*IRWIN AND JOAN JACOBS*
*CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGIES*

# The Martingale Approach for Concentration and Applications in Information Theory, Communications and Coding

**Igal Sason**

**CCIT Report #815**
**October 2012**

Electronics
Computers
Communications

*DEPARTMENT OF ELECTRICAL ENGINEERING*
*TECHNION - ISRAEL INSTITUTE OF TECHNOLOGY, HAIFA 32000, ISRAEL*

CCIT Report #815 October 2012

# The Martingale Approach for Concentration and Applications in Information Theory, Communications and Coding

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

### Abstract

This chapter introduces some concentration inequalities for discrete-time martingales with bounded increments, and it exemplifies some of their potential applications in information theory and related topics. The first part of this chapter introduces briefly discrete-time martingales and the Azuma-Hoeffding & McDiardmid's inequalities which are widely used in this context. It then 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 inter-connections between them. The second part exemplifies the use of these refined inequalities in the context of hypothesis testing, information theory, communications, and coding. The chapter is concluded with a discussion on some directions for further research.

### 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 $\mathbb{P}(|X - \overline{x}| \geq t)$ (or $\mathbb{P}(X - \overline{x} \geq t)$ for a random variable (RV) $X$, where $\overline{x}$ 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" [75]. 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 $\overline{x}$ in a way that the probability of the event $\{|X - \overline{x}| > t\}$ decays exponentially in $t$ (for $t \geq 0$). The foundations in concentration of measures have been introduced, e.g., in [3, Chapter 7], [15, Chapter 2], [16], [42], [47], [48, Chapter 5], [50], [74] and [75]. Concentration inequalities are also at the core of probabilistic analysis of randomized algorithms (see, e.g., [3], [23], [50] and [61]).

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., [74] and [75] with some information-theoretic applications in [40] and [41]), logarithmic-Sobolev inequalities (e.g., [23, Chapter 14], [42, Chapter 5] and [47] with information-theoretic aspects in [37], [38]), transportation-cost inequalities which originated from information theory (e.g., [23, Chapters 12, 13] and [42, Chapter 6]), and the martingale approach (e.g., [3, Chapter 7], [50] with information-theoretic aspects in, e.g., [45], [60], [61], [81]). This chapter 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 [34] 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 [51]; 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 [39] 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 [50, Theorem 3.1]). This inequality was proved (with some possible extensions) via the martingale approach (see [50, 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 [50, Section 3].

The use of the Azuma-Hoeffding inequality was introduced to the computer science literature in [70] 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 [70] 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 [70] 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 [45], [60] and [72]. During the last decade, the Azuma-Hoeffding inequality has been extensively used for proving concentration of measures in coding theory (see, e.g., [61, 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 [77] 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 [80], 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 [81] 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 [81]). This direction of research was further studied in [82], and improved achievable rates have been derived via refined version of the Azuma-Hoeffding inequality.

This chapter is structured as follows: Section II presents briefly discrete-time (sub/ super) martingales, Section III presents the Azuma-Hoeffding inequality and McDiardmid's inequality; these are widely used in proving concentration, and their derivation relies on the martingale approach. Section IV derives some refined versions of the Azuma-Hoeffding inequality, and it considers interconnections between these bounds. Section V considers some connections between the concentration inequalities that are introduced in Section IV 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 VI forms the second part of this work, applying the concentration inequalities from Section IV to information theory and some related topics. This chapter is summarized in Section VII, 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.

In connection to the presentation in this chapter, the reader is referred to [3, Chapter 11], [15, Chapter 2], [16] and [50] as some additional surveys on concentration inequalities for (sub/ super) martingales.

## II. Discrete-Time Martingales

### *A. Martingales*

This sub-section provides a short background on martingales to set definitions and notation (the reader is referred, e.g., to [78] for a nice exposition of discrete-time martingales). 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 $X_0, X_1, \ldots$ of random variables (RVs) and corresponding sub $\sigma$-algebras $\mathcal{F}_0, \mathcal{F}_1, \ldots$ 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 $\Omega$, it is measurable with respect to the $\sigma$-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) $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots$ (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 $X_\infty$ and $\mathcal{F}_\infty$ 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 $i \in \mathbb{N}$, $\mathbb{E}[X_i] = \mathbb{E}\big[\mathbb{E}[X_i | \mathcal{F}_{i-1}]\big] = \mathbb{E}[X_{i-1}]$, 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 $\sigma$-algebras $\{\mathcal{F}_i\}_{i=0}^\infty$, let

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

Then, the sequence $X_0, X_1, \ldots$ 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 $i \in \mathbb{N}$

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

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

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

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. 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 $\mathbb{E}\big[|h(X_i)|\big] < \infty$, 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 $\mathbb{E}\big[|h(X_i)|\big] < \infty$, 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 $\mathbb{E}\big[|h(X_i)|\big] < \infty$, then $\{h(X_i), \mathcal{F}_i\}$ is a sub-martingale.

## III. Two Basic Concentration Inequalities

In the following section, we prove two basic inequalities that are widely used for proving concentration inequalities. Their proofs conveys the main concepts of the martingale approach for proving concentration results. Their presentation also motivates some refinements that are considered later in this chapter, followed by some applications.

### A. The Azuma-Hoeffding Inequality

The Azuma-Hoeffding inequality[1] is a useful concentration inequality for bounded-difference martingales. It was proved in [34] 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.

*Theorem 1:* **[Azuma-Hoeffding 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 $\alpha > 0$,

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

The proof of the Azuma-Hoeffding inequality serves also to present the basic principles on which the martingale approach for proving concentration results is based on. Therefore, we present in the following the proof of this inequality.

*Proof:* For an arbitrary $\alpha > 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha) = \mathbb{P}(X_n - X_0 \geq \alpha) + \mathbb{P}(X_n - X_0 \leq -\alpha). \tag{2}$$

Let $\xi_i \triangleq X_i - X_{i-1}$ for $i = 1, \ldots, n$ designate the jumps of the martingale. Then, it follows by assumption that $|\xi_k| \leq d_k$ and $\mathbb{E}[\xi_k \,|\, \mathcal{F}_{k-1}] = 0$ a.s. for every $k \in \{1, \ldots, n\}$.

From Chernoff's inequality, for every $t \geq 0$,

$$\begin{aligned} &\mathbb{P}(X_n - X_0 \geq \alpha) \\ &= \mathbb{P}\big(\sum_{i=1}^{n} \xi_i \geq \alpha\big) \\ &\leq e^{-\alpha t}\, \mathbb{E}\big[e^{t \sum_{i=1}^{n} \xi_i}\big]. \end{aligned} \tag{3}$$

For every $t \geq 0$

$$\begin{aligned} &\mathbb{E}\left[\exp\left(t \sum_{k=1}^{n} \xi_k\right)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\exp\left(t \sum_{k=1}^{n} \xi_k\right) \,|\, \mathcal{F}_{n-1}\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\exp\left(t \sum_{k=1}^{n-1} \xi_k\right)\, \exp(t\xi_n) \,|\, \mathcal{F}_{n-1}\right]\right] \\ &= \mathbb{E}\left[\exp\left(t \sum_{k=1}^{n-1} \xi_k\right) \mathbb{E}\big[\exp(t\xi_n) \,|\, \mathcal{F}_{n-1}\big]\right] \end{aligned} \tag{4}$$

[1] The Azuma-Hoeffding inequality is also known as Azuma's inequality. Since it is referred numerous times in this chapter, it will be named Azuma's inequality for the sake of brevity.

where the last transition holds since $Y = \exp\bigl(t \sum_{k=1}^{n-1} \xi_k\bigr)$ 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. $\mathbb{E}[XY|\mathcal{F}_{n-1}] = Y\,\mathbb{E}[X|\mathcal{F}_{n-1}]$.

Due to the convexity of the exponential function, and since $|\xi_k| \leq d_k$, then the straight line connecting the end points of the exponential function is below this function over the interval $[-d_k, d_k]$, so a.s. for every $k$

$$\begin{aligned}
&\mathbb{E}\bigl[e^{t\xi_k} \,|\, \mathcal{F}_{k-1}\bigr] \\
&\leq \mathbb{E}\Bigl[\frac{(d_k + \xi_k)e^{td_k} + (\xi_k - d_k)e^{-td_k}}{2d_k} \,|\, \mathcal{F}_{k-1}\Bigr] \\
&= \frac{1}{2}\bigl(e^{td_k} + e^{-td_k}\bigr) \\
&= \cosh(td_k).
\end{aligned} \tag{5}$$

Since, for every integer $m \geq 0$,

$$(2m)! \geq (2m)(2m-2)\ldots 2 = 2^m\, m!$$

then due to the power series expansion of the hyperbolic cosine and exponential functions, we have

$$\cosh(td_k) = \sum_{m=0}^{\infty} \frac{(td_k)^{2m}}{(2m)!} \leq \sum_{m=0}^{\infty} \frac{(td_k)^{2m}}{2^m\, m!} = e^{\frac{t^2\, d_k^2}{2}}$$

which therefore implies that

$$\mathbb{E}\bigl[e^{t\xi_k} \,|\, \mathcal{F}_{k-1}\bigr] \leq e^{\frac{t^2\, d_k^2}{2}}.$$

Hence, by repeatedly using the recursion in (4), it follows that

$$\mathbb{E}\left[\exp\left(t \sum_{k=1}^{n} \xi_k\right)\right] \leq \prod_{k=1}^{n} \exp\left(\frac{t^2\, d_k^2}{2}\right) = \exp\left(\frac{t^2}{2} \sum_{k=1}^{n} d_k^2\right) \tag{6}$$

which then gives from (3) that, for every $t \geq 0$,

$$\mathbb{P}(X_n - X_0 \geq \alpha) \leq \exp\left(-\alpha t + \frac{t^2}{2} \sum_{k=1}^{n} d_k^2\right). \tag{7}$$

An optimization over the free parameter $t \geq 0$ gives that

$$\mathbb{P}(X_n - X_0 \geq \alpha) \leq \exp\left(-\frac{\alpha^2}{2\sum_{k=1}^{n} d_k^2}\right). \tag{8}$$

Since, by assumption, $\{X_k, \mathcal{F}_k\}$ is a martingale with bounded jumps, so is $\{-X_k, \mathcal{F}_k\}$ (with the same bounds on its jumps). This implies that the same bound is also valid for the probability $\mathbb{P}(X_n - X_0 \leq -\alpha)$ and together with (2) it completes the proof of the Azuma-Hoeffding inequality. ■

The proof of the Azuma-Hoeffding inequality will be revisited later in this chapter for the derivation of some refined versions of Azuma's inequality, whose use and advantage will be also exemplified.

*Remark 4:* In [50, 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 $r \geq 0$,

$$\mathbb{P}\left(\left|\sum_{k=1}^{n} Y_k\right| \geq r\right) \leq 2\exp\left(-\frac{2r^2}{\sum_{k=1}^{n}(b_k - a_k)^2}\right). \tag{9}$$

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 (9), it follows that for every $r \geq 0$,

$$\mathbb{P}\left(|X_n - X_0| \geq r\right) \leq 2\exp\left(-\frac{2r^2}{\sum_{k=1}^{n}(b_k - a_k)^2}\right). \tag{10}$$

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 (10).

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 $\alpha \geq 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) \leq 2\exp\left(-\frac{\alpha^2}{2d^2}\right). \tag{11}$$

*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 $k \in \{0, 1, \ldots, \}$, and define the natural filtration $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \ldots$ where

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

is the $\sigma$-algebra that is generated by the random variables $Y_0, \ldots, Y_k$. 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 (11) that

$$\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) \leq 2\exp\left(-\frac{\alpha^2}{2d^2}\right). \tag{12}$$

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 $\alpha \geq 0$,

$$\lim_{n \to \infty} \mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) = 2\, Q\Big(\frac{\alpha}{d}\Big) \tag{13}$$

where

$$Q(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_x^{\infty} \exp\Big(-\frac{t^2}{2}\Big) \mathrm{d}t, \quad \forall\, x \in \mathbb{R} \tag{14}$$

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

$$\frac{1}{\sqrt{2\pi}} \frac{x}{1+x^2} \cdot e^{-\frac{x^2}{2}} < Q(x) < \frac{1}{\sqrt{2\pi}\, x} \cdot e^{-\frac{x^2}{2}}, \;\; \forall\, x > 0 \tag{15}$$

then it follows from (13) that the exponent on the right-hand side of (12) 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 $\alpha \geq 0$

$$\lim_{n \to \infty} \mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) = 2\, Q\left(\frac{\alpha}{\sqrt{\gamma}\, d}\right). \tag{16}$$

From the exponential upper and lower bounds of the Q-function in (15), the right-hand side of (16) 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 IV.

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

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

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., $d_k = d$).

### *B. McDiardmid's Inequality*

The following useful inequality is due to McDiardmid ([49] or [51, Theorem 3.1]), and its original derivation uses the martingale approach for its derivation. We will relate, in the following, the derivation of this inequality to the derivation of the Azuma-Hoeffding inequality (see the previous sub-section).

*Theorem 2:* **[McDiardmid's inequality]** Let $\{X_i\}$ be independent real-valued random variables (not necessarily i.i.d.), and assume that $X_i : \Omega_i \rightarrow \mathbb{R}$ for every $i$. Let $\{\hat{X}_i\}_{i=1}^n$ be independent copies of $\{X_i\}_{i=1}^n$, respectively, and suppose that, for every $k \in \{1, \ldots, n\}$,

$$\big|g(X_1, \ldots, X_{k-1}, X_k, X_{k+1}, \ldots, X_n) - g(X_1, \ldots, X_{k-1}, \hat{X}_k, X_{k+1}, \ldots, X_n)\big| \leq d_k \tag{17}$$

holds a.s. (note that a stronger condition would be to require that the variation of $g$ w.r.t. the $k$-th coordinate of $\underline{x} \in \mathbb{R}^n$ is upper bounded by $d_k$, i.e.,

$$\sup |g(\underline{x}) - g(\underline{x}')| \leq d_k$$

for every $\underline{x}, \underline{x}' \in \mathbb{R}^n$ that differ only in their $k$-th coordinate.) Then, for every $\alpha \geq 0$,

$$\mathbb{P}(\big|g(X_1, \ldots, X_n) - \mathbb{E}\big[g(X_1, \ldots, X_n)\big]\big| \geq \alpha) \leq 2 \exp\left(-\frac{2\alpha^2}{\sum_{k=1}^n d_k^2}\right). \tag{18}$$

*Remark 5:* As we will see from the proof of this inequality, one could use the Azuma-Hoeffding inequality for proving it, but then the exponent will be four times smaller (i.e., the factor 2 in the exponent would have appeared in the denominator instead of appearing in the numerator. Hence, it will be observed from the proof that in the current setting, one gets a gain of a factor of 4 in the exponent.

*Proof:* For $k \in \{1, \ldots, n\}$, let $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$ be the $\sigma$-algebra that is generated by $X_1, \ldots, X_k$ with $\mathcal{F}_0 = \{\emptyset, \Omega\}$ being the minimal sigma-algebra. Define

$$\xi_k \triangleq \mathbb{E}\big[g(X_1, \ldots, X_n) \,|\, \mathcal{F}_k\big] - \mathbb{E}\big[g(X_1, \ldots, X_n) \,|\, \mathcal{F}_{k-1}\big], \quad \forall\, k \in \{1, \ldots, n\}. \tag{19}$$

Note that $\mathcal{F}_0 \subseteq \mathcal{F}_1 \ldots \subseteq \mathcal{F}_n$ is a filtration,

$$\begin{aligned} \mathbb{E}\big[g(X_1, \ldots, X_n) \,|\, \mathcal{F}_0\big] &= \mathbb{E}\big[g(X_1, \ldots, X_n)\big] \\ \mathbb{E}\big[g(X_1, \ldots, X_n) \,|\, \mathcal{F}_n\big] &= g(X_1, \ldots, X_n). \end{aligned} \tag{20}$$

Hence

$$\begin{aligned} & g(X_1, \ldots, X_n) - \mathbb{E}\big[g(X_1, \ldots, X_n)\big] \\ & = \mathbb{E}\big[g(X_1, \ldots, X_n) \,|\, \mathcal{F}_n\big] - \mathbb{E}\big[g(X_1, \ldots, X_n) \,|\, \mathcal{F}_0\big] \\ & = \sum_{k=1}^n \big\{\mathbb{E}\big[g(X_1, \ldots, X_n) \,|\, \mathcal{F}_k\big] - \mathbb{E}\big[g(X_1, \ldots, X_n) \,|\, \mathcal{F}_{k-1}\big]\big\} \\ & = \sum_{k=1}^n \xi_k. \end{aligned} \tag{21}$$

In the following, we need the following lemma:

*Lemma 1:* For every $k \in \{1, \ldots, n\}$, the following properties hold a.s.:

1) $\mathbb{E}[\xi_k \,|\, \mathcal{F}_{k-1}] = 0$, so $\{\xi_k, \mathcal{F}_k\}$ is a martingale-difference and $\xi_k$ is $\mathcal{F}_k$-measurable.
2) $|\xi_k| \leq d_k$
3) $\xi_k \in [a_k, a_k + d_k]$ where $a_k$ is some non-positive $\mathcal{F}_{k-1}$-measurable.

*Proof:* The random variable $\xi_k$ is $\mathcal{F}_k$-measurable since $\mathcal{F}_{k-1} \subseteq \mathcal{F}_k$, and $\xi_k$ is a difference of two functions where one is $\mathcal{F}_k$-measurable and the other one is $\mathcal{F}_{k-1}$-measurable. Furthermore, it is easy to verify that $\mathbb{E}[\xi_k \,|\, \mathcal{F}_{k-1}] = 0$. This verifies the first item. the second item follows from the third item. To prove the third item, let

$$\begin{aligned} \xi_k &= \mathbb{E}\big[g(X_1, \ldots, X_{k-1}, X_k, X_{k+1}, \ldots, X_n) \,|\, \mathcal{F}_k\big] - \mathbb{E}\big[g(X_1, \ldots, X_{k-1}, X_k, X_{k+1}, \ldots, X_n) \,|\, \mathcal{F}_{k-1}\big] \\ \hat{\xi}_k &= \mathbb{E}\big[g(X_1, \ldots, X_{k-1}, \hat{X}_k, X_{k+1}, \ldots, X_n) \,|\, \hat{\mathcal{F}}_k\big] - \mathbb{E}\big[g(X_1, \ldots, X_{k-1}, X_k, X_{k+1}, \ldots, X_n) \,|\, \mathcal{F}_{k-1}\big] \end{aligned}$$

where $\{\hat{X}_i\}_{i=1}^n$ is an independent copy of $\{X_i\}_{i=1}^n$, and we define

$$\hat{\mathcal{F}}_k = \sigma(X_1, \ldots, X_{k-1}, \hat{X}_k).$$

Due to the independence of $X_k$ and $\hat{X}_k$, and since they are also independent of the other RVs then a.s.

$$\begin{aligned}
&|\xi_k - \hat{\xi}_k| \\
&= \big|\mathbb{E}\big[g(X_1, \ldots, X_{k-1}, X_k, X_{k+1}, \ldots, X_n) \,|\, \mathcal{F}_k\big] - \mathbb{E}\big[g(X_1, \ldots, X_{k-1}, \hat{X}_k, X_{k+1}, \ldots, X_n) \,|\, \hat{\mathcal{F}}_k\big]\big| \\
&= \big|\mathbb{E}\big[g(X_1, \ldots, X_{k-1}, X_k, X_{k+1}, \ldots, X_n) - g(X_1, \ldots, X_{k-1}, \hat{X}_k, X_{k+1}, \ldots, X_n) \,|\, \sigma(X_1, \ldots, X_{k-1}, X_k, \hat{X}_k)\big]\big| \\
&\leq \mathbb{E}\big[|g(X_1, \ldots, X_{k-1}, X_k, X_{k+1}, \ldots, X_n) - g(X_1, \ldots, X_{k-1}, \hat{X}_k, X_{k+1}, \ldots, X_n)| \,|\, \sigma(X_1, \ldots, X_{k-1}, X_k, \hat{X}_k)\big] \\
&\leq d_k. \qquad (22)
\end{aligned}$$

Therefore, $|\xi_k - \hat{\xi}_k| \leq d_k$ holds a.s. for every pair of independent copies $X_k$ and $\hat{X}_k$, which are also independent of the other random variables. This implies that $\xi_k$ is a.s. supported on an interval $[a_k, a_k + d_k]$ for some function $a_k = a_k(X_1, \ldots, X_{k-1})$ that is $\mathcal{F}_{k-1}$-measurable (since $X_k$ and $\hat{X}_k$ are independent copies, and $\xi_k - \hat{\xi}_k$ is a difference of $g(X_1, \ldots, X_{k-1}, X_k \ldots, X_n)$ and $g(X_1, \ldots, X_{k-1}, \hat{X}_k \ldots, X_n)$, then this is in essence saying that if a set $\mathcal{S} \subseteq \mathbb{R}$ has the property that the distance between any of its two points is not larger than some $d > 0$, then the set should be included in an interval whose length is $d$). Since also $\mathbb{E}[\xi_k \,|\, \mathcal{F}_{k-1}] = 0$ then a.s. the $\mathcal{F}_{k-1}$-measurable function $a_k$ is non-positive. It is noted that the third item of the lemma is what makes it different from the proof in the Azuma-Hoeffding inequality (which, in that case, implies that $\xi_k \in [-d_k, d_k]$ where the length of the interval is twice large (i.e., $2d_k$).) ■

Let $b_k \triangleq a_k + d_k$. Since $\mathbb{E}[\xi_k \,|\, \mathcal{F}_{k-1}] = 0$ and $\xi_k \in [a_k, b_k]$ with $a_k \leq 0$ and $b_k$ are $\mathcal{F}_{k-1}$-measurable, then

$$\text{Var}(\xi_k \,|\, \mathcal{F}_{k-1}) \leq -a_k b_k \triangleq \sigma_k^2.$$

Applying the convexity of the exponential function gives (similarly to the derivation of the Azuma-Hoeffding inequality, but this time w.r.t. the interval $[a_k, b_k]$ whose length is only $d_k$) that, for every $k \in \{1, \ldots, n\}$

$$\mathbb{E}[e^{t\xi_k} \,|\, \mathcal{F}_{k-1}] \leq \frac{b_k e^{ta_k} - a_k e^{tb_k}}{d_k}. \qquad (23)$$

Let $p_k \triangleq -\frac{a_k}{d_k} \in [0, 1]$, then

$$\begin{aligned}
&\mathbb{E}[e^{t\xi_k} \,|\, \mathcal{F}_{k-1}] \\
&\leq p_k e^{tb_k} + (1 - p_k) e^{ta_k} \\
&= e^{ta_k}\big[1 - p_k + p_k e^{td_k}\big] \\
&= e^{f_k(t)} \qquad (24)
\end{aligned}$$

where $f_k(t) \triangleq ta_k + \ln\big(1 - p_k + p_k e^{td_k}\big)$ for $t \in \mathbb{R}$. Since $f_k(0) = f_k'(0) = 0$ and the geometric mean is less than or equal to the arithmetic mean then, for every $t$,

$$f_k''(t) = \frac{d_k^2 p_k (1 - p_k) e^{td_k}}{(1 - p_k + p_k e^{td_k})^2} \leq \frac{d_k^2}{4}$$

which implies by Taylor's theorem that

$$f_k(t) \leq \frac{t^2 d_k^2}{8}$$

so, from (24),

$$\mathbb{E}[e^{t\xi_k} \,|\, \mathcal{F}_{k-1}] \leq e^{\frac{t^2 d_k^2}{8}}.$$

Similarly to the proof of the Azuma-Hoeffding inequality, by repeatedly using the recursion in (4), the last inequality implies that

$$\mathbb{E}\left[\exp\left(t \sum_{k=1}^n \xi_k\right)\right] \leq \exp\left(\frac{t^2}{8} \sum_{k=1}^n d_k^2\right) \qquad (25)$$

which then gives from (3) that, for every $t \geq 0$,

$$\begin{aligned}&\mathbb{P}(g(X_1, \ldots, X_n) - \mathbb{E}[g(X_1, \ldots, X_n)] \geq \alpha) \\ &= \mathbb{P}(\sum_{k=1}^{n} \xi_k \geq \alpha) \\ &\leq \exp\left(-\alpha t + \frac{t^2}{8} \sum_{k=1}^{n} d_k^2\right). \end{aligned} \tag{26}$$

An optimization over the free parameter $t \geq 0$ gives that

$$\mathbb{P}(g(X_1, \ldots, X_n) - \mathbb{E}[g(X_1, \ldots, X_n)] \geq \alpha) \leq \exp\left(-\frac{2\alpha^2}{\sum_{k=1}^{n} d_k^2}\right). \tag{27}$$

Similarly to the derivation of the Azuma-Hoeffding inequality, this bound is also valid for the probability

$$\mathbb{P}(g(X_1, \ldots, X_n) - \mathbb{E}[g(X_1, \ldots, X_n)] \geq \alpha),$$

which therefore gives the bound in (18). ■

## IV. Refined Versions of The Azuma-Hoeffding Inequality

### A. *First Refinement of Azuma's Inequality*

The following theorem appears in [49] and [21, Corollary 2.4.7].

*Theorem 3:* 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.

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

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

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) \leq 2 \exp\left(-n\, D\Big(\frac{\delta + \gamma}{1 + \gamma} \Big|\Big| \frac{\gamma}{1 + \gamma}\Big)\right) \tag{28}$$

where

$$\gamma \triangleq \frac{\sigma^2}{d^2}, \quad \delta \triangleq \frac{\alpha}{d} \tag{29}$$

and

$$D(p||q) \triangleq p \ln\Big(\frac{p}{q}\Big) + (1 - p) \ln\Big(\frac{1 - p}{1 - q}\Big), \quad \forall\, p, q \in [0, 1] \tag{30}$$

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 (28) is equal to zero.

