Five Proofs of Chernoff's Bound with Applications

Wolfgang Mulzer
Institut für Informatik
Freie Universität Berlin
mulzer@inf.fu-berlin.de

Abstract

We discuss five ways of proving Chernoff's bound and show how they lead to different extensions of the basic bound.

1 Introduction

Chernoff's bound gives an estimate on the probability that a sum of independent Binomial random variables deviates from its expectation [14]. It has many variants and extensions that are known under various names such as *Bernstein's inequality* or *Hoeffding's bound* [4, 14]. Chernoff's bound is one of the most basic and versatile tools in the life of a theoretical computer scientist, with a seemingly endless amount of applications. Almost every contemporary textbook on algorithms or complexity theory contains a statement and a proof of the bound [2, 8, 12, 16], and there are several texts that discuss its various applications in great detail (e.g., the textbooks by Alon and Spencer [1], Dubhashi and Panchonesi [10], Mitzenmacher and Upfal [19], Motwani and Raghavan [21], or the articles by Chung and Lu [6], Hagerup and Rüb [13], or McDiarmid [17]).

In the present survey, we will see five different ways of proving the basic Chernoff bound. The different techniques used in these proofs allow various generalizations and extensions, some of which we will also discuss.

2 The Basic Bound

We begin with a statement of the basic Chernoff bound. For this, we first need a notion from information theory [9]. Let $P = (p_1, \ldots, p_m)$ and $Q = (q_1, \ldots, q_m)$ be two probability distributions on m elements, i.e., $p_i, q_i \in \mathbb{R}$ with $p_i, q_i \geq 0$,

for $i = 1, \ldots, m$, and $\sum_{i=1}^{m} p_i = \sum_{i=1}^{m} q_i = 1$. The *Kullback-Leibler divergence* or *relative entropy* of P and Q is defined as

$$D_{\mathrm{KL}}(P\|Q) := \sum_{i=1}^{m} p_i \ln \frac{p_i}{q_i}.$$

If $m = 2$, i.e., if $P = (p, 1-p)$ and $Q = (q, 1-q)$, we write $D_{\mathrm{KL}}(p\|q)$ for $D_{\mathrm{KL}}((p, 1-p)\|(q, 1-q))$. The Kullback-Leibler divergence measures the distance between the distributions P and Q: it represents the expected loss of efficiency if we encode an m-letter alphabet with distribution P with a code that is optimal for distribution Q. Now, the basic Chernoff bound is as follows:

Theorem 2.1. *Let $n \in \mathbb{N}$, $p \in [0, 1]$, and let $X_1, \ldots, X_n$ be n independent random variables with $X_i \in \{0, 1\}$ and $\Pr[X_i = 1] = p$, for $i = 1, \ldots n$. Set $X := \sum_{i=1}^{n} X_i$. Then, for any $t \in [0, 1-p]$, we have*

$$\Pr[X \geq (p+t)n] \leq e^{-D_{\mathrm{KL}}(p+t\|p)n}.$$

3 Five Proofs for Theorem 2.1

We will now see five different ways of proving Theorem 2.1.

3.1 The Moment Method

The usual textbook proof of Theorem 2.1 uses the exponential function exp and Markov's inequality. It is called the *moment method*, because exp simultaneously encodes all *moments* $X, X^2, X^3, \ldots$ of X. This trick is often attributed to Bernstein [4]. It is very general and can be used to obtain several variants of Theorem 2.1, perhaps most prominently, the Azuma-Hoeffding inequality for martingales with bounded differences [3, 14].

The proof goes as follows. Let $\lambda > 0$ be a parameter to be determined later. We have

$$\Pr[X \geq (p+t)n] = \Pr[\lambda X \geq \lambda(p+t)n] = \Pr\big[e^{\lambda X} \geq e^{\lambda(p+t)n}\big].$$

From Markov's inequality, we obtain

$$\Pr\big[e^{\lambda X} \geq e^{\lambda(p+t)n}\big] \leq \frac{\mathbf{E}[e^{\lambda X}]}{e^{\lambda(p+t)n}}.$$

Now, the independence of the X_i yields

$$\mathbf{E}[e^{\lambda X}] = \mathbf{E}\Big[e^{\lambda \sum_{i=1}^{n} X_i}\Big] = \mathbf{E}\left[\prod_{i=1}^{n} e^{\lambda X_i}\right] = \prod_{i=1}^{n} \mathbf{E}\Big[e^{\lambda X_i}\Big] = (pe^{\lambda} + 1 - p)^n.$$

Thus,

$$\Pr[X > (p+t)n] \leq \Big(\frac{pe^{\lambda} + 1 - p}{e^{\lambda(p+t)}}\Big)^n, \tag{1}$$

for every $\lambda > 0$. Optimizing for λ using calculus, we get that the right hand side is minimized if

$$e^{\lambda} = \frac{(1-p)(p+t)}{p(1-p-t)}.$$

Plugging this into (1), we get

$$\Pr[X > (p+t)n] \leq \left[\Big(\frac{p}{p+t}\Big)^{p+t}\Big(\frac{1-p}{1-p-t}\Big)^{1-p-t}\right]^n = e^{-D_{\mathrm{KL}}(p+t\|p)n},$$

as desired.

