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

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 some concentration inequalities for martingales that include the Azuma-Hoeffding, Bennett, Freedman and McDiarmid inequalities. These inequalities are also specialized for sums of independent and bounded random variables that include the inequalities by Bernstein, Bennett, Hoeffding, and Kearns & Saul. An improvement of the martingale inequalities for some subclasses of martingales (e.g., the conditionally symmetric martingales) is discussed in detail, and some new refined inequalities are derived. The first part of this chapter also considers a geometric interpretation of some of these inequalities, providing an insight on the inter-connections between them. The second part of this chapter exemplifies the potential applications of the considered martingale inequalities in the context of information theory and related topics. The considered applications include binary hypothesis testing, concentration for codes defined on graphs, concentration for OFDM signals, and a use of some martingale inequalities for the derivation of achievable rates under ML decoding and lower bounds on the error exponents for random coding over some linear or non-linear communication channels.

Index Terms

Concentration of measures, error exponents, Fisher information, information divergence, large deviations, martingales.

I. INTRODUCTION

Inequalities providing upper bounds on probabilities of the type $P(|X - \bar{X}| \geq t)$ (or $P(X - \bar{X} \geq t)$ for a random variable (RV) $X$, where $\bar{X}$ denotes the expectation or median of $X$, have been among the main tools of probability theory. These inequalities are known as concentration inequalities, and they have been subject to interesting developments during the last two decades. 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” [77]. The exact meaning of such a statement clearly needs to be clarified rigorously, but it will often mean that such a random variable $X$ concentrates around $\bar{X}$ in a way that the probability of the event $\{|X - \bar{X}| > t\}$ decays exponentially in $t$ (for $t \geq 0$). The foundations in concentration of measures have been introduced, e.g., in [41], [44], [49], [53] and [76].

In recent years, concentration inequalities have been studied and used as a powerful tool in various areas such as convex geometry, functional analysis, statistical physics, dynamical systems, probability (random matrices, Markov processes, random graphs, percolation), information theory and statistics, learning theory and randomized algorithms. Several techniques have been developed so far to prove concentration of measures. These include:

- Talagrand’s inequalities for product measures (see, e.g., [53, Chapter 4], [74, Chapter 6], [76] and [77] with some information-theoretic applications in [39] and [40]).
- The entropy method and logarithmic-Sobolev inequalities (see, e.g., [2], [18, Chapter 14], [41, Chapter 5], [44]–[51] with information-theoretic aspects in, e.g., [34], [35], [37] and [38]). This methodology and its remarkable information-theoretic links will be considered in the next chapter.
- Transportation-cost inequalities that originated from information theory (see, e.g., [18, Chapters 12, 13], [27], [41, Chapter 6], [46], [47], [48]). This methodology and its information-theoretic aspects will be considered in the next chapter, with a discussion on the relation of transportation-cost inequalities to the entropy method and logarithmic-Sobolev inequalities.
- Stein’s method is used to prove concentration inequalities, a.k.a. concentration inequalities with exchangeable pairs (see, e.g., [10], [11] and [12]). This relatively recent framework is not addressed in this work.
- The martingale approach is useful for the derivation of concentration inequalities (see, e.g., [1, Chapter 7], [13], [14], [53] with information-theoretic aspects, e.g., in [20], [32], [43], [62], [63], [67]–[70], [83] and [84]). This chapter mainly considers this last methodology, focusing on concentration inequalities that are derived for discrete-time martingales with bounded jumps.
Let \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) be a function that is characterized by bounded differences whenever the \( n \)-dimensional vectors differ in only one coordinate. A common method for proving bounded concentration of such a function of \( n \) independent RVs, around the expected value \( \mathbb{E}[f] \), is called McDiarmid’s inequality or the 'independent bounded-differences inequality' (see [53, Theorem 3.1]). This inequality was proved (with some possible extensions) via the martingale approach (see [53, 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, e.g., [53, Section 3].

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 [30] who proved this inequality for a sum of independent and bounded random variables, and Azuma [3] 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 [54]; this extension provides sharper concentration results for, e.g., sequences that are related to sampling without replacement.

The use of the Azuma-Hoeffding inequality was introduced to the computer science literature in [71] 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 this number is defined is 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 \( p \). It is noted that the concentration result in [71] 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 [43], [62] and [73]. During the last decade, the Azuma-Hoeffding inequality has been extensively used for proving concentration of measures in coding theory (see, e.g., [63] 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 [80] 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 [82]–[84], the martingale approach was used to derive achievable rates and random coding error exponents for linear and non-linear additive white Gaussian noise channels (with or without memory).

This chapter is structured as follows: Section II presents briefly discrete-time (sub/super) martingales, Section III presents some basic inequalities that are widely used for proving concentration inequalities via the martingale approach. Section IV derives some refined versions of the Azuma-Hoeffding inequality, and it considers interconnections between these concentration inequalities. Section V introduces Freedman’s inequality with a refined version of this inequality, and these inequalities are specialized to get concentration inequalities for sums of independent and bounded random variables. Section VI 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 VII 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 briefly in Section VIII.

In addition to the presentation in this chapter, the reader is also referred to very nice surveys on concentration inequalities via the martingales’ approach in [1, Chapter 11], [13, Chapter 2], [14], [52] and [53]. It is noted that the main focus of this chapter is on information-theoretic aspects which makes the presentation in this chapter different from these other excellent sources for studying concentration inequalities via the martingale approach. The presentation in this chapter is aimed to be self-contained.
II. DISCRETE-TIME MARTINGALES

A. Martingales

This subsection provides a brief revision on martingales to set definitions and notation. We will not need for this chapter any result about martingales beyond the definition and the few basic properties mentioned in the following.

Definition 1: [Discrete-time martingales] Let \((\Omega, \mathcal{F}, \mathbb{P})\) be a probability space, and let \(n \in \mathbb{N}\). A sequence \(\{X_i, \mathcal{F}_i\}_{i=0}^n\), where the \(X_i\)'s are random variables and the \(\mathcal{F}_i\)'s are \(\sigma\)-algebras, is a martingale if the following conditions are satisfied:

1) \(\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}_n\) is a sequence of sub \(\sigma\)-algebras of \(\mathcal{F}\) (the sequence \(\{\mathcal{F}_i\}_{i=0}^n\) is called a filtration); usually, \(\mathcal{F}_0 = \{\emptyset, \Omega\}\) and \(\mathcal{F}_n = \mathcal{F}\).

2) \(X_i \in L^1(\Omega, \mathcal{F}_i, \mathbb{P})\) for every \(i \in \{0, \ldots, n\}\); this means that each \(X_i\) is defined on the same sample space \(\Omega\), it is \(\mathcal{F}_i\)-measurable, and \(\mathbb{E}[|X_i|] = \int_{\Omega} |X_i(\omega)| \, d\mathbb{P}(\omega) < \infty\).

3) For all \(i \in \{1, \ldots, n\}\), the equality \(X_{i-1} = \mathbb{E}[X_i | \mathcal{F}_{i-1}]\) holds almost surely (a.s.).

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

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

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

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

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

Then, the sequence \(X_0, X_1, \ldots, X_n\) forms a martingale (w.r.t. the above filtration) since

1) The RV \(X_i = \mathbb{E}[X | \mathcal{F}_i]\) is \(\mathcal{F}_i\)-measurable, and also \(\mathbb{E}[|X_i|] \leq \mathbb{E}[|X|] < \infty\).

2) By construction \(\{\mathcal{F}_i\}_{i=0}^n\) is a filtration.

3) For every \(i \in \{1, \ldots, n\}\)

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

Remark 3: In continuation to Remark 2, the setting where \(\mathcal{F}_0 = \{\emptyset, \Omega\}\) and \(\mathcal{F}_n = \mathcal{F}\) gives that \(X_0, X_1, \ldots, X_n\) is a martingale sequence with

\[
X_0 = \mathbb{E}[X | \mathcal{F}_0] = \mathbb{E}[X], \quad X_n = \mathbb{E}[X | \mathcal{F}_n] = X \quad \text{a.s.}
\]

In this case, one gets a martingale sequence where the first element is the expected value of \(X\), and the last element 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 expected value. At each step, more and more information about the random variable \(X\) is revealed until its value is known almost surely.

Example 1: Let \(\{U_k\}_{k=1}^k\) be independent random variables on a joint probability space \(\{\Omega, \mathcal{F}, \mathbb{P}\}\), and assume that \(\mathbb{E}[U_k] = 0\) and \(\mathbb{E}[^{|U_k|]} < \infty\) for every \(k\). Let us define

\[
X_k = \sum_{j=1}^k U_j, \quad \forall k \in \{1, \ldots, n\}
\]

with \(X_0 = 0\). Define the natural filtration where \(\mathcal{F}_0 = \{\emptyset, \Omega\}\), and

\[
\mathcal{F}_k = \sigma(X_1, \ldots, X_k)
= \sigma(U_1, \ldots, U_k), \quad \forall k \in \{1, \ldots, n\}.
\]

Note that \(\mathcal{F}_k = \sigma(X_1, \ldots, X_k)\) denotes the minimal \(\sigma\)-algebra that includes all the sets of the form \(\{\omega \in \Omega : (X_1(\omega) \leq \alpha_1, \ldots, X_k(\omega) \leq \alpha_k)\}\) where \(\alpha_j \in \mathbb{R} \cup \{-\infty, +\infty\}\) for \(j \in \{1, \ldots, k\}\). It is easy to verify that \(\{X_k, \mathcal{F}_k\}_{k=0}^n\) is a martingale sequence; this simply implies that all the concentration inequalities that apply to discrete-time martingales (like those introduced in this chapter) can be particularized to concentration inequalities for sums of independent random variables.
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 random process that is both a sub and super-martingale is a martingale, and vice versa. Furthermore, $\{X_i, \mathcal{F}_i\}$ is a sub-martingale if and only if $\{-X_i, \mathcal{F}_i\}$ is a super-martingale. The following properties are direct consequences of Jensen’s inequality for conditional expectations:

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

Example 2: if $\{X_i, \mathcal{F}_i\}$ is a martingale, then $\{|X_i|, \mathcal{F}_i\}$ is a sub-martingale. Furthermore, if $X_i \in L^2(\Omega, \mathcal{F}_i, \mathbb{P})$ then also $\{X_i^2, \mathcal{F}_i\}$ is a sub-martingale. Finally, if $\{X_i, \mathcal{F}_i\}$ is a non-negative sub-martingale and $X_i \in L^2(\Omega, \mathcal{F}_i, \mathbb{P})$ then also $\{X_i^2, \mathcal{F}_i\}$ is a sub-martingale.

III. BASIC CONCENTRATION INEQUALITIES VIA THE MARTINGALE APPROACH

In the following section, some basic inequalities that are widely used for proving concentration inequalities are presented, whose derivation relies on the martingale approach. Their proofs convey the main concepts of the martingale approach for proving concentration. Their presentation also motivates some further refinements that are considered in the continuation of this chapter.

A. The Azuma-Hoeffding Inequality

The Azuma-Hoeffding inequality is a useful concentration inequality for bounded-difference martingales. It was proved in [30] for independent bounded random variables, followed by a discussion on sums of dependent random variables; this inequality was later derived in [3] for the more general setting of bounded-difference martingales. In the following, this inequality is introduced.

Theorem 1: [Azuma-Hoeffding inequality] Let $\{X_k, \mathcal{F}_k\}_{k=0}^n$ be a discrete-parameter real-valued martingale sequence. Suppose that, for every $k \in \{1, \ldots, n\}$, the condition $|X_k - X_{k-1}| \leq d_k$ holds a.s. for a real-valued sequence $\{d_k\}_{k=1}^n$ of non-negative numbers. Then, for every $\alpha > 0$,

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

(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 $\alpha > 0$,

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

(2)

Let $\xi_i \triangleq X_i - X_{i-1}$ for $i = 1, \ldots, n$ designate the jumps of the martingale sequence. 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,

$$\mathbb{P}(X_n - X_0 \geq \alpha) = \mathbb{P}\left(\sum_{i=1}^n \xi_i \geq \alpha\right) \leq e^{-\alpha t} \mathbb{E}\left[\exp\left(t \sum_{i=1}^n \xi_i\right)\right], \quad \forall t \geq 0.$$

(3)

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.
Furthermore,

\[
\mathbb{E} \left[ \exp \left( t \sum_{k=1}^{n} \xi_k \right) \right] = \mathbb{E} \left[ \exp \left( t \sum_{k=1}^{n} \xi_k \right) \mid \mathcal{F}_{n-1} \right] = \mathbb{E} \left[ \exp \left( t \sum_{k=1}^{n-1} \xi_k \right) \mathbb{E}[\exp(t \xi_n) \mid \mathcal{F}_{n-1}] \right]
\]

where the last equality holds since \( Y \triangleq \exp(t \sum_{k=1}^{n-1} \xi_k) \) is \( \mathcal{F}_{n-1} \)-measurable; this holds 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 \( Y \) are both \( \mathcal{F}_{n-1} \)-measurable, and \( \mathbb{E}[XY \mid \mathcal{F}_{n-1}] = Y \mathbb{E}[X \mid \mathcal{F}_{n-1}] \).

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 \([-d_k, d_k] \). Hence, for every \( k \) (note that \( \mathbb{E}[\xi_k \mid \mathcal{F}_{k-1}] = 0 \)),

\[
\mathbb{E} \left[ e^{t \xi_k} \mid \mathcal{F}_{k-1} \right] \leq \frac{(d_k + \xi_k)e^{td_k} + (d_k - \xi_k)e^{-td_k}}{2d_k} \leq \frac{1}{2} (e^{td_k} + e^{-td_k}) = \cosh(td_k).
\]

Since, for every integer \( m \geq 0 \),

\[
(2m)! \geq (2m)(2m - 2) \ldots 2 = 2^m m!
\]

then, due to the power series expansions of the hyperbolic cosine and exponential functions,

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

which therefore implies that

\[
\mathbb{E} \left[ e^{t \xi_k} \mid \mathcal{F}_{k-1} \right] \leq e^{\frac{t^2 d_k^2}{2}}.
\]

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

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

which then gives (see (3)) that

\[
\mathbb{P}(X_n - X_0 \geq \alpha) \leq \exp \left( -\alpha t + \frac{t^2}{2} \sum_{k=1}^{n} d_k^2 \right), \quad \forall t \geq 0.
\]

An optimization over the free parameter \( t \geq 0 \) gives that \( t = \alpha (\sum_{k=1}^{n} d_k^2)^{-1} \), and

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

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 Theorem 1.