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

*Proof:* The proof of this bound starts similarly to the proof of the Azuma-Hoeffding inequality, up to (4). The new ingredient in this proof is Bennett's inequality which replaces the argument of the convexity of the exponential function in the proof of the Azuma-Hoeffding inequality. From Bennett's inequality [10] (see, e.g., [21, Lemma 2.4.1]), if $X$ is a real-valued random variable with $\overline{x} = \mathbb{E}(X)$ and $\mathbb{E}[(X - \overline{x})^2] \leq \sigma^2$ for some $\sigma > 0$, and $X \leq b$ a.s. for some $b \in \mathbb{R}$, then for every $\lambda \geq 0$

$$\mathbb{E}\big[e^{\lambda X}\big] \leq \frac{e^{\lambda \overline{x}} \Big[(b - \overline{x})^2 \exp^{-\frac{\lambda \sigma^2}{b - \overline{x}}} + \sigma^2 e^{\lambda (b - \overline{x})}\Big]}{(b - \overline{x})^2 + \sigma^2}. \tag{31}$$

Applying Bennett's inequality for the conditional law of $\xi_k$ given the $\sigma$-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.

$$\mathbb{E}\left[\exp(t\xi_k) \,|\, \mathcal{F}_{k-1}\right] \leq \frac{\sigma^2 \exp(td) + d^2 \exp\left(-\frac{t\sigma^2}{d}\right)}{d^2 + \sigma^2}. \tag{32}$$

Hence, it follows from (4) and (32) 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 $\gamma$ in (29), this inequality is rewritten as

$$\mathbb{E}\left[\exp\left(t\sum_{k=1}^{n}\xi_k\right)\right] \leq \left(\frac{\gamma\exp(td) + \exp(-\gamma td)}{1+\gamma}\right)^n, \ \forall\, t \geq 0. \tag{33}$$

Let $x \triangleq td$ (so $x \geq 0$). Combining Chernoff's inequality with (33) gives that, for every $\alpha \geq 0$ (where from the definition of $\delta$ in (29), $\alpha t = \delta x$),

$$\begin{aligned}
&\mathbb{P}(X_n - X_0 \geq \alpha n)\\
&\leq \exp(-\alpha n t)\,\mathbb{E}\left[\exp\left(t\sum_{k=1}^{n}\xi_k\right)\right]\\
&\leq \left(\frac{\gamma\exp\bigl((1-\delta)x\bigr) + \exp\bigl(-(\gamma+\delta)x\bigr)}{1+\gamma}\right)^n, \ \forall\, x \geq 0.
\end{aligned} \tag{34}$$

Consider first the case where $\delta = 1$ (i.e., $\alpha = d$), then (34) is particularized to

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

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

$$\mathbb{P}(X_n - X_0 \geq dn) \leq \left(\frac{\gamma}{1+\gamma}\right)^n. \tag{35}$$

Otherwise, if $\delta \in [0,1)$, the minimization of the base of the exponent on the right-hand side of (34) w.r.t. the free non-negative parameter $x$ yields that the optimized value is

$$x = \left(\frac{1}{1+\gamma}\right)\ln\left(\frac{\gamma+\delta}{\gamma(1-\delta)}\right) \tag{36}$$

and its substitution into the right-hand side of (34) gives that, for every $\alpha \geq 0$,

$$\begin{aligned}
&\mathbb{P}(X_n - X_0 \geq \alpha n)\\
&\leq \left[\left(\frac{\gamma+\delta}{\gamma}\right)^{-\frac{\gamma+\delta}{1+\gamma}}(1-\delta)^{-\frac{1-\delta}{1+\gamma}}\right]^n\\
&= \exp\left\{-n\left[\left(\frac{\gamma+\delta}{1+\gamma}\right)\ln\left(\frac{\gamma+\delta}{\gamma}\right) + \left(\frac{1-\delta}{1+\gamma}\right)\ln(1-\delta)\right]\right\}\\
&= \exp\left(-n\,D\left(\frac{\delta+\gamma}{1+\gamma}\Big|\Big|\frac{\gamma}{1+\gamma}\right)\right)
\end{aligned} \tag{37}$$

and the exponent is equal to $+\infty$ if $\delta > 1$ (i.e., if $\alpha > d$). Applying inequality (37) 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 (28). This completes the proof of Theorem 3. ■

*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 $m \in \mathbb{N}$

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

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

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

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

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

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 $x \geq 0$,

$$\mathbb{P}(X_k \geq kx) \leq \exp\left(-\frac{kx^2}{2d^2}\right) \tag{38}$$

independently of the value of $\varepsilon$ (note that $X_0 = 0$ a.s.). The concentration inequality in Theorem 3 enables one to get a better bound: Since a.s., for every $m \in \mathbb{N}$,

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

then from (29)

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

and from (37), for every $x \geq 0$,

$$\mathbb{P}(X_k \geq kx) \leq \exp\left(-k\, D\Big(\frac{x(1-\varepsilon)}{d} + \varepsilon \,||\, \varepsilon\Big)\right). \tag{39}$$

Consider the case where $\varepsilon \to 0$. Then, for arbitrary $x > 0$ and $k \in \mathbb{N}$, Azuma's inequality in (38) provides an upper bound that is strictly positive independently of $\varepsilon$, whereas the one-sided concentration inequality of Theorem 3 implies a bound in (39) that tends to zero. This exemplifies the improvement that is obtained by Theorem 3 in comparison to Azuma's inequality.

*Remark 7:* As was noted, e.g., in [50, 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-B), so it implies that $\big\{\exp(t(X_k - X_0)), \mathcal{F}_k\big\}_{k=0}^{\infty}$ is a sub-martingale for every $t \geq 0$. Hence, by applying Doob's maximal inequality for sub-martingales, it follows that for every $\alpha \geq 0$

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

which coincides with the proof of Theorem 3 with the starting point in (3). This concept applies to all the concentration inequalities derived in this chapter.

*Corollary 1:* In the setting of Theorem 3, for every $\alpha \geq 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) \leq 2 \exp\left(-2n\left(\frac{\delta}{1+\gamma}\right)^2\right). \tag{40}$$

*Proof:* This concentration inequality is a loosened version of Theorem 3. From Pinsker's inequality,

$$D(p||q) \geq \frac{V^2}{2}, \quad \forall\, p, q \in [0,1] \tag{41}$$

where

$$V \triangleq ||(p, 1-p) - (q, 1-q)||_1 = 2|p-q| \tag{42}$$

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

$$D\left(\frac{\delta+\gamma}{1+\gamma}\Big|\Big|\frac{\gamma}{1+\gamma}\right) \geq 2\left(\frac{\delta}{1+\gamma}\right)^2.$$

■

*Remark 8:* As was shown in the proof of Corollary 1, the loosening of the exponential bound in Theorem 3 by using Pinsker's inequality gives inequality (40). Note that (40) 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 (40) is particularized to Azuma's inequality when $\gamma = 1$, and then

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) \leq 2 \exp\left(-\frac{n\delta^2}{2}\right). \tag{43}$$

This is consistent with the observation that if $\gamma = 1$ then, from (29), the requirement in Theorem 3 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 3 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 $k \in \{1, \ldots, n\}$. Then, for every $\alpha \geq 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) \leq 2 \exp\left(-n f(\delta)\right) \tag{44}$$

where

$$f(\delta) = \begin{cases} \ln(2)\left[1 - h_2\left(\frac{1-\delta}{2}\right)\right], & 0 \leq \delta \leq 1 \\ +\infty, & \delta > 1 \end{cases} \tag{45}$$

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 3 (i.e., since there is no constraint on the conditional variance, then one can take $\sigma^2 = d^2$), the corresponding exponent in (28) is equal to

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

since $D(p||\frac{1}{2}) = \ln 2[1 - h_2(p)]$ 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 (45) is given by

$$f(\delta) = \sum_{p=1}^{\infty} \frac{\delta^{2p}}{2p(2p-1)} = \frac{\delta^2}{2} + \frac{\delta^4}{12} + \frac{\delta^6}{30} + \frac{\delta^8}{56} + \frac{\delta^{10}}{90} \ldots \tag{46}$$

which indeed proves the inequality also for $\delta \in (0, 1]$. Figure 1 shows that the two exponents in (43) and (44) 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 (32) with the inequality

$$\mathbb{E}[\exp(t\xi_k)|\mathcal{F}_{k-1}] \leq \frac{1}{2}\big[e^{td} + e^{-td}\big] = \cosh(td) \tag{47}$$

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,

$$\exp(t\xi_k) \leq \frac{1}{2}\left(1 + \frac{\xi_k}{d}\right) e^{td} + \frac{1}{2}\left(1 - \frac{\xi_k}{d}\right) e^{-td} \tag{48}$$

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 (47) indeed follows from (48). Combined with Chernoff's inequality, it yields (after making the substitution $x = td$ where $x \geq 0$) that

$$\mathbb{P}(X_n - X_0 \geq \alpha n) \leq \big(\exp(-\delta x)\cosh(x)\big)^n, \quad \forall\, x \geq 0. \tag{49}$$

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 $\cosh(x) \leq \exp\left(\frac{x^2}{2}\right)$ so the inequality in (49) 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 (49). If $\delta \in [0, 1]$, then by setting to zero the derivative of the logarithm of the right-hand side of (49), it follows that the optimized value is equal to $x = \tanh^{-1}(\delta)$. Substituting this value into the right-hand side of (49) provides the concentration inequality in Corollary 2; to this end, one needs to rely on the identities

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

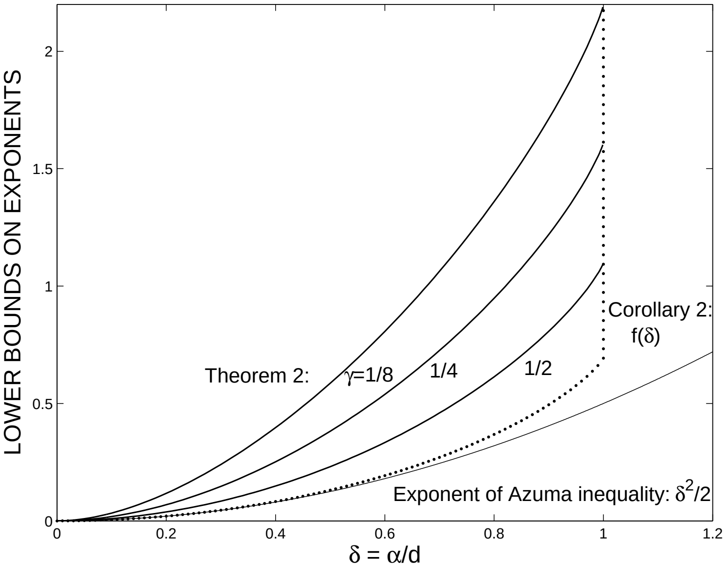

Fig. 1. Plot of the lower bounds on the exponents from Azuma's inequality in (43) and the refined inequalities in Theorem 3 and Corollary 2 (where $f$ is defined in (45)). The pointed line refers to the exponent in Corollary 2, and the three solid lines for $\gamma = \frac{1}{8}, \frac{1}{4}$ and $\frac{1}{2}$ refer to the exponents in Theorem 3.

In the following, a known loosened version of Theorem 3 is re-derived based on Theorem 3.

*Lemma 2:* For every $x, y \in [0,1]$

$$D\left(\frac{x+y}{1+y} \Big|\Big| \frac{y}{1+y}\right) \geq \frac{x^2}{2y}\, B\Big(\frac{x}{y}\Big) \tag{50}$$

where

$$B(u) \triangleq \frac{2[(1+u)\ln(1+u) - u]}{u^2}, \quad \forall\, u > 0. \tag{51}$$

*Proof:* This inequality follows by calculus, and it appears in [21, 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 3. Then, for every $\alpha \geq 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) \leq 2\exp\left(-n\gamma\left[\left(1+\frac{\delta}{\gamma}\right)\ln\left(1+\frac{\delta}{\gamma}\right) - \frac{\delta}{\gamma}\right]\right) \tag{52}$$

where $\gamma, \delta \in [0,1]$ are introduced in (29).

*Proof:* This inequality follows directly by combining inequalities (28) and (50) with the equality in (183). ■

### B. Geometric Interpretation

A common ingredient in proving Azuma's inequality, and Theorem 3 is a derivation of an upper bound on the conditional expectation $\mathbb{E}\big[e^{t\xi_k} \,|\, \mathcal{F}_{k-1}\big]$ for $t \geq 0$ where $\mathbb{E}\big[\xi_k \,|\, \mathcal{F}_{k-1}\big] = 0$, $\text{Var}\big[\xi_k | \mathcal{F}_{k-1}\big] \leq \sigma^2$, 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 3 is based on Bennett's inequality which is applied to the conditional expectation above. The proof of Bennett's inequality (see, e.g., [21, 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 [21, Lemma 2.4.1], it leads to the inequality $y(x) \leq \varphi(x)$ for every $x \in (-\infty, d]$ where $\varphi$ 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 $\alpha \triangleq \big[(1+\gamma)td + 1\big]e^{-t\gamma d} - e^{td}$. 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:

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

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 [21, Lemma 2.4.1]). Applying Bennett's inequality to the conditional law of $\xi_k$ given $\mathcal{F}_{k-1}$ gives (32) (with $\gamma$ in (29)).

### 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 $k \in \mathbb{N}$

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

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

$$\mathbb{P}(|X_n - X_0| \geq n\alpha) \leq 2 \left\{ \inf_{x \geq 0} e^{-\delta x} \left[ 1 + \sum_{l=2}^{m-1} \frac{(\gamma_l - \gamma_m) x^l}{l!} + \gamma_m (e^x - 1 - x) \right] \right\}^n \tag{53}$$

where

$$\delta \triangleq \frac{\alpha}{d}, \quad \gamma_l \triangleq \frac{\mu_l}{d^l}, \;\; \forall\, l = 2, \ldots, m. \tag{54}$$

*Proof:* The starting point of this proof relies on (3) and (4) that were used for the derivation of Theorem 3. From this point, we deviate from the proof of Theorem 3. For every $k \in \mathbb{N}$ and $t \geq 0$

$$\begin{aligned} & \mathbb{E}\big[\exp(t\xi_k)|\mathcal{F}_{k-1}\big] \\ & = 1 + t\mathbb{E}\big[\xi_k|\mathcal{F}_{k-1}\big] + \ldots + \frac{t^{m-1}}{(m-1)!} \cdot \mathbb{E}\big[(\xi_k)^{m-1}|\mathcal{F}_{k-1}\big] \\ & \quad + \mathbb{E}\left[\exp(t\xi_k) - 1 - t\xi_k - \ldots - \frac{t^{m-1}(\xi_k)^{m-1}}{(m-1)!} \,\Big|\, \mathcal{F}_{k-1}\right] \\ & = 1 + t\mathbb{E}\big[\xi_k|\mathcal{F}_{k-1}\big] + \ldots + \frac{t^{m-1}}{(m-1)!} \cdot \mathbb{E}\big[(\xi_k)^{m-1}|\mathcal{F}_{k-1}\big] \\ & \quad + \frac{t^m}{m!} \cdot \mathbb{E}\big[(\xi_k)^m \varphi_m(t\xi_k)|\mathcal{F}_{k-1}\big] \end{aligned} \tag{55}$$

where

$$\varphi_m(y) \triangleq \begin{cases} \frac{m!}{y^m} \left( e^y - \sum_{l=0}^{m-1} \frac{y^l}{l!} \right) & \text{if } y \neq 0 \\ 1 & \text{if } y = 0 \end{cases}. \tag{56}$$

In order to proceed, we need the following lemma:

*Lemma 3:* For every $m \in \mathbb{N}$, the function $\varphi_m$ has the following properties:

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

*Proof:* See Appendix A. ■

*Remark 10:* Note that [28, Lemma 3.1] states that $\varphi_2$ 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 $\varphi_m$ that are stated in Lemma 3; these are sufficient for the continuation of the proof of Theorem 4.

From (55) and Lemma 3, since $\xi_k \leq d$ a.s., then it follows that for an arbitrary $t \geq 0$

$$\varphi_m(t\xi_k) \leq \varphi_m(td), \quad \forall\, k \in \mathbb{N} \tag{57}$$

a.s. (to see this, lets separate the two cases where $\xi_k$ is either non-negative or negative. If $0 \leq \xi_k \leq d$ a.s. then, for $t \geq 0$, inequality (57) holds (a.s.) due to the monotonicity of $\varphi_m$ over $[0, \infty)$. If $\xi_k < 0$ then the second and third properties in Lemma 3 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 (57) is satisfied). Since $m$ is even then $(\xi_k)^m \geq 0$ (note that although Lemma 3 holds in general for every $m \in \mathbb{N}$, this is the point where we need $m$ to be an even number), and

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

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

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

By substituting the last three results on the right-hand side of (55), it follows that for every $t \geq 0$ and every $k \in \mathbb{N}$

$$\mathbb{E}\big[\exp(t\xi_k) | \mathcal{F}_{k-1}\big] \leq 1 + \sum_{l=2}^{m-1} \frac{\gamma_l \, (td)^l}{l!} + \frac{\gamma_m \, (td)^m \, \varphi_m(td)}{m!} \tag{58}$$

so from (4)

$$\mathbb{E}\left[\exp\Big(t \sum_{k=1}^{n} \xi_k\Big)\right] \leq \left(1 + \sum_{l=2}^{m-1} \frac{\gamma_l \, (td)^l}{l!} + \frac{\gamma_m \, (td)^m \, \varphi_m(td)}{m!}\right)^n, \ \ \forall \, t \geq 0. \tag{59}$$

From (3), 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 (29) and (56), for every $\alpha \geq 0$

$$\begin{aligned}
& \mathbb{P}(X_n - X_0 \geq \alpha n) \\
& \leq \left\{ \inf_{x \geq 0} e^{-\delta x} \left(1 + \sum_{l=2}^{m-1} \frac{\gamma_l \, x^l}{l!} + \frac{\gamma_m x^m \, \varphi_m(x)}{m!}\right)\right\}^n \\
& = \left\{ \inf_{x \geq 0} e^{-\delta x} \left[1 + \sum_{l=2}^{m-1} \frac{\gamma_l \, x^l}{l!} + \gamma_m \left(e^x - \sum_{l=0}^{m-1} \frac{x^l}{l!}\right)\right]\right\}^n \\
& = \left\{ \inf_{x \geq 0} e^{-\delta x} \left[1 + \sum_{l=2}^{m-1} \frac{(\gamma_l - \gamma_m) \, x^l}{l!} + \gamma_m \, (e^x - 1 - x)\right]\right\}^n .
\end{aligned} \tag{60}$$

Applying inequality (60) 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 (53) 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. ■

*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 $l \in \{2, \ldots, m\}$. Hence, $\delta \in [0, 1]$, and $\gamma_l \in [0, 1]$ for all values of $l$. Note that, from (29), $\gamma_2 = \gamma$.

*Remark 12:* From the proof of Theorem 4, it follows that the one-sided inequality (60) is satisfied if the martingale $\{X_k, \mathcal{F}_k\}_{k=0}^{n}$ fulfills the following conditions a.s.

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

for some positive number $d > 0$ and a sequence of arbitrary 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 or negative. This remark will be helpful later in this chapter.

*Remark 13:* The infimum in (53) 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 (53), 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, $\lim_{x \to \infty} f(x) = \infty$. 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 $x \to \infty$. 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 3 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 3. Then, for every $\alpha \geq 0$,

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

where $\gamma$ and $\delta$ are introduced in (29), and the exponent in this upper bound gets the following form:

- If $\delta > 1$ then $C(\gamma, \delta) = \infty$.
- If $\delta = 1$ then
$$C(\gamma, \delta) = \frac{1}{\gamma} - \ln\Bigl(\gamma\bigl(e^{\frac{1}{\gamma}} - 1\bigr)\Bigr).$$
- Otherwise, if $\delta \in (0, 1)$, then
$$C(\gamma, \delta) = \delta x - \ln\bigl(1 + \gamma(e^x - 1 - x)\bigr)$$
where $x \in \bigl(0, \frac{1}{\gamma}\bigr)$ is given by
$$x = \frac{1}{\gamma} + \frac{1}{\delta} - 1 - W_0\left(\frac{(1-\delta)e^{\frac{1}{\gamma} + \frac{1}{\delta} - 1}}{\delta}\right) \tag{61}$$
and $W_0$ denotes the principal branch of the Lambert W function [17].

*Proof:* See Appendix B. ■

*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 C. ■

It is of interest to compare the tightness of Theorem 3 and Corollary 4. This leads to the following conclusion:

*Proposition 2:* The concentration inequality in Corollary 4 is looser than Theorem 3.

*Proof:* See Appendix D. ■

*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 3 (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 3 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 VI, 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 3. 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 3 at an early stage, and it cannot be considered as a generalization of Theorem 3 when higher-order moments are available (as is also evidenced in Proposition 2 which demonstrates that, for $m = 2$, Theorem 4 is weaker than Theorem 3).
- Finally, this sufficient condition is particularized in the asymptotic case where $m \to \infty$. 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 3 and 4 are relegated to Section VI (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 $k \in \mathbb{N}$

$$\mu_l = \mathbb{E}[|X_k - X_{k-1}|^l \,|\, \mathcal{F}_{k-1}], \quad \forall\, l = 2, \dots, m. \tag{62}$$

Then, inequality (53) holds with the notation in (54).

*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:* The proof is very technical, and it is omitted for the sake of brevity. ■

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 $\alpha \geq 0$,

$$\begin{aligned} &\mathbb{P}(X_n - X_0 \geq n\alpha) \\ &\leq \left\{ e^{-\delta x} \left[ 1 + \sum_{l=2}^{m-1} \frac{(\gamma_l - \gamma_m) x^l}{l!} + \gamma_m (e^x - 1 - x) \right] \right\}^n \end{aligned} \tag{63}$$

where $\{\gamma_l\}_{l=2}^m$ and $\delta$ are introduced in (54),

$$x = \frac{a+b}{c} - W_0\left(\frac{b}{c} \cdot e^{\frac{a+b}{c}}\right) \tag{64}$$

with $W_0$ that denotes the principal branch of the Lambert W function [17], and

$$a \triangleq \frac{1}{\gamma_2}, \quad b \triangleq \frac{\gamma_m}{\gamma_2}\left(\frac{1}{\delta} - 1\right), \quad c \triangleq \frac{1}{\delta} - b. \tag{65}$$

*Proof:* See Appendix E. ■

*Remark 14:* It is exemplified numerically in Section VI that the replacement of the infimum over $x \geq 0$ on the right-hand side of (53) with the sub-optimal choice of the value of $x$ that is given in (64) and (65) 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 (61).

*Corollary 7:* Under the assumptions of Theorem 3 then, for every $\alpha \geq 0$,

$$\mathbb{P}(X_n - X_0 \geq n\alpha) \leq e^{-nE} \tag{66}$$

where

$$E = E_2(\gamma_2, \delta) \triangleq D\left(\frac{\delta + \gamma_2}{1 + \gamma_2} \Big|\Big| \frac{\gamma_2}{1 + \gamma_2}\right). \tag{67}$$

Also, under the assumptions of Theorem 4 or Corollary 5 then (66) holds for every $\alpha \geq 0$ with

$$\begin{aligned} E &= E_4(\{\gamma_l\}_{l=2}^m, \delta) \\ &\triangleq \sup_{x \geq 0} \left\{ \delta x - \ln\left(1 + \sum_{l=2}^{m-1} \frac{(\gamma_l - \gamma_m)x^l}{l!} + \gamma_m(e^x - 1 - x)\right)\right\} \end{aligned} \tag{68}$$

where $m \geq 2$ is an arbitrary even number. Hence, Theorem 4 or Corollary 5 are better exponentially than Theorem 3 if and only if $E_4 > E_2$.

*Proof:* The proof follows directly from (37) and (60). ■

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

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

with the value of $x$ in (64) and (65). This sufficient condition is exemplified later in Section VI.

### 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, Theorem 3, and also Corollaries 3 and 4 imply that, for every $\alpha \geq 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) \leq 2\exp\Big(-\frac{\delta^2}{2\gamma}\Big)\Big(1 + O\big(n^{-\frac{1}{2}}\big)\Big). \tag{69}$$

Also, under the conditions of Theorem 4, inequality (69) 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 (69)).

*Proof:* See Appendix F. ■

*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., [15, Chapter 2] or [16, 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 3 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.

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

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

$$\mathbb{P}(X_n - X_0 \geq \alpha n) \leq \exp\left(-n\, D\left(\frac{\delta+\gamma}{1+\gamma}\Big|\Big|\frac{\gamma}{1+\gamma}\right)\right) \tag{70}$$

where $\gamma$ and $\delta$ are defined as in (29), and the divergence $D(p||q)$ is introduced in (30). Alternatively, if $\{X_k, \mathcal{F}_k\}_{k=0}^{\infty}$ is a sub-martingale, the same upper bound in (70) 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 3. The only difference is that for a super-martingale, due to its basic property in Section II-B,

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

a.s., where $\xi_k \triangleq X_k - \mathbb{E}[X_k \,|\, \mathcal{F}_{k-1}]$ is $\mathcal{F}_k$-measurable. Hence $\mathbb{P}((X_n - X_0 \geq \alpha n) \leq \mathbb{P}\big(\sum_{k=1}^{n} \xi_k \geq \alpha n\big)$ 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 3 (starting from (3)). 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. ■

## V. Relations of the Refined Inequalities to Some Classical Results in Probability Theory

### A. 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) $\mathbb{E}\big[Y_k \,|\, \mathcal{F}_{k-1}\big] = 0$.

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

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

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., [11, 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 3 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 $\alpha \geq 0$,

$$\lim_{n \to \infty} \mathbb{P}(|X_n - X_0| \geq \alpha \sqrt{n}) = 2\, Q\Big(\frac{\alpha}{\sigma}\Big)$$

where the $Q$ function is introduced in (218).

Based on the notation in (29), the equality $\frac{\alpha}{\sigma} = \frac{\delta}{\sqrt{\gamma}}$ holds, and

$$\lim_{n \to \infty} \mathbb{P}(|X_n - X_0| \geq \alpha \sqrt{n}) = 2\, Q\Big(\frac{\delta}{\sqrt{\gamma}}\Big). \tag{71}$$

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 $\alpha \geq 0$

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

This inequality coincides with the asymptotic result of the inequalities in Proposition 4 (see (69) in the limit where $n \to \infty$), except for the additional factor of 2. Note also that the proof of the concentration inequalities in Proposition 4 (see Appendix F) 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 (15), then it follows from (71) that the exponent in the concentration inequality (69) (i.e., $\frac{\delta^2}{2\gamma}$) cannot be improved under the above assumptions (unless some more information is available).

### *B. Relation between the Law of the Iterated Logarithm (LIL) and Theorem 3*

In this subsection, we discuss the relation between the law of the iterated logarithm (LIL) and Theorem 3.

According to the law of the iterated logarithm (see, e.g., [11, 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

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

and

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

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 $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$ is the $\sigma$-algebra that is generated by the RVs $X_1, \ldots, X_k$ 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 3 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 3, it follows that for every $\alpha \geq 0$

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

where

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

Straightforward calculation shows that

$$\begin{aligned}
& nD\Big(\frac{\delta_n + \gamma}{1+\gamma}\Big|\Big|\frac{\gamma}{1+\gamma}\Big) \\
&= \frac{n\gamma}{1+\gamma}\left[\Big(1 + \frac{\delta_n}{\gamma}\Big)\ \ln\Big(1 + \frac{\delta_n}{\gamma}\Big) + \frac{1}{\gamma}\,(1-\delta_n)\ln(1-\delta_n)\right] \\
&\overset{\text{(a)}}{=} \frac{n\gamma}{1+\gamma}\left[\frac{\delta_n^2}{2}\Big(\frac{1}{\gamma^2} + \frac{1}{\gamma}\Big) + \frac{\delta_n^3}{6}\Big(\frac{1}{\gamma} - \frac{1}{\gamma^3}\Big) + \ldots\right] \\
&= \frac{n\delta_n^2}{2\gamma} - \frac{n\delta_n^3(1-\gamma)}{6\gamma^2} + \ldots \\
&\overset{\text{(b)}}{=} \alpha^2 \ln\ln n\left[1 - \frac{\alpha(c^2-1)}{6c}\sqrt{\frac{\ln\ln n}{n}} + \ldots\right]
\end{aligned} \tag{76}$$

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 $\alpha \geq 0$,

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

and the same bound also applies to $\mathbb{P}\big(S_n \leq -\alpha\sqrt{2n\ln\ln n}\big)$ 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

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

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

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

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 $\theta > 1$):

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

where the last inequality follows from (78) with $\beta_n \to 0$ as $n \to \infty$. 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 $\theta$, 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

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

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

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

Theorem 3 therefore gives inequality (78), and it implies one side in each of the two equalities for the LIL in (72) and (73).

### *C. Relation of Theorems 3 and 4 with the Moderate Deviations Principle*

According to the moderate deviations theorem (see, e.g., [21, 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 $a_n \to 0$ and $na_n \to \infty$ as $n \to \infty$, and let

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

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

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

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

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

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

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

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

It is demonstrated in Appendix G that, in contrast to Azuma's inequality, Theorems 3 and 4 (for every even $m \geq 2$ in Theorem 4) provide upper bounds on the probability

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

which both coincide with the asymptotic limit in (82). The analysis in Appendix G provides another interesting link between Theorems 3 and 4 and a classical result in probability theory, which also emphasizes the significance of the refinements of Azuma's inequality.

### *D. Relation of [50, Lemma 2.8] with Theorem 4 & Corollary 4*

In [50, 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

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

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 (56), it follows that $\varphi(x) = \frac{\varphi_2(x)}{2}$ for every $x \in \mathbb{R}$. Based on [50, 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 $t \geq 0$

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

holds a.s. for every $k \in \mathbb{N}$ (the parameter $\gamma$ was introduced in (29)). The last inequality can be rewritten as

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

This forms a looser bound on the conditional expectation, as compared to (58) with $m = 2$, that gets the form

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

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 [50, 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

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

and derive a closed-form solution of the optimized parameter $t$ in the resulting concentration inequality (see the proof of [50, Theorem 2.7] for the case of independent RVs, and also [50, Theorem 3.15] for the setting of martingales with bounded jumps). However, if for every $k \in \mathbb{N}$, the condition

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

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.

### *E. 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., [31, 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., $\sum_{j \in \mathcal{S}} p_{i,j} \psi(j) = \psi(i), \quad \forall \, i \in \mathcal{S}$), 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 IV for harmonic functions of Markov chains when the function $\psi$ 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 [39]. These bounds were also shown to be essentially identical to the Hoeffding inequality in the special case of i.i.d. RVs (see [39, Remark 1]).