3.2 Chvátal's Method

The following proof of Theorem 2.1 is due to Chvátal [7]. As we will see below, it can be generalized to give tail bounds for the *hypergeometric distribution*. Let $B(n,p)$ be the random variable that gives the number of heads in n independent Bernoulli trials with success probability p. Then,

$$\Pr[B(n,p) = l] = \binom{n}{l} p^l (1-p)^{n-l},$$

for $l = 0, \ldots, n$. Thus, for any $\tau \geq 1$ and $k \geq pn$, we get

$$\begin{aligned}
\Pr[B(n,p) \geq k] &= \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i} \\
&\leq \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i} \underbrace{\tau^{i-k}}_{\geq 1} + \underbrace{\sum_{i=0}^{k-1} \binom{n}{i} p^i (1-p)^{n-i} \tau^{i-k}}_{\geq 0} = \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \tau^{i-k}.
\end{aligned}$$

Using the Binomial theorem, we obtain

$$\begin{aligned}
\Pr[B(n,p) \geq k] &\leq \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \tau^{i-k} \\
&= \tau^{-k} \sum_{i=0}^{n} \binom{n}{i} (p\tau)^i (1-p)^{n-i} = \frac{(p\tau + 1 - p)^n}{\tau^k}.
\end{aligned}$$

If we write $k = (p+t)n$ and $\tau = e^{\lambda}$, we get

$$\Pr[B(n,p) \geq (p+t)n] \leq \Big(\frac{pe^{\lambda} + 1 - p}{e^{\lambda(p+t)}}\Big)^n.$$

This is the same as (1), so we can complete the proof of Theorem 2.1 as in Section 3.1.

3.3 The Impagliazzo-Kabanets Method

The third proof is due to Impagliazzo and Kabanets [15], and it leads to a constructive version of the bound. Let $\lambda \in [0,1]$ be a parameter to be chosen later. Let $I \subseteq \{1,\dots,n\}$ be a random index set obtained by including each element $i \in \{1,\dots,n\}$ with probability λ. We estimate $\Pr\left[\prod_{i\in I} X_i = 1\right]$ in two different ways, where the probability is over the random choice of $X_1,\dots,X_n$ and I.

On the one hand, using the union bound and independence, we have

$$\begin{aligned}
\Pr\Big[\prod_{i\in I} X_i = 1\Big] &\le \sum_{S\subseteq\{1,\dots,n\}} \Pr\Big[I = S \wedge \prod_{i\in S} X_i = 1\Big] \\
&= \sum_{S\subseteq\{1,\dots,n\}} \Pr[I = S] \cdot \prod_{i\in S} \Pr[X_i = 1] \\
&= \sum_{S\subseteq\{1,\dots,n\}} \lambda^{|S|}(1-\lambda)^{n-|S|} \cdot p^{|S|} \\
&= \sum_{s=0}^{n} \binom{n}{s} (\lambda p)^s (1-\lambda)^{n-s} = (\lambda p + 1 - \lambda)^n, \qquad (2)
\end{aligned}$$

by the Binomial theorem. On the other hand, by the law of total probability,

$$\Pr\Big[\prod_{i\in I} X_i = 1\Big] \ge \Pr\Big[\prod_{i\in I} X_i = 1 \mid X \ge (p+t)n\Big] \Pr[X \ge (p+t)n].$$

Now, fix $X_1,\dots,X_n$ with $X \ge (p+t)n$. For the fixed choice of $X_1 = x_1,\dots,X_n = x_n$, the probability $\Pr\left[\prod_{i\in I} x_i = 1\right]$ is exactly the probability that I avoids all the $n - X$ indices i where $x_i = 0$. Thus,

$$\Pr\Big[\prod_{i\in I} x_i = 1\Big] = (1-\lambda)^{n-X} \ge (1-\lambda)^{(1-p-t)n}.$$

Since the bound holds uniformly for every choice of $x_1,\dots,x_n$ with $X \ge (p+t)n$, we get

$$\Pr\Big[\prod_{i\in I} X_i = 1 \mid X \ge (p+t)n\Big] \ge (1-\lambda)^{(1-p-t)n},$$

so

$$\Pr\Big[\prod_{i\in I} X_i = 1\Big] \ge (1-\lambda)^{(1-p-t)n} \Pr[X \ge (p+t)n].$$

Combining with (2),

$$\Pr[X \ge (p+t)n] \le \left(\frac{\lambda p + 1 - \lambda}{(1-\lambda)^{(1-p-t)}}\right)^n. \qquad (3)$$

Using calculus, we get that the right hand side is minimized for $\lambda = t/(1-p)(p+t)$ (note that $\lambda \leq 1$ for $t \leq 1-p$). Plugging this into (3),

$$\Pr[X > (p+t)n] \leq \left[\left(\frac{p}{p+t}\right)^{p+t}\left(\frac{1-p}{1-p-t}\right)^{1-p-t}\right]^n = e^{-D_{\mathrm{KL}}(p+t\|p)n},$$

as desired.

3.4 The Encoding Argument

The next proof stems from discussions with Luc Devroye, Gábor Lugosi, and Pat Morin, and it is inspired by an encoding argument [20]. A similar argument can also be derived from Xinjia Chen's *likelihood ratio method* [5]. Let $\{0,1\}^n$ be the set of all bit strings of length n, and let $w : \{0,1\}^n \to [0,1]$ be a *weight function*. We call w *valid* if $\sum_{x\in\{0,1\}^n} w(x) \leq 1$. The following lemma says that for any probability distribution p_x on $\{0,1\}^n$, a valid weight function is unlikely to be substantially larger than p_x.

