages 531–539, 1997.

[CFM+98] R. Cole, A. Frieze, B.M. Maggs, M. Mitzenmacher, A.W. Richa, R.K. Sitaraman, and E. Upfal. On balls and bins with deletions. In *Proc. of the 2nd International Workshop on Randomization and Approximation Techniques in Computer Science (RANDOM)*, 1998.

[Che52] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–509, 1952.

[CMM+98] R. Cole, B.M. Maggs, F. Meyer auf der Heide, M. Mitzenmacher, A.W. Richa, K. Schröder, R.K. Sitaraman, and B. Vöcking. Randomized protocols for low-congestion circuit routing in multistage interconnection networks. In *Proc. of the 29th ACM Symp. on Theory of Computing (STOC)*, pages 378–388, 1998.

[DP98] D. Dubhashi and A. Panconesi. Concentration of measure for the analysis of randomized algorithms. Unpublished manuscript, accessible via http://www.cs.unibo.it/∼pancones/papers.html, October 20 1998.

[DR94] D. Dubhashi and D. Ranjan. Come correlation inequalities for probabilistic analysis of algorithms. Technical report, MPI-I-94-143, Max Planck Institut für Informatik, Saarbrücken, Germany, 1994.

[EL75] P. Erdős and L. Lovász. Problems and results on 3-chromatic hypergraphs and some related questions. In A. Hajnal, R. Rado, and V.T. Sós, editors, *Infinite and Finite Sets (to Paul Erdős on his 60th birthday)*, pages 609–627. North-Holland, Amsterdam, 1975.

[ES91] P. Erdős and J. Spencer. Lopsided Lovász local lemma and latin transversals. *Discrete Applied Mathematics*, 30:151–154, 1991.

[FGK71] C.M. Fortuin, J. Ginibre, and P.N. Kasteleyn. Correlation inequalities for partially ordered sets. *Communications of Mathematical Physics*, 22:89–103, 1971.

[FK86] Z. Füredi and J. Kahn. On the dimensions of ordered sets of bounded degree. *Order*, 3:15–20, 1986.

[FS98] U. Feige and C. Scheideler. Improved bounds for acyclic job shop scheduling. In *Proc. of the 30th ACM Symp. on Theory of Computing (STOC)*, pages 624–633, 1998.

[GK68] B.V. Gnedenko and A.N. Kolmogorov. *Limit Distributions for Sums of Independent Random Variables*. Addison-Wesley, Reading, 1968.

[Gle75] L. Gleser. On the distribution of the number of successes in independent trials. *Ann. Probab.*, 3:182–188, 1975.

[HMR97] H. Hind, M. Molloy, and B. Reed. Colouring a graph frugally. *Combinatorica*, 17(4):469–482, 1997.

[Hoe63] W. Hoeffding. Probability inequalities for sums of bounded random variables. *American Statistical Association Journal*, 58:13–30, 1963.

[Hof87] M. Hofri. *Probabilistic Analysis of Algorithms*. Springer Verlag, Heidelberg, 1987.

[HR90] T. Hagerup and C. Rüb. A guided tour of Chernoff bounds. *Information Processing Letters*, 33:305–308, 1989/90.

[IM74] K. Ito and H.P. McKean. *Diffusion Processes and their Sample Paths*. Springer Verlag, Heidelberg, 1974.

[JK77] N. L. Johnson and S. Kotz. *Urn Models and Their Application*. John Wiley & Sons, New York, 1977.

[JK82] N.L. Johnson and S. Kotz. *Encyclopedia of Statistical Sciences*. John Wiley, New York, 1982.

[JP83] K. Joag-Dev and F. Proschan. Negative association of random variables, with applications. *The Annals of Statistics*, 11(1):286–295, 1983.

[KLM92] R. Karp, M. Luby, and F. Meyer auf der Heide. Efficient PRAM simulation on a distributed memory machine. In *Proc. of the 24th ACM Symp. on Theory of Computing (STOC)*, pages 318–326, 1992.

[Knu69] D. Knuth. *The Art of Computer Programming, Vol. I.* Addison Wesley, London, 1969.

[LMR94] F. T. Leighton, B. M. Maggs, and S. B. Rao. Packet routing and job-shop scheduling in O(congestion + dilation) steps. *Combinatorica*, 14(2):167–186, 1994.

[LMR99] F. T. Leighton, B. M. Maggs, and A. W. Richa. Fast algorithms for finding o(congestion + dilation) packet routing schedules. *Combinatorica*, 19(2):1–27, 1999.

