Freie Universität Berlin

Masterarbeit am Institut für Informatik der Freien Universität Berlin

Bearbeitungsumfang: 15 LP

A comparative study of some proofs of Chernoff's bound with regard to their applicability for deriving concentration inequalities

Natalia Shenkman

Masterstudiengang für das Lehramt an Gymnasien

Matrikelnummer: 5040685

natalia.shenkman@fu-berlin.de

Themensteller und Erstgutachter: Prof. Dr. Wolfgang Mulzer

Zweitgutachter: PD Dr. Klaus Kriegel

Berlin, 13. Dezember 2018

Selbständigkeitserklärung

Ich erkläre ausdrücklich, dass es sich bei der vorliegenden Arbeit um eine von mir selbständig und ohne Hilfe verfasste Originalarbeit handelt.

Ich bestätige überdies, dass die Arbeit als Ganze oder in Teilen nicht zur Abgeltung anderer Studienleistungen eingereicht worden ist.

Ich erkläre ausdrücklich, dass ich sämtliche in der Arbeit enthaltenen Bezüge auf fremde Quellen als solche kenntlich gemacht habe. Insbesondere bestätige ich, dass ich nach bestem Wissen sowohl bei wörtlich übernommenen Aussagen (Zitaten) als auch bei in eigenen Worten wiedergegebenen Aussagen anderer Autorinnen oder Autoren (Paraphrasen) die Urheberschaft angegeben habe.

Ich nehme zur Kenntnis, dass Arbeiten, welche die Grundsätze der Selbständigkeitserklärung verletzen – insbesondere solche, die Zitate oder Paraphrasen ohne Herkunftsangaben enthalten –, als Plagiat betrachtet werden können.

Ich bestätige mit meiner Unterschrift die Richtigkeit dieser Angaben.

(Datum) (Natalia Shenkman)

Abstract

Concentration inequalities provide bounds on the tail of the distribution of random variables. This thesis focuses on four different techniques used in the literature to derive the most basic tail bound for sums of independent $\{0, 1\}$-valued random variables, commonly referred to as the Chernoff bound. We study their suitability for deriving tail inequalities when the independence assumption as well as the restriction on the values of the random variables are relaxed. As a by-product, we improve the bound in Theorem 3.3 in Impagliazzo and Kabanets (2010).

Zusammenfassung

Konzentrationsungleichungen schätzen die Wahrscheinlichkeit ab, mit der eine Zufallsvariable um mehr als ein vorgegebenes Maß von ihrem Erwartungswert abweicht. Im Fokus der vorliegenden Arbeit stehen vier unterschiedliche Methoden, die in der Literatur dazu verwendet wurden, die klassische Konzentrationsschranke für Summen unabhängiger Bernoulli-verteilter Zufallsvariablen, die als Chernoff Schranke bezeichnet wird, abzuleiten. Sie werden im Hinblick auf ihre Anwendbarkeit untersucht, wenn die Annahme der Unabhängigkeit und die Beschränkung des Wertebereichs der Zufallsvariablen gelockert werden. Dabei wird die Schranke im Theorem 3.3 in Impagliazzo and Kabanets (2010) verschärft.

Contents

1 | Introduction

Thanks to the weak law of large numbers, it is known that any sample average of independent random variables converges in probability towards its expected value. Concentration bounds, also commonly referred to as tail bounds, quantify the rate of decay at which the convergence takes place. There are a multitude of situations in the design and analysis of randomised algorithms (e.g., for routing in networks of parallel processors, for global wiring in gate-arrays, for distributed edge colouring of graphs) where theoretical computer scientists need to resort to such bounds. As a result, the literature abounds with research papers, textbooks as well as reviews on the topic; see Motwani and Raghavan (1995), Mitzenmacher and Upfal (2005), Dubhashi and Panconesi (2009), Alon and Spencer (2008), McDiarmid (1998), Chung and Lu (2006), and Doerr (2018) to name but a few. Very recently, Mulzer (2018) provided a survey on different techniques available in the literature to prove the basic Chernoff bound, that is:

Theorem 1.1 (**Chernoff's bound**). *Let $X_1, \ldots, X_n$ be n independent $\{0,1\}$-valued random variables such that $\mathbb{P}(X_i = 1) = p$ for all $i \in \{1, \ldots, n\}$ and some $p \in [0,1]$. Set $X := \sum_{i=1}^n X_i$. Then, for any $t \in [0, 1-p]$,*

$$\mathbb{P}\Big(X \geq (p+t)n\Big) \leq \left(\left(\frac{1-p}{1-p-t}\right)^{1-p-t}\left(\frac{p}{p+t}\right)^{p+t}\right)^n .$$

Mulzer (2018) also presented three extensions of the Chernoff bound as a means of illustrating the applicability of the aforementioned proof techniques. The purpose of this thesis is to extend his work by considering more general concentration inequalities. To do so, we relax both the independence assumption and the restriction on the values of the X_i's in Theorem 1.1 and choose (to derive the corresponding results) to focus our attention on the classical moment method as well as on three relatively new approaches based, respectively, on combinatorial arguments (see Impagliazzo and Kabanets (2010)), differential privacy analysis (see Steinke and Ullman (2017)), and convex maps (see Pelekis and Ramon (2017)). The reasons why we selected these four methods are that:

a) The classical moment method is very versatile and, as a result, a must for our study. For instance, it is notorious for its usefulness when it comes down to deriving sharp concentration bounds for random variables supported on intervals of different length (see, e.g., Theorem 4.2) or for unbounded random variables.

b) When dealing with an arbitrary dependence between random variables, the usefulness of the moment method is limited. Here, the convex maps approach is more flexible (see, e.g., Theorem 5.2) and, in some cases, leads to tighter bounds.

c) Both the combinatorial and the differential privacy approaches bring new original lines of argumentation.

The reader will find that our work, beside exposing the differences between the four methods and providing at many places a significantly higher level of details compared to what seems to be currently available in the literature, also contains some novel contributions. Most notably, as opposed to Impagliazzo and Kabanets (2010), we were able to generalize the proof of their Theorem 3.1 (which the authors view as one of the main results of the paper) without worsening the bound; see Theorem 3.1. Additionally, we show how Theorem 3.1 can be obtained via the moment and the convex maps methods; see Sections 4.2 and 5.4.

2 | Preliminaries

This chapter, while assuming only a basic knowledge of measure theory, covers most of the mathematical background which is required for our study. It gives a brief summary of several useful properties of the conditional expectation and probability, familiarizes the reader with Markov's inequality and the notion of relative entropy for Bernoulli distributions, and provides two algebraic identities, which come into play in various proofs of concentration inequalities.

2.1 Probabilistic prerequisites

We start with some fundamental facts on probability and expectation conditioned on an event or on a σ-algebra. Since the present work deals solely with random variables that are bounded and, hence, integrable and square-integrable, we limit the scope of the discussion of conditional expectation to integrable or square-integrable random variables. For a more thorough treatment, we refer the reader to the standard textbooks on probability such as Feller (1968, 1971), Jacod and Protter (2004), Çinlar (2011), or Klenke (2014).

2.1.1 Properties of conditional expectation and probability

Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a probability space, where Ω is a non-empty set, $\mathcal{F}$ is a σ-algebra on Ω, and $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$. Recall that the σ-algebra $\mathcal{F}$ is a subset of the power set of Ω which includes Ω and which is closed under complements (with respect to Ω) and under countable unions. The set Ω is referred to as the sample space, and its elements are called outcomes. The elements of $\mathcal{F}$ are called events. Proposition 2.1 lists some basic properties of the probability measure $\mathbb{P}$.

Proposition 2.1 (Properties of the probability measure). *The probability measure* $\mathbb{P}$ *on the measurable space* $(\Omega, \mathcal{F})$ *has the following properties.*

(i) Norming: $\mathbb{P}(\emptyset) = 0$ *and* $\mathbb{P}(\Omega) = 1$.

(ii) Monotonicity: If $A \subseteq B$ *for some* $A, B \in \mathcal{F}$*, then* $\mathbb{P}(A) \leq \mathbb{P}(B)$.

(iii) Boole's inequality for a finite set of events: If $A_i \in \mathcal{F}$ *for all* $i \in \mathscr{S}_n$*, then*

$$\mathbb{P}\Big(\bigcup_{i \in \mathscr{S}_n} A_i \Big) \leq \sum_{i \in \mathscr{S}_n} \mathbb{P}(A_i),$$

where, for each $n \in \mathbb{N}$*, we denote* $\mathscr{S}_n := \{1, \dots, n\}$.

Proof. Norming follows directly from the definition of a probability measure. Properties *(ii)* and *(iii)* are simple consequences of finite additivity and non-negativity of $\mathbb{P}$; see, e.g., the proof of Proposition 3.6 in (Çinlar, 2011, p. 15). □

Let $\mathbb{E}$ denote the expectation operator corresponding to $\mathbb{P}$. Further, let $X \colon \Omega \to \mathbb{R}$ be an integrable random variable defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $A \in \mathcal{F}$ be an event with $\mathbb{P}(A) > 0$. The conditional expectation $\mathbb{E}[X|A]$ of X given A is defined as

$$\mathbb{E}[X|A] := \frac{\mathbb{E}[X \mathbb{1}_A]}{\mathbb{P}(A)}. \tag{2.1}$$

It is the best estimate of $X(\omega)$ given the information that $\omega \in A$.

Proposition 2.2 (Linearity of expectation conditioned on an event). *Let* X *and* Y *be two integrable random variables on* $(\Omega, \mathcal{F}, \mathbb{P})$ *and let* $A \in \mathcal{F}$ *with* $\mathbb{P}(A) > 0$. *Then, for all* $a, b \in \mathbb{R}$*, we have*

$$\mathbb{E}[aX + bY|A] = a\mathbb{E}[X|A] + b\mathbb{E}[Y|A]. \tag{2.2}$$

In particular, $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$.

Proof. Bearing in mind that the expectation operator is an integral operator, linearity of expectation conditioned on an event follows from linearity of integration together with (2.1). Note that linearity of unconditional expectation can also be inferred from (2.2) setting $A = \Omega$. □

The concept of expectation conditioned on an event can be generalized by conditioning on a sub-σ-algebra of $\mathcal{F}$, that is, on a subset of $\mathcal{F}$ which is itself a σ-algebra. We adopt the following definition from (Çinlar, 2011, p. 140ff.); see also Chapter 23 in (Jacod and Protter, 2004, p. 192ff.).

Definition 2.3 (Conditional expectation). *Let X be a real-valued square integrable random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. The conditional expectation $\mathbb{E}[X|\mathcal{G}]$ of X given $\mathcal{G}$ is the unique (up to an almost sure equality) real-valued $\mathcal{G}$-measurable random variable Z on $(\Omega, \mathcal{F}, \mathbb{P})$ satisfying*

$$\mathbb{E}[X\mathbb{1}_A] = \mathbb{E}\left[Z\mathbb{1}_A\right] \quad \forall A \in \mathcal{G}.$$

With regard to Definition 2.3, note that the expectation of a random variable conditioned on a σ-algebra is a random variable, while the expectation of a random variable conditioned on an event is a constant. The relationship between the two notions is as follows. Let $A \in \mathcal{F}$ be an event and let $\mathcal{G} = \left\{\emptyset, A, \Omega\backslash A, \Omega\right\}$, where $\Omega\backslash A$ denotes the complement of A with respect to Ω, be the smallest σ-algebra containing A. Then, for all $\omega \in A$, we have $\mathbb{E}[X|\mathcal{G}](\omega) = \mathbb{E}[X|A]$; cf. Exercise 1.26 in (Çinlar, 2011, p. 148).

Our next proposition provides some useful properties of the conditional expectation.

Proposition 2.4 (Properties of conditional expectation). *Let X and Y be real-valued square-integrable random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. The following statements hold true.*

(i) Linearity: For $a, b \in \mathbb{R}$, we have that $\mathbb{E}[aX + bY|\mathcal{G}] = a\mathbb{E}[X|\mathcal{G}] + b\mathbb{E}[Y|\mathcal{G}]$ almost surely.

(ii) Unbiased estimation: $\mathbb{E}[\mathbb{E}[X|\mathcal{G}]] = \mathbb{E}[X]$.

(iii) Taking out what is known: If Y is $\mathcal{G}$-measurable, then $\mathbb{E}[XY|\mathcal{G}] = Y\mathbb{E}[X|\mathcal{G}]$ almost surely, and, in particular, $\mathbb{E}[Y|\mathcal{G}] = Y$ almost surely.

Proof. Properties (i) and (ii) are shown in, e.g., Theorem 23.3 c) and d) in (Jacod and Protter, 2004, p. 196f.). Property (iii) is contained in, e.g., Theorem 23.7 in (Jacod and Protter, 2004, p. 200), where X, Y, and the product XY are assumed to be integrable. In that respect, we remark that, by virtue of Hölder's inequality, square-integrability of X and Y implies integrability of X, Y, and XY. □

Remark 2.5. *In the sequel, we will encounter expectations conditioned on a random variable or on a collection of random variables. For a real-valued square-integrable random variable X and a collection of random variables $\{Y_i : i \in \mathscr{S}_n\}$ on $(\Omega, \mathcal{F}, \mathbb{P})$, the conditional expectation of X given $\{Y_i : i \in \mathscr{S}_n\}$ is defined as*

$$\mathbb{E}[X \mid Y_1, \ldots, Y_n] := \mathbb{E}[X \mid \mathcal{G}]$$

with $\mathcal{G} = \sigma\Big(\{Y_i : i \in \mathscr{S}_n\}\Big)$ being the σ-algebra generated by the set $\{Y_i : i \in \mathscr{S}_n\}$. Accordingly, all the properties listed in Proposition 2.4 hold true for expectations conditioned on random variables.

Similarly, the elementary notion of probability conditioned on an event can be extended by conditioning on a σ-algebra. Consider again the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. Then, for any $A \in \mathcal{F}$,

$$\mathbb{P}(A \mid \mathcal{G}) := \mathbb{E}[\mathbb{1}_A \mid \mathcal{G}]. \tag{2.3}$$

When $\mathcal{G} = \{\emptyset, \Omega\}$ is a trivial σ-algebra, (2.3) reduces to the basic relationship between probability and expectation, namely $\mathbb{P}(A) = \mathbb{E}[\mathbb{1}_A]$. If $\mathcal{G} = \sigma\Big(\{Y_i : i \in \mathscr{S}_n\}\Big)$ is a σ-algebra generated by a collection of random variables $\{Y_i : i \in \mathscr{S}_n\}$, then

$$\mathbb{P}(A \mid Y_1, \ldots, Y_n) := \mathbb{P}(A \mid \mathcal{G}) \tag{2.4}$$

is referred to as the conditional probability of X given $\{Y_i : i \in \mathscr{S}_n\}$. Notice that, thanks to Proposition 2.4 (*ii*), for any sub-σ-algebra $\mathcal{G}$ of $\mathcal{F}$ and any $A \in \mathcal{F}$, we have

$$\mathbb{P}(A) = \mathbb{E}[\mathbb{1}_A] = \mathbb{E}[\mathbb{E}[\mathbb{1}_A \mid \mathcal{G}]] = \mathbb{E}[\mathbb{P}(A \mid \mathcal{G})]. \tag{2.5}$$

2.1.2 Markov's inequality

Markov's inequality plays an important part in deriving concentration bounds in all but one methods considered in the subsequent chapters. It can be found in most textbooks on probability; see, e.g., Theorem 5.11 in (Klenke, 2014, p. 108).

Theorem 2.6 **(Markov's inequality).** *Let X be a non-negative random variable and let $f\colon [0,\infty) \to [0,\infty)$ be a non-decreasing function. Then, for any $t \geq 0$ such that $f(t) > 0$, we have*

$$\mathbb{P}(X \geq t) \;\leq\; \frac{1}{f(t)}\mathbb{E}[f(X)].$$

Proof. Fix an arbitrary $t \geq 0$ with $f(t) > 0$ and observe that

$$\begin{aligned}\mathbb{E}[f(X)] &= \mathbb{E}\left[f(X)\mathbb{1}_{\{f(X)\geq f(t)\}}\right] + \mathbb{E}\left[f(X)\mathbb{1}_{\{f(X)< f(t)\}}\right] \\ &\geq \mathbb{E}\left[f(X)\mathbb{1}_{\{f(X)\geq f(t)\}}\right] \;\geq\; \mathbb{E}\left[f(t)\mathbb{1}_{\{f(X)\geq f(t)\}}\right] \\ &= f(t)\mathbb{P}\Big(f(X) \geq f(t)\Big) \;=\; f(t)\mathbb{P}(X \geq t),\end{aligned}$$

where the first and the penultimate equalities are due to linearity of expectation, the first inequality is thanks to non-negativity of f, and the last equality to non-decreasingness of f. □

Remark 2.7. *Theorem 2.6 will often be used in the sequel with f as the identity function on $[0,\infty)$. In this case, it reads: For any non-negative random variable X and any $t > 0$, one has $\mathbb{P}(X \geq t) \leq t^{-1}\mathbb{E}[X]$.*

Let us emphasize that the assumption that X in Remark 2.7 (resp. f in Theorem 2.6) is non-negative is crucial, as the following example shows.

Example 2.8. Let $p \in (0,1)$. For all $i \in \mathscr{S}_n$, let $X_i \sim$ *Bernoulli*(p), that is, $\mathbb{P}(X_i = 1) = p = 1 - \mathbb{P}(X_i = 0)$. Set $X := \sum_{i=1}^{n} X_i$. Clearly, the random variable $X - pn$ can take negative values since $\mathbb{P}(X - pn = -pn) = (1-p)^n > 0$. Observe that, on the one hand, for any $t \in (0, 1-p]$, we have

$$\mathbb{P}\Big(X - pn \geq tn\Big) \;=\; \mathbb{P}\Big(X \geq (p+t)n\Big) \;\geq\; \mathbb{P}(X = n) \;=\; p^n \;>\; 0,$$

where the first inequality is due to monotonicity of the probability measure (see Proposition 2.1 (*ii*)) and to the fact that $X \leq n$ almost surely. On the other hand,

$$\frac{1}{tn}\mathbb{E}\Big[X - pn\Big] \;=\; \frac{1}{tn}\left(\sum_{i=1}^{n}\mathbb{E}[X_i] - pn\right) \;=\; \frac{1}{tn}\Big(pn - pn\Big) \;=\; 0,$$

where the first equality is due to linearity of expectation. Overall, we have

$$\mathbb{P}\Big(X - pn \geq tn\Big) \;>\; \frac{1}{tn}\mathbb{E}[X - pn]$$

for all $t \in (0, 1-p]$, which shows that Markov's inequality cannot be applied to random variables that are not non-negative.

Let us consider another example. Assume $n = 2$ and let Y be an independent copy of X. Then the random variable $\max\{X, Y\}$ takes values in $\{0, 1, 2\}$ with the probability mass function

$$\mathbb{P}\Big(\max\{X,Y\} = j\Big) = \begin{cases} (1-p)^4 & \text{if } j = 0, \\ 4p^2(1-p)^2 + 4p(1-p)^3 & \text{if } j = 1, \\ p^4 + 2p^2(1-p)^2 + 4p^3(1-p) & \text{if } j = 2, \end{cases}$$

and the expected value

$$\mathbb{E}\Big[\max\{X,Y\}\Big] = \sum_{j=0}^{2} j\,\mathbb{P}\Big(\max\{X,Y\} = j\Big) = 4p - 4p^2 + 4p^3 - 2p^4.$$

Now check that for $p = 0.5$ and $t \in (3/7, 0.5]$,

$$\mathbb{P}\Big(\max\{X,Y\} - 2p \geq 2t\Big) = \mathbb{P}\Big(\max\{X,Y\} = 2\Big) = 0.4375,$$

while

$$\frac{1}{tn}\mathbb{E}\Big[\max\{X,Y\} - pn\Big] = \frac{0.1875}{t} < 0.4375,$$

which shows once again that Markov's inequality cannot be applied to random variables that are not non-negative.

2.2 Relative entropy for Bernoulli distributions

A detailed discussion of the concept of relative entropy between two probability distributions can be found in (Cover and Thomas, 1991, p. 12ff.). We specialize the definition of the relative entropy between two probability mass functions given in (Cover and Thomas, 1991, p. 18) to the case of Bernoulli distributions.

Definition 2.9 (Relative entropy between two Bernoulli distributions). *Let $p, q \in [0,1]$ and let*

$$P(i) = \begin{cases} p & \textit{if } i = 1, \\ 1-p & \textit{if } i = 0, \end{cases} \qquad Q(i) = \begin{cases} q & \textit{if } i = 1, \\ 1-q & \textit{if } i = 0, \end{cases}$$

be two Bernoulli probability mass functions. The relative entropy of Q with respect to P, or the Kullback-Leibler divergence from Q to P, denoted by $D_{KL}(p\,\|\,q)$, is defined as

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

where we apply the conventions $0\ln 0 = 0$ (since $\lim_{x\to 0^+} x\ln x = 0$) and $\ln\frac{x}{0} = \infty$ for all $x \in (0,1]$. In particular, we have $D_{KL}(p\,\|\,p) = 0$ for all $p \in [0,1]$, $D_{KL}(p\,\|\,0) = \infty$ for all $p \in (0,1]$, and $D_{KL}(p\,\|\,1) = \infty$ for all $p \in [0,1)$.

Note that $D_{KL}(p\,\|\,q) \neq D_{KL}(q\,\|\,p)$ in general. Lemma 2.10 provides some known properties of the relative entropy, which will be useful later.

Lemma 2.10. *Let $p \in [0,1]$. The following statements hold true.*

(i) $D_{KL}(p\,\|\,q) \geq 0$ for all $q \in [0,1]$ with equality if and only if $p = q$.

(ii) $D_{KL}(p+t\,\|\,p) \geq 2t^2$ for all $t \in [0,1-p]$.

Proof. Statement (i) is Theorem 2.6.3 in (Cover and Thomas, 1991, p. 26). Statement (ii) is easily verified when $p \in \{0,1\}$ or $t = 1-p$. Assume $p \in (0,1)$ and inspect the function $f_p\colon [0,1-p) \to \bar{\mathbb{R}}$, $f_p(t) := D_{KL}(p+t\,\|\,p) - 2t^2$. Note that $f_p(0) = 0$ and, hence, that it suffices to show that f_p is non-decreasing on $[0,1-p)$. Compute

$$f_p'(t) = \ln\left(\frac{(p+t)(1-p)}{p(1-p-t)}\right) - 4t$$

and observe that $f_p'(0) = 0$. So, to finish the proof, it is enough to verify that f_p' is non-decreasing on $[0,1-p)$. And, indeed, for all $t \in [0,1-p)$, we have

$$f_p''(t) = \frac{1}{p+t} + \frac{1}{1-p-t} - 4 = \frac{\big(1-2(p+t)\big)^2}{(p+t)(1-p-t)} \;\geq\; 0.$$

Hence the lemma. □

2.3 Some identities

Lemma 2.11 below, which is a special case of the multi-binomial theorem, is always needed to prove Theorem 3.1; see Sections 3.1, 4.2, and 5.4.