### *F. Relations of [16, 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 [16, Theorem 2.23] (or also [15, 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 [16, 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 (29) then $\frac{\alpha^2}{\sigma^2} = \frac{\delta^2}{\gamma}$, and the last inequality gets the form

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

  It therefore follows that [16, Theorem 2.23] implies a concentration inequality of the form in (69). This shows that Proposition 4 can be also regarded as a consequence of [16, Theorem 2.23].
- In the analysis of large deviations, the bound in [16, 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 (29), this inequality is rewritten as

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

It is claimed that the concentration inequality in (86) is looser than Corollary 4. This is a consequence of the proof of [16, 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 chapter (see Theorem 3), where $d_k$ and $\sigma_k^2$ may depend on $k$. In order to show it explicitly, lets compare between the steps of the derivation of the bound in Corollary 4, and the particularization of the derivation of [16, Theorem 2.23] in the special setting where $d_k$ and $\sigma_k^2$ 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 $\alpha \geq 0$,

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

which then leads, after an analytic optimization of the free non-negative parameter $x$ (see Lemma 6 and Appendix B), to the derivation of Corollary 4. On the other hand, the specialization of the proof of [16, 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

$$\begin{aligned} & \mathbb{P}(X_n - X_0 \geq \alpha n) \\ & \leq e^{-n\delta x} e^{n\gamma(e^x - 1 - x)} && (88) \\ & \leq e^{n\left(-\delta x + \frac{\gamma x^2}{1 - \frac{x}{3}}\right)}, \quad \forall\, x \in (0, 3) && (89) \end{aligned}$$

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 [16, Theorem 2.23].

In order to compare quantitatively the exponents of the concentration inequalities in [16, 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 B is positive, and it becomes larger as we let the value of $\gamma \in (0, 1]$ approach zero. Hence, especially for small values of $\gamma$, the loosening of the bound from (87) to (89) is expected to deteriorate more significantly the resulting bound in [16, Theorem 2.23] due to the restriction that $x \in (0, 3)$; this is in contrast to the optimized value of $x$ in Appendix B that may be above 3 for small values of $\gamma$, 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 $\gamma \to 0$, whereas the exponent in (86) tends in this case to $\frac{3}{2}$. To illustrate these differences, Figure 2 plots the exponents of the bounds in Corollary 4 and (86), where the latter refers to [16, Theorem 2.23], for $\gamma = 0.01$ and $0.99$. As is shown in Figure 2, the difference between the exponents of these two bounds is indeed more pronounced when $\gamma$ gets closer to zero.

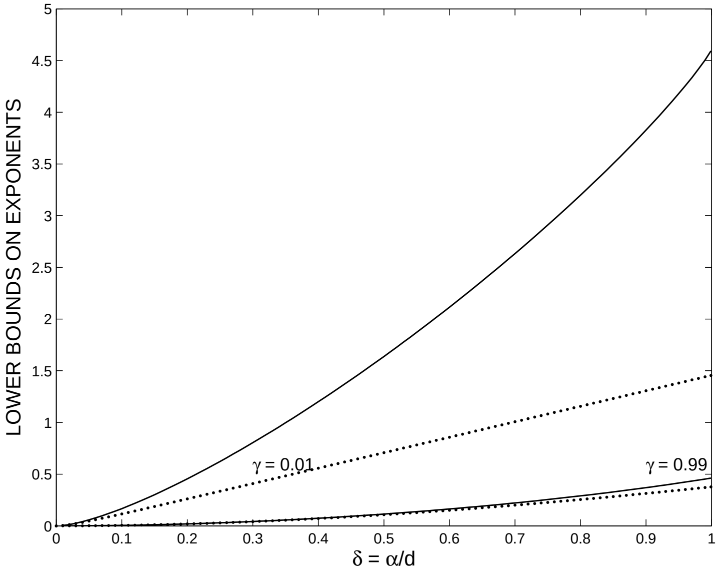

Fig. 2. A comparison of the exponents of the bound in Corollary 4 and the particularized bound (86) from [16, Theorem 2.23]. This comparison is done for both $\gamma = 0.01$ and $0.99$. The solid curves refer to the exponents of the bound in Corollary 4, and the dashed curves refer to the exponents of 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 C 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 [16, Theorem 2.23] essentially scale similarly for small deviations of order $\sqrt{n}$.

## VI. Applications in Information Theory and Related Topics

The refined versions of Azuma's inequality in Section IV are applied in this section to information-theoretic aspects. The results in this section have been presented in part in [65], [66], [67] and [82].

### A. Binary Hypothesis Testing

Binary hypothesis testing for finite alphabet models was analyzed via the method of types, e.g., in [18, Chapter 11] and [19]. 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 $X_1, X_2....$ 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

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

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

$$\lim_{n\to\infty} \frac{L(X_1, \ldots, X_n)}{n} = D(P_1 || P_2) \tag{90}$$

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

$$\lim_{n\to\infty} \frac{L(X_1, \ldots, X_n)}{n} = -D(P_2 || P_1) \tag{91}$$

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 $\overline{\lambda}, \underline{\lambda} \in \mathbb{R}$ that satisfy

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

one decides on hypothesis $H_1$ if

$$L(X_1, \ldots, X_n) > n\overline{\lambda}$$

and on hypothesis $H_2$ if

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

Note that if $\overline{\lambda} = \underline{\lambda} \triangleq \lambda$ then a decision on the two hypotheses is based on comparing the normalized log-likelihood ratio (w.r.t. $n$) to a single threshold $(\lambda)$, and deciding on hypothesis $H_1$ or $H_2$ if it is, respectively, above or below $\lambda$. If $\underline{\lambda} < \overline{\lambda}$ then one decides on $H_1$ or $H_2$ if the normalized log-likelihood ratio is, respectively, above the upper threshold $\overline{\lambda}$ 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

$$\alpha_n^{(1)} \triangleq P_1^n\Big(L(X_1, \ldots, X_n) \leq n\overline{\lambda}\Big) \tag{92}$$

$$\alpha_n^{(2)} \triangleq P_1^n\Big(L(X_1, \ldots, X_n) \leq n\underline{\lambda}\Big) \tag{93}$$

and

$$\beta_n^{(1)} \triangleq P_2^n\Big(L(X_1, \ldots, X_n) \geq n\underline{\lambda}\Big) \tag{94}$$

$$\beta_n^{(2)} \triangleq P_2^n\Big(L(X_1, \ldots, X_n) \geq n\overline{\lambda}\Big) \tag{95}$$

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

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

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

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

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., [21, Theorem 3.4.3] and [35, 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 $\overline{\lambda} = \underline{\lambda} \triangleq \lambda$, the option of erasures does not exist, and $P_{\text{e},n}^{(1)} = P_{\text{e},n}^{(2)} \triangleq P_{\text{e},n}$ is the error probability.

In the considered general case with erasures, let

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

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

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

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

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

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

where the rate function $I$ is given by

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

and

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

The rate function $I$ is convex, lower semi-continuous (l.s.c.) and non-negative (see, e.g., [21] and [35]). Note that

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

where $D_t(P||Q)$ designates Réyni's information divergence of order $t$ [59, Eq. (3.3)], and $I$ in (102) is the Fenchel-Legendre transform of $H$ (see, e.g., [21, Definition 2.2.2]).

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

$$\lim_{n\to\infty} -\frac{\ln P_{\text{e},n}^{(1)}}{n} = \min\Big\{I(\lambda_1), I(\lambda_2) - \lambda_2\Big\} \tag{104}$$

and

$$\lim_{n\to\infty} -\frac{\ln P_{\mathrm{e},n}^{(2)}}{n} = \min\Big\{I(\lambda_2), I(\lambda_1) - \lambda_1\Big\}. \tag{105}$$

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 $P_{\mathrm{e},n}^{(1)} = P_{\mathrm{e},n}^{(2)} \triangleq P_{\mathrm{e},n}$, and its error exponent is equal to

$$\lim_{n\to\infty} -\frac{\ln P_{\mathrm{e},n}}{n} = \min\Big\{I(\lambda), I(\lambda) - \lambda\Big\} \tag{106}$$

which coincides with the error exponent in [21, Theorem 3.4.3] (or [35, Theorem 6.4]). The optimal threshold for obtaining the best error exponent of the error probability $P_{\mathrm{e},n}$ is equal to zero (i.e., $\lambda = 0$); in this case, the exact error exponent is equal to

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

which is the Chernoff information of the probability measures $P_1$ and $P_2$ (see [18, Eq. (11.239)]), and it is symmetric (i.e., $C(P_1, P_2) = C(P_2, P_1)$). Note that, from (102), $I(0) = \sup_{t\in\mathbb{R}}\big(-H(t)\big) = -\inf_{t\in\mathbb{R}}\big(H(t)\big)$; 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 [12] 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 VI-A5 and VI-A6 w.r.t. moderate and small deviations analysis of binary hypothesis testing).

*2) Lower Bound on the Exponents via Theorem 3:* In the following, the tightness of Theorem 3 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 $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \mathcal{F}_n$ is the filtration

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

and

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

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

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

In particular

$$U_0 = nD(P_1||P_2), \tag{109}$$

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

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

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

Let

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

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)

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

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

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

Let

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

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

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 3

$$\begin{aligned} \alpha_n^{(1)} &\triangleq P_1^n\big(L(X_1, \ldots, X_n) \le n\overline{\lambda}\big) \\ &\overset{\text{(a)}}{=} P_1^n(U_n - U_0 \le -\varepsilon_{1,1}\, n) \qquad (116) \\ &\overset{\text{(b)}}{\le} \exp\left(-n\, D\Big(\frac{\delta_{1,1} + \gamma_1}{1+\gamma_1} \Big|\Big| \frac{\gamma_1}{1+\gamma_1}\Big)\right) \qquad (117) \end{aligned}$$

where equality (a) follows from (109), (110) and (114), and inequality (b) follows from Theorem 3 with

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

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

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

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

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

and in particular

$$U_0 = -nD(P_2||P_1), \quad U_n = L(X_1, \ldots, X_n). \tag{120}$$

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

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

Let

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

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

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

Hence, it follows from Theorem 3 that

$$\begin{aligned}
\beta_n^{(1)} &\triangleq P_2^n\big(L(X_1, \dots, X_n) \geq n\underline{\lambda}\big) \\
&= P_2^n(U_n - U_0 \geq \varepsilon_{2,1}\, n) && (124)\\
&\leq \exp\left(-n\, D\Big(\frac{\delta_{2,1} + \gamma_2}{1+\gamma_2} \Big|\Big| \frac{\gamma_2}{1+\gamma_2}\Big)\right) && (125)
\end{aligned}$$

where the equality in (124) holds due to (120) and (114), and (125) follows from Theorem 3 with

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

and $d_2$, $\sigma_2$ 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

$$\lim_{n \to \infty} -\frac{\ln P_{\mathrm{e},n}^{(1)}}{n} \geq \min_{i=1,2} D\Big(\frac{\delta_{i,1} + \gamma_i}{1+\gamma_i} \Big|\Big| \frac{\gamma_i}{1+\gamma_i}\Big). \tag{127}$$

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

$$\lim_{n \to \infty} -\frac{\ln P_{\mathrm{e},n}^{(2)}}{n} \geq \min_{i=1,2} D\Big(\frac{\delta_{i,2} + \gamma_i}{1+\gamma_i} \Big|\Big| \frac{\gamma_i}{1+\gamma_i}\Big) \tag{128}$$

where

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

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

$$\lim_{n \to \infty} -\frac{\ln P_{\mathrm{e},n}}{n} \geq \min_{i=1,2} D\Big(\frac{\delta_i + \gamma_i}{1+\gamma_i} \Big|\Big| \frac{\gamma_i}{1+\gamma_i}\Big) \tag{130}$$

where $\delta_i$ 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., [21, p. 93]), which then yields that (130) is satisfied with

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

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 3, 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 $P_{\text{e},n}^{(1)}$ and $P_{\text{e},n}^{(2)}$.

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

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

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

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

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

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

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

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

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

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

$$\lim_{n\to\infty} -\frac{\ln P_{\text{e},n}^{(j)}}{n} \geq \min_{i=1,2} D\Big(\frac{\delta_{i,j}+\gamma_i}{1+\gamma_i}\Big|\Big|\frac{\gamma_i}{1+\gamma_i}\Big). \tag{137}$$

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

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

with the values of $\delta_1$ and $\delta_2$ 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 3.

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

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

*Lemma 4:*

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

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

*Proof:* The proof follows by elementary calculus. ■

Since $\delta_{i,j} \in [0, 1]$, then (139) and Lemma 4 imply that

$$D\Big(\frac{\delta_{i,j} + \gamma_i}{1 + \gamma_i} \Big|\Big| \frac{\gamma_i}{1 + \gamma_i}\Big) \geq \frac{\delta_{i,j}^2}{2\gamma_i} - \frac{\delta_{i,j}^3}{6\gamma_i^2(1 + \gamma_i)}. \tag{141}$$

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 3 are improved by at least a factor of $\big(\max \gamma_i\big)^{-1}$ 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

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

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 $n \to \infty$. 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 $\Theta$ denotes the parameter set. Assume that $P_\theta$ is differentiable in the parameter $\theta$. Then, the Fisher information is defined as

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

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

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

(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 [18, Eq. (12.364)]).

*Proposition 5:* Under the above assumptions,

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

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

- Let
$$E_{\text{L}}(P_\theta, P_{\theta'}) \triangleq \min_{i=1,2} D\Big(\frac{\delta_i+\gamma_i}{1+\gamma_i} \Big|\Big| \frac{\gamma_i}{1+\gamma_i}\Big) \tag{145}$$
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
$$\lim_{\theta' \to \theta} \frac{E_{\text{L}}(P_\theta, P_{\theta'})}{(\theta-\theta')^2} = \frac{J(\theta)}{8}. \tag{146}$$
- Let
$$\widetilde{E}_{\text{L}}(P_\theta, P_{\theta'}) \triangleq \min_{i=1,2} \frac{\delta_i^2}{2} \tag{147}$$
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,
$$\lim_{\theta' \to \theta} \frac{\widetilde{E}_{\text{L}}(P_\theta, P_{\theta'})}{(\theta-\theta')^2} = \frac{a(\theta)\, J(\theta)}{8} \tag{148}$$
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 H. ■

Proposition 5 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 H).

*Example 5:* Consider the index family of of probability mass functions defined over the binary alphabet $\mathcal{X} = \{0,1\}$:
$$P_\theta(0) = 1-\theta, \;\; P_\theta(1) = \theta, \quad \forall\, \theta \in (0,1).$$
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)
$$C(P_{\theta_1}, P_{\theta_2}), E_{\text{L}}(P_{\theta_1}, P_{\theta_2}) \approx \frac{J(\theta)(\theta_1-\theta_2)^2}{8} = 2.00 \cdot 10^{-4}.$$
Indeed, the exact values of $C(P_{\theta_1}, P_{\theta_2})$ and $E_{\text{L}}(P_{\theta_1}, P_{\theta_2})$ 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 $\overline{\lambda}$ and $\underline{\lambda}$ which are *kept fixed* (i.e., these thresholds do not depend on the block length $n$ of the data sequence) where
$$-D(P_2||P_1) < \underline{\lambda} \leq \overline{\lambda} < D(P_1||P_2).$$
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.,
$$\lim_{n\to\infty} \overline{\lambda}^{(n)} = D(P_1||P_2), \quad \lim_{n\to\infty} \underline{\lambda}^{(n)} = -D(P_2||P_1).$$
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
$$L(X_1, \ldots, X_n) > n\overline{\lambda}^{(n)}$$
and on hypothesis $H_2$ if
$$L(X_1, \ldots, X_n) < n\underline{\lambda}^{(n)}$$

where these upper and lower thresholds are set to

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

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 $\overline{\lambda}$ and $\underline{\lambda}$ are replaced with the above block-length dependent thresholds $\overline{\lambda}^{(n)}$ 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 $\overline{\lambda}^{(n)}$ 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], [32], [56] and [76]).

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 [57] borrowed some tools from a recent work by the same authors in [56].

In [32], 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 [76].

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., [21, Section 3.7]), and in particular the MDP was studied in [20] for continuous-time martingales with bounded jumps.

In light of the discussion in Section V-C on the MDP for i.i.d. RVs and its relation to the concentration inequalities in Section IV (see Appendix G), 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 [57], and it relies on concentration inequalities for martingales. The material in the following was presented in part in [67].

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)

$$\begin{aligned} & P_1^n\bigl(L(X_1,\ldots,X_n) \leq n\overline{\lambda}^{(n)}\bigr) \\ & = P_1^n\bigl(L(X_1,\ldots,X_n) \leq nD(P_1||P_2) - \varepsilon_1 n^{\eta}\bigr) \\ & \leq \exp\left(-nD\biggl(\frac{\delta_1^{(\eta,n)}+\gamma_1}{1+\gamma_1} \,\Big|\Big|\, \frac{\gamma_1}{1+\gamma_1}\biggr)\right) \end{aligned} \tag{149}$$

where

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

with $d_1$ and $\sigma_1^2$ from (112) and (113). From (139), (140) and (150), it follows that

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

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

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

Consequently, in the limit where $n$ tends to infinity,

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

with $\sigma_1^2$ in (113). From the analysis in Section V-C and Appendix G, 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 3).

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

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

that, under hypothesis $H_1$, have zero mean and variance $\sigma_1^2$. Since, by assumption, $\{X_i\}_{i=1}^n$ are i.i.d., then

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

and it follows from the one-sided version of the MDP in (82) that indeed (152) holds with equality. Moreover, Theorem 3 provides, via the inequality in (151), a finite-length result that enhances the asymptotic result for $n \to \infty$.

The second item above follows from the second part of the analysis in Appendix G (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 V-C and Appendix G, 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 $\eta$ as above (note that by reducing the value of $\eta$ for a fixed $n$, the upper and lower thresholds $\overline{\lambda}^{(n)}$ 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 [57], the authors consider moderate deviations analysis for channel coding over memoryless channels. In particular, [4, Theorem 2.2] and [57, 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 3, 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 G, 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 $\sigma_1$ may be significantly smaller than $d_1$).

*6) Second-Order Analysis for Binary Hypothesis Testing:* The moderate deviations analysis in the previous subsection refers to deviations that scale like $n^\eta$ 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 $\sigma_1^2$ (under hypothesis $H_1$), and define the partial sums $S_k = \sum_{i=1}^k Y_i$ for $k \in \{1, \ldots, n\}$ 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 V-A which refers to the relation between the martingale CLT and Proposition 4. Specifically, since from (153),

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

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

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

for a finite block length $n$ (via an analysis that is either related to Theorem 3 or 4) which agrees with the asymptotic result

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

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 [33], [56], and [58], where the variance $\sigma_1^2$ 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 3 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

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

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 $W_{\mathrm{H}}(\underline{x}) = h$. 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 $1 \leq i_1 < \ldots < i_h \leq n$. Based on this notation, the log-likelihood ratio satisfies the equality

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

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

$$\mathcal{F}_k = \sigma(Y_{i_1}, \ldots, Y_{i_k}), \quad k = 1, \ldots, h$$
$$\mathcal{F}_0 = \{\emptyset, \Omega\}$$

and, under hypothesis $H_1$, let

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

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

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

Specifically

$$U_0 = -h\, D\big(P_{Y|X}(\cdot|0) \,||\, P_{Y|X}(\cdot|1)\big) \tag{156}$$

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

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

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

for every $k \in \{1, \ldots, h\}$. 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 $\underline{x} \in \mathcal{C}$ ($W_{\mathrm{H}}(\underline{x}) = h$) over the transmitted all-zero codeword, is equal to

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

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

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

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 $k \in \{1, \ldots, h\}$. We rely in the following on the concentration inequalities of Theorems 3 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 3:* From (158), for every $k \in \{1, \ldots, h\}$

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

holds a.s., where the last equality follows from the definition of the divergence (relative entropy). Based on (159) and the notation in (29), let

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

where $d$ and $\sigma^2$ are introduced in (160) and (161), respectively. Under hypothesis $H_1$, one gets from (159) and Theorem 3 that the pairwise error probability satisfies the upper bound

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

where

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

and $\gamma$, $\delta$ are introduced in (162).

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

$$P_h \leq Z_{\text{B}}^h \tag{165}$$

where the Bhattacharyya parameter $Z_{\text{B}}$ of the binary-input DMC is given by

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

*Example 6:* Consider a binary symmetric channel (BSC) with crossover probability $p$. The Bhattacharyya parameter which corresponds to this channel is $Z_{\text{B}} = \sqrt{4p(1-p)}$. 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

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

and therefore (162) gives that

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

Substituting $\gamma$ and $\delta$ into (164) gives that the base of the exponential bound in (163) is equal to

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

which coincides with the Bhattacharyya parameter for the BSC. This shows that, for the BSC, Theorem 3 implies the Bhattacharyya upper bound on the pairwise error probability.

In general, it is observed numerically that $Z_1 \geq Z_{\text{B}}$ 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 3 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 $k \in \{1, \ldots, h\}$ and $l \in \mathbb{N}$

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

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 $Z_{\text{B}}$ 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 |
|---|---|---|---|---|---|
| $Z_{\text{B}}$ | 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 | 0.4887 | 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 |

and, based on Remark 12, let

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

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

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

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

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

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 $\mathcal{Y} = \{0, 1, \ldots, Q-1\}$, characterized by the following probability transitions:

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

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

$$Z_{\text{B}} = 2\sqrt{p\left[1 - (Q-1)p\right]} + (Q-2)p.$$

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 forth lines in Theorem 4 is consistent with Proposition 2 which indeed assures that Theorem 4 with $m = 2$ is looser than Theorem 3 (hence, indeed $Z_1 < Z_2^{(2)}$ for the considered DMCs). Also, from Example 6, it follows that Theorem 3 coincides with the Battacharyya bound (hence, $Z_1 = Z_{\text{B}}$ for the special case where $Q = 2$, as is

TABLE II

THE BASE $\widetilde{Z}_2^{(m)}$ OF THE EXPONENTIAL BOUND IN (163) AND ITS (TIGHT) UPPER BOUND $\widetilde{Z}_2^{(m)}$ THAT FOLLOWS BY REPLACING THE INFIMUM OPERATION BY THE SUB-OPTIMAL VALUE IN (64) AND (65). 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 |
| $\widetilde{Z}_2^{(10)}$ | 0.3919 | 0.4237 | 0.4553 | 0.4868 | 0.6417 |

indeed verified numerically in Table I). It is interesting to realize from Table I that for the five considered DMCs, the sequence $\{Z_2^{(2)}, Z_2^{(4)}, Z_2^{(6)}, \ldots\}$ 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 3, and which appears to be strictly larger than the corresponding Bhattacharyya parameter of the DMC (except for the BSC, where the equality $Z_1 = Z_{\text{B}}$ 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,

$$\lim_{m \to \infty} Z_2^{(m)} = Z_{\text{B}}$$

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 B for the proof of Corollary 4).

Replacing the infimum over $x \in [0, \infty)$ with the sub-optimal value of $x$ in (64) and (65) 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 (64) and (65) is denoted by $\widetilde{Z}_2^{(m)}$, and the difference between the two values is marginal (see Table II).

### C. Minimum Distance of Binary Linear Block Codes

Consider the ensemble of binary linear block codes of length $n$ and rate $R$. The average value of the normalized minimum distance is equal to

$$\frac{\mathbb{E}[d_{\min}(\mathcal{C})]}{n} = h_2^{-1}(1 - R)$$

where $h_2^{-1}$ designates the inverse of the binary entropy function to the base 2, and the expectation is with respect to the ensemble where the codes are chosen uniformly at random (see [8]).

Let $H$ designate an $n(1-R) \times n$ parity-check matrix of a linear block code $\mathcal{C}$ from this ensemble. The minimum distance of the code is equal to the minimal number of columns in $H$ that are linearly dependent. Note that the minimum distance is a property of the code, and it does not depend on the choice of the particular parity-check matrix which represents the code.

Let us construct a martingale sequence $X_0, \ldots, X_n$ where $X_i$ (for $i = 0, 1, \ldots, n$) is a RV that denotes the minimal number of linearly dependent columns of a parity-check matrix that is chosen uniformly at random from the ensemble, given that we already revealed its first $i$ columns. Based on Remarks 2 and 3, this sequence forms indeed a martingale sequence where the associated filtration of the $\sigma$-algebras $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}_n$ is defined so that $\mathcal{F}_i$ (for $i = 0, 1, \ldots, n$) is the $\sigma$-algebra that is generated by all the sub-sets of $n(1-R) \times n$ binary parity-check matrices whose first $i$ columns are fixed. This martingale sequence satisfies $|X_i - X_{i-1}| \leq 1$ for $i = 1, \ldots, n$ (since if we reveal a new column of $H$, then the minimal number of linearly dependent columns can change by at most 1). Note that the RV $X_0$ is the expected minimum Hamming distance of the ensemble, and $X_n$ is the minimum distance of a particular code from the ensemble (since once we revealed all the $n$ columns of $H$, then the code is known exactly). Hence, by Azuma's inequality

$$\mathbb{P}(|d_{\min}(\mathcal{C}) - \mathbb{E}[d_{\min}(\mathcal{C})]| \geq \alpha\sqrt{n}) \leq 2 \exp\left(-\frac{\alpha^2}{2}\right), \; \forall \, \alpha > 0.$$

This leads to the following theorem:

*Theorem 6:* **[The minimum distance of binary linear block codes]** Let $\mathcal{C}$ be chosen uniformly at random from the ensemble of binary linear block codes of length $n$ and rate $R$. Then for every $\alpha > 0$, with probability at least $1 - 2\exp\left(-\frac{\alpha^2}{2}\right)$, the minimum distance of $\mathcal{C}$ is in the interval

$$[n\, h_2^{-1}(1-R) - \alpha\sqrt{n},\ n\, h_2^{-1}(1-R) + \alpha\sqrt{n}]$$

and it therefore concentrates around its expected value.

Note, however, that some well-known capacity-approaching families of binary linear block codes possess a minimum Hamming distance which grows sub-linearly with the block length $n$. For example, the class of parallel concatenated convolutional (turbo) codes was proved to have a minimum distance which grows at most like the logarithm of the interleaver length [14].

### *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 [29]. 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., [61]). 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 $\lambda_i$ and $\rho_i$ denote the fraction of edges attached, respectively, to variable and parity-check nodes of degree $i$. The LDPC code ensemble is denoted by 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., [64, 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 [25]. 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 [64, Section II-E]. In [64], 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 [64, 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 $\varepsilon$ 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 [25] on cycle-free codes, the lower bound on the cardinality of the fundamental system of cycles in [64, 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 $(n, \lambda, \rho)$ 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 $m = n(1 - R_{\text{d}})$ parity-check nodes, so there are in total $|V(\mathcal{G})| = n(2 - R_{\text{d}})$ nodes. Let $a_{\text{R}}$ 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 $|E(\mathcal{G})| = m a_{\text{R}}$. Therefore, for a code from the $(n, \lambda, \rho)$ LDPC code ensemble, the cardinality of the fundamental system of cycles satisfies the equality

$$\beta(\mathcal{G}) = n\big[(1 - R_{\text{d}})a_{\text{R}} - (2 - R_{\text{d}})\big] + c(\mathcal{G}) \tag{170}$$

where

$$R_{\text{d}} = 1 - \frac{\int_0^1 \rho(x)\,\mathrm{d}x}{\int_0^1 \lambda(x)\,\mathrm{d}x}, \quad a_{\text{R}} = \frac{1}{\int_0^1 \rho(x)\,\mathrm{d}x}$$

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

Let

$$E \triangleq |E(\mathcal{G})| = n(1 - R_{\text{d}})a_{\text{R}} \tag{171}$$

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 $1, \ldots, E$ to the $E$ edges of $\mathcal{G}$. Based on Remarks 2 and 3, lets construct a martingale sequence $X_0, \ldots, X_E$ where $X_i$ (for $i = 0, 1, \ldots, E$) 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 $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}_E$ in this case is defined so that $\mathcal{F}_i$ is the $\sigma$-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 $k = 1, \ldots, E$ (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 $\alpha \geq 0$

$$\begin{aligned} &\mathbb{P}\left(|c(\mathcal{G}) - \mathbb{E}_{\text{LDPC}(n,\lambda,\rho)}[c(\mathcal{G})]| \geq \alpha E\right) \leq 2e^{-f(\alpha)E} \\ \Rightarrow\ &\mathbb{P}\left(|\beta(\mathcal{G}) - \mathbb{E}_{\text{LDPC}(n,\lambda,\rho)}[\beta(\mathcal{G})]| \geq \alpha E\right) \leq 2e^{-f(\alpha)E} \end{aligned} \tag{172}$$

where the last transition follows from (170), and the function $f$ was defined in (45). Hence, for $\alpha > 1$, this probability is zero (since $f(\alpha) = +\infty$ for $\alpha > 1$). Note that, from (170), $\mathbb{E}_{\text{LDPC}(n,\lambda,\rho)}[\beta(\mathcal{G})]$ scales linearly with $n$. The combination of Eqs. (45), (171), (172) gives the following statement:

*Theorem 7:* **[Concentration inequality for the cardinality of the fundamental system of cycles]** Let 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 $\lambda$ and $\rho$. 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:

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

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

*Remark 17:* The loosened version of Theorem 7, which follows from Azuma's inequality, gets the form

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

for every $\alpha \geq 0$, and $\eta$ as defined in Theorem 7. Note, however, that the exponential decay of the two bounds is similar for values of $\alpha$ close to zero (see the exponents in Azuma's inequality and Corollary 2 in Figure 1).

