

On Refined Versions of the Azuma-Hoeffding Inequality with Applications in Information Theory

Igal Sason

Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa 32000, Israel
E-mail: sason@ee.technion.ac.il

Abstract

This paper derives some refined versions of the Azuma-Hoeffding inequality for discrete-parameter martingales with uniformly bounded jumps, and it considers some of their potential applications in information theory and related topics. The first part of this paper derives these refined inequalities, followed by a discussion on their relations to some classical results in probability theory. It also considers a geometric interpretation of some of these inequalities, providing an insight on the interconnections between them. The second part exemplifies the use of these refined inequalities in the context of hypothesis testing, information theory, and communication. The paper is concluded with a discussion on some directions for further research. This work is meant to stimulate the use of some refined versions of the Azuma-Hoeffding inequality in information-theoretic aspects.

Index Terms

Concentration of measures, error exponents, Fisher information, hypothesis testing, information divergence, large deviations, martingales, moderate deviations principle.

I. INTRODUCTION

Inequalities providing upper bounds on probabilities of the type $\mathbb{P}(|X - \bar{x}| \geq t)$ (or $\mathbb{P}(X - \bar{x} \geq t)$) for a random variable (RV) X , where \bar{x} denotes the expectation or median of X , have been among the main tools of probability theory. These inequalities are known as concentration inequalities, and they have been subject to interesting developments in probability theory. Very roughly speaking, the concentration of measure phenomenon can be stated in the following simple way: “A random variable that depends in a smooth way on many independent random variables (but not too much on any of them) is essentially constant” [55]. The exact meaning of such a statement clearly needs to be clarified rigorously, but it will often mean that such a random variable X concentrates around \bar{x} in a way that the probability of the event $\{|X - \bar{x}| > t\}$ decays exponentially in t (for $t \geq 0$). The foundations in concentration of measures have been introduced, e.g., in [3, Chapter 7], [12, Chapter 2], [13], [33], [36], [37, Chapter 5], [38], [54] and [55]. Concentration inequalities are also at the core of probabilistic analysis of randomized algorithms (see, e.g., [3], [19], [38] and [47]).

The Chernoff bounds provide sharp concentration inequalities when the considered RV X can be expressed as a sum of n independent and bounded RVs. However, the situation is clearly more complex for non-product measures where the concentration property may not exist. Several techniques have been developed to prove concentration of measures. Among several methodologies, these include Talagrand’s concentration inequalities for product measures (e.g., [54] and [55] with some information-theoretic applications in [31] and [32]), logarithmic-Sobolev inequalities (e.g., [19, Chapter 14], [33, Chapter 5] and [36] with information-theoretic aspects in [28], [29]), transportation-cost inequalities which originated from information theory (e.g., [19, Chapters 12, 13] and [33, Chapter 6]), and the martingale approach (e.g., [3, Chapter 7], [38] with information-theoretic aspects in, e.g., [35], [46], [47], [60]). This paper mainly considers the last methodology, focusing on discrete-time martingales with bounded jumps.

The Azuma-Hoeffding inequality is by now a well-known methodology that has been often used to prove concentration phenomena for discrete-time martingales whose jumps are bounded almost surely. It is due to Hoeffding [26] who proved this inequality for $X = \sum_{i=1}^n X_i$ where $\{X_i\}$ are independent and bounded RVs, and Azuma [7] later extended it to bounded-difference martingales. It is noted that the Azuma-Hoeffding inequality for

a bounded martingale-difference sequence was extended to centering sequences with bounded differences [39]; this extension provides sharper concentration results for, e.g., sequences that are related to sampling without replacement. Some relative entropy and exponential deviation bounds were derived in [30] for an important class of Markov chains, and these bounds are essentially identical to the Hoeffding inequality in the special case of i.i.d. RVs. A common method for proving concentration of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ of n independent RVs, around the expected value $\mathbb{E}[f]$, where the function f is characterized by bounded differences whenever the n -dimensional vectors differ in only one coordinate, is called McDiarmid's inequality or the 'independent bounded differences inequality' (see [38, Theorem 3.1]). This inequality was proved (with some possible extensions) via the martingale approach (see [38, Section 3.5]). Although the proof of this inequality has some similarity to the proof of the Azuma-Hoeffding inequality, the former inequality is stated under a condition which provides an improvement by a factor of 4 in the exponent. Some of its nice applications to algorithmic discrete mathematics were exemplified in [38, Section 3].

The use of the Azuma-Hoeffding inequality was introduced to the computer science literature in [51] in order to prove concentration, around the expected value, of the chromatic number for random graphs. The chromatic number of a graph is defined to be the minimal number of colors that is required to color all the vertices of this graph so that no two vertices which are connected by an edge have the same color, and the ensemble for which concentration was demonstrated in [51] was the ensemble of random graphs with n vertices such that any ordered pair of vertices in the graph is connected by an edge with a fixed probability p for some $p \in (0, 1)$. It is noted that the concentration result in [51] was established without knowing the expected value over this ensemble. The migration of this bounding inequality into coding theory, especially for exploring some concentration phenomena that are related to the analysis of codes defined on graphs and iterative message-passing decoding algorithms, was initiated in [35], [46] and [52]. During the last decade, the Azuma-Hoeffding inequality has been extensively used for proving concentration of measures in coding theory (see, e.g., [47, Appendix C] and references therein). In general, all these concentration inequalities serve to justify theoretically the ensemble approach of codes defined on graphs. However, much stronger concentration phenomena are observed in practice. The Azuma-Hoeffding inequality was also recently used in [57] for the analysis of probability estimation in the rare-events regime where it was assumed that an observed string is drawn i.i.d. from an unknown distribution, but the alphabet size and the source distribution both scale with the block length (so the empirical distribution does not converge to the true distribution as the block length tends to infinity). In [59], the Azuma-Hoeffding inequality was used to derive achievable rates and random coding error exponents for non-linear additive white Gaussian noise channels. This analysis was followed by another recent work of the same authors [60] who used some other concentration inequalities, for discrete-parameter martingales with bounded jumps, to derive achievable rates and random coding error exponents for non-linear Volterra channels (where their bounding technique can be also applied to intersymbol-interference (ISI) channels, as noted in [60]).

This work derives some refined versions of the Azuma-Hoeffding inequality, and it exemplifies some of their possible applications in information theory and related topics. The paper is structured as follows: Section II presents briefly some background that is essential to the analysis in this work. The core of the paper is divided into two parts. The first part includes Sections III and IV. Section III derives some refined versions of the Azuma-Hoeffding inequality, and it considers interconnections between these bounds. Section IV considers some connections between the concentration inequalities that are introduced in Section III to the method of types, a central limit theorem for martingales, the law of iterated logarithm, the moderate deviations principle for i.i.d. real-valued random variables, and some previously-reported concentration inequalities for discrete-parameter martingales with bounded jumps. Section V forms the second part of this work, applying the concentration inequalities from Section III to information theory and some related topics. This paper is summarized in Section VI, followed by a discussion on some topics, mainly related to information theory and coding, for further research. Various mathematical details of the analysis are relegated to the appendices. This work is meant to stimulate the derivation of some new refined versions of concentration inequalities for martingales with a further consideration of their possible applications in aspects that are related to information theory, communications and coding.

II. PRELIMINARIES

In the following, we present briefly some background that is essential to the analysis in this work, followed by some examples that serve to motivate the continuation of this paper.

A. Martingales

This sub-section provides a short background on martingales to set definitions and notation. We will not use any result about martingales beyond the definition and few basic properties that will be mentioned explicitly.

Definition 1: [Martingale] Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A martingale sequence is a sequence X_0, X_1, \dots of random variables (RVs) and corresponding sub σ -algebras $\mathcal{F}_0, \mathcal{F}_1, \dots$ that satisfy the following conditions:

- 1) $X_i \in \mathbb{L}^1(\Omega, \mathcal{F}_i, \mathbb{P})$ for every i , i.e., each X_i is defined on the same sample space Ω , it is measurable with respect to the σ -algebra \mathcal{F}_i (i.e., X_i is \mathcal{F}_i -measurable) and $\mathbb{E}[|X_i|] = \int_{\Omega} |X_i(\omega)| d\mathbb{P}(\omega) < \infty$.
- 2) $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \dots$ (this sequence is called a filtration).
- 3) For all $i \in \mathbb{N}$, $X_{i-1} = \mathbb{E}[X_i | \mathcal{F}_{i-1}]$ almost surely (a.s.).

In this case, it is written that $\{X_i, \mathcal{F}_i\}_{i=0}^{\infty}$ or $\{X_i, \mathcal{F}_i\}_{i \in \mathbb{N}_0}$ (with $\mathbb{N}_0 \triangleq \mathbb{N} \cup \{0\}$) is a martingale sequence (the inclusion of X_{∞} and \mathcal{F}_{∞} in the martingale is not required here).

Remark 1: Since $\{\mathcal{F}_i\}_{i=0}^{\infty}$ forms a filtration, then it follows from the tower principle for conditional expectations that a.s.

$$X_j = \mathbb{E}[X_i | \mathcal{F}_j], \quad \forall i > j.$$

Also for every $i \in \mathbb{N}$, $\mathbb{E}[X_i] = \mathbb{E}[\mathbb{E}[X_i | \mathcal{F}_{i-1}]] = \mathbb{E}[X_{i-1}]$, so the expectation of a martingale sequence stays constant.

Remark 2: One can generate martingale sequences by the following procedure: Given a RV $X \in \mathbb{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ and an arbitrary filtration of sub σ -algebras $\{\mathcal{F}_i\}_{i=0}^{\infty}$, let

$$X_i = \mathbb{E}[X | \mathcal{F}_i], \quad \forall i \in \{0, 1, \dots\}.$$

Then, the sequence X_0, X_1, \dots forms a martingale since

- 1) The RV $X_i = \mathbb{E}[X | \mathcal{F}_i]$ is \mathcal{F}_i -measurable, and also $\mathbb{E}[|X_i|] \leq \mathbb{E}[|X|] < \infty$ (since conditioning reduces the expectation of the absolute value).
- 2) By construction $\{\mathcal{F}_i\}_{i=0}^{\infty}$ is a filtration.
- 3) For every $i \in \mathbb{N}$

$$\begin{aligned} \mathbb{E}[X_i | \mathcal{F}_{i-1}] &= \mathbb{E}[\mathbb{E}[X | \mathcal{F}_i] | \mathcal{F}_{i-1}] \\ &= \mathbb{E}[X | \mathcal{F}_{i-1}] \quad (\text{since } \mathcal{F}_{i-1} \subseteq \mathcal{F}_i) \\ &= X_{i-1} \quad \text{a.s.} \end{aligned}$$

Remark 3: In continuation to Remark 2, one can choose $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_n = \mathcal{F}$, so that X_0, X_1, \dots, X_n is a martingale sequence where

$$\begin{aligned} X_0 &= \mathbb{E}[X | \mathcal{F}_0] = \mathbb{E}[X] \quad (\text{since } X \text{ is independent of } \mathcal{F}_0) \\ X_n &= \mathbb{E}[X | \mathcal{F}_n] = X \quad \text{a.s.} \quad (\text{since } X \text{ is } \mathcal{F}\text{-measurable}). \end{aligned}$$

In this case, one gets a martingale sequence where the first element is the expected value of X , and the last element of the sequence is X itself (a.s.). This has the following interpretation: At the beginning, one doesn't know anything about X , so it is initially estimated by its expectation. At each step more and more information about X is revealed until one is able to specify it exactly (a.s.).

B. The Azuma-Hoeffding Inequality

The Azuma-Hoeffding inequality¹ is a useful concentration inequality for bounded-difference martingales. It was proved in [26] for independent bounded random variables, followed by a discussion on sums of dependent random variables; this inequality was later derived in [7] for the more general setting of bounded-difference martingales. In the following, this inequality is introduced.

¹The Azuma-Hoeffding inequality is also known as (a.k.a.) Azuma's inequality. Since it is referred numerous times in this paper, it will be named Azuma's inequality for the sake of brevity.

Theorem 1: [Azuma's inequality] Let $\{X_k, \mathcal{F}_k\}_{k=0}^{\infty}$ be a discrete-parameter real-valued martingale sequence such that for every $k \in \mathbb{N}$, the condition $|X_k - X_{k-1}| \leq d_k$ holds a.s. for some non-negative constants $\{d_k\}_{k=1}^{\infty}$. Then, for every $n \in \mathbb{N}$ and $r \geq 0$,

$$\mathbb{P}(|X_n - X_0| \geq r) \leq 2 \exp\left(-\frac{r^2}{2 \sum_{k=1}^n d_k^2}\right). \quad (1)$$

For a proof of Theorem 1 see, e.g., [12, Chapter 2] and [13]. It will be revisited in the next section for the derivation of some refined versions of Azuma's inequality. The reader is referred to [3, Chapter 11], [12, Chapter 2], [13] and [38] for surveys on concentration inequalities for (sub/ super) martingales.

Remark 4: In [38, Theorem 3.13], Azuma's inequality is stated as follows: Let $\{Y_k, \mathcal{F}_k\}_{k=0}^{\infty}$ be a martingale-difference sequence with $Y_0 = 0$ (i.e., Y_k is \mathcal{F}_k -measurable, $\mathbb{E}[|Y_k|] < \infty$ and $\mathbb{E}[Y_k | \mathcal{F}_{k-1}] = 0$ a.s. for every $k \in \mathbb{N}$). Assume that, for every $k \in \mathbb{N}$, there exist numbers $a_k, b_k \in \mathbb{R}$ such that a.s. $a_k \leq Y_k \leq b_k$. Then, for every $r \geq 0$,

$$\mathbb{P}\left(\left|\sum_{k=1}^n Y_k\right| \geq r\right) \leq 2 \exp\left(-\frac{2r^2}{\sum_{k=1}^n (b_k - a_k)^2}\right). \quad (2)$$

Hence, consider a discrete-parameter real-valued martingale sequence $\{X_k, \mathcal{F}_k\}_{k=0}^{\infty}$ where $a_k \leq X_k - X_{k-1} \leq b_k$ a.s. for every $k \in \mathbb{N}$. Let $Y_k \triangleq X_k - X_{k-1}$ for every $k \in \mathbb{N}$. This implies that $\{Y_k, \mathcal{F}_k\}_{k=0}^{\infty}$ is a martingale-difference sequence. From (2), it follows that for every $r \geq 0$,

$$\mathbb{P}(|X_n - X_0| \geq r) \leq 2 \exp\left(-\frac{2r^2}{\sum_{k=1}^n (b_k - a_k)^2}\right). \quad (3)$$

according to the setting in Theorem 1, $a_k = -d_k$ and $b_k = d_k$ for every $k \in \mathbb{N}$, which implies the equivalence between (1) and (3).

As a special case of Theorem 1, let $\{X_k, \mathcal{F}_k\}_{k=0}^{\infty}$ be a martingale sequence, and assume that there exists a constant $d > 0$ such that a.s., for every $k \in \mathbb{N}$, $|X_k - X_{k-1}| \leq d$. Then, for every $n \in \mathbb{N}$ and $\alpha \geq 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) \leq 2 \exp\left(-\frac{\alpha^2}{2d^2}\right). \quad (4)$$

Example 1: Let $\{Y_i\}_{i=0}^{\infty}$ be i.i.d. binary random variables which get the values $\pm d$, for some constant $d > 0$, with equal probability. Let $X_k = \sum_{i=0}^k Y_i$ for $k \in \{0, 1, \dots\}$, and define the natural filtration $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \dots$ where

$$\mathcal{F}_k = \sigma(Y_0, \dots, Y_k), \quad \forall k \in \{0, 1, \dots\}$$

is the σ -algebra that is generated by the random variables Y_0, \dots, Y_k . Note that $\{X_k, \mathcal{F}_k\}_{k=0}^{\infty}$ is a martingale sequence, and (a.s.) $|X_k - X_{k-1}| = |Y_k| = d, \forall k \in \mathbb{N}$. It therefore follows from Azuma's inequality in (4) that

$$\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) \leq 2 \exp\left(-\frac{\alpha^2}{2d^2}\right). \quad (5)$$

for every $\alpha \geq 0$ and $n \in \mathbb{N}$. From the central limit theorem (CLT), since the RVs $\{Y_i\}_{i=0}^{\infty}$ are i.i.d. with zero mean and variance d^2 , then $\frac{1}{\sqrt{n}}(X_n - X_0) = \frac{1}{\sqrt{n}} \sum_{k=1}^n Y_k$ converges in distribution to $\mathcal{N}(0, d^2)$. Therefore, for every $\alpha \geq 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) = 2Q\left(\frac{\alpha}{d}\right) \quad (6)$$

where

$$Q(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_x^{\infty} \exp\left(-\frac{t^2}{2}\right) dt, \quad \forall x \in \mathbb{R} \quad (7)$$

is the probability that a zero-mean and unit-variance Gaussian RV is larger than x . Since the following exponential upper and lower bounds on the Q-function hold

$$\frac{1}{\sqrt{2\pi}} \frac{x}{1+x^2} \cdot e^{-\frac{x^2}{2}} < Q(x) < \frac{1}{\sqrt{2\pi} x} \cdot e^{-\frac{x^2}{2}}, \quad \forall x > 0 \quad (8)$$

then it follows from (6) that the exponent on the right-hand side of (5) is the exact exponent in this example.

Example 2: In continuation to Example 1, let $\gamma \in (0, 1]$, and let us generalize this example by considering the case where the i.i.d. binary RVs $\{Y_i\}_{i=0}^{\infty}$ have the probability law

$$\mathbb{P}(Y_i = +d) = \frac{\gamma}{1 + \gamma}, \quad \mathbb{P}(Y_i = -\gamma d) = \frac{1}{1 + \gamma}.$$

Hence, it follows that the i.i.d. RVs $\{Y_i\}$ have zero mean and variance $\sigma^2 = \gamma d^2$ as in Example 1. Let $\{X_k, \mathcal{F}_k\}_{k=0}^{\infty}$ be defined similarly to Example 1, so that it forms a martingale sequence. Based on the CLT, $\frac{1}{\sqrt{n}}(X_n - X_0) = \frac{1}{\sqrt{n}} \sum_{k=1}^n Y_k$ converges weakly to $\mathcal{N}(0, \gamma d^2)$, so for every $\alpha \geq 0$

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X_0| \geq \alpha \sqrt{n}) = 2Q\left(\frac{\alpha}{\sqrt{\gamma}d}\right). \quad (9)$$

From the exponential upper and lower bounds of the Q-function in (8), the right-hand side of (9) scales exponentially like $e^{-\frac{\alpha^2}{2\gamma d^2}}$. Hence, the exponent in this example is improved by a factor $\frac{1}{\gamma}$ as compared Azuma's inequality (that is the same as in Example 1 since $|X_k - X_{k-1}| \leq d$ for every $k \in \mathbb{N}$). This indicates on the possible refinement of Azuma's inequality by introducing an additional constraint on the second moment. This route was studied extensively in the probability literature, and it is further studied in Section III.

Example 2 serves to motivate the introduction of an additional constraint on the conditional variance of a martingale sequence, i.e., adding an inequality constraint of the form

$$\text{Var}(X_k | \mathcal{F}_{k-1}) = \mathbb{E}[(X_k - X_{k-1})^2 | \mathcal{F}_{k-1}] \leq \gamma d^2$$

where $\gamma \in (0, 1]$ is a constant. Note that since, by assumption $|X_k - X_{k-1}| \leq d$ a.s. for every $k \in \mathbb{N}$, then the additional constraint becomes active when $\gamma < 1$ (i.e., if $\gamma = 1$, then this additional constraint is redundant, and it coincides with the setting of Azuma's inequality with a fixed d_k (i.e., $d_k = d$).

C. Sub/ Super Martingales

Sub and super martingales require the first two conditions in Definition 1, and the equality in the third condition of Definition 1 is relaxed to one of the following inequalities:

- $\mathbb{E}[X_i | \mathcal{F}_{i-1}] \geq X_{i-1}$ holds a.s. for sub-martingales.
- $\mathbb{E}[X_i | \mathcal{F}_{i-1}] \leq X_{i-1}$ holds a.s. for super-martingales.

Clearly, every process that is both a sub and super-martingale is a martingale. Furthermore, $\{X_i, \mathcal{F}_i\}$ is a sub-martingale if and only if $\{-X_i, \mathcal{F}_i\}$ is a super-martingale. The following properties are direct consequences of Jensen's inequality for conditional expectations:

- If $\{X_i, \mathcal{F}_i\}$ is a martingale, h is a convex (concave) function and $\mathbb{E}[|h(X_i)|] < \infty$, then $\{h(X_i), \mathcal{F}_i\}$ is a sub (super) martingale.
- If $\{X_i, \mathcal{F}_i\}$ is a super-martingale, h is monotonic increasing and concave, and $\mathbb{E}[|h(X_i)|] < \infty$, then $\{h(X_i), \mathcal{F}_i\}$ is a super-martingale. Similarly, if $\{X_i, \mathcal{F}_i\}$ is a sub-martingale, h is monotonic increasing and convex, and $\mathbb{E}[|h(X_i)|] < \infty$, then $\{h(X_i), \mathcal{F}_i\}$ is a sub-martingale.

III. REFINED VERSIONS OF AZUMA'S INEQUALITY

A. First Refinement of Azuma's Inequality

Theorem 2: Let $\{X_k, \mathcal{F}_k\}_{k=0}^{\infty}$ be a discrete-parameter real-valued martingale. Assume that, for some constants $d, \sigma > 0$, the following two requirements are satisfied a.s.

$$\begin{aligned} |X_k - X_{k-1}| &\leq d, \\ \text{Var}(X_k | \mathcal{F}_{k-1}) &= \mathbb{E}[(X_k - X_{k-1})^2 | \mathcal{F}_{k-1}] \leq \sigma^2 \end{aligned}$$

for every $k \in \{1, \dots, n\}$. Then, for every $\alpha \geq 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) \leq 2 \exp\left(-n D\left(\frac{\delta + \gamma}{1 + \gamma} \parallel \frac{\gamma}{1 + \gamma}\right)\right) \quad (10)$$

where

$$\gamma \triangleq \frac{\sigma^2}{d^2}, \quad \delta \triangleq \frac{\alpha}{d} \quad (11)$$

and

$$D(p||q) \triangleq p \ln\left(\frac{p}{q}\right) + (1-p) \ln\left(\frac{1-p}{1-q}\right), \quad \forall p, q \in [0, 1] \quad (12)$$

is the divergence (a.k.a. relative entropy or Kullback-Leibler distance) between the two probability distributions $(p, 1-p)$ and $(q, 1-q)$. If $\delta > 1$, then the probability on the left-hand side of (10) is equal to zero.

Remark 5: The idea of the proof of Theorem 2 is essentially similar to the proof of [18, Corollary 2.4.7]. The reasons for introducing a full proof in the following are as follows:

- The geometric interpretation that is associated with the proof of Theorem 2 provides an insight on the underlying connections between this theorem and some other results (e.g., Theorem 3 and Azuma's inequality).
- The technique that is used to derive Theorem 2 serves as a starting point for the derivation of Theorem 4. Then, it is shown that under a certain sufficient condition, the exponent in Theorem 4 is better than the one in Theorem 2. This will be also exemplified numerically.
- Some of the inequalities obtained along the proof of Theorem 2 are meaningful in their own right. They serve to demonstrate, later in this work, the underlying connections of Theorem 2 with some other concentration inequalities. These inequalities are also helpful for some applications discussed in the continuation to this work.
- The proof of Theorem 2 is of interest since it indicates that it is possible to improve the exponent of inequality (10) by imposing some additional assumptions on the conditional distribution of the jumps $\xi_k \triangleq X_k - X_{k-1}$ given \mathcal{F}_{k-1} (see the first item in Section VI-B).
- The inclusion of Theorem 2 and its proof at this stage makes the material self-contained, and enables the use of the same set of notation throughout the paper.

Remark 6: From the above conditions then without any loss of generality, $\sigma^2 \leq d^2$ and therefore $\gamma \in (0, 1]$.

Proof: $X_n - X_0 = \sum_{k=1}^n \xi_k$ where $\xi_k = X_k - X_{k-1}$ for $k = 1, \dots, n$. By assumption $|\xi_k| \leq d$ a.s. for some $d > 0$, and also for $k = 1, \dots, n$

$$\begin{aligned} & \mathbb{E}[\xi_k | \mathcal{F}_{k-1}] \\ &= \mathbb{E}[X_k | \mathcal{F}_{k-1}] - \mathbb{E}[X_{k-1} | \mathcal{F}_{k-1}] \\ &= \mathbb{E}[X_k | \mathcal{F}_{k-1}] - X_{k-1} \quad (\text{since } X_{k-1} \text{ is } \mathcal{F}_{k-1}\text{-measurable}) \\ &= X_{k-1} - X_{k-1} = 0. \end{aligned} \quad (13)$$

Based on Chernoff's inequality, it follows that for every $\alpha \geq 0$

$$\begin{aligned} & \mathbb{P}(X_n - X_0 \geq \alpha n) \\ &= \mathbb{P}\left(\sum_{k=1}^n \xi_k \geq \alpha n\right) \\ &\leq \exp(-\alpha n t) \mathbb{E}\left[\exp\left(t \sum_{k=1}^n \xi_k\right)\right], \quad \forall t \geq 0. \end{aligned} \quad (14)$$

For every $t \geq 0$

$$\begin{aligned} & \mathbb{E}\left[\exp\left(t \sum_{k=1}^n \xi_k\right)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\exp\left(t \sum_{k=1}^n \xi_k\right) | \mathcal{F}_{n-1}\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\exp\left(t \sum_{k=1}^{n-1} \xi_k\right) \exp(t\xi_n) | \mathcal{F}_{n-1}\right]\right] \\ &= \mathbb{E}\left[\exp\left(t \sum_{k=1}^{n-1} \xi_k\right) \mathbb{E}[\exp(t\xi_n) | \mathcal{F}_{n-1}]\right] \end{aligned} \quad (15)$$

where the last transition holds since $Y = \exp(t \sum_{k=1}^{n-1} \xi_k)$ is \mathcal{F}_{n-1} -measurable. The measurability of Y is due to fact that $\xi_k \triangleq X_k - X_{k-1}$ is \mathcal{F}_k -measurable for every $k \in \mathbb{N}$, and $\mathcal{F}_k \subseteq \mathcal{F}_{n-1}$ for $0 \leq k \leq n-1$ since $\{\mathcal{F}_k\}_{k=0}^n$ is a filtration; hence, the RV $\sum_{k=1}^{n-1} \xi_k$ and its exponentiation (Y) are both \mathcal{F}_{n-1} -measurable, and a.s. $\mathbb{E}[XY|\mathcal{F}_{n-1}] = Y \mathbb{E}[X|\mathcal{F}_{n-1}]$.

From Bennett's inequality [9] (see, e.g., [18, Lemma 2.4.1]), if X is a real-valued random variable with $\bar{x} = \mathbb{E}(X)$ and $\mathbb{E}[(X - \bar{x})^2] \leq \sigma^2$ for some $\sigma > 0$, and $X \leq b$ a.s. for some $b \in \mathbb{R}$, then for every $\lambda \geq 0$

$$\mathbb{E}[e^{\lambda X}] \leq \frac{e^{\lambda \bar{x}} \left[(b - \bar{x})^2 \exp^{-\frac{\lambda \sigma^2}{b - \bar{x}}} + \sigma^2 e^{\lambda(b - \bar{x})} \right]}{(b - \bar{x})^2 + \sigma^2}. \quad (16)$$

Applying Bennett's inequality for the conditional law of ξ_k given the σ -algebra \mathcal{F}_{k-1} , since $\mathbb{E}[\xi_k|\mathcal{F}_{k-1}] = 0$, $\text{Var}[\xi_k|\mathcal{F}_{k-1}] \leq \sigma^2$ and $\xi_k \leq d$ a.s. for $k \in \mathbb{N}$, then a.s.

$$\mathbb{E}[\exp(t\xi_k) | \mathcal{F}_{k-1}] \leq \frac{\sigma^2 \exp(td) + d^2 \exp\left(-\frac{t\sigma^2}{d}\right)}{d^2 + \sigma^2}. \quad (17)$$

Hence, it follows from (15) and (17) that, for every $t \geq 0$,

$$\mathbb{E} \left[\exp \left(t \sum_{k=1}^n \xi_k \right) \right] \leq \left(\frac{\sigma^2 \exp(td) + d^2 \exp\left(-\frac{t\sigma^2}{d}\right)}{d^2 + \sigma^2} \right) \mathbb{E} \left[\exp \left(t \sum_{k=1}^{n-1} \xi_k \right) \right]$$

and, by induction, it follows that for every $t \geq 0$

$$\mathbb{E} \left[\exp \left(t \sum_{k=1}^n \xi_k \right) \right] \leq \left(\frac{\sigma^2 \exp(td) + d^2 \exp\left(-\frac{t\sigma^2}{d}\right)}{d^2 + \sigma^2} \right)^n.$$

From the definition of γ in (11), this inequality is rewritten as

$$\mathbb{E} \left[\exp \left(t \sum_{k=1}^n \xi_k \right) \right] \leq \left(\frac{\gamma \exp(td) + \exp(-\gamma td)}{1 + \gamma} \right)^n, \quad \forall t \geq 0. \quad (18)$$

Let $x \triangleq td$ (so $x \geq 0$). Combining (14) with (18) gives that, for every $\alpha \geq 0$ (based on the definition of δ in (11), $\alpha t = \delta x$),

$$\mathbb{P}(X_n - X_0 \geq \alpha n) \leq \left(\frac{\gamma \exp((1 - \delta)x) + \exp(-(\gamma + \delta)x)}{1 + \gamma} \right)^n, \quad \forall x \geq 0. \quad (19)$$

Consider first the case where $\delta = 1$ (i.e., $\alpha = d$), then (19) is particularized to

$$\mathbb{P}(X_n - X_0 \geq dn) \leq \left(\frac{\gamma + \exp(-(\gamma + 1)x)}{1 + \gamma} \right)^n, \quad \forall x \geq 0$$

and the tightest bound within this form is obtained in the limit where $x \rightarrow \infty$. This provides the inequality

$$\mathbb{P}(X_n - X_0 \geq dn) \leq \left(\frac{\gamma}{1 + \gamma} \right)^n. \quad (20)$$

Otherwise, if $\delta \in [0, 1)$, the minimization of the base of the exponent on the right-hand side of (19) w.r.t. the free non-negative parameter x yields that the optimized value is

$$x = \left(\frac{1}{1 + \gamma} \right) \ln \left(\frac{\gamma + \delta}{\gamma(1 - \delta)} \right) \quad (21)$$

and its substitution into the right-hand side of (19) gives that, for every $\alpha \geq 0$,

$$\begin{aligned}
& \mathbb{P}(X_n - X_0 \geq \alpha n) \\
& \leq \left[\left(\frac{\gamma + \delta}{\gamma} \right)^{-\frac{\gamma + \delta}{1 + \gamma}} (1 - \delta)^{-\frac{1 - \delta}{1 + \gamma}} \right]^n \\
& = \exp \left\{ -n \left[\left(\frac{\gamma + \delta}{1 + \gamma} \right) \ln \left(\frac{\gamma + \delta}{\gamma} \right) + \left(\frac{1 - \delta}{1 + \gamma} \right) \ln(1 - \delta) \right] \right\} \\
& = \exp \left(-n D \left(\frac{\delta + \gamma}{1 + \gamma} \parallel \frac{\gamma}{1 + \gamma} \right) \right) \tag{22}
\end{aligned}$$

and the exponent is equal to $+\infty$ if $\delta > 1$ (i.e., if $\alpha > d$). Applying inequality (22) to the martingale $\{-X_k, \mathcal{F}_k\}_{k=0}^\infty$ gives the same upper bound to the other tail-probability $\mathbb{P}(X_n - X_0 \leq -\alpha n)$. The probability of the union of the two disjoint events $\{X_n - X_0 \geq \alpha n\}$ and $\{X_n - X_0 \leq -\alpha n\}$, that is equal to the sum of their probabilities, therefore satisfies the upper bound in (10). This completes the proof of Theorem 2. \blacksquare

Example 3: Let $d > 0$ and $\varepsilon \in (0, \frac{1}{2}]$ be some constants. Consider a discrete-time real-valued martingale $\{X_k, \mathcal{F}_k\}_{k=0}^\infty$ where a.s. $X_0 = 0$, and for every $m \in \mathbb{N}$

$$\begin{aligned}
& \mathbb{P}(X_m - X_{m-1} = d \mid \mathcal{F}_{m-1}) = \varepsilon, \\
& \mathbb{P} \left(X_m - X_{m-1} = -\frac{\varepsilon d}{1 - \varepsilon} \mid \mathcal{F}_{m-1} \right) = 1 - \varepsilon.
\end{aligned}$$

This indeed implies that a.s. for every $m \in \mathbb{N}$

$$\mathbb{E}[X_m - X_{m-1} \mid \mathcal{F}_{m-1}] = \varepsilon d + \left(-\frac{\varepsilon d}{1 - \varepsilon} \right) (1 - \varepsilon) = 0$$

and since X_{m-1} is \mathcal{F}_{m-1} -measurable then a.s.

$$\mathbb{E}[X_m \mid \mathcal{F}_{m-1}] = X_{m-1}.$$

Since $\varepsilon \in (0, \frac{1}{2}]$ then a.s.

$$|X_m - X_{m-1}| \leq \max \left\{ d, \frac{\varepsilon d}{1 - \varepsilon} \right\} = d.$$

From Azuma's inequality, for every $x \geq 0$,

$$\mathbb{P}(X_k \geq kx) \leq \exp \left(-\frac{kx^2}{2d^2} \right) \tag{23}$$

independently of the value of ε (note that $X_0 = 0$ a.s.). The concentration inequality in Theorem 2 enables one to get a better bound: Since a.s., for every $m \in \mathbb{N}$,

$$\mathbb{E}[(X_m - X_{m-1})^2 \mid \mathcal{F}_{m-1}] = d^2 \varepsilon + \left(-\frac{\varepsilon d}{1 - \varepsilon} \right)^2 (1 - \varepsilon) = \frac{d^2 \varepsilon}{1 - \varepsilon}$$

then from (11)

$$\gamma = \frac{\varepsilon}{1 - \varepsilon}, \quad \delta = \frac{x}{d}$$

and from (22), for every $x \geq 0$,

$$\mathbb{P}(X_k \geq kx) \leq \exp \left(-k D \left(\frac{x(1 - \varepsilon)}{d} + \varepsilon \parallel \varepsilon \right) \right). \tag{24}$$

Consider the case where $\varepsilon \rightarrow 0$. Then, for arbitrary $x > 0$ and $k \in \mathbb{N}$, Azuma's inequality in (23) provides an upper bound that is strictly positive independently of ε , whereas the one-sided concentration inequality of Theorem 2 implies a bound in (24) that tends to zero. This exemplifies the improvement that is obtained by Theorem 2 in comparison to Azuma's inequality.

Remark 7: As was noted, e.g., in [38, Section 2], all the concentration inequalities for martingales whose derivation is based on Chernoff's bound can be strengthened to refer to maxima. The reason is that $\{X_k - X_0, \mathcal{F}_k\}_{k=0}^\infty$

is a martingale, and $h(x) = \exp(tx)$ is a convex function on \mathbb{R} for every $t \geq 0$. Recall that a composition of a convex function with a martingale gives a sub-martingale w.r.t. the same filtration (see Section II-C), so it implies that $\{\exp(t(X_k - X_0)), \mathcal{F}_k\}_{k=0}^{\infty}$ is a sub-martingale for every $t \geq 0$. Hence, by applying Doob's maximal inequality for sub-martingales, it follows that for every $\alpha \geq 0$

$$\begin{aligned} & \mathbb{P}\left(\max_{1 \leq k \leq n} X_k - X_0 \geq \alpha n\right) \\ &= \mathbb{P}\left(\max_{1 \leq k \leq n} \exp(t(X_k - X_0)) \geq \exp(\alpha n t)\right) \quad t \geq 0 \\ &\leq \exp(-\alpha n t) \mathbb{E}\left[\exp(t(X_n - X_0))\right] \\ &= \exp(-\alpha n t) \mathbb{E}\left[\exp\left(t \sum_{k=1}^n \xi_k\right)\right] \end{aligned}$$

which coincides with the proof of Theorem 2 with the starting point in (14). This concept applies to all the concentration inequalities derived in this paper.