The proof of this inequality will be revisited later in this chapter for the derivation of some refined versions, whose use and advantage will be also exemplified.
Remark 4: In [53, Theorem 3.13], Azuma’s inequality is stated as follows: Let \( \{Y_k, \mathcal{F}_k\}_{k=0}^n \) 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 \( k \in \{1, \ldots, n\} \)). Assume that, for every \( k \), there exist some numbers \( a_k, b_k \in \mathbb{R} \) such that a.s. \( a_k \leq Y_k \leq b_k \). Then, for every \( r \geq 0 \),
\[
\mathbb{P}\left( \left| \sum_{k=1}^n Y_k \right| \geq r \right) \leq 2 \exp\left( -\frac{2r^2}{\sum_{k=1}^n (b_k-a_k)^2} \right).
\]
(7)

As a consequence of this inequality, consider a discrete-parameter real-valued martingale sequence \( \{X_k, \mathcal{F}_k\}_{k=0}^n \) where \( a_k \leq X_k - X_{k-1} \leq b_k \) a.s. for every \( k \). Let \( Y_k \triangleq X_k - X_{k-1} \) for every \( k \in \{1, \ldots, n\} \), so since \( \{Y_k, \mathcal{F}_k\}_{k=0}^n \) is a martingale-difference sequence and \( \sum_{k=1}^n Y_k = X_n - X_0 \), then
\[
\mathbb{P}\left( |X_n - X_0| \geq r \right) \leq 2 \exp\left( -\frac{2r^2}{\sum_{k=1}^n (b_k-a_k)^2} \right), \quad \forall r > 0.
\]
(8)

Example 3: 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 \( \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \cdots \) 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 that
\[
\mathbb{P}(|X_n - X_0| \geq \alpha \sqrt{n}) \leq 2 \exp\left( -\frac{\alpha^2}{2d^2} \right).
\]
(9)

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 \( \alpha \geq 0 \),
\[
\lim_{n \to \infty} \mathbb{P}(|X_n - X_0| \geq \alpha \sqrt{n}) = 2 Q\left( \frac{\alpha}{d} \right)
\]
(10)

where
\[
Q(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp\left( -\frac{t^2}{2} \right) dt, \quad \forall x \in \mathbb{R}
\]
(11)
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
\[
\frac{1}{\sqrt{2\pi}} \frac{x}{1+x^2} \cdot e^{-\frac{x^2}{2}} < Q(x) < \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}, \quad \forall x > 0
\]
(12)
then it follows from (10) that the exponent on the right-hand side of (9) is the exact exponent in this example.

Example 4: In continuation to Example 3, 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 3. Let \( \{X_k, \mathcal{F}_k\}_{k=0}^\infty \) be defined similarly to Example 3, 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 \( \alpha \geq 0 \)
\[
\lim_{n \to \infty} \mathbb{P}(|X_n - X_0| \geq \alpha \sqrt{n}) = 2 Q\left( \frac{\alpha}{\sqrt{\gamma d^2}} \right).
\]
(13)

From the exponential upper and lower bounds of the Q-function in (12), the right-hand side of (13) scales exponentially like \( e^{-\frac{\alpha^2}{2\gamma^2 d^2}} \). Hence, the exponent in this example is improved by a factor \( \frac{1}{\sqrt{\gamma}} \) as compared Azuma’s inequality (that is the same as in Example 3 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 the focus of Section IV.
B. McDiarmid’s Inequality

The following useful inequality is due to McDiarmid ([52] or [54, 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 preceding subsection).

Theorem 2: [McDiarmid’s inequality] Let \( \{X_i\} \) be independent real-valued random variables (not necessarily i.i.d.), and assume that \( X_i : \Omega \rightarrow \mathbb{R} \) 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 \( k \in \{1, \ldots, n\} \),

\[
|g(X_1, \ldots, X_{k-1}, X_k, X_{k+1}, \ldots, X_n) - g(X_1, \ldots, X_{k-1}, \hat{X}_k, X_{k+1}, \ldots, X_n)| \leq d_k
\]

(14)

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 \( x \in \mathbb{R}^n \) is upper bounded by \( d_k \), i.e.,

\[
\sup |g(x) - g(x')| \leq d_k
\]

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

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

(15)

Remark 5: One can use the Azuma-Hoeffding inequality for a derivation of a concentration inequality in the considered setting. However, the following proof provides in this setting an improvement by a factor of 4 in the exponent of the bound.

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 \( \mathcal{F}_0 = \{\emptyset, \Omega\} \). Define

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

(16)

Note that \( \mathcal{F}_0 \subseteq \mathcal{F}_1 \subset \cdots \subseteq \mathcal{F}_n \) is a filtration, and

\[
\mathbb{E}[g(X_1, \ldots, X_n) | \mathcal{F}_0] = \mathbb{E}[g(X_1, \ldots, X_n)]
\]

\[
\mathbb{E}[g(X_1, \ldots, X_n) | \mathcal{F}_n] = g(X_1, \ldots, X_n).
\]

(17)

Hence, it follows from the last three equalities that

\[
g(X_1, \ldots, X_n) - \mathbb{E}[g(X_1, \ldots, X_n)] = \sum_{k=1}^n \xi_k.
\]

In the following, we need a 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 random variable.

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 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 first and third items. To prove the third item, let

\[
\xi_k = \mathbb{E}[g(X_1, \ldots, X_{k-1}, X_k, X_{k+1}, \ldots, X_n) | \mathcal{F}_k] - \mathbb{E}[g(X_1, \ldots, X_{k-1}, X_k, X_{k+1}, \ldots, X_n) | \mathcal{F}_{k-1}]
\]

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

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.

\[
\begin{align*}
|\xi_k - \hat{\xi}_k| & = |\mathbb{E}[g(X_1, \ldots, X_k, X_k+1, \ldots, X_n) | F_k] - \mathbb{E}[g(X_1, \ldots, X_k, X_k+1, \ldots, X_n) | \hat{F}_k]| \\
& = |\mathbb{E}[g(X_1, \ldots, X_k, X_k+1, \ldots, X_n) - g(X_1, \ldots, X_k, X_k+1, \ldots, X_n) \sigma(X_1, \ldots, X_k, X_k, \hat{X}_k)]| \\
& \leq \mathbb{E}[|g(X_1, \ldots, X_k, X_k+1, \ldots, X_n) - g(X_1, \ldots, X_k, X_k, \hat{X}_k)| \sigma(X_1, \ldots, X_k, X_k, \hat{X}_k)] \\
& \leq d_k.
\end{align*}
\]

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 $F_k$-measurable (since $\hat{X}_k$ and $X_k$ are independent copies, and $\xi_k - \hat{\xi}_k$ is a difference of $g(X_1, \ldots, X_k-1, X_k, X_k+1, \ldots, X_n)$ and $g(X_1, \ldots, X_k-1, \hat{X}_k, X_k+1, \ldots, X_n)$, then this is in essence saying that if a set $S \subseteq \mathbb{R}$ 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 $\mathbb{E}[\xi_k | F_{k-1}] = 0$ then a.s. the $F_{k-1}$-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, it implies that $\xi_k \in [-d_k, d_k]$ where the length of the interval is twice larger (i.e., $2d_k$).)

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

\[
\text{Var}(\xi_k | F_{k-1}) \leq -a_kb_k \triangleq \sigma_k^2.
\]

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 $d_k$) implies that for every $k \in \{1, \ldots, n\}$

\[
\begin{align*}
\mathbb{E}[e^{t\xi_k} | F_{k-1}] & \leq \mathbb{E}
\left[
\left(
\frac{(\xi_k - a_k)e^{tb_k} + (\xi_k + b_k)e^{ta_k}}{d_k}
\right)
\right]
\end{align*}
\]

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

\[
\begin{align*}
\mathbb{E}[e^{t\xi_k} | F_{k-1}] & \leq p_ke^{tb_k} + (1 - p_k)e^{ta_k} \\
& = e^{ta_k} (1 - p_k + p_ke^{td_k}) \\
& = e^{f_k(t)}
\end{align*}
\]

where

\[
f_k(t) \triangleq ta_k + \ln (1 - p_k + p_ke^{td_k}), \quad \forall t \in \mathbb{R}.
\]

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 (19),

\[
\mathbb{E}[e^{t\xi_k} | F_{k-1}] \leq e^{\frac{t^2 d_k^2}{8}}.
\]

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

\[
\mathbb{E} \left[ \exp \left( t \sum_{k=1}^{n} \xi_k \right) \right] \leq \exp \left( \frac{t^2}{8} \sum_{k=1}^{n} d_k^2 \right)
\]
which then gives from (3) that, for every \( t \geq 0 \),
\[
\mathbb{P}(g(X_1, \ldots, X_n) - \mathbb{E}[g(X_1, \ldots, X_n)] \geq \alpha) = \mathbb{P}
\left( \sum_{k=1}^{n} \xi_k \geq \alpha \right) \leq \exp \left( -\alpha t + \frac{t^2}{8} \sum_{k=1}^{n} d_k^2 \right).
\]

(23)

An optimization over the free parameter \( t \geq 0 \) gives that \( t = 4\alpha \left( \sum_{k=1}^{n} d_k^2 \right)^{-1} \), so
\[
\mathbb{P}(g(X_1, \ldots, X_n) - \mathbb{E}[g(X_1, \ldots, X_n)] \geq \alpha) \leq \exp \left( -\frac{2\alpha^2}{\sum_{k=1}^{n} d_k^2} \right).
\]

(24)

By replacing \( g \) with \( -g \), it follows that this bound is also valid for the probability
\[
\mathbb{P}(g(X_1, \ldots, X_n) - \mathbb{E}[g(X_1, \ldots, X_n)] \leq \alpha)
\]
which therefore gives the bound in (15). This completes the proof of Theorem 2.

\[\blacksquare\]

C. Hoeffding’s Inequality, and its Improved Version (the Kearns-Saul Inequality)

In the following, we derive a concentration inequality for sums of independent and bounded random variables as a consequence of McDiarmid’s inequality. This inequality is due to Hoeffding (see [30, Theorem 2]). An improved version of Hoeffding’s inequality, due to Kearns and Saul [33], is also introduced in the following.

Theorem 3 (Hoeffding): Let \( \{U_k\}_{k=1}^{n} \) be a sequence of independent and bounded random variables such that, for every \( k \in \{1, \ldots, n\} \), \( U_k \in [a_k, b_k] \) holds a.s. for some constants \( a_k, b_k \in \mathbb{R} \). Let \( \mu_n \triangleq \sum_{k=1}^{n} \mathbb{E}[U_k] \). Then,
\[
\mathbb{P}\left( \left| \sum_{k=1}^{n} U_k - \mu_n \right| \geq \alpha \sqrt{n} \right) \leq 2 \exp \left( -\frac{2\alpha^2 n}{\sum_{k=1}^{n} (b_k - a_k)^2} \right), \quad \forall \alpha \geq 0.
\]

(25)

Proof: Let \( g(x) \triangleq \sum_{k=1}^{n} x_k \) for every \( x \in \mathbb{R}^n \). Furthermore, let \( X_1, X'_1, \ldots, X_n, X'_n \) be independent random variables such that \( X_k \) and \( X'_k \) are independent copies of \( U_k \) for every \( k \in \{1, \ldots, n\} \). By assumption, it follows that for every \( k \)
\[
|g(X_1, \ldots, X_{k-1}, X_k, X_{k+1}, \ldots, X_n) - g(X_1, \ldots, X_{k-1}, X'_k, X_{k+1}, \ldots, X_n)| = |X_k - X'_k| \leq b_k - a_k
\]
holds a.s., where the last inequality is due to the fact that \( X_k \) and \( X'_k \) are both distributed like \( U_k \), so they are a.s. in the interval \( [a_k, b_k] \). It therefore follows from McDiarmid’s inequality that
\[
\mathbb{P}(\left| g(X_1, \ldots, X_n) - \mathbb{E}[g(X_1, \ldots, X_n)] \right| \geq \alpha \sqrt{n}) \leq 2 \exp \left( -\frac{2\alpha^2 n}{\sum_{k=1}^{n} (b_k - a_k)^2} \right), \quad \forall \alpha \geq 0.
\]

Since
\[
\mathbb{E}[g(X_1, \ldots, X_n)] = \sum_{k=1}^{n} \mathbb{E}[X_k] = \sum_{k=1}^{n} \mathbb{E}[U_k] = \mu_n
\]
and also \( (X_1, \ldots, X_n) \) have the same distribution as \( (U_1, \ldots, U_n) \) (note that the entries of each of these vectors are independent, and \( X_k \) is distributed like \( U_k \)), then
\[
\mathbb{P}(\left| g(U_1, \ldots, U_n) - \mu_n \right| \geq \alpha \sqrt{n}) \leq 2 \exp \left( -\frac{2\alpha^2 n}{\sum_{k=1}^{n} (b_k - a_k)^2} \right), \quad \forall \alpha \geq 0
\]
which is equivalent to (25).

\[\blacksquare\]

An improved version of Hoeffding’s inequality, due to Kearns and Saul [33] is introduced in the following. It is noted that a certain gap in the original proof of the improved inequality in [33] was recently solved in [7] by some tedious calculus. A shorter information-theoretic proof of the same basic inequality that is required for the
derivation of the improved concentration result follows from transportation-cost inequalities, as will be shown in the next chapter (see Section V-C of the next chapter). So, we only state the basic inequality, and use it to derive
the improved version of Hoeffding’s inequality.

To this end, let \( \xi_k \triangleq U_k - \mathbb{E}[U_k] \) for every \( k \in \{1, \ldots, n\} \), so \( \sum_{k=1}^n U_k - \mu_n = \sum_{k=1}^n \xi_k \) with \( \mathbb{E}[\xi_k] = 0 \) and \( \xi_k \in [a_k - \mathbb{E}[U_k], b_k - \mathbb{E}[U_k]] \). Following the argument that is used to derive inequality (19) gives

\[
\mathbb{E} \left[ \exp(t \xi_k) \right] \leq (1 - p_k) \exp(t(a_k - \mathbb{E}[U_k])) + p_k \exp(t(b_k - \mathbb{E}[U_k])) \triangleq \exp(f_k(t))
\]

where \( p_k \in [0, 1] \) is defined by

\[
p_k \triangleq \frac{\mathbb{E}[U_k] - a_k}{b_k - a_k}, \quad \forall k \in \{1, \ldots, n\}.
\]

The derivation of McDiarmid’s inequality (see (21)) gives that for all \( t \in \mathbb{R} \)

\[
f_k(t) \leq \frac{(b_k - a_k)^2}{8}.
\]

The improvement of this bound (see [7, Theorem 4]) gives that for all \( t \in \mathbb{R} \)

\[
f_k(t) \leq \begin{cases} 
\frac{(1-2p_k)(b_k-a_k)^2 t^2}{4 \ln \left( \frac{1}{\sqrt[4]{p_k}} \right)} & \text{if } p_k \neq \frac{1}{2} \\
\frac{(b_k-a_k)^2 t^2}{8} & \text{if } p_k = \frac{1}{2}.
\end{cases}
\]

Note that since

\[
\lim_{p \to \frac{1}{2}} \frac{1 - 2p}{\ln \left( \frac{1}{\sqrt[4]{p}} \right)} = \frac{1}{2}
\]

so the upper bound in (29) is continuous in \( p_k \), and it also improves the bound on \( f_k(t) \) in (28) unless \( p_k = \frac{1}{2} \) (where both bounds coincide in this case). From (29), we have \( f_k(t) \leq c_k t^2 \), for every \( k \in \{1, \ldots, n\} \) and \( t \in \mathbb{R} \), where

\[
c_k \triangleq \begin{cases} 
\frac{(1-2p_k)(b_k-a_k)^2}{4 \ln \left( \frac{1}{\sqrt[4]{p_k}} \right)} & \text{if } p_k \neq \frac{1}{2} \\
\frac{(b_k-a_k)^2}{8} & \text{if } p_k = \frac{1}{2}.
\end{cases}
\]

Hence, Chernoff’s inequality and the similarity of the two one-sided tail bounds give

\[
\mathbb{P} \left( \sum_{k=1}^n U_k - \mu_n \geq \alpha \sqrt{n} \right) \leq 2 \exp(-\alpha \sqrt{n}t) \prod_{k=1}^n \mathbb{E} \left[ \exp(t \xi_k) \right] = 2 \exp(-\alpha t \sqrt{n}) \cdot \exp \left( \sum_{k=1}^n c_k t^2 \right), \quad \forall t \geq 0.
\]

Finally, an optimization over the non-negative free parameter \( t \) leads to the following improved version of Hoeffding’s inequality in [33] (with the recent follow-up in [7]).