Lemma 2.11. *Let $a_i, b_i \in \mathbb{R}$ for all $i \in \mathscr{S}_n$. Then*

$$\sum_{S \subseteq \mathscr{S}_n} \Big(\prod_{i \in S} a_i\Big)\Big(\prod_{j \in \mathscr{S}_n \setminus S} b_j\Big) = \prod_{i=1}^{n} (a_i + b_i). \tag{2.6}$$

Proof. The lemma is proven by induction on n. In the base case $n = 1$, it is readily seen that

$$\sum_{S \subseteq \{1\}} \Big(\prod_{i \in S} a_i\Big)\Big(\prod_{j \in \{1\} \setminus S} b_j\Big) = a_1 + b_1,$$

where we apply the usual convention that an empty product is equal to 1. For the induction step, assume that (2.6) is true for all $n \in \{1, \ldots, k\}$ and some $k \in \mathbb{N}$. Then

$$\begin{aligned}
\sum_{S \subseteq \mathscr{S}_{k+1}} \Big(\prod_{i \in S} a_i\Big)\Big(\prod_{j \in \mathscr{S}_{k+1} \setminus S} b_j\Big) &= \sum_{\substack{S \subseteq \mathscr{S}_{k+1} \\ k+1 \in S}} \Big(\prod_{i \in S} a_i\Big)\Big(\prod_{j \in \mathscr{S}_{k+1} \setminus S} b_j\Big) + \sum_{\substack{S \subseteq \mathscr{S}_{k+1} \\ k+1 \notin S}} \Big(\prod_{i \in S} a_i\Big)\Big(\prod_{j \in \mathscr{S}_{k+1} \setminus S} b_j\Big) \\
&= \sum_{S \subseteq \mathscr{S}_k} a_{k+1} \Big(\prod_{i \in S} a_i\Big)\Big(\prod_{j \in \mathscr{S}_k \setminus S} b_j\Big) + \sum_{S \subseteq \mathscr{S}_k} b_{k+1} \Big(\prod_{i \in S} a_i\Big)\Big(\prod_{j \in \mathscr{S}_k \setminus S} b_j\Big) \\
&= (a_{k+1} + b_{k+1}) \sum_{S \subseteq \mathscr{S}_k} \Big(\prod_{i \in S} a_i\Big)\Big(\prod_{j \in \mathscr{S}_k \setminus S} b_j\Big) \\
&= (a_{k+1} + b_{k+1}) \prod_{i=1}^{k} (a_i + b_i) = \prod_{i=1}^{k+1} (a_i + b_i),
\end{aligned}$$

where the penultimate equality follows from the induction hypothesis. □

In the special case where $b_i = 1 - a_i$ for all $i \in \mathscr{S}_n$, identity (2.6) can be found in (Pelekis and Ramon, 2017, Lemma 2.1).

Corollary 2.12 (Binomial identity). *Let $a, b \in \mathbb{R}$. Then*

$$\sum_{i=0}^{n} \binom{n}{i} a^i\, b^{n-i} = (a+b)^n.$$

Proof. Applying Lemma 2.11 in the penultimate equality, we obtain

$$\sum_{i=0}^{n} \binom{n}{i} a^i\, b^{n-i} = \sum_{i=0}^{n} \sum_{\substack{S \subseteq \mathscr{S}_n \\ |S| = i}} a^i\, b^{n-i} = \sum_{S \subseteq \mathscr{S}_n} a^{|S|}\, b^{n-|S|} = \prod_{i=1}^{n} (a+b) = (a+b)^n,$$

where $|S|$ denotes the cardinality of the set S. □

The following identity is relevant for the convex maps method.

Lemma 2.13. *Let $a_i \in \mathbb{R}$ for all $i \in \mathscr{S}_n$. Then*

$$\sum_{l=1}^{n} l \sum_{\substack{S \subseteq \mathscr{S}_n \\ |S| = l}} \Big(\prod_{i \in S} a_i\Big)\Big(\prod_{j \in \mathscr{S}_n \setminus S} (1 - a_j)\Big) = \sum_{i=1}^{n} a_i. \tag{2.7}$$

Proof. We follow the lines of the proof of Lemma 2.1 in Pelekis and Ramon (2017) which uses an induction on n. In the base case $n = 1$, the claim trivially holds. For the induction step, assume that (2.7) is true for all $n \in \{1, \ldots, k\}$ for some $k \in \mathbb{N}$. Then

$$\begin{aligned}
&\sum_{l=1}^{k+1} l \sum_{\substack{S\subseteq \mathscr{S}_{k+1}\\ |S|=l}} \Big(\prod_{i\in S} a_i\Big)\Big(\prod_{j\in \mathscr{S}_{k+1}\setminus S}(1-a_j)\Big)\\
&= \sum_{l=1}^{k+1} l \sum_{\substack{S\subseteq \mathscr{S}_{k+1}\\ |S|=l,\, k+1\in S}} \Big(\prod_{i\in S} a_i\Big)\Big(\prod_{j\in \mathscr{S}_{k+1}\setminus S}(1-a_j)\Big) + \sum_{l=1}^{k} l \sum_{\substack{S\subseteq \mathscr{S}_{k+1}\\ |S|=l,\, k+1\notin S}} \Big(\prod_{i\in S} a_i\Big)\Big(\prod_{j\in \mathscr{S}_{k+1}\setminus S}(1-a_j)\Big)\\
&= \sum_{l=1}^{k+1} l \sum_{\substack{S\subseteq \mathscr{S}_{k}\\ |S|=l-1}} a_{k+1}\Big(\prod_{i\in S} a_i\Big)\Big(\prod_{j\in \mathscr{S}_{k}\setminus S}(1-a_j)\Big) + \sum_{l=1}^{k} l \sum_{\substack{S\subseteq \mathscr{S}_{k}\\ |S|=l}} (1-a_{k+1})\Big(\prod_{i\in S} a_i\Big)\Big(\prod_{j\in \mathscr{S}_{k}\setminus S}(1-a_j)\Big)\\
&= a_{k+1}\sum_{l=0}^{k}(l+1) \sum_{\substack{S\subseteq \mathscr{S}_{k}\\ |S|=l}} \Big(\prod_{i\in S} a_i\Big)\Big(\prod_{j\in \mathscr{S}_{k}\setminus S}(1-a_j)\Big) + (1-a_{k+1})\sum_{l=1}^{k} l \sum_{\substack{S\subseteq \mathscr{S}_{k}\\ |S|=l}} \Big(\prod_{i\in S} a_i\Big)\Big(\prod_{j\in \mathscr{S}_{k}\setminus S}(1-a_j)\Big)\\
&= (a_{k+1}+1-a_{k+1})\sum_{l=1}^{k} l \sum_{\substack{S\subseteq \mathscr{S}_{k}\\ |S|=l}} \Big(\prod_{i\in S} a_i\Big)\Big(\prod_{j\in \mathscr{S}_{k}\setminus S}(1-a_j)\Big) + a_{k+1}\sum_{l=0}^{k} \sum_{\substack{S\subseteq \mathscr{S}_{k}\\ |S|=l}} \Big(\prod_{i\in S} a_i\Big)\Big(\prod_{j\in \mathscr{S}_{k}\setminus S}(1-a_j)\Big)\\
&= \sum_{i=1}^{k} a_i + a_{k+1} \sum_{S\subseteq \mathscr{S}_{k}} \Big(\prod_{i\in S} a_i\Big)\Big(\prod_{j\in \mathscr{S}_{k}\setminus S}(1-a_j)\Big) \;=\; \sum_{i=1}^{k} a_i + a_{k+1}\prod_{i=1}^{k}(a_i + 1 + a_i) \;=\; \sum_{i=1}^{k+1} a_i,
\end{aligned}$$

where the antepenultimate equality follows from the induction hypothesis and the penultimate equality from Lemma 2.11. This completes the proof. □

3 | A combinatorial approach to Chernoff's bound

The proof technique presented in this chapter builds upon the work of Impagliazzo and Kabanets (2010), who provided a combinatorial proof of the Chernoff-Hoeffding concentration bound for the sum of $[0, 1]$-valued random variables $X_1, \ldots, X_n$ under relaxed independence assumption, namely $\mathbb{E}\left[\prod_{i \in S} X_i\right] \leq \prod_{i \in S} c_i$ for all $S \subseteq \mathscr{S}_n$, where $c_i \in [0, 1]$ for each $i \in \mathscr{S}_n$. More precisely, relying on simple counting arguments, the authors established that there exists a universal constant $\gamma \geq 1$ such that for any $t \in [0, 1 - c]$,

$$\mathbb{P}\Big(X \geq (c + t)n\Big) \leq \gamma e^{-D_{KL}(c+t \| c)n}, \tag{3.1}$$

where $X := \sum_{i=1}^{n} X_i$ and $c := (1/n) \sum_{i=1}^{n} c_i$. Further, they proved that tail inequality (3.1) holds with $\gamma = 1$ when the X_i's are $\{0, 1\}$-valued and with $\gamma = 2$ when the X_i's take values in $[0, 1]$; see Theorem 3.1 and the proof of Theorem 3.3 in Impagliazzo and Kabanets (2010). But we will show that their line of reasoning can be modified to ensure that γ remains equal to 1 even when the X_i's are $[0, 1]$-valued; see Theorem 3.1. Subsequently, we obtain tail bounds for sums of negatively associated random variables and for martingales with bounded differences.

3.1 Tail bounds for sums of bounded variables

To prove Theorem 3.3 in Impagliazzo and Kabanets (2010), the authors, first, reduce the problem to the binary case and, then, argue by contradiction. However, avoiding the contradiction and, instead, using Lemma 2.11 together with arguments similar to those found in the proof of Theorem 3.1 in Impagliazzo and Kabanets (2010) leads to the following result.

Theorem 3.1. *Let $X_1, \ldots, X_n$ be n random variables with values in $[0,1]$. Suppose there are some constants $c_i \in [0,1]$ for each $i \in \mathscr{S}_n$ such that*

$$\mathbb{E}\Big[\prod_{i \in S} X_i\Big] \leq \prod_{i \in S} c_i \quad \forall S \subseteq \mathscr{S}_n. \tag{3.2}$$

Set $X := \sum_{i=1}^n X_i$ and $c := (1/n)\sum_{i=1}^n c_i$. Then, for any $t \in [0, 1-c]$, we have

$$\mathbb{P}\Big(X \geq (c+t)n\Big) \leq e^{-D_{KL}(c+t\|c)n}. \tag{3.3}$$

Proof. We first deal with the case where $c \in (0,1)$ and $t < 1-c$. Consider n random variables $Y_i \sim Bernoulli(X_i)$, $i \in \mathscr{S}_n$, that are conditionally independent given the random variables $X_1, \ldots, X_n$, that is, $Y_1, \ldots, Y_n$ are independent on the σ-algebra generated by the set $\{X_i : i \in \mathscr{S}_n\}$. Further, let $\lambda \in [0,1)$ be arbitrary and let I be a random variable, independent of $X_1, \ldots, X_n$ and of $Y_1, \ldots, Y_n$, taking values in $\{S : S \subseteq \mathscr{S}_n\}$ with the probability mass function

$$\mathbb{P}(I = S) = \lambda^{|S|}(1-\lambda)^{n-|S|} \quad \forall S \subseteq \mathscr{S}_n.$$

Note that the distribution of I is well-defined since

$$\sum_{S \subseteq \mathscr{S}_n} \mathbb{P}(I = S) = \sum_{i=0}^{n} \sum_{\substack{S \subseteq \mathscr{S}_n \\ |S| = i}} \lambda^i (1-\lambda)^{n-i} = \sum_{i=0}^{n} \binom{n}{i} \lambda^i (1-\lambda)^{n-i} = 1,$$

where the last equality is due to the binomial identity (see Corollary 2.12). Fix an arbitrary $t \in [0, 1-c)$. If $\mathbb{P}(X \geq (c+t)n) = 0$, then inequality (3.3) trivially holds. In the case where $\mathbb{P}(X \geq (c+t)n) > 0$, we obtain

$$\begin{aligned}\mathbb{P}\Big(\prod_{i \in I} Y_i = 1\Big) &\geq \mathbb{P}\Big(\Big\{\prod_{i \in I} Y_i = 1\Big\} \cap \Big\{X \geq (c+t)n\Big\}\Big) \\ &= \mathbb{P}\Big(\prod_{i \in I} Y_i = 1 \mid X \geq (c+t)n\Big)\mathbb{P}\Big(X \geq (c+t)n\Big),\end{aligned} \tag{3.4}$$

where the inequality is a consequence of Proposition 2.1 (*ii*). Assuming that the conditional probability in (3.4) is strictly positive (which is shown later on by (3.11)), we may write

$$\mathbb{P}\Big(X \geq (c+t)n\Big) \leq \frac{\mathbb{P}\Big(\prod_{i \in I} Y_i = 1\Big)}{\mathbb{P}\Big(\prod_{i \in I} Y_i = 1 \mid X \geq (c+t)n\Big)}. \tag{3.5}$$

Our next goal is to derive an upper bound on the numerator and a lower bound on the denominator in (3.5). To this end, observe first that for all $S \subseteq \mathscr{S}_n$, one has

$$\mathbb{P}\Big(Y_i = 1\ \forall i \in S\Big) = \mathbb{E}\Big[\mathbb{P}\Big(Y_i = 1\ \forall i \in S \mid X_1, \ldots, X_n\Big)\Big] = \mathbb{E}\Big[\prod_{i \in S} X_i\Big] \leq \prod_{i \in S} c_i, \quad (3.6)$$

where the first equality follows from (2.5). Therefore, we get

$$\begin{aligned}
\mathbb{P}\Big(\prod_{i \in I} Y_i = 1\Big) &= \mathbb{E}\Big[\mathbb{P}\Big(Y_i = 1\ \forall i \in I \mid I\Big)\Big] \\
&= \sum_{S \subseteq \mathscr{S}_n} \mathbb{P}\Big(I = S\Big)\mathbb{P}\Big(Y_i = 1\ \forall i \in I \mid I = S\Big) \\
&= \sum_{S \subseteq \mathscr{S}_n} \lambda^{|S|}(1-\lambda)^{n-|S|}\mathbb{P}\Big(Y_i = 1\ \forall i \in S\Big) \\
&\leq \sum_{S \subseteq \mathscr{S}_n} \lambda^{|S|}(1-\lambda)^{n-|S|} \prod_{i \in S} c_i = \prod_{i=1}^{n}(\lambda c_i + 1 - \lambda) \\
&\leq \left(\frac{1}{n}\sum_{i=1}^{n}(\lambda c_i + 1 - \lambda)\right)^n = (\lambda c + 1 - \lambda)^n, \qquad (3.7)
\end{aligned}$$

where the first equality is again due to (2.5), the first inequality is a consequence of (3.6) together with the fact that $\lambda \in [0, 1)$, the penultimate equality follows from Lemma 2.11 (set $a_i = \lambda c_i$ and $b_i = 1 - \lambda$ for all $i \in \mathscr{S}_n$ in (2.6)), and the last inequality is inferred from the inequality of arithmetic and geometric means.

To obtain a lower bound on the denominator in (3.5), we will first verify that

$$\lambda x + 1 - \lambda \geq (1-\lambda)^{1-x} \quad \forall \lambda \in [0, 1) \text{ and } \forall x \in [0, 1]. \qquad (3.8)$$

For this purpose, consider the function $f\colon [0, 1) \times [0, 1] \to \mathbb{R}$ given by $f(\lambda, x) := \lambda x + 1 - \lambda - (1-\lambda)^{1-x}$. Notice that f is non-decreasing in λ for each $x \in [0, 1]$ since

$$\frac{d}{d\lambda} f(\lambda, x) = (1-x)\Big((1-\lambda)^{-x} - 1\Big) \geq 0.$$

As a result, for each $\lambda \in [0, 1)$ and each $x \in [0, 1]$, we obtain

$$f(\lambda, x) \geq f(0, x) = 1 - 1^{1-x} = 0,$$

which, in turn, gives (3.8). Hence, making use of Remark 2.5 and applying Proposition 2.4 (*ii*) in the second and in the seventh steps, Proposition 2.4 (*iii*) in the third step, (2.3) together with (2.4) in the fourth, linearity of expectation (see Proposition 2.2) in the tenth, Lemma 2.11 in the eleventh, and inequality (3.8) in the twelfth, we deduce

$$\begin{aligned}
\mathbb{P}\left(\Big\{\prod_{i\in I} Y_i = 1\Big\}\cap\Big\{X \geq (c+t)n\Big\}\right) &= \mathbb{E}\Big[\mathbb{1}_{\big\{\prod_{i\in I} Y_i=1\big\}}\mathbb{1}_{\{X\geq(c+t)n\}}\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\Big[\mathbb{1}_{\big\{\prod_{i\in I} Y_i=1\big\}}\mathbb{1}_{\{X\geq(c+t)n\}} \mid X_1,\dots,X_n\Big]\Big] \\
&= \mathbb{E}\Big[\mathbb{1}_{\{X\geq(c+t)n\}}\mathbb{E}\Big[\mathbb{1}_{\big\{\prod_{i\in I} Y_i=1\big\}} \mid X_1,\dots,X_n\Big]\Big] \\
&= \mathbb{E}\Big[\mathbb{1}_{\{X\geq(c+t)n\}}\mathbb{P}\Big(\prod_{i\in I} Y_i = 1 \mid X_1,\dots,X_n\Big)\Big] \\
&= \mathbb{E}\Big[\mathbb{1}_{\{X\geq(c+t)n\}}\mathbb{P}\Big(Y_i = 1\ \forall i\in I \mid X_1,\dots,X_n\Big)\Big] = \mathbb{E}\Big[\mathbb{1}_{\{X\geq(c+t)n\}}\prod_{i\in I} X_i\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\Big[\mathbb{1}_{\{X\geq(c+t)n\}}\prod_{i\in I} X_i \mid I\Big]\Big] = \sum_{S\subseteq\mathscr{S}_n}\mathbb{P}\Big(I=S\Big)\mathbb{E}\Big[\mathbb{1}_{\{X\geq(c+t)n\}}\prod_{i\in I} X_i \mid I = S\Big] \\
&= \sum_{S\subseteq\mathscr{S}_n}\lambda^{|S|}(1-\lambda)^{n-|S|}\mathbb{E}\Big[\mathbb{1}_{\{X\geq(c+t)n\}}\prod_{i\in S} X_i\Big] \\
&= \mathbb{E}\Big[\mathbb{1}_{\{X\geq(c+t)n\}}\sum_{S\subseteq\mathscr{S}_n}\lambda^{|S|}(1-\lambda)^{n-|S|}\prod_{i\in S} X_i\Big] = \mathbb{E}\Big[\mathbb{1}_{\{X\geq(c+t)n\}}\prod_{i=1}^{n}(\lambda X_i + 1 - \lambda)\Big] \\
&\geq \mathbb{E}\Big[\mathbb{1}_{\{X\geq(c+t)n\}}\prod_{i=1}^{n}(1-\lambda)^{1-X_i}\Big] = \mathbb{E}\Big[\mathbb{1}_{\{X\geq(c+t)n\}}(1-\lambda)^{\sum_{i=1}^{n}(1-X_i)}\Big] \\
&= \mathbb{E}\Big[\mathbb{1}_{\{X\geq(c+t)n\}}(1-\lambda)^{n-X}\Big], \qquad (3.9)
\end{aligned}$$

which leads us to

$$\begin{aligned}
\mathbb{P}\left(\prod_{i\in I} Y_i = 1 \mid X \geq (c+t)n\right) &= \frac{\mathbb{P}\Big(\Big\{\prod_{i\in I} Y_i = 1\Big\}\cap\Big\{X\geq(c+t)n\Big\}\Big)}{\mathbb{P}\Big(X\geq(c+t)n\Big)} \\
&\geq \frac{\mathbb{E}\Big[\mathbb{1}_{\{X\geq(c+t)n\}}(1-\lambda)^{n-X}\Big]}{\mathbb{P}\Big(X\geq(c+t)n\Big)} = \mathbb{E}\Big[(1-\lambda)^{n-X} \mid X\geq(c+t)n\Big] \\
&\geq (1-\lambda)^{n-(c+t)n}, \qquad (3.10)
\end{aligned}$$

where the first inequality follows from (3.9) and the last equality from (2.1). In particular, since $\lambda \in [0, 1)$, computation (3.10) implies that

$$\mathbb{P}\left(\prod_{i\in I} Y_i = 1 \mid X \geq (c+t)n\right) \geq (1-\lambda)^{n-(c+t)n} > 0. \qquad (3.11)$$

Plugging the bounds (3.7) and (3.10) in (3.5) and recalling that $\lambda \in [0, 1)$ was chosen arbitrarily, yields

$$\mathbb{P}\Big(X \geq (c+t)n\Big) \leq \left(\frac{\lambda c + 1 - \lambda}{(1-\lambda)^{1-c-t}}\right)^n \qquad \forall \lambda \in [0, 1). \qquad (3.12)$$

To achieve the best possible tail bound, we determine the value of λ which minimizes the right-hand side of the inequality in (3.12). To this end, we define the function $g(\lambda) := (\lambda c + 1 - \lambda)/(1-\lambda)^{1-c-t}$ and compute its derivative

$$g'(\lambda) = \frac{(1-\lambda)^{-c-t}\Big(\lambda(1-c)(c+t) - t\Big)}{(1-\lambda)^{2(1-c-t)}}.$$

It is readily seen that g' has only one zero in the interval $[0,1)$, namely at $\lambda_* := t/((1-c)(c+t)) \in [0,1)$. A straight-forward calculation shows that

$$g(\lambda_*) \;=\; \left(\frac{c}{c+t}\right)^{c+t}\left(\frac{1-c}{1-c-t}\right)^{1-c-t} \;=\; e^{-D_{KL}(c+t\|c)} \;\leq\; 1,$$

where the inequality is due to Lemma 2.10 (i). Taking into account that $g(0) = 1$ and $\lim_{\lambda\to 1} g(\lambda) = \infty$, we conclude that λ_* is the minimum point of g on $[0,1)$. Overall, we have

$$\mathbb{P}\Big(X \geq (c+t)n\Big) \leq \min_{\lambda\in[0,1)} g^n(\lambda) = g^n(\lambda_*) = e^{-D_{KL}(c+t\|c)n}.$$

To complete the proof, it remains to consider the cases where $c = 0$ or $c = 1-t$. Observe that $c = 0$ implies $c_i = 0$ for all $i \in \mathscr{S}_n$, which, in turn, gives $X_i = 0$ almost surely for all $i \in \mathscr{S}_n$. Consequently, we have $\mathbb{P}(X \geq tn) = 1 = e^{-D_{KL}(0\|0)n}$ if $t = 0$ and $\mathbb{P}(X \geq tn) = 0 = e^{-D_{KL}(t\|0)n}$ if $t > 0$; cf. the conventions in Definition 2.9. Lastly, when $c > 0$ and $t = 1-c$, it is easy to check that

$$\begin{aligned}\mathbb{E}\Big[\prod_{i=1}^n X_i\Big] &= \mathbb{E}\Big[\Big(\mathbb{1}_{\{\prod_{i=1}^n X_i = 1\}} + \mathbb{1}_{\{\prod_{i=1}^n X_i < 1\}}\Big)\prod_{i=1}^n X_i\Big] \\ &= \mathbb{P}\Big(\prod_{i=1}^n X_i = 1\Big) + \mathbb{E}\Big[\mathbb{1}_{\{\prod_{i=1}^n X_i < 1\}}\prod_{i=1}^n X_i\Big] \;\geq\; \mathbb{P}\Big(\prod_{i=1}^n X_i = 1\Big),\end{aligned}$$

and, therefore, that

$$\begin{aligned}\mathbb{P}\Big(X \geq (c+t)n\Big) &= \mathbb{P}(X = n) = \mathbb{P}\Big(X_i = 1 \;\forall i \in \mathscr{S}_n\Big) = \mathbb{P}\Big(\prod_{i=1}^n X_i = 1\Big) \\ &\leq \mathbb{E}\Big[\prod_{i=1}^n X_i\Big] \leq \prod_{i=1}^n c_i \;\leq\; \Big(\frac{1}{n}\sum_{i=1}^n c_i\Big)^n \;=\; c^n = e^{-D_{KL}(1\|c)n},\end{aligned}$$

where the last inequality holds true since the geometric mean is at most the arithmetic mean. Hence the theorem. □

The basic Chernoff bound in Theorem 1.1 and all its extensions given in Mulzer (2018) (see Theorems 5.1, 5.2, and 5.5 therein) can readily be deduced from Theorem 3.1, as the following three corollaries show.