Lemma 3.1. *Let $\mathcal{D}$ be a probability distribution on $\{0,1\}^n$ that assigns to each $x \in \{0,1\}^n$ a probability p_x, and let w be a valid weight function. For any $s \geq 1$, we have*

$$\Pr_{x\sim\mathcal{D}}\left[w(x) \geq sp_x\right] \leq 1/s.$$

Proof. Let $Z_s = \{x \in \{0,1\}^n \mid w(x) \geq sp_x\}$. We have

$$\Pr_{x\sim\mathcal{D}}\left[w(x) \geq sp_x\right] = \sum_{\substack{x\in Z_s\\ p_x>0}} p_x \leq \sum_{\substack{x\in Z_s\\ p_x>0}} p_x\frac{w(x)}{sp_x} \leq (1/s)\sum_{x\in Z_s} w(x) \leq 1/s,$$

since $w(x)/sp_x \geq 1$ for $x \in Z_s$, $p_x > 0$, and since w is valid. □

We now show that Lemma 3.1 implies Theorem 2.1. For this, we interpret the sequence $X_1, \ldots, X_n$ as a bit string of length n. This induces a probability distribution $\mathcal{D}$ that assigns to each $x \in \{0,1\}^n$ the probability $p_x = p^{k_x}(1-p)^{n-k_x}$, where k_x denotes the number of 1-bits in x. We define a weight function $w : \{0,1\}^n \to [0,1]$ by $w(x) = (p+t)^{k_x}(1-p-t)^{n-k_x}$, for $x \in \{0,1\}^n$. Then w is valid, since $w(x)$ is the probability that x is generated by setting each bit to 1 independently with probability $p+t$. For $x \in \{0,1\}^n$, we have

$$\frac{w(x)}{p_x} = \left(\frac{p+t}{p}\right)^{k_x}\left(\frac{1-p-t}{1-p}\right)^{n-k_x}.$$

Since $((p+t)/p)((1-p)/(1-p-t)) \geq 1$, it follows that $w(x)/p_x$ is an increasing function of k_x. Hence, if $k_x \geq (p+t)n$, we have

$$\frac{w(x)}{p_x} \geq \left[\left(\frac{p+t}{p}\right)^{p+t}\left(\frac{1-p-t}{1-p}\right)^{1-p-t}\right]^n = e^{D_{\mathrm{KL}}(p+t\|p)n}.$$

We now apply Lemma 3.1 to $\mathcal{D}$ and w to get

$$\Pr[X \geq (p+t)n] = \Pr_{x\sim\mathcal{D}}[k_x \geq (p+t)n] \leq \Pr_{x\sim\mathcal{D}}\left[w(x) \geq p_x e^{D_{\mathrm{KL}}(p+t\|p)n}\right] \leq e^{-D_{\mathrm{KL}}(p+t\|p)n},$$

as claimed in Theorem 2.1.

See the survey [20] for a more thorough discussion of how this proof is related to coding theory.

3.5 A Proof via Differential Privacy

The fifth proof of Chernoff's bound is due to Steinke and Ullman [22], and it uses methods from the theory of differential privacy [11]. Unlike the previous four proofs, it seems to lead to a slightly weaker version of the bound. Let m be a parameter to be determined later. The main idea is to bound the expectation of m independent copies of X.

Lemma 3.2. *Let $m \in \mathbb{N}$, $m \leq e^n$, and let $X^{(1)}, X^{(2)}, \ldots, X^{(m)}$ be m independent copies of X. Then,*

$$\mathbf{E}[\max\{X^{(1)}, X^{(2)}, \ldots, X^{(m)}\}] \leq pn + 5\sqrt{n \ln m}.$$

We will give a proof of Lemma 3.2 below. First, however, we will see how we can use Lemma 3.2 to derive the following weaker version of Theorem 2.1.

Theorem 3.3. *Let $n \in \mathbb{N}$, $p \in [0,1]$, and let $X_1, \ldots, X_n$ be n independent random variables with $X_i \in \{0,1\}$ and $\Pr[X_i = 1] = p$, for $i = 1, \ldots n$. Set $X := \sum_{i=1}^n X_i$. Then, for any $t \in [0, 1-p]$, we have*

$$\Pr[X \geq (p+t)n] \leq e^{-\frac{1}{64}t^2 n}.$$

Proof. Set $\alpha = \Pr[X \geq (p+t)n]$, and let $X^{(1)}, X^{(2)}, \ldots, X^{(m)}$ be m independent copies of X. Then,

$$\Pr\left[\max\{X^{(1)}, X^{(2)}, \ldots, X^{(m)}\} \geq (p+t)n\right] = 1 - (1-\alpha)^m \geq 1 - e^{-\alpha m}. \qquad (4)$$

On the other hand, Markov's inequality gives

$$\begin{aligned}
&\Pr\left[\max\{X^{(1)}, X^{(2)}, \ldots, X^{(m)}\} \geq (p+t)n\right] \\
&= \Pr\left[\max\{X^{(1)}, X^{(2)}, \ldots, X^{(m)}\} - pn \geq tn\right] \\
&\leq \frac{\mathbf{E}[\max\{X^{(1)}, X^{(2)}, \ldots, X^{(m)}\}] - pn}{tn} \leq \frac{5\sqrt{\ln m}}{t\sqrt{n}},
\end{aligned}$$

by Lemma 3.2. Thus, setting $m = \exp\left(\left(\frac{e-1}{5e}\right)^2 t^2 n\right)$, and combining with (4), we get

$$(e-1)/e \geq 1 - e^{-\alpha m} \Leftrightarrow \alpha \leq e^{-\left(\frac{e-1}{5e}\right)^2 t^2 n}.$$

The lemma follows since $\left(\frac{e-1}{5e}\right)^2 \geq \frac{1}{64}$. □

It remains to prove Lemma 3.2. For this, we use an idea from differential privacy. Let $A \in \{0,1\}^{m\times n}$, $A = (a_{ij})$, be an $(m \times n)$-matrix with entries from $\{0,1\}$. For a given parameter $\gamma > 1$, we define a random variable $S_\gamma(A)$ with values in $\{1, \ldots, m\}$ as follows: for $i = 1, \ldots, m$, let $b_i = \sum_{j=1,\ldots,n} a_{ij}$ be the sum of the entries in the i-th row of A. Set

$$C_\gamma(A) = \sum_{i=1}^{m} \gamma^{b_i}.$$

Then, for $i = 1, \ldots, m$, we define

$$\Pr[S_\gamma(A) = i] = \frac{\gamma^{b_i}}{C_\gamma(A)}.$$

The random variable $S_\gamma(A)$ is called a *stable selector* for A (see the work by McSherry and Talwar [18] for more background). The next lemma states two interesting properties for $S_\gamma(A)$. For a matrix $A \in \{0,1\}^{m\times n}$, a vector $\vec{c} \in \{0,1\}^m$, and a number $j \in \{1, \ldots, n\}$ we denote by $(A_{-j}, \vec{c})$ the matrix obtained from A by replacing the j-th column of A with $\vec{c}$.