**Theorem 4 (Kearns-Saul inequality):** Let \( \{U_k\}_{k=1}^n \) be a sequence of independent and bounded random variables such that, for every \( k \in \{1, \ldots, n\} \), \( U_k \in [a_k, b_k] \) holds a.s. for some constants \( a_k, b_k \in \mathbb{R} \). Let \( \mu_n \triangleq \sum_{k=1}^n \mathbb{E}[U_k] \). Then,

\[
\mathbb{P} \left( \sum_{k=1}^n U_k - \mu_n \geq \alpha \sqrt{n} \right) \leq 2 \exp \left( -\frac{\alpha^2 n}{4 \sum_{k=1}^n c_k} \right), \quad \forall \alpha \geq 0.
\]

where \( \{c_k\}_{k=1}^n \) is introduced in (30) with the \( p_k \)'s that are given in (27). Moreover, the exponential bound (32) improves Hoeffding’s inequality, unless \( p_k = \frac{1}{2} \) for every \( k \in \{1, \ldots, n\} \).

The reader is referred to another recent refinement of Hoeffding’s inequality in [25], followed by some numerical comparisons.
IV. Refined Versions of The Azuma-Hoeffding Inequality

Example 4 in the preceding section serves to motivate a derivation of an improved concentration inequality with an additional constraint on the conditional variance of a martingale sequence. In the following, assume that $|X_k - X_{k-1}| \leq d$ holds a.s. for every $k$ (note that $d$ does not depend on $k$, so it is a global bound on the jumps of the martingale). A new condition is added for the derivation of the next concentration inequality, where it is assumed that

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

for some constant $\gamma \in (0, 1]$.

A. A Refinement of the Azuma-Hoeffding Inequality for Discrete-Time Martingales with Bounded Jumps

The following theorem appears in [52] (see also [17, Corollary 2.4.7]).

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

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

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

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

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) \leq 2 \exp\left(-n D\left(\frac{\delta + \gamma}{1 + \gamma} \bigg| \frac{\gamma}{1 + \gamma}\right)\right)$$  (33)

where

$$\gamma \triangleq \frac{\sigma^2}{d^2}, \quad \delta \triangleq \frac{\alpha}{d}$$  (34)

and

$$D(p|q) \triangleq p \ln\left(\frac{p}{q}\right) + (1 - p) \ln\left(\frac{1 - p}{1 - q}\right), \quad \forall p, q \in [0, 1]$$  (35)

is the divergence between the two probability distributions $(p, 1 - p)$ and $(q, 1 - q)$. If $\delta > 1$, then the probability on the left-hand side of (33) is equal to zero.

**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. We introduce in the following a lemma (see, e.g., [17, Lemma 2.4.1]) that is required for the proof of Theorem 5.

**Lemma 2 (Bennett):** Let $X$ be a real-valued random variable with $\overline{X} = \mathbb{E}(X)$ and $\mathbb{E}[(X - \overline{X})^2] \leq \sigma^2$ for some $\sigma > 0$. Furthermore, suppose that $X \leq b$ a.s. for some $b \in \mathbb{R}$. Then, for every $\lambda \geq 0$,

$$\mathbb{E}[e^{\lambda X}] \leq \frac{e^{\lambda \overline{X}} \left[ (b - \overline{X})^2 e^{-\frac{b^2}{2\sigma^2 X}} + \sigma^2 e^{\lambda(b - \overline{X})} \right]}{(b - \overline{X})^2 + \sigma^2}.$$  (36)

**Proof:** The lemma is trivial if $\lambda = 0$, so it is proved in the following for $\lambda > 0$. Let $Y \triangleq \lambda(X - \overline{X})$ for $\lambda > 0$. Then, by assumption, $Y \leq \lambda(b - \overline{X}) \triangleq b_Y$ a.s. and $\text{Var}(Y) \leq \lambda^2 \sigma^2 \triangleq \sigma_Y^2$. It is therefore required to show that if $\mathbb{E}[Y] = 0$, $Y \leq b_Y$, and $\text{Var}(Y) \leq \sigma_Y^2$, then

$$\mathbb{E}[e^Y] \leq \left( \frac{b_Y}{b_Y^2 + \sigma_Y^2} \right) e^{-\frac{\sigma_Y^2}{b_Y^2}} + \left( \frac{\sigma_Y^2}{b_Y^2 + \sigma_Y^2} \right) e^{b_Y}.$$  (37)

Let $Y_0$ be a random variable that gets the two possible values $-\frac{\sigma_Y^2}{b_Y}$ and $b_Y$, where

$$\mathbb{P}\left(Y_0 = -\frac{\sigma_Y^2}{b_Y}\right) = \frac{b_Y^2}{b_Y^2 + \sigma_Y^2}, \quad \mathbb{P}(Y_0 = b_Y) = \frac{\sigma_Y^2}{b_Y^2 + \sigma_Y^2}.$$  (38)

so inequality (37) is equivalent to showing that

$$\mathbb{E}[e^Y] \leq \mathbb{E}[e^{Y_0}].$$  (39)
To that end, let \( \phi \) be the unique parabola where the function
\[
f(y) \triangleq \phi(y) - e^y, \quad \forall y \in \mathbb{R}
\]
is zero at \( y = b_Y \), and \( f(y) = f'(y) = 0 \) at \( y = -\frac{\sigma_Y^2}{b_Y} \). Since \( \phi'' \) is constant then \( \phi''(y) = 0 \) at exactly one value of \( y \), call it \( y_0 \). Furthermore, since \( f(-\frac{\sigma_Y^2}{b_Y}) = f(b_Y) \) (both are equal to zero) then \( f'(y) = 0 \) for some \( y_1 \in (-\frac{\sigma_Y^2}{b_Y}, b_Y) \).

By the same argument, applied to \( f' \) on \( [-\frac{\sigma_Y^2}{b_Y}, y_1] \), it follows that \( y_0 \in (-\frac{\sigma_Y^2}{b_Y}, y_1) \). The function \( f \) is convex on \((-\infty, y_0)\) (since, on this interval, \( f''(y) = \phi''(y) - e^y > \phi''(y_0) - e^{y_0} = \phi''(y_0) - e^{y_0} = f''(y_0) = 0 \)), and its minimal value on this interval is at \( y = -\frac{\sigma_Y^2}{b_Y} \) (since at this point, \( f' \) is zero). Furthermore, \( f \) is concave on \([y_0, \infty)\) and it gets its maximal value on this interval at \( y = y_1 \). It implies that \( f \geq 0 \) on the interval \((-\infty, b_Y)\), so \( \mathbb{E}[f(Y)] \geq 0 \) for any random variable \( Y \) such that \( Y \leq b_Y \) a.s., which therefore gives that
\[
\mathbb{E}[e^Y] \leq \mathbb{E}[\phi(Y)]
\]
with equality if \( \mathbb{P}(Y \in \{-\frac{\sigma_Y^2}{b_Y}, b_Y\}) = 1 \). Since \( f''(y) \geq 0 \) for \( y < y_0 \) then \( \phi''(y) - e^y = f''(y) \geq 0 \), so \( \phi''(0) = \phi''(y) > 0 \) (recall that \( \phi'' \) is constant since \( \phi \) is a parabola). Hence, for any random variable \( Y \) of zero mean, \( \mathbb{E}[f(Y)] \) which only depends on \( \mathbb{E}[Y^2] \) is a non-decreasing function of \( \mathbb{E}[Y^2] \). The random variable \( Y_0 \) that takes values in \( \{-\frac{\sigma_Y^2}{b_Y}, b_Y\} \) and whose distribution is given in (38) is of zero mean and variance \( \mathbb{E}[Y_0^2] = \sigma_Y^2 \),

\[
\mathbb{E}[\phi(Y)] \leq \mathbb{E}[\phi(Y_0)].
\]

Note also that
\[
\mathbb{E}[\phi(Y_0)] = \mathbb{E}[e^{Y_0}]
\]
since \( f(y) = 0 \) (i.e., \( \phi(y) = e^y \)) if \( y = -\frac{\sigma_Y^2}{b_Y} \) or \( b_Y \), and \( Y_0 \) only takes these two values. Combining the last two inequalities with the last equality gives inequality (39), which therefore completes the proof of the lemma. \( \blacksquare \)

Applying Bennett’s inequality in Lemma 2 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.
\[
\mathbb{E}\left[ \exp(t \sum_{k=1}^{n} \xi_k) \mid \mathcal{F}_{k-1} \right] \leq \frac{\sigma^2 \exp(t d) + d^2 \exp\left(-\frac{t \sigma^2}{d}\right)}{d^2 + \sigma^2}. \tag{40}
\]

Hence, it follows from (4) and (40) that, for every \( t \geq 0 \),
\[
\mathbb{E}\left[ \exp\left(t \sum_{k=1}^{n} \xi_k\right) \right] \leq \left( \frac{\sigma^2 \exp(t d) + d^2 \exp\left(-\frac{t \sigma^2}{d}\right)}{d^2 + \sigma^2} \right)^n \mathbb{E}\left[ \exp\left(t \sum_{k=1}^{n-1} \xi_k\right) \right]
\]
and, by induction, it follows that for every \( t \geq 0 \)
\[
\mathbb{E}\left[ \exp\left(t \sum_{k=1}^{n} \xi_k\right) \right] \leq \left( \frac{\sigma^2 \exp(t d) + d^2 \exp\left(-\frac{t \sigma^2}{d}\right)}{d^2 + \sigma^2} \right)^n.
\]

From the definition of \( \gamma \) in (52), this inequality is rewritten as
\[
\mathbb{E}\left[ \exp\left(t \sum_{k=1}^{n} \xi_k\right) \right] \leq \left( \frac{\gamma \exp(t d) + \exp(-\gamma t d)}{1 + \gamma} \right)^n, \quad \forall t \geq 0. \tag{41}
\]

Let \( x \triangleq td \) (so \( x \geq 0 \)). Combining Chernoff’s inequality with (41) gives that, for every \( \alpha \geq 0 \) (where from the definition of \( \delta \) in (52), \( ct = \delta x \))
\[
\mathbb{P}(X_n - X_0 \geq \alpha n) \leq \exp(-\alpha nt) \mathbb{E}\left[ \exp\left(t \sum_{k=1}^{n} \xi_k\right) \right] \leq \left( \frac{\gamma \exp((1-\delta)x) + \exp(-\gamma(\delta + \gamma)x)}{1 + \gamma} \right)^n, \quad \forall x \geq 0. \tag{42}
\]
Consider first the case where \( \delta = 1 \) (i.e., \( \alpha = d \)), then (42) is particularized to
\[
\Pr(X_n - X_0 \geq dn) \leq \left( \frac{\gamma + x}{1 + \gamma} \right)^n, \quad \forall x \geq 0
\]
and the tightest bound within this form is obtained in the limit where \( x \to \infty \). This provides the inequality
\[
\Pr(X_n - X_0 \geq dn) \leq \left( \frac{\gamma}{1 + \gamma} \right)^n.
\tag{43}
\]
Otherwise, if \( \delta \in [0, 1) \), the minimization of the base of the exponent on the right-hand side of (42) w.r.t. the free non-negative parameter \( x \) yields that the optimized value is
\[
x = \left( \frac{1}{1 + \gamma} \right) \ln \left( \frac{\gamma + \delta}{\gamma(1 - \delta)} \right)
\tag{44}
\]
and its substitution into the right-hand side of (42) gives that, for every \( \alpha \geq 0 \),
\[
\Pr(X_n - X_0 \geq \alpha n) \leq \left[ \left( \frac{\gamma + \delta}{\gamma} \right)^{\frac{1 + \delta}{1 + \gamma}} (1 - \delta)^{\frac{1 - \delta}{1 + \gamma}} \right]^n
= \exp \left\{ -n \left[ \left( \frac{\gamma + \delta}{\gamma} \right) \ln \left( \frac{\gamma + \delta}{\gamma(1 - \delta)} \right) + \left( \frac{1 - \delta}{1 + \gamma} \right) \ln(1 - \delta) \right] \right\}
= \exp \left( -n D \left( \frac{\delta + \gamma}{1 + \gamma} \bigg| \bigg| \frac{\gamma}{1 + \gamma} \right) \right)
\tag{45}
\]
and the exponent is equal to \(+\infty\) if \( \delta > 1 \) (i.e., if \( \alpha > d \)). Applying inequality (45) to the martingale \( \{ -X_k, \mathcal{F}_k \}_{k=0}^\infty \) gives the same upper bound to the other tail-probability \( \Pr(X_n - X_0 \leq -\alpha n) \). 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 (33). This completes the proof of Theorem 5. ■

**Example 5:** 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 \( m \in \mathbb{N} \)
\[
\Pr(X_m - X_{m-1} = d \mid \mathcal{F}_{m-1}) = \varepsilon,
\]
\[
\Pr \left( X_m - X_{m-1} = -\frac{\varepsilon d}{1 - \varepsilon} \mid \mathcal{F}_{m-1} \right) = 1 - \varepsilon.
\]
This indeed implies that a.s. for every \( m \in \mathbb{N} \)
\[
\mathbb{E}[X_m - X_{m-1} \mid \mathcal{F}_{m-1}] = \varepsilon d + \left( -\frac{\varepsilon d}{1 - \varepsilon} \right) (1 - \varepsilon) = 0
\]
and since \( X_{m-1} \) is \( \mathcal{F}_{m-1} \)-measurable then a.s.
\[
\mathbb{E}[X_m \mid \mathcal{F}_{m-1}] = X_{m-1}.
\]
Since \( \varepsilon \in (0, \frac{1}{2}] \) then a.s.
\[
|X_m - X_{m-1}| \leq \max \left\{ d, \frac{\varepsilon d}{1 - \varepsilon} \right\} = d.
\]
From Azuma’s inequality, for every \( x \geq 0 \),
\[
\Pr(X_k \geq kx) \leq \exp \left( -\frac{kx^2}{2d^2} \right)
\tag{46}
\]
independently of the value of \( \varepsilon \) (note that \( X_0 = 0 \) a.s.). The concentration inequality in Theorem 5 enables one to get a better bound: Since a.s., for every \( m \in \mathbb{N} \),
\[
\mathbb{E}[(X_m - X_{m-1})^2 \mid \mathcal{F}_{m-1}] = d^2 \varepsilon + \left( -\frac{\varepsilon d}{1 - \varepsilon} \right)^2 (1 - \varepsilon) = \frac{d^2 \varepsilon}{1 - \varepsilon}
\]
then from (52)
\[ \gamma = \frac{\epsilon}{1 - \epsilon}, \quad \delta = \frac{x}{d} \]
and from (45), for every \( x \geq 0 \),
\[ \mathbb{P}(X_k \geq kx) \leq \exp \left( -kD \left( \frac{x(1 - \epsilon)}{d} + \epsilon \| \epsilon \right) \right). \tag{47} \]

Consider the case where \( \epsilon \to 0 \). Then, for arbitrary \( x > 0 \) and \( k \in \mathbb{N} \), Azuma’s inequality in (46) provides an upper bound that is strictly positive independently of \( \epsilon \), whereas the one-sided concentration inequality of Theorem 5 implies a bound in (47) that tends to zero. This exemplifies the improvement that is obtained by Theorem 5 in comparison to Azuma’s inequality.