Corollary 3.2 (Hoeffding's extension). *Let* $X_1, \dots, X_n$ *be* n *independent random variables taking values in* $[0,1]$. *Set* $X := \sum_{i=1}^n X_i$ *and* $c := (1/n)\sum_{i=1}^n \mathbb{E}[X_i]$. *Then, for any* $t \in [0, 1-c]$,

$$\mathbb{P}\Big(X \geq (c+t)n\Big) \leq e^{-D_{KL}(c+t\|c)n}.$$

Proof. The corollary is an immediate consequence of Theorem 3.1 since, by independence of $X_1, \dots, X_n$, we have for all $S \subseteq \mathscr{S}_n$ that $\mathbb{E}\left[\prod_{i\in S} X_i\right] = \prod_{i \in S} \mathbb{E}[X_i]$. □

Corollary 3.3 (Bernoulli trials with dependence). *Let* $X_1, \dots, X_n$ *be* n *random variables with values in* $\{0,1\}$. *Suppose there are some constants* $c_i \in [0,1]$ *for all* $i \in \mathscr{S}_n$ *such that*

$$\mathbb{P}\Big(\prod_{i\in S} X_i = 1\Big) \leq \prod_{i \in S} c_i \quad \forall S \subseteq \mathscr{S}_n. \tag{3.13}$$

Set $X := \sum_{i=1}^n X_i$ *and* $c := (1/n)\sum_{i=1}^n c_i$. *Then, for any* $t \in [0, 1-c]$, *we have*

$$\mathbb{P}\Big(X \geq (c+t)n\Big) \leq e^{-D_{KL}(c+t\|c)n}.$$

Proof. The claim follows immediately from Theorem 3.1 since for all $S \subseteq \mathscr{S}_n$,

$$\mathbb{E}\Big[\prod_{i\in S} X_i\Big] = 1 \cdot \mathbb{P}\Big(\prod_{i \in S} X_i = 1\Big) + 0 \cdot \mathbb{P}\Big(\prod_{i\in S} X_i = 0\Big) \leq \prod_{i\in S} c_i.$$

□

Corollary 3.4 (Hypergeometric distribution). *Let* $N, P, n \in \mathbb{N}_0$ *with* $P, n \leq N$. *Consider the experiment of drawing at random without replacement* n *balls from an urn with* N *balls,* P *of which are red. Let the random variable* X *represent the number of red balls that results from this experiment, that is,* $X \sim Hypergeometric(N, P, n)$. *Set* $p := P/N$. *Then, for any* $t \in [0, 1-p]$,

$$\mathbb{P}\Big(X \geq (p+t)n\Big) \leq e^{-D_{KL}(p+t\|p)n}.$$

Proof. We infer the statement of the corollary from Corollary 3.3. Suppose the n balls in the experiment are drawn from the urn one by one. Introduce the random variables $X_1, \dots, X_n$, where $X_i = 1$ if the ball in the i-th draw is red, and $X_i = 0$ otherwise. Clearly, $\sum_{i=1}^n X_i$ and X are equal in distribution. Hence, to complete the proof, it suffices to establish that $X_1, \dots, X_n$ satisfy (3.13) with $c_i = p$ for all $i \in \mathscr{S}_n$. With this in mind, notice first that the distribution of $(X_1, \dots, X_n)$ is exchangeable, that is, for each permutation $\pi : \{1, \dots, n\} \to \{1, \dots, n\}$, the random vectors $(X_1, \dots, X_n)$ and $(X_{\pi(1)}, \dots, X_{\pi(n)})$ are equal in distribution. This can be seen from the fact that for any $\pi : \{1, \dots, n\} \to \{1, \dots, n\}$ and any $S \subseteq \mathscr{S}_n$,

$$\mathbb{P}\Big(\Big\{X_{\pi(i)} = 1 \;\forall i \in S\Big\} \cap \Big\{X_{\pi(j)} = 0 \;\forall j \in \mathscr{S}_n \backslash S\Big\}\Big) = \frac{\Big(\prod_{i=0}^{|S|-1}(P-i)\Big)\Big(\prod_{j=0}^{n-|S|-1}(N-P-j)\Big)}{N(N-1)\cdots(N-n+1)}.$$

Thus, we obtain for all $\emptyset \neq S \subseteq \mathscr{S}_n$ that

$$\mathbb{P}\Big(\prod_{i\in S} X_i = 1\Big) = \mathbb{P}\Big(\prod_{i=1}^{|S|} X_i = 1\Big) = \frac{P}{N} \cdot \frac{P-1}{N-1} \cdots \frac{P-|S|+1}{N-|S|+1} \leq \Big(\frac{P}{N}\Big)^{|S|} = p^{|S|},$$

where the first equality follows from exchangeability of $(X_1, \dots, X_n)$ and the inequality from $(P-i)/(N-i) \leq P/N$ for all $i = 1, \dots, n-1$, which is a consequence of $P \leq N$. Lastly, we remark that the inequality in (3.13) trivially holds when $S = \emptyset$ since an empty product is equal to the multiplicative identity 1. □

It is not an easy task to modify the proof of Theorem 3.1 in order to obtain good tail bounds for sums of random variables with inhomogeneous ranges, that is, when each X_i, $i \in \mathscr{S}_n$, takes values in an arbitrary finite interval $[a_i, b_i]$. However, for sums of random variables supported on certain finite intervals of the same length, we can show the following.

Theorem 3.5. *Let $\{a_i\}_{i\in\mathscr{S}_n}$ be a sequence of non-positive real constants and let $b \in \mathbb{R}_{>0}$. Further, let $X_1, \dots, X_n$ be n random variables such that $a_i \leq X_i \leq b + a_i$ almost surely for all $i \in \mathscr{S}_n$ and*

$$\mathbb{E}\Big[\prod_{i\in S} X_i\Big] \leq \prod_{i\in S} c_i \quad \forall S \subseteq \mathscr{S}_n$$

for some constants $c_i \in [a_i, b + a_i]$, $i \in \mathscr{S}_n$. Set $X := \sum_{i=1}^n X_i$, $a := (1/n)\sum_{i=1}^n a_i$, and $c := (1/n)\sum_{i=1}^n c_i$. Then, for any $t \in [0, b + a - c]$, we have

$$\mathbb{P}\Big(X \geq (c+t)n\Big) \leq e^{-D_{KL}\left(\frac{c-a+t}{b} \,\|\, \frac{c-a}{b}\right)n}.$$

Proof. Let us introduce the random variables $Y_i := (X_i - a_i)/b$ for all $i \in \mathscr{S}_n$ and verify that for all $S \subseteq \mathscr{S}_n$,

$$\begin{aligned}\mathbb{E}\Big[\prod_{i\in S} Y_i\Big] &= \frac{1}{b^{|S|}}\,\mathbb{E}\Big[\prod_{i\in S}(X_i - a_i)\Big] = \frac{1}{b^{|S|}}\,\mathbb{E}\Big[\sum_{I\subseteq S}\Big(\prod_{i\in I} X_i\Big)\Big(\prod_{j\in S\setminus I}(-a_j)\Big)\Big] \\ &= \frac{1}{b^{|S|}}\sum_{I\subseteq S}\Big(\prod_{j\in S\setminus I}(-a_j)\Big)\,\mathbb{E}\Big[\prod_{i\in I} X_i\Big] \le \frac{1}{b^{|S|}}\sum_{I\subseteq S}\Big(\prod_{j\in S\setminus I}(-a_j)\Big)\Big(\prod_{i\in I} c_i\Big) \\ &= \frac{1}{b^{|S|}}\prod_{i\in S}(c_i - a_i) = \prod_{i\in S}\frac{c_i - a_i}{b},\end{aligned}$$

where the second and the penultimate equalities follow from Lemma 2.11, the third equality is thanks to linearity of expectation, and the inequality uses the fact that $a_i \le 0$ for all $i \in \mathscr{S}_n$. Further, observe that for all $i \in \mathscr{S}_n$, Y_i takes values in $[0,1]$ and $(c_i - a_i)/b \in [0,1]$. Also, note that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{c_i - a_i}{b} = \frac{c-a}{b}.$$

Fix an arbitrary $t \in [0, b+a-c]$. Then $t/b \in [0, 1-(c-a)/b]$ and, hence, by application of Theorem 3.1, we obtain

$$\begin{aligned}\mathbb{P}\Big(X \ge (c+t)n\Big) &= \mathbb{P}\left(\sum_{i=1}^{n}\Big(b\,Y_i + a_i\Big) \ge (c+t)n\right) \\ &= \mathbb{P}\left(\sum_{i=1}^{n} Y_i \ge \Big(\frac{c-a}{b} + \frac{t}{b}\Big)n\right) \le e^{-D_{KL}\left(\frac{c-a+t}{b}\,\|\,\frac{c-a}{b}\right)n}.\end{aligned}$$

□

Our next corollary, which is a trivial consequence of Theorem 3.5, will soon prove itself useful, in particular, for deriving concentration inequalities for martingales.

Corollary 3.6. *Let $a, b \in \mathbb{R}$ such that $a \le 0$ and $a < b$. Further, let $X_1, \dots, X_n$ be n random variables with values in $[a,b]$ satisfying*

$$\mathbb{E}\Big[\prod_{i\in S} X_i\Big] \le \prod_{i\in S} c_i \quad \forall S \subseteq \mathscr{S}_n$$

for some constants $c_i \in [a,b]$, $i \in \mathscr{S}_n$. Set $X := \sum_{i=1}^{n} X_i$ and $c := (1/n)\sum_{i=1}^{n} c_i$. Then, for any $t \in [0, b-c]$, we have

$$\mathbb{P}\Big(X \ge (c+t)n\Big) \le e^{-D_{KL}\left(\frac{c-a+t}{b-a}\,\|\,\frac{c-a}{b-a}\right)n}.$$

Proof. The corollary follows immediately from Theorem 3.5. □

A simple application of Corollary 3.6 is Corollary 3.7. Its cruder bounds in (i) and (ii) can be found in Theorem A.1.1 and Corollary A.1.2 in (Alon and Spencer, 2008, p. 307f.), where they are derived with the help of the classical moment method.

Corollary 3.7. *Let $X_1, \ldots, X_n$ be independent random variables with $\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = 1/2$ for all $i \in \mathscr{S}_n$. Set $X := \sum_{i=1}^n X_i$. Then the following inequalities hold for any $t \in [0, n]$:*

(i) $\mathbb{P}\Big(X \geq t\Big) \leq e^{-D_{KL}\left(\frac{1}{2}+\frac{t}{2n} \,\|\, \frac{1}{2}\right)n} \leq e^{\frac{-t^2}{2n}}$.

(ii) $\mathbb{P}\Big(|X| \geq t\Big) \leq 2e^{-D_{KL}\left(\frac{1}{2}+\frac{t}{2n} \,\|\, \frac{1}{2}\right)n} \leq 2e^{\frac{-t^2}{2n}}$.

Proof. By the independence of the $X_1, \ldots, X_n$, we have

$$\mathbb{E}\Big[\prod_{i \in S} X_i\Big] = \prod_{i \in S} \mathbb{E}\big[X_i\big] = 0 \quad \forall S \subseteq \mathscr{S}_n.$$

Hence, the stronger bound in (i) follows from Corollary 3.6 with $a = -1$, $b = 2$, and $c_i = 0$ for all $i \in \mathscr{S}_n$. The other bound in (i) is due to Lemma 2.10. The inequalities in (ii) follow from (i) together with

$$\begin{aligned}\mathbb{P}\Big(|X| \geq t\Big) &= \mathbb{P}\Big(\{X \geq t\} \cup \{X \leq -t\}\Big) \leq \mathbb{P}\Big(X \geq t\Big) + \mathbb{P}\Big(X \leq -t\Big) \\ &= \mathbb{P}\Big(X \geq t\Big) + \mathbb{P}\Big(-X \geq t\Big) = 2\mathbb{P}\Big(X \geq t\Big),\end{aligned}$$

where the inequality is thanks to Boole's inequality (see Proposition 2.1 (iii)) and the last equality is due to the fact that X_i and $-X_i$ are equal in distribution for all $i \in \mathscr{S}_n$. □

3.2 Tail bounds under negative association

Negatively associated random variables (see Definition 3.8 below) constitute a subclass of random variables satisfying moment condition (3.2) in Theorem 3.1. The restriction to negatively associated random variables enables us to extend Theorem 3.5 to random variables supported on arbitrary finite intervals of the same length and to derive a lower tail bound for their sum. These results were obtained via the moment method in Dubhashi and Desh (1996); see also Janson (2016) and paragraph 3.3 in (Joag-Dev and Proschan, 1983, p. 293).

The concept of negative association for random variables is introduced in Joag-Dev and Proschan (1983), where the following definition can be found.

Definition 3.8 (**Negatively associated random variables**). *Random variables $X_1, \dots, X_n$ are said to be negatively associated if for any pair of non-empty disjoint sets $S_1, S_2 \subseteq \mathscr{S}_n$ and for any pair of real-valued non-decreasing functions f_1 and f_2,*

$$\mathbb{E}\Big[f_1(X_i,\, i \in S_1) \cdot f_2(X_j,\, j \in S_2)\Big] \;\le\; \mathbb{E}\Big[f_1(X_i,\, i \in S_1)\Big]\,\mathbb{E}\Big[f_2(X_j,\, j \in S_2)\Big]. \quad (3.14)$$

Remark 3.9. *Since $-f$ is non-decreasing if f is non-increasing, it is clear that (3.14) holds if f_1 and f_2 are both non-increasing and $X_1, \dots, X_n$ are negatively associated.*

It follows immediately from Definition 3.8 that negatively associated random variables are negatively correlated, namely

$$\mathrm{Cov}[X_i, X_j] \;=\; \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\,\mathbb{E}[X_j] \;\le\; 0$$

for all $i, j \in \mathscr{S}_n$ such that $i \neq j$, where Cov denotes the covariance operator.

Examples of negatively associated random variables include independent random variables, negatively correlated normal random variables, permutation distributions, random sampling without replacement, zero-one random variables whose sum is equal to one, and occupancy numbers in the "balls and bins" experiment; see Joag-Dev and Proschan (1983) and Dubhashi and Desh (1996).

Proposition 3.10 below provides two properties of negatively associated random variables, which are crucial for deriving concentration inequalities. Property (3.15) is stated in Joag-Dev and Proschan (1983), while (3.16) was noticed in Dubhashi and Desh (1996).

Proposition 3.10. *Let $\{b_i\}_{i \in \mathscr{S}_n}$ be a sequence of real constants. Let $X_1, \dots, X_n$ be n negatively associated random variables satisfying $X_i \le b_i$ for all $i \in \mathscr{S}_n$. Then*

$$\mathbb{E}\Big[\prod_{i \in S} g_i(X_i)\Big] \;\le\; \prod_{i \in S} \mathbb{E}[g_i(X_i)] \quad \forall S \subseteq \mathscr{S}_n, \quad (3.15)$$

where $g_1, \dots, g_n$ are arbitrary non-negative non-decreasing functions. Moreover, the random variables $\{b_i - X_i : i \in \mathscr{S}_n\}$ are negatively associated, and, in particular, we have

$$\mathbb{E}\Big[\prod_{i \in S}(b_i - X_i)\Big] \leq \prod_{i \in S} \mathbb{E}[b_i - X_i] \quad \forall S \subseteq \mathscr{S}_n. \tag{3.16}$$

Proof. Statement (3.15) is proven by induction on the cardinality of S. More precisely, (3.15) trivially holds when $|S| \in \{0, 1\}$. Assume it holds for all $S \subseteq \mathscr{S}_n$ with $|S| = k$ for some $k \in \mathbb{N}$, $k < n$. Fix any $S \subseteq \mathscr{S}_n$ with $|S| = k + 1$ and choose an arbitrary element $j \in S$. Then

$$\begin{aligned}\mathbb{E}\Big[\prod_{i \in S} g_i(X_i)\Big] &= \mathbb{E}\Big[g_j(X_j) \prod_{i \in S \setminus \{j\}} g_i(X_i)\Big] \leq \mathbb{E}[g_j(X_j)]\, \mathbb{E}\Big[\prod_{i \in S \setminus \{j\}} g_i(X_i)\Big] \\ &\leq \mathbb{E}[g_j(X_j)] \prod_{i \in S \setminus \{j\}} \mathbb{E}[g_i(X_i)] = \prod_{i \in S} \mathbb{E}[g_i(X_i)],\end{aligned}$$

where the first inequality is due to Definition 3.8 together with the fact that the product of non-negative non-decreasing functions is a non-decreasing function, and the second inequality follows from the induction hypothesis. This gives (3.15).

To justify (3.16), it suffices to show that the random variables $\{b_i - X_i : i \in \mathscr{S}_n\}$ are negatively associated, since, in this case, (3.16) readily follows from (3.15) by noticing that $b_i - X_i \geq 0$ for all $i \in \mathscr{S}_n$. To convince yourself that $\{b_i - X_i : i \in \mathscr{S}_n\}$ are negatively associated, set $g_i(x) := b_i - x$ (which is a non-increasing function on $[0, b_i]$) for each $i \in \mathscr{S}_n$, and verify that

$$\begin{aligned}\mathbb{E}\Big[f_1(b_i - X_i, i \in S_1) \cdot f_2(b_j - X_j, j \in S_2)\Big] &= \mathbb{E}\Big[f_1(g_i(X_i), i \in S_1) \cdot f_2(g_j(X_j), j \in S_2)\Big] \\ &\leq \mathbb{E}\Big[f_1(g_i(X_i), i \in S_1)\Big]\, \mathbb{E}\Big[f_2(g_j(X_j), j \in S_2)\Big] \\ &= \mathbb{E}\Big[f_1(b_i - X_i, i \in S_1)\Big]\, \mathbb{E}\Big[f_2(b_j - X_j, j \in S_2)\Big]\end{aligned}$$

for any pair of non-empty disjoint sets $S_1, S_2 \subseteq \mathscr{S}_n$ and any pair of real-valued non-decreasing functions f_1 and f_2, where the inequality follows from negative association of the $X_1, \dots, X_n$ together with Remark 3.9 and the fact that compositions of non-decreasing functions with non-increasing functions are themselves non-increasing. □

In our next theorem, we make use of Theorem 3.1 to derive upper and lower tail bounds for negatively associated random variables on arbitrary finite intervals of the same length.

Theorem 3.11 (Tail bounds under negative association). *Let $\{a_i\}_{i\in\mathscr{S}_n}$ be a sequence of real constants and let $b \in \mathbb{R}_{>0}$. Further, let $X_1, \ldots, X_n$ be n negatively associated random variables such that $a_i \le X_i \le b + a_i$ for all $i \in \mathscr{S}_n$. Set $X := \sum_{i=1}^n X_i$, $a := (1/n)\sum_{i=1}^n a_i$, and $c := (1/n)\sum_{i=1}^n \mathbb{E}[X_i]$. Then the following inequalities hold.*

(i) Upper tail bound: For any $t \in [0, b+a-c]$, we have

$$\mathbb{P}\Big(X \ge (c+t)n\Big) \;\le\; e^{-D_{KL}\left(\frac{c-a+t}{b} \,\|\, \frac{c-a}{b}\right)n} \;\le\; e^{-2t^2n/b^2}.$$

(ii) Lower tail bound: For any $t \in [0, c-a]$, we have

$$\mathbb{P}\Big(X \le (c-t)n\Big) \;\le\; e^{-D_{KL}\left(\frac{c-a-t}{b} \,\|\, \frac{c-a}{b}\right)n} \;\le\; e^{-2t^2n/b^2}.$$

Moreover, combining the bounds in (i) and (ii) yields for all $t \ge 0$

$$\mathbb{P}\Big(\,|X - cn| \ge tn\Big) \;\le\; 2e^{-2t^2n/b^2}. \tag{3.17}$$

Proof. We follow the lines of the proof of Theorem 3.5. Let $Y_i := (X_i - a_i)/b$ for each $i \in \mathscr{S}_n$. Observe that for all $S \subseteq \mathscr{S}_n$,

$$\mathbb{E}\Big[\prod_{i\in S} Y_i\Big] \;\le\; \prod_{i\in S}\mathbb{E}[Y_i] \;=\; \prod_{i\in S}\left(\frac{\mathbb{E}[X_i]-a_i}{b}\right),$$

where the inequality follows from (3.15) in Proposition 3.10 together with the fact that $Y_i = (X_i - a_i)/b \ge 0$ for all $i \in \mathscr{S}_n$, and the last equality is thanks to linearity of expectation. Also, note that the Y_i's take values in $[0, 1]$ and that

$$\frac{1}{n}\sum_{i=1}^n \frac{\mathbb{E}[X_i]-a_i}{b} = \frac{c-a}{b}.$$

Fix an arbitrary $t \in [0, b+a-c]$. Then $t/b \in [0, 1-(c-a)/b]$, and, hence, by virtue of Theorem 3.1, we obtain

$$\begin{aligned}\mathbb{P}\Big(X \ge (c+t)n\Big) &= \mathbb{P}\left(\sum_{i=1}^n \Big(b\,Y_i + a_i\Big) \ge (c+t)n\right)\\ &= \mathbb{P}\left(\sum_{i=1}^n Y_i \ge \Big(\frac{c-a}{b} + \frac{t}{b}\Big)n\right) \le e^{-D_{KL}\left(\frac{c-a+t}{b} \,\|\, \frac{c-a}{b}\right)n} \;\le\; e^{-2t^2n/b^2},\end{aligned}$$

where the last inequality is a consequence of Lemma 2.10 (ii). This gives (i).