Lemma 3.4. *Let $A \in \{0,1\}^{m\times n}$ be an $m \times n$ matrix with entries in $\{0,1\}$. We have*

- ***Stability**: For every vector $\vec{c} \in \{0,1\}^m$ and every $i \in \{1, \ldots, m\}$,*

$$\gamma^{-2} \Pr[S_\gamma(A_{-j}, \vec{c}) = i] \leq \Pr[S_\gamma(A) = i] \leq \gamma^2 \Pr[S_\gamma(A_{-j}, \vec{c}) = i].$$

- ***Accuracy**: Let b_i be the sum of the i-th row of A. Then,*

$$\mathbf{E}_{i\sim S_\gamma(A)}[b_i] \leq \max_{i=1}^{m} b_i \leq \mathbf{E}_{i\sim S_\gamma(A)}[b_i] + \log_\gamma m.$$

Proof. **Stability**: for $k \in \{1, \dots, m\}$, let b_k be the sum of the k-th row of A, and let $\widetilde{b_k}$ be the sum of the k-th row of $(A_{-j}, \widetilde{c})$. Since A and $(A_{-j}, \widetilde{c})$ differ in one column, and since the entries are from $\{0, 1\}$, we have $\widetilde{b_k} - 1 \leq b_k \leq \widetilde{b_k} + 1$. Hence,

$$\gamma^{-1} C_\gamma(A_{-j}, \vec{c}) \leq C_\gamma(A) \leq \gamma C_\gamma(A_{-j}, \vec{c})$$

and

$$\gamma^{-2} \Pr[S_\gamma(A_{-j}, \vec{c}) = i] \leq \Pr[S_\gamma(A) = i] \leq \gamma^2 \Pr[S_\gamma(A_{-j}, \vec{c}) = i],$$

as claimed.

Accuracy: The inequality $\mathbf{E}_{i \sim S_\gamma(A)}[b_i] \leq \max_{i=1}^m b_i$ is obvious. For the second inequality, we observe that by definition,

$$b_i = \log_\gamma(C_\gamma(A) \Pr[S_\gamma(A) = i]).$$

Thus,

$$\begin{aligned}
\mathbf{E}_{i \sim S_\gamma(A)}[b_i] &= \sum_{i=1}^m \Pr[S_\gamma(A) = i] \log_\gamma(C_\gamma(A) \Pr[S_\gamma(A) = i]) \\
&= \sum_{i=1}^m \Pr[S_\gamma(A) = i] \log_\gamma C_\gamma(A) - \sum_{i=1}^m \Pr[S_\gamma(A) = i] \log_\gamma \frac{1}{\Pr[S_\gamma(A) = i]} \\
&\geq \sum_{i=1}^m \Pr[S_\gamma(A) = i] \log_\gamma \gamma^{\max_{i=1}^m b_i} - \log_\gamma m, \\
&= \max_{i=1}^m b_i - \log_\gamma m,
\end{aligned}$$

since $C_\gamma(A) = \sum_{i=1}^m \gamma^{b_i} \geq \gamma^{\max_{i=1}^m b_i}$ and since $x \mapsto -\log_\gamma(x)$ is a convex function. □

Lemma 3.4 shows that $S_\gamma(A)$ constitutes a reasonable mechanism of estimating the maximum row sum of A without revealing too much information about any single column of A. We can now use Lemma 3.4 to bound the expectation of the maximum of m independent copies of X.

Lemma 3.5. *Let $m \in \mathbb{N}$, and let $X^{(1)}, X^{(2)}, \dots, X^{(m)}$ be m independent copies of X. Then, for any $\gamma > 1$, we have*

$$\mathbf{E}[\max\{X^{(1)}, X^{(2)}, \dots, X^{(m)}\}] \leq \gamma^2 pn + \log_\gamma m.$$

Proof. Let $X_1^{(1)}, X_1^{(2)}, \dots, X_1^{(m)}$ be m independent copies of X_1, let $X_2^{(1)}$, $X_2^{(2)}$, …, $X_2^{(m)}$ be m independent copies of X_2, and so on. We consider the random $m \times n$

matrix $M \in \{0,1\}^{m\times n}$ whose entry in row i and column j is $X_j^{(i)}$. Then, we can write $X^{(i)} = \sum_{j=1}^{n} X_j^{(i)}$, for $i = 1, \dots, m$. By the accuracy claim in Lemma 3.4,

$$\mathbf{E}_M[\max\{X^{(1)}, \dots, X^{(m)}\}] \le \mathbf{E}_{M,i\sim S_\gamma(M)}[X^{(i)}] + \log_\gamma m \tag{5}$$

Now we bound $\mathbf{E}_{M,i\sim S_\gamma(M)}[X^{(i)}]$. We unwrap the expectation for $i \sim S_\gamma(M)$ and get

$$\mathbf{E}_{M,i\sim S_\gamma(M)}[X^{(i)}] = \mathbf{E}_M\Big[\sum_{i=1}^{m} \Pr[S_\gamma(M) = i]X^{(i)}\Big]$$