*Remark 18:* 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 $\varepsilon$ denotes the fractional gap to capacity under belief-propagation decoding (i.e., $R_{\text{d}} = (1 - \varepsilon)C$) [45]. Therefore, for small values of $\alpha$, the exponential decay rate in the inequality of Theorem 7 scales like $\left(\log \frac{1}{\varepsilon}\right)^{-2}$. This large-deviations result complements the result in [64, 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 19:* Consider small deviations from the expected value that scale like $\sqrt{n}$. Note that Corollary 2 is a special case of Theorem 3 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.

### *E. Performance of LDPC Codes under Iterative Message-Passing Decoding*

In the following, we consider ensembles of binary LDPC codes. Following standard notation, let $\lambda_i$ and $\rho_i$ denote the fraction of edges attached, respectively, to variable and parity-check nodes of degree $i$. The LDPC code ensemble that is denoted by LDPC$(n, \lambda, \rho)$ is characterized by the block length $n$ of the codes, and the pair $\lambda(x) \triangleq \sum_i \lambda_i x^{i-1}$ and $\rho(x) \triangleq \sum_i \rho_i x^{i-1}$ which represent, respectively, the left and right degree distributions from the edge perspective.

The following theorem was proved in [61, Appendix C] based on Azuma's inequality:

*Theorem 8:* **[Concentration of the bit error probability around the ensemble average]** Let $\mathcal{C}$, a code chosen uniformly at random from the ensemble LDPC$(n, \lambda, \rho)$, be used for transmission over a memoryless binary-input output-symmetric (MBIOS) channel characterized by its L-density $a_{\text{MBIOS}}$. Assume that the decoder performs $l$ iterations of message-passing decoding, and let $P_{\text{b}}(\mathcal{C}, a_{\text{MBIOS}}, l)$ denote the resulting bit error probability. Then, for every $\delta > 0$, there exists an $\alpha > 0$ where $\alpha = \alpha(\lambda, \rho, \delta, l)$ (*independent of the block length* $n$) such that

$$\mathbb{P}\left(|P_{\text{b}}(\mathcal{C}, a_{\text{MBIOS}}, l) - \mathbb{E}_{\text{LDPC}(n,\lambda,\rho)}[P_{\text{b}}(\mathcal{C}, a_{\text{MBIOS}}, l)]| \geq \delta\right) \leq \exp(-\alpha n)$$

This theorem asserts that all except an exponentially (in the block length) small fraction of codes behave within an arbitrary small $\delta$ from the ensemble average (where $\delta$ is a positive number that can be chosen arbitrarily small). Therefore, assuming a sufficiently large block length, the ensemble average is a good indicator for the performance of individual codes, and it is therefore reasonable to focus on the design and analysis of capacity-approaching ensembles (via the density evolution technique). This theorem is proved in [61, pp. 487–490] based on Azuma's inequality.

### *F. On the Conditional Entropy for LDPC Code Ensembles*

A large-deviation analysis of the conditional entropy for random ensembles of LDPC codes was introduced in [52, Theorem 4] and [54, Theorem 1]. The following theorem is proved in [52, Appendix I] based on Azuma's inequality:

*Theorem 9:* **[Large deviations of the conditional entropy]** Let $\mathcal{C}$ be chosen uniformly at random from the ensemble LDPC$(n, \lambda, \rho)$. Assume that the transmission of the code $\mathcal{C}$ takes place over an MBIOS channel. Let $H(\mathbf{X}|\mathbf{Y})$ designate the conditional entropy of the transmitted codeword $\mathbf{X}$ given the received sequence $\mathbf{Y}$ from the channel. Then for any $\xi > 0$,

$$\mathbb{P}(|H(\mathbf{X}|\mathbf{Y}) - \mathbb{E}_{\text{LDPC}(n,\lambda,\rho)}[H(\mathbf{X}|\mathbf{Y})]| \geq n\,\xi) \leq 2\exp(-nB\xi^2)$$

where $B \triangleq \frac{1}{2(d_{\text{c}}^{\max}+1)^2(1-R_d)}$, $d_{\text{c}}^{\max}$ is the maximal check-node degree, and $R_{\text{d}}$ is the design rate of the ensemble. The conditional entropy scales linearly with $n$, and this inequality considers deviations from the average which also scale linearly with $n$.

In the following, we revisit the proof of Theorem 9 in [52, Appendix I] in order to derive a tightened version of this bound. Based on this proof, let $\mathcal{G}$ be a bipartite graph which represents a code chosen uniformly at random from the ensemble LDPC$(n, \lambda, \rho)$. Define the RV

$$Z = H_{\mathcal{G}}(\mathbf{X}|\mathbf{Y})$$

which forms the conditional entropy when the transmission takes place over an MBIOS channel whose transition probability is given by $P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{n} p_{Y|X}(y_i|x_i)$ where $p_{Y|X}(y|1) = p_{Y|X}(-y|0)$. Fix an arbitrary order for the $m = n(1 - R_{\text{d}})$ parity-check nodes where $R_{\text{d}}$ forms the design rate of the LDPC code ensemble. Let $\{\mathcal{F}_t\}_{t\in\{0,1,\ldots,m\}}$ form a filtration of $\sigma$-algebras $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}_m$ where $\mathcal{F}_t$ (for $t = 0, 1, \ldots, m$) is the $\sigma$-algebra that is generated by all the sub-sets of $m \times n$ parity-check matrices that are characterized by the pair of degree distributions $(\lambda, \rho)$ and whose first $t$ parity-check equations are fixed (for $t = 0$ nothing is fixed, and therefore $\mathcal{F}_0 = \{\emptyset, \Omega\}$ where $\emptyset$ denotes the empty set, and $\Omega$ is the whole sample space of $m \times n$ binary parity-check matrices that are characterized by the pair of degree distributions $(\lambda, \rho)$). Accordingly, based on Remarks 2 and 3, let us define the following martingale sequence

$$Z_t = \mathbb{E}[Z|\mathcal{F}_t] \quad t \in \{0, 1, \ldots, m\}.$$

By construction, $Z_0 = \mathbb{E}[H_{\mathcal{G}}(\mathbf{X}|\mathbf{Y})]$ is the expected value of the conditional entropy for the LDPC code ensemble, and $Z_m$ is the RV that is equal (a.s.) to the conditional entropy of the particular code from the ensemble (see Remark 3). Similarly to [52, Appendix I], we obtain upper bounds on the differences $|Z_{t+1} - Z_t|$ and then rely on Azuma's inequality in Theorem 1.

Without loss of generality, the parity-checks are ordered in [52, Appendix I] by increasing degree. Let $\mathbf{r} = (r_1, r_2, \ldots)$ be the set of parity-check degrees in ascending order, and $\Gamma_i$ be the fraction of parity-check nodes of degree $i$. Hence, the first $m_1 = n(1-R_{\text{d}})\Gamma_{r_1}$ parity-check nodes are of degree $r_1$, the successive $m_2 = n(1-R_{\text{d}})\Gamma_{r_2}$ parity-check nodes are of degree $r_2$, and so on. The $(t+1)$th parity-check will therefore have a well defined degree, to be denoted by $r$. From the proof in [52, Appendix I]

$$|Z_{t+1} - Z_t| \leq (r + 1)\, H_{\mathcal{G}}(\tilde{X}|\mathbf{Y}) \tag{173}$$

where $H_{\mathcal{G}}(\tilde{X}|\mathbf{Y})$ is a RV which designates the conditional entropy of a parity-bit $\tilde{X} = X_{i_1} \oplus \ldots \oplus X_{i_r}$ (i.e., $\tilde{X}$ is equal to the modulo-2 sum of some $r$ bits in the codeword $\mathbf{X}$) given the received sequence $\mathbf{Y}$ at the channel output. The proof in [52, Appendix I] was then completed by upper bounding the parity-check degree $r$ by the maximal parity-check degree $d_{\text{c}}^{\max}$, and also by upper bounding the conditional entropy of the parity-bit $\tilde{X}$ by 1. This gives

$$|Z_{t+1} - Z_t| \leq d_{\text{c}}^{\max} + 1 \quad t = 0, 1, \ldots, m - 1. \tag{174}$$

which then proves Theorem 9 from Azuma's inequality. Note that the $d_i$'s in Theorem 1 are equal to $d_{\text{c}}^{\max} + 1$, and $n$ in Theorem 1 is replaced with the length $m = n(1 - R_{\text{d}})$ of the martingale sequence $\{Z_t\}$ (that is equal to the number of the parity-check nodes in the graph).

In the continuation, we deviate from the proof in [52, Appendix I] in two respects:

- The first difference is related to the upper bound on the conditional entropy $H_{\mathcal{G}}(\tilde{X}|\mathbf{Y})$ in (173) where $\tilde{X}$ is the modulo-2 sum of some $r$ bits of the transmitted codeword $\mathbf{X}$ given the channel output $\mathbf{Y}$. Instead of taking the most trivial upper bound that is equal to 1, as was done in [52, Appendix I], a simple upper bound on the conditional entropy is derived; this bound depends on the parity-check degree $r$ and the channel capacity $C$ (see Proposition 6).
- The second difference is minor, but it proves to be helpful for tightening the large-deviation inequality for LDPC code ensembles that are not right-regular (i.e., the case where the degrees of the parity-check nodes are not fixed to a certain value). Instead of upper bounding the term $r + 1$ on the right-hand side of (173) with $d_{\text{c}}^{\max} + 1$, it is suggested to leave it as is since Azuma's inequality applies to the case where the bounded differences of the martingale sequence are not fixed (see Theorem 1), and since the number of the parity-check nodes of degree $r$ is equal to $n(1-R_{\text{d}})\Gamma_r$. The effect of this simple modification will be shown in Example 10.

The following upper bound is related to the first item above:

*Proposition 6:* Let $\mathcal{G}$ be a bipartite graph which corresponds to a binary linear block code whose transmission takes place over an MBIOS channel. Let $\mathbf{X}$ and $\mathbf{Y}$ designate the transmitted codeword and received sequence at the channel output. Let $\tilde{X} = X_{i_1} \oplus \ldots \oplus X_{i_r}$ be a parity-bit of some $r$ code bits of $\mathbf{X}$. Then, the conditional entropy of $\tilde{X}$ given $\mathbf{Y}$ satisfies

$$H_{\mathcal{G}}(\tilde{X}|\mathbf{Y}) \leq h_2\left(\frac{1 - C^{\frac{r}{2}}}{2}\right). \tag{175}$$

Further, for a binary symmetric channel (BSC) or a binary erasure channel (BEC), this bound can be improved to

$$h_2\left(\frac{1-\left[1-2h_2^{-1}(1-C)\right]^r}{2}\right) \tag{176}$$

and

$$1-C^r \tag{177}$$

respectively, where $h_2^{-1}$ in (176) designates the inverse of the binary entropy function on base 2.

Note that if the MBIOS channel is perfect (i.e., its capacity is $C=1$ bit per channel use) then (175) holds with equality (where both sides of (175) are zero), whereas the trivial upper bound is 1.

*Proof:* Let us upper bound the conditional entropy $H(\tilde{X}\big|\mathbf{Y})$ with $H(\tilde{X}\big|Y_{i_1},\ldots,Y_{i_r})$, where the latter conditioning refers to the intrinsic information for the bits $X_{i_1},\ldots X_{i_r}$ which are used to calculate the parity-bit $\tilde{X}$. Then, from [64, Eq. (17) and Appendix I], the conditional entropy of the bit $\tilde{X}$ given the $n$-length received sequence $\mathbf{Y}$ satisfies the inequality

$$H(\tilde{X}\big|\mathbf{Y}) \leq 1-\frac{1}{2\ln 2}\sum_{p=1}^{\infty}\frac{(g_p)^r}{p(2p-1)} \tag{178}$$

where (see [64, Eq. (19)])

$$g_p \triangleq \int_0^{\infty} a(l)(1+e^{-l})\tanh^{2p}\left(\frac{l}{2}\right) dl, \quad \forall\, p\in\mathbb{N} \tag{179}$$

and $a(\cdot)$ denotes the symmetric *pdf* of the log-likelihood ratio at the output of the MBIOS channel, given that the channel input is equal to zero. From [64, Lemmas 4 and 5], it follows that

$$g_p \geq C^p, \quad \forall\, p\in\mathbb{N}.$$

Substituting this inequality in (178) gives that

$$\begin{aligned} H(\tilde{X}\big|\mathbf{Y}) &\leq 1-\frac{1}{2\ln 2}\sum_{p=1}^{\infty}\frac{C^{pr}}{p(2p-1)} \\ &= h_2\left(\frac{1-C^{\frac{r}{2}}}{2}\right) \end{aligned} \tag{180}$$

where the last equality follows from the power series expansion of the binary entropy function:

$$h_2(x) = 1-\frac{1}{2\ln 2}\sum_{p=1}^{\infty}\frac{(1-2x)^{2p}}{p(2p-1)}, \quad 0\leq x\leq 1. \tag{181}$$

The tightened bound on the conditional entropy for the BSC is obtained from (178) and the equality

$$g_p = \bigl(1-2h_2^{-1}(1-C)\bigr)^{2p}, \quad \forall\, p\in\mathbb{N}$$

which holds for the BSC (see [64, Eq. (97)]). This replaces $C$ on the right-hand side of (180) with $\bigl(1-2h_2^{-1}(1-C)\bigr)^2$, thus leading to the tightened bound in (176).

The tightened result for the BEC holds since from (179)

$$g_p = C, \quad \forall\, p\in\mathbb{N}$$

(see [64, Appendix II]), and a substitution of this equality in (178) gives (176) (note that $\sum_{p=1}^{\infty}\frac{1}{p(2p-1)} = 2\ln 2$). This completes the proof of Proposition 6. ■

From Proposition 6 and (173)

$$|Z_{t+1}-Z_t| \leq (r+1)h_2\left(\frac{1-C^{\frac{r}{2}}}{2}\right) \tag{182}$$

with the corresponding two improvements for the BSC and BEC (where the second term on the right-hand side of (182) is replaced by (176) and (177), respectively). This improves the loosened bound $(d_{\mathrm{c}}^{\max}+1)$ in [52, Appendix I].

From (182) and Theorem 1, we obtain the following tightened version of the large-deviation inequality in Theorem 9.

*Theorem 10:* **[A first tightened large-deviation inequality for the conditional entropy]** Let $\mathcal{C}$ be chosen uniformly at random from the ensemble LDPC$(n, \lambda, \rho)$. Assume that the transmission of the code $\mathcal{C}$ takes place over an MBIOS channel. Let $H(\mathbf{X}|\mathbf{Y})$ designate the conditional entropy of the transmitted codeword $\mathbf{X}$ given the received sequence $\mathbf{Y}$ at the channel output. Then

$$\mathbb{P}(|H(\mathbf{X}|\mathbf{Y}) - \mathbb{E}_{\text{LDPC}(n,\lambda,\rho)}[H(\mathbf{X}|\mathbf{Y})]| \geq n\,\xi) \leq 2\exp(-nB\xi^2)$$

for every $\xi > 0$, and

$$B \triangleq \frac{1}{2(1-R_{\text{d}})\sum_{i=1}^{d_{\text{c}}^{\max}} (i+1)^2 \Gamma_i \left[h_2\left(\frac{1-C^{\frac{i}{2}}}{2}\right)\right]^2} \tag{183}$$

where $d_{\text{c}}^{\max}$ is the maximal check-node degree, $R_{\text{d}}$ is the design rate of the ensemble, and $C$ is the channel capacity (in bits per channel use).

For the BSC and BEC, the parameter $B$ can be improved (increased) to

$$B \triangleq \frac{1}{2(1-R_{\text{d}})\sum_{i=1}^{d_{\text{c}}^{\max}} (i+1)^2 \Gamma_i \left[h_2\left(\frac{1-[1-2h_2^{-1}(1-C)]^i}{2}\right)\right]^2}$$

and

$$B \triangleq \frac{1}{2(1-R_{\text{d}})\sum_{i=1}^{d_{\text{c}}^{\max}} (i+1)^2 \Gamma_i \, (1-C^i)^2} \tag{184}$$

respectively

*Remark 20:* From (183), Theorem 10 indeed yields a stronger large-deviation inequality than Theorem 9.

*Remark 21:* In the limit where $C \to 1$ bit per channel use, it follows from (183) that if $d_{\text{c}}^{\max} < \infty$ then $B \to \infty$. This is in contrast to the value of $B$ in Theorem 9 which does not depend on the channel capacity and is finite. Note that $B$ should be indeed infinity for a perfect channel, and therefore Theorem 10 is tight in this case.

In the case where $d_{\text{c}}^{\max}$ is not finite, we prove the following:

*Lemma 5:* If $d_{\text{c}}^{\max} = \infty$ and $\rho'(1) < \infty$ then $B \to \infty$ in the limit where $C \to 1$.

*Proof:* See Appendix I. ■

This is in contrast to the value of $B$ in Theorem 9 which vanishes when $d_{\text{c}}^{\max} = \infty$, and therefore Theorem 9 is not informative in this case (see Example 10).

*Example 9:* [Comparison of Theorems 9 and 10 for right-regular LDPC code ensembles] In the following, we exemplify the improvement in the tightness of Theorem 10 for right-regular LDPC code ensembles. Consider the case where the communications takes place over a binary-input additive white Gaussian noise channel (BIAWGNC) or a BEC. Let us consider the $(2, 20)$ regular LDPC code ensemble whose design rate is equal to 0.900 bits per channel use. For a BEC, the threshold of the channel bit erasure probability under belief-propagation (BP) decoding is given by

$$p_{\text{BP}} = \inf_{x \in (0,1]} \frac{x}{1-(1-x)^{19}} = 0.0531$$

which corresponds to a channel capacity of $C = 0.9469$ bits per channel use. For the BIAWGNC, the threshold under BP decoding is equal to $\sigma_{\text{BP}} = 0.4156590$. From [61, Example 4.38] which expresses the capacity of the BIAWGNC in terms of the standard deviation $\sigma$ of the Gaussian noise, the minimum capacity of a BIAWGNC over which it is possible to communicate with vanishing bit error probability under BP decoding is $C = 0.9685$ bits per channel use. Accordingly, let us assume that for reliable communications on both channels, the capacity of the BEC and BIAWGNC is set to 0.98 bits per channel use.

Since the considered code ensembles is right-regular (i.e., the parity-check degree is fixed to $d_{\text{c}} = 20$), then $B$ in Theorem 10 is improved by a factor of

$$\frac{1}{\left[h_2\left(\frac{1-C^{\frac{d_{\text{c}}}{2}}}{2}\right)\right]^2} = 5.134.$$

This implies that the inequality in Theorem 10 is satisfied with a block length that is 5.134 times shorter than the block length which corresponds to Theorem 9. For the BEC, the result is improved by a factor of

$$\frac{1}{\left(1-C^{d_\text{c}}\right)^2} = 9.051$$

due to the tightened value of $B$ in (184) as compared to Theorem 9.

*Example 10:* [Comparison of Theorems 9 and 10 for a heavy-tail Poisson distribution (Tornado codes)] In the following, we compare Theorems 9 and 10 for Tornado LDPC code ensembles. This capacity-achieving sequence for the BEC refers to the heavy-tail Poisson distribution, and it was introduced in [45, Section IV], [71] (see also [61, Problem 3.20]). We rely in the following on the analysis in [64, Appendix VI].

Suppose that we wish to design Tornado code ensembles that achieve a fraction $1-\varepsilon$ of the capacity of a BEC under iterative message-passing decoding (where $\varepsilon$ can be set arbitrarily small). Let $p$ designate the bit erasure probability of the channel. The parity-check degree is Poisson distributed, and therefore the maximal degree of the parity-check nodes is infinity. Hence, $B=0$ according to Theorem 9, and this theorem therefore is useless for the considered code ensemble. On the other hand, from Theorem 10

$$\begin{aligned}
&\sum_i (i+1)^2 \Gamma_i \left[h_2\left(\frac{1-C^{\frac{i}{2}}}{2}\right)\right]^2 \\
&\overset{\text{(a)}}{\leq} \sum_i (i+1)^2 \Gamma_i \\
&\overset{\text{(b)}}{=} \frac{\sum_i \rho_i (i+2)}{\int_0^1 \rho(x)\, d\text{x}} + 1 \\
&\overset{\text{(c)}}{=} (\rho'(1)+3) d_\text{c}^\text{avg} + 1 \\
&\overset{\text{(d)}}{=} \left(\frac{\lambda'(0)\rho'(1)}{\lambda_2} + 3\right) d_\text{c}^\text{avg} + 1 \\
&\overset{\text{(e)}}{\leq} \left(\frac{1}{p\lambda_2} + 3\right) d_\text{c}^\text{avg} + 1 \\
&\overset{\text{(f)}}{=} O\left(\log^2(\frac{1}{\varepsilon})\right)
\end{aligned}$$

where inequality (a) holds since the binary entropy function on base 2 is bounded between zero and one, equality (b) holds since

$$\Gamma_i = \frac{\frac{\rho_i}{i}}{\int_0^1 \rho(x)\,\mathrm{d}x}$$

where $\Gamma_i$ and $\rho_i$ denote the fraction of parity-check nodes and the fraction of edges that are connected to parity-check nodes of degree i respectively (and also since $\sum_i \Gamma_i = 1$), equality (c) holds since

$$d_\text{c}^\text{avg} = \frac{1}{\int_0^1 \rho(x)\,\mathrm{d}x}$$

where $d_\text{c}^\text{avg}$ denotes the average parity-check node degree, equality (d) holds since $\lambda'(0) = \lambda_2$, inequality (e) is due to the stability condition for the BEC (where $p\lambda'(0)\rho'(1) < 1$ is a necessary condition for reliable communication on the BEC under BP decoding), and finally equality (f) follows from the analysis in [64, Appendix VI] (an upper bound on $\lambda_2$ is derived in [64, Eq. (120)], and the average parity-check node degree scales like $\log\frac{1}{\varepsilon}$). Hence, from the above chain of inequalities and (183), it follows that for a small gap to capacity, the parameter $B$ in Theorem 10 scales (at least) like

$$O\left(\frac{1}{\log^2(\frac{1}{\varepsilon})}\right).$$

Theorem 10 is therefore useful for the large-deviation analysis of this LDPC code ensemble. As shown above, the parameter $B$ in (183) tends to zero rather slowly as we let the fractional gap $\varepsilon$ tend to zero (which therefore demonstrates a rather fast concentration in Theorem 10).

*Example 11:* This Example forms a direct continuation of Example 9 for the $(n, d_{\text{v}}, d_{\text{c}})$ regular LDPC code ensembles where $d_{\text{v}} = 2$ and $d_{\text{c}} = 20$. With the settings in this example, Theorem 9 gives that

$$\begin{aligned} & \mathbb{P}(|H(\mathbf{X}|\mathbf{Y}) - \mathbb{E}_{\text{LDPC}(n,\lambda,\rho)}[H(\mathbf{X}|\mathbf{Y})]| \geq n\,\xi) \\ & \leq 2\exp(-0.0113 n\xi^2) \end{aligned} \tag{185}$$

for every $\xi > 0$. As was mentioned already in Example 9, the exponential inequalities in Theorem 10 achieve an improvement in the exponent of Theorem 9 by factors 5.134 and 9.051 for the BIAWGNC and BEC, respectively. One therefore obtains via the inequalities in Theorem 10 that for every $\xi > 0$

$$\begin{aligned} & \mathbb{P}(|H(\mathbf{X}|\mathbf{Y}) - \mathbb{E}_{\text{LDPC}(n,\lambda,\rho)}[H(\mathbf{X}|\mathbf{Y})]| \geq n\,\xi) \\ & \leq \begin{cases} 2\exp(-0.0580 n\xi^2) & \text{BIAWGNC} \\ 2\exp(-0.1023 n\xi^2), & \text{BEC} \end{cases} . \end{aligned} \tag{186}$$

### G. Concentration Theorems for LDPC Code Ensembles over ISI channels

Concentration analysis on the number of erroneous variable-to-check messages for random ensembles of LDPC codes was introduced in [46] and [60] for memoryless channels. It was shown that the performance of an individual code from the ensemble concentrates around the expected (average) value over this ensemble when the length of the block length of the code grows and that this average behavior converges to the behavior of the cycle-free case. These results were later generalized in [36] for the case of channels with memory (i.e., for ISI channels). In this section, we revisit the proofs of [36, Theorems 1 and 2] for the case of regular LDPC code ensembles in order to derive an explicit expression for the exponential rate that is related to the concentration inequality. It is then shown that particularizing the expression for memoryless channels provides a tightened concentration inequality as compared to [46] and [60].

*1) The ISI Channel and its message-passing decoding:* In the following, we briefly describe the ISI channel and the graph used for its message-passing decoding. For a detailed description, the reader is referred to [36]. Consider a binary discrete-time ISI channel with a finite memory length, to be denoted by $I$ . The channel's output $Y_t$ at time $t \in \mathbb{Z}$ is given by the equality

$$Y_t = \sum_{i=0}^{I} h_i X_{t-i} + N_t$$

where $X_t \in \{+1, -1\}$ is the channel's input, $h_i$ is the channel's response and $N_t \sim N(0, \sigma^2)$ is an i.i.d. AWGN noise sequence. The information block of length $k$ is coded using a regular $(n, d_{\text{v}}, d_{\text{c}})$ LDPC code, and the resulting $n$ coded bits are converted to $X_t \in \{+1, -1\}$ before transmission over the channel. For decoding we consider the windowed version of the "sum-product" algorithm when applied to ISI channels (see details in [36] and [22]). As in the memoryless case, this is a message passing algorithm. The variable-to-check and check-to-variable messages are computed as in the min-sum algorithm for the memoryless case with the difference that a variable node's message from the channel is not only a function of the the channel output that corresponds to the considered symbol but also a function of $2W$ neighboring channel outputs and $2W$ neighboring variables nodes as illustrated in Fig. 3.

*2) Concentration:* It is proved in this sub-section that for a large $n$, a neighborhood of depth $\ell$ of a variable-to-check node message is tree-like with high probability. Using Azuma's inequality and the later result, it is shown that for most graphs and channel realizations, if $\underline{s}$ is the transmitted codeword, then the probability of a variable-to-check message being erroneous after $\ell$ rounds of message-passing decoding is highly concentrated around its expected value. This expected value is shown to converge to the value of $p^{(\ell)}(\underline{s})$ which corresponds to the cycle-free case. Also, we prove that if the transmitted sequence is i.u.d., then the probability highly concentrated around the value $p^{(\ell)}_{i.u.d.} \equiv \mathbb{E}[p^{(\ell)}(\underline{s})]$.

In the following theorems, we consider an ISI channel and windowed message-passing decoding algorithm, when the code graph is chosen uniformly at random from the ensemble of the graphs with variable and check node degree

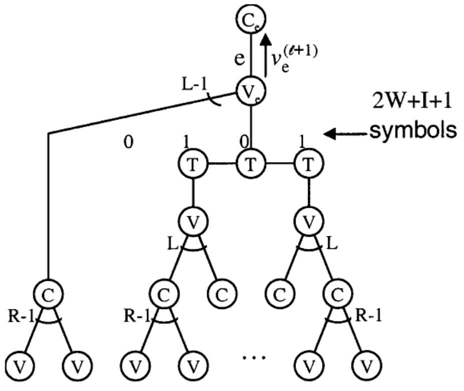

Fig. 3. Message flow neighborhood of depth 1. In this figure $(I, W, d_\mathrm{v} = L, d_\mathrm{c} = R) = (1, 1, 2, 3)$

$d_\mathrm{v}$ and $d_\mathrm{c}$ respectively. Denote $\mathcal{N}_{\vec{e}}^{\ell}$ as the neighborhood of depth $\ell$ of an edge $\vec{e} = (\mathrm{v}, \mathrm{c})$ between a variable-to-check node. Let $N_\mathrm{c}^{\ell}$, $N_\mathrm{v}^{\ell}$ and $N_e^{\ell}$ to the be total number of check nodes, variable nodes and code related edges respectively in this neighborhood. Similarly denote $N_Y^{\ell}$ as the number of variable-to-check node messages in the directed neighborhood of depth $\ell$ of a received value of the channel.