**Remark 6:** As was noted, e.g., in [53, 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 \( \{\exp(t(X_k - X_0)), \mathcal{F}_k\}_{k=0}^\infty \) is a sub-martingale for every \( t \geq 0 \). Hence, by applying Doob’s maximal inequality for sub-martingales, it follows that for every \( \alpha \geq 0 \)
\[ \mathbb{P} \left( \max_{1 \leq k \leq n} X_k - X_0 \geq \alpha n \right) \]
\[ = \mathbb{P} \left( \max_{1 \leq k \leq n} \exp(t(X_k - X_0)) \geq \exp(\alpha nt) \right) \quad \forall t \geq 0 \]
\[ \leq \exp(-\alpha nt) \mathbb{E} \left[ \exp(t(X_n - X_0)) \right] \]
\[ = \exp(-\alpha nt) \mathbb{E} \left[ \exp \left( t \sum_{k=1}^n \xi_k \right) \right] \]
which coincides with the proof of Theorem 5 with the starting point in (3). This concept applies to all the concentration inequalities derived in this chapter.

**Corollary 1:** Let \( \{X_k, \mathcal{F}_k\}_{k=0}^n \) be a discrete-parameter real-valued martingale, and assume that \( |X_k - X_{k-1}| \leq d \) holds a.s. for some constant \( d > 0 \) and for every \( k \in \{1, \ldots, n\} \). Then, for every \( \alpha \geq 0 \),
\[ \mathbb{P}(|X_n - X_0| \geq \alpha n) \leq 2 \exp(-nf(\delta)) \tag{48} \]
where
\[ f(\delta) = \begin{cases} \ln(2) \left[ 1 - h_2 \left( \frac{1-\delta}{2} \right) \right], & 0 \leq \delta \leq 1 \\ +\infty, & \delta > 1 \end{cases} \tag{49} \]
and \( h_2(x) \overset{\Delta}{=} -x \log_2(x) - (1 - x) \log_2(1 - x) \) for \( 0 \leq x \leq 1 \) denotes the binary entropy function on base 2.

**Proof:** By substituting \( \gamma = 1 \) in Theorem 5 (i.e., since there is no constraint on the conditional variance, then one can take \( \sigma^2 = d^2 \)), the corresponding exponent in (33) is equal to
\[ D \left( \frac{1+\delta}{2} \bigg| \frac{1}{2} \right) = f(\delta) \]
since \( D(p||\frac{1}{2}) = \ln 2[1 - h_2(p)] \) for every \( p \in [0,1] \).

**Remark 7:** Corollary 1, which is a special case of Theorem 5 when \( \gamma = 1 \), forms a tightened version of the Azuma-Hoeffding inequality when \( d_k = d \). This can be verified by showing that \( f(\delta) > \frac{\delta^2}{2} \) for every \( \delta > 0 \), which is a direct consequence of Pinsker’s inequality. Figure 1 compares these two exponents, which nearly coincide for \( \delta \leq 0.4 \). Furthermore, the improvement in the exponent of the bound in Theorem 5 is shown in this figure as the value of \( \gamma \in (0,1) \) is reduced; this makes sense, since the additional constraint on the conditional variance in this theorem has a growing effect when the value of \( \gamma \) is decreased.
B. Geometric Interpretation

A common ingredient in proving Azuma’s inequality, and Theorem 5 is a derivation of an upper bound on the conditional expectation $E[e^{\xi_k | F_{k-1}}]$ for $t \geq 0$ where $E[\xi_k | F_{k-1}] = 0$, $\text{Var}[\xi_k | F_{k-1}] \leq \sigma^2$, and $|\xi_k| \leq d$ a.s. for some $\sigma, d > 0$ and for every $k \in \mathbb{N}$. The derivation of Azuma’s inequality and Corollary 1 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 5 is based on Bennett’s inequality which is applied to the conditional expectation above. The proof of Bennett’s inequality (see Lemma 2) is shortly reviewed, while adopting the 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 $(-\gamma d, e^{-\gamma d})$. As is verified in the proof of Lemma 2, it leads to the inequality $y(x) \leq \phi(x)$ for every $x \in (-\infty, d]$ where $\phi$ is the parabola that satisfies the conditions

$$\phi(d) = y(d) = e^{td}, \quad \phi(-\gamma d) = y(-\gamma d) = e^{-\gamma d}, \quad \phi'(-\gamma d) = y'(-\gamma d) = te^{-\gamma d}.$$  

Calculation shows that this parabola admits the form

$$\phi(x) = \frac{(x + \gamma d)e^{td} + (d - x)e^{-\gamma d}}{(1 + \gamma)d} + \frac{\alpha(\gamma d^2 + (1 - \gamma)d x - x^2)}{(1 + \gamma)^2 d^2}$$

where $\alpha \triangleq \frac{1}{(1 + \gamma)td + 1}e^{-\gamma d} - e^{td}$. Since $E[X] = 0$, $E[X^2] = \gamma d^2$ and $X \leq d$ (a.s.), then

$$E[e^{tX}] \leq E[\phi(X)] = \frac{\gamma e^{td} + e^{-\gamma d}}{1 + \gamma} = \frac{E[X^2]e^{td} + d^2 e^{-\frac{\text{Var}[X^2]}{d}}}{d^2 + E[X^2]}$$

which provides a geometric viewpoint to Bennett’s 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 $\frac{1}{1+\gamma}$ and $\frac{\gamma}{1+\gamma}$, respectively. This bound also holds when $E[X^2] \leq \sigma^2$ since the right-hand side of the inequality is a monotonic non-decreasing function of $E[X^2]$ (as it was verified in the proof Lemma 2). Applying Bennett’s inequality to the conditional law of $\xi_k$ given $F_{k-1}$ gives (40) (with $\gamma$ in (52)).
C. Improving the Refined Version of the Azuma-Hoeffding Inequality for Sub-classes of Discrete-Time Martingales

This following subsection derives an exponential deviation inequality that improves the bound in Theorem 5 for conditionally-symmetric discrete-time martingales with bounded increments. This subsection further assumes conditional symmetry of these martingales, as it is defined in the following:

Definition 2: Let \( \{X_k, \mathcal{F}_k\}_{k \in \mathbb{N}_0} \), where \( \mathbb{N}_0 \triangleq \mathbb{N} \cup \{0\} \), be a discrete-time and real-valued martingale, and let \( \xi_k \triangleq X_k - X_{k-1} \) for every \( k \in \mathbb{N} \) designate the jumps of the martingale. Then \( \{X_k, \mathcal{F}_k\}_{k \in \mathbb{N}_0} \) is called a conditionally symmetric martingale if, conditioned on \( \mathcal{F}_{k-1} \), the random variable \( \xi_k \) is symmetrically distributed around zero.

Our goal in this subsection is to demonstrate how the assumption of the conditional symmetry improves the existing deviation inequality in Section IV-A for discrete-time real-valued martingales with bounded increments. The exponent of the new bound is also compared to the exponent of the bound in Theorem 5 without the conditional symmetry assumption. Earlier results, serving as motivation to the discussion in this subsection, appear in [19, Section 4] and [60, Section 6]. The new exponential bounds can be also extended to conditionally symmetric sub or supermartingales, where the construction of these objects is exemplified later in this subsection. Additional results addressing weak-type inequalities, maximal inequalities and ratio inequalities for conditionally symmetric martingales were derived in [58], [59] and [85].

Before we present the new deviation inequality for conditionally symmetric martingales, this discussion is motivated by introducing some constructions of such martingales.

1) Construction of Discrete-Time, Real-Valued and Conditionally Symmetric Sub/ Supermartingales: Before proving the tightened inequalities for discrete-time conditionally symmetric sub/ supermartingales, it is in place to exemplify the construction of these objects.

Example 6: Let \( (\Omega, \mathcal{F}, \mathbb{P}) \) be a probability space, and let \( \{U_k\}_{k \in \mathbb{N}} \subseteq L^1(\Omega, \mathcal{F}, \mathbb{P}) \) be a sequence of independent random variables with zero mean. Let \( \mathcal{F}_k \) be the natural filtration of sub \( \sigma \)-algebras of \( \mathcal{F} \), where \( \mathcal{F}_0 \triangleq \{\emptyset, \Omega\} \) and \( \mathcal{F}_k = \sigma(U_1, \ldots, U_k) \) for \( k \geq 1 \). Furthermore, for \( k \in \mathbb{N} \), let \( A_k \in L^\infty(\Omega, \mathcal{F}_k, \mathbb{P}) \) be an \( \mathcal{F}_{k-1} \)-measurable random variable with a finite essential supremum. Define a new sequence of random variables in \( L^1(\Omega, \mathcal{F}, \mathbb{P}) \)

\[
X_n = \sum_{k=1}^n A_k U_k, \quad \forall n \in \mathbb{N}
\]

and \( X_0 = 0 \). Then, \( \{X_n, \mathcal{F}_n\}_{n \in \mathbb{N}_0} \) is a martingale. Let us assume that the random variables \( \{U_k\}_{k \in \mathbb{N}} \) are symmetrically distributed around zero. Note that \( X_n = X_{n-1} + A_n U_n \) where \( A_n \) is \( \mathcal{F}_{n-1} \)-measurable and \( U_n \) is independent of the \( \sigma \)-algebra \( \mathcal{F}_{n-1} \) (due to the independence of the random variables \( U_1, \ldots, U_n \)). It therefore follows that for every \( n \in \mathbb{N} \), given \( \mathcal{F}_{n-1} \), the random variable \( X_n \) is symmetrically distributed around its conditional expectation \( X_{n-1} \). Hence, the martingale \( \{X_n, \mathcal{F}_n\}_{n \in \mathbb{N}_0} \) is conditionally symmetric.

Example 7: In continuation to Example 6, let \( \{X_n, \mathcal{F}_n\}_{n \in \mathbb{N}_0} \) be a martingale, and define \( Y_0 = 0 \) and

\[
Y_n = \sum_{k=1}^n A_k (X_k - X_{k-1}), \quad \forall n \in \mathbb{N}.
\]

The sequence \( \{Y_n, \mathcal{F}_n\}_{n \in \mathbb{N}_0} \) is a martingale. If \( \{X_n, \mathcal{F}_n\}_{n \in \mathbb{N}_0} \) is a conditionally symmetric martingale then also the martingale \( \{Y_n, \mathcal{F}_n\}_{n \in \mathbb{N}_0} \) is conditionally symmetric (since \( Y_n = Y_{n-1} + A_n (X_n - X_{n-1}) \), and by assumption \( A_n \) is \( \mathcal{F}_{n-1} \)-measurable).

Example 8: In continuation to Example 6, let \( \{U_k\}_{k \in \mathbb{N}} \) be independent random variables with a symmetric distribution around their expected value, and also assume that \( \mathbb{E}(U_k) \leq 0 \) for every \( k \in \mathbb{N} \). Furthermore, let \( A_k \in L^\infty(\Omega, \mathcal{F}_{k-1}, \mathbb{P}) \), and assume that a.s. \( A_k \geq 0 \) for every \( k \in \mathbb{N} \). Let \( \{X_n, \mathcal{F}_n\}_{n \in \mathbb{N}_0} \) be a martingale as defined in Example 6. Note that \( X_n = X_{n-1} + A_n U_n \) where \( A_n \) is non-negative and \( \mathcal{F}_{n-1} \)-measurable, and \( U_n \) is independent of \( \mathcal{F}_{n-1} \) and symmetrically distributed around its average. This implies that \( \{X_n, \mathcal{F}_n\}_{n \in \mathbb{N}_0} \) is a conditionally symmetric supermartingale.

Example 9: In continuation to Examples 7 and 8, let \( \{X_n, \mathcal{F}_n\}_{n \in \mathbb{N}_0} \) be a conditionally symmetric supermartingale. Define \( \{Y_n\}_{n \in \mathbb{N}_0} \) as in Example 7 where \( A_k \) is non-negative a.s. and \( \mathcal{F}_{k-1} \)-measurable for every \( k \in \mathbb{N} \). Then \( \{Y_n, \mathcal{F}_n\}_{n \in \mathbb{N}_0} \) is a conditionally symmetric supermartingale.

Example 10: Consider a standard Brownian motion \( (W_t)_{t \geq 0} \). Define, for some \( T > 0 \), the discrete-time process

\[
X_n = W_{nT}, \quad \mathcal{F}_n = \sigma(\{W_t\}_{0 \leq t \leq nT}), \quad \forall n \in \mathbb{N}_0.
\]
The increments of \((W_t)_{t \geq 0}\) over time intervals \([t_{k-1}, t_k]\) are statistically independent if these intervals do not overlap (except of their endpoints), and they are Gaussian distributed with a zero mean and variance \(t_k - t_{k-1}\). The random variable \(\xi_n \triangleq X_n - X_{n-1}\) is therefore statistically independent of \(\mathcal{F}_{n-1}\), and it is Gaussian distributed with a zero mean and variance \(T\). The martingale \(\{X_n, \mathcal{F}_n\}_{n \in \mathbb{N}_0}\) is therefore conditionally symmetric.

After motivating this discussion with some explicit constructions of discrete-time conditionally symmetric martingales, we introduce a new deviation inequality for this sub-class of martingales, and then show how its derivation follows from the martingale approach that was used earlier for the derivation of Theorem 5. The new deviation inequality for the considered sub-class of discrete-time martingales with bounded increments gets the following form:

**Theorem 6:** Let \(\{X_k, \mathcal{F}_k\}_{k \in \mathbb{N}_0}\) be a discrete-time real-valued and conditionally symmetric martingale. Assume that, for some fixed numbers \(d, \sigma > 0\), the following two requirements are satisfied a.s.