Corollary 1: In the setting of Theorem 2, for every $\alpha \geq 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) \leq 2 \exp\left(-2n \left(\frac{\delta}{1+\gamma}\right)^2\right). \quad (25)$$

Proof: This concentration inequality is a loosened version of Theorem 2. From Pinsker's inequality,

$$D(p||q) \geq \frac{V^2}{2}, \quad \forall p, q \in [0, 1] \quad (26)$$

where

$$V \triangleq \|(p, 1-p) - (q, 1-q)\|_1 = 2|p - q| \quad (27)$$

denotes the L^1 -variational distance between the two probability distributions. Hence, for $\gamma, \delta \in [0, 1]$

$$D\left(\frac{\delta + \gamma}{1 + \gamma} \parallel \frac{\gamma}{1 + \gamma}\right) \geq 2 \left(\frac{\delta}{1 + \gamma}\right)^2. \quad \blacksquare$$

Remark 8: As was shown in the proof of Corollary 1, the loosening of the exponential bound in Theorem 2 by using Pinsker's inequality gives inequality (25). Note that (25) forms a generalization of Azuma's inequality in Theorem 1 for the special case where, for every i , $d_i \triangleq d$ for some $d > 0$. Inequality (25) is particularized to Azuma's inequality when $\gamma = 1$, and then

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) \leq 2 \exp\left(-\frac{n\delta^2}{2}\right). \quad (28)$$

This is consistent with the observation that if $\gamma = 1$ then, from (11), the requirement in Theorem 2 for the conditional variance of the bounded-difference martingale sequence becomes redundant (since if $|X_k - X_{k-1}| \leq d$ a.s. then also $\mathbb{E}[(X_k - X_{k-1})^2 | \mathcal{F}_{k-1}] \leq d^2$). Hence, if $\gamma = 1$, the concentration inequality in Theorem 2 is derived under the same setting as of Azuma's inequality.

Corollary 2: Let $\{X_k, \mathcal{F}_k\}_{k=0}^{\infty}$ be a discrete-parameter real-valued martingale, and assume that for some constant $d > 0$

$$|X_k - X_{k-1}| \leq d$$

a.s. for every $k \in \{1, \dots, n\}$. Then, for every $\alpha \geq 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) \leq 2 \exp(-nf(\delta)) \quad (29)$$

where

$$f(\delta) = \begin{cases} \ln(2) \left[1 - h_2\left(\frac{1-\delta}{2}\right)\right], & 0 \leq \delta \leq 1 \\ +\infty, & \delta > 1 \end{cases} \quad (30)$$

and $h_2(x) \triangleq -x \log_2(x) - (1-x) \log_2(1-x)$ for $0 \leq x \leq 1$ denotes the binary entropy function on base 2.

Proof: By substituting $\gamma = 1$ in Theorem 2 (i.e., since there is no constraint on the conditional variance, then one can take $\sigma^2 = d^2$), the corresponding exponent in (10) is equal to

$$D\left(\frac{1+\delta}{2} \parallel \frac{1}{2}\right) = f(\delta)$$

since $D(p \parallel \frac{1}{2}) = \ln 2[1 - h_2(p)]$ for every $p \in [0, 1]$. ■

Remark 9: Based on Remark 8, and since Corollary 2 is a special case of Corollary 1 when $\gamma = 1$, then it follows that Corollary 2 is a tightened version of Azuma's inequality. This can be verified directly, by showing that $f(\delta) > \frac{\delta^2}{2}$ for every $\delta > 0$. This inequality is trivial for $\delta > 1$ since f is by definition infinity. For $\delta \in (0, 1]$, the power series expansion of f in (30) is given by

$$f(\delta) = \sum_{p=1}^{\infty} \frac{\delta^{2p}}{2p(2p-1)} = \frac{\delta^2}{2} + \frac{\delta^4}{12} + \frac{\delta^6}{30} + \frac{\delta^8}{56} + \frac{\delta^{10}}{90} \dots \quad (31)$$

which indeed proves the inequality also for $\delta \in (0, 1]$. Figure 1 shows that the two exponents in (28) and (29) nearly coincide for $\delta \leq 0.4$. Also, the improvement in the exponent of Corollary 2, as compared to Azuma's inequality, is by factor $2 \ln 2 \approx 1.386$ for $\delta = 1$.

Discussion 1: Corollary 2 can be re-derived by the replacement of Bennett's inequality in (17) with the inequality

$$\mathbb{E}[\exp(t\xi_k) | \mathcal{F}_{k-1}] \leq \frac{1}{2} [e^{td} + e^{-td}] = \cosh(td) \quad (32)$$

that holds a.s. due to the assumption that $|\xi_k| \leq d$ (a.s.) for every k . The geometric interpretation of this inequality is based on the convexity of the exponential function, which implies that its curve is below the line segment that intersects this curve at the two endpoints of the interval $[-d, d]$. Hence,

$$\exp(t\xi_k) \leq \frac{1}{2} \left(1 + \frac{\xi_k}{d}\right) e^{td} + \frac{1}{2} \left(1 - \frac{\xi_k}{d}\right) e^{-td} \quad (33)$$

a.s. for every $k \in \mathbb{N}$ (or vice versa since \mathbb{N} is a countable set). Since, by assumption, $\{X_k, \mathcal{F}_k\}_{k=0}^{\infty}$ is a martingale then $\mathbb{E}[\xi_k | \mathcal{F}_{k-1}] = 0$ a.s. for every $k \in \mathbb{N}$, so (32) indeed follows from (33). Combined with Chernoff's inequality, it yields (after making the substitution $x = td$ where $x \geq 0$) that

$$\mathbb{P}(X_n - X_0 \geq \alpha n) \leq (\exp(-\delta x) \cosh(x))^n, \quad \forall x \geq 0. \quad (34)$$

This inequality leads to the derivation of Azuma's inequality. The difference that makes Corollary 2 be a tightened version of Azuma's inequality is that in the derivation of Azuma's inequality, the hyperbolic cosine is replaced with the bound $\cosh(x) \leq \exp(\frac{x^2}{2})$ so the inequality in (34) is loosened, and then the free parameter $x \geq 0$ is optimized to obtain Azuma's inequality in Theorem 1 for the special case where $d_k \triangleq d$ for every $k \in \mathbb{N}$ (note that Azuma's inequality handles the more general case where d_k is not a fixed value for every k). In the case where $d_k \triangleq d$ for every k , Corollary 2 is obtained by an optimization of the non-negative parameter x in (34). If $\delta \in [0, 1]$, then by setting to zero the derivative of the logarithm of the right-hand side of (34), it follows that the optimized value is equal to $x = \tanh^{-1}(\delta)$. Substituting this value into the right-hand side of (34) provides the concentration inequality in Corollary 2; to this end, one needs to rely on the identities

$$\tanh^{-1}(\delta) = \frac{1}{2} \ln \left(\frac{1+\delta}{1-\delta} \right), \quad \cosh(x) = (1 - \tanh^2(x))^{-\frac{1}{2}}.$$

In the following, a known loosened version of Theorem 2 is re-derived based on Theorem 2.

Lemma 1: For every $x, y \in [0, 1]$

$$D\left(\frac{x+y}{1+y} \parallel \frac{y}{1+y}\right) \geq \frac{x^2}{2y} B\left(\frac{x}{y}\right) \quad (35)$$

where

$$B(u) \triangleq \frac{2[(1+u) \ln(1+u) - u]}{u^2}, \quad \forall u > 0. \quad (36)$$

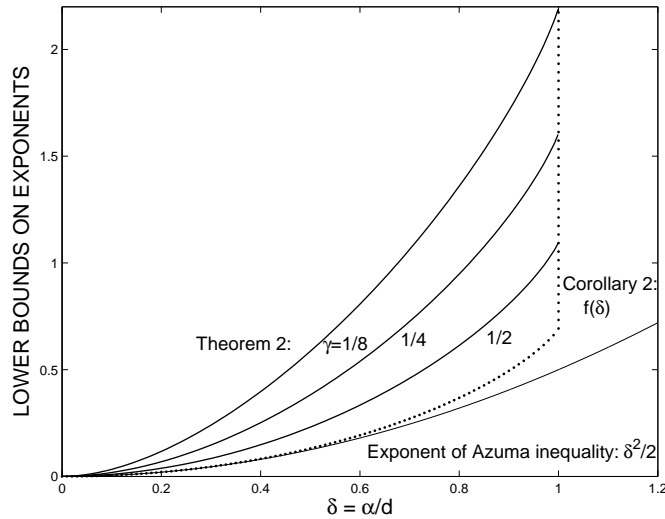


Fig. 1. Plot of the lower bounds on the exponents from Azuma's inequality in (28) and the refined inequalities in Theorem 2 and Corollary 2 (where f is defined in (30)). The pointed line refers to the exponent in Corollary 2, and the three solid lines for $\gamma = \frac{1}{8}, \frac{1}{4}$ and $\frac{1}{2}$ refer to the exponents in Theorem 2.

Proof: This inequality follows by calculus, and it appears in [18, Exercise 2.4.21 (a)]. ■

Corollary 3: Let $\{X_k, \mathcal{F}_k\}_{k=0}^{\infty}$ be a discrete-parameter real-valued martingale that satisfies the conditions in Theorem 2. Then, for every $\alpha \geq 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) \leq 2 \exp \left(-n\gamma \left[\left(1 + \frac{\delta}{\gamma}\right) \ln \left(1 + \frac{\delta}{\gamma}\right) - \frac{\delta}{\gamma} \right] \right) \quad (37)$$

where $\gamma, \delta \in [0, 1]$ are introduced in (11).

Proof: This inequality follows directly by combining inequalities (10) and (35) with the equality in (36). ■

B. Geometric Interpretation

The basic inequality that leads to the derivation of Azuma's inequality (and also its tightened version in Corollary 2) relies on the convexity of the exponential function. Hence, this function is upper bounded over an arbitrary interval by the line segment that intersects the curve of this exponential function at the two endpoints of this interval. Under the additional assumption made in Theorem 2 regarding the conditional variance, one may be motivated by the above geometric viewpoint to improve Azuma's inequality by looking for a suitable parabola that coincides with the exponential function at the two endpoints of the interval, and which forms an improved upper bound to this exponential function over the considered interval (as compared to the upper bound that is obtained by referring to the line segment that intersects the curve of the exponential function at the two endpoints of this interval, see inequality (33)). The analysis that follows from this approach leads to the following theorem.

Theorem 3: Let $\{X_k, \mathcal{F}_k\}_{k=0}^{\infty}$ be a discrete-parameter real-valued martingale that satisfies the conditions in Theorem 2 with some constants $d, \sigma > 0$. Then, for every $\alpha \geq 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) \leq 2 \exp(-nC(\gamma, \delta))$$

where γ and δ are introduced in (11), and the exponent in this bound is defined as follows:

- If $\delta > 1$ then $C(\gamma, \delta) = \infty$.
- If $\delta = 1$ then

$$C(\gamma, \delta) = \ln \left(\frac{4}{1 + \gamma} \right).$$

- Otherwise, if $\delta \in [0, 1)$, then

$$C(\gamma, \delta) = -\ln(u + v)$$

where

$$u \triangleq \left(\frac{1+\gamma}{4} \right) e^{(1-\delta)x}$$

$$v \triangleq \left(\frac{1}{2} + \frac{(1+2x)(1-\gamma)}{4} \right) e^{-(1+\delta)x}.$$

In the above two equalities, $x \in (0, \infty)$ is given by

$$x \triangleq -\frac{1+W_{-1}(w)}{2} - \frac{\gamma+\delta}{(1+\delta)(1-\gamma)}$$

where W_{-1} stands for a branch of the Lambert W function [14], and

$$w \triangleq -\frac{(1+\gamma)(1-\delta)}{(1-\gamma)(1+\delta)} \cdot e^{-1-\frac{2(\gamma+\delta)}{(1+\delta)(1-\gamma)}}.$$

Proof: See Appendix A. ■

As is explained in the following discussion, Theorem 3 is looser than Theorem 2 (though it improves Corollary 2 and Azuma's inequality that are independent of γ). The reason for introducing Theorem 3 here is in order to emphasize the geometric interpretation of the concentration inequalities that were introduced so far, as is discussed in the following.

Discussion 2: A common ingredient in proving Azuma's inequality, and Theorems 2 and 3 is a derivation of an upper bound on the conditional expectation $\mathbb{E}[e^{t\xi_k} | \mathcal{F}_{k-1}]$ for $t \geq 0$ where $\mathbb{E}[\xi_k | \mathcal{F}_{k-1}] = 0$, $\text{Var}[\xi_k | \mathcal{F}_{k-1}] \leq \sigma^2$, and $|\xi_k| \leq d$ a.s. for some $\sigma, d > 0$ and for every $k \in \mathbb{N}$. The derivation of Azuma's inequality and Corollary 2 is based on the line segment that connects the curve of the exponent $y(x) = e^{tx}$ at the endpoints of the interval $[-d, d]$; due to the convexity of y , this chord is above the curve of the exponential function y over the interval $[-d, d]$. The derivation of Theorem 2 is based on Bennett's inequality which is applied to the conditional expectation above. The proof of Bennett's inequality (see, e.g., [18, Lemma 2.4.1]) is shortly reviewed, while adopting its proof to our notation, for the continuation of this discussion. Let X be a random variable with zero mean and variance $E[X^2] = \sigma^2$, and assume that $X \leq d$ a.s. for some $d > 0$. Let $\gamma \triangleq \frac{\sigma^2}{d^2}$. The geometric viewpoint of Bennett's inequality is based on the derivation of an upper bound on the exponential function y over the interval $(-\infty, d]$; this upper bound on y is a parabola that intersects y at the right endpoint (d, e^{td}) and is tangent to the curve of y at the point $(-\gamma d, e^{-t\gamma d})$. As is verified in the proof of [18, Lemma 2.4.1], it leads to the inequality $y(x) \leq \varphi(x)$ for every $x \in (-\infty, d]$ where φ is the parabola that satisfies the conditions

$$\begin{aligned} \varphi(d) &= y(d) = e^{td}, \\ \varphi(-\gamma d) &= y(-\gamma d) = e^{-t\gamma d}, \\ \varphi'(-\gamma d) &= y'(-\gamma d) = te^{-t\gamma d}. \end{aligned}$$

Calculation shows that this parabola admits the form

$$\varphi(x) = \frac{(x+\gamma d)e^{td} + (d-x)e^{-t\gamma d}}{(1+\gamma)d} + \frac{\alpha[\gamma d^2 + (1-\gamma)d x - x^2]}{(1+\gamma)^2 d^2}$$

where $\alpha \triangleq [(1+\gamma)td + 1]e^{-t\gamma d} - e^{td}$. At this point, since $\mathbb{E}[X] = 0$, $\mathbb{E}[X^2] = \gamma d^2$ and $X \leq d$ a.s., then the following bound holds:

$$\begin{aligned} &\mathbb{E}[e^{tX}] \\ &\leq \mathbb{E}[\varphi(X)] \\ &= \frac{\gamma e^{td} + e^{-t\gamma d}}{1+\gamma} + \alpha \left(\frac{\gamma d^2 - \mathbb{E}[X^2]}{(1+\gamma)^2 d^2} \right) \\ &= \frac{\gamma e^{td} + e^{-t\gamma d}}{1+\gamma} \\ &= \frac{\mathbb{E}[X^2]e^{td} + d^2 e^{-\frac{t\mathbb{E}[X^2]}{d}}}{d^2 + \mathbb{E}[X^2]} \end{aligned}$$

which indeed proves Bennett’s inequality in the considered setting, and it also provides a geometric viewpoint to the proof of this inequality. Note that under the above assumption, the bound is achieved with equality when X is a RV that gets the two values $+d$ and $-\gamma d$ with probabilities $\frac{\gamma}{1+\gamma}$ and $\frac{1}{1+\gamma}$, respectively. This bound also holds when $\mathbb{E}[X^2] \leq \sigma^2$ since the right-hand side of the inequality is a monotonic non-decreasing function of $\mathbb{E}[X^2]$ (as it was verified in the proof of [18, Lemma 2.4.1]).

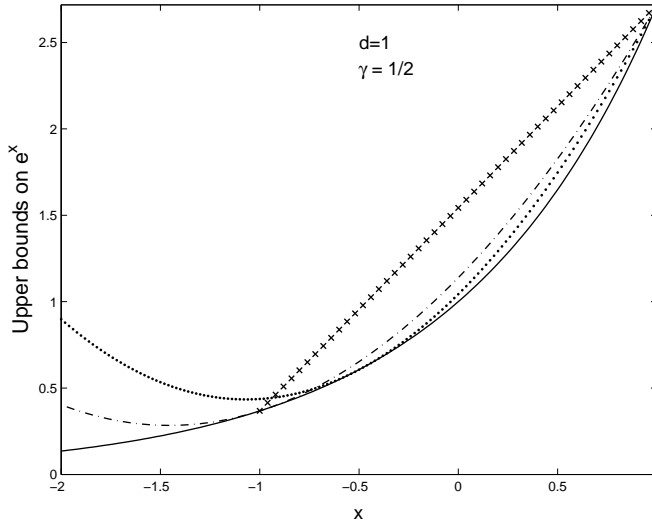


Fig. 2. The function $y = e^x$ (solid line), and the upper bounds on this function that are used to derive Azuma’s inequality and Corollary 2 (the dashed line segment intersecting the exponential function at the endpoints of the interval $[-d, d]$), Theorem 2 (the pointed line for the parabola that coincides with the exponential function at $x = d$ and is tangent to this function at $x = -\gamma d$), and Theorem 3 (the dash-dot line for the parabola that coincides with the exponential function at $x = d$ and is tangent to this function at $x = -d$). These parabolas form upper bounds on the exponential function over $(-\infty, d]$.

Applying Bennett’s inequality to the conditional law of ξ_k given \mathcal{F}_{k-1} gives (17) (with γ in (11)). From this discussion, the parabola that serves for the derivation of Bennett’s inequality is the best one in the sense that it achieves the minimal upper bound on the conditional expectation $\mathbb{E}[e^{t\xi_k} | \mathcal{F}_{k-1}]$ (where $t \geq 0$) with equality for a certain conditional probability distribution. In light of this geometric interpretation, it follows from the proof of Theorem 3 that the concentration inequality in this theorem is looser than the one in Theorem 2. The reason is that the underlying parabola that serves to get an upper bound on the exponential function in Theorem 3 is the parabola that intersects y at $x = d$ and is tangent to the curve of this exponent at $x = -d$; as is illustrated in Figure 2, this parabola forms an upper bound on the exponential function y over the interval $(-\infty, d]$. On the other hand, Theorem 3 refines Azuma’s inequality and Corollary 2 since the chord that connects the curve of the exponential function at the two endpoints of the interval $[-d, d]$ is replaced by a tighter upper bound which is the parabola that coincides with the exponent at the two endpoints of this interval. Figure 2 compares the three considered upper bounds on the exponential function that serve for the derivation of Azuma’s inequality (and Corollary 2), and Theorems 2 and 3. A comparison of the resulting bounds on the exponents of these inequalities and some other bounds that are derived later in this section is shown in Figure 3; it verifies that indeed the exponent of Theorem 2 is superior over the exponent in Theorem 3, but this difference is reduced by increasing the value of $\gamma \in (0, 1]$ (e.g., for $\gamma = \frac{3}{4}$, this difference is already marginal). The reason for this observation is that the two underlying parabolas that serve for the derivation of Theorems 2 and 3 almost coincide when the value of γ is approached to 1 (and they are exactly the same parabola when $\gamma = 1$); in this respect, note that the left tangent point at $x = -\gamma d$ for the parabola that refers to the derivation of Theorem 2 (via Bennet’s inequality) tends to the left endpoint of the interval $[-d, d]$ as $\gamma \rightarrow 1$, and therefore the two parabolas almost coincide for γ close to 1.

C. Another Approach for the Derivation of a Refinement of Azuma’s Inequality

Theorem 4: Let $\{X_k, \mathcal{F}_k\}_{k=0}^\infty$ be a discrete-parameter real-valued martingale, and let $m \in \mathbb{N}$ be an even number. Assume that the following conditions hold a.s. for every $k \in \mathbb{N}$

$$|X_k - X_{k-1}| \leq d,$$

$$\left| \mathbb{E}[(X_k - X_{k-1})^l | \mathcal{F}_{k-1}] \right| \leq \mu_l, \quad l = 2, \dots, m$$

for some $d > 0$ and non-negative numbers $\{\mu_l\}_{l=2}^m$. Then, for every $\alpha \geq 0$,

$$\mathbb{P}(|X_n - X_0| \geq n\alpha) \leq 2 \left\{ \inf_{x \geq 0} e^{-\delta x} \left[1 + \sum_{l=2}^{m-1} \frac{(\gamma_l - \gamma_m)x^l}{l!} + \gamma_m(e^x - 1 - x) \right] \right\}^n \quad (38)$$

where

$$\delta \triangleq \frac{\alpha}{d}, \quad \gamma_l \triangleq \frac{\mu_l}{d^l}, \quad \forall l = 2, \dots, m. \quad (39)$$

Proof: The starting point of this proof relies on (14) and (15) that were used for the derivation of Theorem 2. From this point, we deviate from the proof of Theorem 2. For every $k \in \mathbb{N}$ and $t \geq 0$

$$\begin{aligned} & \mathbb{E}[\exp(t\xi_k) | \mathcal{F}_{k-1}] \\ &= 1 + t\mathbb{E}[\xi_k | \mathcal{F}_{k-1}] + \dots + \frac{t^{m-1}}{(m-1)!} \cdot \mathbb{E}[(\xi_k)^{m-1} | \mathcal{F}_{k-1}] \\ & \quad + \mathbb{E} \left[\exp(t\xi_k) - 1 - t\xi_k - \dots - \frac{t^{m-1}(\xi_k)^{m-1}}{(m-1)!} \mid \mathcal{F}_{k-1} \right] \\ &= 1 + t\mathbb{E}[\xi_k | \mathcal{F}_{k-1}] + \dots + \frac{t^{m-1}}{(m-1)!} \cdot \mathbb{E}[(\xi_k)^{m-1} | \mathcal{F}_{k-1}] \\ & \quad + \frac{t^m}{m!} \cdot \mathbb{E}[(\xi_k)^m \varphi_m(t\xi_k) | \mathcal{F}_{k-1}] \end{aligned} \quad (40)$$

where

$$\varphi_m(y) \triangleq \begin{cases} \frac{m!}{y^m} \left(e^y - \sum_{l=0}^{m-1} \frac{y^l}{l!} \right) & \text{if } y \neq 0 \\ 1 & \text{if } y = 0 \end{cases}. \quad (41)$$

In order to proceed, we need the following lemma:

Lemma 2: For every $m \in \mathbb{N}$, the function φ_m has the following properties:

- 1) $\lim_{y \rightarrow 0} \varphi_m(y) = 1$, so φ_m is a continuous function.
- 2) φ_m is a positive function over the real line.
- 3) φ_m is monotonic increasing over the interval $[0, \infty)$.
- 4) $0 < \varphi_m(y) \leq 1$ for every $y \leq 0$.

Proof: See Appendix B. ■

Remark 10: Note that [21, Lemma 3.1] states that φ_2 is a monotonic increasing and non-negative function over the real line. In general, for $m \in \mathbb{N}$, it is easier to prove the weaker properties of φ_m that are stated in Lemma 2; these are sufficient for the continuation of the proof of Theorem 4.

From (40) and Lemma 2, since $\xi_k \leq d$ a.s., then it follows that for an arbitrary $t \geq 0$

$$\varphi_m(t\xi_k) \leq \varphi_m(td), \quad \forall k \in \mathbb{N} \quad (42)$$

a.s. (to see this, let's separate the two cases where ξ_k is either non-negative or negative. If $0 \leq \xi_k \leq d$ a.s. then, for $t \geq 0$, inequality (42) holds (a.s.) due to the monotonicity of φ_m over $[0, \infty)$. If $\xi_k < 0$ then the second and third properties in Lemma 2 yield that, for $t \geq 0$ and every $k \in \mathbb{N}$,

$$\varphi_m(t\xi_k) \leq 1 = \varphi_m(0) \leq \varphi_m(td),$$

so in both cases inequality (42) is satisfied). Since m is even then $(\xi_k)^m \geq 0$ (note that although Lemma 2 holds in general for every $m \in \mathbb{N}$, this is the point where we need m to be an even number), and

$$\mathbb{E}[(\xi_k)^m \varphi_m(t\xi_k) | \mathcal{F}_{k-1}] \leq \varphi_m(td) \mathbb{E}[(\xi_k)^m | \mathcal{F}_{k-1}], \quad \forall t \geq 0.$$

Also, since $\{X_k, \mathcal{F}_k\}_{k=0}^\infty$ is a martingale then $\mathbb{E}[\xi_k | \mathcal{F}_{k-1}] = 0$, and based on the assumptions of this theorem

$$\mathbb{E}[(\xi_k)^l | \mathcal{F}_{k-1}] \leq \mu_l = d^l \gamma_l, \quad \forall l \in \{2, \dots, m\}.$$

By substituting the last three results on the right-hand side of (40), it follows that for every $t \geq 0$ and every $k \in \mathbb{N}$

$$\mathbb{E}[\exp(t\xi_k)|\mathcal{F}_{k-1}] \leq 1 + \sum_{l=2}^{m-1} \frac{\gamma_l (td)^l}{l!} + \frac{\gamma_m (td)^m \varphi_m(td)}{m!} \quad (43)$$

so from (15)

$$\mathbb{E} \left[\exp \left(t \sum_{k=1}^n \xi_k \right) \right] \leq \left(1 + \sum_{l=2}^{m-1} \frac{\gamma_l (td)^l}{l!} + \frac{\gamma_m (td)^m \varphi_m(td)}{m!} \right)^n, \quad \forall t \geq 0. \quad (44)$$

From (14), if $\alpha \geq 0$ is arbitrary, then for every $t \geq 0$

$$\mathbb{P}(X_n - X_0 \geq \alpha n) \leq \exp(-\alpha n t) \left(1 + \sum_{l=2}^{m-1} \frac{\gamma_l (td)^l}{l!} + \frac{\gamma_m (td)^m \varphi_m(td)}{m!} \right)^n.$$

Let $x \triangleq td$. Then, based on (11) and (41), for every $\alpha \geq 0$

$$\begin{aligned} & \mathbb{P}(X_n - X_0 \geq \alpha n) \\ & \leq \left\{ \inf_{x \geq 0} e^{-\delta x} \left(1 + \sum_{l=2}^{m-1} \frac{\gamma_l x^l}{l!} + \frac{\gamma_m x^m \varphi_m(x)}{m!} \right) \right\}^n \\ & = \left\{ \inf_{x \geq 0} e^{-\delta x} \left[1 + \sum_{l=2}^{m-1} \frac{\gamma_l x^l}{l!} + \gamma_m \left(e^x - \sum_{l=0}^{m-1} \frac{x^l}{l!} \right) \right] \right\}^n \\ & = \left\{ \inf_{x \geq 0} e^{-\delta x} \left[1 + \sum_{l=2}^{m-1} \frac{(\gamma_l - \gamma_m) x^l}{l!} + \gamma_m (e^x - 1 - x) \right] \right\}^n. \end{aligned} \quad (45)$$

Applying inequality (45) to the martingale $\{-X_k, \mathcal{F}_k\}_{k=0}^\infty$ gives the same bound on the probability $\mathbb{P}(X_n - X_0 \leq -\alpha n)$. Finally, the concentration inequality in (38) follows by summing the common upper bound for the probabilities of the two disjoint events $\{X_n - X_0 \geq \alpha n\}$ and $\{X_n - X_0 \leq -\alpha n\}$. This completes the proof of Theorem 4. \blacksquare

Remark 11: Without any loss of generality, it is assumed that $\alpha \in [0, d]$ (as otherwise, the considered probability is zero for $\alpha > d$). Based on the above conditions, it is also assumed that $\mu_l \leq d^l$ for every $l \in \{2, \dots, m\}$. Hence, $\delta \in [0, 1]$, and $\gamma_l \in [0, 1]$ for all values of l . Note that, from (11), $\gamma_2 = \gamma$.

Remark 12: From the proof of Theorem 4, it follows that the one-sided inequality (45) is satisfied if the martingale $\{X_k, \mathcal{F}_k\}_{k=0}^n$ fulfills the following conditions a.s.

$$\begin{aligned} X_k - X_{k-1} & \leq d, \\ \mathbb{E}[(X_k - X_{k-1})^l | \mathcal{F}_{k-1}] & \leq \mu_l, \quad l = 2, \dots, m \end{aligned}$$

for some $d > 0$ and non-negative numbers $\{\mu_l\}_{l=2}^m$. Note that these conditions are weaker than those that are stated in Theorem 4. Under these weaker conditions, $\gamma_l \triangleq \frac{\mu_l}{d^l}$ may be larger than 1. This remark will be helpful later in this paper.

Remark 13: The infimum in (38) of Theorem 4 is attained and thus is a minimum. To show it, let $f(x)$ for $x \in \mathbb{R}^+$ be the base of the exponent in (38), so we need to prove that $L \triangleq \inf_{x \in \mathbb{R}^+} f(x)$ is attained. The infimum is well defined since $f \geq 0$. Moreover, $\lim_{x \rightarrow \infty} f(x) = \infty$. Indeed

$$f(x) = e^{-\delta x} g(x) + \gamma_m e^{(1-\delta)x}, \quad \forall x \in \mathbb{R}^+$$

for some polynomial g , so for $\delta \in (0, 1)$, the first term tends to zero and the second tends to infinity as $x \rightarrow \infty$. This implies that there exists some $A > 0$ such that $f(x) \geq 1$ for every $x \geq A$. As $f(0) = 1$, one can reduce the set over which the infimum of f is taken to the closed interval $[0, A]$. The claim follows from the continuity of f , and since every continuous function over a compact set attains its infimum.

1) *Specialization of Theorem 4 for $m = 2$* : Theorem 4 with $m = 2$ (i.e., when the same conditions as of Theorem 2 hold) is expressible in closed form, as follows:

Corollary 4: Let $\{X_k, \mathcal{F}_k\}_{k=0}^\infty$ be a discrete-parameter real-valued martingale that satisfies a.s. the conditions in Theorem 2. Then, for every $\alpha \geq 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) \leq 2 \exp(-nC(\gamma, \delta))$$

where γ and δ are introduced in (11), and the exponent in this upper bound gets the following form:

- If $\delta > 1$ then $C(\gamma, \delta) = \infty$.
- If $\delta = 1$ then

$$C(\gamma, \delta) = \frac{1}{\gamma} - \ln\left(\gamma(e^{\frac{1}{\gamma}} - 1)\right).$$

- Otherwise, if $\delta \in (0, 1)$, then

$$C(\gamma, \delta) = \delta x - \ln(1 + \gamma(e^x - 1 - x))$$

where $x \in (0, \frac{1}{\gamma})$ is given by

$$x = \frac{1}{\gamma} + \frac{1}{\delta} - 1 - W_0\left(\frac{(1 - \delta)e^{\frac{1}{\gamma} + \frac{1}{\delta} - 1}}{\delta}\right) \quad (46)$$

and W_0 denotes the principal branch of the Lambert W function [14].

Proof: See Appendix C. ■

Proposition 1: If $\gamma < \frac{1}{2}$ then Corollary 4 gives a stronger result than Corollary 2 (and, hence, it is also better than Azuma's inequality).

Proof: See Appendix D. ■

It is of interest to compare the tightness of Theorem 2 and Corollary 4. This leads to the following conclusion:

Proposition 2: The concentration inequality in Corollary 4 is looser than Theorem 2.

Proof: See Appendix E. ■

The statements in Propositions 1 and 2 are illustrated in Figure 3. Sub-plots (a) and (b) in Figure 3 refer to $\gamma \leq \frac{1}{2}$ where the statement in Proposition 1 holds. On the other hand, sub-plots (c) and (d) in Figure 3 refer to higher values of γ , and therefore the statement in Proposition 1 does not apply to these values of γ .

2) *Exploring the Dependence of the Bound in Theorem 4 in Terms of m* : In the previous sub-section, a closed-form expression of Theorem 4 was obtained for the special case where $m = 2$ (see Corollary 4), but also Proposition 2 states that this special case is looser than Theorem 2 (which is also given in closed form). Hence, it is natural to enquire how does the bound in Theorem 4 vary in terms of m (where $m \geq 2$ is even), and if there is a chance to obtain an improvement over Theorem 2 by assigning some even values of $m > 2$ in Theorem 4. Also, due to the closed-form expression in Corollary 4, it would be pleasing to derive from Theorem 4 an inequality that is expressed in closed form for a general even value of $m \geq 2$. The continuation of the study in this sub-section is outlined as follows:

- A loosened version of Theorem 4 is introduced, and it is shown to provide an inequality whose tightness consistently improves by increasing the value of m . For $m = 2$, this loosened version coincides with Theorem 4. Hence, it follows (by introducing this loosened version) that $m = 2$ provides the weakest bound in Theorem 4.
- Inspired by the closed-form expression of the bound in Corollary 4, we derive a closed-form inequality (i.e., a bound that is not subject to numerical optimization) by either loosening Theorem 4 or further loosening its looser version from the previous item. As will be exemplified numerically in Section V, the closed-form expression of the new bound causes to a marginal loosening of Theorem 4. Also, for $m = 2$, it is exactly Theorem 4.
- A necessary and sufficient condition is derived for the case where, for an even $m \geq 4$, Theorem 4 provides a bound that is exponentially advantageous over Theorem 2. Note however that, when $m \geq 4$ in Theorem 4, one needs to calculate conditional moments of the martingale differences that are of higher orders than 2; hence, an improvement in Theorem 4 is obtained at the expense of the need to calculate higher-order conditional moments. Saying this, note that the derivation of Theorem 4 deviates from the proof of Theorem 2 at an early stage, and it cannot be considered as a generalization of Theorem 2 when higher-order moments are

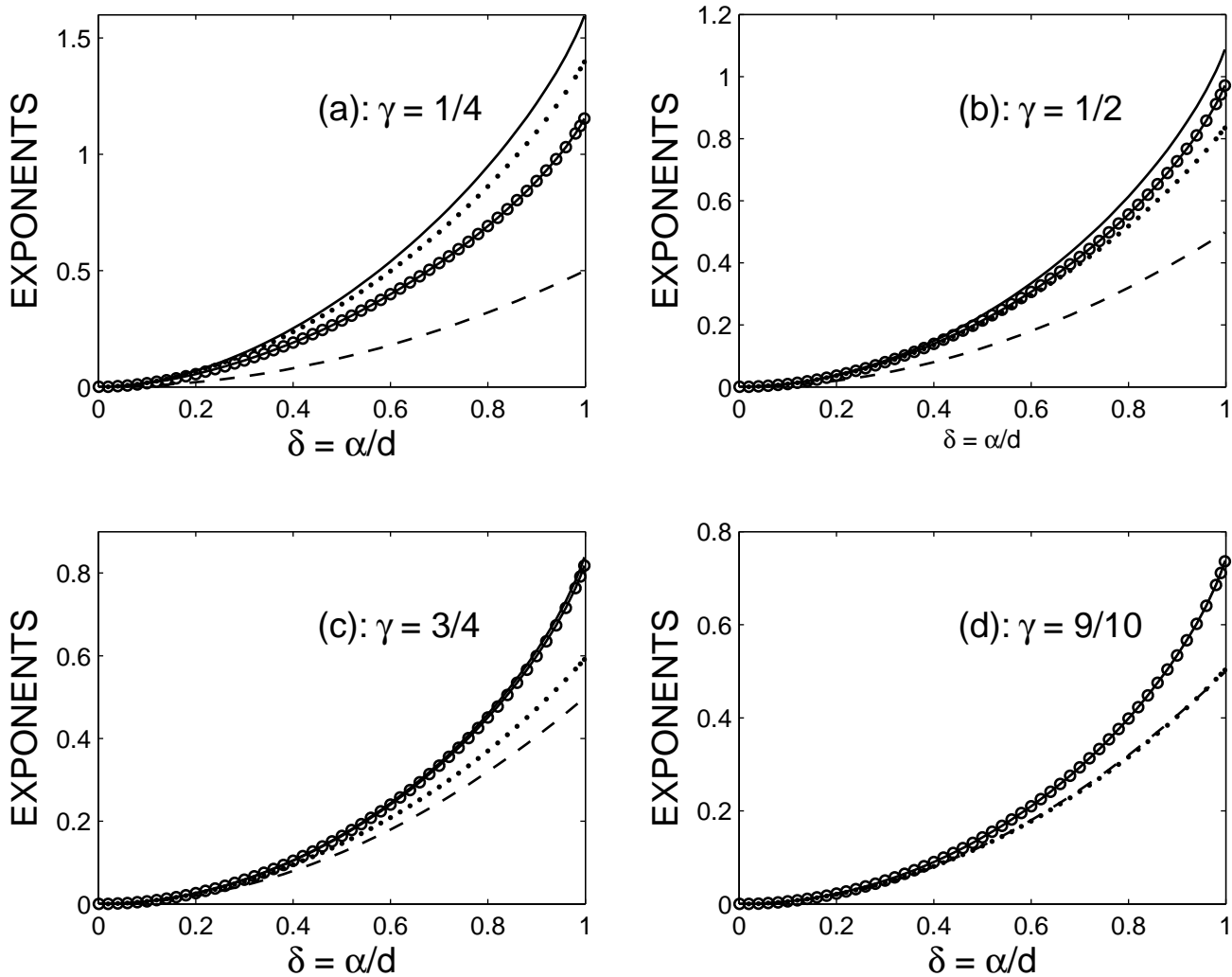


Fig. 3. Plots of the exponents $c \triangleq c(\gamma, \delta)$ for bounds of the form $\mathbb{P}\{|X_n - X_0| \geq n\alpha\} \leq e^{-nc(\gamma, \delta)}$ for an arbitrary $\alpha \geq 0$. The sequence $\{X_k, \mathcal{F}_k\}_{k=0}^\infty$ is a discrete-parameter martingale that satisfies the conditions in Theorem 2 for some positive constants d and σ (without loss of generality, $\sigma \leq d$), and where $\gamma \in (0, 1]$ and $\delta \in [0, 1]$ are introduced in (11). The plots show lower bounds on the exponents according to Azuma's inequality in (28), the bounds in Theorems 2 and 3, and the bound in Corollary 4. The plots are depicted for a fixed value of $\gamma \triangleq \frac{\sigma^2}{d^2}$; the horizontal axes refer to the parameter $\delta \triangleq \frac{\alpha}{d}$, and the vertical axes refer to the lower bounds on the exponents $c(\gamma, \delta)$. The dashed lines refer to Azuma's inequality, the solid lines refer to Theorem 2, the circled lines refer to Theorem 3, and the pointed lines refer to Corollary 4. The subplots (a)-(d) correspond to values of $\gamma = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}$ and $\frac{9}{10}$, respectively.

available (as is also evidenced in Proposition 2 which demonstrates that, for $m = 2$, Theorem 4 is weaker than Theorem 2).