The proof of (ii) is similar. Let $Z_i := (b + a_i - X_i)/b$ for each $i \in \mathscr{S}_n$. Check that for all $S \subseteq \mathscr{S}_n$,

$$\begin{aligned}\mathbb{E}\Big[\prod_{i\in S} Z_i\Big] &= \frac{1}{b^{|S|}}\,\mathbb{E}\Big[\prod_{i\in S}(b + a_i - X_i)\Big] \\ &\le \frac{1}{b^{|S|}}\prod_{i\in S}\mathbb{E}[b + a_i - X_i] = \prod_{i\in S}\left(\frac{b + a_i - \mathbb{E}[X_i]}{b}\right),\end{aligned}$$

where the inequality follows from (3.16) in Proposition 3.10 and the last equality from linearity of expectation. Notice that the Z_i's take values in $[0, 1]$ and that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{b + a_i - \mathbb{E}[X_i]}{b} = \frac{b + a - c}{b}.$$

Fix an arbitrary $t \in [0, c - a]$. Then $t/b \in [0, 1-(b+a-c)/b]$, and, thus, in view of Theorem 3.1, we have

$$\begin{aligned}\mathbb{P}\Big(X \le (c - t)n\Big) &= \mathbb{P}\left(\sum_{i=1}^{n} Z_i \ge \Big(\frac{b + a - c}{b} + \frac{t}{b}\Big)n\right) \\ &\le e^{-D_{KL}\left(\frac{b+a-c+t}{b}\,\|\,\frac{b+a-c}{b}\right)n} \le e^{-2t^2n/b^2},\end{aligned}$$

where the last inequality is a consequence of Lemma 2.10 (ii).

Finally, notice that the cruder bounds in (i) and (ii) hold for any $t \ge 0$ since $\mathbb{P}\Big(X \ge (c+t)n\Big) = 0$ for all $t > b+a-c$ and $\mathbb{P}\Big(X \le (c-t)n\Big) = 0$ for all $t > c-a$. This, together with

$$\mathbb{P}\Big(\,|X - cn| \ge tn\Big) \le \mathbb{P}\Big(X \ge (c + t)n\Big) + \mathbb{P}\Big(X \le (c - t)n\Big),$$

which is due to Boole's inequality (see Proposition 2.1 (iii)), imply (3.17). □

3.3 Inequalities for martingales

Impagliazzo and Kabanets (2010) observed that concentration inequality (3.1) leads to a weaker version of Azuma's inequality for martingales with uniformly bounded differences. We show that, in fact, it implies a stronger result, namely an inequality due to McDiarmid (1989); see Theorem 3.13 below. And, obviously, thanks to Theorem 3.1, we obtain bounds that are better than those of Impagliazzo and Kabanets.

For the convenience of the reader, we first recall the definition of a martingale as well as of a super- and submartingale.

Definition 3.12 (**Discrete (super-, sub-)martingale**). *A sequence $X_0, X_1, \ldots, X_n$ of integrable random variables, that is, $\mathbb{E}[|X_i|] < \infty$ for all $i \in \{0, 1, \ldots, n\}$, is called a supermartingale if*

$$\mathbb{E}\Big[X_i \mid X_0, X_1 \ldots, X_{i-1}\Big] \le X_{i-1} \ a.s. \quad \forall i \in \mathscr{S}_n. \tag{3.18}$$

It is called a submartingale if (3.18) holds with "$\ge$" instead of "$\le$" and a martingale if it is both a super- and a submartingale, i.e., if (3.18) holds with "$=$" instead of "$\le$".

The next theorem, deduced from Corollary 3.6, is a simple generalization of Theorem 6.1 in McDiarmid (1989), where a concentration inequality is obtained for martingales with differences bounded by $-p_i \le X_i - X_{i-1} \le 1 - p_i$ for some $p_i \in (0, 1)$; see also Lemma 3.11 and Theorem 3.12 in McDiarmid (1998).

Theorem 3.13. *Let $\{c_i\}_{i \in \mathscr{S}_n}$ be a sequence of non-negative real constants, $a \in \mathbb{R}$, and $b \in \mathbb{R}_{>0}$ such that $b \ge \max_{i \in \mathscr{S}_n} c_i$. Further, let $a = X_0, X_1, \ldots, X_n$ be a sequence of random variables satisfying $-c_i \le X_i - X_{i-1} \le b - c_i$ almost surely for all $i \in \mathscr{S}_n$. Set $c := (1/n) \sum_{i=1}^n c_i$. Then the following statements hold.*

(i) If $\{X_i\}_{i=0}^n$ is a supermartingale, then for all $t \in [0, (b-c)n]$,

$$\mathbb{P}\Big(X_n - X_0 \ge t\Big) \le e^{-D_{KL}\left(\frac{c}{b} + \frac{t}{nb} \,\|\, \frac{c}{b}\right)n} \le e^{\frac{-2t^2}{nb^2}}.$$

(ii) If $\{X_i\}_{i=0}^n$ is a submartingale, then for all $t \in [0, cn]$,

$$\mathbb{P}\Big(X_n - X_0 \le -t\Big) \le e^{-D_{KL}\left(\frac{c}{b} + \frac{t}{nb} \,\|\, \frac{c}{b}\right)n} \le e^{\frac{-2t^2}{nb^2}}.$$

In addition, if $\{X_i\}_{i=0}^n$ is a martingale, then, for all $t \ge 0$, we have that

$$\mathbb{P}\Big(|X_n - X_0| \ge t\Big) \le 2e^{\frac{-2t^2}{nb^2}}. \tag{3.19}$$

Proof. We will use ideas contained in the proof of Theorem 3.4 in Impagliazzo and Kabanets (2010). Let us first show (i). Define the sequence of random variables $\{Y_i\}_{i \in \mathscr{S}_n}$

via $Y_i := X_i - X_{i-1} + c_i$ for all $i \in \mathscr{S}_n$. Then observe that for all $i \in \mathscr{S}_n$,

$$\begin{aligned}\mathbb{E}\Big[Y_i \mid X_0, \dots, X_{i-1}\Big] &= \mathbb{E}\Big[X_i \mid X_0, X_1 \dots, X_{i-1}\Big] - X_{i-1} + c_i \\ &\leq X_{i-1} - X_{i-1} + c_i = c_i \end{aligned} \tag{3.20}$$

almost surely, where the first equality is due to linearity of conditional expectation (see Proposition 2.4 (i) and Remark 2.5) and the inequality to (3.18). From this, we infer by induction on the cardinality of S that

$$\mathbb{E}\Big[\prod_{i \in S} Y_i\Big] \leq \prod_{i \in S} c_i \quad \forall S \subseteq \mathscr{S}_n. \tag{3.21}$$

More precisely, if $|S| = 0$, that is, $S = \emptyset$, then inequality (3.21) is trivially satisfied. For the induction step, assume that it is satisfied for all $S \subseteq \mathscr{S}_n$ with $|S| \leq k$ for some $0 \leq k < n$. Choose an arbitrary $S \subseteq \mathscr{S}_n$ with $|S| = k+1$ and let j be its largest element. Then

$$\begin{aligned}\mathbb{E}\Big[\prod_{i \in S} Y_i\Big] &= \mathbb{E}\Big[\mathbb{E}\Big[\prod_{i \in S} Y_i \mid X_0, X_1 \dots, X_{j-1}\Big]\Big] = \mathbb{E}\Big[\mathbb{E}\Big[Y_j \prod_{i \in S\setminus\{j\}} Y_i \mid X_0, X_1 \dots, X_{j-1}\Big]\Big] \\ &= \mathbb{E}\Big[\prod_{i \in S\setminus\{j\}} Y_i \, \mathbb{E}\Big[Y_j \mid X_0, X_1 \dots, X_{j-1}\Big]\Big] \leq c_j \, \mathbb{E}\Big[\prod_{i \in S\setminus\{j\}} Y_i\Big] \leq c_j \prod_{i \in S\setminus\{j\}} c_i = \prod_{i \in S} c_i,\end{aligned}$$

where the first equality follows from Proposition 2.4 (ii), the third from Proposition 2.4 (iii), the first inequality from linearity of expectation together with (3.20) and the fact that Y_i's are almost surely non-negative, and the last inequality holds by induction hypothesis. In passing, we remark that since X_0 is a constant random variable, it does not play a role in the conditioning on the collection of random variables $\{X_0, X_1 \dots, X_i\}$. Now, we obtain for all $t \in [0, (b-c)n]$ that

$$\begin{aligned}\mathbb{P}\Big(X_n - X_0 \geq t\Big) &= \mathbb{P}\left(\sum_{i=1}^{n} (X_i - X_{i-1}) \geq t\right) = \mathbb{P}\left(\sum_{i=1}^{n} Y_i \geq \Big(c + t/n\Big)n\right) \\ &\leq e^{-D_{KL}\left(\frac{c+t/n}{b} \,\|\, \frac{c}{b}\right)n} = e^{-D_{KL}\left(\frac{c}{b} + \frac{t}{nb} \,\|\, \frac{c}{b}\right)n} \leq e^{\frac{-2t^2}{nb^2}},\end{aligned}$$

where the first inequality follows by application of Corollary 3.6 bearing in mind that Y_i's take values in $[0, b]$, and the last inequality is due to Lemma 2.10. This gives (i).

Statement (ii) readily follows from (i) by noticing that $\{X_i\}_{i=0}^n$ is a submartingale if and only if $\{-X_i\}_{i=0}^n$ is a supermartingale and that $-(b - c_i) \leq -X_i - (-X_{i-1}) \leq b - (b - c_i)$ almost surely for all $i \in \mathscr{S}_n$.

Finally, if $\{X_n\}_{i=0}^n$ is a martingale, then (i) combined with (ii) imply that (3.19) holds for all $t \geq 0$. More precisely, observe that the probability in (i) is equal to zero for all $t > (b-c)n$ because

$$X_n - X_0 \;=\; \sum_{i=1}^{n} X_i - X_{i-1} \;\leq\; \sum_{i=1}^{n}(b - c_i) \;=\; (b-c)n$$

almost surely, and, analogously, the probability in (ii) is zero for all $t > cn$ because $X_n - X_0 \geq -\sum_{i=1}^n c_i = -cn$ almost surely. Hence, using Boole's inequality (see Proposition 2.1 (iii)), we conclude that

$$\mathbb{P}\Big(|X_n - X_0| \geq t\Big) \;\leq\; \mathbb{P}\Big(X_n - X_0 \geq t\Big) + \mathbb{P}\Big(X_n - X_0 \leq -t\Big) \;\leq\; 2e^{\frac{-2t^2}{nb^2}}\,.$$

This completes the proof. □

As an immediate consequence of Theorem 3.13, we obtain Azuma's inequality for martingales with uniformly bounded differences; see Azuma (1967), Theorem 7.2.1 and Corollary 7.2.2 in (Alon and Spencer, 2008, p. 99f.), and Theorem 4.16 and Corollary 4.17 in (Motwani and Raghavan, 1995, p. 92).

Corollary 3.14. *Let $a \in \mathbb{R}$ and $c \in \mathbb{R}_{>0}$. Let $a = X_0, X_1, \ldots, X_n$ be a sequence of random variables satisfying $|X_i - X_{i-1}| \leq c$ almost surely for all $i \in \mathscr{S}_n$. Then, for all $t \in [0, cn]$, we have*

(i) $\mathbb{P}\Big(X_n - X_0 \geq t\Big) \leq e^{-D_{KL}\left(\frac{1}{2}+\frac{t}{2cn}\,\|\,\frac{1}{2}\right)n} \leq e^{\frac{-t^2}{2nc^2}}$ *if $\{X_i\}_{i=0}^n$ is a supermartingale,*

(ii) $\mathbb{P}\Big(X_n - X_0 \leq -t\Big) \leq e^{-D_{KL}\left(\frac{1}{2}+\frac{t}{2cn}\,\|\,\frac{1}{2}\right)n} \leq e^{\frac{-t^2}{2nc^2}}$ *if $\{X_i\}_{i=0}^n$ is a submartingale.*

In addition, if $\{X_i\}_{i=0}^n$ is a martingale, then, for all $t \in [0, cn]$, we have

$$\mathbb{P}\Big(|X_n - X_0| \geq t\Big) \;\leq\; 2e^{-D_{KL}\left(\frac{1}{2}+\frac{t}{2cn}\,\|\,\frac{1}{2}\right)n} \;\leq\; 2e^{\frac{-t^2}{2nc^2}}\,. \tag{3.22}$$

Proof. Statements (i) and (ii) are easy consequences of Theorem 3.13. If $\{X_n\}_{i=0}^n$ is a martingale, then (3.22) follows from (i) combined with (ii) since by Boole's inequality (see Proposition 2.1 (iii)),

$$\mathbb{P}\Big(|X_n - X_0| \geq t\Big) \;\leq\; \mathbb{P}\Big(X_n - X_0 \geq t\Big) + \mathbb{P}\Big(X_n - X_0 \leq -t\Big).$$

Hence the corollary. □

In Corollary 3.14, we only considered the case where $t \in [0, cn]$. Let us note that by repeated application of the triangular inequality and using the fact that $|X_i - X_{i-1}| \leq c$ almost surely for all $i \in \mathscr{S}_n$, it is readily seen that

$$|X_n - X_0| = \left| \sum_{i=1}^{n} (X_i - X_{i-1}) \right| \leq \sum_{i=1}^{n} |X_i - X_{i-1}| \leq cn$$

almost surely. This means that the tail probabilities in (*i*), (*ii*), and (3.22) in Corollary 3.14 vanish when $t > cn$.

The application of Azuma's inequality, commonly known as the method of bounded differences, is widespread in the literature. For instance, its usefulness for deriving concentration inequalities for the chromatic number of random graphs is discussed in Chapter 4.4 in (Motwani and Raghavan, 1995, p. 83ff.), Chapter 7 in (Alon and Spencer, 2008, p. 97ff.), and Chapter 3.1 in McDiarmid (1998).

4 | The moment method

The classical moment method, thanks to its versatile character, has been extensively used in the literature for deriving concentration inequalities; see some of the most prominent examples in Chernoff (1952), Bennett (1962), Hoeffding (1963), and Bernstein (1964). According to Hoeffding (1963), its origin can be traced back to S. N. Bernstein. The idea behind this approach is simple and can be outlined as follows. Consider a sequence of real-valued random variables $X_1, \ldots, X_n$ and set $X := \sum_{i=1}^n X_i$. To obtain an upper bound on the upper tail probability $\mathbb{P}\Big(X - \mathbb{E}[X] \geq t\Big)$ for an arbitrary $t \geq 0$, let $\lambda > 0$ be arbitrary and compute first

$$\mathbb{P}\Big(X - \mathbb{E}[X] \geq t\Big) = \mathbb{P}\Big(e^{\lambda X} \geq e^{\lambda(\mathbb{E}[X]+t)}\Big) \leq e^{-\lambda(\mathbb{E}[X]+t)}\, \mathbb{E}\left[e^{\lambda X}\right], \tag{4.1}$$

where the inequality follows by application of Markov's inequality (see Remark 2.7). Then, provided that the moment generating function $\mathbb{E}\left[e^{\lambda X}\right]$ is finite in a right neighbourhood of the origin, minimize the right-hand side of (4.1) over $\lambda > 0$, if necessary having substituted $\mathbb{E}\left[e^{\lambda X}\right]$ for a convenient upper bound, which, admittedly, may in some cases prove itself difficult[1].

4.1 Tail bounds for sums of bounded variables

Aiming to derive tail bounds for sums of random variables on intervals of different length, we provide an upper bound on the moment generating function of a random variable supported on an arbitrary finite interval and centered around zero, which we extracted from Hoeffding (1963).

[1]For instance, bounds for independent geometric or exponential random variables that are not necessarily identically distributed appeared in the literature only very recently, presumably due to the intricacy of estimating the moment generating function; see Janson (2018).

Lemma 4.1. *Let X be a random variable satisfying $\mathbb{E}[X] = 0$ and $a \leq X \leq b$ almost surely for some $a, b \in \mathbb{R}$ such that $a < b$. Then, for any $\lambda > 0$,*

$$\mathbb{E}\Big[e^{\lambda X}\Big] \leq e^{\lambda^2 (b-a)^2/8}.$$

Proof. Fix an arbitrary $\lambda > 0$. Observe that $(x-a)/(b-a) \in [0,1]$ for each $x \in [a,b]$ and that

$$x = \frac{x-a}{b-a} b + \Big(1 - \frac{x-a}{b-a}\Big) a.$$

Using convexity of the function $f(x) := e^{\lambda x}$, we obtain for all $x \in [a,b]$ that

$$\begin{aligned} e^{\lambda x} &= f\left(\frac{x-a}{b-a} b + \Big(1 - \frac{x-a}{b-a}\Big) a\right) \\ &\leq \frac{x-a}{b-a} f(b) + \Big(1 - \frac{x-a}{b-a}\Big) f(a) \;=\; \frac{x-a}{b-a} e^{\lambda b} + \frac{b-x}{b-a} e^{\lambda a}. \end{aligned}$$

Define $\tilde{\lambda} := \lambda(b-a)$, $p := -a/(b-a)$, and the function $g(x) := -px + \ln(1-p+pe^x)$. Note that $a \leq X \leq b$ almost surely together with $\mathbb{E}[X] = 0$ imply $a \leq 0$ and $b \geq 0$, so that $p \in [0,1]$ and g is well-defined. This gives

$$\begin{aligned} \mathbb{E}\Big[e^{\lambda X}\Big] &= \frac{\mathbb{E}[X]-a}{b-a} e^{\lambda b} + \frac{b-\mathbb{E}[X]}{b-a} e^{\lambda a} \;=\; \frac{-a}{b-a} e^{\lambda b} + \frac{b}{b-a} e^{\lambda a} \\ &= \frac{-a}{b-a} e^{\lambda(b-a)\left(1+\frac{a}{b-a}\right)} + \Big(1 + \frac{a}{b-a}\Big) e^{\lambda(b-a)\left(\frac{a}{b-a}\right)} \\ &= pe^{\tilde{\lambda}(1-p)} + (1-p)e^{-\tilde{\lambda}p} \;=\; e^{-\tilde{\lambda}p}\Big(1-p+pe^{\tilde{\lambda}}\Big) \;=\; e^{g(\tilde{\lambda})}, \end{aligned}$$

where the first equality is due to linearity of expectation (see Proposition 2.2). Next, we compute the first two derivatives of g:

$$g'(x) = -p + \frac{pe^x}{1-p+pe^x} \quad \text{and} \quad g''(x) = \frac{p(1-p)e^x}{(1-p+pe^x)^2}.$$

By Taylor's theorem, since g' is a continuous function for all $p \in [0,1]$, we may write g as a second order Taylor polynomial with Lagrange reminder. More precisely, for all $x > 0$, there exists $\xi \in (0,x)$ such that

$$g(x) \;=\; g(0) + g'(0)x + \frac{g''(\xi)}{2} x^2.$$

Check that $g(0) = 0$ and $g'(0) = 0$. Moreover, since $g''(x)$ is of the form $u(1-u)$ (where $u = pe^x/(1-p+pe^x)$) with $u(1-u) \leq 1/4$ (which follows from $-(2u-1)^2 \leq 0$), we have that $g''(x) \le 1/4$ for all $x > 0$. Therefore, we conclude that $g(x) \le x^2/8$ for all $x > 0$ and, hence,

$$\mathbb{E}\Big[e^{\lambda X}\Big] = e^{g(\tilde{\lambda})} \le e^{\tilde{\lambda}^2/8} = e^{\lambda^2(b-a)^2/8},$$

which gives the lemma. □

Theorem 4.2 below, due to Dubhashi and Desh (1996), extends Theorem 2 in Hoeffding (1963) for sums of independent random variables to negatively associated random variables. It clearly illustrates the superiority of the moment method; compare Theorem 4.2 to Theorem 3.11.

Theorem 4.2. *Let $a_i, b_i \in \mathbb{R}$ such that $a_i < b_i$ for all $i \in \mathscr{S}_n$. Further, let $X_1, \ldots, X_n$ be n negatively associated random variables satisfying $a_i \le X_i \le b_i$ almost surely for all $i \in \mathscr{S}_n$. Set $X := \sum_{i=1}^n X_i$. Then the following inequalities hold for any $t \ge 0$.*

(i) Upper tail bound: $\mathbb{P}\Big(X - \mathbb{E}[X] \ge t\Big) \le e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2}$.

(ii) Lower tail bound: $\mathbb{P}\Big(X - \mathbb{E}[X] \le -t\Big) \le e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2}$.