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

for every \(k \in \mathbb{N}\). Then, for every \(\alpha \geq 0\) and \(n \in \mathbb{N}\),

\[
P \left( \max_{1 \leq k \leq n} |X_k - X_0| \geq \alpha n \right) \leq 2 \exp\left(-nE(\gamma, \delta)\right)
\]

where

\[
\gamma \triangleq \frac{\sigma^2}{d^2}, \quad \delta \triangleq \frac{\alpha}{d}
\]

and for \(\gamma \in (0, 1)\) and \(\delta \in [0, 1)\)

\[
E(\gamma, \delta) \triangleq \delta x - \ln \left(1 + \gamma [\cosh(x) - 1]\right)
\]

\[
x \triangleq \ln \left(\frac{\delta(1 - \gamma) + \sqrt{\delta^2(1 - \gamma)^2 + \gamma^2(1 - \delta^2)}}{\gamma(1 - \delta)}\right).
\]

If \(\delta > 1\), then the probability on the left-hand side of (51) is zero (so \(E(\gamma, \delta) \triangleq +\infty\)), and \(E(\gamma, 1) = \ln\left(\frac{\gamma}{\gamma}\right)\). Furthermore, the exponent \(E(\gamma, \delta)\) is asymptotically optimal in the sense that there exists a conditionally symmetric martingale, satisfying the conditions in (50) a.s., that attains this exponent in the limit where \(n \to \infty\).

**Remark 8:** From the above conditions, without any loss of generality, \(\sigma^2 \leq d^2\) and therefore \(\gamma \in (0, 1]\). This implies that Theorem 6 characterizes the exponent \(E(\gamma, \delta)\) for all values of \(\gamma\) and \(\delta\).

**Corollary 2:** Let \(\{U_k\}_{k=1}^\infty \in L^2(\Omega, \mathcal{F}, \mathbb{P})\) be i.i.d. and bounded random variables with a symmetric distribution around their mean value. Assume that \(|U_1 - \mathbb{E}[U_1]| \leq d\) a.s. for some \(d > 0\), and \(\text{Var}(U_1) \leq \gamma d^2\) for some \(\gamma \in [0, 1]\). Let \(\{S_n\}\) designate the sequence of partial sums, i.e., \(S_n \triangleq \sum_{k=1}^n U_k\) for every \(n \in \mathbb{N}\). Then, for every \(\alpha \geq 0\),

\[
P \left( \max_{1 \leq k \leq n} |S_k - k \mathbb{E}(U_1)| \geq \alpha n \right) \leq 2 \exp\left(-nE(\gamma, \delta)\right), \quad \forall n \in \mathbb{N}
\]

where \(\delta \triangleq \frac{\alpha}{d}\), and \(E(\gamma, \delta)\) is introduced in (53) and (54).

**Remark 9:** Theorem 6 should be compared to Theorem 5 (see [52, Theorem 6.1] or [17, Corollary 2.4.7]), which does not require the conditional symmetry property. The two exponents in Theorems 6 and 5 are both discontinuous at \(\delta = 1\). This is consistent with the assumption of the bounded jumps that implies that \(P(|X_n - X_0| \geq nd\delta)\) is equal to zero if \(\delta > 1\).

If \(\delta \to 1^-\) then, from (53) and (54), for every \(\gamma \in (0, 1]\),

\[
\lim_{\delta \to 1^-} E(\gamma, \delta) = \lim_{x \to \infty} \left[x - \ln\left(1 + \gamma \cosh(x) - 1\right)\right] = \ln\left(\frac{\gamma}{\gamma}\right).
\]

On the other hand, the right limit at \(\delta = 1\) is infinity since \(E(\gamma, \delta) = +\infty\) for every \(\delta > 1\). The same discontinuity also exists for the exponent in Theorem 5 where the right limit at \(\delta = 1\) is infinity, and the left limit is equal to

\[
\lim_{\delta \to 1^-} D\left(\frac{\delta + \gamma}{1 + \gamma} \middle| \frac{\gamma}{1 + \gamma}\right) = \ln \left(1 + \frac{1}{\gamma}\right)
\]
where the last equality follows from (35). A comparison of the limits in (56) and (57) is consistent with the improvement that is obtained in Theorem 6 as compared to Theorem 5 due to the additional assumption of the conditional symmetry that is relevant if $\gamma \in (0, 1)$. It can be verified that the two exponents coincide if $\gamma = 1$ (which is equivalent to removing the constraint on the conditional variance), and their common value is equal to
\[
\gamma = \frac{2}{\ln(2)}.
\]

Proof: Since $h$ is convex and $\text{supp}(X) = [-d, d]$, then a.s. $h(X^2) \leq h(0) + \frac{1}{2} (h(d^2) - h(0))$. Taking expectations on both sides gives (59), which holds with equality for the symmetric distribution in (60).

Corollary 3: If $X$ is a random variable that satisfies the three requirements in Lemma 3 then, for every $\lambda \in \mathbb{R}$,
\[
\mathbb{E}[\exp(\lambda X)] \leq 1 + \gamma \cosh(\lambda d) - 1
\]
and (61) holds with equality for the symmetric distribution in Lemma 3, independently of the value of $\lambda$.

Proof: For every $\lambda \in \mathbb{R}$, due to the symmetric distribution of $X$, $\mathbb{E}[\exp(\lambda X)] = \mathbb{E}[-\exp(\lambda X)]$. The claim now follows from Lemma 3 since, for every $x \in \mathbb{R}$, $\cosh(\lambda x) = h(x^2)$ where $h(x) = \sum_{n=0}^{\infty} \frac{\lambda^{2n} |x|^{2n}}{(2n)!}$ is a convex function ($h$ is convex since it is a linear combination, with non-negative coefficients, of convex functions), and $h(d^2) = \cosh(\lambda d) \geq 1 = h(0)$.

We continue with the proof of Theorem 6. Under the assumption of this theorem, for every $k \in \mathbb{N}$, the random variable $\xi_k \triangleq X_k - X_{k-1}$ satisfies a.s. $\mathbb{E}[\xi_k | \mathcal{F}_{k-1}] = 0$ and $\mathbb{E}[(\xi_k)^2 | \mathcal{F}_{k-1}] \leq \sigma^2$. Applying Corollary 3 for the conditional law of $\xi_k$ given $\mathcal{F}_{k-1}$, it follows that for every $k \in \mathbb{N}$ and $t \in \mathbb{R}$
\[
\mathbb{E}[\exp(t \xi_k) | \mathcal{F}_{k-1}] \leq 1 + \gamma \cosh(\lambda d) - 1
\]
holds a.s., and therefore it follows from (4) and (62) that for every $t \in \mathbb{R}$
\[
\mathbb{E}[\exp(t \sum_{k=1}^{n} \xi_k)] \leq \left(1 + \gamma \cosh(\lambda d) - 1\right)^n.
\]

By applying the maximal inequality for submartingales, then for every $\alpha \geq 0$ and $n \in \mathbb{N}$
\[
\mathbb{P} \left( \max_{1 \leq k \leq n} (X_k - X_0) \geq \alpha n \right) \\
= \mathbb{P} \left( \max_{1 \leq k \leq n} \exp(t(X_k - X_0)) \geq \exp(\alpha n) \right) \\
\leq \exp(-\alpha t) \mathbb{E} \left[ \exp(t(X_n - X_0)) \right] \\
= \exp(-\alpha t) \mathbb{E} \left[ \exp \left( \sum_{k=1}^{n} \xi_k \right) \right]
\]

(64)
Therefore, from (64), for every $t \geq 0$,
\[
\mathbb{P}\left(\max_{1 \leq k \leq n} (X_k - X_0) \geq \alpha n\right) \leq \exp\left(-\alpha nt\left(1 + \gamma \left[\cosh(td) - 1\right]\right)^n\right).\tag{65}
\]

From (52) and a replacement of $td$ with $x$, then for an arbitrary $\alpha \geq 0$ and $n \in \mathbb{N}$
\[
\mathbb{P}\left(\max_{1 \leq k \leq n} (X_k - X_0) \geq \alpha n\right) \leq \inf_{x \geq 0} \left\{\exp\left(-n\left[\delta x - \ln\left(1 + \gamma \left[\cosh(x) - 1\right]\right)\right]\right)\right\}.	ag{66}
\]

Applying (66) to the martingale $\{-E[X_k] \mid k \in \mathbb{N}_0\}$ gives the same bound on $\mathbb{P}(\min_{1 \leq k \leq n} (X_k - X_0) \leq -\alpha n)$ for an arbitrary $\alpha \geq 0$. The union bound implies that
\[
\mathbb{P}\left(\max_{1 \leq k \leq n} |X_k - X_0| \geq \alpha n\right) \leq \mathbb{P}\left(\max_{1 \leq k \leq n} (X_k - X_0) \geq \alpha n\right) + \mathbb{P}\left(\min_{1 \leq k \leq n} (X_k - X_0) \leq -\alpha n\right).	ag{67}
\]

This doubles the bound on the right-hand side of (66), thus proving the exponential bound in Theorem 6.

**Proof for the asymptotic optimality of the exponents in Theorems 6 and 5:** In the following, we show that under the conditions of Theorem 6, the exponent $E(\gamma, \delta)$ in (53) and (54) is asymptotically optimal. To show this, let $d > 0$ and $\gamma \in (0, 1]$, and let $U_1, U_2, \ldots$ be i.i.d. random variables whose probability distribution is given by
\[
\mathbb{P}(U_i = d) = \mathbb{P}(U_i = -d) = \frac{\gamma}{2}, \quad \mathbb{P}(U_i = 0) = 1 - \gamma, \quad \forall i \in \mathbb{N}.	ag{68}
\]

Consider the particular case of the conditionally symmetric martingale $\{X_n, \mathcal{F}_n\}_{n \in \mathbb{N}_0}$ in Example 6 (see Section IV-C1) where $X_n \triangleq \sum_{i=1}^{n} U_i$ for $n \in \mathbb{N}$, and $X_0 \triangleq 0$. It follows that $|X_n - X_{n-1}| \leq d$ and $\text{Var}(X_n|\mathcal{F}_{n-1}) = \gamma d^2$ a.s. for every $n \in \mathbb{N}$. From Cramér’s theorem in $\mathbb{R}$, for every $\alpha \geq \mathbb{E}[U_1] = 0$,
\[
\lim_{n \to \infty} \frac{1}{n} \ln \mathbb{P}(X_n - X_0 \geq \alpha n) = \lim_{n \to \infty} \frac{1}{n} \ln \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} U_i \geq \alpha\right) = -I(\alpha)
\]
where the rate function is given by
\[
I(\alpha) = \sup_{t \geq 0} \left\{t \alpha - \ln \mathbb{E}[\exp(tU_1)]\right\}
\]
(see, e.g., [17, Theorem 2.2.3] and [17, Lemma 2.2.5(b)] for the restriction of the supremum to the interval $[0, \infty)$). From (68) and (122), for every $\alpha \geq 0$,
\[
I(\alpha) = \sup_{t \geq 0} \left\{t \alpha - \ln\left(1 + \gamma \left[\cosh(td) - 1\right]\right)\right\}
\]
but it is equivalent to the optimized exponent on the right-hand side of (65), giving the exponent of the bound in Theorem 6. Hence, $I(\alpha) = E(\gamma, \delta)$ in (53) and (54). This proves that the exponent of the bound in Theorem 6 is indeed asymptotically optimal in the sense that there exists a discrete-time, real-valued and conditionally symmetric martingale, satisfying the conditions in (50) a.s., that attains this exponent in the limit where $n \to \infty$. The proof for the asymptotic optimality of the exponent in Theorem 5 (see the right-hand side of (33)) is similar to the proof for Theorem 6, except that the i.i.d. random variables $U_1, U_2, \ldots$ are now distributed as follows:
\[
\mathbb{P}(U_i = d) = \frac{\gamma}{1 + \gamma}, \quad \mathbb{P}(U_i = -d) = \frac{1}{1 + \gamma}, \quad \forall i \in \mathbb{N}
\]
and, as before, the martingale $\{X_n, \mathcal{F}_n\}_{n \in \mathbb{N}_0}$ is defined by $X_n = \sum_{i=1}^{n} U_i$ and $\mathcal{F}_n = \sigma(U_1, \ldots, U_n)$ for every $n \in \mathbb{N}$ with $X_0 = 0$ and $\mathcal{F}_0 = \{\emptyset, \Omega\}$ (in this case, it is not a conditionally symmetric martingale unless $\gamma = 1$).

Theorem 6 provides an improvement over the bound in Theorem 5 for conditionally symmetric martingales with bounded jumps. The bounds in Theorems 6 and 5 depend on the conditional variance of the martingale, but they do not take into consideration conditional moments of higher orders. The following bound generalizes the bound in Theorem 6, but it does not admit in general a closed-form expression.
These events correspond to small deviations. This is in contrast to events of the form \( \Pr(\{X_n \geq n^{1/2} \delta \}) \) for some \( \delta > 0 \) and positive numbers \( \{\mu_2, \mu_4, \ldots, \mu_m\} \). Then, for every \( \alpha \geq 0 \) and \( n \in \mathbb{N} \),

\[
\Pr\left( \max_{1 \leq k \leq n} |X_k - X_0| \geq \alpha n \right) \leq 2 \min_{x \geq 0} \left[ 1 + \sum_{l=1}^{m-1} \frac{\gamma_l x^{2l}}{(2l)!} + \gamma_m (\cosh(x) - 1) \right]^{n/2} \tag{71}
\]

where

\[
\delta \triangleq \frac{\alpha}{d}, \quad \gamma_{2l} \triangleq \frac{\mu_{2l}}{d^{2l}}, \quad \forall l \in \left\{1, \ldots, \frac{m}{2} \right\}. \tag{72}
\]

**Proof:** The starting point of the proof of Theorem 7 relies on (64) and (4). For every \( k \in \mathbb{N} \) and \( t \in \mathbb{R} \), since \( \Pr(\{\xi^{2l-1}_k \mid \mathcal{F}_{k-1}\}) = 0 \) for every \( l \in \mathbb{N} \) (due to the conditionally symmetry property of the martingale),

\[
\Pr\left[ \exp(t \xi_k) \mid \mathcal{F}_{k-1} \right] = 1 + \sum_{l=1}^{\infty} \frac{t^{2l} \Pr(\{\xi^{2l}_k \mid \mathcal{F}_{k-1}\})}{(2l)!} + \sum_{l=1}^{\infty} \frac{t^{2l} \Pr(\{\xi^{2l}_k \mid \mathcal{F}_{k-1}\})}{(2l)!} 
\]

\[
\leq 1 + \sum_{l=1}^{\infty} \frac{t^{2l} \gamma_{2l}}{(2l)!} + \sum_{l=1}^{\infty} \frac{t^{2l} \gamma_m}{(2l)!} 
\]

\[
= 1 + \sum_{l=1}^{\infty} \frac{t^{2l} (\gamma_{2l} - \gamma_m)}{(2l)!} + \gamma_m (\cosh(td) - 1) \tag{73}
\]

where the inequality above holds since \( \sum_{l=1}^{\infty} \frac{\gamma_{2l}}{(2l)!} \leq 1 \) a.s., so that \( 0 \leq \ldots \leq \gamma_m \leq \ldots \leq \gamma_4 \leq \gamma_2 \leq 1 \), and the last equality in (73) holds since \( \cosh(x) = \sum_{n=0}^{\infty} \frac{x^{2n}}{(2n)!} \) for every \( x \in \mathbb{R} \). Therefore, from (4),

\[
\Pr\left( \exp\left( t \sum_{k=1}^{n} \xi_k \right) \right) \leq \left( 1 + \sum_{l=1}^{\infty} \frac{t^{2l} \gamma_{2l} - \gamma_m}{(2l)!} + \gamma_m \cosh(td) - 1 \right)^n \tag{74}
\]

for an arbitrary \( t \in \mathbb{R} \). The inequality then follows from (64). This completes the proof of Theorem 7.