*Theorem 11:* **[Probability of a neighborhood of depth $\ell$ of a variable-to-check node message to be tree-like for channels with ISI]** Define $P_{\bar{t}}^{\ell} \equiv \Pr\left\{\mathcal{N}_{\vec{e}}^{\ell} \text{ not a tree}\right\}$ as the probability that $\mathcal{N}_{\vec{e}}^{\ell}$ is not tree-like. Then, there exists a positive constant $\gamma(d_v, d_c, \ell) = N_v^{\ell^2} + \frac{d_c}{d_v} N_v^{\ell^2}$ *such that* $P_{\bar{t}}^{\ell} \leq \frac{\gamma}{n}$.

*Proof:* This proof follows from the proof in [60] and extends it to the case of ISI channels. Consider a neighborhood $\mathcal{N}_{\vec{e}}^{\ell}$ of fixed depth $\ell$ . Note that at each level the graph expands by factor $\alpha \equiv (d_\mathrm{v} - 1 + 2W d_\mathrm{v})(d_\mathrm{c} - 1)$, therefore there are in total

$$N_\mathrm{v}^{\ell} = 1 + [(d_\mathrm{v} - 1)(d_\mathrm{c} - 1) + 2W(1 + d_\mathrm{v}(d_\mathrm{c} - 1))] \sum_{i=0}^{\ell-1} \alpha^i$$

variable nodes and

$$N_\mathrm{c}^{\ell} = 1 + (d_\mathrm{v} - 1 + 2W d_\mathrm{v}) \sum_{i=0}^{\ell-1} \alpha^i$$

check nodes in this neighborhood.

In order to lower bound $P_{\bar{t}}^{\ell}$ we can upper bound $P_t^{\ell} = 1 - P_{\bar{t}}^{\ell}$. This is done by factorizing $P_t^{\ell}$ as

$$P_t^{\ell} = \Pr\left\{\mathcal{N}_{\vec{e}}^{0} \text{ is tree}\right\} \prod_{\ell^*=0}^{\ell-1} \Pr\left\{\mathcal{N}_{\vec{e}}^{\ell^*+1} \text{ is tree} | \mathcal{N}_{\vec{e}}^{\ell^*} \text{ is tree}\right\} \tag{187}$$

and bounding each factor. For $\ell^* = 0$ we have a single edge which is a tree, therefore $\Pr\left\{\mathcal{N}_{\vec{e}}^{0} \text{ is tree}\right\} = 1$. To bound $\Pr\left\{\mathcal{N}_{\vec{e}}^{\ell^*+1} \text{ is tree} | \mathcal{N}_{\vec{e}}^{\ell^*} \text{ is tree}\right\}$ we assume that $\mathcal{N}_{\vec{e}}^{\ell^*}$ is tree-like and reveal the code related edges (variable-to-check node or vice versa, as opposed to the channel related edges which are predetermined) one at a time. If in this process (of revealing the $\ell^* + 1$-th level of the tree) no loops are created then $\mathcal{N}_{\vec{e}}^{\ell^*+1}$ is also a tree. We start by revealing the leaves of a variable node . As opposed to the case with no ISI, where each variable node has only $d_\mathrm{v} - 1$ direct paths to check nodes from the next level, here also $2W d_\mathrm{v}$ indirect paths through trellis nodes exist (i.e, variable-trellis-variable-check). Since the edges connected to the trellis nodes are predetermined then also the indirect path requires the revelation of a single variable-to-check node edge. Assume that $k$ additional edges have been revealed at this stage without creating a loop. The next revealed edge is chosen among $m d_\mathrm{c} - k - N_\mathrm{c}^{\ell^*}$ edges and it does not create a loop if it is connected to one of the $(m - k - N_\mathrm{c}^{\ell^*})$ un-explored check nodes. Since each un-explored check node has $d_{\mathrm{c}}$ edges then the probability for not creating a loop is given by $\frac{(m-k-N_{\mathrm{c}}^{\ell^*})d_{\mathrm{c}}}{md_{\mathrm{c}}-k-N_{\mathrm{c}}^{\ell^*}}$. For large $n$ we have

$$\begin{aligned}&\frac{(m-k-N_{\mathrm{c}}^{\ell^*})d_{\mathrm{c}}}{md_{\mathrm{c}}-k-N_{\mathrm{c}}^{\ell^*}}\\&=\frac{(N_{\mathrm{c}}^{\ell^*}+k)(d_{\mathrm{c}}-1)}{md_{\mathrm{c}}-k-N_{\mathrm{c}}^{\ell^*}}\\&\geq 1-\frac{N_{\mathrm{c}}^{\ell}}{m}\end{aligned} \tag{188}$$

We have $N_{\mathrm{c}}^{\ell^*+1}-N_{\mathrm{c}}^{\ell^*}$ edges to reveal (one for each check node), therefore, the probability that revealing all variable node leaves does not creates a loop, given $\mathcal{N}_{\vec{e}}^{\ell^*}$ is tree-like is lower bounded by $\left(1-\frac{N_{\mathrm{c}}^{\ell}}{m}\right)^{N_{\mathrm{c}}^{\ell^*+1}-N_{\mathrm{c}}^{\ell^*}}$. Next, we reveal the outgoing edges of the check node leaves one at a time (here only $d_{\mathrm{c}}$ direct paths exist, as in the case without ISI). Assuming $k$ variable nodes have been revealed without creating a loop, then the probability that the next revealed edge does no create a loop is $\frac{(n-k-N_{\mathrm{v}}^{\ell^*})d_{\mathrm{v}}}{nd_{\mathrm{v}}-k-N_{\mathrm{v}}^{\ell^*}}$. For large $n$ we have

$$\begin{aligned}&\frac{(n-k-N_{\mathrm{v}}^{\ell^*})d_{\mathrm{v}}}{nd_{\mathrm{v}}-k-N_{\mathrm{v}}^{\ell^*}}\\&=\frac{(N_{\mathrm{v}}^{\ell^*}+k)(d_{\mathrm{v}}-1)}{nd_{\mathrm{v}}-k-N_{\mathrm{c}}^{\ell^*}}\\&\geq 1-\frac{N_{\mathrm{v}}^{\ell}}{n}.\end{aligned} \tag{189}$$

We have $N_{\mathrm{v}}^{\ell^*+1}-N_{\mathrm{v}}^{\ell^*}$ edges to reveal (one for each variable node), therefore, the probability that revealing all check node leaves does not creates loop, given the neighborhood is tree-like so far is lower bounded by $\left(1-\frac{N_{\mathrm{v}}^{\ell}}{n}\right)^{N_{\mathrm{v}}^{\ell^*+1}-N_{\mathrm{v}}^{\ell^*}}$. Combining (187), (189), (188) and $P_t^\ell=1-P_{\bar{t}}^\ell$ we have

$$P_{\bar{t}}^\ell\leq 1-\left(1-\frac{N_{\mathrm{v}}^{\ell}}{n}\right)^{N_{\mathrm{v}}^{\ell}}\left(1-\frac{N_{\mathrm{c}}^{\ell}}{m}\right)^{N_{\mathrm{c}}^{\ell}}$$

Thus, for $n$ sufficiently large

$$P_{\bar{t}}^\ell\leq\frac{N_{\mathrm{v}}^{\ell 2}+\frac{d_{\mathrm{c}}}{d_{\mathrm{v}}}N_{\mathrm{c}}^{\ell 2}}{n}\equiv\frac{\gamma}{n}$$

■

*Theorem 12:* **[Concentration of the number of erroneous variable-to-check messages for channels with ISI]** Let $\underline{s}$ be the transmitted codeword. Let $Z^\ell(\underline{s})$ be the number of erroneous variable-to-check messages after $\ell$ rounds of the windowed message-passing decoding algorithm when the code graph is chosen uniformly at random from the ensemble of the graphs with variable and check node degree $d_{\mathrm{v}}$ and $d_{\mathrm{c}}$ respectively. Let $p^{(\ell)}(\underline{s})$ be the expected fraction of incorrect messages passed along an edge with a tree-like directed neighborhood of depth $\ell$. Then, there exist positive constants $\beta(d_{\mathrm{v}},d_{\mathrm{c}},\ell)=\frac{d_{\mathrm{v}}^2}{8(4d_{\mathrm{v}}(N_e^\ell)^2+(N_Y^\ell)^2)}$ and $\gamma(d_{\mathrm{v}},d_{\mathrm{c}},\ell)=N_{\mathrm{v}}^{\ell 2}+\frac{d_{\mathrm{c}}}{d_{\mathrm{v}}}N_{\mathrm{c}}^{\ell 2}$ such that

[**Concentration around expectation**] For any $\epsilon>0$ we have

$$\Pr\left\{\left|\frac{Z^\ell(\underline{s})}{nd_v}-\frac{\mathbb{E}[Z^\ell(\underline{s})]}{nd_v}\right|>\epsilon/2\right\}\leq 2e^{-\beta\epsilon^2 n} \tag{190}$$

[**Convergence to cycle-free case**] For any $\epsilon>0$ and $n>\frac{2\gamma}{\epsilon}$ we have

$$\left|\frac{\mathbb{E}[Z^\ell(\underline{s})]}{nd_v}-p^{(\ell)}(\underline{s})\right|<\epsilon/2 \tag{191}$$

[**Concentration around cycle-free case**] For any $\epsilon>0$ and $n>\frac{2\gamma}{\epsilon}$ we have

$$\Pr\left\{\left|\frac{Z^\ell(\underline{s})}{nd_v}-p^{(\ell)}(\underline{s})\right|>\epsilon\right\}\leq 2e^{-\beta\epsilon^2 n} \tag{192}$$

*Proof:* First note that for $n > \frac{2\gamma}{\epsilon}$ the following inequality holds

$$\begin{aligned} &\Pr\left\{\left|\frac{Z^\ell(\underline{s})}{nd_\text{v}} - p^{(\ell)}(\underline{s})\right| > \epsilon\right\} \\ &\leq \Pr\left\{\left|\frac{Z^\ell(\underline{s})}{nd_\text{v}} - \frac{\mathbb{E}[Z^\ell(\underline{s})]}{nd_\text{v}}\right| > \epsilon/2\right\} + \Pr\left\{\left|\frac{\mathbb{E}[Z^\ell(\underline{s})]}{nd_\text{v}} - p^{(\ell)}(\underline{s})\right| > \epsilon/2\right\} \end{aligned} \tag{193}$$

If inequality (191) holds, then $\Pr\left\{\left|\frac{Z^\ell(\underline{s})}{nd_\text{v}} - p^{(\ell)}(\underline{s})\right| > \epsilon/2\right\} = 0$, therefore using (193) we deduce that (192) follows from (190) and (191). We start by proving (190). For a deterministic sequence $\underline{s}$ the random variable $Z^\ell(\underline{s})$ denotes the number of incorrect variable-to-check node messages among all $nd_\text{v}$ variable-to-check node messages passed in the $\ell$th iteration for a particular graph $\mathcal{G}$ and decoder's input $\underline{Y}$. Let us form a Doob's martingale by first exposing the $nd_\text{v}$ edges of the graph one by one and then exposing the $n$ received values $Y_i$ one by one. For $i = 0, ...n(d_\text{v}+1)$, define the RV $\tilde{Z}_i = \mathbb{E}[Z^\ell(\underline{s})|a_1, ...a_i]$, where the sequence $a$ is the sequence of the $nd_\text{v}$ variable-to-check node edges of the graph followed by the sequence of the $n$ received values. Note that it is a martingale sequence where $\tilde{Z}_0 = \mathbb{E}[Z^\ell(\underline{s})]$ and $\tilde{Z}_{n(d_\text{v}+1)} = Z^\ell(\underline{s})$. We can use Azuma's inequality if we can bound the sequence of differences $|\tilde{Z}_{i+1} - \tilde{Z}_i| \leq d_i$.

We now consider the effect of exposing an edge of the graph. Consider two graphs $\mathcal{G}$ and $\tilde{\mathcal{G}}$ whose edges are identical except an exchange of the endpoint of two edges. A variable-to-check message is affected by this change if one (or both) of the edges is in its directed neighborhood of depth $\ell$.

Consider a neighborhood of depth $\ell$ of a variable-to-check node message. Since at each level the graph expands by factor $\alpha \equiv (d_\text{v} - 1 + 2Wd_\text{v})(d_\text{c} - 1)$ then there are, in total

$$N_e^\ell = 1 + d_\text{c}(d_\text{v} - 1 + 2Wd_\text{v}) \sum_{i=0}^{\ell-1} \alpha^i$$

edges related to the code structure (variable-to-check node edges or vice versa) in the neighborhood $\mathcal{N}_\ell^{\vec{e}}$. By symmetry the two edges can affect at most $2N_e^\ell$ neighborhoods (Alternatively we could directly sum the number of variable-to-check node edges in a neighborhood of a variable-to-check node edge and in a neighborhood of a check-to-variable node edge). The change in the number of incorrect variable-to-check node messages is bounded by the case that each change in the neighborhood of a message introduces an error. In a similar manner, when we reveal a received value, then variable-to-check node messages whose directed neighborhood include that channel input can be affected. We consider a neighborhood of depth $\ell$ of a received value. By counting, it can be shown that this neighborhood includes

$$N_Y^\ell = d_\text{v}(2W + 1) \sum_{i=0}^{\ell-1} \alpha^i$$

variable-to-check node edges. Therefore a change in a received value can affect up to $N_Y^\ell$ variable-to-check node messages. We conclude that $d_i \leq 2N_e^\ell$ for the first $d_\text{v}n$ exposures and $d_i \leq N_Y^\ell$ for the last $n$ exposures. By applying Azuma's inequality we get

$$\Pr\left\{\left|\frac{Z^\ell(\underline{s})}{nd_\text{v}} - \frac{\mathbb{E}[Z^\ell(\underline{s})}{nd_\text{v}}]\right| > \epsilon/2\right\} \leq e^{-\frac{(nd_\text{v}\epsilon/2)^2}{2[nd_\text{v}(2N_e^\ell)^2 + n(N_Y^\ell)^2]}}$$

By comparing the result to (190), we get an expression for $\beta$

$$\frac{1}{\beta} = 8\left(4d_\text{v}(N_e^\ell)^2 + (N_Y^\ell)^2\right)/d_\text{v}^2$$

Next, we prove inequality (191), again it is adopted from [60] and [36]. Let $\mathbb{E}[Z_i^\ell(\underline{s})], i \in [nd_\text{v}]$ be the expected number of incorrect messages passed along edge $\overrightarrow{e}_i$, where the average is over all graphs and all received values. Then by linearity of expectation and by symmetry

$$\mathbb{E}[Z^\ell(\underline{s})] = \sum_{i\in[nd_\text{v}]} \mathbb{E}[Z_i^\ell(\underline{s})] = nd_\text{v}\mathbb{E}[Z_1^\ell(\underline{s})]. \tag{194}$$

Furthermore

$$\mathbb{E}[Z_1^\ell(\underline{s})] = \mathbb{E}[Z_1^\ell(\underline{s})|\mathcal{N}_{\vec{e}}^\ell \text{ is tree}]P_t^\ell + \mathbb{E}[Z_1^\ell(\underline{s})|\mathcal{N}_{\vec{e}}^\ell \text{ not a tree}]P_{\bar{t}}^\ell$$

As shown in Theorem 11, $P_{\bar{t}}^\ell \leq \frac{\gamma}{n}$ where $\gamma$ is a positive constant independent of $n$. Furthermore, we have $\mathbb{E}[Z_1^\ell(\underline{s}) \,|\, \text{neighborhood is tree}] = p^{(\ell)}(\underline{s})$ and by definition $0 \leq \mathbb{E}[Z_1^\ell(\underline{s})|\text{neighborhood not a tree}] \leq 1$. Hence

$$\begin{aligned}
\mathbb{E}[Z_1^\ell(\underline{s})] &\leq (1 - P_{\bar{t}}^\ell)p^{(\ell)}(\underline{s}) + P_{\bar{t}}^\ell \leq p^{(\ell)}(\underline{s}) + P_{\bar{t}}^\ell \\
\mathbb{E}[Z_1^\ell(\underline{s})] &\geq (1 - P_{\bar{t}}^\ell)p^{(\ell)}(\underline{s}) \geq p^{(\ell)}(\underline{s}) - P_{\bar{t}}^\ell .
\end{aligned} \tag{195}$$

Using (194), (195) and $P_{\bar{t}}^\ell \leq \frac{\gamma}{n}$ we get that

$$\left| \frac{\mathbb{E}[Z^\ell(\underline{s})]}{nd_\text{v}} - p^{(\ell)}(\underline{s}) \right| \leq P_{\bar{t}}^\ell \leq \frac{\gamma}{n}$$

It follows that if $n > \frac{2\gamma}{\epsilon}$ then (191) holds. ■

*Discussion 2:* The concentration result proved above is a generalization of the results given in [60] for the memoryless case. One can degenerate the expression $\frac{1}{\beta} = 8\left(4d_\text{v}(N_e^\ell)^2 + (N_Y^\ell)^2\right)/d_\text{v}^2$ to the memoryless case by setting $W = 0$ and $I = 0$. Since we used exact expressions for $N_e^\ell$ and $N_Y^\ell$ in the proof, we can expect a tighter bound as compared to the earlier result $\frac{1}{\beta_\text{old}} = 544 d_\text{v}^{2\ell-1} d_\text{c}^{2\ell}$ given in [60]. For example for $(d_\text{v}, d_\text{c}, \ell) = (3, 4, 10)$ we get an improvement by a factor of about 1 million. However even with this improved expression, the required size of $n$ according to our proof can be absurdly large. This is because the proof is very pessimistic. We assume that any change in an edge or the decoder's input will introduce an error in every message it affects. This is especially pessimistic if large $\ell$ is considered, since as $\ell$ grows each message is a function of many edges and received values (since the neighborhood grows with $\ell$). However in practice, the probability that changing a single edge or input will change the message is close to zero for long codes.

*Theorem 13:* Let $\underline{s}$ be a random sequence of i.u.d. binary variables $S_1, S_2....S_n$. Let $Z^\ell(\underline{S})$ be the number of erroneous variable-to-check messages after $\ell$ rounds of the windowed message-passing decoding algorithm when the code graph is chosen uniformly at random from the ensemble of the graphs with variable and check node degree $d_\text{v}$ and $d_\text{c}$ respectively. Let $p_{i.u.d.}^{(\ell)} \equiv \mathbb{E}[p^{(\ell)}(\underline{s})]$ be the expected fraction of incorrect messages passed along an edge with a tree-like directed neighborhood of depth $\ell$. Then, there exist positive constants $\beta' = \beta(d_v, d_c, \ell)$, and $\gamma = \gamma(d_v, d_c, \ell)$ such that for any $\epsilon > 0$ and $n > \frac{2\gamma}{\epsilon}$ we have

$$\Pr\left\{ \left| \frac{Z^\ell(\underline{S})}{nd_v} - p_{i.u.d.}^{(\ell)} \right| > \epsilon \right\} \leq 4e^{-\beta'\epsilon^2 n} \tag{196}$$

Furthermore, $p_{i.u.d.}^{(\ell)}$ is equal to the error probability when all neighborhood types are equally probable.

*Proof:* The proof follows closely the one presented in [36]. First, note that the following chain of inequalities hold

$$\begin{aligned}
&\Pr\left\{ \left| \frac{Z^\ell(\underline{S})}{nd_\text{v}} - p_\text{i.u.d.}^{(\ell)} \right| > \epsilon \right\} \\
&= \sum_{j=1}^{2^n} 2^{-n} \Pr\left\{ \left| \frac{Z^\ell(\underline{s_j})}{nd_\text{v}} - p_\text{i.u.d.}^{(\ell)} \right| > \epsilon \right\} \\
&\leq \sum_{j=1}^{2^n} 2^{-n} \Pr\left\{ \left| \frac{Z^\ell(\underline{s_j})}{nd_\text{v}} - p^{(\ell)}(\underline{s_j}) \right| > \epsilon/2 \right\} + \sum_{j=1}^{2^n} 2^{-n} \Pr\left\{ \left| p^{(\ell)}(\underline{s_j}) - p_\text{i.u.d.}^{(\ell)} \right| > \epsilon/2 \right\} \\
&\leq \sum_{j=1}^{2^n} 2^{-n} \cdot 2e^{-\beta\epsilon^2 n/4} + \Pr\left\{ \left| p^{(\ell)}(\underline{S}) - p_\text{i.u.d.}^{(\ell)} \right| > \epsilon/2 \right\} \\
&= 2e^{-\beta\epsilon^2 n/4} + \Pr\left\{ \left| p^{(\ell)}(\underline{S}) - p_{i.u.d.}^{(\ell)} \right| > \epsilon/2 \right\}
\end{aligned} \tag{197}$$

To bound the second term in the last line we shall use Azuma's inequality. Let us form a Doob's martingale by exposing the $n$ received symbols one by one. For $t = 1, ..., n$, define the RV $M_t = \mathbb{E}[p^\ell(\underline{S})|S_1, S_2, ..., S_t]$. Note that $M_0 = \mathbb{E}[p^\ell(\underline{S})] = p_{i.u.d.}^{(\ell)}$ and $M_n = \mathbb{E}[p^\ell(\underline{S})|S_1, S_2, ..., S_n] = p^\ell(\underline{S})$. In order to use Azuma's inequality we shall show that the sequence of differences is bounded $|M_{t+1} - M_t| \leq d_t$. Since the channel has ISI of degree I, then exposing a single channel input affects I channel output (which are the received values for the decoder). A variable-to-check node message is affected only if one of the affected received values are in its neighborhood. Therefore, changing a channel input can affect at most $IN_Y^\ell$ variable-to-check node messages among the $nd_{\text{v}}$ messages in the graph. Thus $|M_{t+1} - M_t| \leq \frac{IN_Y^\ell}{nd_{\text{v}}}$, and by using Azuma's inequality we have

$$\Pr\left\{\left|p^{(\ell)}(\underline{S}) - p^{(\ell)}_{i.u.d.}\right| > \epsilon/2\right\} \leq 2e^{-\delta\epsilon^2 n} \tag{198}$$

where $\delta = \frac{1}{8}\left(\frac{d_{\text{v}}}{IN_Y^\ell}\right)^2$. Combining (198), (197) and comparing it to (196) gives that $\beta' = \min(\beta, \delta)$.

Next, we get an expression for $p^{(\ell)}_{i.u.d.}$ and show it is equal to the error probability when all neighborhood types are equally probable. In Fig. 3, a depth 1 message-flow neighborhood is shown. The row of bits ”0101” given above the trellis section represent the binary symbols of the codeword $\underline{S}$ corresponding to the trellis nodes that influence the message flow. Since the channel has ISI memory of length $I$, there are $2W + I + 1$ binary symbols of that influence the message flow. we call this sequence of bits a neighborhood type. For example, Fig. 3, the neighborhood type is $\theta = [0101]$. We expand this definition to a depth $\ell$ neighborhood by cascading the bits of each sub-neighborhood of depth $\ell$. Since at each level, the graph expends by factor $\alpha \equiv (d_{\text{v}} - 1 + 2Wd_{\text{v}})(d_{\text{c}} - 1)$ then there are exactly $2^{N(\ell)}$ possible types of message flow neighborhoods of depth $\ell$, where

$$N(\ell) = (2W + I + 1)\sum_{i=0}^{\ell-1} \alpha^i = (2W + I + 1)\frac{\alpha^\ell - 1}{\alpha - 1}$$

We can now define

$$\pi^\ell_{\underline{\theta}} = \Pr\left(\text{tree delivers incorrect message}|\text{tree type } \theta\right)$$

and

$$P(\underline{\theta}|\underline{s}) = \Pr\left(\text{tree type } \theta|\text{transmitted sequence} = \underline{s}\right)$$

Therefore we can express $p^{(\ell)}(\underline{S})$ as

$$p^{(\ell)}(\underline{S}) = \sum_{i=0}^{2^{N(\ell)}} \pi^{(\ell)}_{\underline{\theta}_i} \Pr(\underline{\theta}_i|\underline{S}).$$

Next, recognize that if $\underline{S}$ is an i.u.d. sequence, all neighborhood types are equally probable, i.e. $\Pr\left(\underline{\theta}|\underline{S}\right) = 2^{-N(\ell)}$. Using this we have

$$\begin{aligned}
\mathbb{E}[p^{(\ell)}(\underline{S})] &= \sum_{j=0}^{2^n} 2^{-n} p^{(\ell)}(\underline{s_j}) \\
&= \sum_{j=0}^{2^n} 2^{-n} \sum_{i=0}^{2^{N(\ell)}} \pi^{(\ell)}_{\underline{\theta}_i} \Pr(\underline{\theta}_i|\underline{s_j}) \\
&= \sum_{j=0}^{2^{N(\ell)}} \pi^{(\ell)}_{\underline{\theta}_i} \sum_{i=0}^{2^n} 2^{-n} \Pr(\underline{\theta}_i|\underline{s_j}) \\
&= \sum_{i=0}^{2^{N(\ell)}} \pi^{(\ell)}_{\underline{\theta}_i} \Pr(\underline{\theta}_i|\underline{S}) \\
&= \sum_{i=0}^{2^{N(\ell)}} \pi^{(\ell)}_{\underline{\theta}_i} 2^{-N(\ell)}
\end{aligned}$$

The last term is equal to the error probability when all neighborhood types are equally probable. Since $\mathbb{E}[p^{(\ell)}(\underline{S})] = p^{(\ell)}_{i.u.d.}$ the theorem is proved.

■

### *H. Expansion of Random Regular Bipartite Graphs*

Azuma's inequality is useful for analyzing the expansion of random bipartite graphs. The following theorem was introduced in [72, Theorem 25]. It is stated and proved here slightly more precisely, in the sense of characterizing the relation between the deviation from the expected value and the exponential convergence rate of the resulting probability.

*Theorem 14:* **[Expansion of random regular bipartite graphs]** Let $\mathcal{G}$ be chosen uniformly at random from the regular ensemble LDPC$(n, x^{l-1}, x^{r-1})$. Let $\alpha \in (0,1)$ and $\delta > 0$ be fixed. Then, with probability at least $1 - \exp(-\delta n)$, all sets of $\alpha n$ variables in $\mathcal{G}$ have a number of neighbors that is at least

$$n \left[ \frac{l\big(1-(1-\alpha)^r\big)}{r} - \sqrt{2l\alpha\left(h(\alpha)+\delta\right)} \right] \tag{199}$$

where $h$ designates the binary entropy function to the natural base (i.e., $h(x) = -x\ln(x) - (1-x)\ln(1-x)$ for $x \in [0,1]$).

*Proof:* The proof starts by looking at the expected number of neighbors, and then exposing one neighbor at a time to bound the probability that the number of neighbors deviates significantly from this mean.

Note that the number of expected neighbors of $\alpha n$ variable nodes is equal to

$$\frac{nl\big(1-(1-\alpha)^r\big)}{r}$$

since for each of the $\frac{nl}{r}$ check nodes, the probability that it has at least one edge in the subset of $n\alpha$ chosen variable nodes is $1-(1-\alpha)^r$. Let us form a martingale sequence to estimate, via Azuma's inequality, the probability that the actual number of neighbors deviates by a certain amount from this expected value.