Moreover, combining the bounds in (i) and (ii) yields for all $t \ge 0$

$$\mathbb{P}\Big(|X - \mathbb{E}[X]| \ge t\Big) \le 2e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2}. \tag{4.2}$$

Proof. We first prove (i). Note that the inequality in (i) trivially holds when $t = 0$. Fix an arbitrary $t > 0$. Let $\lambda > 0$ be arbitrary. By Markov's inequality, we have

$$\mathbb{P}\Big(X - \mathbb{E}[X] \ge t\Big) = \mathbb{P}\Big(e^{\lambda(X - \mathbb{E}[X])} \ge e^{\lambda t}\Big) \le e^{-\lambda t}\,\mathbb{E}\Big[e^{\lambda(X - \mathbb{E}[X])}\Big].$$

Further, verify that

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right] &= e^{-\lambda \mathbb{E}[X]}\,\mathbb{E}\Big[\prod_{i=1}^n e^{\lambda X_i}\Big] \le e^{-\lambda \mathbb{E}[X]} \prod_{i=1}^n \mathbb{E}\left[e^{\lambda X_i}\right] \\
&= \prod_{i=1}^n \mathbb{E}\left[e^{\lambda(X_i - \mathbb{E}[X_i])}\right] \le \prod_{i=1}^n e^{\lambda^2(b_i - a_i)^2/8} = e^{\lambda^2 \sum_{i=1}^n (b_i - a_i)^2/8},
\end{aligned}$$

where we apply linearity of expectation in the first and in the third steps, property (3.15) in the second step, and Lemma 4.1 in the fourth step. Therefore, we obtain for all $\lambda > 0$ that

$$\mathbb{P}\Big(X - \mathbb{E}[X] \ge t\Big) \le e^{-\lambda t + \lambda^2 \sum_{i=1}^n (b_i - a_i)^2/8}.$$

Define the function $f(\lambda) := e^{-\lambda t + \lambda^2 \sum_{i=1}^n (b_i - a_i)^2/8}$. A straight-forward computation shows that

$$f'(\lambda) = \left(-t + \frac{1}{4}\lambda \sum_{i=1}^{n} (b_i - a_i)^2\right) e^{-\lambda t + \lambda^2 \sum_{i=1}^n (b_i - a_i)^2/8}$$

and that $f'' > 0$, implying that f attains its minimum at $\lambda_* := 4t / \sum_{i=1}^n (b_i - a_i)^2 > 0$. Hence,

$$\mathbb{P}\Big(X - \mathbb{E}[X] \geq t\Big) \leq \min_{\lambda \in (0,\infty)} f(\lambda) = f(\lambda_*) = e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2}, \tag{4.3}$$

which gives the upper tail bound in (i). Replacing each X_i with $b_i - X_i$, we obtain for all $t \geq 0$

$$\begin{aligned} \mathbb{P}\Big(X - \mathbb{E}[X] \leq -t\Big) &= \mathbb{P}\Big(-X + \mathbb{E}[X] \geq t\Big) \\ &= \mathbb{P}\left(\sum_{i=1}^{n} (b_i - X_i) - \mathbb{E}\left[\sum_{i=1}^{n} (b_i - X_i)\right] \geq t\right) \\ &\leq e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2}, \end{aligned} \tag{4.4}$$

where the second equality is due to linearity of expectation, and the inequality follows from (i) since $0 \leq b_i - X_i \leq b_i - a_i$ for all $i \in \mathscr{S}_n$ and since, in view of Proposition 3.10, the random variables $\Big\{b_i - X_i : i \in \mathscr{S}_n\Big\}$ are negatively associated. This gives (ii). Finally, applying Boole's inequality and combining the bounds in (4.3) and (4.4) leads to

$$\mathbb{P}\Big(|X - \mathbb{E}[X]| \geq t\Big) \leq \mathbb{P}\Big(X - \mathbb{E}[X] \geq t\Big) + \mathbb{P}\Big(X - \mathbb{E}[X] \leq -t\Big) \leq 2e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2}$$

for all $t \geq 0$. This concludes the proof. □

To extend Theorem 4.2 to random variables satisfying the moment condition

$$\mathbb{E}\Big[\prod_{i \in S} X_i\Big] \leq \prod_{i \in S} c_i \quad \forall S \subseteq \mathscr{S}_n$$

is still an open problem. However, proving Theorem 3.1 with the help of the moment method is very simple. This is the subject of the next section.

4.2 An alternative proof of Theorem 3.1

Proof of Theorem 3.1. We first tackle the case where $c \in (0,1)$ and $t \in (0, 1-c)$. Fix an arbitrary $t \in (0, 1-c)$ and let $\lambda > 0$ be arbitrary. Thanks to Markov's inequality, we have

$$\mathbb{P}\Big(X \geq (c+t)n\Big) = \mathbb{P}\Big(e^{\lambda X} \geq e^{\lambda(c+t)n}\Big) \leq e^{-\lambda(c+t)n}\,\mathbb{E}\left[e^{\lambda X}\right].$$

Using convexity of the function $f(x) := e^{\lambda x}$, we obtain for all $x \in [0,1]$ that

$$e^{\lambda x} \;=\; f\Big(x\cdot 1 + (1-x)\cdot 0\Big) \;\leq\; xf(1) + (1-x)f(0) \;=\; 1 + (e^{\lambda}-1)x. \tag{4.5}$$

From this, we infer

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda X}\right] &= \mathbb{E}\Big[\prod_{i=1}^{n} e^{\lambda X_i}\Big] \leq \mathbb{E}\Big[\prod_{i=1}^{n}\Big(1+(e^{\lambda}-1)X_i\Big)\Big] \\
&= \mathbb{E}\Big[\sum_{S\subseteq \mathscr{S}_n}(e^{\lambda}-1)^{|S|}\prod_{i\in S}X_i\Big] = \sum_{S\subseteq \mathscr{S}_n}(e^{\lambda}-1)^{|S|}\mathbb{E}\Big[\prod_{i\in S}X_i\Big] \\
&\leq \sum_{S\subseteq \mathscr{S}_n}(e^{\lambda}-1)^{|S|}\prod_{i\in S}c_i = \prod_{i=1}^{n}\Big(1+(e^{\lambda}-1)c_i\Big) \\
&\leq \left(\frac{1}{n}\sum_{i=1}^{n}\Big(1+(e^{\lambda}-1)c_i\Big)\right)^{n} = \Big(1+(e^{\lambda}-1)c\Big)^{n},
\end{aligned} \tag{4.6}$$

where in the second step we apply (4.5), in the third and in the sixth steps Lemma 2.11, in the fourth step linearity of expectation, in the fifth the inequalities in (3.2), and in the seventh the inequality of arithmetic and geometric means. Overall, we get

$$\mathbb{P}\Big(X \geq (c+t)n\Big) \;\leq\; e^{-\lambda(c+t)n}\Big(1+(e^{\lambda}-1)c\Big)^{n} \tag{4.7}$$

for any $\lambda > 0$. Next, we determine the value of λ which minimizes the right-hand side of (4.7). Consider the function $g(\lambda) := e^{-\lambda(c+t)}(ce^{\lambda}+1-c)$ and compute

$$g'(\lambda) \;=\; \frac{e^{\lambda(c+t)}\Big(e^{\lambda}c(1-c-t) - (1-c)(c+t)\Big)}{e^{2\lambda(c+t)}}.$$

Clearly, g' has only one zero in the interval $(0,\infty)$, namely at

$$\lambda_* := \ln\frac{(1-c)(c+t)}{c(1-c-t)} > 0.$$

A straight-forward calculation shows that $g(\lambda_*) = e^{-D_{KL}(c+t\|c)} < 1$, where the strict inequality is an immediate consequence of Lemma 2.10 (i) together with the fact that $t > 0$. Noting that $g(0) = 1$ and $\lim_{\lambda\to\infty} g(\lambda) = \infty$ since $t < 1 - c$, we deduce that λ_* is the minimum point of g on $(0, \infty)$. Hence, we conclude that

$$\mathbb{P}\Big(X \geq (c+t)n\Big) \leq \min_{\lambda\in(0,\infty)} g^n(\lambda) = g^n(\lambda_*) = e^{-D_{KL}(c+t\|c)n}.$$

To prove the theorem in the case where $c = 0$ or $t = 1 - c$, we use exactly the same argumentation as the one found in the last paragraph of the proof of Theorem 3.1 in Section 3.1. In the remaining case $c \in (0, 1)$ and $t = 0$, the statement of the theorem trivially holds since, by virtue of Lemma 2.10 (*i*), $e^{-D_{KL}(c\|c)n} = 1$. □

4.3 Inequalities for martingales

Thanks to Lemma 4.1, Theorem 3.13 and Corollary 3.14 can be generalized to the case where martingale differences are supported on finite intervals of different length.

Theorem 4.3. *Let $a \in \mathbb{R}$ and let $a_i, b_i \in \mathbb{R}$ be such that $a_i < b_i$ for all $i \in \mathscr{I}_n$. Further, let $a = X_0, X_1, \ldots, X_n$ be a martingale satisfying $a_i \leq X_i - X_{i-1} \leq b_i$ almost surely for all $i \in \mathscr{I}_n$. Then the following inequalities hold for any $t \geq 0$.*

(i) Upper tail bound: $\mathbb{P}\Big(X_n - X_0 \geq t\Big) \leq e^{-2t^2/\sum_{i=1}^n (b_i-a_i)^2}$.

(ii) Lower tail bound: $\mathbb{P}\Big(X_n - X_0 \leq -t\Big) \leq e^{-2t^2/\sum_{i=1}^n (b_i-a_i)^2}$.

Moreover, combining the bounds in (i) and (ii) yields for all $t \geq 0$

$$\mathbb{P}\Big(|X_n - X_0| \geq t\Big) \leq 2e^{-2t^2/\sum_{i=1}^n (b_i-a_i)^2}.$$

Proof. Our argumentation draws ideas from the proof of Theorem 4.2 as well as from the proofs in Chapter 3.4 in McDiarmid (1998) (where even more general results for martingales are provided). Let $\lambda > 0$ be arbitrary. By virtue of Markov's inequality,

$$\begin{aligned}\mathbb{P}\Big(X_n - X_0 \geq t\Big) &= \mathbb{P}\Big(\sum_{i=1}^n (X_i - X_{i-1}) \geq t\Big) \\ &= \mathbb{P}\Big(e^{\lambda\sum_{i=1}^n (X_i - X_{i-1})} \geq e^{\lambda t}\Big) \leq e^{-\lambda t}\,\mathbb{E}\Big[e^{\lambda\sum_{i=1}^n (X_i - X_{i-1})}\Big]. \qquad (4.8)\end{aligned}$$

In the proof of Lemma 4.1, we have shown that for any $x \in [\alpha, \beta]$, where α, β are real constants such that $\alpha < \beta$,

$$e^{\lambda x} \leq \frac{x-\alpha}{\beta-\alpha}e^{\lambda\beta} + \frac{\beta - x}{\beta-\alpha}e^{\lambda\alpha},$$

and that for any $\alpha \leq 0$ and $\beta \geq 0$ with $\alpha < \beta$,

$$\frac{-\alpha}{\beta-\alpha}e^{\lambda\beta} + \frac{\beta}{\beta-\alpha}e^{\lambda\alpha} \;\leq\; e^{\lambda^2(\beta-\alpha)^2/8}.$$

Using this and taking into account that for all $i \in \mathscr{S}_n$, $a_i \leq 0$ and $b_i \geq 0$ (since $a_i \leq X_i - X_{i-1} \leq b_i$ almost surely and $\{X_i\}_{i=0}^n$ is a martingale), we verify for all $i \in \mathscr{S}_n$ that

$$\begin{aligned}
&\mathbb{E}\Big[e^{\lambda(X_i-X_{i-1})} \mid X_0, X_1, \ldots, X_{i-1}\Big] \\
&\qquad \leq \mathbb{E}\Big[\frac{X_i - X_{i-1} - a_i}{b_i - a_i}e^{\lambda b_i} + \frac{b_i - (X_i - X_{i-1})}{b_i - a_i}e^{\lambda a_i} \mid X_0, X_1, \ldots, X_{i-1}\Big] \\
&\qquad = \frac{1}{b_i - a_i}e^{\lambda b_i}\Big(\mathbb{E}\big[X_i \mid X_0, X_1, \ldots, X_{i-1}\big] - X_{i-1} - a_i\Big) \\
&\qquad\quad + \frac{1}{b_i - a_i}e^{\lambda a_i}\Big(b_i - \mathbb{E}\big[X_i \mid X_0, X_1, \ldots, X_{i-1}\big] + X_{i-1}\Big) \\
&\qquad = \frac{-a_i}{b_i - a_i}e^{\lambda b_i} + \frac{b_i}{b_i - a_i}e^{\lambda a_i} \;\leq\; e^{\lambda^2(b_i-a_i)^2/8}
\end{aligned} \tag{4.9}$$

almost surely, where the first equality follows by linearity of conditional expectation and the second by the martingale property of $\{X_i\}_{i=0}^n$. Arguing iteratively and applying Proposition 2.4 (*ii*) in the first step, Proposition 2.4 (*iii*) in the second, and linearity of expecation together with (4.9) in the third, we obtain

$$\begin{aligned}
\mathbb{E}\Big[e^{\lambda\sum_{i=1}^n(X_i-X_{i-1})}\Big] &= \mathbb{E}\Big[\mathbb{E}\Big[e^{\lambda(X_n-X_{n-1})}\prod_{i=1}^{n-1}e^{\lambda(X_i-X_{i-1})} \mid X_0, X_1, \ldots, X_{n-1}\Big]\Big] \\
&= \mathbb{E}\Big[\prod_{i=1}^{n-1}e^{\lambda(X_i-X_{i-1})}\mathbb{E}\Big[e^{\lambda(X_n-X_{n-1})} \mid X_0, X_1, \ldots, X_{n-1}\Big]\Big] \\
&\leq e^{\lambda^2(b_n-a_n)^2/8}\,\mathbb{E}\Big[\prod_{i=1}^{n-1}e^{\lambda(X_i-X_{i-1})}\Big] \;\leq\; \prod_{i=1}^{n}e^{\lambda^2(b_i-a_i)^2/8}.
\end{aligned}$$

Plugging this estimate in (4.8) yields

$$\mathbb{P}\Big(X_n - X_0 \geq t\Big) \;\leq\; e^{-\lambda t + \lambda^2\sum_{i=1}^n(b_i-a_i)^2/8}. \tag{4.10}$$

We have already minimized the right-hand side of (4.10) in the proof of Theorem 4.2. Thus, bearing in mind that $\{-X_i\}_{i=0}^n$ is itself a martingale, we may complete the proof as for Theorem 4.2. □

Remark 4.4. *In the case where the random variables $X_1, \ldots, X_n$ are independent, Theorem 4.2 can be deduced from Theorem 4.3; see (McDiarmid, 1998, p. 24), where a* *similar observation was made. To convince yourself, consider n independent random variables $X_1, \ldots, X_n$ satisfying $a_i \le X_i \le b_i$ almost surely for all $i \in \mathscr{S}_n$ and set $X := \sum_{i=1}^n X_i$. Define the sequence of random variables $Y_0, Y_1, \ldots, Y_n$ via $Y_0 := 0$ and*

$$Y_i := Y_{i-1} + X_i - \mathbb{E}[X_i] \quad \forall i \in \mathscr{S}_n.$$

It is easily seen that $\{Y_i\}_{i=0}^n$ is a martingale. More precisely, thanks to linearity of conditional expectation and to the independence of the X_i's, one has for all $i \in \mathscr{S}_n$,

$$\mathbb{E}\Big[Y_i \mid Y_0, Y_1 \ldots, Y_{i-1}\Big] \;=\; Y_{i-1} + \mathbb{E}[X_i] - \mathbb{E}[X_i] \;=\; Y_{i-1}.$$

Hence, calling to mind that $a_i - \mathbb{E}[X_i] \le Y_i - Y_{i-1} \le b_i - \mathbb{E}[X_i]$ for all $i \in \mathscr{S}_n$ and that

$$Y_n - Y_0 \;=\; \sum_{i=0}^n (Y_i - Y_{i-1}) \;=\; \sum_{i=0}^n (X_i - \mathbb{E}[X_i]) \;=\; X - \mathbb{E}[X],$$

we obtain Theorem 4.2 from Theorem 4.3.

In contrast to the proof of Theorem 3.13, the proof of Theorem 4.3 cannot be easily adapted to give tail inequalities for super- and submartingales without weakening the assumption on the values of the increments $X_i - X_{i-1}$. This is the topic of the next theorem, which is a variant of the well-known Azuma's inequality; see Azuma (1967) and also Theorem 4.16 in (Motwani and Raghavan, 1995, p. 92).

Theorem 4.5 (**Azuma's inequality**). *Let $a \in \mathbb{R}$ and let $c_i \in \mathbb{R}_{>0}$ for all $i \in \mathscr{S}_n$. Let $a = X_0, X_1, \ldots, X_n$ be a sequence of random variables satisfying $|X_i - X_{i-1}| \le c_i$ almost surely for all $i \in \mathscr{S}_n$. Then, for all $t \ge 0$, we have*

(i) $\mathbb{P}\Big(X_n - X_0 \ge t\Big) \;\le\; e^{-t^2/\left(2\sum_{i=1}^n c_i^2\right)}$ *if $\{X_i\}_{i=0}^n$ is a supermartingale,*

(ii) $\mathbb{P}\Big(X_n - X_0 \le -t\Big) \;\le\; e^{-t^2/\left(2\sum_{i=1}^n c_i^2\right)}$ *if $\{X_i\}_{i=0}^n$ is a submartingale.*

In addition, if $\{X_i\}_{i=0}^n$ is a martingale, then, for all $t \ge 0$, we have

$$\mathbb{P}\Big(|X_n - X_0| \ge t\Big) \;\le\; 2e^{-t^2/\left(2\sum_{i=1}^n c_i^2\right)}. \tag{4.11}$$

Proof. The tail bound in (i) is obtained as in the proof of Theorem 4.3 with $a_i = -c_i$ and $b_i = c_i$. Thanks to this simplification, we can now exploit the supermartingale property of $\{X_i\}_{i=0}^n$ in computation (4.9). More precisely, we have

$$\begin{aligned}
&\mathbb{E}\Big[e^{\lambda(X_i - X_{i-1})} \mid X_0, X_1, \dots, X_{i-1}\Big] \\
&\quad\le \mathbb{E}\Big[\frac{X_i - X_{i-1} + c_i}{2c_i} e^{\lambda c_i} + \frac{c_i - (X_i - X_{i-1})}{2c_i} e^{-\lambda c_i} \mid X_0, X_1, \dots, X_{i-1}\Big] \\
&\quad= \Big(\frac{1}{2c_i} e^{\lambda c_i} - \frac{1}{2c_i} e^{-\lambda c_i}\Big)\Big(\mathbb{E}\big[X_i \mid X_0, X_1, \dots, X_{i-1}\big] - X_{i-1}\Big) + \frac{1}{2} e^{\lambda c_i} + \frac{1}{2} e^{-\lambda c_i} \\
&\quad\le \frac{1}{2} e^{\lambda c_i} + \frac{1}{2} e^{-\lambda c_i} \;\le\; e^{\lambda^2 (2c_i)^2/8} \;=\; e^{\lambda^2 c_i^2/2}
\end{aligned}$$

almost surely, where the first equality is due to linearity of conditional expectation and the penultimate inequality to supermartingale property of $\{X_i\}_{i=0}^n$. Statement (ii) readily follows from (i) since $\{X_i\}_{i=0}^n$ is a submartingale if and only if $\{-X_i\}_{i=0}^n$ is a supermartingale. In addition, if $\{X_i\}_{i=0}^n$ is a martingale, then the correctness of the bounds in (i), (ii), and (4.11) follows directly from Theorem 4.3 (setting $a_i = -c_i$ and $b_i = c_i$ for all $i \in \mathscr{S}_n$). $\square$

5 | The convex maps method

In the present chapter, we discuss a method for deriving concentration inequalities recently presented[1] by Pelekis and Ramon (2017). The authors take advantage of Lemma 2.13 (which allows a sum of n reals in the interval $[0,1]$ to be expressed as a convex combination of n integers in $\mathscr{S}_n$) to reduce the problem of obtaining an upper bound on a tail probability of a sum of $[0,1]$-valued random variables to the simpler problem of computing an upper bound on the expectation of a convex function of a $\{0,1,\ldots,n\}$-valued random variable. The key benefit of this technique, compared to the methods we have considered so far, is that it allows to derive concentration inequalities for sums of random variables for which no assumption is made regarding their dependence structure. In particular, this includes weakly dependent, e.g., k-wise independent, random variables.

5.1 Tail bounds for sums of bounded variables

Proposition 5.1 below is the main technical result in the work of Pelekis and Ramon (2017) and our starting point.

Proposition 5.1. *Let $X_1,\ldots,X_n$ be n random variables taking values in $[0,1]$ and let $f\colon [0,\infty)\to[0,\infty)$ be a non-decreasing and convex function. Set $X := \sum_{i=1}^n X_i$. Then, for all $t\geq 0$ such that $f(t+\mathbb{E}[X])>0$, we have*

$$\mathbb{P}\Big(X-\mathbb{E}[X]\geq t\Big) \;\leq\; \frac{1}{f(t+\mathbb{E}[X])}\mathbb{E}\left[f(Z)\right],$$

where Z is a random variable with values in $\{0,1,\ldots,n\}$ defined by the probability

[1] A similar approach was pursued in Schmidt et al. (1995).

mass function

$$\mathbb{P}(Z=j) = \sum_{\substack{S\subseteq \mathscr{S}_n\\ |S|=j}} \mathbb{E}\left[\left(\prod_{i\in S} X_i\right)\left(\prod_{l\in\mathscr{S}_n\setminus S}(1-X_l)\right)\right] \qquad \forall j\in\{0,1,\dots,n\}. \tag{5.1}$$

Before proving the proposition, let us note that (5.1) indeed defines a proper probability mass function since

$$\begin{aligned}\sum_{j=0}^{n}\mathbb{P}(Z=j) &= \sum_{S\subseteq\mathscr{S}_n}\mathbb{E}\left[\left(\prod_{i\in S} X_i\right)\left(\prod_{l\in\mathscr{S}_n\setminus S}(1-X_l)\right)\right]\\ &= \mathbb{E}\left[\sum_{S\subseteq\mathscr{S}_n}\left(\prod_{i\in S} X_i\right)\left(\prod_{l\in\mathscr{S}_n\setminus S}(1-X_l)\right)\right] = \mathbb{E}\left[\prod_{i=1}^{n}(X_i+1-X_i)\right] = 1,\end{aligned}$$

where the second equality follows by linearity of expectation and the penultimate equality by Lemma 2.11.