- Finally, this sufficient condition is particularized in the asymptotic case where $m \rightarrow \infty$. It is of interest since the tightness of the loosened version of Theorem 4 from the first item is improved by increasing the value of m .

The analysis that is related to the above outline is presented in the following. Numerical results that are related to the comparison of Theorems 2 and 4 are relegated to Section V (while considered in a certain communication-theoretic context).

Corollary 5: Let $\{X_k, \mathcal{F}_k\}_{k=0}^n$ be a discrete-parameter real-valued martingale, and let $m \in \mathbb{N}$ be an even number. Assume that $|X_k - X_{k-1}| \leq d$ holds a.s. for every $k \in \mathbb{N}$, and that there exists a (non-negative) sequence $\{\mu_l\}_{l=2}^m$ so that for every $k \in \mathbb{N}$

$$\mu_l = \mathbb{E}[|X_k - X_{k-1}|^l | \mathcal{F}_{k-1}], \quad \forall l = 2, \dots, m. \quad (47)$$

Then, inequality (38) holds with the notation in (39).

Proof: This corollary is a consequence of Theorem 4 since

$$|\mathbb{E}[(X_k - X_{k-1})^l | \mathcal{F}_{k-1}]| \leq \mathbb{E}[|X_k - X_{k-1}|^l | \mathcal{F}_{k-1}].$$

Proposition 3: Theorem 4 and Corollary 5 coincide for $m = 2$ (hence, Corollary 5 provides in this case the result stated in Corollary 4). Furthermore, the bound in Corollary 5 improves as the even value of $m \in \mathbb{N}$ is increased. ■

Proof: See Appendix F. ■

Inspired by the closed-form inequality that follows from Theorem 4 for $m = 2$ (see Corollary 4), a closed-form inequality is suggested in the following by either loosening Theorem 4 or Corollary 5. It generalizes the result in Corollary 4, and it coincides with Theorem 4 and Corollary 5 for $m = 2$.

Corollary 6: Under the conditions of Corollary 5 then, for every $\alpha \geq 0$,

$$\begin{aligned} & \mathbb{P}(X_n - X_0 \geq n\alpha) \\ & \leq \left\{ e^{-\delta x} \left[1 + \sum_{l=2}^{m-1} \frac{(\gamma_l - \gamma_m)x^l}{l!} + \gamma_m(e^x - 1 - x) \right] \right\}^n \end{aligned} \quad (48)$$

where $\{\gamma_l\}_{l=2}^m$ and δ are introduced in (39),

$$x = \frac{a+b}{c} - W_0\left(\frac{b}{c} \cdot e^{\frac{a+b}{c}}\right) \quad (49)$$

with W_0 that denotes the principal branch of the Lambert W function [14], and

$$a \triangleq \frac{1}{\gamma_2}, \quad b \triangleq \frac{\gamma_m}{\gamma_2} \left(\frac{1}{\delta} - 1 \right), \quad c \triangleq \frac{1}{\delta} - b. \quad (50)$$

Proof: See Appendix G. ■

Remark 14: It is exemplified numerically in Section V that the replacement of the infimum over $x \geq 0$ on the right-hand side of (38) with the sub-optimal choice of the value of x that is given in (49) and (50) implies a marginal loosening in the exponent of the bound. Note also that, for $m = 2$, this value of x is optimal since it coincides with the exact value in (46).

Corollary 7: Under the assumptions of Theorem 2 then, for every $\alpha \geq 0$,

$$\mathbb{P}(X_n - X_0 \geq n\alpha) \leq e^{-nE} \quad (51)$$

where

$$E = E_2(\gamma_2, \delta) \triangleq D \left(\frac{\delta + \gamma_2}{1 + \gamma_2} \parallel \frac{\gamma_2}{1 + \gamma_2} \right). \quad (52)$$

Also, under the assumptions of Theorem 4 or Corollary 5 then (51) holds for every $\alpha \geq 0$ with

$$\begin{aligned} E &= E_4(\{\gamma_l\}_{l=2}^m, \delta) \\ &\triangleq \sup_{x \geq 0} \left\{ \delta x - \ln \left(1 + \sum_{l=2}^{m-1} \frac{(\gamma_l - \gamma_m)x^l}{l!} + \gamma_m(e^x - 1 - x) \right) \right\} \end{aligned} \quad (53)$$

where $m \geq 2$ is an arbitrary even number. Hence, Theorem 4 or Corollary 5 are better exponentially than Theorem 2 if and only if $E_4 > E_2$.

Proof: The proof follows directly from (22) and (45). ■

Remark 15: In order to avoid the operation of taking the supremum over $x \in [0, \infty)$, it is sufficient to first check if $\tilde{E}_4 > E_2$ where

$$\tilde{E}_4 \triangleq \delta x - \ln \left(1 + \sum_{l=2}^{m-1} \frac{(\gamma_l - \gamma_m)x^l}{l!} + \gamma_m(e^x - 1 - x) \right)$$

with the value of x in (49) and (50). This sufficient condition is exemplified later in Section V.

D. Concentration Inequalities for Small Deviations

In the following, we consider the probability of the events $\{|X_n - X_0| \geq \alpha\sqrt{n}\}$ for an arbitrary $\alpha \geq 0$. These events correspond to small deviations. This is in contrast to events of the form $\{|X_n - X_0| \geq \alpha n\}$, whose probabilities were analyzed earlier in this section, referring to large deviations.

Proposition 4: Let $\{X_k, \mathcal{F}_k\}_{k=0}^\infty$ be a discrete-parameter real-valued martingale. Then, Theorems 2 and 3, and also Corollaries 3 and 4 imply that, for every $\alpha \geq 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) \leq 2 \exp\left(-\frac{\delta^2}{2\gamma}\right) \left(1 + O(n^{-\frac{1}{2}})\right). \quad (54)$$

Also, under the conditions of Theorem 4, inequality (54) holds for every even $m \geq 2$ (so the conditional moments of higher order than 2 do not improve, via Theorem 4, the scaling of the upper bound in (54)).

Proof: See Appendix H. ■

Remark 16: From Proposition 4, all the upper bounds on $\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n})$ (for an arbitrary $\alpha \geq 0$) improve the exponent of Azuma's inequality by a factor of $\frac{1}{\gamma}$.

E. Inequalities for Sub and Super Martingales

Upper bounds on the probability $\mathbb{P}(X_n - X_0 \geq r)$ for $r \geq 0$, earlier derived in this section for martingales, can be adapted to super-martingales (similarly to, e.g., [12, Chapter 2] or [13, Section 2.7]). Alternatively, replacing $\{X_k, \mathcal{F}_k\}_{k=0}^n$ with $\{-X_k, \mathcal{F}_k\}_{k=0}^n$ provides upper bounds on the probability $\mathbb{P}(X_n - X_0 \leq -r)$ for sub-martingales. For example, the adaptation of Theorem 2 to sub and super martingales gives the following theorem:

Theorem 5: Let $\{X_k, \mathcal{F}_k\}_{k=0}^\infty$ be a discrete-parameter real-valued super-martingale. Assume that, for some constants $d, \sigma > 0$, the following two requirements are satisfied a.s.

$$\begin{aligned} X_k - \mathbb{E}[X_k | \mathcal{F}_{k-1}] &\leq d, \\ \text{Var}(X_k | \mathcal{F}_{k-1}) &\triangleq \mathbb{E}\left[(X_k - \mathbb{E}[X_k | \mathcal{F}_{k-1}])^2 | \mathcal{F}_{k-1}\right] \leq \sigma^2 \end{aligned}$$

for every $k \in \{1, \dots, n\}$. Then, for every $\alpha \geq 0$,

$$\mathbb{P}(X_n - X_0 \geq \alpha n) \leq \exp\left(-n D\left(\frac{\delta + \gamma}{1 + \gamma} \parallel \frac{\gamma}{1 + \gamma}\right)\right) \quad (55)$$

where γ and δ are defined as in (11), and the divergence $D(p||q)$ is introduced in (12). Alternatively, if $\{X_k, \mathcal{F}_k\}_{k=0}^\infty$ is a sub-martingale, the same upper bound in (55) holds for the probability $\mathbb{P}(X_n - X_0 \leq -\alpha n)$. If $\delta > 1$, then these two probabilities are equal to zero.

Proof: The proof of this theorem is similar to the proof of Theorem 2. The only difference is that for a super-martingale, due to its basic property in Section II-C,

$$X_n - X_0 = \sum_{k=1}^n (X_k - X_{k-1}) \leq \sum_{k=1}^n \xi_k$$

a.s., where $\xi_k \triangleq X_k - \mathbb{E}[X_k | \mathcal{F}_{k-1}]$ is \mathcal{F}_k -measurable. Hence $\mathbb{P}((X_n - X_0) \geq \alpha n) \leq \mathbb{P}(\sum_{k=1}^n \xi_k \geq \alpha n)$ where a.s. $\xi_k \leq d$, $\mathbb{E}[\xi_k | \mathcal{F}_{k-1}] = 0$, and $\text{Var}(\xi_k | \mathcal{F}_{k-1}) \leq \sigma^2$. The continuation of the proof coincides with the proof of Theorem 2 (starting from (14)). The other inequality for sub-martingales holds due to the fact that if $\{X_k, \mathcal{F}_k\}$ is a sub-martingale then $\{-X_k, \mathcal{F}_k\}$ is a super-martingale. ■

IV. RELATIONS OF THE REFINED INEQUALITIES TO SOME CLASSICAL RESULTS IN PROBABILITY THEORY

A. Relation of Theorem 2 to the Method of Types

Consider a sequence of i.i.d. RVs X_1, X_2, \dots that are Bernoulli(p) distributed (i.e., for every $i \in \mathbb{N}$, $\mathbb{P}(X_i = 1) = p$ and $\mathbb{P}(X_i = 0) = 1 - p$). According to the method of types (see, e.g., [15, Section 11.1]), it follows that for every $n \in \mathbb{N}$ and $r \geq p$

$$\frac{e^{-nD(r||p)}}{n+1} \leq \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i \geq r\right) \leq e^{-nD(r||p)} \quad (56)$$

where the divergence $D(r||p)$ is given in (12), and therefore

$$\lim_{n \rightarrow \infty} \frac{1}{n} \ln \mathbb{P} \left(\frac{1}{n} \sum_{i=1}^n X_i \geq r \right) = -D(r||p), \quad \forall r \geq p \quad (57)$$

gives the exact exponent. This equality can be obtained as a particular case of Cramér's theorem in \mathbb{R} where the rate function of $X \sim \text{Bernoulli}(p)$ is given by

$$I(x) = \begin{cases} D(x||p) & \text{if } x \in [0, 1] \\ +\infty & \text{otherwise} \end{cases}$$

(for Cramér's theorem in \mathbb{R} see, e.g., [18, Section 2.2.1 and Exercise 2.2.23] and [27, Section 1.3]).

In the following, it is shown that Theorem 2 gives in the considered setting the upper bound on the right-hand side of (56), and it therefore provides the exact exponent in (57). To this end, consider the filtration where $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and

$$\mathcal{F}_n = \sigma(X_1, \dots, X_n), \quad \forall n \in \mathbb{N}$$

and let the sequence of RVs $\{S_n\}_{n=0}^\infty$ be defined as $S_0 = 0$, and

$$S_n = \sum_{i=1}^n X_i - np, \quad \forall n \in \mathbb{N}. \quad (58)$$

It is easy to verify that $\{S_n, \mathcal{F}_n\}_{n=0}^\infty$ is a martingale, and for every $n \in \mathbb{N}$

$$\begin{aligned} |S_n - S_{n-1}| &= |X_n - p| \leq \max\{p, 1 - p\}, \\ \text{Var}(S_n | \mathcal{F}_{n-1}) &= \mathbb{E}[(X_n - p)^2] = p(1 - p). \end{aligned}$$

Consider the case where $p \leq \frac{1}{2}$. Then, from the notation of Theorem 2

$$\sigma^2 = p(1 - p), \quad d = 1 - p.$$

Therefore, it follows from Theorem 2 that for every $\alpha \geq 0$

$$\mathbb{P}(S_n \geq n\alpha) \leq \exp \left(-n D \left(\frac{\delta + \gamma}{1 + \gamma} \middle| \middle| \frac{\gamma}{1 + \gamma} \right) \right) \quad (59)$$

where

$$\gamma = \frac{p}{1 - p}, \quad \delta = \frac{\alpha}{1 - p}. \quad (60)$$

Substituting (60) into (59) gives that for every $\alpha \geq 0$

$$\mathbb{P}(S_n \geq n\alpha) \leq \exp(-nD(\alpha + p||p)). \quad (61)$$

Let $r \triangleq \alpha + p$ (where $r \geq p \iff \alpha \geq 0$). The substitution of (58) into the left-hand side of (61) implies that (61) coincides with the upper bound on the right-hand side of (56). Hence, Theorem 2 gives indeed the exact exponent in (57) for the case of i.i.d. RVs that are Bernoulli(p) distributed with $p \in [0, \frac{1}{2}]$.

The method of types gives that a similar one-sided version of inequality (56) holds for every $r \leq p$, and therefore

$$\lim_{n \rightarrow \infty} \frac{1}{n} \ln \mathbb{P} \left(\frac{1}{n} \sum_{i=1}^n X_i \leq r \right) = -D(r||p), \quad \forall r \leq p. \quad (62)$$

For the case where $p \geq \frac{1}{2}$, let $Y_i \triangleq 1 - X_i$ for every $i \in \mathbb{N}$. From Theorem 2, for every $\alpha \geq 0$,

$$\begin{aligned} & \mathbb{P} \left(\sum_{i=1}^n X_i \leq n(p - \alpha) \right) \\ &= \mathbb{P} \left(\sum_{i=1}^n Y_i \geq n(\alpha + 1 - p) \right) \\ &\stackrel{(a)}{\leq} \exp(-nD(\alpha + 1 - p||1 - p)) \\ &\stackrel{(b)}{=} \exp(-nD(p - \alpha||p)) \end{aligned} \quad (63)$$

where inequality (a) follows from inequality (61) since the i.i.d. RVs $\{Y_i\}_{i \in \mathbb{N}}$ are Bernoulli($1 - p$) distributed ($1 - p \leq \frac{1}{2}$), and equality (b) is satisfied since $D(1 - x || 1 - y) = D(x || y)$ (see (12)). The substitution $r \triangleq p - \alpha$ (so $r \leq p \iff \alpha \geq 0$) in (63) gives the same exponent as on the right-hand side of (62), so Theorem 2 also gives the exact exponent in (62) for i.i.d. RVs that are Bernoulli(p) distributed with $p \in [\frac{1}{2}, 1]$.

B. Relations of [18, Corollary 2.4.7] with Theorem 2 and Proposition 4

According to [18, Corollary 2.4.7], suppose $v > 0$ and a sequence of real-valued RVs $\{Y_n\}_{n=1}^\infty$ satisfies a.s.

- $Y_n \leq 1$ for every $n \in \mathbb{N}$.
- $\mathbb{E}[Y_n | S_{n-1}] = 0$ and $\mathbb{E}[Y_n^2 | S_{n-1}] \leq v$ for

$$S_n \triangleq \sum_{j=1}^n Y_j, \quad S_0 = 0.$$

Then, for every $\lambda \geq 0$,

$$\mathbb{E}[\exp(\lambda S_n)] \leq \left(\frac{v \exp(\lambda) + \exp(-\lambda v)}{1 + \lambda} \right)^n. \quad (64)$$

Moreover, for every $x \geq 0$

$$\mathbb{P}\left(\frac{S_n}{n} \geq x\right) \leq \exp\left(-nD\left(\frac{x+v}{1+v} \middle| \middle| \frac{v}{1+v}\right)\right) \quad (65)$$

and, for every $y \geq 0$,

$$\mathbb{P}\left(\frac{S_n}{\sqrt{n}} \geq y\right) \leq \exp\left(-\frac{2y^2}{(1+v)^2}\right). \quad (66)$$

In the following, we show that [18, Corollary 2.4.7] is closely related to Theorem 2 in this paper. To this end, let $\{X_k, \mathcal{F}_k\}_{k=0}^\infty$ be a discrete-parameter real-valued martingale where $X_k - X_{k-1} \leq d$ a.s. for every $k \in \mathbb{N}$. Let us define the martingale-difference sequence $\{Y_k, \mathcal{F}_k\}_{k=0}^\infty$ where

$$Y_k \triangleq \frac{X_k - X_{k-1}}{d}, \quad \forall k \in \mathbb{N}$$

and $Y_0 \triangleq 0$. Based on the assumptions in Theorem 2, it follows from (11) that $Y_k \leq 1$ a.s. for every $k \in \mathbb{N}$, and

$$\mathbb{E}[Y_k | \mathcal{F}_{k-1}] = 0, \quad \mathbb{E}[Y_k^2 | \mathcal{F}_{k-1}] \leq \frac{\sigma^2}{d^2} = \gamma.$$

Hence, by definition, $\xi_k \triangleq X_k - X_{k-1}$ satisfies the equality $\xi_k = dY_k$ for every $k \in \mathbb{N}$. From (18), with $t = \frac{\lambda}{d}$ and $\gamma = v$, it follows that for every $\lambda \geq 0$

$$\begin{aligned} & \mathbb{E}[\exp(\lambda S_n)] \\ &= \mathbb{E}\left[\exp\left(\frac{\lambda}{d} \sum_{k=1}^n \xi_k\right)\right] \\ &\leq \left(\frac{v \exp(\lambda) + \exp(-v\lambda)}{1 + v}\right)^n \end{aligned}$$

which then coincides with (64). It is noted that in Theorem 2 it was required that $|X_k - X_{k-1}| \leq d$ whereas, due to [18, Corollary 2.4.7], it is enough that $X_k - X_{k-1} \leq d$. In fact, this relaxation is possible due to the use of Bennett's inequality which only requires that $\xi_k \leq d$. The only reason it was stated in Theorem 2 with the absolute value was simply because we wanted to get without any loss of generality that $\gamma \leq 1$ (due the second requirement on the conditional variance). Finally, since

$$\frac{S_n}{n} = \frac{X_n - X_0}{nd},$$

then it follows from Theorem 2 that for every $x \geq 0$

$$\begin{aligned} & \mathbb{P}\left(\frac{S_n}{n} \geq x\right) \\ &= \mathbb{P}(X_n - X_0 \geq nxd) \\ &\leq \exp\left(-nD\left(\frac{x+v}{1+v} \parallel \frac{v}{1+v}\right)\right) \end{aligned} \quad (67)$$

where, from (11), the correspondence between Theorem 2 and [18, Corollary 2.4.7] is that $\gamma = v$ and $\delta = x$. This shows the relation between Theorem 2 and Eqs. (64) and (65) (respectively, Eqs. (2.4.8) and (2.4.9) in [18]).

We show in the following that Proposition 4 suggests an improvement over the bound in (66) (that is introduced in [18, Eq. (2.4.10)]). To see this, note that from Proposition 4 (see (54)), then for every $y \geq 0$,

$$\begin{aligned} & \mathbb{P}\left(\frac{S_n}{\sqrt{n}} \geq y\right) \\ &= \mathbb{P}(X_n - X_0 \geq yd\sqrt{n}) \\ &\leq \exp\left(-\frac{y^2}{2v}\right) \left(1 + O\left(\frac{1}{\sqrt{n}}\right)\right) \end{aligned} \quad (68)$$

where the term on the right-hand side of (68) that scales like $O\left(\frac{1}{\sqrt{n}}\right)$ is expressed explicitly in terms of n for each concentration inequality that was derived in Section III (see the proof of Proposition 4 in Appendix H). The improvement of the exponent (68) over the exponent in [18, Eq. (2.4.10)] (see (66)) holds since

$$\frac{y^2}{2v} - \frac{2y^2}{(1+v)^2} = \frac{y^2}{2v} \left(\frac{v-1}{v+1}\right)^2 \geq 0$$

with equality if and only if $v = 1$. Note that this improvement is especially pronounced if $v \ll 1$; in the limit where v tends to zero then the improved exponent $\left(\frac{y^2}{2v}\right)$ tends to $+\infty$, whereas the other exponent (*i.e.*, $\frac{2y^2}{(1+v)^2}$) stays bounded.

C. Relations of [18, Execrie 2.4.21(b)], [21, Theorem 1.6] and [53, Theorem 1] with Corollary 3 and Proposition 4

The following theorem was introduced in [21, Theorem 1.6] and [53, Theorem 1] (and in [18, Execrie 2.4.21(b)] with the weaker condition below).

Theorem 6: Let $\{S_n, \mathcal{F}_n\}_{n=0}^\infty$ be a discrete-parameter real-valued martingale such that $S_0 = 0$, and $Y_k \triangleq S_k - S_{k-1} \leq 1$ a.s. for every $k \in \mathbb{N}$. Let us define the random variables

$$Q_n \triangleq \sum_{j=1}^n \mathbb{E}(Y_j^2 | \mathcal{F}_{j-1}) \quad (69)$$

where $Q_0 \triangleq 0$. Then for every $z, r > 0$

$$\mathbb{P}(S_n \geq z, Q_n \leq r) \leq \exp\left(-\frac{z^2}{2r} \cdot B\left(\frac{z}{r}\right)\right) \quad (70)$$

where B was introduced in (36).

Proposition 5: Let $\{X_k, \mathcal{F}_k\}_{k=0}^\infty$ be a discrete-parameter real-valued martingale. Then, Theorem 6 implies the results in Corollary 3 and inequality (54) in Proposition 4.

Proof: See Appendix I. ■

D. Relation between the Martingale Central Limit Theorem (CLT) and Proposition 4

In this subsection, we discuss the relation between the martingale CLT and the concentration inequalities for discrete-parameter martingales in Proposition 4.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Given a filtration $\{\mathcal{F}_k\}$, then $\{Y_k, \mathcal{F}_k\}_{k=0}^\infty$ is said to be a martingale-difference sequence if, for every k ,

- 1) Y_k is \mathcal{F}_k -measurable,
- 2) $\mathbb{E}[|Y_k|] < \infty$,
- 3) $\mathbb{E}[Y_k | \mathcal{F}_{k-1}] = 0$.

Let

$$S_n = \sum_{k=1}^n Y_k, \quad \forall n \in \mathbb{N}$$

and $S_0 = 0$, then $\{S_k, \mathcal{F}_k\}_{k=0}^\infty$ is a martingale. Assume that the sequence of RVs $\{Y_k\}$ is bounded, i.e., there exists a constant d such that $|Y_k| \leq d$ a.s., and furthermore, assume that the limit

$$\sigma^2 \triangleq \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \mathbb{E}[Y_k^2 | \mathcal{F}_{k-1}]$$

exists in probability and is positive. The martingale CLT asserts that, under the above conditions, $\frac{S_n}{\sqrt{n}}$ converges in distribution (i.e., weakly converges) to the Gaussian distribution $\mathcal{N}(0, \sigma^2)$. It is denoted by $\frac{S_n}{\sqrt{n}} \Rightarrow \mathcal{N}(0, \sigma^2)$. We note that there exist more general versions of this statement (see, e.g., [10, pp. 475–478]).

Let $\{X_k, \mathcal{F}_k\}_{k=0}^\infty$ be a discrete-parameter real-valued martingale with bounded jumps, and assume that there exists a constant d so that a.s. for every $k \in \mathbb{N}$

$$|X_k - X_{k-1}| \leq d, \quad \forall k \in \mathbb{N}.$$

Define, for every $k \in \mathbb{N}$,

$$Y_k \triangleq X_k - X_{k-1}$$

and $Y_0 \triangleq 0$, so $\{Y_k, \mathcal{F}_k\}_{k=0}^\infty$ is a martingale-difference sequence, and $|Y_k| \leq d$ a.s. for every $k \in \mathbb{N} \cup \{0\}$. Furthermore, for every $n \in \mathbb{N}$,

$$S_n \triangleq \sum_{k=1}^n Y_k = X_n - X_0.$$

Under the assumptions in Theorem 2 and its subsequences, for every $k \in \mathbb{N}$, one gets a.s. that

$$\mathbb{E}[Y_k^2 | \mathcal{F}_{k-1}] = \mathbb{E}[(X_k - X_{k-1})^2 | \mathcal{F}_{k-1}] \leq \sigma^2.$$

Lets assume that this inequality holds a.s. with equality. It follows from the martingale CLT that

$$\frac{X_n - X_0}{\sqrt{n}} \Rightarrow \mathcal{N}(0, \sigma^2)$$

and therefore, for every $\alpha \geq 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X_0| \geq \alpha \sqrt{n}) = 2Q\left(\frac{\alpha}{\sigma}\right)$$

where the Q function is introduced in (7).

Based on the notation in (11), the equality $\frac{\alpha}{\sigma} = \frac{\delta}{\sqrt{\gamma}}$ holds, and

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X_0| \geq \alpha \sqrt{n}) = 2Q\left(\frac{\delta}{\sqrt{\gamma}}\right). \quad (71)$$

Since, for every $x \geq 0$,

$$Q(x) \leq \frac{1}{2} \exp\left(-\frac{x^2}{2}\right)$$

then it follows that for every $\alpha \geq 0$

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) \leq \exp\left(-\frac{\delta^2}{2\gamma}\right).$$

This inequality coincides with the asymptotic result of the inequalities in Proposition 4 (see (54) in the limit where $n \rightarrow \infty$), except for the additional factor of 2. Note also that the proof of the concentration inequalities in Proposition 4 (see Appendix H) provides inequalities that are informative for finite n , and not only in the asymptotic case where n tends to infinity. Furthermore, due to the exponential upper and lower bounds of the Q-function in (8), then it follows from (71) that the exponent in the concentration inequality (54) (i.e., $\frac{\delta^2}{2\gamma}$) cannot be improved under the above assumptions (unless some more information is available).

E. Relation between the Law of the Iterated Logarithm (LIL) and Theorem 2

In this subsection, we discuss the relation between the law of the iterated logarithm (LIL) and Theorem 2.

According to the law of the iterated logarithm (see, e.g., [10, Theorem 9.5]) if $\{X_k\}_{k=1}^{\infty}$ are i.i.d. real-valued RVs with zero mean and unit variance, and $S_n \triangleq \sum_{i=1}^n X_i$ for every $n \in \mathbb{N}$, then

$$\limsup_{n \rightarrow \infty} \frac{S_n}{\sqrt{2n \ln \ln n}} = 1 \quad \text{a.s.} \quad (72)$$

and

$$\liminf_{n \rightarrow \infty} \frac{S_n}{\sqrt{2n \ln \ln n}} = -1 \quad \text{a.s.} \quad (73)$$

Eqs. (72) and (73) assert, respectively, that for every $\varepsilon > 0$, along almost any realization,

$$S_n > (1 - \varepsilon)\sqrt{2n \ln \ln n}$$

and

$$S_n < -(1 - \varepsilon)\sqrt{2n \ln \ln n}$$

are satisfied infinitely often (i.o.). On the other hand, Eqs. (72) and (73) imply that along almost any realization, each of the two inequalities

$$S_n > (1 + \varepsilon)\sqrt{2n \ln \ln n}$$

and

$$S_n < -(1 + \varepsilon)\sqrt{2n \ln \ln n}$$

is satisfied for a finite number of values of n .

Let $\{X_k\}_{k=1}^{\infty}$ be i.i.d. real-valued RVs, defined over the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with $\mathbb{E}[X_1] = 0$ and $\mathbb{E}[X_1^2] = 1$.

Let us define the natural filtration where $\mathcal{F}_0 = \{\emptyset, \Omega\}$, and $\mathcal{F}_k = \sigma(X_1, \dots, X_k)$ is the σ -algebra that is generated by the RVs X_1, \dots, X_k for every $k \in \mathbb{N}$. Let $S_0 = 0$ and S_n be defined as above for every $n \in \mathbb{N}$. It is straightforward to verify by Definition 1 that $\{S_n, \mathcal{F}_n\}_{n=0}^{\infty}$ is a martingale.

In order to apply Theorem 2 to the considered case, let us assume that the RVs $\{X_k\}_{k=1}^{\infty}$ are uniformly bounded, i.e., it is assumed that there exists a constant c such that $|X_k| \leq c$ a.s. for every $k \in \mathbb{N}$. Since $\mathbb{E}[X_1^2] = 1$ then $c \geq 1$. This assumption implies that the martingale $\{S_n, \mathcal{F}_n\}_{n=0}^{\infty}$ has bounded jumps, and for every $n \in \mathbb{N}$

$$|S_n - S_{n-1}| \leq c \quad \text{a.s.}$$

Moreover, due to the independence of the RVs $\{X_k\}_{k=1}^{\infty}$, then

$$\text{Var}(S_n | \mathcal{F}_{n-1}) = \mathbb{E}(X_n^2 | \mathcal{F}_{n-1}) = \mathbb{E}(X_n^2) = 1 \quad \text{a.s.}$$

From Theorem 2, it follows that for every $\alpha \geq 0$

$$\mathbb{P}\left(S_n \geq \alpha\sqrt{2n \ln \ln n}\right) \leq \exp\left(-nD\left(\frac{\delta_n + \gamma}{1 + \gamma} \parallel \frac{\gamma}{1 + \gamma}\right)\right) \quad (74)$$

where

$$\delta_n \triangleq \frac{\alpha}{c} \sqrt{\frac{2 \ln \ln n}{n}}, \quad \gamma \triangleq \frac{1}{c^2}. \quad (75)$$

Straightforward calculation shows that

$$\begin{aligned} & nD\left(\frac{\delta_n + \gamma}{1 + \gamma} \parallel \frac{\gamma}{1 + \gamma}\right) \\ &= \frac{n\gamma}{1 + \gamma} \left[\left(1 + \frac{\delta_n}{\gamma}\right) \ln\left(1 + \frac{\delta_n}{\gamma}\right) + \frac{1}{\gamma} (1 - \delta_n) \ln(1 - \delta_n) \right] \\ &\stackrel{(a)}{=} \frac{n\gamma}{1 + \gamma} \left[\frac{\delta_n^2}{2} \left(\frac{1}{\gamma^2} + \frac{1}{\gamma}\right) + \frac{\delta_n^3}{6} \left(\frac{1}{\gamma} - \frac{1}{\gamma^3}\right) + \dots \right] \\ &= \frac{n\delta_n^2}{2\gamma} - \frac{n\delta_n^3(1 - \gamma)}{6\gamma^2} + \dots \\ &\stackrel{(b)}{=} \alpha^2 \ln \ln n \left[1 - \frac{\alpha(c^2 - 1)}{6c} \sqrt{\frac{\ln \ln n}{n}} + \dots \right] \end{aligned} \quad (76)$$

where equality (a) follows from the power series expansion

$$(1 + u) \ln(1 + u) = u + \sum_{k=2}^{\infty} \frac{(-u)^k}{k(k-1)}, \quad -1 < u \leq 1$$

and equality (b) follows from (75). A substitution of (76) into (74) gives that, for every $\alpha \geq 0$,

$$\mathbb{P}\left(S_n \geq \alpha \sqrt{2n \ln \ln n}\right) \leq (\ln n)^{-\alpha^2 \left[1 + O\left(\sqrt{\frac{\ln \ln n}{n}}\right)\right]} \quad (77)$$

and the same bound also applies to $\mathbb{P}(S_n \leq -\alpha \sqrt{2n \ln \ln n})$ for $\alpha \geq 0$. This provides complementary information to the limits in (72) and (73) that are provided by the LIL. From Remark 7, which follows from Doob's maximal inequality for sub-martingales, the inequality in (77) can be strengthened to

$$\mathbb{P}\left(\max_{1 \leq k \leq n} S_k \geq \alpha \sqrt{2n \ln \ln n}\right) \leq (\ln n)^{-\alpha^2 \left[1 + O\left(\sqrt{\frac{\ln \ln n}{n}}\right)\right]}. \quad (78)$$

It is shown in the following that (78) and the first Borel-Cantelli lemma can serve to prove one part of (72). Using this approach, it is shown that if $\alpha > 1$, then the probability that $S_n > \alpha \sqrt{2n \ln \ln n}$ i.o. is zero. To this end, let $\theta > 1$ be set arbitrarily, and define

$$A_n = \bigcup_{k: \theta^{n-1} \leq k \leq \theta^n} \left\{ S_k \geq \alpha \sqrt{2k \ln \ln k} \right\}$$

for every $n \in \mathbb{N}$. Hence, the union of these sets is

$$A \triangleq \bigcup_{n \in \mathbb{N}} A_n = \bigcup_{k \in \mathbb{N}} \left\{ S_k \geq \alpha \sqrt{2k \ln \ln k} \right\}$$

The following inequalities hold (since $\theta > 1$):

$$\begin{aligned} \mathbb{P}(A_n) &\leq \mathbb{P}\left(\max_{\theta^{n-1} \leq k \leq \theta^n} S_k \geq \alpha \sqrt{2\theta^{n-1} \ln \ln(\theta^{n-1})}\right) \\ &= \mathbb{P}\left(\max_{\theta^{n-1} \leq k \leq \theta^n} S_k \geq \frac{\alpha}{\sqrt{\theta}} \sqrt{2\theta^n \ln \ln(\theta^{n-1})}\right) \\ &\leq \mathbb{P}\left(\max_{1 \leq k \leq \theta^n} S_k \geq \frac{\alpha}{\sqrt{\theta}} \sqrt{2\theta^n \ln \ln(\theta^{n-1})}\right) \\ &\leq (n \ln \theta)^{-\frac{\alpha^2}{\theta} (1 + \beta_n)} \end{aligned} \quad (79)$$

where the last inequality follows from (78) with $\beta_n \rightarrow 0$ as $n \rightarrow \infty$. Since

$$\sum_{n=1}^{\infty} n^{-\frac{\alpha^2}{\theta}} < \infty, \quad \forall \alpha > \sqrt{\theta}$$

then it follows from the first Borel-Cantelli lemma that $\mathbb{P}(A \text{ i.o.}) = 0$ for all $\alpha > \sqrt{\theta}$. But the event A does not depend on θ , and $\theta > 1$ can be made arbitrarily close to 1. This asserts that $\mathbb{P}(A \text{ i.o.}) = 0$ for every $\alpha > 1$, or equivalently

$$\limsup_{n \rightarrow \infty} \frac{S_n}{\sqrt{2n \ln \ln n}} \leq 1 \quad \text{a.s.}$$

Similarly, by replacing $\{X_i\}$ with $\{-X_i\}$, it follows that

$$\liminf_{n \rightarrow \infty} \frac{S_n}{\sqrt{2n \ln \ln n}} \geq -1 \quad \text{a.s.}$$

Theorem 2 therefore gives inequality (78), and it implies one side in each of the two equalities for the LIL in (72) and (73).

