A Simple Proof of Chernoff's Bound*

Pat Morin[1] and Wolfgang Mulzer[2]

1 **School of Computer Science, Carleton University, Canada**
`morin@scs.carleton.ca`
2 **Institut für Informatik, Freie Universität Berlin, Germany**
`mulzer@inf.fu-berlin.de`

Abstract

We present a simple proof of Chernoff's bound inspired by coding theory. The proof is elementary and does not require any calculus.

1998 ACM Subject Classification F.2.2 Nonnumerical Algorithms and Problems; G.3 Probability and Statistics

Keywords and phrases Chernoff's bound, Encoding Argument

Digital Object Identifier 10.4230/OASIcs.CVIT.2016.23

1 Introduction

Chernoff's bound gives an estimate on the probability that a sum of independent Binomial random variables deviates from its expectation [10]. It has many variants and extensions that are known under various names such as *Bernstein's inequality* or *Hoeffding's bound* [3, 10]. 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 of applications. Almost every contemporary textbook on algorithms or complexity theory contains a statement and a proof of the bound [2, 5, 8, 11], and there are several texts that discuss its various applications in great detail (see, e.g., the textbooks by Alon and Spencer [1], Dubhashi and Panchonesi [7], Mitzenmacher and Upfal [13], Motwani and Raghavan [15], or the articles by Chung and Lu [4], Hagerup and Rüb [9], or McDiarmid [12]).

We give a simple proof of Chernoff's bound that is inspired by coding theory. The proof relies on a weighted version of Markov's inequality and does not need any calculus. It is derived from ideas discussed with Luc Devroye and Gábor Lugosi at the Ninth Annual Probability, Combinatorics and Geometry Workshop, held April 4–11, 2014, at McGill University's Bellairs Institute. A broader discussion of coding theoretic arguments in theoretical computer science can be found in the survey [14].

2 The Chernoff Bound

We begin with a statement of the basic Chernoff bound. For this, we need a notion from information theory [6]. Let $p, q \in [0, 1]$. The *Kullback-Leibler divergence* or *relative entropy* of the probability distributions $(p, 1-p)$ and $(q, 1-q)$ on two elements is defined as

$$D_{\mathrm{KL}}(p\|q) := p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q}.$$

* Supported in part by DFG Grants MU 3501/1 and MU 3501/2.

42nd Conference on Very Important Topics (CVIT 2016).
Editors: John Q. Open and Joan R. Acces; Article No. 23; pp. 23:1–23:3
Open Access Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

The Kullback-Leibler divergence measures the distance between the distributions $(p, 1-p)$ and $(q, 1-q)$: it represents the expected loss of efficiency if we encode a bit string where a 0-bit has probability p and a 1-bit has probability $1-p$ with a code that is optimal for the case that a 0-bit has probability q and a 1-bit has probability $1-q$. Now, the basic Chernoff bound is as follows:

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

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

Many other, perhaps more familiar, bounds can be derived from Theorem 2.1; see the survey [16] for more details.

3 The New Proof

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, a weighted version of Markov's inequality, 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*

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

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

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

since $w(x)/s p_x \geq 1$ 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 $X_1, \dots, X_n$ 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.

Acknowledgements. The proof in this paper was derived from arguments discussed with Luc Devroye and Gábor Lugosi at the Ninth Annual Probability, Combinatorics and Geometry Workshop, held April 4–11, 2014, at McGill University's Bellairs Institute. The authors would like to thank them and all the other participants of the workshop for inspiring discussions and for providing a great research atmosphere.

References

1 Noga Alon and Joel Spencer. *The Probabilistic Method*. Wiley-Interscience, 2016.
2 Sanjeev Arora and Boaz Barak. *Computational Complexity - A Modern Approach*. Cambridge University Press, 2009.
3 Sergei Natanovich Bernstein. *Sobranie Sochinenii [Collected Works]*. Nauka, Moscow, 1964.
4 Fan R. K. Chung and Lincoln Lu. Concentration inequalities and martingale inequalities: A survey. *Internet Mathematics*, 3(1):79–127, 2006.
5 Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT Press, 3rd edition, 2009.
6 Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience, 2en edition, 2006.
7 Devdatt P. Dubhashi and Alessandro Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, 2009.
8 Oded Goldreich. *Computational complexity - a conceptual perspective*. Cambridge University Press, 2008.
9 Torben Hagerup and Christine Rüb. A guided tour of Chernoff bounds. *Inform. Process. Lett.*, 33(6):305–308, 1990.
10 Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.
11 Jon M. Kleinberg and Éva Tardos. *Algorithm design*. Addison-Wesley, 2006.
12 Colin McDiarmid. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, volume 16 of *Algorithms Combin.*, pages 195–248. Springer-Verlag, 1998.
13 Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge University Press, 2nd edition, 2017.
14 Pat Morin, Wolfgang Mulzer, and Tommy Reddad. Encoding arguments. *ACM Comput. Surv.*, 50(3):46:1–46:36, 2017.
15 Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.
16 Wolfgang Mulzer. Five proofs of Chernoff's bound with applications. `unpublished manuscript`, 2017.