Proof of Proposition 5.1. We follow the lines of the proof of Theorem 1.2 in Pelekis and Ramon (2017). By virtue of Markov's inequality, we deduce for all $t\geq 0$ satisfying $f(t+\mathbb{E}[X])>0$ that

$$\mathbb{P}\Big(X-\mathbb{E}[X]\geq t\Big) \leq \frac{1}{f(t+\mathbb{E}[X])}\mathbb{E}[f(X)].$$

Noticing that

$$\sum_{j=0}^{n}\sum_{\substack{S\subseteq\mathscr{S}_n\\ |S|=j}}\left(\prod_{i\in S} X_i\right)\left(\prod_{l\in\mathscr{S}_n\setminus S}(1-X_l)\right) = 1$$

thanks to Lemma 2.11, we make use of Lemma 2.13 to write X as a convex combination of the integers in $\{0,1,\dots,n\}$, namely

$$X = \sum_{j=0}^{n} j \sum_{\substack{S\subseteq\mathscr{S}_n\\ |S|=j}}\left(\prod_{i\in S} X_i\right)\left(\prod_{l\in\mathscr{S}_n\setminus S}(1-X_l)\right).$$

Hence, we infer that

$$\begin{aligned}\mathbb{E}[f(X)] &= \mathbb{E}\left[f\left(\sum_{j=0}^{n} j\sum_{\substack{S\subseteq\mathscr{S}_n\\ |S|=j}}\left(\prod_{i\in S} X_i\right)\left(\prod_{l\in\mathscr{S}_n\setminus S}(1-X_l)\right)\right)\right]\\ &\leq \mathbb{E}\left[\sum_{j=0}^{n} f(j)\sum_{\substack{S\subseteq\mathscr{S}_n\\ |S|=j}}\left(\prod_{i\in S} X_i\right)\left(\prod_{l\in\mathscr{S}_n\setminus S}(1-X_l)\right)\right]\\ &= \sum_{j=0}^{n} f(j)\mathbb{P}(Z=j) = \mathbb{E}[f(Z)],\end{aligned}$$

where the inequality follows by application of Jensen's inequality to the convex function f, and the penultimate equality is due to linearity of expectation. $\square$

Pelekis and Ramon (2017) applied Proposition 5.1 to derive a tail bound for sums of $[0,1]$-valued random variables without any restriction on the dependence structure; see Theorem 5.2 below. This was originally obtained by Schmidt et al. (1995) with the help of certain symmetric multilinear polynomials. In the special case where $X_i \sim Bernoulli(p_i)$ for all $i \in \mathscr{S}_n$, an alternative proof of this result relying on elementary combinatorial arguments can be found in Linial and Luria (2014).

Theorem 5.2. *Let $X_1, \ldots, X_n$ be n random variables taking values in $[0,1]$. Set $X := \sum_{i=1}^n X_i$. Then, for any $t \geq 0$ and any $k \in \mathbb{N}_0$ such that $k < t + \mathbb{E}[X] + 1$, we have*

$$\mathbb{P}\Big(X - \mathbb{E}[X] \geq t\Big) \leq \frac{1}{\binom{t+\mathbb{E}[X]}{k}} \sum_{\substack{S \subseteq \mathscr{S}_n \\ |S|=k}} \mathbb{E}\left[\prod_{i \in S} X_i\right], \tag{5.2}$$

where for any $x \in \mathbb{R}_{\geq 0}$ and any $k \in \mathbb{N}$ satisfying $k < x + 1$, we denote

$$\binom{x}{k} := \frac{x(x-1)\cdots(x-k+1)}{k!} \quad \text{and} \quad \binom{x}{0} := 1.$$

Proof. We follow the lines of the proof of Theorem 3.2 in Pelekis and Ramon (2017) but fill the gap left in order to justify the convexity of the generalized binomial coefficient. Let $t \geq 0$ be arbitrary. For $k = 0$, the inequality in (5.2) trivially holds since an empty product is equal to 1. Fix an arbitrary $k \in \mathbb{N}$ such that $k < t + \mathbb{E}[X] + 1$, which is always possible whenever $t + \mathbb{E}[X] > 0$. We aim at applying Proposition 5.1 to the function $f_k \colon [0,\infty) \to [0,\infty)$ defined by[2]

$$f_k(x) = \begin{cases} 0 & \text{if } x \leq k-1, \\ \frac{x(x-1)\cdots(x-k+1)}{k!} & \text{if } x > k-1. \end{cases}$$

To this end, observe first that f_k is non-decreasing and that $f_k(t + \mathbb{E}[X]) > 0$. Moreover, f_k is convex. To justify this, it suffices to show that $f_k''(x) \geq 0$ for all

[2] To understand the reasoning behind the choice of the function f_k, we advise the interested reader to consult Section 2.1 in Schmidt et al. (1995).

$x > k-1$. And, indeed, if $x > k-1$, then

$$f_k'(x) = \frac{1}{k!}\exp\left(\sum_{i=0}^{k-1}\ln(x-i)\right)\sum_{i=0}^{k-1}\frac{1}{x-i} = f_k(x)\sum_{i=0}^{k-1}\frac{1}{x-i}$$

and

$$\begin{aligned} f_k''(x) &= f_k'(x)\sum_{i=0}^{k-1}\frac{1}{x-i} - f_k(x)\sum_{i=0}^{k-1}\frac{1}{(x-i)^2} \\ &= f_k(x)\left(\left(\sum_{i=0}^{k-1}\frac{1}{x-i}\right)^2 - \sum_{i=0}^{k-1}\frac{1}{(x-i)^2}\right) \geq 0. \end{aligned}$$

Hence, we infer from Proposition 5.1 that

$$\mathbb{P}\Big(X - \mathbb{E}[X] \geq t\Big) \leq \frac{1}{f_k\left(t + \mathbb{E}[X]\right)}\sum_{j=0}^{n} f_k(j)\mathbb{P}(Z=j),$$

where Z is a random variable taking values in $\{0,1,\ldots,n\}$ defined by the probability mass function in (5.1). Now, observe that

$$\begin{aligned} \sum_{j=0}^{n} f_k(j)\mathbb{P}(Z=j) &= \sum_{j=0}^{n} f_k(j)\sum_{\substack{S\subseteq\mathscr{S}_n\\|S|=j}}\mathbb{E}\left[\left(\prod_{i\in S}X_i\right)\left(\prod_{l\in\mathscr{S}_n\setminus S}(1-X_l)\right)\right] \\ &= \sum_{j=k}^{n}\binom{j}{k}\sum_{\substack{S\subseteq\mathscr{S}_n\\|S|=j}}\mathbb{E}\left[\left(\prod_{i\in S}X_i\right)\left(\prod_{l\in\mathscr{S}_n\setminus S}(1-X_l)\right)\right] \\ &= \sum_{\substack{T\subseteq\mathscr{S}_n\\|T|=k}}\sum_{\substack{S\subseteq\mathscr{S}_n\\S\supseteq T}}\mathbb{E}\left[\left(\prod_{i\in S}X_i\right)\left(\prod_{l\in\mathscr{S}_n\setminus S}(1-X_l)\right)\right] = \sum_{\substack{T\subseteq\mathscr{S}_n\\|T|=k}}\mathbb{E}\left[\prod_{i\in T}X_i\right], \end{aligned}$$

where the second equality follows by noticing that $f_k(j) = 0$ for all $j \in \{0,1,\ldots,k-1\}$ and $f_k(j) = \binom{j}{k}$ for all $j \in \{k, k+1, \ldots, n\}$, the penultimate equality uses the fact that there are for each $S \subseteq \mathscr{S}_n$ with cardinality $|S| = j$, $j \geq k$, precisely $\binom{j}{k}$ sets $T \subseteq S$ with $|T| = k$, and the last equality is due to linearity of expecation together with the fact that for all $T \subseteq \mathscr{S}_n$,

$$\sum_{\substack{S\subseteq\mathscr{S}_n\\S\supseteq T}}\left(\prod_{i\in S}X_i\right)\left(\prod_{l\in\mathscr{S}_n\setminus S}(1-X_l)\right) = \left(\prod_{i\in T}X_i\right)\sum_{S\subseteq\mathscr{S}_n\setminus T}\left(\prod_{i\in S}X_i\right)\left(\prod_{l\in(\mathscr{S}_n\setminus T)\setminus S}(1-X_l)\right) = \prod_{i\in T}X_i,$$

where the last equality is a consequence of Lemma 2.11. Hence the theorem. □

In the case where the random variables $X_1,\ldots,X_n$ are independent with values in $\{0,1\}$, Schmidt et al. (1995) demonstrated that Theorem 5.2 leads to a sharper bound than the one in Theorem 1.1. Further, as pointed out in Pelekis and Ramon (2017), Theorem 5.2 reduces to Markov's inequality when $k = 1$. Finally, note that replacing each X_i with $1 - X_i$ (as in the proof of Theorem 4.2) readily gives the lower tail bound

$$\mathbb{P}\Big(X - \mathbb{E}[X] \leq -t\Big) \leq \frac{1}{\binom{t+n-\mathbb{E}[X]}{k}} \sum_{\substack{S \subseteq \mathscr{I}_n \\ |S|=k}} \mathbb{E}\left[\prod_{i \in S}(1 - X_i)\right]$$

for all $t \geq 0$ and any $k \in \mathbb{N}_0$ such that $k < t + n - \mathbb{E}[X] + 1$.

5.2 Tail bounds under k-wise independence

In this section, we will use Theorem 5.2 to derive tail bounds for sums of k-wise independent random variables. For the convenience of the reader, we start with some elementary facts about k-wise independent random variables.

Definition 5.3 (k-**wise independence**). *For each $i \in \mathscr{I}_n$, let X_i be a random variable with values in $R_i \subseteq \mathbb{R}$. We refer to $X_1, \ldots, X_n$ as k-wise independent for some $k \in \mathscr{I}_n$ if for any set $S \subseteq \mathscr{I}_n$ with $|S| \leq k$ and any values $x_i \in R_i$,*

$$\mathbb{P}\left(\bigcap_{i \in S} \{X_i \leq x_i\}\right) = \prod_{i \in S} \mathbb{P}(X_i \leq x_i).$$

Some examples of k-wise independent random variables and of their applicability for derandomization of randomized algorithms can be found in, e.g., Chapter 16.2 in (Alon and Spencer, 2008, p. 280ff.) and Chapter 3.5 in (Vadhan, 2012, p. 62ff.). Any random variables $X_1, \ldots, X_n$ are, trivially, 1-wise independent, while 2-wise independence simply means that $X_1, \ldots, X_n$ are pairwise (but not necessarily mutually) independent. Also, k-wise independence implies k_*-wise independence for all $k_* \in \{1, \ldots, k\}$, and, clearly, if $X_1, \ldots, X_n$ are k-wise independent, then for all sets $S \subseteq \mathscr{I}_n$ with $|S| \leq k$,

$$\mathbb{E}\left[\prod_{i \in S} X_i\right] = \prod_{i \in S} \mathbb{E}[X_i].$$

Theorem 5.5 below, a tail inequality for sums of k-wise independent random variables, is originally due to Schmidt et al. (1995). Pelekis and Ramon (2017) infer, without proving it, that this theorem is readily deducible from Theorem 5.2. However, a closer look shows that an additional technical result from Schmidt et al. (1995) is required, namely Lemma 5.4.

Lemma 5.4. *Let $\{a_i\}_{i\in\mathscr{S}_n}$ be a sequence of non-negative real constants. Then, for any $k \in \mathscr{S}_n$, we have*

$$\sum_{\substack{S\subseteq\mathscr{S}_n\\|S|=k}} \prod_{i\in S} a_i \leq \binom{n}{k}\left(\frac{1}{n}\sum_{i=1}^{n} a_i\right)^k. \tag{5.3}$$

In particular, when $k = n$, inequality (5.3) reduces to the inequality of arithmetic and geometric means.

Proof. We simply fill the gaps in the proof of Lemma 2 in Schmidt et al. (1995). If $k = 1$, then (5.3) trivially holds. Fix an arbitrary $k \in \{2, \ldots, n\}$ and consider the function $f_k : [0,\infty)^n \to [0,\infty)$,

$$f_k(x_1,\ldots,x_n) = \sum_{\substack{S\subseteq\mathscr{S}_n\\|S|=k}} \prod_{i\in S} x_i.$$

To conclude, it suffices to show that the function $f_k(x_1,\ldots,x_n)$ is maximized at $\left(\frac{1}{n}\sum_{i=1}^n a_i,\ldots,\frac{1}{n}\sum_{i=1}^n a_i\right)$ when subject to the constraint $\sum_{i=1}^n x_i = \sum_{i=1}^n a_i$. For this purpose, suppose that $x_l < x_j$, where $l, j \in \mathscr{S}_n$, $l \neq j$, for some vector $(x_1,\ldots,x_n) \in [0,\infty)^n$ satisfying $\sum_{i=1}^n x_i = \sum_{i=1}^n a_i$. Let $\epsilon \in (0, x_j - x_l)$ be arbitrary. Define the vector $(y_1,\ldots,y_n) \in [0,\infty)^n$ via $y_l = x_l + \epsilon$, $y_j = x_j - \epsilon$, and $y_i = x_i$ for all $i \in \mathscr{S}_n\backslash\{l,j\}$. Clearly, $\sum_{i=1}^n y_i = \sum_{i=1}^n x_i = \sum_{i=1}^n a_i$. Moreover, we have

$$\begin{aligned}
f_k(y_1,\ldots,y_n) &= \sum_{\substack{S\subseteq\mathscr{S}_n\\|S|=k;\, l,j\notin S}} \prod_{i\in S} x_i + (x_l+\epsilon)\sum_{\substack{S\subseteq\mathscr{S}_n\\|S|=k;\, l\in S,\, j\notin S}} \prod_{i\in S\backslash\{l\}} x_i + (x_j-\epsilon)\sum_{\substack{S\subseteq\mathscr{S}_n\\|S|=k;\, l\notin S,\, j\in S}} \prod_{i\in S\backslash\{j\}} x_i \\
&\quad + (x_l+\epsilon)(x_j-\epsilon)\sum_{\substack{S\subseteq\mathscr{S}_n\\|S|=k;\, l,j\in S}} \prod_{i\in S\backslash\{l,j\}} x_i \\
&= f_k(x_1,\ldots,x_n) + \epsilon\sum_{\substack{S\subseteq\mathscr{S}_n\\|S|=k-1;\, l,j\notin S}} \prod_{i\in S} x_i - \epsilon\sum_{\substack{S\subseteq\mathscr{S}_n\\|S|=k-1;\, l,j\notin S}} \prod_{i\in S} x_i \\
&\quad + \left(\epsilon x_j - \epsilon x_l - \epsilon^2\right)\sum_{\substack{S\subseteq\mathscr{S}_n\\|S|=k-2;\, l,j\notin S}} \prod_{i\in S} x_i \\
&= f_k(x_1,\ldots,x_n) + \left((x_j-x_l)\epsilon - \epsilon^2\right)\sum_{\substack{S\subseteq\mathscr{S}_n\\|S|=k-2;\, l,j\notin S}} \prod_{i\in S} x_i \geq f_k(x_1,\ldots,x_n),
\end{aligned}$$

where the inequality follows from $x_j - x_l > \epsilon$. Repeating the argument with $\epsilon = (x_j - x_l)/2$ leads us to the conclusion that $\left(\frac{1}{n}\sum_{i=1}^n a_i, \ldots, \frac{1}{n}\sum_{i=1}^n a_i\right)$ maximizes $f_k(x_1, \ldots, x_n)$ subject to the constraint $\sum_{i=1}^n x_i = \sum_{i=1}^n a_i$. □

Theorem 5.5 (Upper tail bound under k-wise independence). *Let $X_1, \ldots, X_n$ be n random variables with values in $[0,1]$ such that at least one X_i, $i \in \mathscr{S}_n$, does not concentrate all its mass at 1. Set $X := \sum_{i=1}^n X_i$. Further, let $t \in (0, n - \mathbb{E}[X]]$ be arbitrary and set $k := \left\lceil \frac{tn}{n-\mathbb{E}[X]} \right\rceil$, where $\lceil \cdot \rceil$ denotes the ceiling function. If $X_1, \ldots, X_n$ are k-wise independent, then*

$$\mathbb{P}\left(X - \mathbb{E}[X] \geq t\right) \;\leq\; \binom{n}{k}\left(\frac{1}{n}\,\mathbb{E}[X]\right)^k \Big/ \binom{t + \mathbb{E}[X]}{k}. \tag{5.4}$$

Proof. Note first that $n - \mathbb{E}[X] > 0$ because at least one X_i is not almost surely 1. Further, it can be easily inferred from $t \leq n - \mathbb{E}[X]$ that $\frac{tn}{(n-\mathbb{E}[X])(t+\mathbb{E}[X])} \leq 1$ and, consequently, that

$$k \;<\; \frac{tn}{n - \mathbb{E}[X]} + 1 \;=\; \frac{tn(t + \mathbb{E}[X])}{(n - \mathbb{E}[X])(t + \mathbb{E}[X])} + 1 \;\leq\; t + \mathbb{E}[X] + 1.$$

Therefore, we may apply Theorem 5.2 to obtain

$$\mathbb{P}\Big(X - \mathbb{E}[X] \geq t\Big) \leq \frac{1}{\binom{t+\mathbb{E}[X]}{k}} \sum_{\substack{S \subseteq \mathscr{S}_n \\ |S|=k}} \mathbb{E}\left[\prod_{i \in S} X_i\right].$$

Now, observe that the k-wise independence of $X_1, \ldots, X_n$ implies

$$\sum_{\substack{S \subseteq \mathscr{S}_n \\ |S|=k}} \mathbb{E}\left[\prod_{i \in S} X_i\right] \;=\; \sum_{\substack{S \subseteq \mathscr{S}_n \\ |S|=k}} \prod_{i \in S} \mathbb{E}\left[X_i\right] \;\leq\; \binom{n}{k}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[X_i\right]\right)^k \;=\; \binom{n}{k}\left(\frac{1}{n}\,\mathbb{E}[X]\right)^k,$$

where the inequality is due to Lemma 5.4. Hence the theorem. □

Notice that Theorem 5.5, just like Theorem 5.2, reduces to Markov's inequality when $k = 1$. For the reader wondering about the choice of the integer k in the statement of Theorem 5.5, we have the following explanation.

Remark 5.6. *While the proof of Theorem 5.5 shows that tail inequality (5.4) holds true for any $k < t + \mathbb{E}[X] + 1$ whenever $X_1, \ldots, X_n$ are k-wise independent, the bound $\binom{n}{k}\left(\frac{1}{n}\,\mathbb{E}[X]\right)^k \Big/ \binom{t+\mathbb{E}[X]}{k}$ is minimized at $k = k_* := \left\lceil \frac{tn}{n-\mathbb{E}[X]} \right\rceil$. To convince yourself of this fact, verify that, for any $k \in \mathbb{N}_0$ satisfying $k < t + \mathbb{E}[X]$, the term*

$$\frac{\binom{n}{k+1}\left(\frac{1}{n}\,\mathbb{E}[X]\right)^{k+1} \Big/ \binom{t+\mathbb{E}[X]}{k+1}}{\binom{n}{k}\left(\frac{1}{n}\,\mathbb{E}[X]\right)^{k} \Big/ \binom{t+\mathbb{E}[X]}{k}} \;=\; \frac{(n-k)\mathbb{E}[X]}{(t + \mathbb{E}[X] - k)n}$$

is less than, equal to, or greater than 1 *if and only if* k *is less than, equal to, or greater than* $\frac{tn}{n-\mathbb{E}[X]}$*, respectively; see (Schmidt et al., 1995, p. 6), where a similar observation was made in a related context. Since* $k_* - 1 < \frac{tn}{n-\mathbb{E}[X]}$ *and* $k_* \geq \frac{tn}{n-\mathbb{E}[X]}$*, we conclude that* $\binom{n}{k}\left(\frac{1}{n}\,\mathbb{E}[X]\right)^k / \binom{t+\mathbb{E}[X]}{k}$ *is minimized at* $k = k_*$.

For completeness, we provide a lower tail inequality for sums of k-wise independent random variables, which is a simple consequence of Theorem 5.5.

Corollary 5.7 (Lower tail bound under k-wise independence). *Let* $X_1, \ldots, X_n$ *be* n *random variables with values in* $[0, 1]$ *such that at least one* X_i*,* $i \in \mathscr{S}_n$*, does not concentrate all its mass at* 0*. Set* $X := \sum_{i=1}^n X_i$*. Further, let* $t \in (0, \mathbb{E}[X]]$ *be arbitrary and set* $k := \left\lceil \frac{tn}{\mathbb{E}[X]} \right\rceil$*. If* $X_1, \ldots, X_n$ *are* k*-wise independent, then*

$$\mathbb{P}\left(X - \mathbb{E}[X] \leq -t\right) \leq \binom{n}{k}\left(1 - \frac{1}{n}\,\mathbb{E}[X]\right)^k \Big/ \binom{t + n - \mathbb{E}[X]}{k}.$$

Proof. The corollary follows from Theorem 5.5 by replacing each X_i with $1 - X_i$ and noticing that if $\{X_i \colon i \in \mathscr{S}_n\}$ are k-wise independent, then $\{1 - X_i \colon i \in \mathscr{S}_n\}$ are also k-wise independent (cf. Propositions 5.8 and 5.9 in (Çinlar, 2011, p. 85)). □

Remark 5.8. *The results given in Sections 5.1 and 5.2 can easily be adapted, thanks to a scaling argument similar to the one found in the proof of Theorem 3.5, to the case where the random variables are supported on arbitrary finite intervals of the same length.*

5.3 An alternative to Theorem 3.1

From the proof of Theorem 3.9 in Pelekis and Ramon (2017), we can extract the following alternative to Theorem 3.1.

Theorem 5.9. *Let* $X_1, \ldots, X_n$ *be* n *random variables with values in* $[0, 1]$*. Suppose there are some constants* $c_i \in [0, 1]$ *for each* $i \in \mathscr{S}_n$ *such that*

$$\mathbb{E}\Big[\prod_{i \in S} X_i\Big] \leq \prod_{i \in S} c_i \quad \forall S \subseteq \mathscr{S}_n. \tag{5.5}$$

Set $X := \sum_{i=1}^n X_i$ *and* $c := (1/n)\sum_{i=1}^n c_i$. *Let* $t \in [0, 1-c]$ *be arbitrary and set* $k_* := \left\lceil \frac{tn}{1-c} \right\rceil \mathbb{1}_{\{c<1\}}$. *Then*

$$\mathbb{P}\Big(X \geq (c+t)n\Big) \leq \min_{\substack{k\in\mathbb{N}_0 \\ k<(c+t)n+1}} \binom{n}{k} c^k \Big/ \binom{(c+t)n}{k} = \binom{n}{k_*} c^{k_*} \Big/ \binom{(c+t)n}{k_*}. \tag{5.6}$$

Proof. By application of Theorem 5.2, we obtain for all $k \in \mathbb{N}$ with $k < (c+t)n+1$

$$\begin{aligned}
\mathbb{P}\Big(X \geq (c+t)n\Big) &\leq \frac{1}{\binom{(c+t)n}{k}} \sum_{\substack{S\subseteq \mathscr{I}_n \\ |S|=k}} \mathbb{E}\Big[\prod_{i\in S} X_i\Big] \\
&\leq \frac{1}{\binom{(c+t)n}{k}} \sum_{\substack{S\subseteq \mathscr{I}_n \\ |S|=k}} \prod_{i\in S} c_i \leq \binom{n}{k} c^k \Big/ \binom{(c+t)n}{k},
\end{aligned}$$

where the penultimate inequality is a consequence of (5.5), and the last inequality is due to Lemma 5.4. This, together with the fact that $\binom{n}{0}c^0\big/\binom{(c+t)n}{0} = 1$, where we apply the usual convention $0^0 = 1$ if $c = 0$, yield the inequality in (5.6).