Let $\widetilde{M}$ be an independent copy of M. Denote the entry in the i-th row and j-th column of $\widetilde{M}$ by $\widetilde{X}_j^{(i)}$, and set $\widetilde{X}^{(i)} = \sum_{j=1}^{n} \widetilde{X}_j^{(i)}$, for $i = 1, \dots, m$. By the stability claim in Lemma 3.4, for every $j \in \{1, \dots, n\}$,

$$\mathbf{E}_M\Big[\sum_{i=1}^{m} \Pr\left[S_\gamma(M) = i\right]X^{(i)}\Big] \le \gamma^2 \mathbf{E}_{M,\widetilde{M}}\Big[\sum_{i=1}^{m} \Pr\left[S_\gamma(M_{-j}, \widetilde{M}_j) = i\right]X^{(i)}\Big].$$

Since the random variables $X_j^{(i)}$, $\widetilde{X}_j^{(i)}$, $1 \le i \le m$, $1 \le j \le n$, are independent, the pairs $((M_{-j}, \widetilde{M}_j), X_j^{(i)})$ and $(M, \widetilde{X}_j^{(i)})$ have the same distribution. Therefore, we can write

$$\begin{aligned}
\mathbf{E}_M\Big[\sum_{i=1}^{m} \Pr\left[S_\gamma(M)) = i\right]X^{(i)}\Big] &= \mathbf{E}_M\Big[\sum_{i=1}^{m}\sum_{j=1}^{n} \Pr\left[S_\gamma(M)) = i\right]X_j^{(i)}\Big] \\
&\le \gamma^2 \mathbf{E}_{M,\widetilde{M}}\Big[\sum_{j=1}^{n}\sum_{i=1}^{m} \Pr\left[S_\gamma(M_{-j}, \widetilde{M}_j) = i\right]X_j^{(i)}\Big] \\
&= \gamma^2 \mathbf{E}_{M,\widetilde{M}}\Big[\sum_{j=1}^{n}\sum_{i=1}^{m} \Pr\left[S_\gamma(M) = i\right]\widetilde{X}_j^{(i)}\Big] \\
&= \gamma^2 \mathbf{E}_M\Big[\sum_{i=1}^{m} \Pr\left[S_\gamma(M) = i\right]\mathbf{E}_{\widetilde{M}}[\widetilde{X}^{(i)}]\Big] \\
&= \gamma^2 \mathbf{E}_M\Big[\sum_{i=1}^{m} \Pr\left[S_\gamma(M) = i\right]pn\Big] = \gamma^2 pn.
\end{aligned}$$

We can conclude the lemma by plugging this bound into (5). □

To obtain Lemma 3.2, we set $\gamma = 1 + \frac{\sqrt{\ln m}}{\sqrt{n}}$. Now, Lemma 3.5 gives

$$\begin{aligned}
\mathbf{E}[\max\{X^{(1)}, X^{(2)}, \dots, X^{(m)}\}] &\le \left(1 + \frac{\sqrt{\ln m}}{\sqrt{n}}\right)^2 pn + \frac{\ln m}{\ln\left(1 + \frac{\sqrt{\ln m}}{\sqrt{n}}\right)} \\
&\le \left(1 + \frac{3\sqrt{\ln m}}{\sqrt{n}}\right) pn + \frac{\ln m}{\frac{\sqrt{\ln m}}{2\sqrt{n}}},
\end{aligned}$$

since $\frac{\sqrt{\ln m}}{\sqrt{n}} \leq 1$ by our assumption $m \leq e^n$ and $\ln(1+x) \geq x/2$, for $x \in [0,1]$. Hence, using $pn \leq n$,

$$\mathbf{E}[\max\{X^{(1)}, X^{(2)}, \ldots, X^{(m)}\}] \leq pn + 5\sqrt{n \ln m},$$

as desired.

4 Useful Consequences

We now show several useful consequences of Theorem 2.1. These results can be derived directly from Theorem 2.1, and therefore they also hold for variants of the theorem with slightly different assumptions.

4.1 The Lower Tail

First, we show that an analogous bound holds for the lower tail probability $\Pr[X \leq (p-t)n]$.

Corollary 4.1. *Let $X_1, \ldots, X_n$ be independent random variables with $X_i \in \{0,1\}$ and $\Pr[X_i = 1] = p$, for $i = 1, \ldots n$. Set $X := \sum_{i=1}^n X_i$. Then, for any $t \in [0,p]$, we have*

$$\Pr[X \leq (p-t)n] \leq e^{-D_{\mathrm{KL}}(p-t\|p)n}.$$

Proof.

$$\Pr[X \leq (p-t)n] = \Pr[n - X \geq n - (p-t)n] = \Pr[X' \geq (1-p+t)n],$$

where $X' = \sum_{i=1}^n X_i'$ with independent random variables $X_i' \in \{0,1\}$ such that $\Pr[X_i' = 1] = 1-p$. The result follows from $D_{\mathrm{KL}}(1-p+t\|1-p) = D_{\mathrm{KL}}(p-t\|p)$. □

4.2 Multiplicative version

Next, we derive a multiplicative variant of Theorem 2.1. This well-known version of the bound can be found in the classic text by Motwani and Raghavan [21].

Corollary 4.2. *Let $X_1, \ldots, X_n$ be independent random variables with $X_i \in \{0,1\}$ and $\Pr[X_i = 1] = p$, for $i = 1, \ldots n$. Set $X := \sum_{i=1}^n X_i$ and $\mu = pn$. Then, for any $\delta \geq 0$, we have*

$$\Pr[X \geq (1+\delta)\mu] \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}, \text{ and}$$

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

Proof. Setting $t = \delta\mu/n$ in Theorem 2.1 yields

$$\begin{aligned}\Pr[X \geq (1+\delta)\mu] &\leq \exp\left(-n\left[p(1+\delta)\ln(1+\delta) + p\left(\frac{1-p}{p} - \delta\right)\ln\left(1 - \delta\frac{p}{1-p}\right)\right]\right) \\ &= \left(\frac{(1-\delta p/(1-p))^{\delta-(1-p)/p}}{(1+\delta)^{1+\delta}}\right)^{\mu} \\ &\leq \left(\frac{e^{-\delta^2 p/(1-p)+\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu} \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}.\end{aligned}$$