F. Relation of Theorems 2 and 4 with the Moderate Deviations Principle

According to the moderate deviations theorem (see, e.g., [18, Theorem 3.7.1]) in \mathbb{R} , let $\{X_i\}_{i=1}^n$ be a sequence of real-valued i.i.d. RVs such that $\Lambda_X(\lambda) = \mathbb{E}[e^{\lambda X_i}] < \infty$ in some neighborhood of zero, and also assume that $\mathbb{E}[X_i] = 0$ and $\sigma^2 = \text{Var}(X_i) > 0$. Let $\{a_n\}_{n=1}^{\infty}$ be a non-negative sequence such that $a_n \rightarrow 0$ and $na_n \rightarrow \infty$ as $n \rightarrow \infty$, and let

$$Z_n \triangleq \sqrt{\frac{a_n}{n}} \sum_{i=1}^n X_i, \quad \forall n \in \mathbb{N}. \quad (80)$$

Then, for every measurable set $\Gamma \subseteq \mathbb{R}$,

$$\begin{aligned} & -\frac{1}{2\sigma^2} \inf_{x \in \Gamma^0} x^2 \\ & \leq \liminf_{n \rightarrow \infty} a_n \ln \mathbb{P}(Z_n \in \Gamma) \\ & \leq \limsup_{n \rightarrow \infty} a_n \ln \mathbb{P}(Z_n \in \Gamma) \\ & \leq -\frac{1}{2\sigma^2} \inf_{x \in \bar{\Gamma}} x^2 \end{aligned} \quad (81)$$

where Γ^0 and $\bar{\Gamma}$ designate, respectively, the interior and closure sets of Γ .

Let $\eta \in (\frac{1}{2}, 1)$ be an arbitrary fixed number, and let $\{a_n\}_{n=1}^{\infty}$ be the non-negative sequence

$$a_n = n^{1-2\eta}, \quad \forall n \in \mathbb{N}$$

so that $a_n \rightarrow 0$ and $na_n \rightarrow \infty$ as $n \rightarrow \infty$. Let $\alpha \in \mathbb{R}^+$, and $\Gamma \triangleq (-\infty, -\alpha] \cup [\alpha, \infty)$. Note that, from (80),

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| \geq \alpha n^\eta\right) = \mathbb{P}(Z_n \in \Gamma)$$

so from the moderate deviations principle (MDP), for every $\alpha \geq 0$,

$$\lim_{n \rightarrow \infty} n^{1-2\eta} \ln \mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| \geq \alpha n^\eta\right) = -\frac{\alpha^2}{2\sigma^2}. \quad (82)$$

It is demonstrated in Appendix J that, in contrast to Azuma's inequality, Theorems 2 and 4 (for every even $m \geq 2$ in Theorem 4) provide upper bounds on the probability

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| \geq \alpha n^\eta\right), \quad \forall n \in \mathbb{N}, \alpha \geq 0$$

which both coincide with the asymptotic limit in (82). The analysis in Appendix J provides another interesting link between Theorems 2 and 4 and a classical result in probability theory, which also emphasizes the significance of the refinements of Azuma's inequality.

G. Relation of [38, Lemma 2.8] with Theorem 4 & Corollary 4

In [38, Lemma 2.8], it is proved that if X is a random variable that satisfies $\mathbb{E}[X] = 0$ and $X \leq d$ a.s. (for some $d > 0$), then

$$\mathbb{E}[e^X] \leq \exp(\varphi(d) \text{Var}(X)) \quad (83)$$

where

$$\varphi(x) = \begin{cases} \frac{\exp(x)-1-x}{x^2} & \text{if } x \neq 0 \\ \frac{1}{2} & \text{if } x = 0 \end{cases}.$$

From (41), it follows that $\varphi(x) = \frac{\varphi_2(x)}{2}$ for every $x \in \mathbb{R}$. Based on [38, Lemma 2.8], it follows that if $\{\xi_k, \mathcal{F}_k\}$ is a difference-martingale sequence (i.e., for every $k \in \mathbb{N}$,

$$\mathbb{E}[\xi_k | \mathcal{F}_{k-1}] = 0$$

a.s.), and $\xi_k \leq d$ a.s. for some $d > 0$, then for an arbitrary $t \geq 0$

$$\mathbb{E}[\exp(t\xi_k) | \mathcal{F}_{k-1}] \leq \exp\left(\frac{\gamma (td)^2 \varphi_2(td)}{2}\right)$$

holds a.s. for every $k \in \mathbb{N}$ (the parameter γ was introduced in (11)). The last inequality can be rewritten as

$$\mathbb{E}[\exp(t\xi_k) | \mathcal{F}_{k-1}] \leq \exp(\gamma (\exp(td) - 1 - td)), \quad t \geq 0. \quad (84)$$

This forms a looser bound on the conditional expectation, as compared to (43) with $m = 2$, that gets the form

$$\mathbb{E}[\exp(t\xi_k) | \mathcal{F}_{k-1}] \leq 1 + \gamma (\exp(td) - 1 - td), \quad t \geq 0. \quad (85)$$

The improvement in (85) over (84) follows since $e^x \geq 1 + x$ for $x \geq 0$ with equality if and only if $x = 0$. Note that the proof of [38, Lemma 2.8] shows that indeed the right-hand side of (85) forms an upper bound on the above conditional expectation, whereas it is loosened to the bound on the right-hand side of (84) in order to handle the case where

$$\frac{1}{n} \sum_{k=1}^n \mathbb{E}[(\xi_k)^2 | \mathcal{F}_{k-1}] \leq \sigma^2$$

and derive a closed-form solution of the optimized parameter t in the resulting concentration inequality (see the proof of [38, Theorem 2.7] for the case of independent RVs, and also [38, Theorem 3.15] for the setting of martingales with bounded jumps). However, if for every $k \in \mathbb{N}$, the condition

$$\mathbb{E}[(\xi_k)^2 | \mathcal{F}_{k-1}] \leq \sigma^2$$

holds a.s., then the proof of Corollary 4 shows that a closed-form solution of the non-negative free parameter t is obtained. More on the consequence of the difference between the bounds in (84) and (85) is considered in the next sub-section.

H. Relation of the Concentration Inequalities for Martingales to Discrete-Time Markov Chains

A striking well-known relation between discrete-time Markov chains and martingales is the following (see, e.g., [23, p. 473]): Let $\{X_n\}_{n \in \mathbb{N}_0}$ ($\mathbb{N}_0 \triangleq \mathbb{N} \cup \{0\}$) be a discrete-time Markov chain taking values in a countable state space \mathcal{S} with transition matrix \mathbf{P} , and let the function $\psi : \mathcal{S} \rightarrow \mathcal{S}$ be harmonic (i.e., $\sum_{j \in \mathcal{S}} p_{i,j} \psi(j) = \psi(i)$, $\forall i \in \mathcal{S}$), and assume that $E[|\psi(X_n)|] < \infty$ for every n . Then, $\{Y_n, \mathcal{F}_n\}_{n \in \mathbb{N}_0}$ is a martingale where $Y_n \triangleq \psi(X_n)$ and $\{\mathcal{F}_n\}_{n \in \mathbb{N}_0}$ is the natural filtration. This relation, which follows directly from the Markov property, enables to apply the concentration inequalities in Section III for harmonic functions of Markov chains when the function ψ is bounded (so that the jumps of the martingale sequence are uniformly bounded).

Exponential deviation bounds for an important class of Markov chains, called Doeblin chains (they are characterized by an exponentially fast convergence to the equilibrium, uniformly in the initial condition) were derived in [30]. These bounds were also shown to be essentially identical to the Hoeffding inequality in the special case of i.i.d. RVs (see [30, Remark 1]).

I. Relations of [13, Theorem 2.23] with Corollary 4 and Proposition 4

In the following, we consider the relation between the inequalities in Corollary 4 and Proposition 4 to the particularized form of [13, Theorem 2.23] (or also [12, Theorem 2.23]) in the setting where $d_k = d$ and $\sigma_k^2 = \sigma^2$ are fixed for every $k \in \mathbb{N}$. The resulting exponents of these concentration inequalities are also compared.

Let $\alpha \geq 0$ be an arbitrary non-negative number.

- In the analysis of small deviations, the bound in [13, Theorem 2.23] is particularized to

$$\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) \leq 2 \exp\left(-\frac{\alpha^2 n}{2n\sigma^2 + \frac{2d\alpha\sqrt{n}}{3}}\right).$$

From the notation in (11) then $\frac{\alpha^2}{\sigma^2} = \frac{\delta^2}{\gamma}$, and the last inequality gets the form

$$\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) \leq 2 \exp\left(-\frac{\delta^2}{2\gamma}\right) \left(1 + O\left(\frac{1}{\sqrt{n}}\right)\right).$$

It therefore follows that [13, Theorem 2.23] implies a concentration inequality of the form in (54). This shows that Proposition 4 can be also regarded as a consequence of [13, Theorem 2.23].

- In the analysis of large deviations, the bound in [13, Theorem 2.23] is particularized to

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) \leq 2 \exp\left(-\frac{\alpha^2 n}{2\sigma^2 + \frac{2d\alpha}{3}}\right).$$

From the notation in (11), this inequality is rewritten as

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) \leq 2 \exp\left(-\frac{\delta^2 n}{2\gamma + \frac{2\delta}{3}}\right). \quad (86)$$

It is claimed that the concentration inequality in (86) is looser than Corollary 4. This is a consequence of the proof of [13, Theorem 2.23] where the derived concentration inequality is loosened in order to handle the more general case, as compared to the setting in this paper (see Theorem 2), where d_k and σ_k^2 may depend on k . In order to show it explicitly, let's compare between the steps of the derivation of the bound in Corollary 4, and the particularization of the derivation of [13, Theorem 2.23] in the special setting where d_k and σ_k^2 are independent of k . This comparison is considered in the following. The derivation of the concentration inequality in Corollary 4 follows by substituting $m = 2$ in the proof of Theorem 4. It then follows that, for every $\alpha \geq 0$,

$$\mathbb{P}(X_n - X_0 \geq \alpha n) \leq e^{-n\delta x} \left(1 + \gamma(e^x - 1 - x)\right)^n, \quad \forall x \geq 0 \quad (87)$$

which then leads, after an analytic optimization of the free non-negative parameter x (see Lemma 6 and Appendix C), to the derivation of Corollary 4. On the other hand, the specialization of the proof of [13, Theorem 2.23] to the case where $d_k = d$ and $\sigma_k^2 = \sigma^2$ for every $k \in \mathbb{N}$ is equivalent to a further loosening of (87) to the bound

$$\begin{aligned} & \mathbb{P}(X_n - X_0 \geq \alpha n) \\ & \leq e^{-n\delta x} e^{n\gamma(e^x - 1 - x)} \end{aligned} \quad (88)$$

$$\leq e^{n\left(-\delta x + \frac{\gamma x^2}{1 - \frac{\gamma}{3}}\right)}, \quad \forall x \in (0, 3) \quad (89)$$

and then choosing an optimal $x \in (0, 3)$. This indeed shows that Corollary 4 provides a concentration inequality that is more tight than the bound in [13, Theorem 2.23].

In order to compare quantitatively the exponents of the concentration inequalities in [13, Theorem 2.23] and Corollary 4, let us revisit the derivation of the upper bounds on the probability of the events $\{|X_n - X_0| \geq \alpha n\}$ where $\alpha \geq 0$ is arbitrary. The optimized value of x that is obtained in Appendix C is positive, and it becomes larger as we let the value of $\gamma \in (0, 1]$ approach zero. Hence, especially for small values of γ , the loosening of the bound from (87) to (89) is expected to deteriorate more significantly the resulting bound in [13, Theorem 2.23] due to the restriction that $x \in (0, 3)$; this is in contrast to the optimized value of x in Appendix C that may be above 3 for small values of γ , and it lies in general between 0 and $\frac{1}{\gamma}$. Note also that at $\delta = 1$, the exponent in Corollary 4 tends to infinity in the limit where $\gamma \rightarrow 0$, whereas the exponent in (86) tends in this case to $\frac{3}{2}$. To

illustrate these differences, Figure 4 plots the exponents of the bounds in Corollary 4 and (86), where the latter refers to [13, Theorem 2.23], for $\gamma = 0.01$ and 0.99 . As is shown in Figure 4, the difference between the exponents of these two bounds is indeed more pronounced when γ gets closer to zero.

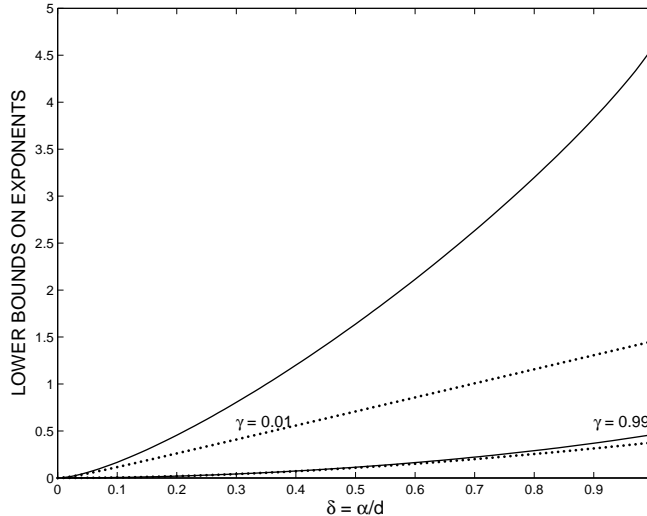


Fig. 4. A comparison of the exponents of the bound in Corollary 4 and the particularized bound (86) from [13, Theorem 2.23]. This comparison is done for both $\gamma = 0.01$ and 0.99 . The solid curves refer to the exponents of the bound in Corollary 4, and the dashed curves refer to the exponents of the looser bound in (86). The upper pair of curves refers to the exponents for $\gamma = 0.01$, and the lower pair of curves (that approximately coincide) refers to the exponents for $\gamma = 0.99$.

Consider, on the other hand, the probability of an event $\{|X_n - X_0| \geq \alpha\sqrt{n}\}$ where $\alpha \geq 0$ is arbitrary. It is shown in Appendix D that the optimized value of x for the bound in Corollary 4 (and its generalized version in Theorem 4) scales like $\frac{1}{\sqrt{n}}$. Hence, it is approximately zero for $n \gg 1$, and $u \triangleq \gamma(e^x - 1 - x) \approx \frac{\gamma x^2}{2}$ scales like $\frac{1}{n}$. It therefore follows that $(1 + u)^n \approx e^{nu}$ for $n \gg 1$. Moreover, the restriction on x to be less than 3 in (89) does not affect the tightness of the bound in this case since the optimized value of x is anyway close to zero. This explains the observation that the two bounds in Proposition 4 and [13, Theorem 2.23] essentially scale similarly for small deviations of order \sqrt{n} .

V. APPLICATIONS IN INFORMATION THEORY AND RELATED TOPICS

The refined versions of Azuma's inequality in Section III are applied in this section to information-theoretic aspects.

A. Binary Hypothesis Testing

Binary hypothesis testing for finite alphabet models was analyzed via the method of types, e.g., in [15, Chapter 11] and [16]. It is assumed that the data sequence is of a fixed length (n), and one wishes to make the optimal decision based on the received sequence and the Neyman-Pearson ratio test.

Let the RVs X_1, X_2, \dots be i.i.d. $\sim Q$, and consider two hypotheses:

- $H_1 : Q = P_1$.
- $H_2 : Q = P_2$.

For the simplicity of the analysis, let us assume that the RVs are discrete, and take their values on a finite alphabet \mathcal{X} where $P_1(x), P_2(x) > 0$ for every $x \in \mathcal{X}$.

In the following, let

$$L(X_1, \dots, X_n) \triangleq \ln \frac{P_1^n(X_1, \dots, X_n)}{P_2^n(X_1, \dots, X_n)} = \sum_{i=1}^n \ln \frac{P_1(X_i)}{P_2(X_i)}$$

designate the log-likelihood ratio. By the strong law of large numbers (SLLN), if hypothesis H_1 is true, then a.s.

$$\lim_{n \rightarrow \infty} \frac{L(X_1, \dots, X_n)}{n} = D(P_1 || P_2) \quad (90)$$

and otherwise, if hypothesis H_2 is true, then a.s.

$$\lim_{n \rightarrow \infty} \frac{L(X_1, \dots, X_n)}{n} = -D(P_2 || P_1) \quad (91)$$

where the above assumptions on the probability mass functions P_1 and P_2 imply that the relative entropies, $D(P_1 || P_2)$ and $D(P_2 || P_1)$, are both finite. Consider the case where for some fixed constants $\bar{\lambda}, \underline{\lambda} \in \mathbb{R}$ that satisfy

$$-D(P_2 || P_1) < \underline{\lambda} \leq \bar{\lambda} < D(P_1 || P_2)$$

one decides on hypothesis H_1 if

$$L(X_1, \dots, X_n) > n\bar{\lambda}$$

and on hypothesis H_2 if

$$L(X_1, \dots, X_n) < n\underline{\lambda}.$$

Note that if $\bar{\lambda} = \underline{\lambda} \triangleq \lambda$ then a decision on the two hypotheses is based on comparing the normalized log-likelihood ratio (w.r.t. n) to a single threshold (λ), and deciding on hypothesis H_1 or H_2 if it is, respectively, above or below λ . If $\underline{\lambda} < \bar{\lambda}$ then one decides on H_1 or H_2 if the normalized log-likelihood ratio is, respectively, above the upper threshold $\bar{\lambda}$ or below the lower threshold $\underline{\lambda}$. Otherwise, if the normalized log-likelihood ratio is between the upper and lower thresholds, then an erasure is declared and no decision is taken in this case.

Let

$$\alpha_n^{(1)} \triangleq P_1^n \left(L(X_1, \dots, X_n) \leq n\bar{\lambda} \right) \quad (92)$$

$$\alpha_n^{(2)} \triangleq P_1^n \left(L(X_1, \dots, X_n) \leq n\underline{\lambda} \right) \quad (93)$$

and

$$\beta_n^{(1)} \triangleq P_2^n \left(L(X_1, \dots, X_n) \geq n\underline{\lambda} \right) \quad (94)$$

$$\beta_n^{(2)} \triangleq P_2^n \left(L(X_1, \dots, X_n) \geq n\bar{\lambda} \right) \quad (95)$$

then $\alpha_n^{(1)}$ and $\beta_n^{(1)}$ are the probabilities of either making an error or declaring an erasure under, respectively, hypotheses H_1 and H_2 ; similarly, $\alpha_n^{(2)}$ and $\beta_n^{(2)}$ are the probabilities of making an error under hypotheses H_1 and H_2 , respectively.

Let $\pi_1, \pi_2 \in (0, 1)$ denote the a-priori probabilities of the hypotheses H_1 and H_2 , respectively, so

$$P_{e,n}^{(1)} = \pi_1 \alpha_n^{(1)} + \pi_2 \beta_n^{(1)} \quad (96)$$

is the probability of having either an error or an erasure, and

$$P_{e,n}^{(2)} = \pi_1 \alpha_n^{(2)} + \pi_2 \beta_n^{(2)} \quad (97)$$

is the probability of error.

1) *Exact Exponents:* When we let n tend to infinity, the exact exponents of $\alpha_n^{(j)}$ and $\beta_n^{(j)}$ ($j = 1, 2$) are derived via Cramér's theorem. The resulting exponents form a straightforward generalization of, e.g., [18, Theorem 3.4.3] and [27, Theorem 6.4] that addresses the case where the decision is made based on a single threshold of the log-likelihood ratio. In this particular case where $\bar{\lambda} = \underline{\lambda} \triangleq \lambda$, the option of erasures does not exist, and $P_{e,n}^{(1)} = P_{e,n}^{(2)} \triangleq P_{e,n}$ is the error probability.

In the considered general case with erasures, let

$$\lambda_1 \triangleq -\bar{\lambda}, \quad \lambda_2 \triangleq -\underline{\lambda}$$

then Cramér's theorem on \mathbb{R} yields that the exact exponents of $\alpha_n^{(1)}$, $\alpha_n^{(2)}$, $\beta_n^{(1)}$ and $\beta_n^{(2)}$ are given by

$$\lim_{n \rightarrow \infty} -\frac{\ln \alpha_n^{(1)}}{n} = I(\lambda_1) \quad (98)$$

$$\lim_{n \rightarrow \infty} -\frac{\ln \alpha_n^{(2)}}{n} = I(\lambda_2) \quad (99)$$

$$\lim_{n \rightarrow \infty} -\frac{\ln \beta_n^{(1)}}{n} = I(\lambda_2) - \lambda_2 \quad (100)$$

$$\lim_{n \rightarrow \infty} -\frac{\ln \beta_n^{(2)}}{n} = I(\lambda_1) - \lambda_1 \quad (101)$$

where the rate function I is given by

$$I(r) \triangleq \sup_{t \in \mathbb{R}} (tr - H(t)) \quad (102)$$

and

$$H(t) = \ln \left(\sum_{x \in \mathcal{X}} P_1(x)^{1-t} P_2(x)^t \right), \quad \forall t \in \mathbb{R}. \quad (103)$$

The rate function I is convex, lower semi-continuous (l.s.c.) and non-negative (see, e.g., [18] and [27]). Note that

$$H(t) = (t-1)D_t(P_2||P_1)$$

where $D_t(P||Q)$ designates Rényi's information divergence of order t [45, Eq. (3.3)], and I in (102) is the Fenchel-Legendre transform of H (see, e.g., [18, Definition 2.2.2]).

From (96)–(101), the exact exponents of $P_{e,n}^{(1)}$ and $P_{e,n}^{(2)}$ are equal to

$$\lim_{n \rightarrow \infty} -\frac{\ln P_{e,n}^{(1)}}{n} = \min \left\{ I(\lambda_1), I(\lambda_2) - \lambda_2 \right\} \quad (104)$$

and

$$\lim_{n \rightarrow \infty} -\frac{\ln P_{e,n}^{(2)}}{n} = \min \left\{ I(\lambda_2), I(\lambda_1) - \lambda_1 \right\}. \quad (105)$$

For the case where the decision is based on a single threshold for the log-likelihood ratio (i.e., $\lambda_1 = \lambda_2 \triangleq \lambda$), then $P_{e,n}^{(1)} = P_{e,n}^{(2)} \triangleq P_{e,n}$, and its error exponent is equal to

$$\lim_{n \rightarrow \infty} -\frac{\ln P_{e,n}}{n} = \min \left\{ I(\lambda), I(\lambda) - \lambda \right\} \quad (106)$$

which coincides with the error exponent in [18, Theorem 3.4.3] (or [27, Theorem 6.4]). The optimal threshold for obtaining the best error exponent of the error probability $P_{e,n}$ is equal to zero (i.e., $\lambda = 0$); in this case, the exact error exponent is equal to

$$\begin{aligned} I(0) &= - \min_{0 \leq t \leq 1} \ln \left(\sum_{x \in \mathcal{X}} P_1(x)^{1-t} P_2(x)^t \right) \\ &\triangleq C(P_1, P_2) \end{aligned} \quad (107)$$

which is the Chernoff information of the probability measures P_1 and P_2 (see [15, Eq. (11.239)]), and it is symmetric (i.e., $C(P_1, P_2) = C(P_2, P_1)$). Note that, from (102), $I(0) = \sup_{t \in \mathbb{R}} (-H(t)) = -\inf_{t \in \mathbb{R}} (H(t))$; the minimization in (107) over the interval $[0, 1]$ (instead of taking the infimum of H over \mathbb{R}) is due to the fact that $H(0) = H(1) = 0$ and the function H in (103) is convex, so it is enough to restrict the infimum of H to the closed interval $[0, 1]$ for which it turns to be a minimum.

Paper [11] forms a classical paper that considers binary hypothesis testing from an information-theoretic point of view, and it derives the error exponents of binary hypothesis testers in analogy to optimum channel codes via the use of relative entropy measures. We will further explore on this kind of analogy in the continuation to this section (see later Sections V-A5 and V-A6 w.r.t. moderate and small deviations analysis of binary hypothesis testing).

2) *Lower Bound on the Exponents via Theorem 2:* In the following, the tightness of Theorem 2 is examined by using it for the derivation of lower bounds on the error exponent and the exponent of the event of having either an error or an erasure. These results will be compared in the next sub-section to the exact exponents from the previous sub-section.

We first derive a lower bound on the exponent of $\alpha_n^{(1)}$. Under hypothesis H_1 , let us construct the martingale sequence $\{U_k, \mathcal{F}_k\}_{k=0}^n$ where $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \dots \subseteq \mathcal{F}_n$ is the filtration

$$\mathcal{F}_0 = \{\emptyset, \Omega\}, \quad \mathcal{F}_k = \sigma(X_1, \dots, X_k), \quad \forall k \in \{1, \dots, n\}$$

and

$$U_k = \mathbb{E}_{P_1^n} [L(X_1, \dots, X_n) \mid \mathcal{F}_k]. \quad (108)$$

For every $k \in \{0, \dots, n\}$

$$\begin{aligned} U_k &= \mathbb{E}_{P_1^n} \left[\sum_{i=1}^n \ln \frac{P_1(X_i)}{P_2(X_i)} \mid \mathcal{F}_k \right] \\ &= \sum_{i=1}^k \ln \frac{P_1(X_i)}{P_2(X_i)} + \sum_{i=k+1}^n \mathbb{E}_{P_1^n} \left[\ln \frac{P_1(X_i)}{P_2(X_i)} \right] \\ &= \sum_{i=1}^k \ln \frac{P_1(X_i)}{P_2(X_i)} + (n-k)D(P_1 \parallel P_2). \end{aligned}$$

In particular

$$U_0 = nD(P_1 \parallel P_2), \quad (109)$$

$$U_n = \sum_{i=1}^n \ln \frac{P_1(X_i)}{P_2(X_i)} = L(X_1, \dots, X_n) \quad (110)$$

and, for every $k \in \{1, \dots, n\}$,

$$U_k - U_{k-1} = \ln \frac{P_1(X_k)}{P_2(X_k)} - D(P_1 \parallel P_2). \quad (111)$$

Let

$$d_1 \triangleq \max_{x \in \mathcal{X}} \left| \ln \frac{P_1(x)}{P_2(x)} - D(P_1 \parallel P_2) \right| \quad (112)$$

so $d_1 < \infty$ since by assumption the alphabet set \mathcal{X} is finite, and $P_1(x), P_2(x) > 0$ for every $x \in \mathcal{X}$. From (111) and (112)

$$|U_k - U_{k-1}| \leq d_1$$

holds a.s. for every $k \in \{1, \dots, n\}$, and due to the statistical independence of the RVs in the sequence $\{X_i\}$

$$\begin{aligned} &\mathbb{E}_{P_1^n} [(U_k - U_{k-1})^2 \mid \mathcal{F}_{k-1}] \\ &= \mathbb{E}_{P_1} \left[\left(\ln \frac{P_1(X_k)}{P_2(X_k)} - D(P_1 \parallel P_2) \right)^2 \right] \\ &= \sum_{x \in \mathcal{X}} \left\{ P_1(x) \left(\ln \frac{P_1(x)}{P_2(x)} - D(P_1 \parallel P_2) \right)^2 \right\} \\ &\triangleq \sigma_1^2. \end{aligned} \quad (113)$$

Let

$$\varepsilon_{1,1} = D(P_1 \parallel P_2) - \bar{\lambda}, \quad \varepsilon_{2,1} = D(P_2 \parallel P_1) + \underline{\lambda} \quad (114)$$

$$\varepsilon_{1,2} = D(P_1 \parallel P_2) - \underline{\lambda}, \quad \varepsilon_{2,2} = D(P_2 \parallel P_1) + \bar{\lambda} \quad (115)$$

The probability of making an erroneous decision on hypothesis H_2 or declaring an erasure under the hypothesis H_1 is equal to $\alpha_n^{(1)}$, and from Theorem 2

$$\begin{aligned} \alpha_n^{(1)} &\triangleq P_1^n(L(X_1, \dots, X_n) \leq n\bar{\lambda}) \\ &\stackrel{(a)}{=} P_1^n(U_n - U_0 \leq -\varepsilon_{1,1} n) \end{aligned} \quad (116)$$

$$\leq \exp\left(-n D\left(\frac{\delta_{1,1} + \gamma_1}{1 + \gamma_1} \parallel \frac{\gamma_1}{1 + \gamma_1}\right)\right) \quad (117)$$

where equality (a) follows from (109), (110) and (114), and inequality (b) follows from Theorem 2 with

$$\gamma_1 \triangleq \frac{\sigma_1^2}{d_1^2}, \quad \delta_{1,1} \triangleq \frac{\varepsilon_{1,1}}{d_1}. \quad (118)$$

Note that if $\varepsilon_{1,1} > d_1$ then it follows from (111) and (112) that $\alpha_n^{(1)}$ is zero; in this case $\delta_{1,1} > 1$, so the divergence in (117) is infinity and the upper bound is also equal to zero. Hence, it is assumed without loss of generality that $\delta_{1,1} \in [0, 1]$.

Similarly to (108), under hypothesis H_2 , let us define the martingale sequence $\{U_k, \mathcal{F}_k\}_{k=0}^n$ with the same filtration and

$$U_k = \mathbb{E}_{P_2^n}[L(X_1, \dots, X_n) \mid \mathcal{F}_k], \quad \forall k \in \{0, \dots, n\}. \quad (119)$$

For every $k \in \{0, \dots, n\}$

$$U_k = \sum_{i=1}^k \ln \frac{P_1(X_i)}{P_2(X_i)} - (n - k)D(P_2 \parallel P_1)$$

and in particular

$$U_0 = -nD(P_2 \parallel P_1), \quad U_n = L(X_1, \dots, X_n). \quad (120)$$

For every $k \in \{1, \dots, n\}$,

$$U_k - U_{k-1} = \ln \frac{P_1(X_k)}{P_2(X_k)} + D(P_2 \parallel P_1). \quad (121)$$

Let

$$d_2 \triangleq \max_{x \in \mathcal{X}} \left| \ln \frac{P_2(x)}{P_1(x)} - D(P_2 \parallel P_1) \right| \quad (122)$$

then, the jumps of the latter martingale sequence are uniformly bounded by d_2 and, similarly to (113), for every $k \in \{1, \dots, n\}$

$$\begin{aligned} &\mathbb{E}_{P_2^n}[(U_k - U_{k-1})^2 \mid \mathcal{F}_{k-1}] \\ &= \sum_{x \in \mathcal{X}} \left\{ P_2(x) \left(\ln \frac{P_2(x)}{P_1(x)} - D(P_2 \parallel P_1) \right)^2 \right\} \\ &\triangleq \sigma_2^2. \end{aligned} \quad (123)$$

Hence, it follows from Theorem 2 that

$$\begin{aligned} \beta_n^{(1)} &\triangleq P_2^n(L(X_1, \dots, X_n) \geq n\underline{\lambda}) \\ &= P_2^n(U_n - U_0 \geq \varepsilon_{2,1} n) \end{aligned} \quad (124)$$

$$\leq \exp\left(-n D\left(\frac{\delta_{2,1} + \gamma_2}{1 + \gamma_2} \parallel \frac{\gamma_2}{1 + \gamma_2}\right)\right) \quad (125)$$

where the equality in (124) holds due to (120) and (114), and (125) follows from Theorem 2 with

$$\gamma_2 \triangleq \frac{\sigma_2^2}{d_2^2}, \quad \delta_{2,1} \triangleq \frac{\varepsilon_{2,1}}{d_2} \quad (126)$$

and d_2, σ_2 are introduced, respectively, in (122) and (123).

From (96), (117) and (125), the exponent of the probability of either having an error or an erasure is lower bounded by

$$\lim_{n \rightarrow \infty} -\frac{\ln P_{e,n}^{(1)}}{n} \geq \min_{i=1,2} D\left(\frac{\delta_{i,1} + \gamma_i}{1 + \gamma_i} \parallel \frac{\gamma_i}{1 + \gamma_i}\right). \quad (127)$$

Similarly to the above analysis, one gets from (97) and (115) that the error exponent is lower bounded by

$$\lim_{n \rightarrow \infty} -\frac{\ln P_{e,n}^{(2)}}{n} \geq \min_{i=1,2} D\left(\frac{\delta_{i,2} + \gamma_i}{1 + \gamma_i} \parallel \frac{\gamma_i}{1 + \gamma_i}\right) \quad (128)$$

where

$$\delta_{1,2} \triangleq \frac{\varepsilon_{1,2}}{d_1}, \quad \delta_{2,2} \triangleq \frac{\varepsilon_{2,2}}{d_2}. \quad (129)$$

For the case of a single threshold (i.e., $\bar{\lambda} = \underline{\lambda} \triangleq \lambda$) then (127) and (128) coincide, and one obtains that the error exponent satisfies

$$\lim_{n \rightarrow \infty} -\frac{\ln P_{e,n}}{n} \geq \min_{i=1,2} D\left(\frac{\delta_i + \gamma_i}{1 + \gamma_i} \parallel \frac{\gamma_i}{1 + \gamma_i}\right) \quad (130)$$

where δ_i is the common value of $\delta_{i,1}$ and $\delta_{i,2}$ (for $i = 1, 2$). In this special case, the zero threshold is optimal (see, e.g., [18, p. 93]), which then yields that (130) is satisfied with

$$\delta_1 = \frac{D(P_1 \parallel P_2)}{d_1}, \quad \delta_2 = \frac{D(P_2 \parallel P_1)}{d_2} \quad (131)$$

with d_1 and d_2 from (112) and (122), respectively. The right-hand side of (130) forms a lower bound on Chernoff information which is the exact error exponent for this special case.

3) *Comparison of the Lower Bounds on the Exponents with those that Follow from Azuma's Inequality:* The lower bounds on the error exponent and the exponent of the probability of having either errors or erasures, that were derived in the previous sub-section via Theorem 2, are compared in the following to the loosened lower bounds on these exponents that follow from Azuma's inequality.

We first obtain upper bounds on $\alpha_n^{(1)}$, $\alpha_n^{(2)}$, $\beta_n^{(1)}$ and $\beta_n^{(2)}$ via Azuma's inequality, and then use them to derive lower bounds on the exponents of $P_{e,n}^{(1)}$ and $P_{e,n}^{(2)}$.

From (111), (112), (116), (118), and Azuma's inequality

$$\alpha_n^{(1)} \leq \exp\left(-\frac{\delta_{1,1}^2 n}{2}\right) \quad (132)$$

and, similarly, from (121), (122), (124), (126), and Azuma's inequality

$$\beta_n^{(1)} \leq \exp\left(-\frac{\delta_{2,1}^2 n}{2}\right). \quad (133)$$

From (93), (95), (115), (129) and Azuma's inequality

$$\alpha_n^{(2)} \leq \exp\left(-\frac{\delta_{1,2}^2 n}{2}\right) \quad (134)$$

$$\beta_n^{(2)} \leq \exp\left(-\frac{\delta_{2,2}^2 n}{2}\right). \quad (135)$$

Therefore, it follows from (96), (97) and (132)–(135) that the resulting lower bounds on the exponents of $P_{e,n}^{(1)}$ and $P_{e,n}^{(2)}$ are

$$\lim_{n \rightarrow \infty} -\frac{\ln P_{e,n}^{(j)}}{n} \geq \min_{i=1,2} \frac{\delta_{i,j}^2}{2}, \quad j = 1, 2 \quad (136)$$

as compared to (127) and (128) which give, for $j = 1, 2$,

$$\lim_{n \rightarrow \infty} -\frac{\ln P_{e,n}^{(j)}}{n} \geq \min_{i=1,2} D\left(\frac{\delta_{i,j} + \gamma_i}{1 + \gamma_i} \parallel \frac{\gamma_i}{1 + \gamma_i}\right). \quad (137)$$

For the specific case of a zero threshold, the lower bound on the error exponent which follows from Azuma's inequality is given by

$$\lim_{n \rightarrow \infty} -\frac{\ln P_{e,n}^{(j)}}{n} \geq \min_{i=1,2} \frac{\delta_i^2}{2} \quad (138)$$

with the values of δ_1 and δ_2 in (131).

The lower bounds on the exponents in (136) and (137) are compared in the following. Note that the lower bounds in (136) are loosened as compared to those in (137) since they follow, respectively, from Azuma's inequality and its improvement in Theorem 2.