Let $\mathcal{V}$ denote the set of $n\alpha$ nodes. This set has $n\alpha l$ outgoing edges. Let us reveal the destination of each of these edges one at a time. More precisely, let $S_i$ be the RV denoting the check-node socket which the $i$-th edge is connected to, where $i \in \{1, \ldots, n\alpha l\}$. Let $X(\mathcal{G})$ be a RV which denotes the number of neighbors of a chosen set of $n\alpha$ variable nodes in a bipartite graph $\mathcal{G}$ from the ensemble, and define for $i = 0, \ldots, n\alpha l$

$$X_i = \mathbb{E}[X(\mathcal{G}) | S_1, \ldots, S_{i-1}].$$

Note that it is a martingale sequence where $X_0 = \mathbb{E}[X(\mathcal{G})]$ and $X_{n\alpha l} = X(\mathcal{G})$. Also for every $i \in \{1, \ldots, n\alpha l\}$, we have $|X_i - X_{i-1}| \leq 1$ since every time only one check-node socket is revealed, so the number of neighbors of the chosen set of variable nodes cannot change by more than 1 at every single time. Thus, by the one-sided Azuma's inequality derived in Section III-A

$$\mathbb{P}\big(\mathbb{E}[X(\mathcal{G})] - X(\mathcal{G}) \geq \lambda\sqrt{l\alpha n}\big) \leq \exp\big(-\frac{\lambda^2}{2}\big), \quad \forall\, \lambda > 0.$$

Since there are $\binom{n}{n\alpha}$ choices for the set $\mathcal{V}$ then, from the union bound, the event that there exists a set of size $n\alpha$ whose number of neighbors is less than $\mathbb{E}[X(\mathcal{G})] - \lambda\sqrt{l\alpha n}$ occurs with probability that is at most $\binom{n}{n\alpha}\exp\big(-\frac{\lambda^2}{2}\big)$. Since $\binom{n}{n\alpha} \leq e^{nh(\alpha)}$, then we get the loosened bound

$$\exp\Big(nh(\alpha) - \frac{\lambda^2}{2}\Big).$$

Finally, the choice $\lambda = \sqrt{2n\big(h(\alpha)+\delta\big)}$ gives the required result. ■

*Remark 22:* It is noted that Theorem 14 uniformly improves the statement in [61, Problem C.4] for every $\delta > 0$. This holds even in the case where $\alpha \to 1$ (i.e., when considering a set that includes almost all the variable nodes in the bipartite graph, and whose number of neighbors is expected to be close to $\frac{nl}{r}$). The expression which appears there, instead of (199), is given by

$$n \left[ \frac{l\big(1-(1-\alpha)^r\big)}{r} - \sqrt{2l\alpha\, h(\alpha)} - \delta\sqrt{\frac{l\alpha}{2h(\alpha)}} \right]$$

so it tends to $-\infty$ (for every $\delta > 0$) in the case where $\alpha \to 1$, whereas (199) tends nearly to $\frac{nl}{r}$ (for small $\delta > 0$) as expected.

### *I. 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 steams 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. [53, 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 [43], 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 [43], [62, Section 4] and [79]).

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

$$s(t) = \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} X_i \exp\Bigl(\frac{j\,2\pi i t}{T}\Bigr), \quad 0 \leq t \leq T. \tag{200}$$

Lets assume that $X_0, \ldots, X_{n-1}$ 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.,

$$\frac{1}{T} \int_0^T |s(t)|^2 dt = 1. \tag{201}$$

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

$$\text{CF}_n(s) \triangleq \max_{0 \leq t \leq T} |s(t)|. \tag{202}$$

From [62, Section 4] and [79], it follows that the CF scales with high probability like $\sqrt{\ln n}$ for large $n$. In [43, Theorem 3 and Corollary 5], a concentration inequality was derived for the CF of OFDM signals. It states that for an arbitrary $c \geq 2.5$

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

*Remark 23:* The analysis used to derive this rather strong concentration inequality (see [43, Appendix C]) requires some assumptions on the distribution of the $X_i$'s (see the two conditions in [43, Theorem 3] followed by [43, 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 [43, 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. This material was presented in part in [66].

*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

$$Y_i = \mathbb{E}[\,\text{CF}_n(s)\,|\,X_0, \ldots, X_{i-1}], \quad i = 0, \ldots, n \tag{203}$$

Based on a standard construction of martingales, $\{Y_i, \mathcal{F}_i\}_{i=0}^n$ is a martingale where $\mathcal{F}_i$ is the $\sigma$-algebra that is generated by the first $i$ symbols $(X_0, \ldots, X_{i-1})$ in (200). Hence, $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}_n$ is a filtration. This martingale has also bounded jumps, and

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

for $i \in \{1, \ldots, n\}$ since revealing the additional $i$-th coordinate $X_i$ affects the CF, as is defined in (202), by at most $\frac{2}{\sqrt{n}}$ (see the first part of Appendix J). It therefore follows from Azuma's inequality that, for every $\alpha > 0$,

$$\mathbb{P}(|\text{CF}_n(s) - \mathbb{E}[\text{CF}_n(s)]| \geq \alpha) \leq 2\exp\left(-\frac{\alpha^2}{8}\right) \tag{204}$$

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 (203), Appendix J gives that a.s.

$$|Y_i - Y_{i-1}| \leq \frac{2}{\sqrt{n}}, \quad \mathbb{E}\big[(Y_i - Y_{i-1})^2 | \mathcal{F}_{i-1}\big] \leq \frac{2}{n} \tag{205}$$

for every $i \in \{1, \ldots, n\}$. Note that the conditioning on the $\sigma$-algebra $\mathcal{F}_{i-1}$ is equivalent to the conditioning on the symbols $X_0, \ldots, X_{i-2}$, 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 $\alpha > 0$

$$\begin{aligned}
&\mathbb{P}(|\text{CF}_n(s) - \mathbb{E}[\text{CF}_n(s)]| \geq \alpha) \\
&= \mathbb{P}(|Y_n - Y_0| \geq \alpha) \\
&= \mathbb{P}(|Z_n - Z_0| \geq \alpha\sqrt{n}) \\
&\leq 2\exp\left(-\frac{\alpha^2}{4}\left(1 + O\Big(\frac{1}{\sqrt{n}}\Big)\right)\right)
\end{aligned} \tag{206}$$

(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 (204) via Azuma's inequality, and the term which scales like $O\Big(\frac{1}{\sqrt{n}}\Big)$ on the right-hand side of (206) is expressed explicitly for finite $n$ (see Appendix F).

*3) A Concentration Inequality via Talagrand's Method:* In his seminal paper [74], 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., [21, Section 2.4.2], [50, Section 4] and [74]). 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.,

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

where $I$ stands for the indicator function.

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

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

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

Hence, $d_{\text{H}}(\mathbf{x}, \mathbf{y}) = \sqrt{n}\, d_a(\mathbf{x}, \mathbf{y})$ for $a = \big(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\big)$.

The following is a special form of Talagrand's inequalities ([50, Chapter 4], [74], [75]).

*Theorem 15 (Talagrand's inequality):* Let the random vector $\mathbf{X} = (X_1, \ldots, X_n)$ 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 $f : A \to \mathbb{R}$ satisfy the condition that, for every $\mathbf{x} \in A$, there exists a non-negative, normalized $n$-length vector $a = a(x)$ such that

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \sigma d_a(\mathbf{x}, \mathbf{y}), \quad \forall\, \mathbf{y} \in A \tag{207}$$

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

$$\mathbb{P}(|f(X) - m| \geq \alpha) \leq 4\exp\left(-\frac{\alpha^2}{4\sigma^2}\right) \tag{208}$$

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 (208) holds if the condition in (207) is replaced by

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \sigma d_a(\mathbf{x}, \mathbf{y}), \quad \forall\, \mathbf{y} \in A. \tag{209}$$

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 $X_0, Y_0, \ldots, X_{n-1}, Y_{n-1}$ 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

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

where

$$a \triangleq \Bigl(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\Bigr) \tag{210}$$

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 15 implies that, for every $\alpha \geq 0$,

$$\mathbb{P}(|\text{CF}_n(s) - m_n| \geq \alpha) \leq 4 \exp\Bigl(-\frac{\alpha^2}{16}\Bigr) \tag{211}$$

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 (211), 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 (211)

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

■

*Remark 24:* This result applies in general to an arbitrary function $f$ satisfying the condition in (207), where Talagrand's inequality in (208) implies that (see, e.g., [50, Lemma 4.6])

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

*4) Establishing Concentration via McDiardmid's Inequality:* McDiardmid's inequality (see Theorem 2) is applied in the following to prove a concentration inequality for the crest factor of OFDM signals. To this end, let us define

$$\begin{aligned} U &\triangleq \max_{0 \leq t \leq T} \bigl| s(t; X_0, \ldots, X_{i-1}, X_i, \ldots, X_{n-1}) \bigr| \\ V &\triangleq \max_{0 \leq t \leq T} \bigl| s(t; X_0, \ldots, X'_{i-1}, X_i, \ldots, X_{n-1}) \bigr| \end{aligned}$$

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

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

where the last inequality holds since $|X_{i-1}| = |X'_{i-1}| = 1$. Hence, McDiarmid's inequality in Theorem 2 implies that, for every $\alpha \geq 0$,

$$\mathbb{P}(|\mathrm{CF}_n(s) - \mathbb{E}[\mathrm{CF}_n(s)]| \geq \alpha) \leq 2\exp\Bigl(-\frac{\alpha^2}{2}\Bigr) \tag{212}$$

which demonstrates concentration of this measure around its expected value. By comparing (211) with (212), it follows that McDiarmid's inequality provides an improvement in the exponent. 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 [43] and [79]), 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.

### *J. Random Coding Theorems via Martingale Inequalities*

The following sub-section establishes new error exponents and achievable rates of random coding, for channels with and without memory, under maximum-likelihood (ML) decoding. The analysis relies on some exponential inequalities for martingales with bounded jumps. The characteristics of these coding theorems are exemplified in special cases of interest that include non-linear channels. The material in this sub-section is based on [80], [81] and [82] (and mainly on the latest improvements of these achievable rates in [82]).

Random coding theorems address the average error probability of an ensemble of codebooks as a function of the code rate $R$, the block length $N$, and the channel statistics. It is assumed that the codewords are chosen randomly, subject to some possible constraints, and the codebook is known to the encoder and decoder.

Nonlinear effects are typically encountered in wireless communication systems and optical fibers, which degrade the quality of the information transmission. In satellite communication systems, the amplifiers located on board satellites typically operate at or near the saturation region in order to conserve energy. Saturation nonlinearities of amplifiers introduce nonlinear distortion in the transmitted signals. Similarly, power amplifiers in mobile terminals are designed to operate in a nonlinear region in order to obtain high power efficiency in mobile cellular communications. Gigabit optical fiber communication channels typically exhibit linear and nonlinear distortion as a result of non-ideal transmitter, fiber, receiver and optical amplifier components. Nonlinear communication channels can be represented by Volterra models [9, Chapter 14].

Significant degradation in performance may result in the mismatched regime. However, in the following, it is assumed that both the transmitter and the receiver know the exact probability law of the channel.

We start the presentation by writing explicitly the martingale inequalities that we rely on, derived earlier along the derivation of the concentration inequalities in this chapter.

*1) Martingale inequalities:*

- The first martingale inequality is a known result (see [21, Corollary 2.4.7] and [49]) that will be useful later in this paper.

  *Theorem 16:* Let $\{X_k, \mathcal{F}_k\}_{k=0}^n$, for some $n \in \mathbb{N}$, be a discrete-parameter, real-valued martingale with bounded jumps. Let

  $$\xi_k \triangleq X_k - X_{k-1}, \quad \forall\, k \in \{1, \dots, n\}$$

  designate the jumps of the martingale. Assume that, for some constants $d, \sigma > 0$, the following two requirements

  $$\xi_k \leq d, \quad \text{Var}(\xi_k | \mathcal{F}_{k-1}) \leq \sigma^2$$

  hold almost surely (a.s.) for every $k \in \{1, \dots, n\}$. Let $\gamma \triangleq \frac{\sigma^2}{d^2}$. Then, for every $t \geq 0$,

  $$\mathbb{E}\biggl[\exp\biggl(t \sum_{k=1}^n \xi_k\biggr)\biggr] \leq \left(\frac{e^{-\gamma t d} + \gamma e^{td}}{1+\gamma}\right)^n. \tag{213}$$

  The proof of this theorem relies on Bennett's inequality (see [10] and [21, Lemma 2.4.1], and it was presented earlier in this chapter for the derivation of the first refinement of the Azuma-Hoeffding inequality.

- Second inequality: The following theorem presents a new martingale inequality that will be useful later in this sub-section.

  *Theorem 17:* Let $\{X_k, \mathcal{F}_k\}_{k=0}^n$, for some $n \in \mathbb{N}$, be a discrete-time, real-valued martingale with bounded jumps. Let

  $$\xi_k \triangleq X_k - X_{k-1}, \quad \forall\, k \in \{1, \dots, n\}$$

  and let $m \in \mathbb{N}$ be an even number, $d > 0$ be a positive number, and $\{\mu_l\}_{l=2}^m$ be a sequence of numbers such that

  $$\xi_k \leq d, \tag{214}$$

  $$\mathbb{E}\bigl[(\xi_k)^l \,|\, \mathcal{F}_{k-1}\bigr] \leq \mu_l, \quad \forall\, l \in \{2, \dots, m\} \tag{215}$$

  holds a.s. for every $k \in \{1, \dots, n\}$. Furthermore, let

  $$\gamma_l \triangleq \frac{\mu_l}{d^l}, \quad \forall\, l \in \{2, \dots, m\}. \tag{216}$$

  Then, for every $t \geq 0$,

  $$\mathbb{E}\biggl[\exp\biggl(t \sum_{k=1}^n \xi_k\biggr)\biggr] \leq \left(1 + \sum_{l=2}^{m-1} \frac{(\gamma_l - \gamma_m)\,(td)^l}{l!} + \gamma_m(e^{td} - 1 - td)\right)^n. \tag{217}$$

*2) Achievable Rates under ML Decoding:* The goal of this sub-section is to derive achievable rates in the random coding setting under ML decoding. We first review briefly the analysis in [81] for the derivation of the upper bound on the ML decoding error probability. This review is necessary in order to make the beginning of the derivation of this bound more accurate, and to correct along the way some inaccuracies that appear in [81, Section II]. After the first stage of this analysis, we proceed by improving the resulting error exponents and their corresponding achievable rates via the application of the martingale inequalities in the previous sub-section.

Consider an ensemble of block codes $\mathbf{C}$ of length $N$ and rate $R$. Let $\mathcal{C} \in \mathbf{C}$ be a codebook in the ensemble. The number of codewords in $\mathcal{C}$ is $M = \lceil \exp(NR) \rceil$. The codewords of a codebook $\mathcal{C}$ are assumed to be independent, and the symbols in each codeword are assumed to be i.i.d. with an arbitrary probability distribution $P$. An ML decoding error occurs if, given the transmitted message $m$ and the received vector $\mathbf{y}$, there exists another message $m' \neq m$ such that

$$||\mathbf{y} - D\mathbf{u}_{m'}||_2 \leq ||\mathbf{y} - D\mathbf{u}_m||_2.$$

The union bound for an AWGN channel implies that

$$P_{\text{e}|m}(\mathcal{C}) \leq \sum_{m' \neq m} Q\left(\frac{\|D\mathbf{u}_m - D\mathbf{u}_{m'}\|_2}{2\sigma_\nu}\right)$$

where

$$Q(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp\Bigl(-\frac{t^2}{2}\Bigr) \mathrm{d}t, \quad \forall\, x \in \mathbb{R} \tag{218}$$

is the complementary Gaussian cumulative distribution function. By using the inequality $Q(x) \leq \frac{1}{2} \exp\bigl(-\frac{x^2}{2}\bigr)$ for $x \geq 0$, it gives the loosened bound (by also ignoring the factor of one-half in the bound of $Q$)

$$P_{\text{e}|m}(\mathcal{C}) \leq \sum_{m' \neq m} \exp\left(-\frac{\|D\mathbf{u}_m - D\mathbf{u}_{m'}\|_2^2}{8\sigma_\nu^2}\right).$$

At this stage, let us introduce a new parameter $\rho \in [0, 1]$, and write

$$P_{\text{e}|m}(\mathcal{C}) \leq \sum_{m' \neq m} \exp\left(-\frac{\rho\, \|D\mathbf{u}_m - D\mathbf{u}_{m'}\|_2^2}{8\sigma_\nu^2}\right).$$

Note that at this stage, the introduction of the additional parameter $\rho$ is useless as its optimal value is $\rho_{\text{opt}} = 1$. The average ML decoding error probability over the code ensemble therefore satisfies

$$\overline{P}_{\text{e}|m} \leq \mathbb{E}\left[\sum_{m' \neq m} \exp\left(-\frac{\rho\, \|D\mathbf{u}_m - D\mathbf{u}_{m'}\|_2^2}{8\sigma_\nu^2}\right)\right]$$

and the average ML decoding error probability over the code ensemble and the transmitted message satisfies

$$\overline{P}_{\text{e}} \leq (M-1)\, \mathbb{E}\left[\exp\left(-\frac{\rho\, \|D\mathbf{u} - D\widetilde{\mathbf{u}}\|_2^2}{8\sigma_\nu^2}\right)\right] \tag{219}$$

where the expectation is taken over two randomly chosen codewords $\mathbf{u}$ and $\widetilde{\mathbf{u}}$ where these codewords are independent, and their symbols are i.i.d. with a probability distribution $P$.

Consider a filtration $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}_N$ where the sub $\sigma$-algebra $\mathcal{F}_i$ is given by

$$\mathcal{F}_i \triangleq \sigma(U_1, \widetilde{U}_1, \ldots, U_i, \widetilde{U}_i), \quad \forall\, i \in \{1, \ldots, N\} \tag{220}$$

for two randomly selected codewords $\mathbf{u} = (u_1, \ldots, u_N)$, and $\widetilde{\mathbf{u}} = (\tilde{u}_1, \ldots, \tilde{u}_N)$ from the codebook; $\mathcal{F}_i$ is the minimal $\sigma$-algebra that is generated by the first $i$ coordinates of these two codewords. In particular, let $\mathcal{F}_0 \triangleq \{\emptyset, \Omega\}$ be the trivial $\sigma$-algebra. Furthermore, define the discrete-time martingale $\{X_k, \mathcal{F}_k\}_{k=0}^N$ by

$$X_k = \mathbb{E}[||D\mathbf{u} - D\widetilde{\mathbf{u}}||_2^2 \,|\, \mathcal{F}_k] \tag{221}$$

designates the conditional expectation of the squared Euclidean distance between the distorted codewords $D\mathbf{u}$ and $D\widetilde{\mathbf{u}}$ given the first $i$ coordinates of the two codewords $\mathbf{u}$ and $\widetilde{\mathbf{u}}$. The first and last elements of this martingale sequence are, respectively, equal to

$$X_0 = \mathbb{E}\,[||D\mathbf{u} - D\widetilde{\mathbf{u}}||_2^2], \quad X_N = ||D\mathbf{u} - D\widetilde{\mathbf{u}}||_2^2. \tag{222}$$

Furthermore, following earlier notation, let $\xi_k = X_k - X_{k-1}$ be the jumps of the martingale, then

$$\sum_{k=1}^{N} \xi_k = X_N - X_0 = ||D\mathbf{u} - D\widetilde{\mathbf{u}}||_2^2 - \mathbb{E}\,[||D\mathbf{u} - D\widetilde{\mathbf{u}}||_2^2]$$

and the substitution of the last equality into (219) gives that

$$\overline{P}_{\mathrm{e}} \leq \exp(NR)\, \exp\left(-\frac{\rho\,\mathbb{E}\big[||D\mathbf{u} - D\widetilde{\mathbf{u}}|||_2^2\big]}{8\sigma_\nu^2}\right) \mathbb{E}\left[\exp\left(-\frac{\rho}{8\sigma^2} \cdot \sum_{k=1}^{N} \xi_k\right)\right]. \tag{223}$$

Since the codewords are independent and their symbols are i.i.d., then it follows that

$$\begin{aligned}
&\mathbb{E}||D\mathbf{u} - D\widetilde{\mathbf{u}}|||_2^2 \\
&= \sum_{k=1}^{N} \mathbb{E}\Big[\big([D\mathbf{u}]_k - [D\widetilde{\mathbf{u}}]_k\big)^2\Big] \\
&= \sum_{k=1}^{N} \mathrm{Var}\,\big([D\mathbf{u}]_k - [D\widetilde{\mathbf{u}}]_k\big) \\
&= 2\sum_{k=1}^{N} \mathrm{Var}\,\big([D\mathbf{u}]_k\big) \\
&= 2\left(\sum_{k=1}^{q-1} \mathrm{Var}\,\big([D\mathbf{u}]_k\big) + \sum_{k=q}^{N} \mathrm{Var}\,\big([D\mathbf{u}]_k\big)\right).
\end{aligned}$$

Due to the channel model (see Eq. (240)) and the assumption that the symbols $\{u_i\}$ are i.i.d., it follows that $\mathrm{Var}\,\big([D\mathbf{u}]_k\big)$ is fixed for $k = q, \ldots, N$. Let $D_{\mathrm{v}}(P)$ designate this common value of the variance (i.e., $D_{\mathrm{v}}(P) = \mathrm{Var}\,\big([D\mathbf{u}]_k\big)$ for $k \geq q$), then

$$\mathbb{E}||D\mathbf{u} - D\widetilde{\mathbf{u}}|||_2^2 = 2\left(\sum_{k=1}^{q-1} \mathrm{Var}\,\big([D\mathbf{u}]_k\big) + (N - q + 1)D_{\mathrm{v}}(P)\right).$$

Let

$$C_\rho(P) \triangleq \exp\left\{-\frac{\rho}{8\sigma_\nu^2}\,\left(\sum_{k=1}^{q-1} \mathrm{Var}\,\big([D\mathbf{u}]_k\big) - (q-1)D_{\mathrm{v}}(P)\right)\right\}$$

which is a bounded constant, under the assumption that $||\mathbf{u}||_\infty \leq K < +\infty$ holds a.s. for some $K > 0$, and it is independent of the block length $N$. This therefore implies that the ML decoding error probability satisfies

$$\overline{P}_{\mathrm{e}} \leq C_\rho(P)\, \exp\left\{-N\Big(\frac{\rho\, D_{\mathrm{v}}(P)}{4\sigma_\nu^2} - R\Big)\right\} \mathbb{E}\left[\exp\left(\frac{\rho}{8\sigma_\nu^2} \cdot \sum_{k=1}^{N} Z_k\right)\right], \quad \forall\, \rho \in [0,1]. \tag{224}$$

where $Z_k \triangleq -\xi_k$, so $\{Z_k, \mathcal{F}_k\}$ is a martingale-difference that corresponds to the jumps of the martingale $\{-X_k, \mathcal{F}_k\}$. From (221), it follows that the martingale-difference sequence $\{Z_k, \mathcal{F}_k\}$ is given by

$$\begin{aligned}
Z_k &= X_{k-1} - X_k \\
&= \mathbb{E}[||D\mathbf{u} - D\widetilde{\mathbf{u}}||_2^2 \,|\, \mathcal{F}_{k-1}] - \mathbb{E}[||D\mathbf{u} - D\widetilde{\mathbf{u}}||_2^2 \,|\, \mathcal{F}_k].
\end{aligned} \tag{225}$$

For the derivation of improved achievable rates and error exponents (as compared to [81]), the two martingale inequalities presented earlier in this sub-section are applied to the obtain two possible exponential upper bounds (in terms of $N$) on the last term on the right-hand side of (224).

Let us assume that the essential supremum of the channel input is finite a.s. (i.e., $||u||_\infty$ is bounded a.s.). Based on the upper bound on the ML decoding error probability in (224), combined with the exponential martingale inequalities that are introduced in Theorems 16 and 17, one obtains the following bounds:

1) *First Bounding Technique*: From Theorem 16, if

$$Z_k \leq d, \quad \text{Var}(Z_k \,|\, \mathcal{F}_{k-1}) \leq \sigma^2$$

holds a.s. for every $k \geq 1$, and $\gamma_2 \triangleq \frac{\sigma^2}{d^2}$, then it follows from (224) that for every $\rho \in [0,1]$

$$\overline{P}_{\text{e}} \leq C_\rho(P) \, \exp\left\{-N\left(\frac{\rho \, D_{\text{v}}(P)}{4\sigma_\nu^2} - R\right)\right\} \, \left(\frac{\exp\left(-\frac{\rho\,\gamma_2\,d}{8\sigma_\nu^2}\right) + \gamma_2 \exp\left(\frac{\rho\,d}{8\sigma_\nu^2}\right)}{1+\gamma_2}\right)^N.$$

Therefore, the maximal achievable rate that follows from this bound is given by

$$R_1(\sigma_\nu^2) \triangleq \max_P \max_{\rho\in[0,1]} \left\{\frac{\rho\, D_{\text{v}}(P)}{4\sigma_\nu^2} - \ln\left(\frac{\exp\left(-\frac{\rho\,\gamma_2\,d}{8\sigma_\nu^2}\right) + \gamma_2 \exp\left(\frac{\rho\,d}{8\sigma_\nu^2}\right)}{1+\gamma_2}\right)\right\} \tag{226}$$

where the double maximization is performed over the input distribution $P$ and the parameter $\rho \in [0,1]$. The inner maximization in (226) can be expressed in closed form, leading to the following simplified expression:

$$R_1(\sigma_\nu^2) = \max_P \begin{cases} D\left(\left(\frac{\gamma_2}{1+\gamma_2} + \frac{2\mathcal{D}_{\text{v}}(P)}{d(1+\gamma_2)}\right) \,\Big|\Big|\, \frac{\gamma_2}{1+\gamma_2}\right), & \text{if } \; D_{\text{v}}(P) < \frac{\gamma_2\, d\left(\exp\left(\frac{d(1+\gamma_2)}{8\sigma_\nu^2}\right)-1\right)}{2\left(1+\gamma_2 \exp\left(\frac{d(1+\gamma_2)}{8\sigma_\nu^2}\right)\right)} \\ \frac{D_{\text{v}}(P)}{4\sigma_\nu^2} - \ln\left(\frac{\exp\left(-\frac{\gamma_2\,d}{8\sigma_\nu^2}\right)+\gamma_2\exp\left(\frac{d}{8\sigma_\nu^2}\right)}{1+\gamma_2}\right), & \text{otherwise} \end{cases} \tag{227}$$

where

$$D(p||q) \triangleq p \ln\left(\frac{p}{q}\right) + (1-p)\ln\left(\frac{1-p}{1-q}\right), \quad \forall\, p, q \in (0,1) \tag{228}$$

denotes the Kullback-Leibler distance (a.k.a. divergence or relative entropy) between the two probability distributions $(p, 1-p)$ and $(q, 1-q)$.

2) *Second Bounding Technique* Based on the combination of Theorem 17 and Eq. (224), we derive in the following a second achievable rate for random coding under ML decoding. Referring to the martingale-difference sequence $\{Z_k, \mathcal{F}_k\}_{k=1}^N$ in Eqs. (220) and (225), one obtains from Eq. (224) that if for some even number $m \in \mathbb{N}$

$$Z_k \leq d, \qquad \mathbb{E}\big[(Z_k)^l \,|\, \mathcal{F}_{k-1}\big] \leq \mu_l, \;\; \forall\, l \in \{2, \ldots, m\}$$

hold a.s. for some positive constant $d > 0$ and a sequence $\{\mu_l\}_{l=2}^m$, and

$$\gamma_l \triangleq \frac{\mu_l}{d^l} \quad \forall\, l \in \{2, \ldots, m\},$$

then the average error probability satisfies, for every $\rho \in [0,1]$,

$$\overline{P}_{\text{e}} \leq C_\rho(P) \, \exp\left\{-N\left(\frac{\rho\, D_{\text{v}}(P)}{4\sigma_\nu^2} - R\right)\right\} \, \left[1 + \sum_{l=2}^{m-1} \frac{\gamma_l - \gamma_m}{l!} \left(\frac{\rho d}{8\sigma_\nu^2}\right)^l + \gamma_m \left(\exp\Big(\frac{\rho\, d}{8\,\sigma_\nu^2}\Big) - 1 - \frac{\rho\, d}{8\,\sigma_\nu^2}\right)\right]^N.$$

This gives the following achievable rate, for an arbitrary even number $m \in \mathbb{N}$,

$$R_2(\sigma_\nu^2) \triangleq \max_P \max_{\rho\in[0,1]} \left\{\frac{\rho\, D_{\text{v}}(P)}{4\sigma_\nu^2} - \ln\left(1 + \sum_{l=2}^{m-1} \frac{\gamma_l - \gamma_m}{l!}\left(\frac{\rho d}{8\sigma_\nu^2}\right)^l + \gamma_m\left(\exp\Big(\frac{\rho\, d}{8\,\sigma_\nu^2}\Big) - 1 - \frac{\rho\, d}{8\,\sigma_\nu^2}\right)\right)\right\} \tag{229}$$

where, similarly to (226), the double maximization in (229) is performed over the input distribution $P$ and the parameter $\rho \in [0,1]$.

*3) Achievable Rates for Random Coding:* In the following, the achievable rates for random coding over various linear and non-linear channels (with and without memory) are exemplified. In order to assess the tightness of the bounds, we start with a simple example where the mutual information for the given input distribution is known, so that its gap can be estimated (since we use here the union bound, it would have been in place also to compare the achievable rate with the cutoff rate).