Setting $t = \delta\mu/n$ in Corollary 4.1 yields

$$\begin{aligned}\Pr[X \leq (1-\delta)\mu] &\leq \exp\left(-n\left[p(1-\delta)\ln(1-\delta) + p\left(\frac{1-p}{p} + \delta\right)\ln\left(1 + \delta\frac{p}{1-p}\right)\right]\right) \\ &= \left(\frac{(1+\delta p/(1-p))^{-\delta-(1-p)/p}}{(1-\delta)^{1-\delta}}\right)^{\mu} \\ &\leq \left(\frac{e^{-\delta^2 p/(1-p)-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu} \leq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu}.\end{aligned}$$

□

4.3 Useful Variants

The next few corollaries give some handy variants of the bound that are often more manageable in practice. First, we give a simple bound for the multiplicative lower tail.

Corollary 4.3. *Let $X_1, \ldots, X_n$ be independent random variables with $X_i \in \{0, 1\}$ and $\Pr[X_i = 1] = p$, for $i = 1, \ldots n$. Set $X := \sum_{i=1}^{n} X_i$ and $\mu = pn$. Then, for any $\delta \in (0, 1)$, we have*

$$\Pr[X \leq (1-\delta)\mu] \leq e^{-\delta^2\mu/2}.$$

Proof. By Corollary 4.2

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

Using the power series expansion of $\ln(1-\delta)$, we get

$$(1-\delta)\ln(1-\delta) = -(1-\delta)\sum_{i=1}^{\infty}\frac{\delta^i}{i} = -\delta + \sum_{i=2}^{\infty}\frac{\delta^i}{(i-1)i} \geq -\delta + \delta^2/2.$$

Thus,

$$\Pr[X \leq (1-\delta)\mu] \leq e^{[-\delta+\delta-\delta^2/2]\mu} = e^{-\delta^2\mu/2},$$

as claimed. □

An only slightly more complicated bound can be found for the multiplicative upper tail.

Corollary 4.4. *Let $X_1, \dots, X_n$ be independent random variables with $X_i \in \{0, 1\}$ and $\Pr[X_i = 1] = p$, for $i = 1, \dots n$. Set $X := \sum_{i=1}^{n} X_i$ and $\mu = pn$. Then, for any $\delta \geq 0$, we have*

$$\Pr[X \geq (1 + \delta)\mu] \leq e^{-\min\{\delta^2, \delta\}\mu/4}.$$

Proof. We may assume that $(1 + \delta)p \leq 1$. Then, Theorem 2.1 gives

$$\Pr[X \geq (1 + \delta)pn] \leq e^{-D_{\mathrm{KL}}((1+\delta)p \| p)n}.$$

Define $f(\delta) := D_{\mathrm{KL}}((1 + \delta)p \| p)$. Then,

$$f'(\delta) = p \ln(1 + \delta) - p \ln(1 - \delta p/(1 - p))$$

and

$$f''(\delta) = \frac{p}{(1 + \delta)(1 - p - \delta p)} \geq \frac{p}{1 + \delta}.$$

By Taylor's theorem, we have

$$f(\delta) = f(0) + \delta f'(0) + \frac{\delta^2}{2} f''(\xi),$$

for some $\xi \in [0, \delta]$. Since $f(0) = f'(0) = 0$, it follows that

$$f(\delta) = \frac{\delta^2}{2} f''(\xi) \geq \frac{\delta^2 p}{2(1 + \xi)} \geq \frac{\delta^2 p}{2(1 + \delta)}.$$

For $\delta \geq 1$, we have $\delta/(1 + \delta) \geq 1/2$, for $\delta < 1$, we have $1/(\delta + 1) \geq 1/2$. This gives, for all $\delta \geq 0$,

$$f(\delta) \geq \min\{\delta^2, \delta\} p/4,$$

and the claim follows. □

The following corollary combines the two bounds. This variant can be found, e.g., in the book by Arora and Barak [2].

Corollary 4.5. *Let $X_1, \dots, X_n$ be independent random variables with $X_i \in \{0, 1\}$ and $\Pr[X_i = 1] = p$, for $i = 1, \dots n$. Set $X := \sum_{i=1}^{n} X_i$ and $\mu = pn$. Then, for any $\delta > 0$, we have*

$$\Pr[|X - \mu| \geq \delta\mu] \leq 2e^{-\min\{\delta^2, \delta\}\mu/4}.$$

Proof. Combine Corollaries 4.3 and 4.4. □

The following corollary, which appears, e.g., in the book by Motwani and Raghavan [21], is also sometimes useful.

Corollary 4.6. *Let $X_1, \dots, X_n$ be independent random variables with $X_i \in \{0, 1\}$ and $\Pr[X_i = 1] = p$, for $i = 1, \dots n$. Set $X := \sum_{i=1}^{n} X_i$ and $\mu = pn$. For $t \geq 2e\mu$, we have*

$$\Pr[X \geq t] \leq 2^{-t}.$$

Proof. By Corollary 4.2

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

For $\delta \geq 2e - 1$, the denominator in the right hand side is at least $2e$, and the claim follows. □

5 Generalizations

We mention a few generalizations of the proof techniques for Section 3. Since the consequences from Section 4 are based on simple algebraic manipulation of the bounds, the same consequences also hold for the generalized settings.

5.1 Hoeffding-Extension

The moment method (Section 3.1) yields many generalizations of Theorem 2.1. The following result is known as *Hoeffding's extension* [14]. It shows that the X_i can actually be chosen to be continuous with varying expectations.