The divergence in the exponent of (137) is equal to

$$\begin{aligned} & D\left(\frac{\delta_{i,j} + \gamma_i}{1 + \gamma_i} \parallel \frac{\gamma_i}{1 + \gamma_i}\right) \\ &= \left(\frac{\delta_{i,j} + \gamma_i}{1 + \gamma_i}\right) \ln\left(1 + \frac{\delta_{i,j}}{\gamma_i}\right) + \left(\frac{1 - \delta_{i,j}}{1 + \gamma_i}\right) \ln(1 - \delta_{i,j}) \\ &= \frac{\gamma_i}{1 + \gamma_i} \left[\left(1 + \frac{\delta_{i,j}}{\gamma_i}\right) \ln\left(1 + \frac{\delta_{i,j}}{\gamma_i}\right) + \frac{(1 - \delta_{i,j}) \ln(1 - \delta_{i,j})}{\gamma_i} \right]. \end{aligned} \quad (139)$$

Lemma 3:

$$(1 + u) \ln(1 + u) \geq \begin{cases} u + \frac{u^2}{2}, & u \in [-1, 0] \\ u + \frac{u^2}{2} - \frac{u^3}{6}, & u \geq 0 \end{cases} \quad (140)$$

where at $u = -1$, the left-hand side is defined to be zero (it is the limit of this function when $u \rightarrow -1$ from above).

Proof: The proof follows by elementary calculus. ■

Since $\delta_{i,j} \in [0, 1]$, then (139) and Lemma 3 imply that

$$D\left(\frac{\delta_{i,j} + \gamma_i}{1 + \gamma_i} \parallel \frac{\gamma_i}{1 + \gamma_i}\right) \geq \frac{\delta_{i,j}^2}{2\gamma_i} - \frac{\delta_{i,j}^3}{6\gamma_i^2(1 + \gamma_i)}. \quad (141)$$

Hence, by comparing (136) with the combination of (137) and (141), then it follows that (up to a second-order approximation) the lower bounds on the exponents that were derived via Theorem 2 are improved by at least a factor of $(\max_i \gamma_i)^{-1}$ as compared to those that follow from Azuma's inequality.

Example 4: Consider two probability measures P_1 and P_2 where

$$P_1(0) = P_2(1) = 0.4, \quad P_1(1) = P_2(0) = 0.6,$$

and the case of a single threshold of the log-likelihood ratio that is set to zero (i.e., $\lambda = 0$). The exact error exponent in this case is Chernoff information that is equal to

$$C(P_1, P_2) = 2.04 \cdot 10^{-2}.$$

The improved lower bound on the error exponent in (130) and (131) is equal to $1.77 \cdot 10^{-2}$, whereas the loosened lower bound in (138) is equal to $1.39 \cdot 10^{-2}$. In this case $\gamma_1 = \frac{2}{3}$ and $\gamma_2 = \frac{7}{9}$, so the improvement in the lower bound on the error exponent is indeed by a factor of approximately

$$\left(\max_i \gamma_i\right)^{-1} = \frac{9}{7}.$$

Note that, from (117), (125) and (132)–(135), these are lower bounds on the error exponents for any finite block length n , and not only asymptotically in the limit where $n \rightarrow \infty$. The operational meaning of this example is that the improved lower bound on the error exponent assures that a fixed error probability can be obtained based on a sequence of i.i.d. RVs whose length is reduced by 22.2% as compared to the loosened bound which follows from Azuma's inequality.

4) *Comparison of the Exact and Lower Bounds on the Error Exponents, Followed by a Relation to Fisher Information:* In the following, we compare the exact and lower bounds on the error exponents. Consider the case where there is a single threshold on the log-likelihood ratio (i.e., referring to the case where the erasure option is not provided) that is set to zero. The exact error exponent in this case is given by the Chernoff information (see (107)), and it will be compared to the two lower bounds on the error exponents that were derived in the previous two subsections.

Let $\{P_\theta\}_{\theta \in \Theta}$, denote an indexed family of probability mass functions where Θ denotes the parameter set. Assume that P_θ is differentiable in the parameter θ . Then, the Fisher information is defined as

$$J(\theta) \triangleq \mathbb{E}_\theta \left[\frac{\partial}{\partial \theta} \ln P_\theta(x) \right]^2 \quad (142)$$

where the expectation is w.r.t. the probability mass function P_θ . The divergence and Fisher information are two related information measures, satisfying the equality

$$\lim_{\theta' \rightarrow \theta} \frac{D(P_\theta || P_{\theta'})}{(\theta - \theta')^2} = \frac{J(\theta)}{2} \quad (143)$$

(note that if it was a relative entropy to base 2 then the right-hand side of (143) would have been divided by $\ln 2$, and be equal to $\frac{J(\theta)}{\ln 4}$ as in [15, Eq. (12.364)]).

Proposition 6: Under the above assumptions,

- The Chernoff information and Fisher information are related information measures that satisfy the equality

$$\lim_{\theta' \rightarrow \theta} \frac{C(P_\theta, P_{\theta'})}{(\theta - \theta')^2} = \frac{J(\theta)}{8}. \quad (144)$$

- Let

$$E_L(P_\theta, P_{\theta'}) \triangleq \min_{i=1,2} D\left(\frac{\delta_i + \gamma_i}{1 + \gamma_i} \parallel \frac{\gamma_i}{1 + \gamma_i}\right) \quad (145)$$

be the lower bound on the error exponent in (130) which corresponds to $P_1 \triangleq P_\theta$ and $P_2 \triangleq P_{\theta'}$, then also

$$\lim_{\theta' \rightarrow \theta} \frac{E_L(P_\theta, P_{\theta'})}{(\theta - \theta')^2} = \frac{J(\theta)}{8}. \quad (146)$$

- Let

$$\tilde{E}_L(P_\theta, P_{\theta'}) \triangleq \min_{i=1,2} \frac{\delta_i^2}{2} \quad (147)$$

be the loosened lower bound on the error exponent in (138) which refers to $P_1 \triangleq P_\theta$ and $P_2 \triangleq P_{\theta'}$. Then,

$$\lim_{\theta' \rightarrow \theta} \frac{\tilde{E}_L(P_\theta, P_{\theta'})}{(\theta - \theta')^2} = \frac{a(\theta) J(\theta)}{8} \quad (148)$$

for some deterministic function a bounded in $[0, 1]$, and there exists an indexed family of probability mass functions for which $a(\theta)$ can be made arbitrarily close to zero for any fixed value of $\theta \in \Theta$.

Proof: See Appendix K. ■

Proposition 6 shows that, in the considered setting, the refined lower bound on the error exponent provides the correct behavior of the error exponent for a binary hypothesis testing when the relative entropy between the pair of probability mass functions that characterize the two hypotheses tends to zero. This stays in contrast to the loosened error exponent, which follows from Azuma's inequality, whose scaling may differ significantly from the correct exponent (for a concrete example, see the last part of the proof in Appendix K).

Example 5: Consider the index family of of probability mass functions defined over the binary alphabet $\mathcal{X} = \{0, 1\}$:

$$P_\theta(0) = 1 - \theta, \quad P_\theta(1) = \theta, \quad \forall \theta \in (0, 1).$$

From (142), the Fisher information is equal to

$$J(\theta) = \frac{1}{\theta} + \frac{1}{1 - \theta}$$

and, at the point $\theta = 0.5$, $J(\theta) = 4$. Let $\theta_1 = 0.51$ and $\theta_2 = 0.49$, so from (144) and (146)

$$C(P_{\theta_1}, P_{\theta_2}), E_L(P_{\theta_1}, P_{\theta_2}) \approx \frac{J(\theta)(\theta_1 - \theta_2)^2}{8} = 2.00 \cdot 10^{-4}.$$

Indeed, the exact values of $C(P_{\theta_1}, P_{\theta_2})$ and $E_L(P_{\theta_1}, P_{\theta_2})$ are $2.000 \cdot 10^{-4}$ and $1.997 \cdot 10^{-4}$, respectively.

5) *Moderate Deviations Analysis for Binary Hypothesis Testing*: So far, we have discussed large deviations analysis for binary hypothesis testing, and compared the exact error exponents with lower bounds that follow from refined versions of Azuma's inequality.

Based on the asymptotic results in (90) and (91), which hold a.s. under hypotheses H_1 and H_2 respectively, the large deviations analysis refers to upper and lower thresholds $\bar{\lambda}$ and $\underline{\lambda}$ which are *kept fixed* (i.e., these thresholds do not depend on the block length n of the data sequence) where

$$-D(P_2||P_1) < \underline{\lambda} \leq \bar{\lambda} < D(P_1||P_2).$$

Suppose that instead of having some fixed upper and lower thresholds, one is interested to set these thresholds such that as the block length n tends to infinity, they tend to their asymptotic limits in (90) and (91), i.e.,

$$\lim_{n \rightarrow \infty} \bar{\lambda}^{(n)} = D(P_1||P_2), \quad \lim_{n \rightarrow \infty} \underline{\lambda}^{(n)} = -D(P_2||P_1).$$

Specifically, let $\eta \in (\frac{1}{2}, 1)$, and $\varepsilon_1, \varepsilon_2 > 0$ be arbitrary fixed numbers, and consider the case where one decides on hypothesis H_1 if

$$L(X_1, \dots, X_n) > n\bar{\lambda}^{(n)}$$

and on hypothesis H_2 if

$$L(X_1, \dots, X_n) < n\underline{\lambda}^{(n)}$$

where these upper and lower thresholds are set to

$$\begin{aligned} \bar{\lambda}^{(n)} &= D(P_1||P_2) - \varepsilon_1 n^{-(1-\eta)} \\ \underline{\lambda}^{(n)} &= -D(P_2||P_1) + \varepsilon_2 n^{-(1-\eta)} \end{aligned}$$

so that they approach, respectively, the relative entropies $D(P_1||P_2)$ and $-D(P_2||P_1)$ in the asymptotic case where the block length n of the data sequence tends to infinity. Accordingly, the conditional probabilities in (92)–(95) are modified so that the fixed thresholds $\bar{\lambda}$ and $\underline{\lambda}$ are replaced with the above block-length dependent thresholds $\bar{\lambda}^{(n)}$ and $\underline{\lambda}^{(n)}$, respectively. The moderate deviations analysis for binary hypothesis testing studies the probability of an error event and also the probability of the event of either making an erroneous decision or making no decision (i.e., declaring an erasure) under the two hypotheses. Particularly, we also study the asymptotic scaling of these probability under either H_1 and H_2 when simultaneously the block length of the input sequence n tends to infinity, and the thresholds $\bar{\lambda}^{(n)}$ and $\underline{\lambda}^{(n)}$ tend to $D(P_1||P_2)$ and $-D(P_2||P_1)$, respectively (which are the asymptotic limits in (90) and (91), respectively, when the block length tends to infinity).

Before proceeding to the moderate deviations analysis of binary hypothesis testing, the related literature in the context of information-theoretic problems is shortly reviewed. The moderate deviations analysis in the context of source and channel coding has recently attracted some interest among information theorists (see [1], [4], [24], [42] and [56]).

Moderate deviations were analyzed in [1, Section 4.3] for a channel model that gets noisier as the block length is increased. Due to the dependence of the channel parameter in the block length, the usual notion of capacity for these channels is zero. Hence, the issue of increasing the block length for the considered type of degrading channels was examined in [1, Section 4.3] via moderate deviations analysis when the number of codewords increases sub-exponentially with the block length. In another recent work [4], the moderate deviations behavior of channel coding for discrete memoryless channels was studied by Altug and Wagner with a derivation of direct and converse results which explicitly characterize the rate function of the moderate deviations principle (MDP). In [4], the authors studied the interplay between the probability of error, code rate and block length when the communication takes place over discrete memoryless channels, having the interest to figure out how the decoding error probability of the best code scales when simultaneously the block length tends to infinity and the code rate approaches the channel capacity. The novelty in the setup of their analysis was the consideration of the scenario mentioned above, in

contrast to the case where the rate is kept fixed below capacity, and the study is reduced to a characterization of the dependence between the two remaining parameters (i.e., the block length n and the average/ maximal error probability of the best code). As opposed to the latter case when the code rate is kept fixed, which then corresponds to large deviations analysis and characterizes the error exponents as a function of the rate, the analysis in [4] (via the introduction of direct and converse theorems) demonstrated a sub-exponential scaling of the maximal error probability in the considered moderate deviations regime. This work was followed by a work by Polynaskiy and Verdú where they show that a DMC satisfies the MDP if and only if its channel dispersion is non-zero, and also that the AWGN channel satisfies the MDP with a constant that is equal to the channel dispersion. The approach used in [4] was based on the method of types, whereas the approach used in [43] borrowed some tools from a recent work by the same authors in [42].

In [24], the moderate deviations analysis of the Slepian-Wolf problem for lossless source coding was studied. More recently, moderate deviations analysis for lossy source coding of stationary memoryless sources was studied in [56].

These works, including the following discussion, indicate a recent interest in moderate deviations analysis in the context of information-theoretic problems. In the literature on probability theory, the moderate deviations analysis was extensively studied (see, e.g., [18, Section 3.7]), and in particular the MDP was studied in [17] for continuous-time martingales with bounded jumps.

In light of the discussion in Section IV-F on the MDP for i.i.d. RVs and its relation to the concentration inequalities in Section III (see Appendix J), and also motivated by the recent works on moderate-deviations analysis for information-theoretic aspects, we consider in the following moderate deviations analysis for binary hypothesis testing. Our approach for this kind of analysis is different from [4] and [43], and it relies on concentration inequalities for martingales.

In the following, we analyze the probability of a joint error and erasure event under hypothesis H_1 , i.e., derive an upper bound on $\alpha_n^{(1)}$ in (92). The same kind of analysis can be adapted easily for the other probabilities in (93)–(95). As mentioned earlier, let $\varepsilon_1 > 0$ and $\eta \in (\frac{1}{2}, 1)$ be two arbitrarily fixed numbers. Then, under hypothesis H_1 , it follows that similarly to (116)–(118)

$$\begin{aligned} & P_1^n(L(X_1, \dots, X_n) \leq n\bar{\lambda}^{(n)}) \\ &= P_1^n(L(X_1, \dots, X_n) \leq nD(P_1||P_2) - \varepsilon_1 n^\eta) \\ &\leq \exp\left(-nD\left(\frac{\delta_1^{(\eta,n)} + \gamma_1}{1 + \gamma_1} \parallel \frac{\gamma_1}{1 + \gamma_1}\right)\right) \end{aligned} \quad (149)$$

where

$$\delta_1^{(\eta,n)} \triangleq \frac{\varepsilon_1 n^{-(1-\eta)}}{d_1}, \quad \gamma_1 \triangleq \frac{\sigma_1^2}{d_1^2} \quad (150)$$

with d_1 and σ_1^2 from (112) and (113). From (139), (140) and (150), it follows that

$$\begin{aligned} & D\left(\frac{\delta_1^{(\eta,n)} + \gamma_1}{1 + \gamma_1} \parallel \frac{\gamma_1}{1 + \gamma_1}\right) \\ &= \frac{\gamma_1}{1 + \gamma_1} \left[\left(1 + \frac{\delta_1^{(\eta,n)}}{\gamma_1}\right) \ln\left(1 + \frac{\delta_1^{(\eta,n)}}{\gamma_1}\right) + \frac{(1 - \delta_1^{(\eta,n)}) \ln(1 - \delta_1^{(\eta,n)})}{\gamma_1} \right] \\ &\geq \frac{\gamma_1}{1 + \gamma_1} \left[\left(\frac{\delta_1^{(\eta,n)}}{\gamma_1} + \frac{(\delta_1^{(\eta,n)})^2}{2\gamma_1^2} - \frac{(\delta_1^{(\eta,n)})^3}{6\gamma_1^3}\right) + \frac{1}{\gamma_1} \left(-\delta_1^{(\eta,n)} + \frac{(\delta_1^{(\eta,n)})^2}{2}\right) \right] \\ &= \frac{(\delta_1^{(\eta,n)})^2}{2\gamma_1} \left(1 - \frac{\delta_1^{(\eta,n)}}{3\gamma_1(1 + \gamma_1)}\right) \\ &= \frac{\varepsilon_1^2 n^{-2(1-\eta)}}{2\sigma_1^2} \left(1 - \frac{\varepsilon_1 d_1}{3\sigma_1^2(1 + \gamma_1)} \frac{1}{n^{1-\eta}}\right) \end{aligned}$$

provided that $\delta_1^{(\eta, n)} < 1$ (which holds for $n \geq n_0$ for some $n_0 \triangleq n_0(\eta, \varepsilon_1, d_1) \in \mathbb{N}$ that is determined from (150)). By substituting this lower bound on the divergence into (149), it follows that

$$\begin{aligned} \alpha_n^{(1)} &= P_1^n(L(X_1, \dots, X_n) \leq nD(P_1||P_2) - \varepsilon_1 n^\eta) \\ &\leq \exp\left(-\frac{\varepsilon_1^2 n^{2\eta-1}}{2\sigma_1^2} \left(1 - \frac{\varepsilon_1 d_1}{3\sigma_1^2(1+\gamma_1)} \frac{1}{n^{1-\eta}}\right)\right). \end{aligned} \quad (151)$$

Consequently, in the limit where n tends to infinity,

$$\lim_{n \rightarrow \infty} n^{1-2\eta} \ln \alpha_n^{(1)} \leq -\frac{\varepsilon_1^2}{2\sigma_1^2} \quad (152)$$

with σ_1^2 in (113). From the analysis in Section IV-F and Appendix J, the following things hold:

- The inequality for the asymptotic limit in (152) holds in fact with equality.
- The same asymptotic result also follows from Theorem 4 for every even-valued $m \geq 2$ (instead of Theorem 2).

To verify these statements, consider the real-valued sequence of i.i.d. RVs

$$Y_i \triangleq \ln\left(\frac{P_1(X_i)}{P_2(X_i)}\right) - D(P_1||P_2), \quad i = 1, \dots, n$$

that, under hypothesis H_1 , have zero mean and variance σ_1^2 . Since, by assumption, $\{X_i\}_{i=1}^n$ are i.i.d., then

$$L(X_1, \dots, X_n) - nD(P_1||P_2) = \sum_{i=1}^n Y_i, \quad (153)$$

and it follows from the one-sided version of the MDP in (82) that indeed (152) holds with equality. Moreover, Theorem 2 provides, via the inequality in (151), a finite-length result that enhances the asymptotic result for $n \rightarrow \infty$. The second item above follows from the second part of the analysis in Appendix J (i.e., the part of analysis in this appendix that follows from Theorem 4).

In the considered setting of moderate deviations analysis for binary hypothesis testing, the upper bound on the probability $\alpha_n^{(1)}$ in (151), which refers to the probability of either making an error or declaring an erasure (i.e., making no decision) under the hypothesis H_1 , decays to zero sub-exponentially with the length n of the sequence. As mentioned above, based on the analysis in Section IV-F and Appendix J, the asymptotic upper bound in (152) is tight. A completely similar moderate-deviations analysis can be also performed under the hypothesis H_2 . Hence, a sub-exponential scaling of the probability $\beta_n^{(1)}$ in (94) of either making an error or declaring an erasure (where the lower threshold $\underline{\lambda}$ is replaced with $\underline{\lambda}^{(n)}$) also holds under the hypothesis H_2 . These two sub-exponential decays to zero for the probabilities $\alpha_n^{(1)}$ and $\beta_n^{(1)}$, under hypothesis H_1 or H_2 respectively, improve as the value of $\eta \in (\frac{1}{2}, 1)$ is increased. On the other hand, the two *exponential decays* to zero of the probabilities of error (i.e., $\alpha_n^{(2)}$ and $\beta_n^{(2)}$ under hypothesis H_1 or H_2 , respectively) improve as the value of $\eta \in (\frac{1}{2}, 1)$ is decreased; this is due to the fact that, for a fixed value of n , the margin which serves to protect us from making an error (either under hypothesis H_1 or H_2) is increased by decreasing the value of η as above (note that by reducing the value of η for a fixed n , the upper and lower thresholds $\bar{\lambda}^{(n)}$ and $\underline{\lambda}^{(n)}$ are made closer to $D(P_1||P_2)$ from below and to $-D(P_2||P_1)$ from above, respectively, which therefore increases the margin that is used for protecting one from making an erroneous decision). This shows the existence of a tradeoff, in the choice of the parameter $\eta \in (\frac{1}{2}, 1)$, between the probability of error and the joint probability of error and erasure under either hypothesis H_1 or H_2 (where this tradeoff exists symmetrically for each of the two hypotheses).

In [4] and [43], the authors consider moderate deviations analysis for channel coding over memoryless channels. In particular, [4, Theorem 2.2] and [43, Theorem 6] indicate on a tight lower bound (i.e., a converse) to the asymptotic result in (152) for binary hypothesis testing. This tight converse is indeed consistent with the asymptotic result of the MDP in (82) for real-valued i.i.d. random variables, which implies that the asymptotic upper bound in (152), obtained via the martingale approach with the refined version of Azuma's inequality in Theorem 2, holds indeed with equality. Note that this equality does not follow from Azuma's inequality, so its refinement was essential for obtaining this equality. The reason is that, due to Appendix J, the upper bound in (152) that is equal to $-\frac{\varepsilon_1^2}{2\sigma_1^2}$ is replaced via Azuma's inequality by the looser bound $-\frac{\varepsilon_1^2}{2d_1^2}$ (note that, from (112) and (113), $\sigma_1 \leq d_1$ where σ_1 may be significantly smaller than d_1).

6) *Second-Order Analysis for Binary Hypothesis Testing*: The moderate deviations analysis in the previous sub-section refers to deviations that scale like n^η for $\eta \in (\frac{1}{2}, 1)$. Let us consider now the case of $\eta = \frac{1}{2}$ which corresponds to small deviations. To this end, refer to the real-valued sequence of i.i.d. RVs $\{Y_i\}_{i=1}^n$ with zero mean and variance σ_1^2 (under hypothesis H_1), and define the partial sums $S_k = \sum_{i=1}^k Y_i$ for $k \in \{1, \dots, n\}$ with $S_0 = 0$. This implies that $\{S_k, \mathcal{F}_k\}_{k=0}^n$ is a martingale sequence. At this point, it links the current discussion on binary hypothesis testing to Section IV-D which refers to the relation between the martingale CLT and Proposition 4. Specifically, since from (153),

$$S_n - S_0 = L(X_1, \dots, X_n) - nD(P_1||P_2)$$

then from the proof of Proposition 4, one gets an upper bound on the probability

$$P_1^n(L(X_1, \dots, X_n) \leq nD(P_1||P_2) - \varepsilon_1\sqrt{n})$$

for a finite block length n (via an analysis that is either related to Theorem 2 or 4) which agrees with the asymptotic result

$$\lim_{n \rightarrow \infty} \ln P_1^n(L(X_1, \dots, X_n) \leq nD(P_1||P_2) - \varepsilon_1\sqrt{n}) = -\frac{\varepsilon_1^2}{2\sigma_1^2}.$$

Referring to small deviations analysis and the CLT, it shows a duality between these kind of results and recent works on second-order analysis for channel coding (see [25], [42], and [44], where the variance σ_1^2 in (113) is replaced with the channel dispersion that is defined to be the variance of the mutual information RV between the channel input and output, and is a property of the communication channel solely).

B. Pairwise Error Probability for Linear Block Codes over Binary-Input Output-Symmetric DMCs

In this sub-section, the tightness of Theorems 2 and 4 is studied by the derivation of upper bounds on the pairwise error probability under maximum-likelihood (ML) decoding when the transmission takes place over a discrete memoryless channel (DMC).

Let \mathcal{C} be a binary linear block code of block length n , and assume that the codewords are a-priori equi-probable. Consider the case where the communication takes place over a binary-input output-symmetric DMC whose input alphabet is $\mathcal{X} = \{0, 1\}$, and its output alphabet \mathcal{Y} is finite.

In the following, boldface letters denote vectors, regular letters with sub-scripts denote individual elements of vectors, capital letters represent RVs, and lower-case letters denote individual realizations of the corresponding RVs. Let

$$P_{\mathbf{Y}|\mathbf{X}}(\underline{y}|\underline{x}) = \prod_{i=1}^n P_{Y_i|X_i}(y_i|x_i)$$

be the transition probability of the DMC, where due to the symmetry assumption

$$P_{Y|X}(y|0) = P_{Y|X}(-y|1), \quad \forall y \in \mathcal{Y}.$$

It is also assumed in the following that $P_{Y|X}(y|x) > 0$ for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Due to the linearity of the code and the symmetry of the DMC, the decoding error probability is independent of the transmitted codeword, so it is assumed without any loss of generality that the all-zero codeword is transmitted. In the following, we consider the pairwise error probability when the competitive codeword $\underline{x} \in \mathcal{C}$ has a Hamming weight that is equal to h , and denote it by $W_H(\underline{x}) = h$. Let $P_{\mathbf{Y}}$ denote the probability distribution of the channel output.

In order to derive upper bounds on the pairwise error probability, let us define the following two hypotheses:

- $H_1 : P_{\mathbf{Y}}(\underline{y}) = \prod_{i=1}^n P_{Y|X}(y_i|0), \quad \forall \underline{y} \in \mathcal{Y}^n,$
- $H_2 : P_{\mathbf{Y}}(\underline{y}) = \prod_{i=1}^n P_{Y|X}(y_i|x_i), \quad \forall \underline{y} \in \mathcal{Y}^n$

which correspond, respectively, to the transmission of the all-zero codeword and the competitive codeword $\underline{x} \in \mathcal{C}$.

Under hypothesis H_1 , the considered pairwise error event under ML decoding occurs if and only if

$$\sum_{i=1}^n \ln \left(\frac{P_{Y|X}(y_i|x_i)}{P_{Y|X}(y_i|0)} \right) \geq 0.$$

Let $\{i_k\}_{k=1}^h$ be the h indices of the coordinates of \underline{x} where $x_i = 1$, ordered such that $1 \leq i_1 < \dots < i_h \leq n$. Based on this notation, the log-likelihood ratio satisfies the equality

$$\sum_{i=1}^n \ln \left(\frac{P_{Y|X}(y_i|x_i)}{P_{Y|X}(y_i|0)} \right) = \sum_{m=1}^h \ln \left(\frac{P_{Y|X}(y_{i_m}|1)}{P_{Y|X}(y_{i_m}|0)} \right). \quad (154)$$

For the continuation of the analysis in this sub-section, let us define the martingale sequence $\{U_k, \mathcal{F}_k\}_{k=0}^n$ with the filtration

$$\begin{aligned} \mathcal{F}_k &= \sigma(Y_{i_1}, \dots, Y_{i_k}), \quad k = 1, \dots, h \\ \mathcal{F}_0 &= \{\emptyset, \Omega\} \end{aligned}$$

and, under hypothesis H_1 , let

$$U_k = \mathbb{E} \left[\sum_{m=1}^h \ln \left(\frac{P_{Y|X}(Y_{i_m}|1)}{P_{Y|X}(Y_{i_m}|0)} \right) \mid \mathcal{F}_k \right], \quad \forall k \in \{0, 1, \dots, h\}.$$

Since, under hypothesis H_1 , the RVs Y_{i_1}, \dots, Y_{i_h} are statistically independent, then for $k \in \{0, 1, \dots, h\}$

$$\begin{aligned} U_k &= \sum_{m=1}^k \ln \left(\frac{P_{Y|X}(Y_{i_m}|1)}{P_{Y|X}(Y_{i_m}|0)} \right) \\ &\quad + (h-k) \sum_{y \in \mathcal{Y}} P_{Y|X}(y|0) \ln \left(\frac{P_{Y|X}(y|1)}{P_{Y|X}(y|0)} \right) \\ &= \sum_{m=1}^k \ln \left(\frac{P_{Y|X}(Y_{i_m}|1)}{P_{Y|X}(Y_{i_m}|0)} \right) \\ &\quad - (h-k) D(P_{Y|X}(\cdot|0) \parallel P_{Y|X}(\cdot|1)). \end{aligned} \quad (155)$$

Specifically

$$U_0 = -h D(P_{Y|X}(\cdot|0) \parallel P_{Y|X}(\cdot|1)) \quad (156)$$

$$U_h = \sum_{i=1}^n \ln \left(\frac{P_{Y|X}(Y_i|x_i)}{P_{Y|X}(Y_i|0)} \right) \quad (157)$$

where the last equality follows from (154) and (155), and the differences of the martingale sequence are given by

$$\begin{aligned} \xi_k &\triangleq U_k - U_{k-1} \\ &= \ln \left(\frac{P_{Y|X}(Y_{i_k}|1)}{P_{Y|X}(Y_{i_k}|0)} \right) + D(P_{Y|X}(\cdot|0) \parallel P_{Y|X}(\cdot|1)) \end{aligned} \quad (158)$$

for every $k \in \{1, \dots, h\}$. Note that, under hypothesis H_1 , indeed $\mathbb{E}[\xi_k | \mathcal{F}_{k-1}] = 0$.

The probability of a pairwise error event, where the ML decoder prefers a competitive codeword $\underline{x} \in \mathcal{C}$ ($W_H(\underline{x}) = h$) over the transmitted all-zero codeword, is equal to

$$\begin{aligned} P_h &\triangleq \mathbb{P}(U_h > 0 \mid H_1) \\ &= \mathbb{P}(U_h - U_0 > h D(P_{Y|X}(\cdot|0) \parallel P_{Y|X}(\cdot|1)) \mid H_1). \end{aligned} \quad (159)$$

It therefore follows that a.s. for every $k \in \{1, \dots, h\}$

$$\begin{aligned} |\xi_k| &\leq \max_{y \in \mathcal{Y}} \left| \ln \left(\frac{P_{Y|X}(y|1)}{P_{Y|X}(y|0)} \right) \right| + D(P_{Y|X}(\cdot|0) \parallel P_{Y|X}(\cdot|1)) \\ &\triangleq d < \infty \end{aligned} \quad (160)$$

which is indeed finite since, by assumption, the alphabet \mathcal{Y} is finite and $P_{Y|X}(y|x) > 0$ for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Note that, in fact, taking an absolute value in the maximization of the logarithm on the right-hand side of (160) is

redundant due to the channel symmetry, and also due to the equality $\sum_y P_{Y|X}(y|0) = \sum_y P_{Y|X}(y|1) = 1$ (so that it follows, from this equality, that there exists an element $y \in \mathcal{Y}$ such that $P_{Y|X}(y|1) \geq P_{Y|X}(y|0)$).

As an interim conclusion, $\{U_k, \mathcal{F}_k\}_{k=0}^h$ is a martingale sequence with bounded jumps, and $|U_k - U_{k-1}| \leq d$ holds a.s. for every $k \in \{1, \dots, h\}$. We rely in the following on the concentration inequalities of Theorems 2 and 4 to obtain, via (158)–(160), upper bounds on the pairwise error probability. The tightness of these bounds will be examined numerically, and they will be compared to the Bhattacharyya upper bound.

1) *Analysis Related to Theorem 2:* From (158), for every $k \in \{1, \dots, h\}$

$$\begin{aligned} & \mathbb{E}[\xi_k^2 | \mathcal{F}_{k-1}] \\ &= \sum_{y \in \mathcal{Y}} P_{Y|X}(y|0) \left[\ln \left(\frac{P_{Y|X}(y|1)}{P_{Y|X}(y|0)} \right) + D(P_{Y|X}(\cdot|0) \| P_{Y|X}(\cdot|1)) \right]^2 \\ &= \sum_{y \in \mathcal{Y}} P_{Y|X}(y|0) \left[\ln \left(\frac{P_{Y|X}(y|1)}{P_{Y|X}(y|0)} \right) \right]^2 \\ &\quad - \left[D(P_{Y|X}(\cdot|0) \| P_{Y|X}(\cdot|1)) \right]^2 \triangleq \sigma^2 \end{aligned} \tag{161}$$

holds a.s., where the last equality follows from the definition of the divergence (relative entropy). Based on (159) and the notation in (11), let

$$\gamma = \frac{\sigma^2}{d^2}, \quad \delta \triangleq \frac{D(P_{Y|X}(\cdot|0) \| P_{Y|X}(\cdot|1))}{d} \tag{162}$$

where d and σ^2 are introduced in (160) and (161), respectively. Under hypothesis H_1 , one gets from (159) and Theorem 2 that the pairwise error probability satisfies the upper bound

$$P_h \leq Z_1^h \tag{163}$$

where

$$Z_1 \triangleq \exp \left(-D \left(\frac{\delta + \gamma}{1 + \gamma} \parallel \frac{\gamma}{1 + \gamma} \right) \right) \tag{164}$$

and γ, δ are introduced in (162).

In the following, we compare the exponential bound in (163) with the Bhattacharyya bound

$$P_h \leq Z_B^h \tag{165}$$

where the Bhattacharyya parameter Z_B of the binary-input DMC is given by

$$Z_B \triangleq \sum_{y \in \mathcal{Y}} \sqrt{P_{Y|X}(y|0) P_{Y|X}(y|1)}. \tag{166}$$

Example 6: Consider a binary symmetric channel (BSC) with crossover probability p . The Bhattacharyya parameter which corresponds to this channel is $Z_B = \sqrt{4p(1-p)}$. In the following, Z_1 from (164) is calculated for comparison. Without loss of generality, assume that $p \leq \frac{1}{2}$. Straightforward calculation shows that

$$\begin{aligned} d &= 2(1-p) \ln \left(\frac{1-p}{p} \right) \\ \sigma^2 &= 4p(1-p) \left[\ln \left(\frac{1-p}{p} \right) \right]^2 \\ D(P_{Y|X}(\cdot|0) \| P_{Y|X}(\cdot|1)) &= (1-2p) \ln \left(\frac{1-p}{p} \right) \end{aligned}$$

and therefore (162) gives that

$$\gamma = \frac{p}{1-p}, \quad \delta = \frac{1-2p}{2(1-p)}.$$

Substituting γ and δ into (164) gives that the base of the exponential bound in (163) is equal to

$$Z_1 = \exp \left(-D \left(\frac{1}{2} \parallel p \right) \right) = \sqrt{4p(1-p)}$$

which coincides with the Bhattacharyya parameter for the BSC. This shows that, for the BSC, Theorem 2 implies the Bhattacharyya upper bound on the pairwise error probability.

In general, it is observed numerically that $Z_1 \geq Z_B$ for binary-input output-symmetric DMCs with an equality for the BSC (this will be exemplified after introducing the bound on the pairwise error probability which follows from Theorem 4). This implies that Theorem 2 yields in general a looser bound than the Bhattacharyya upper bound in the context of the pairwise error probability for DMCs.

2) *Analysis Related to Theorem 4:* In the following, a parallel upper bound on the pairwise error probability is derived from Remark 12 on Theorem 4, and the martingale sequence $\{U_k, \mathcal{F}_k\}_{k=0}^h$. Under hypothesis H_1 (i.e., the assumption that the all-zero codeword is transmitted), (158) implies that the conditional expectation of $(U_k - U_{k-1})^l$ given \mathcal{F}_{k-1} is equal (a.s.) to the un-conditional expectation where l is an arbitrary natural number. Also, it follows from (158) that for every $k \in \{1, \dots, h\}$ and $l \in \mathbb{N}$

$$\begin{aligned} & \mathbb{E}[(U_k - U_{k-1})^l | \mathcal{F}_{k-1}] \\ &= (-1)^l \mathbb{E} \left[\left(\ln \left(\frac{P_{Y|X}(Y|0)}{P_{Y|X}(Y|1)} \right) - D(P_{Y|X}(\cdot|0) \| P_{Y|X}(\cdot|1)) \right)^l \right] \end{aligned}$$

and, from the requirement that the sequence $\{\mu_l\}$ is non-negative, let (based on Remark 12)

$$\mu_l \triangleq \max \left\{ 0, (-1)^l \mathbb{E} \left[\left(\ln \left(\frac{P_{Y|X}(Y|0)}{P_{Y|X}(Y|1)} \right) - D(P_{Y|X}(\cdot|0) \| P_{Y|X}(\cdot|1)) \right)^l \right] \right\}. \quad (167)$$

for every $l \in \mathbb{N}$ (for even-valued l , there is no need to take the maximization with zero). Based on the notation used in the context of Remark 12, let

$$\gamma_l \triangleq \frac{\mu_l}{d^l}, \quad l = 2, 3, \dots$$

and δ be the same parameter as in (162). Note that the equality $\gamma_2 = \gamma$ holds for the parameter γ in (162). Then, Remark 12 on Theorem 4 yields that for every even-valued $m \geq 2$

$$P_h \leq (Z_2^{(m)})^h \quad (168)$$

where

$$Z_2^{(m)} \triangleq \inf_{x \geq 0} \left\{ e^{-\delta x} \left[1 + \sum_{l=2}^{m-1} \frac{(\gamma_l - \gamma_m)x^l}{l!} + \gamma_m(e^x - 1 - x) \right] \right\}.$$

Example 7: In the following example, the bases of the two exponential bounds on the pairwise error probability in (163) and (168) are compared to the corresponding Bhattacharyya parameter (see (166)) for some binary-input output-symmetric DMCs.