1) *Binary-Input AWGN Channel*: Consider the case of a binary-input AWGN channel where

$$Y_k = U_k + \nu_k$$

where $U_i = \pm A$ for some constant $A > 0$ is a binary input, and $\nu_i \sim \mathcal{N}(0, \sigma_\nu^2)$ is an additive Gaussian noise with zero mean and variance $\sigma_\nu^2$. Since the codewords $\mathbf{U} = (U_1, \ldots, U_N)$ and $\widetilde{\mathbf{U}} = (\widetilde{U}_1, \ldots, \widetilde{U}_N)$ are independent and their symbols are i.i.d., let

$$P(U_k = A) = P(\widetilde{U}_k = A) = \alpha, \quad P(U_k = -A) = P(\widetilde{U}_k = -A) = 1 - \alpha$$

for some $\alpha \in [0,1]$. Since the channel is memoryless and the all the symbols are i.i.d. then one gets from (220) and (225) that

$$\begin{aligned}
Z_k &= \mathbb{E}[||\mathbf{U} - \widetilde{\mathbf{U}}||_2^2 \,|\, \mathcal{F}_{k-1}] - \mathbb{E}[||\mathbf{U} - \widetilde{\mathbf{U}}||_2^2 \,|\, \mathcal{F}_k] \\
&= \left[\sum_{j=1}^{k-1} (U_j - \widetilde{U}_j)^2 + \sum_{j=k}^{N} \mathbb{E}\big[(U_j - \widetilde{U}_j)^2\big]\right] - \left[\sum_{j=1}^{k} (U_j - \widetilde{U}_j)^2 + \sum_{j=k+1}^{N} \mathbb{E}\big[(U_j - \widetilde{U}_j)^2\big]\right] \\
&= \mathbb{E}[(U_k - \widetilde{U}_k)^2] - (U_k - \widetilde{U}_k)^2 \\
&= \alpha(1-\alpha)(-2A)^2 + \alpha(1-\alpha)(2A)^2 - (U_k - \widetilde{U}_k)^2 \\
&= 8\alpha(1-\alpha)A^2 - (U_k - \widetilde{U}_k)^2.
\end{aligned}$$

Hence, for every $k$,

$$Z_k \leq 8\alpha(1-\alpha)A^2 \triangleq d. \tag{230}$$

Furthermore, for every $k, l \in \mathbb{N}$, due to the above properties

$$\begin{aligned}
&\mathbb{E}\big[(Z_k)^l \,|\, \mathcal{F}_{k-1}\big] \\
&= \mathbb{E}\big[(Z_k)^l\big] \\
&= \mathbb{E}\Big[\big(8\alpha(1-\alpha)A^2 - (U_k - \widetilde{U}_k)^2\big)^l\Big] \\
&= \big[1 - 2\alpha(1-\alpha)\big]\big(8\alpha(1-\alpha)A^2\big)^l + 2\alpha(1-\alpha)\,\big(8\alpha(1-\alpha)A^2 - 4A^2\big)^l \triangleq \mu_l
\end{aligned} \tag{231}$$

and therefore, from (230) and (231), for every $l \in \mathbb{N}$

$$\gamma_l \triangleq \frac{\mu_l}{d^l} = \big[1 - 2\alpha(1-\alpha)\big]\left[1 + (-1)^l \left(\frac{1 - 2\alpha(1-\alpha)}{2\alpha(1-\alpha)}\right)^{l-1}\right]. \tag{232}$$

Let us now rely on the two achievable rates for random coding in Eqs. (227) and (229), and apply them to the binary-input AWGN channel. Due to the channel symmetry, the considered input distribution is symmetric (i.e., $\alpha = \frac{1}{2}$ and $P = (\frac{1}{2}, \frac{1}{2})$). In this case, we obtain from (230) and (232) that

$$D_{\text{v}}(P) = \text{Var}(U_k) = A^2, \quad d = 2A^2, \qquad \gamma_l = \frac{1 + (-1)^l}{2}, \ \ \forall\, l \in \mathbb{N}. \tag{233}$$

Based on the first bounding technique that leads to the achievable rate in Eq. (227), since the first condition in this equation cannot hold for the set of parameters in (233) then the achievable rate in this equation is equal to

$$R_1(\sigma_\nu^2) = \frac{A^2}{4\sigma_\nu^2} - \ln\cosh\Big(\frac{A^2}{4\sigma_\nu^2}\Big)$$

in units of nats per channel use. Let $\text{SNR} \triangleq \frac{A^2}{\sigma_\nu^2}$ designate the signal to noise ratio, then the first achievable rate gets the form

$$R'_1(\text{SNR}) = \frac{\text{SNR}}{4} - \ln\cosh\left(\frac{\text{SNR}}{4}\right). \tag{234}$$

It is observed here that the optimal value of $\rho$ in (227) is equal to 1 (i.e., $\rho^\star = 1$).

Let us compare it in the following with the achievable rate that follows from (229). Let $m \in \mathbb{N}$ be an even number. Since, from (233), $\gamma_l = 1$ for all even values of $l \in \mathbb{N}$ and $\gamma_l = 0$ for all odd values of $l \in \mathbb{N}$, then

$$\begin{aligned}
&1 + \sum_{l=2}^{m-1} \frac{\gamma_l - \gamma_m}{l!} \left(\frac{\rho d}{8\sigma_\nu^2}\right)^l + \gamma_m \left(\exp\Bigl(\frac{\rho\, d}{8\,\sigma_\nu^2}\Bigr) - 1 - \frac{\rho\, d}{8\,\sigma_\nu^2}\right) \\
&= 1 - \sum_{l=1}^{\frac{m}{2}-1} \frac{1}{(2l+1)!} \left(\frac{\rho d}{8\sigma_\nu^2}\right)^{2l+1} + \left(\exp\Bigl(\frac{\rho\, d}{8\,\sigma_\nu^2}\Bigr) - 1 - \frac{\rho\, d}{8\,\sigma_\nu^2}\right)
\end{aligned} \tag{235}$$

Since the infinite sum $\sum_{l=1}^{\frac{m}{2}-1} \frac{1}{(2l+1)!} \left(\frac{\rho d}{8\sigma_\nu^2}\right)^{2l+1}$ is monotonically increasing with $m$ (where $m$ is even and $\rho \in [0,1]$), then from (229), the best achievable rate within this form is obtained in the limit where $m$ is even and $m \to \infty$. In this asymptotic case one gets

$$\begin{aligned}
&\lim_{m\to\infty} \left(1 + \sum_{l=2}^{m-1} \frac{\gamma_l - \gamma_m}{l!} \left(\frac{\rho d}{8\sigma_\nu^2}\right)^l + \gamma_m \left(\exp\Bigl(\frac{\rho\, d}{8\,\sigma_\nu^2}\Bigr) - 1 - \frac{\rho\, d}{8\,\sigma_\nu^2}\right)\right) \\
&\overset{\text{(a)}}{=} 1 - \sum_{l=1}^{\infty} \frac{1}{(2l+1)!} \left(\frac{\rho d}{8\sigma_\nu^2}\right)^{2l+1} + \left(\exp\Bigl(\frac{\rho\, d}{8\,\sigma_\nu^2}\Bigr) - 1 - \frac{\rho\, d}{8\,\sigma_\nu^2}\right) \\
&\overset{\text{(b)}}{=} 1 - \left(\sinh\Bigl(\frac{\rho\, d}{8\,\sigma_\nu^2}\Bigr) - \frac{\rho\, d}{8\,\sigma_\nu^2}\right) + \left(\exp\Bigl(\frac{\rho\, d}{8\,\sigma_\nu^2}\Bigr) - 1 - \frac{\rho\, d}{8\,\sigma_\nu^2}\right) \\
&\overset{\text{(c)}}{=} \cosh\Bigl(\frac{\rho\, d}{8\,\sigma_\nu^2}\Bigr)
\end{aligned} \tag{236}$$

where equality (a) follows from (235), equality (b) holds since $\sinh(x) = \sum_{l=0}^{\infty} \frac{x^{2l+1}}{(2l+1)!}$ for $x \in \mathbb{R}$, and equality (c) holds since $\sinh(x) + \cosh(x) = \exp(x)$. Therefore, the achievable rate in (229) gives (from (233), $\frac{d}{8\sigma_\nu^2} = \frac{A^2}{4\sigma_\nu^2}$)

$$R_2(\sigma_\nu^2) = \max_{\rho\in[0,1]} \left(\frac{\rho A^2}{4\sigma_\nu^2} - \ln\cosh\Bigl(\frac{\rho A^2}{4\sigma_\nu^2}\Bigr)\right).$$

Since the function $f(x) \triangleq x - \ln\cosh(x)$ for $x \in \mathbb{R}$ is monotonic increasing (note that $f'(x) = 1 - \tanh(x) \geq 0$), then the optimal value of $\rho \in [0,1]$ is equal to 1, and therefore the best achievable rate that follows from the second bounding technique in Eq. (229) is equal to

$$R_2(\sigma_\nu^2) = \frac{A^2}{4\sigma_\nu^2} - \ln\cosh\Bigl(\frac{A^2}{4\sigma_\nu^2}\Bigr)$$

in units of nats per channel use, and it is obtained in the asymptotic case where we let the even number $m$ tend to infinity. Finally, setting $\text{SNR} = \frac{A^2}{\sigma_\nu^2}$, gives the achievable rate in (234), so the first and second achievable rates for the binary-input AWGN channel coincide, i.e.,

$$R'_1(\text{SNR}) = R'_2(\text{SNR}) = \frac{\text{SNR}}{4} - \ln\cosh\left(\frac{\text{SNR}}{4}\right). \tag{237}$$

Note that this common rate tends to zero as we let the signal to noise ratio tend to zero, and it tends to $\ln 2$ nats per channel use (i.e., 1 bit per channel use) as we let the signal to noise ratio tend to infinity.

In the considered setting of random coding, in order to exemplify the tightness of the achievable rate in (237), it is compared in the following with the symmetric i.i.d. mutual information of the binary-input AWGN channel. The mutual information for this channel (in units of nats per channel use) is given by (see, e.g., [61, Example 4.38 on p. 194])

$$\begin{aligned} C(\text{SNR}) &= \ln 2 + (2\ \text{SNR} - 1)\, Q(\sqrt{\text{SNR}}) - \sqrt{\frac{2\,\text{SNR}}{\pi}}\ \exp\Big(-\frac{\text{SNR}}{2}\Big) \\ &\quad + \sum_{i=1}^{\infty} \left\{ \frac{(-1)^i}{i(i+1)} \cdot \exp(2i(i+1)\ \text{SNR})\ Q\big((1+2i)\sqrt{\text{SNR}}\big) \right\} \end{aligned} \tag{238}$$

where the $Q$-function that appears in the infinite series on the right-hand side of (238) is the complementary Gaussian cumulative distribution function in (218). Furthermore, this infinite series has a fast convergence where the absolute value of its $n$-th remainder is bounded by the $(n+1)$-th term of the series, which scales like $\frac{1}{n^3}$ (due to a basic theorem on infinite series of the form $\sum_{n\in\mathbb{N}}(-1)^n\, a_n$ where $\{a_n\}$ is a positive and monotonically decreasing sequence; the theorem states that the $n$-th remainder of the series is upper bounded in absolute value by $a_{n+1}$).

The comparison between the mutual information of the binary-input AWGN channel with a symmetric i.i.d. input distribution and the common achievable rate in (237) that follows from the martingale approach is shown in Figure 4.

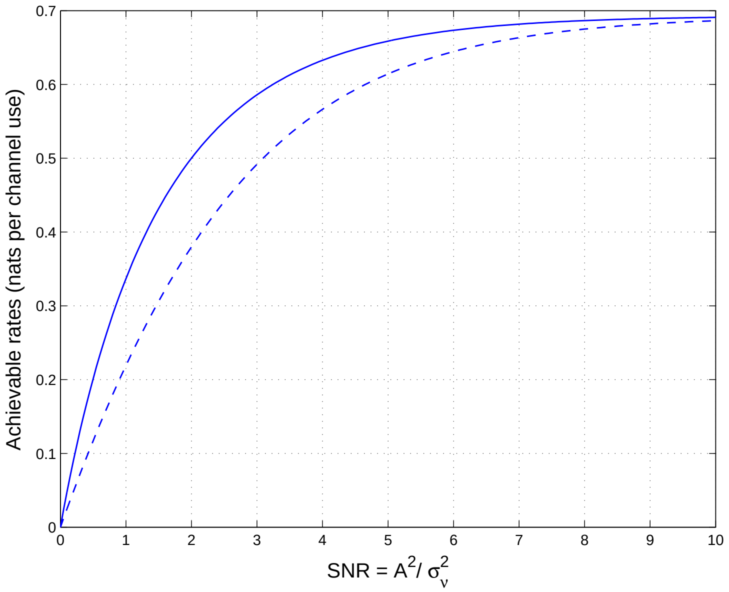

Fig. 4. A comparison between the symmetric i.i.d. mutual information of the binary-input AWGN channel (solid line) and the common achievable rate in (237) (dashed line) that follows from the martingale approach in this sub-section.

From the discussion in this sub-section, the first and second bounding techniques in Section VI-J2 lead to the same achievable rate (see (237)) in the setup of random coding and ML decoding where we assume a symmetric input distribution (i.e., $P(\pm A) = \frac{1}{2}$). But this is due to the fact that, from (233), the sequence $\{\gamma_l\}_{l\geq 2}$ is equal to zero for odd indices of $l$ and it is equal to 1 for even values of $l$ (see the derivation of (235) and (236)). Note, however, that the second bounding technique may provide tighter bounds than the first one (which follows from Bennett's inequality) due to the knowledge of $\{\gamma_l\}$ for $l > 2$. This approach was exemplified in Table I in the context of the pairwise error probability (under ML decoding) for some binary-input discrete memoryless channels.

2) *Nonlinear Channels with Memory - Third-Order Volterra Channels*: The channel model is first presented in the following (see Figure 5). We refer in the following to a discrete-time channel model of nonlinear Volterra channels where the input-output channel model is given by

$$y_i = [D\mathbf{u}]_i + \nu_i \tag{239}$$

TABLE III
KERNELS OF THE 3RD ORDER VOLTERRA SYSTEM $D_1$ WITH MEMORY 2

| kernel | $h_1(0)$ | $h_1(1)$ | $h_1(2)$ | $h_2(0,0)$ | $h_2(1,1)$ | $h_2(0,1)$ |
|---|---|---|---|---|---|---|
| value | 1.0 | 0.5 | $-0.8$ | 1.0 | $-0.3$ | 0.6 |

| kernel | $h_3(0,0,0)$ | $h_3(1,1,1)$ | $h_3(0,0,1)$ | $h_3(0,1,1)$ |
|---|---|---|---|---|
| value | 1.0 | $-0.5$ | 1.2 | 0.8 |

| kernel | $h_3(0,1,2)$ |
|---|---|
| value | 0.6 |

where $i$ is the time index. Volterra's operator $D$ of order $L$ and memory $q$ is given by

$$[D\mathbf{u}]_i = h_0 + \sum_{j=1}^{L} \sum_{i_1=0}^{q} \ldots \sum_{i_j=0}^{q} h_j(i_1, \ldots, i_j) u_{i-i_1} \ldots u_{i-i_j}. \tag{240}$$

and $\boldsymbol{\nu}$ is an additive Gaussian noise vector with i.i.d. components $\nu_i \sim \mathcal{N}(0, \sigma_\nu^2)$.

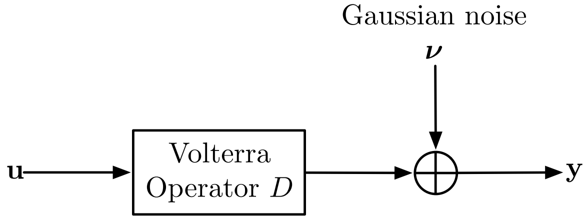

Fig. 5. The discrete-time Volterra non-linear channel model in Eqs. (239) and (240) where the channel input and output are $\{U_i\}$ and $\{Y_i\}$, respectively, and the additive noise samples $\{\nu_i\}$, which are added to the distorted input, are i.i.d. with zero mean and variance $\sigma_\nu^2$.

Under the same setup of the previous subsection regarding the channel input characteristics, we consider next the transmission of information over the Volterra system $D_1$ of order $L = 3$ and memory $q = 2$, whose kernels are depicted in Table III. Such system models are used in the base-band representation of nonlinear narrow-band communication channels. Due to complexity of the channel model, the calculation of the achievable rates provided earlier in this sub-section requires the numerical calculation of the parameters $d$ and $\sigma^2$ and thus of $\gamma_2$ for the martingale $\{Z_i, \mathcal{F}_i\}_{i=0}^N$. In order to achieve this goal, we have to calculate $|Z_i - Z_{i-1}|$ and $\text{Var}(Z_i|\mathcal{F}_{i-1})$ for all possible combinations of the input samples which contribute to the aforementioned expressions. Thus, the analytic calculation of $d$ and $\gamma_l$ increases as the system's memory $q$ increases. Numerical results are provided in Figure 6 for the case where $\sigma_\nu^2 = 1$. The new achievable rates $R_1^{(2)}(D_1, A, \sigma_\nu^2)$ and $R_2(D_1, A, \sigma_\nu^2)$, which depend on the channel input parameter $A$, are compared to the achievable rate provided in [81, Fig.2] and are shown to be larger than the latter.

To conclude, improvements of the achievable rates in the low SNR regime are expected to be obtained via existing improvements to Bennett's inequality (see [26] and [27]), combined with a possible tightening of the union bound under ML decoding (see, e.g., [63]). This direction of research is studied in [69].

## VII. 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 chapter derives some classical concentration inequalities 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

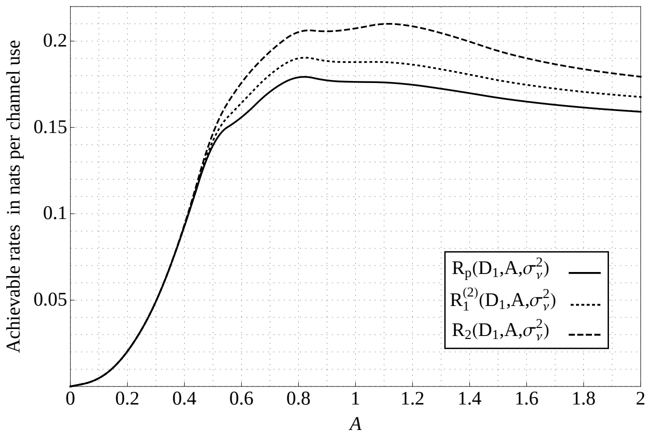

Fig. 6. Comparison of the achievable rates in this sub-section $R_1(D_1, A, \sigma_\nu^2)$ and $R_2^{(2)}(D_1, A, \sigma_\nu^2)$ (where $m = 2$) with the bound $R_p(D_1, A, \sigma_\nu^2)$ of [81, Fig.2] for the nonlinear channel with kernels depicted in Table III and noise variance $\sigma_\nu^2 = 1$. Rates are expressed in nats per channel use.

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 martingale approach for establishing concentration in information and communication-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 the discussion in this chapter.

- *Possible refinements of Theorem 3*: The proof of the concentration inequality in Theorem 3 relies on Bennett's inequality (31). 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 (32)). As was noted in [10, 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 3, consider a martingale-difference sequence $\{\xi_k, \mathcal{F}_k\}_{k=0}^n$ where, e.g., $\xi_k$ 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 3 under such an assumption. This direction has been recently studied in [68], and it calls for suitable applications.
- *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 $\mathrm{GF}(q)$, 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*: 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. Some results on this direction are presented in [24] by refining the proper constants that follow from the Azuma-Hoeffding inequality for the studied applications.
- *Martingale-based Inequalities Related to Exponential Bounds on Error Probability with Feedback*: As a follow-up to [55, Section 3.3] and [58, Theorem 11], an analysis that relies on the refined versions of Azuma's inequality in Section IV (with the standard adaptation of these inequalities to sub-martingales) has the potential to provide further results in this direction.

## Appendix A
## Proof of Lemma 3

The first and third properties of $\varphi_m$ 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 $\lim_{y \to 0} \varphi_m(y) = 1$, and it follows from the above power series expansion that $\varphi_m$ is strictly monotonic increasing over the interval $[0, \infty)$. The fourth property of $\varphi_m$ 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 $\varphi_m$ follows by combining the third and fourth properties.

## Appendix B
## 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:

$$\mathbb{P}(|X_n - X_0| \geq n\alpha) \leq 2 \left\{ \inf_{x \geq 0} e^{-\delta x} \Big[ 1 + \gamma(e^x - 1 - x) \Big] \right\}^n \tag{241}$$

where $\gamma = \gamma_2$ according to the notation in (29).

By differentiating the logarithm of the right-hand side of (241) w.r.t. $x$ (where $x \geq 0$) and setting this derivative to zero, it follows that

$$\frac{1 - \gamma x}{\gamma(e^x - 1)} = \frac{1 - \delta}{\delta}. \tag{242}$$

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 $x \to \infty$. In the former case where $x = \frac{1}{\gamma}$, the resulting bound in (241) is equal to

$$\exp\left[-n\left(\frac{1}{\gamma} - \ln\Bigl(\gamma\bigl(e^{\frac{1}{\gamma}} - 1\bigr)\Bigr)\right)\right]. \tag{243}$$

In the latter case where $x \to \infty$, the resulting bound in (241) when $\delta = 1$ is equal to

$$\begin{aligned}&\lim_{x\to\infty} e^{-nx}\bigl(1 + \gamma(e^x - 1 - x)\bigr)^n \\ &= \lim_{x\to\infty}\Bigl(e^{-x} + \gamma\bigl(1 - (1+x)e^{-x}\bigr)\Bigr)^n \\ &= \gamma^n.\end{aligned}$$

Hence, since for $\gamma \in (0, 1)$

$$\begin{aligned}\ln\Bigl(\frac{1}{\gamma}\Bigr) &= \frac{1}{\gamma} - \ln\Bigl(\gamma e^{\frac{1}{\gamma}}\Bigr) \\ &< \frac{1}{\gamma} - \ln\Bigl(\gamma\bigl(e^{\frac{1}{\gamma}} - 1\bigr)\Bigr)\end{aligned}$$

then the optimized value is $x = \frac{1}{\gamma}$, and the resulting bound in the case where $\delta = 1$ is equal to (243).

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 (242) has a unique solution, and it lies in $(0, \frac{1}{\gamma})$. This solution is given in (61).

*Proof:* Consider equation (242), 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 (242) is negative for $x < 0$ and $x > \frac{1}{\gamma}$. Since the function $t$ is continuous on the interval $(0, \frac{1}{\gamma}]$ and

$$t\left(\frac{1}{\gamma}\right) = 0, \qquad \lim_{x\to 0^+} t(x) = +\infty$$

then there is a solution $x \in \bigl(0, \frac{1}{\gamma}\bigr)$. Moreover, the function $t$ is monotonic decreasing in the interval $\bigl(0, \frac{1}{\gamma}\bigr]$ (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. (242) can be expressed in the form

$$\frac{a - x}{e^x - 1} = b \tag{244}$$

where

$$a \triangleq \frac{1}{\gamma}, \; b \triangleq \frac{1 - \delta}{\delta} \tag{245}$$

are both positive. The substitution $u = a + b - x$ in (244) gives

$$ue^u = be^{a+b}$$

whose solution is, by definition, given by $u = W_0\left(be^{a+b}\right)$ where $W_0$ denotes the principal branch of the multi-valued Lambert W function [17]. 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

$$\begin{aligned}x &= a + b - u \\ &= a + b - W_0\left(be^{a+b}\right)\end{aligned} \tag{246}$$

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 $W_0\left(be^{a+b}\right) < a+b$ and it was earlier proved in this appendix that the positive solution $x$ of (242) is unique). Finally, the substitution of (245) into (246) gives (61). This completes the proof of Lemma 6. ■

The bound in (241) is given by

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) \leq 2\exp\biggl(-n\Bigl[\delta x - \ln\bigl(1+\gamma(e^x - 1 - x)\bigr)\Bigr]\biggr) \tag{247}$$

with the value of $x$ in (61).

## Appendix C
## Proof of Proposition 1

Lets compare the right-hand sides of (49) and (241) that refer to Corollaries 2 and 4, respectively. Proposition 1 follows by showing that if $\gamma \leq \frac{1}{2}$

$$1 + \gamma(\exp(x) - 1 - x) < \cosh(x), \quad \forall\, x > 0. \tag{248}$$

To this end, define

$$f(x) \triangleq \cosh(x) - \gamma\bigl(\exp(x) - 1 - x\bigr), \quad \forall\, x \geq 0.$$

If $\gamma \leq \frac{1}{2}$, then for every $x > 0$

$$\begin{aligned} f'(x) &= \sinh(x) - \gamma\bigl(\exp(x) - 1\bigr) \\ &= \left(\frac{1}{2} - \gamma\right)\exp(x) + \gamma - \frac{\exp(-x)}{2} \\ &> \left(\frac{1}{2} - \gamma\right) + \gamma - \frac{1}{2} = 0 \end{aligned}$$

so, since $f$ is monotonic increasing on $[0, \infty)$ and $f(0) = 0$, then $f(x) > 0$ for every $x > 0$. This validates (248), and it therefore completes the proof of Proposition 1.

## Appendix D
## Proof of Proposition 2

*Lemma 7:* For every $\gamma, x > 0$

$$\frac{\gamma e^x + e^{-\gamma x}}{1+\gamma} < 1 + \gamma(e^x - 1 - x). \tag{249}$$

*Proof:* Let $\gamma$ be an arbitrary positive number, and define the function

$$f_\gamma(x) \triangleq \frac{\gamma e^x + e^{-\gamma x}}{1+\gamma} - \left[1 + \gamma(e^x - 1 - x)\right], \quad x \geq 0.$$

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 (249). ■

This claim in Proposition 2 follows directly from Lemma 7, and the two inequalities in (33) and (59) with $m = 2$. In the case where $m = 2$, the right-hand side of (59) is equal to

$$\big(1 + \gamma(e^{td} - 1 - td)\big)^n.$$

Note that (33) and (59) with $m = 2$ were used to derive, respectively, Theorem 3 and Corollary 4 (based on Chernoff's bound). The conclusion follows by substituting $x \triangleq td$ on the right-hand sides of (33) and (59) with $m = 2$ (so that $x \geq 0$ since $t \geq 0$ and $d > 0$, and (249) turns from an inequality if $x > 0$ into an equality if $x = 0$).

## Appendix E
## Proof of Corollary 6

A minimization of the logarithm of the exponential bound on the right-hand side of (63) gives the equation

$$\frac{\displaystyle\sum_{l=2}^{m-1} \frac{(\gamma_l - \gamma_m)x^{l-1}}{(l-1)!} + \gamma_m(e^x - 1)}{\displaystyle 1 + \sum_{l=2}^{m-1} \frac{(\gamma_l - \gamma_m)x^l}{l!} + \gamma_m(e^x - 1 - x)} = \delta$$

and after standard algebraic operations, it gives the equation

$$\begin{aligned} &\gamma_m \left(\frac{1}{\delta} - 1\right)(e^x - 1 - x) + \frac{\gamma_2 x}{\delta} \\ &+ \sum_{l=2}^{m-1} \left\{ \left[\frac{\gamma_{l+1}}{\delta} - \gamma_l - \gamma_m\Big(\frac{1}{\delta} - 1\Big)\right] \frac{x^l}{l!} \right\} - 1 = 0. \end{aligned} \tag{250}$$

As we have seen in the proof of Corollary 4 (see Appendix B), 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 (53) 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 $\gamma_l$ 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 $\gamma_2$, then it gives the equation $a + b - cx = be^x$ with $a, b$ and $c$ from (65). 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 [17]. 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 (64), 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 (53) with the value of $x$ that is given in (64) and (65). For $m = 2$ where the sum on the left-hand side of (250) that was later neglected is anyway zero, this forms indeed the exact optimal value of $x$ (so that it coincides with Eq. (61) in Corollary 4).

## APPENDIX F
## 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 3 and 4, and also Corollaries 3 and 4 imply (69). For the sake of brevity, we introduce in the following the analysis that is related to Theorem 3. The others are technical as well.