Theorem 5.1. *Let $X_1, \dots, X_n$ be independent random variables with $X_i \in [0, 1]$ and $\mathbf{E}[X_i] = p_i$. Set $X := \sum_{i=1}^{n} X_i$ and $p := (1/n)\sum_{i=1}^{n} p_i$. Then, for any $t \in [0, 1-p]$, we have*

$$\Pr[X \geq (p+t)n] \leq e^{-D_{\mathrm{KL}}(p+t\|p)n}.$$

Proof. Let $\lambda > 0$ a parameter to be determined later. As before, Markov's inequality yields

$$\Pr[e^{\lambda X} \geq e^{\lambda(p+t)n}] \leq \frac{\mathbf{E}[e^{\lambda X}]}{e^{\lambda(p+t)n}}.$$

Using independence, we get

$$\mathbf{E}[e^{\lambda X}] = \mathbf{E}\Big[e^{\lambda \sum_{i=1}^{n} X_i}\Big] = \prod_{i=1}^{n} \mathbf{E}\Big[e^{\lambda X_i}\Big]. \tag{6}$$

Now we need to estimate $\mathbf{E}[e^{\lambda X_i}]$. The function $z \mapsto e^{\lambda z}$ is convex, so $e^{\lambda z} \leq (1-z)e^{0\cdot\lambda} + ze^{1\cdot\lambda}$ for $z \in [0, 1]$. Hence,

$$\mathbf{E}[e^{\lambda X_i}] \leq \mathbf{E}[1 - X_i + X_i e^{\lambda}] = 1 - p_i + p_i e^{\lambda}.$$

Going back to (6),

$$\mathbf{E}[e^{\lambda X}] \leq \prod_{i=1}^{n}(1 - p_i + p_i e^{\lambda}).$$

Using the arithmetic-geometric mean inequality $\prod_{i=1}^{n} x_i \leq ((1/n)\sum_{i=1}^{n} x_i)^n$, for $x_i \geq 0$, this is

$$\mathbf{E}[e^{\lambda X}] \leq (1 - p + pe^{\lambda})^n.$$

From here we continue as in Section 3.1. □

5.2 Hypergeometric Distribution

Chvátals proof [7] from Section 3.2 generalizes to the *hypergeometric* distribution. We emphasize once again that this means that all the corollaries from Section 4 also apply to this case.

Theorem 5.2. *Suppose we have an urn with N balls, P of which are red. We randomly draw n balls from the urn without replacement. Let $H(N, P, n)$ denote the number of red balls in the sample. Set $p := P/N$. Then, for any $t \in [0, 1 - p]$, we have*

$$\Pr\left[H(N, P, n) \geq (p + t)n\right] \leq e^{-D_{\mathrm{KL}}(p+t\|p)n}.$$

Proof. It is well known that

$$\Pr[H(N, P, n) = l] = \binom{P}{l}\binom{N - p}{n - l}\binom{N}{l}^{-1},$$

for $l = 0, \ldots, n$.

Claim 5.3. *For every $j \in \{0, \ldots, n\}$, we have*

$$\binom{N}{n}^{-1} \sum_{i=j}^{n} \binom{P}{i}\binom{N - P}{n - i}\binom{i}{j} \leq \binom{n}{j} p^j.$$

Proof. Consider the following random experiment: take a random permutation of the N balls in the urn. Let S be the sequence of the first n elements in the permutation. Let X be the number of j-subsets of S that contain only red balls. We compute $\mathbf{E}[X]$ in two different ways. On the one hand,

$$\mathbf{E}[X] = \sum_{i=j}^{n} \Pr[\text{S contains } i \text{ red balls}]\binom{i}{j} = \sum_{i=j}^{n} \binom{N}{n}^{-1}\binom{P}{i}\binom{N - P}{n - i}\binom{i}{j}. \tag{7}$$

On the other hand, let $I \subseteq \{1, \ldots, n\}$ with $|I| = j$. Then the probability that all the balls in the positions indexed by I are red is

$$\frac{P}{N} \cdot \frac{P-1}{N-1} \cdot \cdots \cdot \frac{P-j+1}{N-j+1} \leq \left(\frac{P}{N}\right)^j = p^j.$$

Thus, by linearity of expectation $\mathbf{E}[X] \leq \binom{n}{j} p^j$. Together with (7), the claim follows. □

Claim 5.4. *For every $\tau \geq 1$, we have*

$$\binom{N}{n}^{-1} \sum_{i=0}^{n} \binom{P}{i}\binom{N-P}{n-i} \tau^i \leq (1 + (\tau - 1)p)^n.$$

Proof. Using Claim 5.3 and the Binomial theorem (twice),

$$\begin{aligned}
\binom{N}{n}^{-1} \sum_{i=0}^{n} \binom{P}{i}\binom{N-P}{n-i} \tau^i &= \binom{N}{n}^{-1} \sum_{i=0}^{n} \binom{P}{i}\binom{N-P}{n-i} (1 - (\tau - 1))^i \\
&= \binom{N}{n}^{-1} \sum_{i=0}^{n} \binom{P}{i}\binom{N-P}{n-i} \sum_{j=0}^{i} \binom{i}{j} (\tau - 1)^j \\
&= \binom{N}{n}^{-1} \sum_{j=0}^{n} (\tau - 1)^j \sum_{i=j}^{n} \binom{P}{i}\binom{N-P}{n-i}\binom{i}{j} \\
&\leq \sum_{j=0}^{n} \binom{n}{j} ((\tau - 1)p)^j = (1 + (\tau - 1)p)^n,
\end{aligned}$$

as claimed. □

Thus, for any $\tau \geq 1$ and $k \geq pn$, we get as before

$$\begin{aligned}
\Pr[H(N, P, n) \geq k] &= \binom{N}{n}^{-1} \sum_{i=k}^{n} \binom{P}{i}\binom{N-P}{n-i} \\
&\leq \binom{N}{n}^{-1} \sum_{i=0}^{n} \binom{P}{i}\binom{N-P}{n-i} \tau^{i-k} \leq \frac{(p\tau + 1 - p)^n}{\tau^k},
\end{aligned}$$

by Claim 5.4. From here the proof proceeds as in Section 3.2. □

5.3 Negative Correlations

The proof by Impagliazzo and Kabanets [15] from Section 3.3 can be used to relax the independence assumption. It now suffices that the random variables are *negatively correlated*.