For a integer-valued $Q \geq 2$, let $P_{Y|X}^{(Q)}$ be a binary-input output-symmetric DMC with input alphabet $\mathcal{X} = \{0, 1\}$ and output alphabet $\mathcal{Y} = \{0, 1, \dots, Q-1\}$, characterized by the following probability transitions:

$$\begin{aligned} P_{Y|X}^{(Q)}(0|0) &= P_{Y|X}^{(Q)}(Q-1|1) = 1 - (Q-1)p, \\ P_{Y|X}^{(Q)}(1|0) &= \dots = P_{Y|X}^{(Q)}(Q-1|0) = p \\ P_{Y|X}^{(Q)}(0|1) &= \dots = P_{Y|X}^{(Q)}(Q-2|1) = p \end{aligned} \quad (169)$$

where $0 < p < \frac{1}{Q-1}$. The considered exponential bounds are exemplified in the following for the case where $p = 0.04$ and $Q = 2, 3, 4, 5, 10$. The bases of the exponential bounds in (163) and (168) are compared in Table I to the corresponding Bhattacharyya parameters of these five DMCs that, from (166), is equal to

$$Z_B = 2\sqrt{p[1 - (Q-1)p]} + (Q-2)p.$$

TABLE I

THE BASES OF THE EXPONENTIAL BOUNDS Z_1 AND $Z_2^{(m)}$ IN (163) AND (168) (FOR AN EVEN-VALUED $m \geq 2$), RESPECTIVELY. THE BASES OF THESE EXPONENTIAL BOUNDS ARE COMPARED TO THE BHATTACHARYYA PARAMETER Z_B IN (166) FOR THE FIVE DMC CHANNELS IN (169) WITH $p = 0.04$ AND $|\mathcal{Y}| = Q = 2, 3, 4, 5, 10$.

Q	2	3	4	5	10
Z_B	0.3919	0.4237	0.4552	0.4866	0.6400
Z_1	0.3919	0.4424	0.4879	0.5297	0.7012
$Z_2^{(2)}$	0.3967	0.4484	0.4950	0.5377	0.7102
$Z_2^{(4)}$	0.3919	0.4247	0.4570	0.4877	0.6421
$Z_2^{(6)}$	0.3919	0.4237	0.4553	0.4867	0.6400
$Z_2^{(8)}$	0.3919	0.4237	0.4552	0.4866	0.6400
$Z_2^{(10)}$	0.3919	0.4237	0.4552	0.4866	0.6400

TABLE II

THE BASE $\tilde{Z}_2^{(m)}$ OF THE EXPONENTIAL BOUND IN (163) AND ITS (TIGHT) UPPER BOUND $\tilde{Z}_2^{(m)}$ THAT FOLLOWS BY REPLACING THE INFIMUM OPERATION BY THE SUB-OPTIMAL VALUE IN (49) AND (50). THE FIVE DMCs ARE THE SAME AS IN (169) AND TABLE I.

Q	2	3	4	5	10
$Z_2^{(10)}$	0.3919	0.4237	0.4552	0.4866	0.6400
$\tilde{Z}_2^{(10)}$	0.3919	0.4237	0.4553	0.4868	0.6417

As is shown in Table I, the choice of $m = 2$ gives the worst upper bound in Theorem 4 (since $Z_2^{(2)} \geq Z_2^{(m)}$ for every even-valued $m \geq 2$). This is consistent with Corollary 3. Moreover, the comparison of the third and fourth lines in Theorem 4 is consistent with Proposition 2 which indeed assures that Theorem 4 with $m = 2$ is looser than Theorem 2 (hence, indeed $Z_1 < Z_2^{(2)}$ for the considered DMCs). Also, from Example 6, it follows that Theorem 2 coincides with the Bhattacharyya bound (hence, $Z_1 = Z_B$ for the special case where $Q = 2$, as is indeed verified numerically in Table I). It is interesting to realize from Table I that for the five considered DMCs, the sequence $\{Z_2^{(2)}, Z_2^{(4)}, Z_2^{(6)}, \dots\}$ converges very fast, and the limit is equal to the Bhattacharyya parameter for all the examined cases. This stays in contrast to the exponential base Z_1 that was derived from Theorem 2, and which appears to be strictly larger than the corresponding Bhattacharyya parameter of the DMC (except for the BSC, where the equality $Z_1 = Z_B$ holds, as is shown in Example 6).

Example 7 leads to the following conjecture:

Conjecture 1: For the martingale sequence $\{U_k, \mathcal{F}_k\}_{k=0}^h$ introduced in this sub-section,

$$\lim_{m \rightarrow \infty} Z_2^{(m)} = Z_B$$

and this convergence is quadratic.

Example 8: The base $Z_2^{(m)}$ of the exponential bound in (168) involves an operation of taking an infimum over the interval $[0, \infty)$. This operation is performed numerically in general, except for the special case where $m = 2$ for which a closed-form solution exists (see Appendix C for the proof of Corollary 4).

Replacing the infimum over $x \in [0, \infty)$ with the sub-optimal value of x in (49) and (50) gives an upper bound on the respective exponential base of the bound (note that due to the analysis, this sub-optimal value turns to be optimal in the special case where $m = 2$). The upper bound on $Z_2^{(m)}$ which follows by replacing the infimum with the sub-optimal value in (49) and (50) is denoted by $\tilde{Z}_2^{(m)}$, and the difference between the two values is marginal (see Table II).

C. Concentration of the Crest-Factor for OFDM Signals

Orthogonal-frequency-division-multiplexing (OFDM) is a modulation that converts a high-rate data stream into a number of low-rate streams that are transmitted over parallel narrow-band channels. OFDM is widely used in several international standards for digital audio and video broadcasting, and for wireless local area networks. For

a textbook providing a survey on OFDM, see e.g. [40, Chapter 19]. One of the problems of OFDM signals is that the peak amplitude of the signal can be significantly higher than the average amplitude. This issue makes the transmission of OFDM signals sensitive to non-linear devices in the communication path such as digital to analog converters, mixers and high-power amplifiers. As a result of this drawback, it increases the symbol error rate and it also reduces the power efficiency of OFDM signals as compared to single-carrier systems. Commonly, the impact of nonlinearities is described by the distribution of the crest-factor (CF) of the transmitted signal [34], but its calculation involves time-consuming simulations even for a small number of sub-carriers. The expected value of the CF for OFDM signals is known to scale like the logarithm of the number of sub-carriers of the OFDM signal (see [34], [48, Section 4] and [58]).

Given an n -length codeword $\{X_i\}_{i=0}^{n-1}$, a single OFDM baseband symbol is described by

$$s(t) = \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} X_i \exp\left(\frac{j 2\pi i t}{T}\right), \quad 0 \leq t \leq T. \quad (170)$$

Lets assume that X_0, \dots, X_{n-1} are complex RVs, and that a.s. $|X_i| = 1$ (these RVs should not be necessarily independent). Since the sub-carriers are orthonormal over $[0, T]$, then the signal power over the interval $[0, T]$ is 1 a.s., i.e.,

$$\frac{1}{T} \int_0^T |s(t)|^2 dt = 1. \quad (171)$$

The CF of the signal s , composed of n sub-carriers, is defined as

$$\text{CF}_n(s) \triangleq \max_{0 \leq t \leq T} |s(t)|. \quad (172)$$

From [48, Section 4] and [58], it follows that the CF scales with high probability like $\sqrt{\ln n}$ for large n . In [34, Theorem 3 and Corollary 5], a concentration inequality was derived for the CF of OFDM signals. It states that for an arbitrary $c \geq 2.5$

$$\mathbb{P}\left(\left|\text{CF}_n(s) - \sqrt{\ln n}\right| < \frac{c \ln \ln n}{\sqrt{\ln n}}\right) = 1 - O\left(\frac{1}{(\ln n)^4}\right).$$

Remark 17: The analysis used to derive this rather strong concentration inequality (see [34, Appendix C]) requires some assumptions on the distribution of the X_i 's (see the two conditions in [34, Theorem 3] followed by [34, Corollary 5]). These requirements are not needed in the following analysis, and the derivation of concentration inequalities that are introduced in this subsection are much more simple and provide some insight to the problem, though they lead to weaker concentration result than in [34, Theorem 3].

In the following, Azuma's inequality and a refined version of this inequality are considered under the assumption that $\{X_j\}_{j=0}^{n-1}$ are independent complex-valued random variables with magnitude 1, attaining the M points of an M -ary PSK constellation with equal probability.

1) *Establishing Concentration of the Crest-Factor via Azuma's Inequality:* In the following, Azuma's inequality is used to derive a concentration result. Let us define

$$Y_i = \mathbb{E}[\text{CF}_n(s) | X_0, \dots, X_{i-1}], \quad i = 0, \dots, n \quad (173)$$

Based on a standard construction of martingales, $\{Y_i, \mathcal{F}_i\}_{i=0}^n$ is a martingale where \mathcal{F}_i is the σ -algebra that is generated by the first i symbols (X_0, \dots, X_{i-1}) in (170). Hence, $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \dots \subseteq \mathcal{F}_n$ is a filtration. This martingale has also bounded jumps, and

$$|Y_i - Y_{i-1}| \leq \frac{2}{\sqrt{n}}$$

for $i \in \{1, \dots, n\}$ since revealing the additional i -th coordinate X_i affects the CF, as is defined in (172), by at most $\frac{2}{\sqrt{n}}$ (see the first part of Appendix L). It therefore follows from Azuma's inequality that, for every $\alpha > 0$,

$$\mathbb{P}(|\text{CF}_n(s) - \mathbb{E}[\text{CF}_n(s)]| \geq \alpha) \leq 2 \exp\left(-\frac{\alpha^2}{8}\right) \quad (174)$$

which demonstrates concentration around the expected value.

2) *Establishing Concentration of the Crest-Factor via the Refined Version of Azuma's Inequality in Proposition 4:* In the following, we rely on Proposition 4 to derive an improved concentration result. For the martingale sequence $\{Y_i\}_{i=0}^n$ in (173), Appendix L gives that a.s.

$$|Y_i - Y_{i-1}| \leq \frac{2}{\sqrt{n}}, \quad \mathbb{E}[(Y_i - Y_{i-1})^2 | \mathcal{F}_{i-1}] \leq \frac{2}{n} \quad (175)$$

for every $i \in \{1, \dots, n\}$. Note that the conditioning on the σ -algebra \mathcal{F}_{i-1} is equivalent to the conditioning on the symbols X_0, \dots, X_{i-2} , and there is no conditioning for $i = 1$. Further, let $Z_i = \sqrt{n}Y_i$ for $0 \leq i \leq n$. Proposition 4 therefore implies that for an arbitrary $\alpha > 0$

$$\begin{aligned} & \mathbb{P}(|\text{CF}_n(s) - \mathbb{E}[\text{CF}_n(s)]| \geq \alpha) \\ &= \mathbb{P}(|Y_n - Y_0| \geq \alpha) \\ &= \mathbb{P}(|Z_n - Z_0| \geq \alpha\sqrt{n}) \\ &\leq 2 \exp\left(-\frac{\alpha^2}{4} \left(1 + O\left(\frac{1}{\sqrt{n}}\right)\right)\right) \end{aligned} \quad (176)$$

(since $\delta = \frac{\alpha}{2}$ and $\gamma = \frac{1}{2}$ in the setting of Proposition 4). Note that the exponent in the last inequality is doubled as compared to the bound that was obtained in (174) via Azuma's inequality, and the term which scales like $O\left(\frac{1}{\sqrt{n}}\right)$ on the right-hand side of (176) is expressed explicitly for finite n (see Appendix H).

3) *A Concentration Inequality via Talagrand's Method:* In his seminal paper [54], Talagrand introduced an approach for proving concentration inequalities in product spaces. It forms a powerful probabilistic tool for establishing concentration results for coordinate-wise Lipschitz functions of independent random variables (see, e.g., [18, Section 2.4.2], [38, Section 4] and [54]). This approach is used in the following to derive a concentration result of the crest factor around its median, and it also enables to derive an upper bound on the distance between the median and the expected value. We provide in the following definitions that will be required for introducing a special form of Talagrand's inequalities. Afterwards, this inequality will be applied to obtain a concentration result for the crest factor of OFDM signals.

Definition 2 (Hamming distance): Let \mathbf{x}, \mathbf{y} be two n -length vectors. The Hamming distance between \mathbf{x} and \mathbf{y} is the number of coordinates where \mathbf{x} and \mathbf{y} disagree, i.e.,

$$d_H(\mathbf{x}, \mathbf{y}) \triangleq \sum_{i=1}^n I_{\{x_i \neq y_i\}}$$

where I stands for the indicator function.

The following suggests a generalization and normalization of the previous distance metric.

Definition 3: Let $\mathbf{a} = (a_1, \dots, a_n) \in \mathbb{R}_+^n$ (i.e., \mathbf{a} is a non-negative vector) satisfy $\|\mathbf{a}\|^2 = \sum_{i=1}^n (a_i)^2 = 1$. Then, define

$$d_a(\mathbf{x}, \mathbf{y}) \triangleq \sum_{i=1}^n a_i I_{\{x_i \neq y_i\}}.$$

Hence, $d_H(\mathbf{x}, \mathbf{y}) = \sqrt{n} d_a(\mathbf{x}, \mathbf{y})$ for $\mathbf{a} = \left(\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}}\right)$.

The following is a special form of Talagrand's inequalities ([38, Chapter 4], [54], [55]).

Theorem 7 (Talagrand's inequality): Let the random vector $\mathbf{X} = (X_1, \dots, X_n)$ be a vector of independent random variables with X_k taking values in a set A_k , and let $A \triangleq \prod_{k=1}^n A_k$. Let $f : A \rightarrow \mathbb{R}$ satisfy the condition that, for every $\mathbf{x} \in A$, there exists a non-negative, normalized n -length vector $\mathbf{a} = \mathbf{a}(\mathbf{x})$ such that

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \sigma d_a(\mathbf{x}, \mathbf{y}), \quad \forall \mathbf{y} \in A \quad (177)$$

for some fixed value $\sigma > 0$. Then, for every $\alpha \geq 0$,

$$\mathbb{P}(|f(X) - m| \geq \alpha) \leq 4 \exp\left(-\frac{\alpha^2}{4\sigma^2}\right) \quad (178)$$

where m is the median of $f(X)$ (i.e., $\mathbb{P}(f(X) \leq m) \geq \frac{1}{2}$ and $\mathbb{P}(f(X) \geq m) \geq \frac{1}{2}$). The same conclusion in (178) holds if the condition in (177) is replaced by

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \sigma d_a(\mathbf{x}, \mathbf{y}), \quad \forall \mathbf{y} \in A. \quad (179)$$

At this stage, we are ready to apply Talagrand's inequality to prove a concentration inequality for the crest factor of OFDM signals. As before, let us assume that $X_0, Y_0, \dots, X_{n-1}, Y_{n-1}$ are i.i.d. bounded complex RVs, and also assume for simplicity that $|X_i| = |Y_i| = 1$. In order to apply Talagrand's inequality to prove concentration, note that

$$\begin{aligned}
& \left| \max_{0 \leq t \leq T} |s(t; X_0, \dots, X_{n-1})| - \max_{0 \leq t \leq T} |s(t; Y_0, \dots, Y_{n-1})| \right| \\
& \leq \max_{0 \leq t \leq T} |s(t; X_0, \dots, X_{n-1}) - s(t; Y_0, \dots, Y_{n-1})| \\
& \leq \frac{1}{\sqrt{n}} \left| \sum_{i=0}^{n-1} (X_i - Y_i) \exp\left(\frac{j 2\pi i t}{T}\right) \right| \\
& \leq \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} |X_i - Y_i| \\
& \leq \frac{2}{\sqrt{n}} \sum_{i=0}^{n-1} I_{\{x_i \neq y_i\}} \\
& = 2d_a(X, Y)
\end{aligned}$$

where

$$a \triangleq \left(\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}}\right) \quad (180)$$

is a non-negative unit-vector of length n (note that a in this case is independent of x). Hence, Talagrand's inequality in Theorem 7 implies that, for every $\alpha \geq 0$,

$$\mathbb{P}(|\text{CF}_n(s) - m_n| \geq \alpha) \leq 4 \exp\left(-\frac{\alpha^2}{16}\right) \quad (181)$$

where m_n is the median of the crest factor for OFDM signals that are composed of n sub-carriers. This inequality demonstrates the concentration of this measure around its median. As a simple consequence of (181), one obtains the following result.

Corollary 8: The median and expected value of the crest factor differ by at most a constant, independently of the number of sub-carriers n .

Proof: By the concentration inequality in (181)

$$\begin{aligned}
|\mathbb{E}[\text{CF}_n(s)] - m_n| & \leq \mathbb{E}|\text{CF}_n(s) - m_n| \\
& = \int_0^\infty \mathbb{P}(|\text{CF}_n(s) - m_n| \geq \alpha) d\alpha \\
& \leq \int_0^\infty 4 \exp\left(-\frac{\alpha^2}{16}\right) d\alpha \\
& = 8\sqrt{\pi}.
\end{aligned}$$

■

Remark 18: This result applies in general to an arbitrary function f satisfying the condition in (177), where Talagrand's inequality in (178) implies that (see, e.g., [38, Lemma 4.6])

$$|\mathbb{E}[f(X)] - m| \leq 4\sigma\sqrt{\pi}.$$

4) *Establishing Concentration via McDiarmid's Inequality:* For the sake of self-completeness, we first state McDiarmid's inequality (see [38, Theorem 3.1]).

Theorem 8: Let $\mathbf{X} = (X_1, \dots, X_n)$ be a vector of independent random variables with X_k taking values in a set A_k for each k . Suppose that a real-valued function f , defined on $\prod_k A_k$, satisfies $|f(\mathbf{x}) - f(\mathbf{x}')| \leq c_k \in \mathbb{R}^+$ whenever the vectors \mathbf{x} and \mathbf{x}' differ only in the k -th coordinate. Let $\mu \triangleq \mathbb{E}[f(X)]$ be the expected value of $f(X)$. Then,

$$\mathbb{P}(|f(X) - \mu| \geq \alpha) \leq 2 \exp\left(-\frac{2\alpha^2}{\sum_k c_k^2}\right), \quad \forall \alpha \geq 0.$$

For a proof via martingales, see [38, Section 3.5].

Remark 19: In the special case where the condition for the function f in Theorem 7 (Talagrand's inequality) is satisfied with the additional property that the vector a on the right-hand side of (177) is *independent* of x (i.e., the value of this vector is fixed), then the concentration inequality in (178) follows from McDiarmid's inequality. To verify this observation, the reader is referred to [38, Theorem 3.6] followed by the discussion in [38, p. 211] (leading to [38, Eqs. (3.12) and (3.13)]).

In the following, McDiarmid's inequality is applied to prove a concentration inequality for the crest factor of OFDM signals. To this end, let us define

$$U \triangleq \max_{0 \leq t \leq T} |s(t; X_0, \dots, X_{i-1}, X_i, \dots, X_{n-1})|$$

$$V \triangleq \max_{0 \leq t \leq T} |s(t; X_0, \dots, X'_{i-1}, X_i, \dots, X_{n-1})|$$

where the two vectors $(X_0, \dots, X_{i-1}, X_i, \dots, X_{n-1})$ and $(X_0, \dots, X'_{i-1}, X_i, \dots, X_{n-1})$ may only differ in their i -th coordinate. This then implies that

$$\begin{aligned} |U - V| &\leq \max_{0 \leq t \leq T} |s(t; X_0, \dots, X_{i-1}, X_i, \dots, X_{n-1}) \\ &\quad - s(t; X_0, \dots, X'_{i-1}, X_i, \dots, X_{n-1})| \\ &= \max_{0 \leq t \leq T} \frac{1}{\sqrt{n}} \left| (X_{i-1} - X'_{i-1}) \exp\left(\frac{j 2\pi i t}{T}\right) \right| \\ &= \frac{|X_{i-1} - X'_{i-1}|}{\sqrt{n}} \leq \frac{2}{\sqrt{n}} \end{aligned}$$

where the last inequality holds since $|X_{i-1}| = |X'_{i-1}| = 1$. Hence, McDiarmid's inequality in Theorem 8 implies that, for every $\alpha \geq 0$,

$$\mathbb{P}(|\text{CF}_n(s) - \mathbb{E}[\text{CF}_n(s)]| \geq \alpha) \leq 2 \exp\left(-\frac{\alpha^2}{2}\right) \quad (182)$$

which demonstrates concentration of this measure around its expected value. By comparing (181) with (182), it follows that McDiarmid's inequality provides an improvement in the exponent. This is consistent with Remark 19 and the fixed value of the non-negative normalized vector in (180). The improvement of McDiarmid's inequality is by a factor of 4 in the exponent as compared to Azuma's inequality, and by a factor of 2 as compared to the refined version of Azuma's inequality in Proposition 4.

5) *Summary:* This subsection derives four concentration inequalities for the crest-factor (CF) of OFDM signals under the assumption that the symbols are independent. The first two concentration inequalities rely on Azuma's inequality and a refined version of it, and the last two concentration inequalities are based on Talagrand's and McDiarmid's inequalities. Although these concentration results are weaker than some existing results from the literature (see [34] and [58]), they establish concentration in a rather simple way and provide some insight to the problem. The use of these bounding techniques, in the context of concentration for OFDM signals, seems to be new. McDiarmid's inequality improves the exponent of Azuma's inequality by a factor of 4, and the exponent of the refined version of Azuma's inequality from Proposition 4 by a factor of 2. Note however that Proposition 4 may be in general tighter than McDiarmid's inequality (if $\gamma < \frac{1}{4}$ in the setting of Proposition 4). It also follows from Talagrand's method that the median and expected value of the CF differ by at most a constant, independently of the number of sub-carriers.

D. Concentration of the Cardinality of the Fundamental System of Cycles for LDPC Code Ensembles

Low-density parity-check (LDPC) codes are linear block codes that are represented by sparse parity-check matrices [22]. A sparse parity-check matrix enables to represent the corresponding linear block code by a sparse bipartite graph, and to use this graphical representation for implementing low-complexity iterative message-passing decoding. The low-complexity decoding algorithms used for LDPC codes and some of their variants are remarkable

in that they achieve rates close to the Shannon capacity limit for properly designed code ensembles (see, e.g., [47]). As a result of their remarkable performance under practical decoding algorithms, these coding techniques have revolutionized the field of channel coding and they have been incorporated in various digital communication standards during the last decade.

In the following, we consider ensembles of binary LDPC codes. The codes are represented by bipartite graphs where the variable nodes are located on the left side of the graph, and the parity-check nodes are on the right. The parity-check equations that define the linear code are represented by edges connecting each check node with the variable nodes that are involved in the corresponding parity-check equation. The bipartite graphs representing these codes are sparse in the sense that the number of edges in the graph scales linearly with the block length n of the code. Following standard notation, let λ_i and ρ_i denote the fraction of edges attached, respectively, to variable and parity-check nodes of degree i . The LDPC code ensemble is denoted by $\text{LDPC}(n, \lambda, \rho)$ where n is the block length of the codes, and the pair $\lambda(x) \triangleq \sum_i \lambda_i x^{i-1}$ and $\rho(x) \triangleq \sum_i \rho_i x^{i-1}$ represents, respectively, the left and right degree distributions of the ensemble from the edge perspective. For a short summary of preliminary material on binary LDPC code ensembles see, e.g., [49, Section II-A].

It is well known that linear block codes which can be represented by cycle-free bipartite (Tanner) graphs have poor performance even under ML decoding [20]. The bipartite graphs of capacity-approaching LDPC codes should therefore have cycles. For analyzing this issue, we focused on the notion of "the cardinality of the fundamental system of cycles of bipartite graphs". For the required preliminary material, the reader is referred to [49, Section II-E]. In [49], we address the following question:

Question: Consider an LDPC ensemble whose transmission takes place over a memoryless binary-input output symmetric channel, and refer to the bipartite graphs which represent codes from this ensemble where every code is chosen uniformly at random from the ensemble. How does the average cardinality of the fundamental system of cycles of these bipartite graphs scale as a function of the achievable gap to capacity ?

In light of this question, an information-theoretic lower bound on the average cardinality of the fundamental system of cycles was derived in [49, Corollary 1]. This bound was expressed in terms of the achievable gap to capacity (even under ML decoding) when the communication takes place over a memoryless binary-input output-symmetric channel. More explicitly, it was shown that if ε designates the gap in rate to capacity, then the number of fundamental cycles should grow at least like $\log \frac{1}{\varepsilon}$. Hence, this lower bound remains unbounded as the gap to capacity tends to zero. Consistently with the study in [20] on cycle-free codes, the lower bound on the cardinality of the fundamental system of cycles in [49, Corollary 1] shows quantitatively the necessity of cycles in bipartite graphs which represent good LDPC code ensembles. As a continuation to this work, we present in the following a large-deviations analysis with respect to the cardinality of the fundamental system of cycles for LDPC code ensembles.

Let the triple (n, λ, ρ) represent an LDPC code ensemble, and let \mathcal{G} be a bipartite graph that corresponds to a code from this ensemble. Then, the cardinality of the fundamental system of cycles of \mathcal{G} , denoted by $\beta(\mathcal{G})$, is equal to

$$\beta(\mathcal{G}) = |E(\mathcal{G})| - |V(\mathcal{G})| + c(\mathcal{G})$$

where $E(\mathcal{G})$, $V(\mathcal{G})$ and $c(\mathcal{G})$ denote the edges, vertices and components of \mathcal{G} , respectively, and $|A|$ denotes the number of elements of a (finite) set A . Note that for such a bipartite graph \mathcal{G} , there are n variable nodes and $m = n(1 - R_d)$ parity-check nodes, so there are in total $|V(\mathcal{G})| = n(2 - R_d)$ nodes. Let a_R designate the average right degree (i.e., the average degree of the parity-check nodes), then the number of edges in \mathcal{G} is given by $|E(\mathcal{G})| = ma_R$. Therefore, for a code from the (n, λ, ρ) LDPC code ensemble, the cardinality of the fundamental system of cycles satisfies the equality

$$\beta(\mathcal{G}) = n[(1 - R_d)a_R - (2 - R_d)] + c(\mathcal{G}) \quad (183)$$

where

$$R_d = 1 - \frac{\int_0^1 \rho(x) dx}{\int_0^1 \lambda(x) dx}, \quad a_R = \frac{1}{\int_0^1 \rho(x) dx}$$

denote, respectively, the design rate and average right degree of the ensemble.

Let

$$E \triangleq |E(\mathcal{G})| = n(1 - R_d)a_R \quad (184)$$

denote the number of edges of an arbitrary bipartite graph \mathcal{G} from the ensemble (where we refer interchangeably to codes and to the bipartite graphs that represent these codes from the considered ensemble). Let us arbitrarily assign numbers $1, \dots, E$ to the E edges of \mathcal{G} . Based on Remarks 2 and 3, let's construct a martingale sequence X_0, \dots, X_E where X_i (for $i = 0, 1, \dots, E$) is a RV that denotes the conditional expected number of components of a bipartite graph \mathcal{G} , chosen uniformly at random from the ensemble, given that the first i edges of the graph \mathcal{G} are revealed. Note that the corresponding filtration $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \dots \subseteq \mathcal{F}_E$ in this case is defined so that \mathcal{F}_i is the σ -algebra that is generated by all the sets of bipartite graphs from the considered ensemble whose first i edges are fixed. For this martingale sequence

$$X_0 = \mathbb{E}_{\text{LDPC}(n,\lambda,\rho)}[\beta(\mathcal{G})], \quad X_E = \beta(\mathcal{G})$$

and (a.s.) $|X_k - X_{k-1}| \leq 1$ for $k = 1, \dots, E$ (since by revealing a new edge of \mathcal{G} , the number of components in this graph can change by at most 1). By Corollary 2, it follows that for every $\alpha \geq 0$

$$\begin{aligned} \mathbb{P}(|c(\mathcal{G}) - \mathbb{E}_{\text{LDPC}(n,\lambda,\rho)}[c(\mathcal{G})]| \geq \alpha E) &\leq 2e^{-f(\alpha)E} \\ \Rightarrow \mathbb{P}(|\beta(\mathcal{G}) - \mathbb{E}_{\text{LDPC}(n,\lambda,\rho)}[\beta(\mathcal{G})]| \geq \alpha E) &\leq 2e^{-f(\alpha)E} \end{aligned} \quad (185)$$

where the last transition follows from (183), and the function f was defined in (30). Hence, for $\alpha > 1$, this probability is zero (since $f(\alpha) = +\infty$ for $\alpha > 1$). Note that, from (183), $\mathbb{E}_{\text{LDPC}(n,\lambda,\rho)}[\beta(\mathcal{G})]$ scales linearly with n . The combination of Eqs. (30), (184), (185) gives the following statement:

Theorem 9: [Concentration inequality for the cardinality of the fundamental system of cycles] Let $\text{LDPC}(n, \lambda, \rho)$ be the LDPC code ensemble that is characterized by a block length n , and a pair of degree distributions (from the edge perspective) of λ and ρ . Let \mathcal{G} be a bipartite graph chosen uniformly at random from this ensemble. Then, for every $\alpha \geq 0$, the cardinality of the fundamental system of cycles of \mathcal{G} , denoted by $\beta(\mathcal{G})$, satisfies the following inequality:

$$\mathbb{P}(|\beta(\mathcal{G}) - \mathbb{E}_{\text{LDPC}(n,\lambda,\rho)}[\beta(\mathcal{G})]| \geq \alpha n) \leq 2 \cdot 2^{-[1-h_2(\frac{1-\eta}{2})]n}$$

where h_2 designates the binary entropy function to the base 2, $\eta \triangleq \frac{\alpha}{(1-R_d)a_R}$, and R_d and a_R designate, respectively, the design rate and average right degree of the ensemble. Consequently, if $\eta > 1$, this probability is zero.

Remark 20: The loosened version of Theorem 9, which follows from Azuma's inequality, gets the form

$$\mathbb{P}(|\beta(\mathcal{G}) - \mathbb{E}_{\text{LDPC}(n,\lambda,\rho)}[\beta(\mathcal{G})]| \geq \alpha n) \leq 2e^{-\frac{\eta^2 n}{2}}$$

for every $\alpha \geq 0$, and η as defined in Theorem 9. Note, however, that the exponential decay of the two bounds is similar for values of α close to zero (see the exponents in Azuma's inequality and Corollary 2 in Figure 1).

Remark 21: For various capacity-achieving sequences of LDPC code ensembles on the binary erasure channel, the average right degree scales like $\log \frac{1}{\varepsilon}$ where ε denotes the fractional gap to capacity under belief-propagation decoding (i.e., $R_d = (1 - \varepsilon)C$) [35]. Therefore, for small values of α , the exponential decay rate in the inequality of Theorem 9 scales like $(\log \frac{1}{\varepsilon})^{-2}$. This large-deviations result complements the result in [49, Corollary 1] which provides a lower bound on the average cardinality of the fundamental system of cycles that scales like $\log \frac{1}{\varepsilon}$.

Remark 22: Consider small deviations from the expected value that scale like \sqrt{n} . Note that Corollary 2 is a special case of Theorem 2 when $\gamma = 1$ (i.e., when only an upper bound on the jumps of the martingale sequence is available, but there is no non-trivial upper bound on the conditional variance). Hence, it follows from Proposition 4 that Corollary 2 does not provide in this case any improvement in the exponent of the concentration inequality (as compared to Azuma's inequality) when small deviations are considered.

VI. SUMMARY AND OUTLOOK

This section provides a short summary of this work, followed by a discussion on some directions for further research.

A. Summary

This paper derives some refined versions of the Azuma-Hoeffding inequality (see [7] and [26]) for discrete-parameter martingales with uniformly bounded jumps, and it considers some of their applications in information theory and related topics. The first part is focused on the derivation of these refined inequalities, followed by a discussion on their relations to some classical results in probability theory. Along this discussion, these inequalities are linked to the method of types, martingale central limit theorem, law of iterated logarithm, moderate deviations principle, and to some reported concentration inequalities from the literature. The second part of this work exemplifies these refined inequalities in the context of hypothesis testing and information theory, communication, and coding theory. The interconnections between the concentration inequalities that are analyzed in the first part of this work (including some geometric interpretation w.r.t. some of these inequalities) are studied, and the conclusions of this study serve for the discussion on information-theoretic aspects related to these concentration inequalities in the second part of this work. Rather than covering a large number of applications, we chose to exemplify the use of the concentration inequalities by considering several applications carefully, which also provide some insight on these concentration inequalities. Several more applications and information-theoretic aspects are outlined shortly in the next sub-section, as a continuation to this work. It is aimed to stimulate the use of some refined versions of concentration inequalities that are based on the martingale approach in information-theoretic aspects.

B. Topics for Further Research

We gather here what we consider to be the most interesting directions for future work as a follow-up to this paper.

- *Possible refinements of Theorem 2:* The proof of the concentration inequality in Theorem 2 relies on Bennett's inequality (16). This inequality is applied to a martingale-difference sequence where it is assumed that the jumps of the martingale are uniformly upper bounded, and a global upper bound on their conditional variances is available (see (17)). As was noted in [9, p. 44] with respect to the derivation of Bennett's inequality: "*The above analysis may be extended when more information about the distribution of the component random variables is available.*" Hence, in the context of the proof of Theorem 2, consider a martingale-difference sequence $\{\xi_k, \mathcal{F}_k\}_{k=0}^n$ where, e.g., ξ_k is conditionally symmetrically distributed around zero given \mathcal{F}_{k-1} (for $k = 1, \dots, n$). This additional property enables to obtain a tightened version of Bennett's inequality, and accordingly to improve the exponent of the concentration inequality in Theorem 2 under such an assumption. This direction has been recently studied in [50].
- *Perspectives on the achievable rates and random coding error exponents for linear ISI and non-linear Volterra channels:* Martingale-based concentration inequalities were recently applied in [60] to obtain lower bounds on the error exponents, and upper bounds on the achievable rates of random codes whose transmission takes place over nonlinear Volterra channels. Performance analysis of random coding over these channels is of theoretical and practical interest since various wireless communication channels exhibit non-linear behavior (e.g., the satellite amplifier operates near its saturation point, and exhibits highly non-linear characteristics). For background on digital transmission over non-linear Volterra channels see, e.g., [8, Chapter 14]. The concentration inequalities in Section III can be applied to improve the recent bounds of the work in [60]. To this end, note that the jumps and the conditional variance of the martingale in [60] are uniformly bounded (see [60, Eq. (22)], followed by the refined analysis in [60, Section IV]). Hence, inequality (43) for the special case where $m = 2$ (serving for the derivation of Corollary 4) provides an improvement to the analysis in [60, Eq. (38)] (since $e^x > 1 + x$ for $x > 0$). Furthermore, based on Proposition 2, a further improvement to this analysis can be obtained by using, instead of (44) with $m = 2$, the inequality in (18) (that was used to derive Theorem 2). Based on the analysis in Section III-C, yet another improvement to this analysis can be obtained by relying on (44) for even values of m larger than 2. This provides an improvement to the analysis of the lower bound on the random coding exponents under ML decoding where the communication takes place over a non-linear Volterra channel; respectively, it also improves the upper bounds on the maximal achievable rates of random coding under ML decoding. As was noted in [60, Section V], the same kind of analysis can be applied to the special case where the communication takes place over a stationary, causal and linear intersymbol-interference (ISI) channel. This direction has been recently studied in [61].

- *Channel polarization*: Channel polarization was recently introduced by Arikan [5] to develop a channel coding scheme, called polar codes, that was demonstrated to be a capacity-achieving coding scheme for memoryless symmetric channels under sequential decoding, and with a feasible encoding and decoding complexity. The fundamental concept of channel polarization was introduced in [5, Theorem 1], and it was proved via the convergence theorem for martingales. This analysis was strengthened in [6] where the key to this analysis is in [6, Observation 1]; it stated that the random processes that keep track of the mutual information and Bhattacharyya parameter arising in the course of the channel polarization are, respectively, a martingale and a super-martingale. Since both random processes are bounded (so they fit the setting in Theorem 5), it is of interest to consider the applicability of concentration inequalities for refining the martingale-based analysis of channel polarization for finite block-lengths. A martingale approach to optimize the kernel of polar codes for q -ary input channels (where q is a prime number) has been studied in [2] by maximizing the spread of the polar martingale for noise additive channels. It shows that over $\text{GF}(q)$, for $q > 2$ that is prime, the martingale spread can be significantly increased as compared to the original kernel in [5], leading in some cases to remarkable improvements in the performance of polar codes even with small to moderate block lengths. The study in [2] stimulates further research of the issue of optimizing the polar kernels by following the martingale approach, and possibly some concentration inequalities introduced in our work.
- *Message-passing decoding for graph-based codes*: A great simplification in the analysis of codes defined on graphs under iterative message-passing decoding is obtained by considering the asymptotic performance of ensembles instead of the performance of specific codes. The theoretical justification of this approach is based on Azuma's concentration inequality and a definition of a proper martingale that enables to assert that all except an exponentially (in the block length) small fraction of codes behave within an arbitrary small δ from the ensemble average. This important concentration result was proved by Richardson and Urbanke (see [47, pp. 487–490]). It implies that for a sufficiently large block length, the ensemble average is a good indicator for the performance of individual codes from this ensemble, and it therefore seems a reasonable route to focus on the design and analysis of capacity-approaching ensembles (by density evolution [47]). Some more concentration inequalities for codes defined on graphs and iterative decoding algorithms were derived in the coding literature during the last decade (see [47] and references therein). The concentration inequalities which have been proved in the setting of iterative message-passing decoding so far rely on Azuma's inequality. They are rather loose, and much stronger concentration phenomena are observed in practice for moderate to large block lengths. Therefore, to date, these concentration inequalities serve mostly to justify theoretically the ensemble approach, but they are not tight bounds for finite block lengths. It is of interest to apply martingale-based concentration inequalities, which improve the exponent of Azuma's inequality, to obtain better concentration results. To this end, one needs to tackle the problem of evaluating (or efficiently bounding) the conditional variance of the related martingales.
- *Martingale-based Inequalities Related to Exponential Bounds on Error Probability with Feedback*: As a follow-up to [41, Section 3.3] and [44, Theorem 11], an analysis that relies on the refined versions of Azuma's inequality in Section III (with the standard adaptation of these inequalities to sub-martingales) has the potential to provide further results in this direction.