The equality in (5.6) follows from a similar argument to the one used in Remark 5.6. More precisely, it trivially holds when $c = 1$ (in which case $t = k_* = 0$). Assume that $c < 1$. From $t \leq 1-c$, we infer that $t \leq (c+t)(1-c)$ and, accordingly, that

$$k_* = \left\lceil \frac{tn}{1-c} \right\rceil < \frac{tn}{1-c} + 1 \leq \frac{(c+t)tn}{(c+t)(1-c)} + 1 \leq (c+t)n + 1.$$

Further, verify that, for any $k \in \mathbb{N}_0$ satisfying $k < (c+t)n$, the term

$$\frac{\binom{n}{k+1} c^{k+1} \big/ \binom{(c+t)n}{k+1}}{\binom{n}{k} c^k \big/ \binom{(c+t)n}{k}} = \frac{(n-k)c}{(c+t)n-k}$$

is less than, equal to, or greater than 1 if and only if k is less than, equal to, or greater than $\frac{tn}{1-c}$, respectively. Since $k_* - 1 < \frac{tn}{1-c}$ and $k_* \geq \frac{tn}{1-c}$, we conclude that $\binom{n}{k}c^k\big/\binom{(c+t)n}{k}$ is minimized at $k = k_*$. This gives the theorem. □

Pelekis and Ramon (2017) show that the tail bound in Theorem 5.9 is sharper than the one in Theorem 3.1. Unfortunately, the authors use a spurious argument to justify their claim. This is the subject of the next remark.

Remark 5.10. *Pelekis and Ramon (2017) argue in the proof of their Theorem 3.9 that the bound* $\binom{n}{k_*}c^{k_*}\big/\binom{(c+t)n}{k_*}$, *where* c, t, *and* k_* *are given as in the statement of Theorem 5.9, is less than or equal to the Chernoff-Hoeffding bound* $e^{-D_{KL}(c+t\|c)n}$. *In* *the degenerate cases where $c \in \{0,1\}$ or $t = 0$, this is trivially true. It is also true when $t = 1-c$, because then $k_* = n$ and $e^{-D_{KL}(1\|c)n} = c^n$. In the case where $c \in (0,1)$ and $t \in (0, 1-c)$, Pelekis and Ramon consider the function $Q\colon [0, k_*] \to (0, \infty)$,*

$$Q_{n,t,c}(x) = c^x \left(\frac{1}{c+t}\right)^x \left(\frac{n}{n-x}\right)^{n-x} \left(\frac{(c+t)n - x}{(c+t)n}\right)^{(c+t)n-x},$$

and deduce their claim from the observations that $\binom{n}{k} c^k \big/ \binom{(c+t)n}{k} \le Q_{n,t,c}(k)$ for all $k \in \mathbb{N}_0$ satisfying $k \le k_*$, that $Q_{n,t,c}\left(\frac{tn}{1-c}\right) = e^{-D_{KL}(c+t\|c)n}$, and that $Q_{n,t,c}$ is non-increasing for $x \le k_* = \left\lceil \frac{tn}{1-c} \right\rceil$. However, the latter observation is wrong. More precisely, by looking at the sign of the first derivative

$$Q'_{n,t,c}(x) = Q_{n,t,c}(x) \ln\left(\frac{(n-x)c}{(c+t)n - x}\right),$$

we can tell that $Q_{n,t,c}$ is strictly decreasing for $x < \frac{tn}{1-c}$ and strictly increasing for $x > \frac{tn}{1-c}$. This shows that, in the general case when $\frac{tn}{1-c}$ is not an integer, the proof of Pelekis and Ramon is invalid. The argument used by Pelekis and Ramon goes back to Schmidt et al. (1995), who mistakenly state in the proof of their Theorem 2 that $Q_{n,t,c}$ is non-increasing for $x \le k_*$ (although they need the non-increasingness of $Q_{n,t,c}$ only for $x \le k_* - 1$).

The tail bound in Theorem 5.9 is presumably sharper than the one in Theorem 3.1. In the special case where $(c+t)n$ is integral, this is established in Section 2.1 in Schmidt et al. (1995). In the general case, a rigorous proof is yet to be given.

5.4 A third proof of Theorem 3.1

In this section, we show that using the machinery of Pelekis and Ramon (2017), it is easy to obtain yet another proof of Theorem 3.1.

Proof of Theorem 3.1. We first deal with the case where $c \in (0,1)$ and $t \in (0, 1-c)$. Fix any $t \in (0, 1-c)$ and let $\lambda > 0$ be arbitrary. Define $f : [0,\infty) \to [0,\infty)$, $f(x) = e^{\lambda x}$. Notice that $f\big((c+t)n\big) > 0$ and that

$$(c+t)n - \mathbb{E}[X] \;=\; tn + \sum_{i=1}^{n} \big(c_i - \mathbb{E}[X_i]\big) \;\ge\; tn \;>\; 0,$$

where the penultimate inequality is due to the assumption $\mathbb{E}[X_i] \leq c_i$ for all $i \in \mathscr{S}_n$. Therefore, keeping in mind that f is non-decreasing and convex, we apply Proposition 5.1 to obtain

$$\mathbb{P}\Big(X \geq (c+t)n\Big) = \mathbb{P}\Big(X - \mathbb{E}[X] \geq (c+t)n - \mathbb{E}[X]\Big) \leq e^{-\lambda(c+t)n}\,\mathbb{E}[e^{\lambda Z}],$$

where Z is a $\{0, 1, \dots, n\}$-valued random variable defined by the probability mass function in (5.1). Further, observe that

$$\begin{aligned}
\mathbb{E}[e^{\lambda Z}] &= \sum_{j=0}^{n} e^{\lambda j} \sum_{\substack{S \subseteq \mathscr{S}_n \\ |S|=j}} \mathbb{E}\left[\left(\prod_{i \in S} X_i\right)\left(\prod_{l \in \mathscr{S}_n \setminus S} (1 - X_l)\right)\right] \\
&= \sum_{j=0}^{n} \sum_{\substack{S \subseteq \mathscr{S}_n \\ |S|=j}} \mathbb{E}\left[\left(\prod_{i \in S} e^{\lambda} X_i\right)\left(\prod_{l \in \mathscr{S}_n \setminus S} (1 - X_l)\right)\right] \\
&= \mathbb{E}\left[\sum_{S \subseteq \mathscr{S}_n} \left(\prod_{i \in S} e^{\lambda} X_i\right)\left(\prod_{l \in \mathscr{S}_n \setminus S} (1 - X_l)\right)\right] = \mathbb{E}\left[\prod_{i=1}^{n} \Big(1 + (e^{\lambda} - 1) X_i\Big)\right],
\end{aligned}$$

where the second and the third equalities are thanks to linearity of expectation, and the last equality follows from Lemma 2.11. From here, we simply follow the proof of Theorem 3.1 in Section 4.2, starting from the last expression in the first line of computation (4.6). □

6 | Concentration inequalities via differential privacy

Recently, Steinke and Ullman (2017) demonstrated that differencial privacy[1], a statistical technique introduced in Dwork et al. (2006), can also be applied to derive exponentially decaying concentration bounds (alas considerably weaker than those obtained with the previous three methods). The underlying idea is to reduce the problem of obtaining an upper bound on the tail probability $\mathbb{P}\big(X - \mathbb{E}[X] \geq t\big)$ to the problem of computing an upper bound on the expectation of the random variable $\max_{j \in \mathscr{S}_m}\big\{0, X^{(j)} - \mathbb{E}[X]\big\}$ with parameter $m \in \mathbb{N}$, where $X^{(1)}, \dots, X^{(m)}$ are m independent copies of X. That way, the so-called randomized stable selection procedure can be brought to bear on the matter at hand. The purpose of this chapter is to illustrate the method by considering tail bounds for sums of independent random variables. For a derivation of, e.g., McDiarmid's inequality, we refer the interested reader to (Steinke and Ullman, 2017, Chapter 3.2) and Bassily et al. (2016).

6.1 Randomized stable selection procedure

Let $\mathcal{S}_{\mathbf{A},t}$ be an $\mathscr{S}_m$-valued random variable with parameters $\mathbf{A} = (\mathbf{A}_{i,j}) \in [0,1]^{n \times m}$ and $t > 0$, and with probability mass function

$$\mathbb{P}\big(\mathcal{S}_{\mathbf{A},t} = j\big) = \frac{1}{g(\mathbf{A},t)} \exp\left(\frac{t}{2}\sum_{i=1}^{n} \mathbf{A}_{i,j}\right) > 0 \quad \forall j \in \mathscr{S}_m, \tag{6.1}$$

[1]For a comprehensive treatment of the subject, we refer the reader to McSherry and Talwar (2007) and Dwork and Roth (2014).

where the function g is given by

$$g(\mathbf{A}, t) := \sum_{k=1}^{m} \exp\left(\frac{t}{2}\sum_{i=1}^{n} \mathbf{A}_{i,k}\right). \tag{6.2}$$

It is easily seen that $\sum_{j=1}^{m} \mathbb{P}(\mathcal{S}_{\mathbf{A},t} = j) = 1$, meaning that the distribution of $\mathcal{S}_{\mathbf{A},t}$ is well-defined. The random variable $\mathcal{S}_{\mathbf{A},t}$ is referred to as a stable selector for $\mathbf{A}$ with stability parameter t. Lemma 6.1 below, which stems from Steinke and Ullman (2017), establishes two useful properties of $\mathcal{S}_{\mathbf{A},t}$.

Lemma 6.1. *Let $\mathbf{A} \in [0,1]^{n\times m}$, $t > 0$, and let $\mathcal{S}_{\mathbf{A},t}$ be the random variable defined by the probability mass function (6.1). The following statements hold true.*

(i) Stability: For all row vectors $\mathbf{a} \in [0,1]^{1\times m}$ and all $i \in \mathscr{S}_n$, $j \in \mathscr{S}_m$, one has

$$\mathbb{P}\Big(\mathcal{S}_{\mathbf{A},t} = j\Big) \leq e^t\, \mathbb{P}\Big(\mathcal{S}_{(\mathbf{A}_{-i},\mathbf{a}),t} = j\Big), \tag{6.3}$$

where $(\mathbf{A}_{-i}, \mathbf{a}) \in [0,1]^{n\times m}$ is the matrix $\mathbf{A}$ with i-th row being replaced by $\mathbf{a}$.

(ii) Accuracy:

$$\mathbb{E}\left[\sum_{l=1}^{n} \mathbf{A}_{l,\mathcal{S}_{\mathbf{A},t}}\right] \geq \max_{j\in\mathscr{S}_m} \sum_{l=1}^{n} \mathbf{A}_{l,j} - \frac{2\ln(m)}{t}.$$

Before proving the lemma, we note that inequality (6.3) is equivalent to the two-sided inequality in Claim 2.2 in the proof of Lemma 2.1 in Steinke and Ullman (2017), where the left inequality can be inferred from the right inequality and vice versa. Notice also, as pointed out in Steinke and Ullman (2017), that the parameter t manages the trade-off between stability and accuracy. The former makes sure that the probability mass function of $\mathcal{S}_{\mathbf{A},t}$ does not change much when a single row of the matrix $\mathbf{A}$ is being replaced, while the latter ensures that, in expectation, the sum of the elements in the column selected by $\mathcal{S}_{\mathbf{A},t}$ is close to sum of the "largest column".

Proof of Lemma 6.1. We follow the lines of the proof of Lemma 2.1 in Steinke and Ullman (2017). To verify (i), fix some arbitrary row vector $\mathbf{a} \in [0,1]^{1\times m}$, $i \in \mathscr{S}_n$, and $j \in \mathscr{S}_m$. Observe that

$$\begin{aligned}\frac{\exp\left(\frac{t}{2}\sum_{l=1}^{n} \mathbf{A}_{l,j}\right)}{\exp\left(\frac{t}{2}\sum_{l=1}^{n} (\mathbf{A}_{-i},\mathbf{a})_{l,j}\right)} &= \exp\left(\frac{t}{2}\sum_{l=1}^{n}\Big(\mathbf{A}_{l,j} - (\mathbf{A}_{-i},\mathbf{a})_{l,j}\Big)\right)\\ &= \exp\left(\frac{t}{2}\Big(\mathbf{A}_{i,j} - (\mathbf{A}_{-i},\mathbf{a})_{i,j}\Big)\right) \in \Big[e^{-\frac{t}{2}}, e^{\frac{t}{2}}\Big]\end{aligned} \tag{6.4}$$

because, by definition, $\mathbf{A}_{l,j} = (\mathbf{A}_{-i}, \mathbf{a})_{l,j}$ unless $l = i$, and because the entries of both matrices $\mathbf{A}$ and $(\mathbf{A}_{-i}, \mathbf{a})$ are in $[0,1]$, meaning that $-1 \leq \mathbf{A}_{i,j} - (\mathbf{A}_{-i}, \mathbf{a}_i)_{i,j} \leq 1$. From this, we infer that

$$g(\mathbf{A}, t) = \sum_{k=1}^{m} \exp\left(\frac{t}{2}\sum_{l=1}^{n} \mathbf{A}_{l,k}\right) \geq \sum_{k=1}^{m} e^{-\frac{t}{2}} \exp\left(\frac{t}{2}\sum_{l=1}^{n} (\mathbf{A}_{-i}, \mathbf{a})_{l,k}\right) = e^{-\frac{t}{2}} g\Big((\mathbf{A}_{-i}, \mathbf{a}), t\Big).$$

Combining the above with (6.4) yields

$$\frac{\mathbb{P}\Big(\mathcal{S}_{\mathbf{A},t} = j\Big)}{\mathbb{P}\Big(\mathcal{S}_{(\mathbf{A}_{-i},\mathbf{a}),t} = j\Big)} = \frac{g\left((\mathbf{A}_{-i}, \mathbf{a}), t\right)}{g(\mathbf{A}, t)} \frac{\exp\left(\frac{t}{2}\sum_{l=1}^{n} \mathbf{A}_{l,j}\right)}{\exp\left(\frac{t}{2}\sum_{l=1}^{n} (\mathbf{A}_{-i}, \mathbf{a})_{l,j}\right)} \leq e^{\frac{t}{2}} e^{\frac{t}{2}} = e^{t},$$

which, in turn, gives (i).

To prove (ii), note first that, trivially, (6.1) implies

$$\sum_{l=1}^{n} \mathbf{A}_{l,k} = \frac{2}{t}\Big(\ln g(\mathbf{A}, t) + \ln \mathbb{P}\Big(\mathcal{S}_{\mathbf{A},t} = k\Big)\Big) \qquad \forall k \in \mathscr{S}_m,$$

and that (6.2) implies

$$g(\mathbf{A}, t) = \sum_{k=1}^{m} \exp\left(\frac{t}{2}\sum_{l=1}^{n} \mathbf{A}_{l,k}\right) \geq \max_{k\in\mathscr{S}_m} \exp\left(\frac{t}{2}\sum_{l=1}^{n} \mathbf{A}_{l,k}\right) = \exp\left(\frac{t}{2}\max_{k\in\mathscr{S}_m}\sum_{l=1}^{n} \mathbf{A}_{l,k}\right).$$

Hence, we conclude that

$$\begin{aligned}
\mathbb{E}\left[\sum_{l=1}^{n} \mathbf{A}_{l,\mathcal{S}_{\mathbf{A},t}}\right] &= \sum_{k=1}^{m} \mathbb{P}\Big(\mathcal{S}_{\mathbf{A},t} = k\Big) \sum_{l=1}^{n} \mathbf{A}_{l,k} \\
&= \frac{2}{t}\sum_{k=1}^{m} \mathbb{P}\Big(\mathcal{S}_{\mathbf{A},t} = k\Big)\Big(\ln g(\mathbf{A}, t) + \ln \mathbb{P}\Big(\mathcal{S}_{\mathbf{A},t} = k\Big)\Big) \\
&= \frac{2}{t}\Big(\ln g(\mathbf{A}, t) + \sum_{k=1}^{m} \mathbb{P}\Big(\mathcal{S}_{\mathbf{A},t} = k\Big) \ln \mathbb{P}\Big(\mathcal{S}_{\mathbf{A},t} = k\Big)\Big) \\
&\geq \max_{k\in\mathscr{S}_m} \sum_{l=1}^{n} \mathbf{A}_{l,k} - \frac{2\ln(m)}{t},
\end{aligned}$$

where, to justify the inequality, we use the fact that for each probability mass function p on $\mathscr{S}_m$, one has $\sum_{k=1}^{m} p(k) \ln p(k) \geq -\ln(m)$; see, e.g., Theorem 2.6.4 in (Cover and Thomas, 1991, p. 27). □

The parameter $\mathbf{A}$ of the stable selector $\mathcal{S}_{\mathbf{A},t}$ can be randomized by a random matrix $\mathbf{X}$. This is helpful to prove the following.

Lemma 6.2. *Let $\mathbf{X}$ be a random $n \times m$ matrix with values in $[0,1]^{n\times m}$ and independent row vectors $(\mathbf{X}_{1,1}, \ldots, \mathbf{X}_{1,m}), \ldots, (\mathbf{X}_{n,1}, \ldots, \mathbf{X}_{n,m})$. Then, for all $t > 0$, we have*

$$\mathbb{E}\left[\max_{j\in\mathscr{S}_m}\sum_{i=1}^{n} \mathbf{X}_{i,j}\right] \leq e^{t} \max_{j\in\mathscr{S}_m} \mathbb{E}\left[\sum_{i=1}^{n} \mathbf{X}_{i,j}\right] + \frac{2\ln(m)}{t}.$$

Proof. The proof of the lemma is along the lines of the proof of Claim 2.4 in Steinke and Ullman (2017). Let $\widetilde{\mathbf{X}}$ be an independent copy of $\mathbf{X}$. For each $i \in \mathscr{S}_n$, let $\widetilde{\mathbf{x}}_i$ denote the i-th row vector of the matrix $\widetilde{\mathbf{X}}$. Since the row vectors of $\mathbf{X}$ are independent[2], it follows that $(\mathbf{X}_{-i}, \widetilde{\mathbf{x}}_i)$ and $\mathbf{X}_{i,k}$ have the same joint distribution as $\mathbf{X}$ and $\widetilde{\mathbf{X}}_{i,k}$ for all $i \in \mathscr{S}_n$ and all $k \in \mathscr{S}_m$. This leads us to

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n} \mathbf{X}_{i,\mathcal{S}_{\mathbf{X},t}}\right] &= \sum_{i=1}^{n} \mathbb{E}\left[\mathbf{X}_{i,\mathcal{S}_{\mathbf{X},t}}\right] = \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}\left[\mathbf{X}_{i,\mathcal{S}_{\mathbf{X},t}} \,|\, \mathbf{X}\right]\right] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[\sum_{k=1}^{m} \mathbf{X}_{i,k} \frac{1}{g(\mathbf{X},t)} \exp\left(\frac{t}{2}\sum_{l=1}^{n} \mathbf{X}_{l,k}\right)\right] \\
&\leq \sum_{i=1}^{n} \mathbb{E}\left[\sum_{k=1}^{m} \mathbf{X}_{i,k} \frac{e^t}{g\left((\mathbf{X}_{-i}, \widetilde{\mathbf{x}}_i), t\right)} \exp\left(\frac{t}{2}\sum_{l=1}^{n} (\mathbf{X}_{-i}, \widetilde{\mathbf{x}}_i)_{l,k}\right)\right] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[\sum_{k=1}^{m} \widetilde{\mathbf{X}}_{i,k} \frac{e^t}{g\left(\mathbf{X}, t\right)} \exp\left(\frac{t}{2}\sum_{l=1}^{n} \mathbf{X}_{l,k}\right)\right] \\
&= e^t \sum_{k=1}^{m} \mathbb{E}\left[\frac{1}{g\left(\mathbf{X}, t\right)} \exp\left(\frac{t}{2}\sum_{l=1}^{n} \mathbf{X}_{l,k}\right)\right] \mathbb{E}\left[\sum_{i=1}^{n} \widetilde{\mathbf{X}}_{i,k}\right] \\
&\leq e^t \sum_{k=1}^{m} \mathbb{E}\left[\frac{1}{g\left(\mathbf{X}, t\right)} \exp\left(\frac{t}{2}\sum_{l=1}^{n} \mathbf{X}_{l,k}\right)\right] \max_{j \in \mathscr{S}_m} \mathbb{E}\left[\sum_{i=1}^{n} \widetilde{\mathbf{X}}_{i,j}\right] \\
&= e^t \max_{j \in \mathscr{S}_m} \mathbb{E}\left[\sum_{i=1}^{n} \mathbf{X}_{i,j}\right] \mathbb{E}\left[\sum_{k=1}^{m} \frac{1}{g\left(\mathbf{X}, t\right)} \exp\left(\frac{t}{2}\sum_{l=1}^{n} \mathbf{X}_{l,k}\right)\right] \\
&= e^t \max_{j \in \mathscr{S}_m} \mathbb{E}\left[\sum_{i=1}^{n} \mathbf{X}_{i,j}\right], \qquad (6.5)
\end{aligned}$$

where we apply Proposition 2.4 (ii) together with Remark 2.5 in the second step, Lemma 6.1 (i) in the fourth, the fact that $(\mathbf{X}_{-i}, \widetilde{\mathbf{x}}_i)$ and $\mathbf{X}_{i,k}$ have the same joint distribution as $\mathbf{X}$ and $\widetilde{\mathbf{X}}_{i,k}$ in the fifth, linearity of expectation together with independence of $\mathbf{X}$ and $\widetilde{\mathbf{X}}$ in the sixth, the fact that $\mathbf{X}$ and $\widetilde{\mathbf{X}}$ are identically distributed in the penultimate step, and (6.2) in the last one. Hence, we conclude that

$$\mathbb{E}\left[\max_{j \in \mathscr{S}_m} \sum_{i=1}^{n} \mathbf{X}_{i,j}\right] \leq \mathbb{E}\left[\sum_{i=1}^{n} \mathbf{X}_{i,\mathcal{S}_{\mathbf{X},t}}\right] + \frac{2\ln(m)}{t} \leq e^t \max_{j \in \mathscr{S}_m} \mathbb{E}\left[\sum_{i=1}^{n} \mathbf{X}_{i,j}\right] + \frac{2\ln(m)}{t},$$

where the first inequality is thanks to Lemma 6.1 (ii) and the second to (6.5). □

[2]The independence of the row vectors of $\mathbf{X}$ allows us to replace the i-th row of $\mathbf{X}$ with the vector $\widetilde{\mathbf{x}}_i$ without changing the distribution of the matrix. Lemma 6.2 will be applied to derive a tail inequality for the sum of n independent random variables $X_1, \dots, X_n$. For this purpose, each column vector of $\mathbf{X}$ will be an independent copy of the vector $(X_1, \dots, X_n)^\top$. And it is precisely the independence of the random variables $X_1, \dots, X_n$ that will guarantee the independence of the row vectors of $\mathbf{X}$.