Theorem 5.5. *Let $X_1, \dots, X_n$ be random variables with $X_i \in \{0, 1\}$. Suppose there exist $p_i \in [0, 1]$, $i = 1, \dots, n$, such that for every index set $I \subseteq \{1, \dots, n\}$, we have $\Pr\left[\prod_{i \in I} X_i = 1\right] \leq \prod_{i \in I} p_i$. Set $X := \sum_{i=1}^{n} X_i$ and $p := (1/n) \sum_{i=1}^{n} p_i$. Then, for any $t \in [0, 1 - p]$, we have*

$$\Pr[X \geq (p + t)n] \leq e^{-D_{\mathrm{KL}}(p+t \| p)n}.$$

Proof. Let $\lambda \in [0, 1]$ be a parameter to be chosen later. Let $I \subseteq \{1, \dots, n\}$ be a random index set obtained by including each element $i \in \{1, \dots, n\}$ with probability λ. As before, we estimate the probability $\Pr\left[\prod_{i \in I} X_i = 1\right]$ in two different ways, where the probability is over the random choice of $X_1, \dots, X_n$ and I. Similarly to before,

$$\begin{aligned}\Pr\Big[\prod_{i \in I} X_i = 1\Big] = \Pr\Big[\prod_{i \in I} X_i = 1\Big] &\leq \sum_{S \subseteq \{1,\dots,n\}} \Pr\Big[I = S \wedge \prod_{i \in S} X_i = 1\Big] \\ &\leq \sum_{S \subseteq \{1,\dots,n\}} \Pr[I = S] \cdot \Pr\Big[\prod_{i \in S} X_i = 1\Big] \leq \sum_{S \subseteq \{1,\dots,n\}} \lambda^{|S|}(1 - \lambda)^{n-|S|} \cdot \prod_{i \in S} p_i. \quad (8)\end{aligned}$$

We define n independent random variables $Z_1, \dots, Z_n$ as follows: for $i = 1, \dots, n$, with probability $1 - \lambda$, we set $Z_i = 1$, and with probability λ, we set $Z_i = p_i$. By (8), and using independence and the arithmetic-geometric mean inequality.

$$\Pr\Big[\prod_{i \in I} X_i = 1\Big] = \mathbf{E}\Big[\prod_{i=1}^{n} Z_i\Big] = \prod_{i=1}^{n} \mathbf{E}[Z_i] = \prod_{i=1}^{n} (1 - \lambda + p_i\lambda) \leq (1 - \lambda + p\lambda)^n. \quad (9)$$

The proof of the lower bound remains unchanged and yields

$$\Pr\Big[\prod_{i \in I} X_i = 1\Big] \geq (1 - \lambda)^{(1-p-t)n} \Pr[X \geq (p + t)n],$$

as before. Combining with (9) and optimizing for λ finishes the proof, see Section 3.3. □

Acknowledgments. This survey is based on lecture notes for a class on advanced algorithms at Freie Universität Berlin. I would like to thank all the students who took this class for their interest and participation. I would also like to thank Jonathan Ullman for valuable comments that improved this survey. Partially supported by ERC StG 757609.

References

[1] N. Alon and J. Spencer. *The Probabilistic Method*. Wiley-Interscience, 2016.

[2] S. Arora and B. Barak. *Computational Complexity – A Modern Approach*. Cambridge University Press, 2009.

[3] K. Azuma. Weighted sums of certain dependent random variables. *Tôhoku Math. J. (2)*, 19:357–367, 1967.

[4] S. N. Bernstein. *Sobranie Sochinenii [Collected Works]*. Nauka, Moscow, 1964.

[5] X. Chen. A likelihood ratio approach for probabilistic inequalities. `arXiv: 1308.4123`, 2013.

[6] F. R. K. Chung and L. Lu. Concentration inequalities and martingale inequalities: A survey. *Internet Mathematics*, 3(1):79–127, 2006.

[7] V. Chvátal. The tail of the hypergeometric distribution. *Discrete Mathematics*, 25(3):285–287, 1979.

[8] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press, 3rd edition, 2009.

[9] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley-Interscience, 2en edition, 2006.

[10] D. P. Dubhashi and A. Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, 2009.

[11] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[12] O. Goldreich. *Computational complexity – a conceptual perspective*. Cambridge University Press, 2008.

[13] T. Hagerup and C. Rüb. A guided tour of Chernoff bounds. *Inform. Process. Lett.*, 33(6):305–308, 1990.

[14] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

[15] R. Impagliazzo and V. Kabanets. Constructive proofs of concentration bounds. In *Proc. 13th Int. Conf. Approx. (APPROX) and 14th Int. Conf. Rand. Comb. Opt. (RANDOM)*, pages 617–631, 2010.

[16] J. M. Kleinberg and É. Tardos. *Algorithm design*. Addison-Wesley, 2006.

[17] C. McDiarmid. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, volume 16 of *Algorithms Combin.*, pages 195–248. Springer-Verlag, 1998.

[18] F. McSherry and K. Talwar. Mechanism design via differential privacy. In *Proc. 48th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, pages 94–103, 2007.

[19] M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge University Press, 2nd edition, 2017.

[20] P. Morin, W. Mulzer, and T. Reddad. Encoding arguments. *ACM Comput. Surv.*, 50(3):46:1–46:36, 2017.

[21] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

[22] T. Steinke and J. Ullman. Subgaussian tail bounds via stability arguments. `arXiv:1701.03493`, 2017.