APPENDIX A PROOF OF THEOREM 3

Consider a discrete-parameter martingale $\{X_k, \mathcal{F}_k\}_{k=0}^{\infty}$ that satisfies the assumptions of Theorem 2 (a.s.) for some fixed constants $d, \sigma > 0$. Due to the convexity of the exponential function, this function is upper bounded over an arbitrary interval by the line segment that intersects the curve of the exponential function at the two endpoints of this interval. The improvement made in the derivation of Theorem 3 relies on a specification of the tightest parabola that coincides with the exponential function at the endpoints of the interval $[-d, d]$, and is above this exponential function over the interval $(-d, d)$. Let $\xi_k \triangleq X_k - X_{k-1}$ for every $k \in \mathbb{N}$. This correspondingly improves the upper bound on $\mathbb{E}[\exp(t\xi_k) | \mathcal{F}_k]$ for $t \geq 0$ as compared to the simple upper bound that refers to the line segment that connects the exponential function at the endpoints of the interval $[-d, d]$. The calculation of the considered parabola leads to the following lemma:

Lemma 4: Let d be an arbitrary positive number. Then, for every $x \in (-\infty, d]$

$$e^x \leq \left(\frac{d+x}{2d} \right) e^d + \left(\frac{d-x}{2d} \right) e^{-d} - \frac{\sinh(d) - de^{-d}}{2} \left(1 - \left(\frac{x}{d} \right)^2 \right).$$

Moreover, this is the tightest parabola that coincides with the exponential function at the endpoints of the interval $[-d, d]$, and is above the exponential function over the interval $(-d, d)$.

Proof: The proof follows by calculus, and the details are omitted for the sake of brevity. \blacksquare

Since by definition $\xi_k = X_k - X_{k-1}$, for every $k \in \mathbb{N}$, then $X_n - X_0 = \sum_{k=1}^n \xi_k$. By the first assumption of Theorem 2, $|\xi_k| \leq d$ a.s. for every k . By Lemma 4, for every $t \geq 0$,

$$e^{t\xi_k} \leq \left(\frac{td + t\xi_k}{2td} \right) e^{td} + \left(\frac{td - t\xi_k}{2td} \right) e^{-td} - \left(\frac{\sinh(td) - tde^{-td}}{2} \right) \left(1 - \left(\frac{\xi_k}{d} \right)^2 \right) \quad (186)$$

a.s. for every $k \in \mathbb{N}$. The assumptions of Theorem 2 on the martingale sequence $\{X_k, \mathcal{F}_k\}_{k=0}^\infty$ yield that a.s.

$$\mathbb{E}[\xi_k | \mathcal{F}_{k-1}] = 0 \quad (187)$$

$$\mathbb{E}[\xi_k^2 | \mathcal{F}_{k-1}] \leq \sigma^2 \quad (188)$$

$$|\xi_k| \leq d \quad (189)$$

for every k , where without any loss of generality σ can be determined such that $\sigma \leq d$. From (186) and (188) then, for every $t \geq 0$,

$$\mathbb{E}[e^{t\xi_k} | \mathcal{F}_{k-1}] \leq \cosh(td) - \frac{(1-\gamma)(\sinh(td) - tde^{-td})}{2}$$

where $\gamma \triangleq \frac{\sigma^2}{d^2}$ is introduced in (11). From (15) and the last inequality, then for every $t \geq 0$

$$\mathbb{E} \left[\exp \left(t \sum_{k=1}^n \xi_k \right) \right] \leq \left[\cosh(td) - \frac{(1-\gamma)(\sinh(td) - tde^{-td})}{2} \right]^n.$$

From (14) and the last inequality, then for an arbitrary $\alpha \geq 0$

$$\begin{aligned} & \mathbb{P}(X_n - X_0 \geq \alpha n) \\ & \leq \left\{ e^{-\alpha t} \left[\cosh(td) - \frac{(1-\gamma)(\sinh(td) - tde^{-td})}{2} \right] \right\}^n \\ & = \left\{ \left(\frac{1+\gamma}{4} \right) e^{(d-\alpha)t} + \left[\frac{1}{2} + \frac{(1+2td)(1-\gamma)}{4} \right] e^{-(\alpha+d)t} \right\}^n \end{aligned}$$

for every $t \geq 0$. In the following, the value of the non-negative parameter t that obtains the tightest exponential bound within this form is calculated. Let $\delta \triangleq \frac{\alpha}{d}$ as in (11).

If $\delta > 1$ then $\mathbb{P}(X_n - X_0 \geq \alpha n) = 0$, and the exponent of the bound is therefore set to infinity. In the continuation, we consider the case where $\delta \leq 1$. Based on the notation in (11) and the substitution $x = td$ (where $x \geq 0$ is a free parameter), the last inequality admits the equivalent form

$$\mathbb{P}(X_n - X_0 \geq \alpha n) \leq \left\{ \frac{(1+\gamma)e^{(1-\delta)x} + [2 + (1-\gamma)(1+2x)]e^{-(1+\delta)x}}{4} \right\}^n \quad (190)$$

where the non-negative parameter x is subject to optimization in order to get the tightest bound within this form.

In the particular case where $\delta = 1$ (note also that $\gamma \leq 1$), then the tightest bound in (190) is obtained in the limit where we let x tend to infinity. This gives

$$\mathbb{P}(X_n - X_0 \geq dn) \leq \left(\frac{1+\gamma}{4} \right)^n. \quad (191)$$

In the following, we derive a closed-form expression for the optimized bound in (190) for the case where $\delta \in [0, 1)$. In this case, by differentiating the base of the exponential bound on the right-hand side of (190) (w.r.t. the free non-negative parameter x) and setting this derivative to zero, one gets the following equation:

$$\left\{ (1+\delta)[2 + (1+2x)(1-\gamma)] - 2(1-\gamma) \right\} e^{-2x} = (1+\gamma)(1-\delta). \quad (192)$$

Let us first examine the special case where $\sigma = d$ (i.e., $\gamma = 1$). In this case, one gets from the assumptions in Theorem 2 that the requirement on the conditional variance is redundant which then implies the same requirement of a bounded-difference martingale that was used to derive the tightened Azuma's inequality in Corollary 2. Indeed, in the case where $\gamma = 1$, equation (192) is simplified, and its solution is $x = \tanh^{-1}(\delta)$. The substitution of this value of x and $\gamma = 1$ into the right-hand side of (190) gives the bound

$$\begin{aligned} & \mathbb{P}(X_n - X_0 \geq \alpha n) \\ & \leq \left(\frac{e^{(1-\delta)x} + e^{-(1+\delta)x}}{2} \right)^n \\ & = (e^{-\delta x} \cosh(x))^n \\ & = \exp \left\{ -n \ln(2) \left[1 - h_2 \left(\frac{1-\delta}{2} \right) \right] \right\} \end{aligned}$$

which indeed coincides with Corollary 2.

The following lemma asserts the existence and uniqueness of a non-negative solution of equation (192), and it also provides a closed-form expression for this solution.

Lemma 5: If $\gamma, \delta \in (0, 1)$, then there exists a unique non-negative solution to equation (192), and it is equal to

$$x = -\frac{1 + W_{-1}(w)}{2} - \frac{\gamma + \delta}{(1 + \delta)(1 - \gamma)} \quad (193)$$

where W_{-1} stands for a branch of the Lambert W function [14], and

$$w \triangleq -\frac{(1 + \gamma)(1 - \delta)}{(1 - \gamma)(1 + \delta)} \cdot e^{-1 - \frac{2(\gamma + \delta)}{(1 + \delta)(1 - \gamma)}}.$$

Proof: Equation (192) can be rewritten in the form

$$(a + bx)e^{-2x} = c \quad (194)$$

where

$$a \triangleq 1 + \gamma + (3 - \gamma)\delta, \quad b \triangleq 2(1 - \gamma)(1 + \delta), \quad c \triangleq (1 + \gamma)(1 - \delta).$$

Lets define the function

$$f(x) = (a + bx)e^{-2x} - c, \quad \forall x \in \mathbb{R}^+$$

then, since $\gamma, \delta \in (0, 1)$,

$$f(0) = a - c = 4\delta > 0, \quad \lim_{x \rightarrow \infty} f(x) = -c < 0$$

so, it follows from the intermediate-value theorem that there exists a solution $x \in (0, \infty)$ to the equation $f(x) = 0$; this value of x is also a solution of equation (192). This assures the existence of a positive solution of equation (192). In order to prove the uniqueness of this positive solution, note that

$$\begin{aligned} f'(x) &= [(b - 2a) - 2bx]e^{-2x} \\ &= -4[\gamma + \delta + (1 + \delta)(1 - \gamma)x]e^{-2x} \\ &< 0, \quad \forall x > 0 \end{aligned}$$

which yields that f is a monotonic decreasing function over the interval $(0, \infty)$, so the positive solution of the equation $f(x) = 0$ (or equivalently equation (192)) is unique.

In order to solve equation (192), we rely on the equivalent simplified form in (194). The substitution $z = -2(x + \frac{a}{b})$ transforms it to the equation

$$ze^z = -\frac{2c}{b} \cdot e^{-\frac{2a}{b}}.$$

Its solution is, by definition, expressed in terms of the Lambert W-function [14]:

$$\begin{aligned} z &= W\left(-\frac{2c}{b} \cdot e^{-\frac{2a}{b}}\right) \\ &= W(w) \end{aligned}$$

where

$$w \triangleq -\frac{1}{e} \frac{(1+\gamma)(1-\delta)}{(1-\gamma)(1+\delta)} \cdot e^{-\frac{2(\gamma+\delta)}{(1+\delta)(1-\gamma)}}.$$

The inverse transformation gives

$$\begin{aligned} x &= -\frac{z}{2} - \frac{a}{b} \\ &= -\frac{1+z}{2} - \frac{\gamma+\delta}{(1+\delta)(1-\gamma)}. \end{aligned}$$

Since $w \in (-\frac{1}{e}, 0)$, then the branch $W = W_{-1}$ of the multi-valued Lambert W-function ensures that x is a real positive number as required. This completes the proof of Lemma 5. \blacksquare

Putting all the pieces of this proof together, it completes the proof of Theorem 3.

APPENDIX B PROOF OF LEMMA 2

The first and third properties of φ_m follow from the power series expansion of the exponential function where

$$\varphi_m(y) = \frac{m!}{y^m} \sum_{l=m}^{\infty} \frac{y^l}{l!} = \sum_{l=0}^{\infty} \frac{m!y^l}{(m+l)!}, \quad \forall y \in \mathbb{R}.$$

From its absolute convergence then $\lim_{y \rightarrow 0} \varphi_m(y) = 1$, and it follows from the above power series expansion that φ_m is strictly monotonic increasing over the interval $[0, \infty)$. The fourth property of φ_m holds since

$$\varphi_m(y) = \frac{m!}{y^m} \cdot R_{m-1}(y)$$

where R_{m-1} is the remainder of the Taylor approximation of order $m-1$ for the exponential function $f(y) = e^y$. Hence, for every $y \leq 0$,

$$\varphi_m(y) = f^{(m)}(\xi) = e^\xi$$

for some $\xi \in [y, 0]$, so $0 < \varphi_m(y) \leq 1$. The second property of φ_m follows by combining the third and fourth properties.

APPENDIX C PROOF OF COROLLARY 4

The proof of Corollary 4 is based on the specialization of Theorem 4 for $m = 2$. This gives that, for every $\alpha \geq 0$, the following concentration inequality holds:

$$\mathbb{P}(|X_n - X_0| \geq n\alpha) \leq 2 \left\{ \inf_{x \geq 0} e^{-\delta x} \left[1 + \gamma(e^x - 1 - x) \right] \right\}^n \quad (195)$$

where $\gamma = \gamma_2$ according to the notation in (11).

By differentiating the logarithm of the right-hand side of (195) w.r.t. x (where $x \geq 0$) and setting this derivative to zero, it follows that

$$\frac{1 - \gamma x}{\gamma(e^x - 1)} = \frac{1 - \delta}{\delta}. \quad (196)$$

Let us first consider the case where $\delta = 1$. In this case, this equation is satisfied either if $x = \frac{1}{\gamma}$ or in the limit where $x \rightarrow \infty$. In the former case where $x = \frac{1}{\gamma}$, the resulting bound in (195) is equal to

$$\exp \left[-n \left(\frac{1}{\gamma} - \ln \left(\gamma \left(e^{\frac{1}{\gamma}} - 1 \right) \right) \right) \right]. \quad (197)$$

In the latter case where $x \rightarrow \infty$, the resulting bound in (195) when $\delta = 1$ is equal to

$$\begin{aligned} & \lim_{x \rightarrow \infty} e^{-nx} (1 + \gamma(e^x - 1 - x))^n \\ &= \lim_{x \rightarrow \infty} \left(e^{-x} + \gamma(1 - (1+x)e^{-x}) \right)^n \\ &= \gamma^n. \end{aligned}$$

Hence, since for $\gamma \in (0, 1)$

$$\begin{aligned} \ln\left(\frac{1}{\gamma}\right) &= \frac{1}{\gamma} - \ln\left(\gamma e^{\frac{1}{\gamma}}\right) \\ &< \frac{1}{\gamma} - \ln\left(\gamma(e^{\frac{1}{\gamma}} - 1)\right) \end{aligned}$$

then the optimized value is $x = \frac{1}{\gamma}$, and the resulting bound in the case where $\delta = 1$ is equal to (197).

Let us consider now the case where $0 < \delta < 1$ (the case where $\delta = 0$ is trivial). In the following lemma, the existence and uniqueness of a solution of this equation is assured, and a closed-form expression for this solution is provided.

Lemma 6: If $\delta \in (0, 1)$, then equation (196) has a unique solution, and it lies in $(0, \frac{1}{\gamma})$. This solution is given in (46).

Proof: Consider equation (196), and note that the right-hand side of this equation is positive for $\delta \in (0, 1)$. The function

$$t(x) = \frac{1 - \gamma x}{\gamma(e^x - 1)}, \quad x \in \mathbb{R}$$

on the left-hand side of (196) is negative for $x < 0$ and $x > \frac{1}{\gamma}$. Since the function t is continuous on the interval $(0, \frac{1}{\gamma}]$ and

$$t\left(\frac{1}{\gamma}\right) = 0, \quad \lim_{x \rightarrow 0^+} t(x) = +\infty$$

then there is a solution $x \in (0, \frac{1}{\gamma})$. Moreover, the function t is monotonic decreasing in the interval $(0, \frac{1}{\gamma}]$ (the numerator of t is monotonic decreasing and the denominator of t is monotonic increasing and both are positive in this interval). This implies the existence and uniqueness of the solution, which lies in the interval $(0, \frac{1}{\gamma})$. In the following, a closed-form expression of this solution is derived. Note that Eq. (196) can be expressed in the form

$$\frac{a - x}{e^x - 1} = b \tag{198}$$

where

$$a \triangleq \frac{1}{\gamma}, \quad b \triangleq \frac{1 - \delta}{\delta} \tag{199}$$

are both positive. The substitution $u = a + b - x$ in (198) gives

$$ue^u = be^{a+b}$$

whose solution is, by definition, given by $u = W_0(be^{a+b})$ where W_0 denotes the principal branch of the multi-valued Lambert W function [14]. Since $a, b > 0$ then $be^{a+b} > 0$, so that the principal branch of W is the only one which is a real number. In the following, it will be confirmed that the selection of this branch also implies that $x > 0$ as required. By the inverse transformation one gets

$$\begin{aligned} x &= a + b - u \\ &= a + b - W_0\left(be^{a+b}\right) \end{aligned} \tag{200}$$

Hence, the selection of this branch for W indeed ensures that x is the positive solution we are looking for (since $a, b > 0$, then it readily follows from the definition of the Lambert W function that $W_0(be^{a+b}) < a + b$ and it was earlier proved in this appendix that the positive solution x of (196) is unique). Finally, the substitution of (199) into (200) gives (46). This completes the proof of Lemma 6. ■

The bound in (195) is given by

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) \leq 2 \exp\left(-n \left[\delta x - \ln(1 + \gamma(e^x - 1 - x))\right]\right) \tag{201}$$

with the value of x in (46).

APPENDIX D
PROOF OF PROPOSITION 1

Lets compare the right-hand sides of (34) and (195) that refer to Corollaries 2 and 4, respectively. Proposition 1 follows by showing that if $\gamma \leq \frac{1}{2}$

$$1 + \gamma(\exp(x) - 1 - x) < \cosh(x), \quad \forall x > 0. \quad (202)$$

To this end, define

$$f(x) \triangleq \cosh(x) - \gamma(\exp(x) - 1 - x), \quad \forall x \geq 0.$$

If $\gamma \leq \frac{1}{2}$, then for every $x > 0$

$$\begin{aligned} f'(x) &= \sinh(x) - \gamma(\exp(x) - 1) \\ &= \left(\frac{1}{2} - \gamma\right) \exp(x) + \gamma - \frac{\exp(-x)}{2} \\ &> \left(\frac{1}{2} - \gamma\right) + \gamma - \frac{1}{2} = 0 \end{aligned}$$

so, since f is monotonic increasing on $[0, \infty)$ and $f(0) = 0$, then $f(x) > 0$ for every $x > 0$. This validates (202), and it therefore completes the proof of Proposition 1.

APPENDIX E
PROOF OF PROPOSITION 2

Lemma 7: For every $\gamma, x > 0$

$$\frac{\gamma e^x + e^{-\gamma x}}{1 + \gamma} < 1 + \gamma(e^x - 1 - x). \quad (203)$$

Proof: Let γ be an arbitrary positive number, and define the function

$$f_\gamma(x) \triangleq \frac{\gamma e^x + e^{-\gamma x}}{1 + \gamma} - [1 + \gamma(e^x - 1 - x)], \quad x \geq 0.$$

Then, $f_\gamma(0) = 0$, and the first derivative is equal to

$$f'_\gamma(x) = \gamma \left(1 - \frac{\gamma e^x + e^{-\gamma x}}{1 + \gamma} \right).$$

From the convexity of the exponential function $y(u) = e^u$, then for every $x > 0$

$$\begin{aligned} \frac{\gamma e^x + e^{-\gamma x}}{1 + \gamma} &= \left(\frac{\gamma}{1 + \gamma}\right) y(x) + \left(\frac{1}{1 + \gamma}\right) y(-\gamma x) \\ &> y\left(\frac{\gamma}{1 + \gamma} \cdot x + \frac{1}{1 + \gamma} \cdot (-\gamma x)\right) \\ &= y(0) = 1 \end{aligned}$$

so, it follows that $f'_\gamma(x) < 0$ for every $x > 0$. Since $f_\gamma(0) = 0$ and the first derivative is negative over $(0, \infty)$, then $f_\gamma(x) < 0$ for every $x > 0$. This completes the proof of inequality (203). ■

This claim in Proposition 2 follows directly from Lemma 7, and the two inequalities in (18) and (44) with $m = 2$. In the case where $m = 2$, the right-hand side of (44) is equal to

$$(1 + \gamma(e^{td} - 1 - td))^n.$$

Note that (18) and (44) with $m = 2$ were used to derive, respectively, Theorem 2 and Corollary 4 (based on Chernoff's bound). The conclusion follows by substituting $x \triangleq td$ on the right-hand sides of (18) and (44) with $m = 2$ (so that $x \geq 0$ since $t \geq 0$ and $d > 0$, and (203) turns from an inequality if $x > 0$ into an equality if $x = 0$).

APPENDIX F
PROOF OF PROPOSITION 3

For the case where $m = 2$, the conditions in Theorem 4 are identical to Corollary 5. Hence, since Corollary 4 follows as a particular case of Theorem 4 for $m = 2$, then Corollary 5 implies the result in Corollary 4.

The sequence $\{\gamma_l\}_{l=2}^m$ is monotonic non-increasing and non-negative. This follows from the assumption where a.s.

$$|X_k - X_{k-1}| \leq d, \quad \forall k \in \mathbb{N}$$

and the definition of the sequences $\{\gamma_l\}_{l=2}^m$ and $\{\mu_l\}_{l=2}^m$ in (39) and (47), respectively. Hence, for every l

$$\begin{aligned} 0 \leq \gamma_{l+1} &= \mathbb{E} \left[\left(\frac{|X_{k+1} - X_k|}{d} \right)^{l+1} \middle| \mathcal{F}_{k-1} \right] \\ &\leq \mathbb{E} \left[\left(\frac{|X_{k+1} - X_k|}{d} \right)^l \middle| \mathcal{F}_{k-1} \right] = \gamma_l \leq 1. \end{aligned}$$

Since $\{\gamma_l\}_{l=2}^\infty$ is monotonic non-increasing and non-negative (note also that γ_l is independent of the value of m) then it is a convergent sequence.

Referring to the base of the exponential bound on the right-hand side of (38), for an arbitrary $x \geq 0$ and an even $m \in \mathbb{N}$, we want to show that the value of the non-negative base of this exponential bound is decreased by increasing m to $m + 2$; hence, it implies that the bound in Corollary 5 is improved as the value of m is increased (where $m \in \mathbb{N}$ is even). To show this, note that for an arbitrary $x \geq 0$ and an even $m \in \mathbb{N}$,

$$\begin{aligned} &\left[1 + \sum_{l=2}^{m-1} \frac{(\gamma_l - \gamma_m)x^l}{l!} + \gamma_m(e^x - 1 - x) \right] \\ &\quad - \left[1 + \sum_{l=2}^{m+1} \frac{(\gamma_l - \gamma_{m+2})x^l}{l!} + \gamma_{m+2}(e^x - 1 - x) \right] \\ &= \sum_{l=2}^{m-1} \frac{(\gamma_{m+2} - \gamma_m)x^l}{l!} - \frac{(\gamma_m - \gamma_{m+2})x^m}{m!} \\ &\quad - \frac{(\gamma_{m+1} - \gamma_{m+2})x^{m+1}}{(m+1)!} + (\gamma_m - \gamma_{m+2})(e^x - 1 - x) \\ &= (\gamma_m - \gamma_{m+2}) \left(e^x - \sum_{l=0}^m \frac{x^l}{l!} \right) - \frac{(\gamma_{m+1} - \gamma_{m+2})x^{m+1}}{(m+1)!} \\ &= (\gamma_m - \gamma_{m+2}) \sum_{l=m+1}^{\infty} \frac{x^l}{l!} - \frac{(\gamma_{m+1} - \gamma_{m+2})x^{m+1}}{(m+1)!} \\ &= (\gamma_m - \gamma_{m+2}) \left[\frac{x^{m+1}}{(m+1)!} + \sum_{l=m+2}^{\infty} \frac{x^l}{l!} \right] - \frac{(\gamma_{m+1} - \gamma_{m+2})x^{m+1}}{(m+1)!} \\ &= \frac{(\gamma_m - \gamma_{m+1})x^{m+1}}{(m+1)!} + (\gamma_m - \gamma_{m+2}) \sum_{l=m+2}^{\infty} \frac{x^l}{l!}. \end{aligned}$$

Since $\gamma_m - \gamma_{m+1} \geq 0$ and $\gamma_m - \gamma_{m+2} \geq 0$ (due to the monotonicity of the sequence $\{\gamma_l\}_{l=2}^\infty$) and $x \geq 0$, then the above difference is non-negative and therefore

$$\begin{aligned} &1 + \sum_{l=2}^{m-1} \frac{(\gamma_l - \gamma_m)x^l}{l!} + \gamma_m(e^x - 1 - x) \\ &\geq 1 + \sum_{l=2}^{m+1} \frac{(\gamma_l - \gamma_{m+2})x^l}{l!} + \gamma_{m+2}(e^x - 1 - x). \end{aligned}$$

Multiplying the two sides of this inequality by $e^{-\delta x}$ and taking the infimum w.r.t. $x \geq 0$ gives that, for every $m \in \mathbb{N}$ that is even,

$$\begin{aligned} & \inf_{x \geq 0} \left\{ e^{-\delta x} \left[1 + \sum_{l=2}^{m-1} \frac{(\gamma_l - \gamma_m)x^l}{l!} + \gamma_m(e^x - 1 - x) \right] \right\} \\ & \geq \inf_{x \geq 0} \left\{ e^{-\delta x} \left[1 + \sum_{l=2}^{m+1} \frac{(\gamma_l - \gamma_{m+2})x^l}{l!} + \gamma_{m+2}(e^x - 1 - x) \right] \right\}. \end{aligned}$$

Note that the right-hand side of this inequality is non-negative due to inequality (38) where the right-hand side is a non-negative upper bound. Hence, it follows that the bound in Corollary 5 improves as the value of the even number $m \in \mathbb{N}$ is increased. This completes the proof of Proposition 3.

APPENDIX G PROOF OF COROLLARY 6

A minimization of the logarithm of the exponential bound on the right-hand side of (48) gives the equation

$$\frac{\sum_{l=2}^{m-1} \frac{(\gamma_l - \gamma_m)x^{l-1}}{(l-1)!} + \gamma_m(e^x - 1)}{1 + \sum_{l=2}^{m-1} \frac{(\gamma_l - \gamma_m)x^l}{l!} + \gamma_m(e^x - 1 - x)} = \delta$$

and after standard algebraic operations, it gives the equation

$$\begin{aligned} & \gamma_m \left(\frac{1}{\delta} - 1 \right) (e^x - 1 - x) + \frac{\gamma_2 x}{\delta} \\ & + \sum_{l=2}^{m-1} \left\{ \left[\frac{\gamma_{l+1}}{\delta} - \gamma_l - \gamma_m \left(\frac{1}{\delta} - 1 \right) \right] \frac{x^l}{l!} \right\} - 1 = 0. \end{aligned} \quad (204)$$

As we have seen in the proof of Corollary 4 (see Appendix C), the solution of this equation can be expressed in a closed-form for $m = 2$, but in general, a closed-form solution to this equation is not available. A sub-optimal value of x on the right-hand side of (38) is obtained by neglecting the sum that appears in the second line of this equation (the rationality for this approximation is that $\{\gamma_l\}$ was observed to converge very fast, so it was verified numerically that γ_l stays almost constant starting from a small value of l). Note that the operation of $\inf_{x \geq 0}$ can be loosened by taking an arbitrary non-negative value of x ; hence, in particular, x will be chosen in the following to satisfy the equation

$$\gamma_m \left(\frac{1}{\delta} - 1 \right) (e^x - 1 - x) + \frac{\gamma_2 x}{\delta} = 1.$$

By dividing both sides of the equation by γ_2 , then it gives the equation $a + b - cx = be^x$ with a, b and c from (50). This equation can be written in the form

$$\left(\frac{a+b}{c} - x \right) e^{-x} = \frac{b}{c}.$$

Substituting $u \triangleq \frac{a+b}{c} - x$ gives the equation

$$ue^u = \frac{b}{c} \cdot e^{\frac{a+b}{c}}$$

whose solution is given by

$$u = W_0 \left(\frac{b}{c} \cdot e^{\frac{a+b}{c}} \right)$$

where W_0 denotes the principal branch of the Lambert W function [14]. The inverse transformation back to x gives that

$$x = \frac{a+b}{c} - W_0 \left(\frac{b}{c} \cdot e^{\frac{a+b}{c}} \right).$$

This justifies the choice of x in (49), and it provides a loosening of either Theorem 4 or Corollary 5 by replacing the operation of the infimum over the non-negative values of x on the right-hand side of (38) with the value of x that is given in (49) and (50). For $m = 2$ where the sum on the left-hand side of (204) that was later neglected is anyway zero, this forms indeed the exact optimal value of x (so that it coincide with (193) in Corollary 4).

APPENDIX H PROOF OF PROPOSITION 4

Let $\{X_k, \mathcal{F}_k\}_{k=0}^{\infty}$ be a discrete-parameter martingale. We prove in the following that Theorems 2 and 4, and also Corollaries 3 and 4 imply (54). Since Theorem 3 is looser than Theorem 2, and it was introduced in Section III in order to highlight a geometric interpretation, we skip the proof that also Theorem 3 implies the same scaling as in (54).

A. Analysis Related to Theorem 2

Let $\{X_k, \mathcal{F}_k\}_{k=0}^{\infty}$ be a discrete-parameter martingale that satisfies the conditions in Theorem 2. From (10)

$$\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) \leq 2 \exp\left(-n D\left(\frac{\delta' + \gamma}{1 + \gamma} \parallel \frac{\gamma}{1 + \gamma}\right)\right) \quad (205)$$

where from (11)

$$\delta' \triangleq \frac{\alpha}{\sqrt{n}} = \frac{\delta}{\sqrt{n}}. \quad (206)$$

From the right-hand side of (205)

$$\begin{aligned} & D\left(\frac{\delta' + \gamma}{1 + \gamma} \parallel \frac{\gamma}{1 + \gamma}\right) \\ &= \frac{\gamma}{1 + \gamma} \left[\left(1 + \frac{\delta}{\gamma\sqrt{n}}\right) \ln\left(1 + \frac{\delta}{\gamma\sqrt{n}}\right) + \frac{1}{\gamma} \left(1 - \frac{\delta}{\sqrt{n}}\right) \ln\left(1 - \frac{\delta}{\sqrt{n}}\right) \right]. \end{aligned} \quad (207)$$

From the equality

$$(1 + u) \ln(1 + u) = u + \sum_{k=2}^{\infty} \frac{(-u)^k}{k(k-1)}, \quad -1 < u \leq 1$$

then it follows from (207) that for every $n > \frac{\delta^2}{\gamma^2}$

$$\begin{aligned} nD\left(\frac{\delta' + \gamma}{1 + \gamma} \parallel \frac{\gamma}{1 + \gamma}\right) &= \frac{\delta^2}{2\gamma} - \frac{\delta^3(1 - \gamma)}{6\gamma^2} \frac{1}{\sqrt{n}} + \dots \\ &= \frac{\delta^2}{2\gamma} + O\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

Substituting this into the exponent on the right-hand side of (205) gives (54).

B. Analysis Related to Corollary 3

Let $\{X_k, \mathcal{F}_k\}_{k=0}^{\infty}$ be a discrete-parameter martingale that satisfies the conditions in Theorem 2. From Corollary 3, it follows that for every $\alpha \geq 0$

$$\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) \leq 2 \exp\left(-n\gamma \left[\left(1 + \frac{\delta'}{\gamma}\right) \ln\left(1 + \frac{\delta'}{\gamma}\right) - \frac{\delta'}{\gamma} \right]\right)$$

where δ' is introduced in (206). By substituting (36) and (206) into the last inequality, it follows that for every $\alpha \geq 0$

$$\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) \leq 2 \exp\left(-\frac{\delta^2}{2\gamma} B\left(\frac{\delta}{\gamma\sqrt{n}}\right)\right). \quad (208)$$

The power series expansion around zero of the function B in (36) is given by

$$B(u) = \sum_{k=0}^{\infty} \frac{2(-1)^k u^k}{(k+1)(k+2)}, \quad |u| \leq 1.$$

Therefore, if $n \gg 1$, substituting this equality in (208) implies that for every $\alpha \geq 0$

$$\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) \leq 2 \exp\left(-\frac{\delta^2}{2\gamma}\right) \left(1 + O\left(\frac{1}{\sqrt{n}}\right)\right)$$

which gives the inequality in (54).

C. Analysis Related to Theorem 4

From (38), for every $\alpha \geq 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) \leq 2 \left\{ \inf_{x \geq 0} e^{-\delta'x} \left[1 + \sum_{l=2}^{m-1} \frac{(\gamma_l - \gamma_m)x^l}{l!} + \gamma_m(e^x - 1 - x) \right] \right\}^n$$

where

$$\delta' = \frac{\alpha}{\sqrt{n}} = \frac{\delta}{\sqrt{n}}, \quad \gamma_l \triangleq \frac{\mu_l}{d^l}, \quad \forall l = 2, \dots, m.$$

The optimization of the free non-negative parameter x in the above upper bound is obtained by minimizing the exponent of this bound. This gives the equation

$$\frac{d}{dx} \left\{ -\delta'x + \ln \left[1 + \sum_{l=2}^{m-1} \frac{(\gamma_l - \gamma_m)x^l}{l!} + \gamma_m(e^x - 1 - x) \right] \right\} = 0$$

that is equivalent to the equation

$$\frac{(\gamma_2 - \gamma_m)x + \gamma_m(e^x - 1) + \sum_{l=3}^{m-1} \frac{(\gamma_l - \gamma_m)x^{l-1}}{(l-1)!}}{1 + \sum_{l=2}^{m-1} \frac{(\gamma_l - \gamma_m)x^l}{l!} + \gamma_m(e^x - 1 - x)} = \frac{\delta}{\sqrt{n}}. \quad (209)$$

Note that if $n \gg 1$, then $\delta' \approx 0$ but the values of $\{\gamma_l\}_{l=2}^m$ stay fixed. Hence, in this case, the solution of (209) is approximately zero. As in the previous analysis, we start with an approximate analysis to get an approximate solution of (209). Since the above upper bound is valid for every $x \geq 0$, we then perform an exact analysis with the approximated value of x that forms a solution of the optimization equation in (209).

For $x \approx 0$, we calculate a first order approximation of the left-hand side of (209). Note that

$$\begin{aligned} & (\gamma_2 - \gamma_m)x + \gamma_m(e^x - 1) + \sum_{l=3}^{m-1} \frac{(\gamma_l - \gamma_m)x^{l-1}}{(l-1)!} \\ &= \gamma_2x + O(x^2), \end{aligned}$$

and

$$1 + \sum_{l=2}^{m-1} \frac{(\gamma_l - \gamma_m)x^l}{l!} + \gamma_m(e^x - 1 - x) = 1 + O(x^2)$$

so the left-hand side of (209) is equal to $\gamma_2x + O(x^2)$. Hence, if $n \gg 1$, then $x \approx \frac{\delta}{\gamma_2\sqrt{n}}$ is an approximated solution of (209). Following this approximation, we chose sub-optimally the value of x to be

$$x = \frac{\delta}{\gamma_2\sqrt{n}} \quad (210)$$

and proceed with an exact analysis that relies on the concentration inequality in Theorem 4. Substituting $\delta' = \frac{\delta}{\sqrt{n}}$, and the replacement of the infimum in the upper bound by the value of the bound at x in (210) gives that, for every $\alpha \geq 0$,

$$\begin{aligned}
& \mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) \\
& \leq 2 \left\{ e^{-\delta'x} \left[1 + \sum_{l=2}^{m-1} \frac{(\gamma_l - \gamma_m)x^l}{l!} + \gamma_m(e^x - 1 - x) \right] \right\}^n \\
& = 2 \exp\left(-\frac{\delta^2}{\gamma_2}\right) \left[1 + \sum_{l=2}^{m-1} \frac{\gamma_l x^l}{l!} + \gamma_m \left(e^x - \sum_{l=0}^{m-1} \frac{x^l}{l!} \right) \right]^n \\
& = 2 \exp\left(-\frac{\delta^2}{\gamma_2}\right) \left[1 + \frac{\gamma_2 x^2}{2} + O(x^3) \right]^n \\
& = 2 \exp\left(-\frac{\delta^2}{\gamma_2}\right) \left[1 + \frac{\delta^2}{2\gamma_2 n} + O\left(n^{-\frac{3}{2}}\right) \right]^n \\
& = 2 \exp\left(-\frac{\delta^2}{2\gamma_2}\right) \left(1 + O\left(n^{-\frac{1}{2}}\right) \right).
\end{aligned}$$

Note that $\gamma_2 = \gamma$ in (11) and Theorem 4. This proves (54) via the concentration inequality in Theorem 4 for every $m \geq 2$.