6.2 Tail bounds for sums of independent bounded random variables

Lemma 6.3 provides an upper bound on the expectation of the random variable $\max\left\{0, X^{(1)} - \mathbb{E}[X], \ldots, X^{(m)} - \mathbb{E}[X]\right\}$, where $X^{(1)}, \ldots, X^{(m)}$ are m independent copies of the sum X of n independent $[0,1]$-valued random variables. A comparison with Proposition 2.5 in Steinke and Ullman (2017) reveals an improvement of the bound by a multiplicative factor of $\left(2\sqrt{\ln(3/2)}\right)^{-1}$. This was achieved by identifying a better interval over which a function of interest is minimized.

Lemma 6.3. *Let $X_1, \ldots, X_n$ be n independent $[0,1]$-valued random variables. Set $X := \sum_{i=1}^{n} X_i$. Further, let $m \in \mathbb{N}$ and let $X^{(1)}, \ldots X^{(m)}$ be m independent copies of X. Then*

$$\mathbb{E}\left[\max_{j \in \mathscr{S}_m}\left\{0, X^{(j)} - \mathbb{E}[X]\right\}\right] \le \frac{2}{\sqrt{\ln(3/2)}}\sqrt{n\ln(m+1)} < 3.1409\sqrt{n\ln(m+1)}.$$

Proof. We follow the lines of the proof of Proposition 2.5 in Steinke and Ullman (2017). Observe first that, since $\max_{j \in \mathscr{S}_m}\left\{0, X^{(j)} - \mathbb{E}[X]\right\} \le n$, the statement of the lemma trivially holds if $m > e^{n\ln(3/2)/4} - 1$, which is always the case if $n \le 6$. Therefore, we assume $n \ge 7$ and $m \le e^{n\ln(3/2)/4} - 1$. Let $\mathbf{X}$ be a random $n \times (m+1)$ matrix such that its column vectors $(\mathbf{X}_{1,j}, \ldots, \mathbf{X}_{n,j})^\top$, $j \in \mathscr{S}_m$, are independent and distributed as the vector $(X_1, \ldots, X_n)^\top$ and such that its entries in the last column are given by $\mathbf{X}_{i,m+1} = \mathbb{E}[X_i]$ for all $i \in \mathscr{S}_n$. Clearly, since $X_1, \ldots, X_n$ are independent, all row vectors of $\mathbf{X}$ are independent. Hence, by application of Lemma 6.2, we obtain for all $t > 0$ that

$$\mathbb{E}\left[\max_{j \in \mathscr{S}_{m+1}} \sum_{i=1}^{n} \mathbf{X}_{i,j}\right] \le e^t \max_{j \in \mathscr{S}_{m+1}} \mathbb{E}\left[\sum_{i=1}^{n} \mathbf{X}_{i,j}\right] + \frac{2\ln(m+1)}{t} = e^t\,\mathbb{E}\left[X\right] + \frac{2\ln(m+1)}{t}.$$

Subtracting $\mathbb{E}[X]$ from both sides and calling to mind that for all $j \in \mathscr{S}_m$, $\sum_{i=1}^{n} \mathbf{X}_{i,j}$ has the same distribution as the random variable $X^{(j)}$, we get

$$\mathbb{E}\left[\max_{j \in \mathscr{S}_m}\left\{0, X^{(j)} - \mathbb{E}[X]\right\}\right] \le (e^t - 1)\mathbb{E}\left[X\right] + \frac{2\ln(m+1)}{t} \le (e^t - 1)n + \frac{2\ln(m+1)}{t},$$

where the last inequality is thanks to the estimate $\mathbb{E}[X] \le n$. Now, notice that

$$e^t - 1 \le \frac{t}{2\ln(3/2)} \quad \forall t \in \left(0, \ln(3/2)\right]. \tag{6.6}$$

To convince yourself that (6.6) is true, consider the function $f(t) := e^t - 1 - t/(2\ln(3/2))$ and verify that $f(0) = 0$, $f(\ln(3/2)) = 0$, and $f''(t) = e^t > 0$, meaning that $f(t) \le 0$ for all $t \in (0, \ln(3/2)]$, which, in turn, gives (6.6). This yields

$$\mathbb{E}\Big[\max_{j \in \mathscr{S}_m} \big\{0, X^{(j)} - \mathbb{E}[X]\big\}\Big] \le \frac{tn}{2\ln(3/2)} + \frac{2\ln(m+1)}{t} \qquad \forall t \in \big(0, \ln(3/2)\big]. \tag{6.7}$$

To minimize the right-hand side of inequality (6.7), let $g(t) := tn/(2\ln(3/2)) + 2\ln(m+1)/t$ and compute

$$g'(t) = \frac{n}{2\ln(3/2)} - \frac{2\ln(m+1)}{t^2} \qquad \text{and} \qquad g''(t) = \frac{4\ln(m+1)}{t^3}.$$

Hence, we infer that g attains its minimum at $t_* := 2\sqrt{\ln(3/2)\ln(m+1)/n}$. From our assumption $m \le e^{n\ln(3/2)/4} - 1$, or, equivalently, $\ln(m+1) \le n\ln(3/2)/4$, we deduce that $t_* \le \ln(3/2)$. To finish the proof, check that

$$g(t_*) = \frac{2n\sqrt{\ln(3/2)\ln(m+1)}}{2\ln(3/2)\sqrt{n}} + \frac{2\ln(m+1)\sqrt{n}}{2\sqrt{\ln(3/2)\ln(m+1)}} = \frac{2}{\sqrt{\ln(3/2)}}\sqrt{n\ln(m+1)}.$$

□

Remark 6.4. *If X is the sum of n independent $[0,1]$-valued random variables and $X^{(1)}, \dots, X^{(m)}$ are m independent copies of X, then*

$$\mathbb{E}\Big[\max_{j \in \mathscr{S}_m} X^{(j)}\Big] \le \mathbb{E}[X] + \frac{2}{\sqrt{\ln(3/2)}}\sqrt{n\ln m} < \mathbb{E}[X] + 3.1409\sqrt{n\ln m}. \tag{6.8}$$

The proof of (6.8) is almost identical to the one of Lemma 6.3. The only difference is that, this time, we leave the $(m+1)$-st column out of the random matrix $\mathbf{X}$. The result in (6.8) is interesting since it improves Lemma 3.2 in Mulzer (2018).

Our next theorem presents a tail bound for the sum of independent $[0,1]$-valued random variables. It slightly improves Theorem 2.6 in Steinke and Ullman (2017). The gain in accuracy is obtained thanks to both the sharper bound in Lemma 6.4 and an idea found in the proof of Theorem 3.3 in Mulzer (2018).

Theorem 6.5. *Let $X_1, \dots, X_n$ be n independent $[0,1]$-valued random variables. Set $X := \sum_{i=1}^n X_i$. Then, for any $t \ge 0$, we have*

$$\mathbb{P}\big(X - \mathbb{E}[X] \ge t\big) \le e^{1 - \frac{t^2}{29n}}. \tag{6.9}$$

Proof. Observe first that (6.9) trivially holds if $t \le \sqrt{29n}$. Fix an arbitrary $t > \sqrt{29n}$. Let $m \in \mathbb{N}$ be arbitrary. Let $X^{(1)}, \dots, X^{(m)}$ be m independent copies of X. On the one hand, we have

$$\begin{aligned}
\mathbb{P}\Big(\max_{j \in \mathscr{S}_m} \big\{0, X^{(j)} - \mathbb{E}[X]\big\} \ge t \Big) &= 1 - \mathbb{P}\Big(\max_{j \in \mathscr{S}_m} \big\{0, X^{(j)} - \mathbb{E}[X]\big\} < t \Big) \\
&= 1 - \mathbb{P}\Big(X^{(j)} - \mathbb{E}[X] < t \ \ \forall j \in \mathscr{S}_m \Big) = 1 - \prod_{i=1}^{m} \mathbb{P}\Big(X^{(j)} - \mathbb{E}[X] < t \Big) \\
&= 1 - \Big(1 - \mathbb{P}\big(X - \mathbb{E}[X] \ge t \big) \Big)^m \ge 1 - e^{-m \mathbb{P}(X - \mathbb{E}[X] \ge t)}, \qquad (6.10)
\end{aligned}$$

where the third equality follows from the independence of the $X^{(j)}$'s, the fourth from the fact that all $X^{(j)}$'s are distributed as X, and the inequality holds true since for all $x \in [0,1]$, one has $1 - x \le e^{-x}$ (check that $e^{-x} - 1 + x$ is minimized at $x = 0$). On the other hand, applying Markov's inequality in the first step and Lemma 6.3 in the second, we deduce that

$$\mathbb{P}\Big(\max_{j \in \mathscr{S}_m} \big\{0, X^{(j)} - \mathbb{E}[X]\big\} \ge t \Big) \le \frac{1}{t} \mathbb{E}\Big[\max_{j \in \mathscr{S}_m} \big\{0, X^{(j)} - \mathbb{E}[X]\big\} \Big] \le \frac{2\sqrt{n \ln(m+1)}}{t \sqrt{\ln(3/2)}}.$$

Combining this with (6.10) yields

$$\mathbb{P}\Big(X - \mathbb{E}[X] \ge t \Big) \le -\frac{1}{m} \ln \left(1 - \frac{2\sqrt{n \ln(m+1)}}{t\sqrt{\ln(3/2)}} \right) \qquad \forall m \in \mathbb{N}. \qquad (6.11)$$

Since the right-hand side of (6.11) is not well suited for an analytical minimization, we are going to pick a convenient value for m. For notational convenience, set $c := 2\sqrt{n} \big/ \big(t \sqrt{\ln(3/2)} \big)$. Now, choose

$$m_* := \left\lfloor \exp\left(\frac{(1 - e^{-1})^2}{c^2} \right) - 1 \right\rfloor,$$

where $\lfloor \cdot \rfloor$ denotes the floor function, and where $m_* \ge 2$ since $t > \sqrt{29n}$. Inserting $m = m_*$ in the inequality in (6.11), gives

$$\begin{aligned}
\mathbb{P}\Big(X - \mathbb{E}[X] \ge t \Big) &\le -\frac{1}{m_*} \ln \Big(1 - c\sqrt{\ln(m_* + 1)} \Big) \le -\frac{1}{m_*} \ln e^{-1} \\
&< \frac{1}{\exp\left(\frac{(1-e^{-1})^2}{c^2} \right) - 2} \le \exp\left(1 - \frac{(1 - e^{-1})^2}{c^2} \right) \\
&= \exp\left(1 - \frac{(1 - e^{-1})^2 \ln(3/2)\, t^2}{4n} \right) \le e^{1 - \frac{t^2}{29n}}. \qquad (6.12)
\end{aligned}$$

The fourth inequality in (6.12) is due to both the fact that $1/(e^x - 2) \leq e^{1-x}$ for all $x \geq \ln(2/(e-1)) + 1$ (because $e^{x-1}(e-1) - 2$ is non-decreasing in x) and the fact that $(1-e^{-1})^2/c^2 \geq \ln(2/(e-1)) + 1$, which is true since $t > \sqrt{29n}$. □

Remark 6.6. *To derive the tail bound for X in Theorem 6.5, we used the inequality*

$$\mathbb{P}\Big(X - \mathbb{E}[X] \geq t\Big) \;\leq\; -\frac{1}{m}\ln\left(1 - \mathbb{P}\Big(\max_{j\in\mathscr{S}_m}\Big\{0, X^{(j)} - \mathbb{E}[X]\Big\} \geq t\Big)\right) \qquad \forall t > 0$$

and applied Markov's inequality in order to obtain an upper bound on the probability $\mathbb{P}\Big(\max_{j\in\mathscr{S}_m}\Big\{0, X^{(j)} - \mathbb{E}[X]\Big\} \geq t\Big)$. Instead, it might be tempting, in view of the facts that bound (6.8) in Remark 6.4 on $\mathbb{E}\Big[\max_{j\in\mathscr{S}_m} X^{(j)}\Big] - \mathbb{E}[X]$ is lower than the bound in Lemma 6.3 on $\mathbb{E}\Big[\max_{j\in\mathscr{S}_m}\Big\{0, X^{(j)} - \mathbb{E}[X]\Big\}\Big]$ and that

$$\mathbb{P}\left(\max_{j\in\mathscr{S}_m}\Big\{0, X^{(j)} - \mathbb{E}[X]\Big\} \geq t\right) \;=\; \mathbb{P}\left(\max_{j\in\mathscr{S}_m} X^{(j)} - \mathbb{E}[X] \geq t\right) \qquad \forall t > 0,$$

to apply Markov's inequality to $\mathbb{P}\Big(\max_{j\in\mathscr{S}_m} X^{(j)} - \mathbb{E}[X] \geq t\Big)$ aiming to improve the tail bound for X. However, this is not possible since $\max_{j\in\mathscr{S}_m} X^{(j)} - \mathbb{E}[X]$ is usually not non-negative; see Example 2.8

The arguments leading to Theorem 6.5 cannot be easily modified to derive tail bounds for sums of dependent random variables or to deal with inhomogeneous ranges in a non-trivial manner. However, using a simple scaling argument, we obtain the following extension of Theorem 6.5.

Corollary 6.7. *Let $\{a_i\}_{i\in\mathscr{S}_n}$ be a sequence of real constants and let $b \in \mathbb{R}_{>0}$. Further, let $X_1, \dots, X_n$ be n independent random variables such that $a_i \leq X_i \leq b + a_i$ almost surely for all $i \in \mathscr{S}_n$. Set $X := \sum_{i=1}^n X_i$. Then, for any $t \geq 0$,*

(i) $\mathbb{P}\Big(X - \mathbb{E}[X] \geq t\Big) \;\leq\; e^{1-\frac{t^2}{29nb^2}}$.

(ii) $\mathbb{P}\Big(|X - \mathbb{E}[X]| \geq t\Big) \leq 2e^{1-\frac{t^2}{29nb^2}}$.

Proof. For every $i \in \mathscr{S}_n$, define the random variable $Y_i := (X_i - a_i)/b$. Obviously, Y_i's are independent and take values in $[0,1]$. Hence, applying Theorem 6.5 and making use of linearity of expectation, we deduce for all $t \geq 0$ that

$$\begin{aligned}\mathbb{P}\Big(X - \mathbb{E}[X] \geq t\Big) &= \mathbb{P}\left(\sum_{i=1}^n \Big(bY_i + a_i\Big) - \sum_{i=1}^n \Big(b\mathbb{E}[Y_i] + a_i\Big) \geq t\right) \\ &= \mathbb{P}\left(\sum_{i=1}^n Y_i - \mathbb{E}\left[\sum_{i=1}^n Y_i\right] \geq \frac{t}{b}\right) \;\leq\; e^{1-\frac{t^2}{29nb^2}}. \qquad (6.13)\end{aligned}$$

Replacing each X_i with $-X_i$ and noticing that $(-b-a_i) \le -X_i \le b+(-b-a_i)$ almost surely for all $i \in \mathscr{S}_n$, we obtain for all $t \ge 0$

$$\mathbb{P}\Big(X - \mathbb{E}[X] \le -t\Big) \;\le\; e^{1-\frac{t^2}{29nb^2}}. \tag{6.14}$$

In view of Boole's inequality, we can combine the bounds in (6.13) and (6.14) to get

$$\mathbb{P}\Big(|X - \mathbb{E}[X]| \ge t\Big) \;\le\; \mathbb{P}\Big(X - \mathbb{E}[X] \ge t\Big) + \mathbb{P}\Big(X - \mathbb{E}[X] \le -t\Big) \;\le\; 2e^{1-\frac{t^2}{29nb^2}}$$

for all $t \ge 0$. Hence the corollary. $\square$

7 | Conclusion

The suitability of four different proof techniques for deriving concentration inequalities for sums of bounded random variables with various dependence structures and possibly distinct ranges was investigated in the present thesis. It has emerged that:

- Except for the moment method, none of the considered approaches seem to be adequate to prove sharp concentration bounds when the random variables are supported on finite intervals of different length; compare, e.g., Theorem 3.13 to Theorem 4.3.

- Only the convex map method can be applied to derive tail bounds for arbitrarily dependent, e.g., k-wise independent, random variables. Interestingly, these bounds can be of a different type than the typical exponential Chernoff-type bounds obtained with other methods; compare Theorem 3.1 to Theorem 5.9. Understanding how they relate to each other remains an open problem; see Remark 5.10 and the discussion thereafter.

- Despite its limitation in comparison to the moment and the convex maps methods, the combinatorial approach could still be shown to perform better than originally thought; compare Theorem 3.3 in Impagliazzo and Kabanets (2010) to Theorem 3.1.

- The method based on differential privacy clearly does not compare favorably to the other approaches we considered. Not only does it apply exclusively to independent random variables but it also leads to weaker bounds regardless of our efforts; compare Theorem 2.6 in Steinke and Ullman (2017) to Theorem 6.5 (to see our slight improvement) and to the Chernoff-Hoeffding bound in Corollary 3.2.

Bibliography

Alon, N. and Spencer, J.H. (2008). *The probabilistic method, 3rd ed.* Wiley-Interscience, New York.

Azuma, K. (1967). *Weighted sums of certain dependent random variables*, Tohoku Mathematical Journal **19**, pp. 357–367.

Bassily, R., Nissim, K., Smith, A., Steinke, T., Stemmer, U., and Ullman, J. (2016). *Algorithmic stability for adaptive data analysis*, Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, pp. 1046–1059.

Bennett, G. (1962). *Probability inequalities for the sum of independent random variables*, Journal of the American Statistical Association **57**, pp. 33–45.

Bernstein, S.N. (1964). *Collected works, vol. IV.* Nauka, Moscow (in Russian).

Chernoff, H. (1952). *A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations*, Annals of Mathematical Statistics **23**, pp. 493–507.

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

Çinlar, E. (2011). *Probability and stochastics.* Graduate Texts in Mathematics **261**. Springer, New York.

Cover, T.M. and Thomas, J.A. (1991). *Elements of information theory, 2nd ed.* Wiley-Interscience, New York.

Doerr, B. (2018). *Probabilistic tools for the analysis of randomized optimization heuristics.* Available at https://arxiv.org/pdf/1801.06733v3.pdf.

Dubhashi, D. and Desh, R. (1996). *Balls and bins: A study in negative dependence*, Basic Research in Computer Science Report Series **RS-96-26**.

Dubhashi, D. and Panconesi, A. (2009). *Concentration of Measure for the Analysis of Randomized Algorithms.* Cambridge University Press, Cambridge, New York.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). *Calibrating noise to sensitivity in private data analysis.* In: Halevi, S. and Rabin, T. (eds) Theory of Cryptography Conference. Lecture Notes in Computer Science **3876**, pp. 265–284. Springer, Berlin, Heidelberg.

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

Feller, W. (1968). *An introduction to probability theory and its applications, vol. I, 3rd ed.* Wiley Series in Probability and Statistics, John Wiley & Sons, New York.

Feller, W. (1971). *An introduction to probability theory and its applications, vol. II, 2nd ed.* Wiley Series in Probability and Statistics, John Wiley & Sons, New York.

Gut, A. (2005). *Probability: A graduate course.* Springer Texts in Statistics, Springer, New York.

Hoeffding, W. (1963). *Probability inequalities for sums of bounded random variables*, Journal of the American Statistical Association **58**, pp. 13–30.

Impagliazzo, R. and Kabanets, V. (2010). *Constructive proofs of concentration bounds.* In: Serna, M., Shaltie, R., Jansen, K., Rolim, J. (eds) Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques. RANDOM 2010, APPROX 2010. Lecture Notes in Computer Science **6302**, pp. 617–631. Springer, Berlin, Heidelberg.

Jacod, J. and Protter, P. (2004). *Probability essentials.* Springer, Berlin, Heidelberg.

Janson, S. (2016). *Large deviation inequalities for sums of indicator variables.* Available at https://arxiv.org/abs/1609.00533.

Janson, S. (2018). *Tail bounds for sums of geometric and exponential variables*, Statistics & Probability Letters **135**, pp. 1–6.

Joag-Dev, K. and Proschan, F. (1983). *Negative association of random variables with applications*, Annals of Statistics **11:1**, pp. 286–295.

Klenke, A. (2014). *Probability theory: A comprehensive course, 2nd ed.* Springer, London.

Linial, N. and Luria, Z. (2014). *Chernoff's inequality – a very elementary proof.* Available at https://arxiv.org/abs/1403.7739.

McDiarmid, C. (1989). *On the method of bounded differences*, London Mathematical Society Lecture Note Series **141**, pp. 148–188.

McDiarmid, C. (1998). *Concentration.* In: Habib, M., McDiarmid, C., Ramirez-Alfonsin, J., Reed, B. (eds) Probabilistic Methods for Algorithmic Discrete Mathematics. Algorithms and Combinatorics **16**, pp. 195–248. Springer, Berlin, Heidelberg.

McSherry, F. and Talwar, K. (2007). *Mechanism design via differential privacy*, Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science, pp. 94–103.

Mitzenmacher, M. and Upfal, E. (2005). *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis.* Cambridge University Press, Cambridge, New York.

Motwani, R. and Raghavan, P. (1995). *Randomized algorithms.* Cambridge University Press, Cambridge, New York.

Mulzer, W. (2018). *Five proofs of Chernoff's bound with applications*, Bulletin of the EATCS (BEATCS) **124**, (February 2018).

Pelekis, C. and Ramon, J. (2017). *Hoeffding's inequality for sums of dependent random variables*, Mediterranean Journal of Mathematics **14:243**.

Schmidt, J.P., Siegel, A., and Srinivasan, A. (1995). *Chernoff–Hoeffding bounds for applications with limited independence*, SIAM Journal on Discrete Mathematics **8:2**, pp. 223–250.

Steinke, T. and Ullman, J. (2017). *Subgaussian tail bounds via stability arguments.* Available at https://arxiv.org/abs/1701.03493v2.

Vadhan, S.P. (2012). *Pseudorandomness*, Foundations and Trends in Theoretical Computer Science **7:1–3**, pp. 1–336. Available at http://people.seas.harvard.edu/~salil/pseudorandomness.