Let $\{X_k, \mathcal{F}_k\}_{k=0}^{\infty}$ be a discrete-parameter martingale that satisfies the conditions in Theorem 3. From (28)

$$\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) \leq 2\exp\left(-n\, D\left(\frac{\delta' + \gamma}{1+\gamma} \Big|\Big| \frac{\gamma}{1+\gamma}\right)\right) \tag{251}$$

where from (29)

$$\delta' \triangleq \frac{\frac{\alpha}{\sqrt{n}}}{d} = \frac{\delta}{\sqrt{n}}\,. \tag{252}$$

From the right-hand side of (251)

$$\begin{aligned} & D\left(\frac{\delta' + \gamma}{1+\gamma} \Big|\Big| \frac{\gamma}{1+\gamma}\right) \\ & = \frac{\gamma}{1+\gamma}\left[\left(1 + \frac{\delta}{\gamma\sqrt{n}}\right)\ln\left(1 + \frac{\delta}{\gamma\sqrt{n}}\right) + \frac{1}{\gamma}\left(1 - \frac{\delta}{\sqrt{n}}\right)\ln\left(1 - \frac{\delta}{\sqrt{n}}\right)\right]. \end{aligned} \tag{253}$$

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 (253) that for every $n > \frac{\delta^2}{\gamma^2}$

$$\begin{aligned} nD\left(\frac{\delta' + \gamma}{1+\gamma} \Big|\Big| \frac{\gamma}{1+\gamma}\right) &= \frac{\delta^2}{2\gamma} - \frac{\delta^3(1-\gamma)}{6\gamma^2}\,\frac{1}{\sqrt{n}} + \ldots \\ &= \frac{\delta^2}{2\gamma} + O\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

Substituting this into the exponent on the right-hand side of (251) gives (69).

## APPENDIX G
## ANALYSIS RELATED TO THE MODERATE DEVIATIONS PRINCIPLE IN SECTION V-C

It is demonstrated in the following that, in contrast to Azuma's inequality, both Theorems 3 and 4 provide upper bounds on

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

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

$$S_k \triangleq \sum_{i=1}^{k} X_i, \quad \mathcal{F}_k \triangleq \sigma(X_1, \ldots, X_k)$$

for every $k \in \{1, \ldots, n\}$ 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 $k \in \{1, \ldots, n\}$. Hence it follows from Azuma's inequality that, for every $\alpha \geq 0$,

$$\mathbb{P}\left(|S_n| \geq \alpha n^{\eta}\right) \leq 2\exp\left(-\frac{\alpha^2 n^{2\eta-1}}{2d^2}\right)$$

and therefore

$$\lim_{n\to\infty} n^{1-2\eta}\, \ln \mathbb{P}\big(|S_n| \geq \alpha n^{\eta}\big) \leq -\frac{\alpha^2}{2d^2}. \tag{254}$$

This differs from the limit in (82) where $\sigma^2$ 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 3:* The analysis here is a slight modification of the analysis in Appendix F with the required adaptation of the calculations for $\eta \in (\frac{1}{2}, 1)$. It follows from Theorem 3 that, for every $\alpha \geq 0$,

$$\mathbb{P}(|S_n| \geq \alpha n^{\eta}) \leq 2 \exp\left(-n\, D\left(\frac{\delta' + \gamma}{1+\gamma} \Big|\Big| \frac{\gamma}{1+\gamma}\right)\right)$$

where $\gamma$ is introduced in (29), and $\delta'$ in (252) is replaced with

$$\delta' \triangleq \frac{\frac{\alpha}{n^{1-\eta}}}{d} = \delta n^{-(1-\eta)} \tag{255}$$

due to the definition of $\delta$ in (29). Following the same analysis as in Appendix F, it follows that for every $n \in \mathbb{N}$

$$\mathbb{P}(|S_n| \geq \alpha n^{\eta}) \leq 2 \exp\left(-\frac{\delta^2 n^{2\eta-1}}{2\gamma}\left[1 + \frac{\alpha(1-\gamma)}{3\gamma d} \cdot n^{-(1-\eta)} + \ldots\right]\right)$$

and therefore (since, from (29), $\frac{\delta^2}{\gamma} = \frac{\alpha^2}{\sigma^2}$)

$$\lim_{n\to\infty} n^{1-2\eta} \, \ln \mathbb{P}\big(|S_n| \geq \alpha n^{\eta}\big) \leq -\frac{\alpha^2}{2\sigma^2}. \tag{256}$$

Hence, this upper bound coincides with the exact asymptotic result in (82). It can be shown that the same conclusion also follows from Theorem 4.

## Appendix H
## Proof of Proposition 5

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

$$C(P_\theta, P_{\theta'}) \geq E_{\mathrm{L}}(P_\theta, P_{\theta'}) \geq \min_{i=1,2} \left\{ \frac{\delta_i^2}{2\gamma_i} - \frac{\delta_i^3}{6\gamma_i^2(1+\gamma_i)} \right\}. \tag{257}$$

The left-hand side of (257) holds since $E_{\mathrm{L}}$ 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 (257) follows from Lemma 4 (see (141)) and the definition of $E_{\mathrm{L}}$ 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),

$$\varepsilon_1 \triangleq D(P_\theta || P_{\theta'}), \quad \varepsilon_2 \triangleq D(P'_\theta || P_\theta). \tag{258}$$

The term on the left-hand side of (257) 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 (257) and the last inequality that

$$C(P_\theta, P_{\theta'}) \geq E_{\mathrm{L}}(P_\theta, P_{\theta'}) \geq \min_{i=1,2} \left\{ \frac{\varepsilon_i^2}{2\sigma_i^2}\left(1 - \frac{\varepsilon_i d_i}{3}\right) \right\}. \tag{259}$$

Based on the continuity assumption of the indexed family $\{P_\theta\}_{\theta \in \Theta}$, then it follows from (258) that

$$\lim_{\theta' \to \theta} \varepsilon_i = 0, \quad \forall\, i \in \{1, 2\}$$

and also, from (112) and (122) with $P_1$ and $P_2$ replaced by $P_\theta$ and $P'_\theta$ respectively, then

$$\lim_{\theta' \to \theta} d_i = 0, \quad \forall\, i \in \{1, 2\}.$$

It therefore follows from (144) and (259) that

$$\frac{J(\theta)}{8} \geq \lim_{\theta'\to\theta} \frac{E_{\mathrm{L}}(P_\theta, P_{\theta'})}{(\theta-\theta')^2} \geq \lim_{\theta'\to\theta} \min_{i=1,2} \left\{ \frac{\varepsilon_i^2}{2\sigma_i^2(\theta-\theta')^2} \right\}. \tag{260}$$

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 $\frac{J(\theta)}{8}$.

$$\begin{aligned}
&\lim_{\theta'\to\theta} \frac{\varepsilon_1^2}{2\sigma_1^2(\theta-\theta')^2} \\
&\overset{\text{(a)}}{=} \lim_{\theta'\to\theta} \frac{D(P_\theta||P_{\theta'})^2}{2\sigma_1^2(\theta-\theta')^2} \\
&\overset{\text{(b)}}{=} \frac{J(\theta)}{4} \lim_{\theta'\to\theta} \frac{D(P_\theta||P_{\theta'})}{\sigma_1^2} \\
&\overset{\text{(c)}}{=} \frac{J(\theta)}{4} \lim_{\theta'\to\theta} \frac{D(P_\theta||P_{\theta'})}{\sum_{x\in\mathcal{X}} P_\theta(x) \left(\ln \frac{P_\theta(x)}{P_{\theta'}(x)} - D(P_\theta||P_{\theta'})\right)^2} \\
&\overset{\text{(d)}}{=} \frac{J(\theta)}{4} \lim_{\theta'\to\theta} \frac{D(P_\theta||P_{\theta'})}{\sum_{x\in\mathcal{X}} P_\theta(x) \left(\ln \frac{P_\theta(x)}{P_{\theta'}(x)}\right)^2 - D(P_\theta||P_{\theta'})^2} \\
&\overset{\text{(e)}}{=} \frac{J(\theta)^2}{8} \lim_{\theta'\to\theta} \frac{(\theta-\theta')^2}{\sum_{x\in\mathcal{X}} P_\theta(x) \left(\ln \frac{P_\theta(x)}{P_{\theta'}(x)}\right)^2 - D(P_\theta||P_{\theta'})^2} \\
&\overset{\text{(f)}}{=} \frac{J(\theta)^2}{8} \lim_{\theta'\to\theta} \frac{(\theta-\theta')^2}{\sum_{x\in\mathcal{X}} P_\theta(x) \left(\ln \frac{P_\theta(x)}{P_{\theta'}(x)}\right)^2} \\
&\overset{\text{(g)}}{=} \frac{J(\theta)}{8}
\end{aligned} \tag{261}$$

where equality (a) follows from (258), 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

$$\lim_{\theta'\to\theta} \frac{\varepsilon_2^2}{2\sigma_2^2(\theta-\theta')^2} = \frac{J(\theta)}{8}$$

so

$$\lim_{\theta'\to\theta} \min_{i=1,2} \left\{ \frac{\varepsilon_i^2}{2\sigma_i^2(\theta-\theta')^2} \right\} = \frac{J(\theta)}{8}.$$

Hence, it follows from (260) that $\lim_{\theta'\to\theta} \frac{E_{\mathrm{L}}(P_\theta,P_{\theta'})}{(\theta-\theta')^2} = \frac{J(\theta)}{8}$. This completes the proof of (146).

We prove now equation (148). From (112), (122), (131) and (147) then

$$\widetilde{E}_{\mathrm{L}}(P_\theta, P_{\theta'}) = \min_{i=1,2} \frac{\varepsilon_i^2}{2d_i^2}$$

with $\varepsilon_1$ and $\varepsilon_2$ in (258). Hence,

$$\lim_{\theta'\to\theta} \frac{\widetilde{E}_{\mathrm{L}}(P_\theta, P_{\theta'})}{(\theta'-\theta)^2} \leq \lim_{\theta'\to\theta} \frac{\varepsilon_1^2}{2d_1^2(\theta'-\theta)^2}$$

and from (261) and the last inequality, it follows that

$$\begin{aligned}
&\lim_{\theta'\to\theta} \frac{\widetilde{E}_{\mathrm{L}}(P_\theta, P_{\theta'})}{(\theta'-\theta)^2} \\
&\leq \frac{J(\theta)}{8} \lim_{\theta'\to\theta} \frac{\sigma_1^2}{d_1^2} \\
&\overset{\text{(a)}}{=} \frac{J(\theta)}{8} \lim_{\theta'\to\theta} \frac{\sum_{x\in\mathcal{X}} P_\theta(x) \left(\ln \frac{P_\theta(x)}{P_{\theta'}(x)} - D(P_\theta||P_{\theta'})\right)^2}{\left(\max_{x\in\mathcal{X}} \left|\ln \frac{P_\theta(x)}{P_{\theta'}(x)} - D(P_\theta||P_{\theta'})\right|\right)^2}.
\end{aligned} \tag{262}$$

It is clear that the second term on the right-hand side of (262) 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 (262) 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

$$\lim_{\theta'\to\theta} \frac{\sum_{x\in\mathcal{X}} P_\theta(x) \left(\ln \frac{P_\theta(x)}{P_{\theta'}(x)} - D(P_\theta||P_{\theta'})\right)^2}{\left(\max_{x\in\mathcal{X}} \left|\ln \frac{P_\theta(x)}{P_{\theta'}(x)} - D(P_\theta||P_{\theta'})\right|\right)^2} = (1-\alpha)\theta$$

so, for any $\theta \in \mathbb{R}^+$, the above limit can be made arbitrarily close to zero by choosing $\alpha$ close enough to 1. This completes the proof of (148), and also the proof of Proposition 5.

## APPENDIX I
## PROOF OF LEMMA 5

In order to prove Lemma 5, one needs to show that if $\rho'(1) < \infty$ then

$$\lim_{C\to 1} \sum_{i=1}^{\infty} (i+1)^2 \Gamma_i \left[h_2\left(\frac{1-C^{\frac{i}{2}}}{2}\right)\right]^2 = 0 \tag{263}$$

which then yields from (183) that $B \to \infty$ in the limit where $C \to 1$.

By the assumption in Lemma 5 where $\rho'(1) < \infty$ then $\sum_{i=1}^\infty i\rho_i < \infty$, and therefore it follows from the Cauchy-Schwarz inequality that

$$\sum_{i=1}^{\infty} \frac{\rho_i}{i} \geq \frac{1}{\sum_{i=1}^\infty i\rho_i} > 0.$$

Hence, the *average* degree of the parity-check nodes is finite

$$d_{\mathrm{c}}^{\mathrm{avg}} = \frac{1}{\sum_{i=1}^\infty \frac{\rho_i}{i}} < \infty.$$

The infinite sum $\sum_{i=1}^\infty (i+1)^2 \Gamma_i$ converges under the above assumption since

$$\begin{aligned}
&\sum_{i=1}^{\infty} (i+1)^2 \Gamma_i \\
&= \sum_{i=1}^{\infty} i^2 \Gamma_i + 2\sum_{i=1}^{\infty} i\Gamma_i + \sum_i \Gamma_i \\
&= d_{\mathrm{c}}^{\mathrm{avg}} \left(\sum_{i=1}^{\infty} i\rho_i + 2\right) + 1 < \infty.
\end{aligned}$$

where the last equality holds since

$$\begin{aligned}\Gamma_i &= \frac{\frac{\rho_i}{i}}{\int_0^1 \rho(x)\,\mathrm{d}x} \\ &= d_{\mathrm{c}}^{\mathrm{avg}} \left(\frac{\rho_i}{i}\right), \quad \forall\, i \in \mathbb{N}.\end{aligned}$$

The infinite series in (263) therefore uniformly converges for $C \in [0,1]$, hence, the order of the limit and the infinite sum can be exchanged. Every term of the infinite series in (263) converges to zero in the limit where $C \to 1$, hence the limit in (263) is zero. This completes the proof of Lemma 5.

## APPENDIX J
## PROOF OF THE PROPERTIES IN (205) FOR OFDM SIGNALS

Consider an OFDM signal from Section VI-I. The sequence in (203) is a martingale due to basic properties of martingales. From (202), for every $i \in \{0,\dots,n\}$

$$Y_i = \mathbb{E}\Big[\max_{0\le t\le T}\big|s(t;X_0,\dots,X_{n-1})\big| \,\Big|\, X_0,\dots,X_{i-1}\Big].$$

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 $1 \le i \le n$,

$$\begin{aligned}Y_{i-1} &= \mathbb{E}\Big[\max_{0\le t\le T}\big|s(t;X_0,\dots,X'_{i-1},X_i,\dots,X_{n-1})\big| \,\Big|\, X_0,\dots,X_{i-2}\Big] \\ &= \mathbb{E}\Big[\max_{0\le t\le T}\big|s(t;X_0,\dots,X'_{i-1},X_i,\dots,X_{n-1})\big| \,\Big|\, X_0,\dots,X_{i-2},X_{i-1}\Big].\end{aligned}$$

Since $|\mathbb{E}(Z)| \le \mathbb{E}(|Z|)$, then for $i \in \{1,\dots,n\}$

$$|Y_i - Y_{i-1}| \le \mathbb{E}_{X'_{i-1},X_i,\dots,X_{n-1}}\Big[|U-V| \,\Big|\, X_0,\dots,X_{i-1}\Big] \tag{264}$$

where

$$\begin{aligned}U &\triangleq \max_{0\le t\le T}\big|s(t;X_0,\dots,X_{i-1},X_i,\dots,X_{n-1})\big| \\ V &\triangleq \max_{0\le t\le T}\big|s(t;X_0,\dots,X'_{i-1},X_i,\dots,X_{n-1})\big|.\end{aligned}$$

From (200)

$$\begin{aligned}|U-V| &\le \max_{0\le t\le T}\big|s(t;X_0,\dots,X_{i-1},X_i,\dots,X_{n-1}) - s(t;X_0,\dots,X'_{i-1},X_i,\dots,X_{n-1})\big| \\ &= \max_{0\le t\le T}\frac{1}{\sqrt{n}}\left|\big(X_{i-1}-X'_{i-1}\big)\exp\Big(\frac{j\,2\pi i t}{T}\Big)\right| \\ &= \frac{|X_{i-1}-X'_{i-1}|}{\sqrt{n}}. \end{aligned} \tag{265}$$

By assumption, $|X_{i-1}| = |X'_{i-1}| = 1$, and therefore a.s.

$$|X_{i-1}-X'_{i-1}| \le 2 \Longrightarrow |Y_i - Y_{i-1}| \le \frac{2}{\sqrt{n}}.$$

In the following, an upper bound on the conditional variance $\mathrm{Var}(Y_i \,|\, \mathcal{F}_{i-1}) = \mathbb{E}\big[(Y_i - Y_{i-1})^2 \,|\, \mathcal{F}_{i-1}\big]$ is obtained. Since $\big(\mathbb{E}(Z)\big)^2 \le \mathbb{E}(Z^2)$ for a real-valued RV $Z$, then from (264) and (265)

$$\mathbb{E}\big[(Y_i - Y_{i-1})^2 \,|\mathcal{F}_{i-1}\big] \le \frac{1}{n}\cdot \mathbb{E}_{X'_{i-1}}\big[|X_{i-1}-X'_{i-1}|^2 \,|\, \mathcal{F}_i\big]$$

where $\mathcal{F}_i$ is the $\sigma$-algebra that is generated by $X_0, \ldots, X_{i-1}$. Due to symmetry of the PSK constellation, it follows that

$$\begin{aligned}
&\mathbb{E}\big[(Y_i - Y_{i-1})^2 \,|\, \mathcal{F}_{i-1}\big] \\
&\leq \frac{1}{n}\, \mathbb{E}_{X'_{i-1}} \big[|X_{i-1} - X'_{i-1}|^2 \,|\, \mathcal{F}_i\big] \\
&= \frac{1}{n}\, \mathbb{E}\big[|X_{i-1} - X'_{i-1}|^2 \,|\, X_0, \ldots, X_{i-1}\big] \\
&= \frac{1}{n}\, \mathbb{E}\big[|X_{i-1} - X'_{i-1}|^2 \,|\, X_{i-1}\big] \\
&= \frac{1}{n}\, \mathbb{E}\Big[|X_{i-1} - X'_{i-1}|^2 \,|\, X_{i-1} = e^{\frac{j\pi}{M}}\Big] \\
&= \frac{1}{nM} \sum_{l=0}^{M-1} \Big| e^{\frac{j\pi}{M}} - e^{\frac{j(2l+1)\pi}{M}} \Big|^2 \\
&= \frac{4}{nM} \sum_{l=1}^{M-1} \sin^2\Big(\frac{\pi l}{M}\Big) = \frac{2}{n}.
\end{aligned}$$

## Acknowledgment

The feedback of I. Emre Telatar on this chapter, which in particular motivated the refinement of the analysis in Section V-B, is gratefully acknowledged. We are thankful to Nicholas Kalouptsidis for suggesting Remark 13, and also for shortening the proof in Appendix A. Yuri Polyanskiy and Sergio Verdú are acknowledged for notifying the author on their work in [57]. Kostis Xenoulis and Nicholas Kalouptsidis are acknowledged for their collaboration on random coding theorems, that served for the writing of Section VI-J. Ronen Eshel is gratefully acknowledged for collaboration on concentration for LDPC code ensembles and iterative message-passing decoding, especially for the material in Section VI-G that relies on his work in [24]. The hospitality of the Bernoulli inter-faculty center at EPFL (in Lausanne, Switzerland) during August 2011 was helpful for the writing of this chapter, and it 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," *Proceedings of the 2012 International Zurich Seminar (IZS 2012)*, pp. 44–47, 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] A. Barg and G. D. Forney, "Random codes: minimum distances and error exponents," *IEEE Trans. on Information Theory*, vol. 48, no. 9, pp. 2568–2573, September 2002.

[9] S. Benedetto and E. Biglieri, *Principles of Digital Transmission with Wireless Applications*, Kluwer Academic/Plenum Publishers, 1999.

[10] 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.

[11] P. Billingsley, *Probability and Measure*, Wiley Series in Probability and Mathematical Statistics, Third Edition, 1995.

[12] R. E. Blahut, "Hypothesis testing and information theory," *IEEE Trans. on Information Theory*, vol. 20, no. 4, pp. 405–417, July 1974.

[13] S. Boyd, L. O. Chua and C. A. Desoer, "Analytical foundations of Volterra series," *IMA Journal of Mathematical Control & Information*, vol. 1, pp. 243–282, 1984.

[14] M. Breiling, "A logarithmic upper bound on the minimum distance of turbo codes," *IEEE Trans. on Information Theory*, vol. 50, pp. 1692–1710, August 2004.

[15] F. Chung and L. Lu, *Complex Graphs and Networks*, *Regional Conference Series in Mathematics*, vol. 107, 2006.

[16] 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.

[17] 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.

[18] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley and Sons, second edition, 2006.

[19] 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.

[20] A. Dembo, "Moderate deviations for martingales with bounded jumps," *Electronic Communications in Probability*, vol. 1, no. 3, pp. 11–17, March 1996.

[21] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, Springer, second edition, 1997.

[22] J. Douillard, M. Jezequel, C. Berrou, A. Picart, P. Didier, and A. Glavieux, "Iterative correction of intersymbol interference: Turbo-equalization," Europ. Trans. Commun., vol. 6, pp. 507-511, Sept. 1995.

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

[24] R. Eshel, *Aspects of Convex Optimization and Concentration in Coding*, MSc thesis, Department of Electrical Engineering, Technion - Israel Institute of Technology, February 2012. [Online]. Available: http://webee.technion.ac.il/people/sason/Ronen_MSc_presentation.pdf.

[25] 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.

[26] X. Fan, I. Grama and Q. Liu, "Hoeffding's inequality for supermartingales," November 2011. [Online]. Available: http://arxiv.org/abs/1109.4359.

[27] X. Fan, I. Grama and Q. Liu, "The missing factor in Bennett's inequality," June 2012. [Online]. Available: http://http://arxiv.org/abs/1206.2592.

[28] D. Freedman, "On tail probabilities for martingales," *Annals of Probability*, vol. 3, no. 1, pp. 100–118, January 1975.

[29] R. G. Gallager, *Low-Density Parity-Check Codes*, Cambridge, MA: MIT Press, 1963.

[30] R. M. Gray, "Toeplitz and Circulant Matrices," *Foundations and Trends in Communication and Information Theory*, vol. 2, no. 3, pp. 155–239, 2006.

[31] G. Grimmett and D. Stirzaker, *Probability and Random Processes*, Oxford University Press, third edition, 2001.

[32] 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.

[33] 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.

[34] 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.

[35] F. den Hollander, *Large Deviations*, Fields Institute Monographs, American Mathematical Society, 2000.

[36] A. Kavćic, X. Ma and M. Mitzenmacher, "Binary intersymbol interference channels: Gallager bounds, density evolution, and code performance bounds,"*IEEE Trans. on Information Theory*, vol. 49, no. 7, pp. 1636-1652, July 2003.

[37] 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.

[38] 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.

[39] I. Kontoyiannis, L. A. Latras-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.

[40] 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.

[41] 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.

[42] M. Ledoux, *The Concentration of Measure Phenomenon*, Mathematical Surveys and Monographs, vol. 89, American Mathematical Society, 2001.

[43] 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.

[44] M. G. Luby, M. Mitzenmacher, M. A. Shokrollahi, and D. A. Spielman, "Improved low density parity check codes using irregular graphs and belief propagation" *Proceedings 1998 IEEE International Symposium on Information Theory*, Cambridge, MA, Aug. 1998, p. 117.

[45] 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.

[46] M. G. Luby, M. Mitzenmacher, M. A. Shokrollahi and D. A. Spielman, "Improved low-density parity-check codes using irregular graphs," *IEEE Trans. on Information Theory*, vol. 47, no. 2, pp. 585–598, February 2001.

[47] G. Lugosi, *Concentration of Measure Inequalities*, lecture notes, June 2009. [Online]. Available: http://www.econ.upf.edu/~lugosi/anu.pdf.

[48] P. Massart, *Concentration Inequalities and Model Selection*, Lecture Notes in Mathematics, vol. 1896, Springer, 2007.

[49] C. McDiarmid, "On the method of bounded differences," *Surveys in Combinatorics*, vol. 141, pp. 148–188, Cambridge University Press, Cambridge, 1989.

[50] C. McDiarmid, "Concentration," *Probabilistic Methods for Algorithmic Discrete Mathematics*, pp. 195–248, Springer, 1998.

[51] C. McDiarmid, "Centering sequences with bounded differences," *Combinatorics, Probability and Computing*, vol. 6, no. 1, pp. 79–86, March 1997.

[52] C. Méasson, A. Montanari and R. Urbanke, "Maxwell construction: The hidden bridge between iterative and maximum apposteriori decoding," *IEEE Trans. on Information Theory*, vol. 54, pp. 5277–5307, December 2008.

[53] A. F. Molisch, *Wireless Communications*, John Wiley and Sons, 2005.

[54] A. Montanari, "Tight bounds for LDPC and LDGM codes under MAP decoding," *IEEE Trans. on Information Theory*, vol. 51, no. 9, pp. 3247–3261, September 2005.

[55] B. Nakiboğlu, *Exponential Bounds on Error Probability with Feedback*, PhD dissertation, MIT, Boston, USA, February 2011.

[56] 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.

[57] 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.

[58] 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.

[59] A. Réyni, "On measures of entropy and information," *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 547–561, California, USA, 1961.

[60] 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.

[61] T. J. Richardson and R. Urbanke, *Modern Coding Theory*, Cambridge University Press, 2008.

[62] 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.

[63] I. Sason and S. Shamai, *Performance Analysis of Linear Codes under Maximum-Likelihood Decoding: A Tutorial,* published in the *Foundations and Trends in Communications and Information Theory*, vol. 3, no. 1–2, pp. 1–222, NOW Publishers, Delft, the Netherlands, July 2006.

[64] 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.

[65] I. Sason and R. Eshel, "On concentration of measures for LDPC code ensembles," *Proceedings of the 2011 IEEE International Symposium on Information Theory (ISIT 2011)*, pp. 1273–1277, Saint Petersburg, Russia, August 2011.

[66] I. Sason, "On the concentration of the crest factor for OFDM signals," *Proceedings of the 8th International Symposium on Wireless Communication Systems (ISWCS '11)*, pp. 784–788, Aachen, Germany, November 2011.

[67] I. Sason, "Moderate deviations analysis of binary hypothesis testing," *Proceedings of the 2012 IEEE International Symposium on Information Theory (ISIT 2012)*, pp. 826–830, MIT, Boston, MA, USA, July 2012.

[68] I. Sason, "Tightened exponential bounds for discrete-time conditionally symmetric martingales with bounded increments," *Proceedings of the 2012 International Workshop on Applied Probability (IWAP 2012)*, p. 59, Jerusalem, Israel, June 11–14, 2012. Full version of this work available in http://arxiv.org/abs/1201.0533.

[69] I. Sason, K. Xenoulis and N. Kalouptsidis, "New achievable rates for nonlinear Volterra channels via martingale inequalities," *in preparation*.

[70] E. Shamir and J. Spencer, "Sharp concentration of the chromatic number on random graphs," *Combinatorica*, vol. 7, no. 1, pp. 121–129, 1987.

[71] A. Shokrollahi, "Capacity-achieving sequences," *IMA Volume in Mathematics and its Applications*, vol. 123, pp. 153–166, 2000.

[72] M. Sipser and D. A. Spielman, "Expander codes," *IEEE Trans. on Information Theory*, vol. 42, no. 6, pp. 1710–1722, November 1996.

[73] 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.

[74] M. Talagrand, "Concentration of measure and isoperimteric inequalities in product spaces," *Publications Mathématiques de l'I.H.E.S*, vol. 81, pp. 73–205, 1995.

[75] M. Talagrand, "A new look at independence," *Annals of Probability*, vol. 24, no. 1, pp. 1–34, January 1996.

[76] V. Y. F. Tan, "Moderate-deviations of lossy source coding for discrete and Gaussian sources," *Proceedings of the 2012 IEEE International Symposium on Information Theory (ISIT 2012)*, pp. 925–929, MIT, Boston, MA, USA, July 2012.

[77] 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.

[78] D. Williams, *Probability with Martingales*, Cambridge University Press, 1991.

[79] 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.

[80] 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.

[81] 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.

[82] K. Xenoulis, N. Kalouptsidis and I. Sason, "New achievable rates for nonlinear Volterra channels via martingale inequalities," *Proceedings of the 2012 IEEE International Symposium on Information Theory (ISIT 2012)*, pp. 1430–1434, MIT, Boston, MA, USA, July 2012.