**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 1:** Let \( \{X_k, \mathcal{F}_k\} \) be a discrete-parameter real-valued martingale. Then, Theorem 5 implies that for every \( \alpha \geq 0 \)

\[
\Pr(|X_n - X_0| \geq \alpha \sqrt{n}) \leq 2 \exp\left( -\frac{\alpha^2}{2} \right) \left( 1 + O(n^{-1/2}) \right). \tag{75}
\]

**Proof:** See Appendix A.

**Remark 10:** From Proposition 1, the upper bound on \( \Pr(|X_n - X_0| \geq \alpha \sqrt{n}) \) (for an arbitrary \( \alpha \geq 0 \)) improves the exponent of Azuma’s inequality by a factor of \( \frac{\sqrt{2}}{\gamma} \).
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., [13, Chapter 2] or [14, 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 5 to sub and super martingales gives the following inequality:

**Corollary 4:** 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.

$$X_k - \mathbb{E}[X_k | \mathcal{F}_{k-1}] \leq d,$$

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

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

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

(76)

where $\gamma$ and $\delta$ are defined as in (52), and the divergence $D(p||q)$ is introduced in (35). Alternatively, if $\{X_k, \mathcal{F}_k\}_{k=0}^\infty$ is a sub-martingale, the same upper bound in (76) 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 corollary is similar to the proof of Theorem 5. 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 $\mathbb{P}(X_n - X_0 \geq \alpha n) \leq \mathbb{P}(\sum_{k=1}^n \xi_k \geq \alpha n)$ where a.s. $\xi_k \leq d$, $\mathbb{E}[\xi_k | \mathcal{F}_{k-1}] = 0$, and $\text{Var}(\xi_k | \mathcal{F}_{k-1}) \leq \sigma^2$. The continuation of the proof coincides with the proof of Theorem 5 (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. Freedman’s Inequality and a Refined Version

We consider in the following a different type of exponential inequalities for discrete-time martingales with bounded jumps, which is a classical inequality that dates back to Freedman [24]. Freedman’s inequality is refined in the following to conditionally symmetric martingales with bounded jumps (see [69]). Furthermore, these two inequalities are specialized to two concentration inequalities for sums of independent and bounded random variables.

**Theorem 8:** Let $\{X_n, \mathcal{F}_n\}_{n \in \mathbb{N}_0}$ be a discrete-time real-valued and conditionally symmetric martingale. Assume that there exists a fixed number $d > 0$ such that $\xi_k \triangleq X_k - X_{k-1} \leq d$ a.s. for every $k \in \mathbb{N}$. Let

$$Q_n \triangleq \sum_{k=1}^n \mathbb{E}[\xi_k^2 | \mathcal{F}_{k-1}]$$

(77)

with $Q_0 \triangleq 0$, be the predictable quadratic variation of the martingale up to time $n$. Then, for every $z, r > 0$,

$$\mathbb{P}\left(\max_{1 \leq k \leq n} (X_k - X_0) \geq z, Q_n \leq r \text{ for some } n \in \mathbb{N}\right) \leq \exp\left(-\frac{z^2}{2r} \cdot C\left(\frac{zd}{r}\right)\right)$$

(78)

where

$$C(u) \triangleq \frac{2[u \sinh^{-1}(u) - \sqrt{1 + u^2} + 1]}{u^2}, \quad \forall u > 0.$$  

(79)

Theorem 8 should be compared to Freedman’s inequality in [24, Theorem 1.6] (see also [17, Exercise 2.4.21(b)]) that was stated without the requirement for the conditional symmetry of the martingale. It provides the following result:
Theorem 9: Let \( \{X_n, \mathcal{F}_n\}_{n \in \mathbb{N}_0} \) be a discrete-time real-valued martingale. Assume that there exists a fixed number \( d > 0 \) such that \( \xi_k \overset{\text{def}}{=} X_k - X_{k-1} \leq d \) a.s. for every \( k \in \mathbb{N} \). Then, for every \( z, r > 0 \),
\[
\mathbb{P} \left( \max_{1 \leq k \leq n} (X_k - X_0) \geq z, Q_n \leq r \text{ for some } n \in \mathbb{N} \right) \leq \exp \left( -\frac{z^2}{2r} \cdot B \left( \frac{zd}{r} \right) \right)
\]
where
\[
B(u) \overset{\text{def}}{=} \frac{2[(1 + u) \ln(1 + u) - u]}{u^2}, \quad \forall u > 0.
\]

The proof of [24, Theorem 1.6] is modified in the following by using Bennett’s inequality for the derivation of the original bound in Theorem 9 (without the conditional symmetry requirement). Furthermore, this modified proof serves to derive the improved bound in Theorem 8 under the conditional symmetry assumption of the martingale sequence.

We provide in the following a combined proof of Theorems 8 and 9.

Proof: The proof of Theorem 8 relies on the proof of Freedman’s inequality in Theorem 9, where the latter dates back to Freedman’s paper (see [24, Theorem 1.6], and also [17, Exercise 2.4.21(b)]). The original proof of Theorem 9 (see [24, Section 3]) is modified in a way that facilitates to realize how the bound can be improved for conditionally symmetric martingales with bounded jumps. This improvement is obtained via the refinement in (62) of Bennett’s inequality for conditionally symmetric distributions. Furthermore, the following revisited proof of Theorem 9 simplifies the derivation of the new and improved bound in Theorem 8 for the considered subclass of martingales.

Without any loss of generality, let’s assume that \( d = 1 \) (otherwise, \( \{X_k\} \) and \( z \) are divided by \( d \), and \( \{Q_k\} \) and \( r \) are divided by \( d^2 \); this normalization extends the bound to the case of an arbitrary \( d > 0 \)). Let \( S_n \overset{\text{def}}{=} X_n - X_0 \) for every \( n \in \mathbb{N}_0 \), then \( \{S_n, \mathcal{F}_n\}_{n \in \mathbb{N}_0} \) is a martingale with \( S_0 = 0 \). The proof starts by introducing two lemmas.

Lemma 4: Under the assumptions of Theorem 9, let
\[
U_n \overset{\text{def}}{=} \exp(\lambda S_n - \theta Q_n), \quad \forall n \in \{0, 1, \ldots\}
\]
where \( \lambda \geq 0 \) and \( \theta \geq e^\lambda - \lambda - 1 \) are arbitrary constants. Then, \( \{U_n, \mathcal{F}_n\}_{n \in \mathbb{N}_0} \) is a supermartingale.

Proof: \( U_n \) in (82) is \( \mathcal{F}_n \)-measurable (since \( Q_n \) in (77) is \( \mathcal{F}_{n-1} \)-measurable, where \( \mathcal{F}_{n-1} \subseteq \mathcal{F}_n \), and \( S_n \) is \( \mathcal{F}_n \)-measurable), \( Q_n \) and \( U_n \) are non-negative random variables, and \( S_n = \sum_{k=1}^n \xi_k \leq n \) a.s. (since \( \xi_k \leq 1 \) and \( S_0 = 0 \)). It therefore follows that \( 0 \leq U_n \leq e^{\lambda n} \) a.s. for \( \lambda, \theta \geq 0 \), so \( U_n \in L^1(\Omega, \mathcal{F}_n, \mathbb{P}) \). It is required to show that \( \mathbb{E}[U_n | \mathcal{F}_{n-1}] \leq U_{n-1} \) holds a.s. for every \( n \in \mathbb{N} \), under the above assumptions on the parameters \( \lambda \) and \( \theta \) in (82).

\[
\begin{align*}
\mathbb{E}[U_n | \mathcal{F}_{n-1}] \\
\overset{(a)}{=} \exp(-\theta Q_n) \mathbb{E}[\exp(\lambda S_{n-1}) | \mathcal{F}_{n-1}] \\
\overset{(b)}{=} \exp(\lambda S_{n-1}) \mathbb{E}[\exp(-\theta (Q_{n-1} + \mathbb{E}[\xi_n^2 | \mathcal{F}_{n-1}]) | \mathcal{F}_{n-1}] \\
\overset{(c)}{=} U_{n-1} \left( \frac{\mathbb{E}[\exp(\lambda S_n) | \mathcal{F}_{n-1}]}{\mathbb{E}[\exp(\theta \mathbb{E}[\xi_n^2 | \mathcal{F}_{n-1}])]} \right)
\end{align*}
\]

where (a) follows from (82) and because \( Q_n \) and \( S_{n-1} \) are \( \mathcal{F}_{n-1} \)-measurable and \( S_n = S_{n-1} + \xi_n \), (b) follows from (77), and (c) follows from (82).

A modification of the original proof of Lemma 4 (see [24, Section 3]) is suggested in the following, which then enables to improve the bound in Theorem 9 for real-valued, discrete-time, conditionally symmetric martingales with bounded jumps. This leads to the improved bound in Theorem 8 for the considered subclass of martingales.

Since by assumption \( \xi_n \leq 1 \) and \( \mathbb{E}[\xi_n | \mathcal{F}_{n-1}] = 0 \) a.s., then applying Bennett’s inequality in (40) to the conditional expectation of \( e^{\lambda \xi_n} \) given \( \mathcal{F}_{n-1} \) (recall that \( \lambda \geq 0 \) gives
\[
\mathbb{E}[\exp(\lambda \xi_n) | \mathcal{F}_{n-1}] \leq \frac{\exp(-\lambda \mathbb{E}[\xi_n | \mathcal{F}_{n-1}]) + \mathbb{E}[\xi_n^2 | \mathcal{F}_{n-1}] \exp(\lambda)}{1 + \mathbb{E}[\xi_n^2 | \mathcal{F}_{n-1}]}
\]
which therefore implies from (83) and the last inequality that
\[
\mathbb{E}[U_n | \mathcal{F}_{n-1}] \leq U_{n-1} \left( \frac{\exp \left( (\lambda + \theta) \mathbb{E}[\xi_n^2 | \mathcal{F}_{n-1}] \right)}{1 + \mathbb{E}[\xi_n^2 | \mathcal{F}_{n-1}]} + \frac{\mathbb{E}[\xi_n^2 | \mathcal{F}_{n-1}] \exp(\lambda - \theta \mathbb{E}[\xi_n^2 | \mathcal{F}_{n-1}])}{1 + \mathbb{E}[\xi_n^2 | \mathcal{F}_{n-1}]} \right).
\]
In order to prove that \( \mathbb{E}[U_n | \mathcal{F}_{n-1}] \leq U_{n-1} \) a.s., it is sufficient to prove that the second term on the right-hand side of (84) is a.s. less than or equal to 1. To this end, lets find the condition on \( \lambda, \theta \geq 0 \) such that for every \( \alpha \geq 0 \)

\[
\left( \frac{1}{1 + \alpha} \exp(-\alpha(\lambda + \theta)) \right) + \left( \frac{\alpha}{1 + \alpha} \right) \exp(\alpha \lambda - \alpha \theta) \leq 1 \tag{85}
\]

which then assures that the second term on the right-hand side of (84) is less than or equal to 1 a.s. as required.

**Lemma 5:** If \( \lambda \geq 0 \) and \( \theta \geq \exp(\lambda) - \lambda - 1 \) then the condition in (85) is satisfied for every \( \alpha \geq 0 \).

**Proof:** This claim follows by calculus, showing that the function

\[
g(\alpha) = (1 + \alpha) \exp(\alpha \theta) - \alpha \exp(\lambda) - \exp(-\alpha \lambda), \quad \forall \alpha \geq 0
\]

is non-negative on \( \mathbb{R}_+ \) if \( \lambda \geq 0 \) and \( \theta \geq \exp(\lambda) - \lambda - 1 \).

From (84) and Lemma 5, it follows that \( \{U_n, \mathcal{F}_n\}_{n \in \mathbb{N}_0} \) is a supermartingale if \( \lambda \geq 0 \) and \( \theta \geq \exp(\lambda) - \lambda - 1 \). This completes the proof of Lemma 4.

At this point, we start to discuss in parallel the derivation of the tightened bound in Theorem 8 for conditionally symmetric martingales. As before, it is assumed without any loss of generality that \( d = 1 \).

**Lemma 6:** Under the additional assumption of the conditional symmetry in Theorem 8, then \( \{U_n, \mathcal{F}_n\}_{n \in \mathbb{N}_0} \) in (82) is a supermartingale if \( \lambda \geq 0 \) and \( \theta \geq \cosh(\lambda) - 1 \) are arbitrary constants.

**Proof:** By assumption \( \xi_n = S_n - S_{n-1} \leq 1 \) a.s., and \( \xi_n \) is conditionally symmetric around zero, given \( \mathcal{F}_{n-1} \), for every \( n \in \mathbb{N} \). By applying Corollary 3 to the conditional expectation of \( \exp(\lambda \xi_n) \) given \( \mathcal{F}_{n-1} \), for every \( \lambda \geq 0 \),

\[
\mathbb{E}[\exp(\lambda \xi_n) | \mathcal{F}_{n-1}] \leq 1 + \mathbb{E}[\xi_n^2 | \mathcal{F}_{n-1}] \left( \cosh(\lambda) - 1 \right). \tag{86}
\]

Hence, combining (83) and (86) gives

\[
\mathbb{E}[U_n | \mathcal{F}_{n-1}] \leq U_{n-1} \left( \frac{1 + \mathbb{E}[\xi_n^2 | \mathcal{F}_{n-1}] \left( \cosh(\lambda) - 1 \right)}{\exp(\theta \mathbb{E}[\xi_n^2 | \mathcal{F}_{n-1}])} \right). \tag{87}
\]

Let \( \lambda \geq 0 \). Since \( \mathbb{E}[\xi_n^2 | \mathcal{F}_{n-1}] \geq 0 \) a.s. then in order to ensure that \( \{U_n, \mathcal{F}_n\}_{n \in \mathbb{N}_0} \) forms a supermartingale, it is sufficient (based on (87)) that the following condition holds:

\[
\frac{1 + \alpha(\cosh(\lambda) - 1)}{\exp(\theta \alpha)} \leq 1, \quad \forall \alpha \geq 0. \tag{88}
\]

Calculus shows that, for \( \lambda \geq 0 \), the condition in (88) is satisfied if and only if

\[
\theta \geq \cosh(\lambda) - 1 \triangleq \theta_{\min}(\lambda). \tag{89}
\]

From (87), \( \{U_n, \mathcal{F}_n\}_{n \in \mathbb{N}_0} \) is a supermartingale if \( \lambda \geq 0 \) and \( \theta \geq \theta_{\min}(\lambda) \). This proves Lemma 6.