[Loé77] M. Loéve. *Probability theory*. Springer Verlag, New York, 1977.

[McD89] C. McDiarmid. On the method of bounded differences. In J. Siemons, editor, *Surveys in Combinatorics*. London Mathematical Society Lecture Note Series 141, Cambridge University Press, 1989.

[McD98] McDiarmid. Concentration. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, editors, *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 195–247. Springer Verlag, Berlin, 1998.

[Mik93] V.G. Mikhailov. On a refinement of the Central Limit Theorem for sums of independent random indicators. *Theory of Probability and Its Applications*, 38(3):479–489, 1993.

[Mol98] M. Molloy. The probabilistic method. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, editors, *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 1–35. Springer Verlag, Berlin, 1998.

[MP97] M. Merkle and L. Petrovic. Inequalities for sums of independent geometrical random variables. *Aequationes Mathematicae*, 54:173–180, 1997.

[MR98] M. Molloy and B. Reed. Further algorithmic aspects of the local lemma. In *Proc. of the 30th ACM Symp. on Theory of Computing (STOC)*, pages 524–529, 1998.

[Nav97] P. Naveau. Comparison between the Chernoff and factorial moment bounds for discrete random variables. *The American Statistician*, 51(1):40–41, 1997.

[Pet95] V.V. Petrov. *Limit Theorems of Probability Theory*. Clarendon Press, Oxford, 1995.

[PN95] T. K. Philips and R. Nelson. The moment bound is tighter than Chernoff's bound for positive tail estimates. *American Statistical Association*, 49(2):175–178, 1995.

[Pre85] E.L. Presman. Approximation in variation of the distribution of a sum of independent Bernoulli variables with a Poisson law. *Theory of Probability and Its Applications*, 30(2):417–422, 1985.

[PS92] A. Panconesi and A. Srinivasan. Fast randomized algorithms for distributed edge coloring. In *Proc. of the 11th IEEE Symp. on Principles of Distributed Computing (PODC)*, pages 251–262, 1992.

[Ree98] B. Reed. ω, Δ, and χ. *Journal of Graph Theory*, 27(4):177–212, 1998.

[RS98] J. Radhakrishnan and A. Srinivasan. Improved bounds and algorithms for hypergraph two-coloring. In *Proc. of the 39th IEEE Symp. on Foundations of Computer Science (FOCS)*, pages 684–693, 1998.

[RT96] Y. Rabani and E. Tardos. Distributed packet switching in arbitrary networks. In *Proc. of the 28th ACM Symp. on Theory of Computing (STOC)*, pages 366–375, 1996.

[Sch93] C. Scheideler. Analysis of shared memory simulations that use several hash functions (in German). Master's thesis, Dept. of Mathematics and Computer Science, Paderborn University, Germany, June 1993.

[Sch98] C. Scheideler. *Universal Routing Strategies for Interconnection Networks*. Springer Verlag, Heidelberg, 1998.

[She85] Shearer. On a problem of Spencer. *Combinatorica*, 5:241–245, 1985.

[Shi96] A.N. Shiryaev. *Probability*. Springer Verlag, New York, 1996.

[Spe95] J. Spencer. Probabilistic methods. In R. Graham, M. Grötschel, and L. Lovász, editors, *Handbook of Combinatorics*, chapter 33. Elsevier Science, Amsterdam, 1995.

[Sri96] A. Srinivasan. An extension of the Lovász local lemma, and its applications to integer programming. In *Proc. of the 7th ACM Symp. on Discrete Algorithms (SODA)*, pages 6–15, 1996.

[SSS95] J. Schmidt, A. Siegel, and A. Srinivasan. Chernoff-Hoeffding bounds for applications with limited independence. *SIAM Journal on Discrete Mathematics*, 8(2):223–250, 1995.

[SV99] C. Scheideler and B. Vöcking. From static to dynamic routing: efficient transformations of store-and-forward protocols. In *Proc. of the 31st ACM Symp. on Theory of Computing (STOC)*, pages 215–224, 1999.

[Tal95a] M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Institut Des Hautes Études Scientifiques, Publications Mathématiques*, 81:73–205, 1995.

[Tal95b] M. Talagrand. The missing factor in Hoeffding's inequalities. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 31(4):689–702, 1995.

[Val82] L.G. Valiant. A scheme for fast parallel communication. *SIAM Journal on Computing*, 11(2):350–361, 1982.