D. Analysis Related to Corollary 4

The concentration inequality in (54) was proved to be a consequence of Theorem 4 for an arbitrary even $m \geq 2$. Since Corollary 4 is a special case of Theorem 4 for $m = 2$, then it follows that (54) results in from Corollary 4 without a need for a separate analysis.

APPENDIX I PROOF OF PROPOSITION 5

Consider the setting of Theorem 2 where $\{X_k, \mathcal{F}_k\}_{k=0}^\infty$ is a discrete-parameter real-valued martingale such that

$$|X_k - X_{k-1}| \leq d, \quad \mathbb{E}[(X_k - X_{k-1})^2 | \mathcal{F}_{k-1}] \leq \sigma^2$$

a.s. for every $k \in \mathbb{N}$. Let $S_k \triangleq \frac{X_k - X_0}{d}$, so $\{S_k, \mathcal{F}_k\}_{k=0}^\infty$ is a martingale sequence with $S_0 = 0$. Also, a.s.

$$\begin{aligned}
Y_k & \triangleq S_k - S_{k-1} = \frac{X_k - X_{k-1}}{d} \leq 1 \\
Q_n & \triangleq \sum_{j=1}^n \mathbb{E}(Y_j^2 | \mathcal{F}_{j-1}) \leq \frac{n\sigma^2}{d^2} = \gamma n
\end{aligned}$$

where $\gamma \triangleq \frac{\sigma^2}{d^2}$ is introduced in (11). Hence, $Q_n \leq \gamma n$ a.s., and it follows from Theorem 6 that for every $\alpha \geq 0$

$$\begin{aligned}
& \mathbb{P}(X_n - X_0 \geq \alpha n) \\
& = \mathbb{P}(S_n \geq \delta n, Q_n \leq \gamma n) \\
& \leq \exp\left(-\frac{\delta^2 n}{2\gamma} B\left(\frac{\delta}{\gamma}\right)\right)
\end{aligned}$$

where the last inequality follows from (70). From the definition of the function B in (36) then, for every $\alpha \geq 0$,

$$\mathbb{P}(X_n - X_0 \geq \alpha n) \leq \exp\left(-n\gamma \left[\left(1 + \frac{\delta}{\gamma}\right) \ln\left(1 + \frac{\delta}{\gamma}\right) - \frac{\delta}{\gamma} \right]\right).$$

By applying the last inequality to the martingale sequence $\{-X_k, \mathcal{F}_k\}_{k=0}^\infty$, the same inequality also holds as an upper bound on $\mathbb{P}(X_n - X_0 \leq -\alpha n)$ for an arbitrary $\alpha \geq 0$. Finally, inequality (37) holds due to the equality

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) = \mathbb{P}(X_n - X_0 \geq \alpha n) + \mathbb{P}(X_n - X_0 \leq -\alpha n).$$

This shows that Corollary 3 is a consequence of Theorem 6. Similarly, it follows from (70) that for every $\alpha \geq 0$,

$$\mathbb{P}(X_n - X_0 \geq \alpha\sqrt{n}) \leq \exp\left(-\frac{\delta^2}{2\gamma} B\left(\frac{\delta}{\gamma\sqrt{n}}\right)\right).$$

The latter inequality coincides with (208). As is shown in Appendix H, (54) follows from (208), so a concentration inequality of the form of (54) in Proposition 4 follows as a consequence of Theorem 6. This completes the proof of Proposition 5.

APPENDIX J

ANALYSIS RELATED TO THE MODERATE DEVIATIONS PRINCIPLE IN SECTION IV-F

It is demonstrated in the following that, in contrast to Azuma's inequality, both Theorems 2 and 4 provide upper bounds on

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| \geq \alpha n^\eta\right), \quad \forall \alpha \geq 0$$

which coincide with the exact asymptotic limit in (82). It is proved under the further assumption that there exists some constant $d > 0$ such that $|X_k| \leq d$ a.s. for every $k \in \mathbb{N}$. Let us define the martingale sequence $\{S_k, \mathcal{F}_k\}_{k=0}^n$ where

$$S_k \triangleq \sum_{i=1}^k X_i, \quad \mathcal{F}_k \triangleq \sigma(X_1, \dots, X_k)$$

for every $k \in \{1, \dots, n\}$ with $S_0 = 0$ and $\mathcal{F}_0 = \{\emptyset, \mathcal{F}\}$.

1) *Analysis related to Azuma's inequality:* The martingale sequence $\{S_k, \mathcal{F}_k\}_{k=0}^n$ has uniformly bounded jumps, where $|S_k - S_{k-1}| = |X_k| \leq d$ a.s. for every $k \in \{1, \dots, n\}$. Hence it follows from Azuma's inequality that, for every $\alpha \geq 0$,

$$\mathbb{P}(|S_n| \geq \alpha n^\eta) \leq 2 \exp\left(-\frac{\alpha^2 n^{2\eta-1}}{2d^2}\right)$$

and therefore

$$\lim_{n \rightarrow \infty} n^{1-2\eta} \ln \mathbb{P}(|S_n| \geq \alpha n^\eta) \leq -\frac{\alpha^2}{2d^2}. \quad (211)$$

This differs from the limit in (82) where σ^2 is replaced by d^2 , so Azuma's inequality does not provide the asymptotic limit in (82) (unless $\sigma^2 = d^2$, i.e., $|X_k| = d$ a.s. for every k).

2) *Analysis related to Theorem 2:* The analysis here is a slight modification of the analysis in Appendix H-A with the required adaptation of the calculations for $\eta \in (\frac{1}{2}, 1)$. It follows from Theorem 2 that, for every $\alpha \geq 0$,

$$\mathbb{P}(|S_n| \geq \alpha n^\eta) \leq 2 \exp\left(-n D\left(\frac{\delta' + \gamma}{1 + \gamma} \parallel \frac{\gamma}{1 + \gamma}\right)\right)$$

where γ is introduced in (11), and δ' in (206) is replaced with

$$\delta' \triangleq \frac{\alpha}{n^{1-\eta}} = \delta n^{-(1-\eta)} \quad (212)$$

due to the definition of δ in (11). Following the same analysis as in Appendix H-A, it follows that for every $n \in \mathbb{N}$

$$\mathbb{P}(|S_n| \geq \alpha n^\eta) \leq 2 \exp\left(-\frac{\delta^2 n^{2\eta-1}}{2\gamma} \left[1 + \frac{\alpha(1-\gamma)}{3\gamma d} \cdot n^{-(1-\eta)} + \dots\right]\right)$$

and therefore (since, from (11), $\frac{\delta^2}{\gamma} = \frac{\alpha^2}{\sigma^2}$)

$$\lim_{n \rightarrow \infty} n^{1-2\eta} \ln \mathbb{P}(|S_n| \geq \alpha n^\eta) \leq -\frac{\alpha^2}{2\sigma^2}. \quad (213)$$

Hence, this upper bound coincides with the exact asymptotic result in (82).

3) *Analysis related to Theorem 4*: It is shown in the following that Theorem 4 coincides with the exact asymptotic result in (82) for an arbitrary even number $m \geq 2$. To this end, it is sufficient to prove it w.r.t. the looser version in Corollary 5. Due to Proposition 3, the tightness of the bound in Corollary 5 is improved by increasing the even value of $m \geq 2$. It is therefore enough to show that choosing $m = 2$, which provides the weakest bound in Corollary 5, is already good enough to get the exact asymptotic result in (82). But, for $m = 2$, Theorem 4 and Corollary 5 coincide (and both imply the inequality in Corollary 4). This suggests to follow the analysis in Appendix H-C for $m = 2$ with the slight required modification of this analysis for $\eta \in (\frac{1}{2}, 1)$ (instead of the case where it is one-half as in Appendix H-C). In the following, we refer to the martingale sequence $\{S_k, \mathcal{F}_k\}_{k=1}^n$ as above. Based on the analysis in Appendix H-C, the sub-optimal value of x in (210) is modified to $x = \left(\frac{\delta}{\gamma}\right) n^{-(1-\eta)}$ where, from (11), $\gamma = \frac{\sigma^2}{d^2}$ and $\delta = \frac{\alpha}{d}$. Hence, $x = \frac{\delta'}{\gamma}$ with δ' in (212). Following the analysis in Appendix H-C for the special case of $m = 2$ with the sub-optimal x as above, then for every $\alpha \geq 0$

$$\begin{aligned}
& \mathbb{P}(|S_n| \geq \alpha n^\eta) \\
& \leq 2 \left\{ e^{-\delta'x} \left[1 + \gamma(e^x - 1 - x) \right] \right\}^n \\
& = 2 \exp\left(-\frac{n\delta'^2}{\gamma}\right) \left(1 + \frac{\gamma x^2}{2} + \frac{\gamma x^3}{6} + O(x^4) \right)^n \\
& = 2 \exp\left(-\frac{\alpha^2 n^{2\eta-1}}{\gamma d^2}\right) \left(1 + \frac{\alpha^2 n^{-2(1-\eta)}}{2\gamma d^2} + \frac{\alpha^3 n^{-3(1-\eta)}}{6\gamma^2 d^3} + \dots \right)^n \\
& \leq 2 \exp\left(-\frac{\alpha^2 n^{2\eta-1}}{\sigma^2}\right) \exp\left(\frac{\alpha^2 n^{2\eta-1} (1 + O(n^{-(1-\eta)}))}{2\sigma^2}\right) \\
& \leq 2 \exp\left(-\frac{\alpha^2 n^{2\eta-1} (1 + O(n^{-(1-\eta)}))}{2\sigma^2}\right)
\end{aligned}$$

so the asymptotic result in (213) also follows in this case, thus coinciding with the exact asymptotic result in (82).

APPENDIX K PROOF OF PROPOSITION 6

The proof of (144) is based on calculus, and it is similar to the proof of the limit in (143) that relates the divergence and Fisher information. For the proof of (146), note that

$$C(P_\theta, P_{\theta'}) \geq E_L(P_\theta, P_{\theta'}) \geq \min_{i=1,2} \left\{ \frac{\delta_i^2}{2\gamma_i} - \frac{\delta_i^3}{6\gamma_i^2(1+\gamma_i)} \right\}. \quad (214)$$

The left-hand side of (214) holds since E_L is a lower bound on the error exponent, and the exact value of this error exponent is the Chernoff information. The right-hand side of (214) follows from Lemma 3 (see (141)) and the definition of E_L in (145). By definition $\gamma_i \triangleq \frac{\sigma_i^2}{d_i^2}$ and $\delta_i \triangleq \frac{\varepsilon_i}{d_i}$ where, based on (131),

$$\varepsilon_1 \triangleq D(P_\theta || P_{\theta'}), \quad \varepsilon_2 \triangleq D(P_{\theta'} || P_\theta). \quad (215)$$

The term on the left-hand side of (214) therefore satisfies

$$\begin{aligned}
& \frac{\delta_i^2}{2\gamma_i} - \frac{\delta_i^3}{6\gamma_i^2(1+\gamma_i)} \\
& = \frac{\varepsilon_i^2}{2\sigma_i^2} - \frac{\varepsilon_i^3 d_i^3}{6\sigma_i^2(\sigma_i^2 + d_i^2)} \\
& \geq \frac{\varepsilon_i^2}{2\sigma_i^2} \left(1 - \frac{\varepsilon_i d_i}{3} \right)
\end{aligned}$$

so it follows from (214) and the last inequality that

$$C(P_\theta, P_{\theta'}) \geq E_L(P_\theta, P_{\theta'}) \geq \min_{i=1,2} \left\{ \frac{\varepsilon_i^2}{2\sigma_i^2} \left(1 - \frac{\varepsilon_i d_i}{3} \right) \right\}. \quad (216)$$

Based on the continuity assumption of the indexed family $\{P_\theta\}_{\theta \in \Theta}$, then it follows from (215) that

$$\lim_{\theta' \rightarrow \theta} \varepsilon_i = 0, \quad \forall i \in \{1, 2\}$$

and also, from (112) and (122) with P_1 and P_2 replaced by P_θ and $P_{\theta'}$ respectively, then

$$\lim_{\theta' \rightarrow \theta} d_i = 0, \quad \forall i \in \{1, 2\}.$$

It therefore follows from (144) and (216) that

$$\frac{J(\theta)}{8} \geq \lim_{\theta' \rightarrow \theta} \frac{E_L(P_\theta, P_{\theta'})}{(\theta - \theta')^2} \geq \lim_{\theta' \rightarrow \theta} \min_{i=1,2} \left\{ \frac{\varepsilon_i^2}{2\sigma_i^2(\theta - \theta')^2} \right\}. \quad (217)$$

The idea is to show that the limit on the right-hand side of this inequality is $\frac{J(\theta)}{8}$ (same as the left-hand side), and hence, the limit of the middle term is also $\frac{J(\theta)}{8}$.

$$\begin{aligned} & \lim_{\theta' \rightarrow \theta} \frac{\varepsilon_1^2}{2\sigma_1^2(\theta - \theta')^2} \\ & \stackrel{(a)}{=} \lim_{\theta' \rightarrow \theta} \frac{D(P_\theta \| P_{\theta'})^2}{2\sigma_1^2(\theta - \theta')^2} \\ & \stackrel{(b)}{=} \frac{J(\theta)}{4} \lim_{\theta' \rightarrow \theta} \frac{D(P_\theta \| P_{\theta'})}{\sigma_1^2} \\ & \stackrel{(c)}{=} \frac{J(\theta)}{4} \lim_{\theta' \rightarrow \theta} \frac{D(P_\theta \| P_{\theta'})}{\sum_{x \in \mathcal{X}} P_\theta(x) \left(\ln \frac{P_\theta(x)}{P_{\theta'}(x)} - D(P_\theta \| P_{\theta'}) \right)^2} \\ & \stackrel{(d)}{=} \frac{J(\theta)}{4} \lim_{\theta' \rightarrow \theta} \frac{D(P_\theta \| P_{\theta'})}{\sum_{x \in \mathcal{X}} P_\theta(x) \left(\ln \frac{P_\theta(x)}{P_{\theta'}(x)} \right)^2 - D(P_\theta \| P_{\theta'})^2} \\ & \stackrel{(e)}{=} \frac{J(\theta)^2}{8} \lim_{\theta' \rightarrow \theta} \frac{(\theta - \theta')^2}{\sum_{x \in \mathcal{X}} P_\theta(x) \left(\ln \frac{P_\theta(x)}{P_{\theta'}(x)} \right)^2 - D(P_\theta \| P_{\theta'})^2} \\ & \stackrel{(f)}{=} \frac{J(\theta)^2}{8} \lim_{\theta' \rightarrow \theta} \frac{(\theta - \theta')^2}{\sum_{x \in \mathcal{X}} P_\theta(x) \left(\ln \frac{P_\theta(x)}{P_{\theta'}(x)} \right)^2} \\ & \stackrel{(g)}{=} \frac{J(\theta)}{8} \end{aligned} \quad (218)$$

where equality (a) follows from (215), equalities (b), (e) and (f) follow from (143), equality (c) follows from (113) with $P_1 = P_\theta$ and $P_2 = P_{\theta'}$, equality (d) follows from the definition of the divergence, and equality (g) follows by calculus (the required limit is calculated by using L'Hôpital's rule twice) and from the definition of Fisher information in (142). Similarly, also

$$\lim_{\theta' \rightarrow \theta} \frac{\varepsilon_2^2}{2\sigma_2^2(\theta - \theta')^2} = \frac{J(\theta)}{8}$$

so

$$\lim_{\theta' \rightarrow \theta} \min_{i=1,2} \left\{ \frac{\varepsilon_i^2}{2\sigma_i^2(\theta - \theta')^2} \right\} = \frac{J(\theta)}{8}.$$

Hence, it follows from (217) that $\lim_{\theta' \rightarrow \theta} \frac{E_L(P_\theta, P_{\theta'})}{(\theta - \theta')^2} = \frac{J(\theta)}{8}$. This completes the proof of (146).

We prove now equation (148). From (112), (122), (131) and (147) then

$$\tilde{E}_L(P_\theta, P_{\theta'}) = \min_{i=1,2} \frac{\varepsilon_i^2}{2d_i^2}$$

with ε_1 and ε_2 in (215). Hence,

$$\lim_{\theta' \rightarrow \theta} \frac{\tilde{E}_L(P_\theta, P_{\theta'})}{(\theta' - \theta)^2} \leq \lim_{\theta' \rightarrow \theta} \frac{\varepsilon_1^2}{2d_1^2(\theta' - \theta)^2}$$

and from (218) and the last inequality, it follows that

$$\begin{aligned} & \lim_{\theta' \rightarrow \theta} \frac{\tilde{E}_L(P_\theta, P_{\theta'})}{(\theta' - \theta)^2} \\ & \leq \frac{J(\theta)}{8} \lim_{\theta' \rightarrow \theta} \frac{\sigma_1^2}{d_1^2} \\ & \stackrel{(a)}{=} \frac{J(\theta)}{8} \lim_{\theta' \rightarrow \theta} \frac{\sum_{x \in \mathcal{X}} P_\theta(x) \left(\ln \frac{P_\theta(x)}{P_{\theta'}(x)} - D(P_\theta \| P_{\theta'}) \right)^2}{\left(\max_{x \in \mathcal{X}} \left| \ln \frac{P_\theta(x)}{P_{\theta'}(x)} - D(P_\theta \| P_{\theta'}) \right| \right)^2}. \end{aligned} \quad (219)$$

It is clear that the second term on the right-hand side of (219) is bounded between zero and one (if the limit exists). This limit can be made arbitrarily small, i.e., there exists an indexed family of probability mass functions $\{P_\theta\}_{\theta \in \Theta}$ for which the second term on the right-hand side of (219) can be made arbitrarily close to zero. For a concrete example, let $\alpha \in (0, 1)$ be fixed, and $\theta \in \mathbb{R}^+$ be a parameter that defines the following indexed family of probability mass functions over the ternary alphabet $\mathcal{X} = \{0, 1, 2\}$:

$$P_\theta(0) = \frac{\theta(1 - \alpha)}{1 + \theta}, \quad P_\theta(1) = \alpha, \quad P_\theta(2) = \frac{1 - \alpha}{1 + \theta}.$$

Then, it follows by calculus that for this indexed family

$$\lim_{\theta' \rightarrow \theta} \frac{\sum_{x \in \mathcal{X}} P_\theta(x) \left(\ln \frac{P_\theta(x)}{P_{\theta'}(x)} - D(P_\theta \| P_{\theta'}) \right)^2}{\left(\max_{x \in \mathcal{X}} \left| \ln \frac{P_\theta(x)}{P_{\theta'}(x)} - D(P_\theta \| P_{\theta'}) \right| \right)^2} = (1 - \alpha)\theta$$

so, for any $\theta \in \mathbb{R}^+$, the above limit can be made arbitrarily close to zero by choosing α close enough to 1. This completes the proof of (148), and also the proof of Proposition 6.

APPENDIX L

PROOF OF THE PROPERTIES IN (175) FOR OFDM SIGNALS

Consider an OFDM signal from Section V-C. The sequence in (173) is a martingale due to basic properties of martingales. From (172), for every $i \in \{0, \dots, n\}$

$$Y_i = \mathbb{E} \left[\max_{0 \leq t \leq T} |s(t; X_0, \dots, X_{n-1})| \mid X_0, \dots, X_{i-1} \right].$$

The conditional expectation for the RV Y_{i-1} refers to the case where only X_0, \dots, X_{i-2} are revealed. Let X'_{i-1} and X_{i-1} be independent copies, which are also independent of $X_0, \dots, X_{i-2}, X_i, \dots, X_{n-1}$. Then, for every $1 \leq i \leq n$,

$$\begin{aligned} Y_{i-1} &= \mathbb{E} \left[\max_{0 \leq t \leq T} |s(t; X_0, \dots, X'_{i-1}, X_i, \dots, X_{n-1})| \mid X_0, \dots, X_{i-2} \right] \\ &= \mathbb{E} \left[\max_{0 \leq t \leq T} |s(t; X_0, \dots, X'_{i-1}, X_i, \dots, X_{n-1})| \mid X_0, \dots, X_{i-2}, X_{i-1} \right]. \end{aligned}$$

Since $|\mathbb{E}(Z)| \leq \mathbb{E}(|Z|)$, then for $i \in \{1, \dots, n\}$

$$|Y_i - Y_{i-1}| \leq \mathbb{E}_{X'_{i-1}, X_i, \dots, X_{n-1}} \left[|U - V| \mid X_0, \dots, X_{i-1} \right] \quad (220)$$

where

$$\begin{aligned} U &\triangleq \max_{0 \leq t \leq T} |s(t; X_0, \dots, X_{i-1}, X_i, \dots, X_{n-1})| \\ V &\triangleq \max_{0 \leq t \leq T} |s(t; X_0, \dots, X'_{i-1}, X_i, \dots, X_{n-1})|. \end{aligned}$$

From (170)

$$\begin{aligned}
|U - V| &\leq \max_{0 \leq t \leq T} |s(t; X_0, \dots, X_{i-1}, X_i, \dots, X_{n-1}) - s(t; X_0, \dots, X'_{i-1}, X_i, \dots, X_{n-1})| \\
&= \max_{0 \leq t \leq T} \frac{1}{\sqrt{n}} \left| (X_{i-1} - X'_{i-1}) \exp\left(\frac{j 2\pi i t}{T}\right) \right| \\
&= \frac{|X_{i-1} - X'_{i-1}|}{\sqrt{n}}.
\end{aligned} \tag{221}$$

By assumption, $|X_{i-1}| = |X'_{i-1}| = 1$, and therefore a.s.

$$|X_{i-1} - X'_{i-1}| \leq 2 \implies |Y_i - Y_{i-1}| \leq \frac{2}{\sqrt{n}}.$$

In the following, an upper bound on the conditional variance $\text{Var}(Y_i | \mathcal{F}_{i-1}) = \mathbb{E}[(Y_i - Y_{i-1})^2 | \mathcal{F}_{i-1}]$ is obtained. Since $(\mathbb{E}(Z))^2 \leq \mathbb{E}(Z^2)$ for a real-valued RV Z , then from (220) and (221)

$$\mathbb{E}[(Y_i - Y_{i-1})^2 | \mathcal{F}_{i-1}] \leq \frac{1}{n} \cdot \mathbb{E}_{X'_{i-1}}[|X_{i-1} - X'_{i-1}|^2 | \mathcal{F}_i]$$

where \mathcal{F}_i is the σ -algebra that is generated by X_0, \dots, X_{i-1} . Due to symmetry of the PSK constellation, it follows that

$$\begin{aligned}
&\mathbb{E}[(Y_i - Y_{i-1})^2 | \mathcal{F}_{i-1}] \\
&\leq \frac{1}{n} \mathbb{E}_{X'_{i-1}}[|X_{i-1} - X'_{i-1}|^2 | \mathcal{F}_i] \\
&= \frac{1}{n} \mathbb{E}[|X_{i-1} - X'_{i-1}|^2 | X_0, \dots, X_{i-1}] \\
&= \frac{1}{n} \mathbb{E}[|X_{i-1} - X'_{i-1}|^2 | X_{i-1}] \\
&= \frac{1}{n} \mathbb{E}[|X_{i-1} - X'_{i-1}|^2 | X_{i-1} = e^{\frac{j\pi}{M}}] \\
&= \frac{1}{nM} \sum_{l=0}^{M-1} \left| e^{\frac{j\pi}{M}} - e^{\frac{j(2l+1)\pi}{M}} \right|^2 \\
&= \frac{4}{nM} \sum_{l=1}^{M-1} \sin^2\left(\frac{\pi l}{M}\right) = \frac{2}{n}
\end{aligned}$$

To clarify the last equality, note that if $x \in \mathbb{R}$ and $m \in \mathbb{N}$

$$\begin{aligned}
\sum_{l=1}^m \sin^2(lx) &= \frac{1}{2} \sum_{l=0}^m (1 - \cos(2lx)) \\
&= \frac{m+1}{2} - \frac{1}{2} \text{Re} \left\{ \sum_{l=0}^m e^{j2lx} \right\} \\
&= \frac{m+1}{2} - \frac{1}{2} \text{Re} \left\{ \frac{1 - e^{2j(m+1)x}}{1 - e^{2jx}} \right\} \\
&= \frac{m+1}{2} - \frac{1}{2} \text{Re} \left\{ \frac{(e^{j(m+1)x} - e^{-j(m+1)x}) e^{jmx}}{e^{jx} - e^{-jx}} \right\} \\
&= \frac{m+1}{2} - \frac{1}{2} \text{Re} \left\{ \frac{\sin((m+1)x) e^{jmx}}{\sin(x)} \right\} \\
&= \frac{m+1}{2} - \frac{\sin((m+1)x) \cos(mx)}{2 \sin(x)}
\end{aligned}$$

so, it follows that $\sum_{l=1}^{M-1} \sin^2\left(\frac{\pi l}{M}\right) = \frac{M}{2}$.

ACKNOWLEDGMENT

The feedback of I. Emre Telatar on the manuscript, which in particular motivated the author to refine the analysis in Section IV-E, is gratefully acknowledged. The author thanks Nicholas Kalouptsidis for suggesting Remark 13, and also for shortening the proof in Appendix B. Yuri Polyanskiy and Sergio Verdú are acknowledged for notifying the author on their work in [43]. The author is thankful to Nicholas Kalouptsidis, Leor Kehaty, and Kostis Xenoulis for detecting some typos. The hospitality of the Bernoulli inter-faculty center at EPFL (in Lausanne, Switzerland) during August 2011 is gratefully acknowledged.

REFERENCES

- [1] E. Abbe, *Local to Global Geometric Methods in Information Theory*, Ph.D. dissertation, MIT, Boston, MA, USA, June 2008.
- [2] E. Abbe, “Polar martingales of maximal spread,” to be presented at the *2012 International Zurich Seminar (IZS 2012)*, Zurich, Switzerland, March 2012.
- [3] N. Alon and J. H. Spencer, *The Probabilistic Method*, Wiley Series in Discrete Mathematics and Optimization, Third Edition, 2008.
- [4] Y. Altuğ and A. B. Wagner, “Moderate deviations analysis of channel coding: discrete memoryless case,” *Proceedings of the 2010 IEEE International Symposium on Information Theory*, pp. 265–269, Austin, Texas, USA, June 2010.
- [5] E. Arikan, “Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels,” *IEEE Trans. on Information Theory*, vol. 55, no. 7, pp. 3051–3073, July 2009.
- [6] E. Arikan and E. Telatar, “On the rate of channel polarization,” *Proceedings of the 2009 IEEE International Symposium on Information Theory*, pp. 1493–1495, Seoul, South Korea, July 2009 (extended version available in http://arxiv.org/PS_cache/arxiv/pdf/0807/0807.3806v3.pdf).
- [7] K. Azuma, “Weighted sums of certain dependent random variables,” *Tohoku Mathematical Journal*, vol. 19, pp. 357–367, 1967.
- [8] S. Benedetto and E. Biglieri, *Principles of Digital Transmission with Wireless Applications*, Kluwer Academic/Plenum Publishers, 1999.
- [9] G. Bennett, “Probability inequalities for the sum of independent random variables,” *Journal of the American Statistical Association*, vol. 57, no. 297, pp. 33–45, March 1962.
- [10] P. Billingsley, *Probability and Measure*, Wiley Series in Probability and Mathematical Statistics, Third Edition, 1995.
- [11] R. E. Blahut, “Hypothesis testing and information theory,” *IEEE Trans. on Information Theory*, vol. 20, no. 4, pp. 405–417, July 1974.
- [12] F. Chung and L. Lu, *Complex Graphs and Networks*, *Regional Conference Series in Mathematics*, vol. 107, 2006.
- [13] F. Chung and L. Lu, “Concentration inequalities and martingale inequalities: a survey,” *Internet Mathematics*, vol. 3, no. 1, pp. 79–127, March 2006. [Online]. Available: <http://www.ucsd.edu/~fan/wp/concen.pdf>.
- [14] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey and D. E. Knuth, “On the Lambert W function,” *Advances in Computational Mathematics*, vol. 5, pp. 329–359, May 1996.
- [15] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley and Sons, second edition, 2006.
- [16] I. Csiszár and P. C. Shields, *Information Theory and Statistics: A Tutorial*, *Foundations and Trends in Communications and Information Theory*, vol. 1, no. 4, pp. 417–528, 2004.
- [17] A. Dembo, “Moderate deviations for martingales with bounded jumps,” *Electronic Communications in Probability*, vol. 1, no. 3, pp. 11–17, March 1996.
- [18] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, Springer, second edition, 1997.
- [19] D. P. Dubashi and A. Panconesi, *Concentration of Measure for the Analysis of Randomized Algorithms*, Cambridge University Press, 2009.
- [20] T. Etzion, A. Trachtenberg and A. Vardy, “Which codes have cycle-free Tanner graphs ?,” *IEEE Trans. on Information Theory*, vol. 45, no. 6, pp. 2173–2181, September 1999.
- [21] D. Freedman, “On tail probabilities for martingales,” *Annals of Probability*, vol. 3, no. 1, pp. 100–118, January 1975.

- [22] R. G. Gallager, *Low-Density Parity-Check Codes*, Cambridge, MA: MIT Press, 1963.
- [23] G. Grimmett and D. Stirzaker, *Probability and Random Processes*, Oxford University Press, third edition, 2001.
- [24] D. He, L. A. Lastras-Montaño, E. Yang, A. Jagmohan and J. Chen, “On the redundancy of Slepian-Wolf coding,” *IEEE Trans. on Information Theory*, vol. 55, no. 12, pp. 5607–5627, December 2009.
- [25] M. Hayashi, “Information spectrum approach to second-order coding rate in channel coding,” *IEEE Trans. on Information Theory*, vol. 55, no. 11, pp. 4947–4966, November 2009.
- [26] W. Hoeffding, “Probability inequalities for sums of bounded random variables,” *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, March 1963.
- [27] F. den Hollander, *Large Deviations*, Fields Institute Monographs, American Mathematical Society, 2000.
- [28] C. P. Kistos and N. K. Tavoularis, “Logarithmic Sobolev inequalities for information measures,” *IEEE Trans. on Information Theory*, vol. 55, no. 6, pp. 2554–2561, June 2009.
- [29] I. Kontoyiannis and M. Madiman, “Entropy, compound Poisson approximation, log-Sobolev inequalities and measure concentration,” *Proceedings of the 2004 IEEE Information Theory Workshop*, pp. 71–75, San Antonio, Texas, USA, October 2004.
- [30] I. Kontoyiannis, L. A. Lastras-Montaño and S. P. Meyn, “Relative entropy and exponential deviation bounds for general Markov chains,” *Proceedings of the 2005 IEEE International Symposium on Information Theory*, pp. 1563–1567, Adelaide, Australia, September 2005.
- [31] S. B. Korada and N. Macris, “On the concentration of the capacity for a code division multiple access system,” *Proceedings of the 2007 IEEE International Symposium on Information Theory*, pp. 2801–2805, Nice, France, June 2007.
- [32] S. B. Korada and N. Macris, “Tight bounds on the capacity of binary input random CDMA systems,” *IEEE Trans. on Information Theory*, vol. 56, no. 11, pp. 5590–5613, November 2010.
- [33] M. Ledoux, *The Concentration of Measure Phenomenon*, Mathematical Surveys and Monographs, vol. 89, American Mathematical Society, 2001.
- [34] S. Litsyn and G. Wunder, “Generalized bounds on the crest-factor distribution of OFDM signals with applications to code design,” *IEEE Trans. on Information Theory*, vol. 52, pp. 992–1006, March 2006.
- [35] M. G. Luby, M. Mitzenmacher, M. A. Shokrollahi and D. A. Spielman, “Efficient erasure-correcting codes,” *IEEE Trans. on Information Theory*, vol. 47, no. 2, pp. 569–584, February 2001.
- [36] G. Lugosi, *Concentration of Measure Inequalities*, lecture notes, June 2009. [Online]. Available: <http://www.econ.upf.edu/~lugosi/anu.pdf>.
- [37] P. Massart, *Concentration Inequalities and Model Selection*, Lecture Notes in Mathematics, vol. 1896, Springer, 2007.
- [38] C. McDiarmid, “Concentration,” *Probabilistic Methods for Algorithmic Discrete Mathematics*, pp. 195–248, Springer, 1998.
- [39] C. McDiarmid, “Centering sequences with bounded differences,” *Combinatorics, Probability and Computing*, vol. 6, no. 1, pp. 79–86, March 1997.
- [40] A. F. Molisch, *Wireless Communications*, John Wiley and Sons, 2005.
- [41] B. Nakiboğlu, *Exponential Bounds on Error Probability with Feedback*, PhD dissertation, MIT, Boston, USA, February 2011.
- [42] Y. Polyanskiy, H. V. Poor, and S. Verdú, “Channel coding rate in finite blocklength regime,” *IEEE Trans. on Information Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.
- [43] Y. Polyanskiy and S. Verdú, “Channel dispersion and moderate deviations limits of memoryless channels,” *Proceedings Forty-Eighth Annual Allerton Conference*, pp. 1334–1339, Illinois, USA, October 2010.
- [44] Y. Polyanskiy, H. V. Poor, and S. Verdú, “Feedback in the non-asymptotic regime,” *IEEE Trans. on Information Theory*, vol. 57, no. 8, pp. 4903–4925, August 2011.
- [45] A. Rényi, “On measures of entropy and information,” *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 547–561, California, USA, 1961.
- [46] T. J. Richardson and R. Urbanke, “The capacity of low-density parity-check codes under message-passing decoding,” *IEEE Trans. on Information Theory*, vol. 47, no. 2, pp. 599–618, February 2001.
- [47] T. J. Richardson and R. Urbanke, *Modern Coding Theory*, Cambridge University Press, 2008.

- [48] R. Salem and A. Zygmund, "Some properties of trigonometric series whose terms have random signs," *Acta Mathematica*, vol. 91, no. 1, pp. 245–301, 1954.
- [49] I. Sason, "On universal properties of capacity-approaching LDPC code ensembles," *IEEE Trans. on Information Theory*, vol. 55, no. 7, pp. 2956–2990, July 2009.
- [50] I. Sason, "Tightened exponential bounds for discrete-time, conditionally symmetric martingales with bounded jumps," submitted to the *Electronic Journal of Probability*, January 2012. [Online]. Available: <http://arxiv.org/abs/1201.0533>.
- [51] E. Shamir and J. Spencer, "Sharp concentration of the chromatic number on random graphs," *Combinatorica*, vol. 7, no. 1, pp. 121–129, 1987.
- [52] M. Sipser and D. A. Spielman, "Expander codes," *IEEE Trans. on Information Theory*, vol. 42, no. 6, pp. 1710–1722, November 1996.
- [53] W. L. Steiger, "A best possible Kolmogoroff-type inequality for martingales and a characteristic property," *Annals of Mathematical Statistics*, vol. 40, no. 3, pp. 764–769, June 1969.
- [54] M. Talagrand, "Concentration of measure and isoperimetric inequalities in product spaces," *Publications Mathématiques de l'I.H.E.S.*, vol. 81, pp. 73–205, 1995.
- [55] M. Talagrand, "A new look at independence," *Annals of Probability*, vol. 24, no. 1, pp. 1–34, January 1996.
- [56] V. Y. F. Tan, "Moderate-deviations of lossy source coding for discrete and Gaussian sources," <http://arxiv.org/abs/1111.2217>, November 2011.
- [57] A. B. Wagner, P. Viswanath and S. R. Kulkarni, "Probability estimation in the rare-events regime," *IEEE Trans. on Information Theory*, vol. 57, no. 6, pp. 3207–3229, June 2011.
- [58] G. Wunder and H. Boche, "New results on the statistical distribution of the crest-factor of OFDM signals," *IEEE Trans. on Information Theory*, vol. 49, no. 2, pp. 488–494, February 2003.
- [59] K. Xenoulis and N. Kalouptsidis, "On the random coding exponent of nonlinear Gaussian channels," *Proceedings of the 2009 IEEE International Workshop on Information Theory*, pp. 32–36, Volos, Greece, June 2009.
- [60] K. Xenoulis and N. Kalouptsidis, "Achievable rates for nonlinear Volterra channels," *IEEE Trans. on Information Theory*, vol. 57, no. 3, pp. 1237–1248, March 2011.
- [61] K. Xenoulis, N. Kalouptsidis and I. Sason, "New achievable rates for nonlinear Volterra channels via martingale inequalities," submitted to the *2012 IEEE International Symposium on Information Theory (ISIT 2012)*, January 2012.