Hence, due to the assumption of the conditional symmetry of the martingale in Theorem 8, the set of parameters for which \( \{U_n, \mathcal{F}_n\} \) is a supermartingale was extended. This follows from a comparison of Lemma 4 and 6 where indeed \( \exp(\lambda) - 1 - \lambda \geq \theta_{\min}(\lambda) \geq 0 \) for every \( \lambda \geq 0 \).

Let \( z, r > 0, \lambda \geq 0 \) and either \( \theta \geq \cosh(\lambda) - 1 \) or \( \theta \geq \exp(\lambda) - \lambda - 1 \) with or without assuming the conditional symmetry property, respectively (see Lemma 4 and 6). In the following, we rely on Doob’s sampling theorem. To this end, let \( M \in \mathbb{N} \), and define two stopping times adapted to \( \{\mathcal{F}_n\} \). The first stopping time is \( \alpha = 0 \), and the second stopping time \( \beta \) is the minimal value of \( n \in \{0, \ldots, M\} \) (if any) such that \( S_n \geq z \) and \( Q_n \leq r \) (note that \( S_n \) is \( \mathcal{F}_n \)-measurable and \( Q_n \) is \( \mathcal{F}_{n-1} \)-measurable, so the event \( \{\beta \leq n\} \) is \( \mathcal{F}_n \)-measurable); if such a value of \( n \) does not exist, let \( \beta \triangleq M \). Hence \( \alpha \leq \beta \) are two bounded stopping times. From Lemma 4 or 6, \( \{U_n, \mathcal{F}_n\}_{n \in \mathbb{N}_0} \) is a supermartingale for the corresponding set of parameters \( \lambda \) and \( \theta \), and from Doob’s sampling theorem

\[
\mathbb{E}[U_\beta] \leq \mathbb{E}[U_0] = 1 \tag{90}
\]
\( (S_0 = Q_0 = 0, \text{ so from (82), } U_0 = 1 \text{ a.s.}). \) Hence, it implies the following chain of inequalities:

\[
\begin{align*}
\mathbb{P}(\exists n \in \mathbb{N} : S_n \geq z, Q_n \leq r) & \equiv \mathbb{P}(S_\beta \geq z, Q_\beta \leq r) \\
& \leq \mathbb{P}(\lambda S_\beta - \theta Q_\beta \geq \lambda z - \theta r) \\
& \leq \mathbb{E}[\exp(\lambda S_\beta - \theta Q_\beta)] \\
& \leq \frac{\mathbb{E}[U_\beta]}{\exp(\lambda z - \theta r)} \\
& \leq \exp(-\lambda z - \theta r)) \quad (91)
\end{align*}
\]

where equality (a) follows from the definition of the stopping time \( \beta \in \{0, \ldots, M\} \), (b) holds since \( \lambda, \theta \geq 0 \), (c) follows from Chernoff's bound, (d) follows from the definition in (82), and finally (e) follows from (90). Since (91) holds for every \( M \in \mathbb{N} \), then from the continuity theorem for non-decreasing events and (91)

\[
\begin{align*}
\mathbb{P}(\exists n \in \mathbb{N} : S_n \geq z, Q_n \leq r) & = \lim_{M \to \infty} \mathbb{P}(\exists n \leq M : S_n \geq z, Q_n \leq r) \\
& \leq \exp(-\lambda z - \theta r)) \quad (92)
\end{align*}
\]

The choice of the non-negative parameter \( \theta \) as the minimal value for which (92) is valid provides the tightest bound within this form. Hence, without assuming the conditional symmetry property for the martingale \( \{X_n, \mathcal{F}_n\} \), let (see Lemma 4) \( \theta = \exp(\lambda) - \lambda - 1 \). This gives that for every \( z, r > 0 \),

\[
\mathbb{P}(\exists n \in \mathbb{N} : S_n \geq z, Q_n \leq r) \leq \exp\left(-\left[\lambda z - (\exp(\lambda) - \lambda - 1)r\right]\right), \quad \forall \lambda \geq 0.
\]

The minimization w.r.t. \( \lambda \) gives that \( \lambda = \ln(1 + \frac{z}{r}) \), and its substitution in the bound yields that

\[
\begin{align*}
\mathbb{P}(\exists n \in \mathbb{N} : S_n \geq z, Q_n \leq r) & \leq \exp\left(-\frac{z^2}{2r} \cdot B\left(\frac{z}{r}\right)\right) \\
& \quad (93)
\end{align*}
\]

where the function \( B \) is introduced in (81).

Furthermore, under the assumption that the martingale \( \{X_n, \mathcal{F}_n\}_{n \in \mathbb{N}_0} \) is conditionally symmetric, let \( \theta = \theta_{\min}(\lambda) \) (see Lemma 6) for obtaining the tightest bound in (92) for a fixed \( \lambda \geq 0 \). This gives the inequality

\[
\mathbb{P}(\exists n \in \mathbb{N} : S_n \geq z, Q_n \leq r) \leq \exp\left(-\left[\lambda z - r\theta_{\min}(\lambda)\right]\right), \quad \forall \lambda \geq 0.
\]

The optimized \( \lambda \) is equal to \( \lambda = \sinh^{-1}\left(\frac{z}{r}\right) \). Its substitution in (89) gives that \( \theta_{\min}(\lambda) = \sqrt{1 + \frac{z^2}{r^2}} - 1 \), and

\[
\begin{align*}
\mathbb{P}(\exists n \in \mathbb{N} : S_n \geq z, Q_n \leq r) & \leq \exp\left(-\frac{z^2}{2r} \cdot C\left(\frac{z}{r}\right)\right) \\
& \quad (94)
\end{align*}
\]

where the function \( C \) is introduced in (79).

Finally, the proof of Theorems 8 and 9 is completed by showing that the following equality holds:

\[
A \triangleq \{\exists n \in \mathbb{N} : S_n \geq z, Q_n \leq r\} = \{\exists n \in \mathbb{N} : \max_{1 \leq k \leq n} S_k \geq z, Q_n \leq r\} \triangleq B.
\]

Clearly \( A \subseteq B \), so one needs to show that \( B \subseteq A \). To this end, assume that event \( B \) is satisfied. Then, there exists some \( n \in \mathbb{N} \) and \( k \in \{1, \ldots, n\} \) such that \( S_k \geq z \) and \( Q_n \leq r \). Since the predictable quadratic variation process \( \{Q_n\}_{n \in \mathbb{N}_0} \) in (77) is monotonic non-decreasing, then it implies that \( S_k \geq z \) and \( Q_k \leq r \); therefore, event \( A \) is also satisfied and \( B \subseteq A \). The combination of (94) and (95) completes the proof of Theorem 8, and respectively the combination of (93) and (95) completes the proof of Theorem 9.

\[ \blacksquare \]
Let \( \phi \) where \( \phi \) where end, the following lower bound on \( \phi \) that are symmetrically distributed around zero. This leads to the following result:

**Corollary 5:** Let \( \{U_i\}_{i=1}^n \) be i.i.d. and bounded random variables such that \( \mathbb{E}[U_i] = 0, \mathbb{E}[U_i^2] = \sigma^2, \) and \( |U_i| \leq d \) a.s. for some constant \( d > 0 \). Then, the following inequality holds:

\[
\mathbb{P} \left( \left| \sum_{i=1}^n U_i \right| \geq \alpha \right) \leq 2 \exp \left( -\frac{n\sigma^2}{d^2} \cdot \phi_1 \left( \frac{\alpha d}{n\sigma^2} \right) \right), \quad \forall \alpha > 0
\]  

(96)

where \( \phi_1(x) \triangleq (1 + x) \ln(1 + x) - x \) for every \( x > 0 \). Furthermore, if the i.i.d. and bounded random variables \( \{U_i\}_{i=1}^n \) have a symmetric distribution around zero, then the bound in (96) can be improved to

\[
\mathbb{P} \left( \left| \sum_{i=1}^n U_i \right| \geq \alpha \right) \leq 2 \exp \left( -\frac{n\sigma^2}{d^2} \cdot \phi_2 \left( \frac{\alpha d}{n\sigma^2} \right) \right), \quad \forall \alpha > 0
\]

(97)

where \( \phi_2(x) \triangleq x \sinh^{-1}(x) - \sqrt{1 + x^2} + 1 \) for every \( x > 0 \).

**Proof:** Inequality (96) follows from Freedman’s inequality in Theorem 9, and inequality (97) follows from the refinement of Freedman’s inequality for conditionally symmetric martingales in Theorem 8. These two theorems are applied here to the martingale sequence \( \{X_k, \mathcal{F}_k\}_{k=0}^n \) where \( X_k = \sum_{i=1}^k U_i \) and \( \mathcal{F}_k = \sigma(U_1, \ldots, U_k) \) for every \( k \in \{1, \ldots, n\} \), and \( X_0 = 0, \mathcal{F}_0 = \{\emptyset, \Omega\} \). The corresponding predictable quadratic variation of the martingale up to time \( n \) for this special case of a sum of i.i.d. random variables is \( Q_n = \sum_{i=1}^n \mathbb{E}[U_i^2] = n\sigma^2 \). The result now follows by taking \( z = n\sigma^2 \) in inequalities (78) and (80) (with the related functions that are introduced in (81) and (79), respectively). Note that the same bound holds for the two one-sided tail inequalities, giving the factor 2 on the right-hand sides of (96) and (97).

**Remark 11:** Bennett’s concentration inequality in (96) can be loosened to obtain Bernstein’s inequality. To this end, the following lower bound on \( \phi_1 \) is used:

\[
\phi_1(x) \geq \frac{x^2}{2 + \frac{2x}{3}}, \quad \forall x > 0
\]

This gives the inequality

\[
\mathbb{P} \left( \left| \sum_{i=1}^n U_i \right| \geq \alpha \right) \leq 2 \exp \left( -\frac{\alpha^2}{2n\sigma^2 + \frac{2\alpha d}{3}} \right), \quad \forall \alpha > 0
\]

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

A. Relation between the Martingale Central Limit Theorem (CLT) and Proposition 1

In this subsection, we discuss the relation between the martingale CLT and the concentration inequalities for discrete-parameter martingales in Proposition 1.

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

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

Let

\[
S_n = \sum_{k=1}^n Y_k, \quad \forall n \in \mathbb{N}
\]
and \( S_0 = 0 \), then \( \{S_k, F_k\}_{k=0}^\infty \) 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

\[ \sigma^2 \triangleq \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} \mathbb{E} [Y_k^2 | F_{k-1}] \]

exists in probability and is positive. The martingale CLT asserts that, under the above conditions, \( \frac{S_n}{\sqrt{n}} \) converges in distribution (i.e., weakly converges) to the Gaussian distribution \( N(0, \sigma^2) \). It is denoted by \( \frac{S_n}{\sqrt{n}} \Rightarrow N(0, \sigma^2) \).

Let \( \{X_k, F_k\}_{k=0}^\infty \) 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 \( \{Y_k, F_k\}_{k=0}^\infty \) 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 5 and its subsequences, for every \( k \in \mathbb{N} \), one gets a.s. that

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

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

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

and therefore, for every \( \alpha \geq 0 \),

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

where the \( Q \) function is introduced in (206).

Based on the notation in (52), the equality \( \frac{\alpha}{\sigma} = \frac{\delta}{\sqrt{\gamma}} \) holds, and

\[ \lim_{n \to \infty} \mathbb{P} (|X_n - X_0| \geq \alpha \sqrt{n}) = 2 Q \left( \frac{\delta}{\sqrt{\gamma}} \right) . \tag{98} \]

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 1 (see (75) 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 1 (see Appendix A) 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 (12), then it follows from (98) that the exponent in the concentration inequality (75) (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 5

In this subsection, we discuss the relation between the law of the iterated logarithm (LIL) and Theorem 5.

According to the law of the iterated logarithm (see, e.g., [8, Theorem 9.5]) if \( \{X_k\}_{k=1}^\infty \) are i.i.d. real-valued RVs with zero mean and unit variance, and \( S_n = \sum_{i=1}^n X_i \) for every \( n \in \mathbb{N} \), then

\[
\limsup_{n \to \infty} \frac{S_n}{\sqrt{2n \ln \ln n}} = 1 \quad \text{a.s. (99)}
\]

and

\[
\liminf_{n \to \infty} \frac{S_n}{\sqrt{2n \ln \ln n}} = -1 \quad \text{a.s. (100)}
\]

Eqs. (99) and (100) 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. (99) and (100) 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 5 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 5, it follows that for every \( \alpha \geq 0 \)

\[
\mathbb{P} \left( S_n \geq \alpha \sqrt{2n \ln \ln n} \right) \leq \exp \left( -nD\left( \frac{\delta_n + \gamma}{1 + \gamma} \bigg| \bigg| \frac{\gamma}{1 + \gamma} \right) \right) \quad (101)
\]

where

\[
\delta_n \triangleq \frac{\alpha}{c} \sqrt{\frac{2 \ln \ln n}{n}}, \quad \gamma \triangleq \frac{1}{c^2} \quad (102)
\]
Straightforward calculation shows that

\[ nD(\frac{\delta_n + \gamma}{1 + \gamma} \bigg| \frac{\gamma}{1 + \gamma}) = \frac{n\gamma}{1 + \gamma} \left[ \left( 1 + \frac{\delta_n}{\gamma} \right) \ln(1 + \frac{\delta_n}{\gamma}) + \frac{1}{\gamma} (1 - \delta_n) \ln(1 - \delta_n) \right] \]

\[ \equiv \frac{n\gamma}{1 + \gamma} \left[ \frac{\delta_n^2}{2} \left( \frac{1}{\gamma^2} + \frac{1}{\gamma} \right) + \frac{\delta_n^3}{6} \left( \frac{1}{\gamma} - \frac{1}{\gamma^3} \right) + \ldots \right] \]

\[ = n\frac{\delta_n^2}{2\gamma} - \frac{n\delta_n^3(1 - \gamma)}{6\gamma^2} + \ldots \]

\[ \equiv \alpha^2 \ln \ln n \left[ 1 - \frac{\alpha(c^2 - 1)}{6c} \sqrt{\ln \ln n} + \ldots \right] \quad (103) \]

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 (102). A substitution of (103) into (101) gives that, for every \( \alpha \geq 0, \)

\[ \mathbb{P}\left( S_n \geq \alpha \sqrt{2n \ln \ln n} \right) \leq (\ln n)^{-\alpha^2} \left[ 1 + O\left( \frac{\ln \ln n}{n} \right) \right] \quad (104) \]

and the same bound also applies to \( \mathbb{P}\left( S_n \leq -\alpha \sqrt{2n \ln \ln n} \right) \) for \( \alpha \geq 0. \) This provides complementary information to the limits in (99) and (100) that are provided by the LIL. From Remark 6, which follows from Doob’s maximal inequality for sub-martingales, the inequality in (104) can be strengthened to

\[ \mathbb{P}\left( \max_{1 \leq k \leq n} S_k \geq \alpha \sqrt{2n \ln \ln n} \right) \leq (\ln n)^{-\alpha^2} \left[ 1 + O\left( \frac{\ln \ln n}{n} \right) \right]. \quad (105) \]

It is shown in the following that (105) and the first Borel-Cantelli lemma can serve to prove one part of (99). Using this approach, it is shown that if \( \alpha > 1, \) then the probability that \( S_n > \alpha \sqrt{2n \ln \ln n} \) i.o. is zero. To this end, let \( \theta > 1 \) be set arbitrarily, and define

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

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

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

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

\[ \mathbb{P}(A_n) \leq \mathbb{P}\left( \max_{\theta^{n-1} \leq k \leq \theta^n} S_k \geq \alpha \sqrt{2\theta^{n-1} \ln(\theta^{n-1})} \right) \]

\[ = \mathbb{P}\left( \max_{\theta^{n-1} \leq k \leq \theta^n} S_k \geq \frac{\alpha}{\sqrt{\theta}} \sqrt{2\theta^n \ln(\theta^{n-1})} \right) \]

\[ \leq \mathbb{P}\left( \max_{1 \leq k \leq \theta^n} S_k \geq \frac{\alpha}{\sqrt{\theta}} \sqrt{2\theta^n \ln(\theta^{n-1})} \right) \]

\[ \leq (n \ln \theta)^{-\frac{\alpha^2}{2}} (1 + \beta_n) \quad (106) \]

where the last inequality follows from (105) with \( \beta_n \to 0 \) as \( n \to \infty. \) Since

\[ \sum_{n=1}^{\infty} n^{-\frac{\alpha^2}{2}} < \infty, \quad \forall \alpha > \sqrt{\theta} \]
then it follows from the first Borel-Cantelli lemma that $P(A \text{ i.o.}) = 0$ 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 $\limsup_{n \to \infty} S_n \leq 1$ a.s.

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

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

Theorem 5 therefore gives inequality (105), and it implies one side in each of the two equalities for the LIL in (99) and (100).

C. Relation of Theorem 5 with the Moderate Deviations Principle

According to the moderate deviations theorem (see, e.g., [17, 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 $\Lambda_X(\lambda) = \mathbb{E}[e^{\lambda X}] < \infty$ 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

$$Z_n \triangleq \sqrt{\frac{a_n}{n}} \sum_{i=1}^n X_i, \quad \forall \, n \in \mathbb{N}. \quad (107)$$

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

$$-\frac{1}{2\sigma^2} \inf_{x \in \Gamma^0} x^2 \leq \liminf_{n \to \infty} a_n \ln \mathbb{P}(Z_n \in \Gamma) \leq \limsup_{n \to \infty} a_n \ln \mathbb{P}(Z_n \in \Gamma) \leq -\frac{1}{2\sigma^2} \inf_{x \in \overline{\Gamma}} x^2 \quad (108)$$

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

Let $\eta \in \left(\frac{1}{2}, 1\right)$ 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 (107),

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

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

$$\lim_{n \to \infty} n^{1-2\eta} \ln \mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| \geq \alpha n^\eta\right) = -\frac{\alpha^2}{2\sigma^2}. \quad (109)$$

It is demonstrated in Appendix B that, in contrast to Azuma’s inequality, Theorem 5 provides an upper bound on the probability

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

which coincides with the asymptotic limit in (109). The analysis in Appendix B provides another interesting link between Theorem 5 and a classical result in probability theory, which also emphasizes the significance of the refinements of Azuma’s inequality.
D. 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., [29, 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 \( S \) with transition matrix \( P \), and let the function \( \psi : S \rightarrow S \) be harmonic (i.e., \( \sum_{j \in S} P_{i,j} \psi(j) = \psi(i) \), \( \forall i \in S \)), 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 [36]. These bounds were also shown to be essentially identical to the Hoeffding inequality in the special case of i.i.d. RVs (see [36, Remark 1]).

VII. APPLICATIONS IN INFORMATION THEORY AND RELATED TOPICS

A. Binary Hypothesis Testing

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

Let the RVs \( X_1, X_2, \ldots \) 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

\[
L(X_1, \ldots, X_n) \triangleq \ln \frac{P^n_1(X_1, \ldots, X_n)}{P^n_2(X_1, \ldots, X_n)} = \sum_{i=1}^{n} \ln \frac{P_1(X_i)}{P_2(X_i)}
\]

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

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

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

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

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 \( \lambda, \lambda' \in \mathbb{R} \) that satisfy

\[
-D(P_2||P_1) < \lambda < \lambda' < D(P_1||P_2)
\]

one decides on hypothesis \( H_1 \) if

\[
L(X_1, \ldots, X_n) > n\lambda
\]

and on hypothesis \( H_2 \) if

\[
L(X_1, \ldots, X_n) < n\lambda'.
\]

Note that if \( \lambda' = \lambda \triangleq \lambda \) 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 \( \lambda < \lambda' \) then one decides on \( H_1 \) or \( H_2 \) if the normalized log-likelihood ratio is, respectively, above the upper threshold \( \lambda' \) or below the lower threshold \( \lambda \). 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
\[ \alpha_n^{(1)} \equiv P_1^n \left( L(X_1, \ldots, X_n) \leq n\bar{\lambda} \right) \]  
(112)
\[ \alpha_n^{(2)} \equiv P_1^n \left( L(X_1, \ldots, X_n) \leq n\bar{\lambda} \right) \]  
(113)
and
\[ \beta_n^{(1)} \equiv P_2^n \left( L(X_1, \ldots, X_n) \geq n\bar{\lambda} \right) \]  
(114)
\[ \beta_n^{(2)} \equiv P_2^n \left( L(X_1, \ldots, X_n) \geq n\bar{\lambda} \right) \]  
(115)
then \( \alpha_n^{(1)} \) and \( \beta_n^{(1)} \) are the probabilities of either making an error or declaring an erasure under, respectively, hypotheses \( H_1 \) and \( H_2 \); similarly, \( \alpha_n^{(2)} \) and \( \beta_n^{(2)} \) are the probabilities of making an error under hypotheses \( H_1 \) and \( H_2 \), respectively.

Let \( \pi_1, \pi_2 \in (0, 1) \) denote the a-priori probabilities of the hypotheses \( H_1 \) and \( H_2 \), respectively, so
\[ P_{e,n}^{(1)} = \pi_1 \alpha_n^{(1)} + \pi_2 \beta_n^{(1)} \]  
(116)
is the probability of having either an error or an erasure, and
\[ P_{e,n}^{(2)} = \pi_1 \alpha_n^{(2)} + \pi_2 \beta_n^{(2)} \]  
(117)
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., [17, Theorem 3.4.3] and [31, 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 \( \lambda = \lambda \equiv \bar{\lambda} \), the option of erasures does not exist, and \( P_{e,n}^{(1)} = P_{e,n}^{(2)} \equiv P_{e,n} \) is the error probability.

In the considered general case with erasures, let
\[ \lambda_1 \equiv -\bar{\lambda}, \quad \lambda_2 \equiv -\bar{\lambda} \]  
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
\[ \lim_{n \to \infty} -\frac{\ln \alpha_n^{(1)}}{n} = I(\lambda_1) \]  
(118)
\[ \lim_{n \to \infty} -\frac{\ln \alpha_n^{(2)}}{n} = I(\lambda_2) \]  
(119)
\[ \lim_{n \to \infty} -\frac{\ln \beta_n^{(1)}}{n} = I(\lambda_2) - \lambda_2 \]  
(120)
\[ \lim_{n \to \infty} -\frac{\ln \beta_n^{(2)}}{n} = I(\lambda_1) - \lambda_1 \]  
(121)
where the rate function \( I \) is given by
\[ I(r) \equiv \sup_{t \in \mathbb{R}} \left( tr - H(t) \right) \]  
(122)
and
\[ H(t) = \ln \left( \sum_{x \in \mathcal{X}} P_1(x)^{1-t} P_2(x)^t \right), \quad \forall t \in \mathbb{R}. \]  
(123)
The rate function \( I \) is convex, lower semi-continuous (l.s.c.) and non-negative (see, e.g., [17] and [31]). Note that
\[ H(t) = (t - 1)D_t(P_2||P_1) \]  
where \( D_t(P||Q) \) designates Rényi’s information divergence of order \( t \) [61, Eq. (3.3)], and \( I \) in (122) is the Fenchel-Legendre transform of \( H \) (see, e.g., [17, Definition 2.2.2]).
From (116)–(121), the exact exponents of $P_{e,n}^{(1)}$ and $P_{e,n}^{(2)}$ are equal to

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

(124)

and

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

(125)

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

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

(126)

which coincides with the error exponent in [17, Theorem 3.4.3] (or [31, Theorem 6.4]). The optimal threshold for obtaining the best error exponent of the error probability $P_{e,n}$ is equal to zero (i.e., $\lambda = 0$); in this case, the exact error exponent is equal to

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

(127)

which is the Chernoff information of the probability measures $P_1$ and $P_2$ (see [15, Eq. (11.239)]), and it is symmetric (i.e., $C(P_1, P_2) = C(P_2, P_1)$). Note that, from (122), $I(0) = \sup_{t \in \mathbb{R}} (-H(t)) = -\inf_{t \in \mathbb{R}} (H(t))$; the minimization in (127) 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 (123) 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.

2) Lower Bound on the Exponents via Theorem 5: In the following, the tightness of Theorem 5 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 subsection to the exact exponents from the previous subsection.

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

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

and

$$U_k = \mathbb{E}_{P^n} \left[ L(X_1, \ldots, X_n) \mid \mathcal{F}_k \right].$$

(128)

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

$$U_k = \mathbb{E}_{P^n} \left[ \sum_{i=1}^{n} \ln \frac{P_1(X_i)}{P_2(X_i)} \mid \mathcal{F}_k \right]$$

$$= \sum_{i=1}^{k} \ln \frac{P_1(X_i)}{P_2(X_i)} + \sum_{i=k+1}^{n} \mathbb{E}_{P^n} \left[ \ln \frac{P_1(X_i)}{P_2(X_i)} \right]$$

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

In particular

$$U_0 = nD(P_1 || P_2),$$

(129)

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

(130)
and, for every $k \in \{1, \ldots, n\}$,
\[
U_k - U_{k-1} = \ln \frac{P_1(X_k)}{P_2(X_k)} - D(P_1 \| P_2).
\tag{131}
\]
Let
\[
d_1 \triangleq \max_{x \in \mathcal{X}} \left| \ln \frac{P_1(x)}{P_2(x)} - D(P_1 \| P_2) \right|
\tag{132}
\]
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 (131) and (132)
\[
|U_k - U_{k-1}| \leq d_1
\]
holds a.s. for every $k \in \{1, \ldots, n\}$, and due to the statistical independence of the RVs in the sequence $\{X_i\}$
\[
\mathbb{E}_{P_1}[(U_k - U_{k-1})^2 \mid \mathcal{F}_{k-1}]
\]
\[
= \mathbb{E}_{P_1} \left[ \left( \ln \frac{P_1(X_k)}{P_2(X_k)} - D(P_1 \| P_2) \right)^2 \right]
\]
\[
= \sum_{x \in \mathcal{X}} P_1(x) \left( \ln \frac{P_1(x)}{P_2(x)} - D(P_1 \| P_2) \right)^2 \]
\[
\triangleq \sigma_1^2.
\tag{133}
\]
Let
\[
\varepsilon_{1,1} = D(P_1 \| P_2) - \underline{\lambda}, \quad \varepsilon_{2,1} = D(P_2 \| P_1) + \underline{\lambda}
\tag{134}
\]
\[
\varepsilon_{1,2} = D(P_1 \| P_2) - \overline{\lambda}, \quad \varepsilon_{2,2} = D(P_2 \| P_1) + \overline{\lambda}
\tag{135}
\]
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 5
\[
\alpha_n^{(1)} \triangleq P_1^n \left( L(X_1, \ldots, X_n) \leq n\overline{\lambda} \right)
\triangleq P_1^n (U_n - U_0 \leq -\varepsilon_{1,1} n)
\tag{136}
\]
\[
\leq \exp \left( -n D \left( \frac{\delta_{1,1} + \gamma_1}{1 + \gamma_1} \left\| \frac{\gamma_1}{1 + \gamma_1} \right\| \right) \right)
\tag{137}
\]
where equality (a) follows from (129), (130) and (134), and inequality (b) follows from Theorem 5 with
\[
\gamma_1 \triangleq \frac{\sigma_1^2}{d_1}, \quad \delta_{1,1} \triangleq \frac{\varepsilon_{1,1}}{d_1}
\tag{138}
\]
Note that if $\varepsilon_{1,1} > d_1$ then it follows from (131) and (132) that $\alpha_n^{(1)}$ is zero; in this case $\delta_{1,1} > 1$, so the divergence in (137) 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 (128), under hypothesis $H_2$, let us define the martingale sequence $\{U_k, \mathcal{F}_k\}_{k=0}^n$ with the same filtration and
\[
U_k = \mathbb{E}_{P_2^n} [L(X_1, \ldots, X_n) \mid \mathcal{F}_k], \quad \forall k \in \{0, \ldots, n\}.
\tag{139}
\]
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
\[
U_0 = -n D(P_2 \| P_1), \quad U_n = L(X_1, \ldots, X_n).
\tag{140}
\]
For every $k \in \{1, \ldots, n\}$,
\[
U_k - U_{k-1} = \ln \frac{P_1(X_k)}{P_2(X_k)} + D(P_2 \| P_1).
\tag{141}
\]
Let
\[ d_2 \triangleq \max_{x \in \mathcal{X}} \left| \ln \frac{P_2(x)}{P_1(x)} - D(P_2 || P_1) \right| \]  
then, the jumps of the latter martingale sequence are uniformly bounded by \( d_2 \) and, similarly to (133), for every \( k \in \{1, \ldots, n\} \)

\[
\mathbb{E}_{P_n^k} \left[ (U_k - U_{k-1})^2 \mid \mathcal{F}_{k-1} \right] = \sum_{x \in \mathcal{X}} \left\{ P_2(x) \left( \ln \frac{P_2(x)}{P_1(x)} - D(P_2 || P_1) \right)^2 \right\} \triangleq \sigma_2^2.
\]  
Hence, it follows from Theorem 5 that

\[
\beta_n^{(1)} \triangleq P_n^2(L(X_1, \ldots, X_n) \geq n\lambda) = P_n^2(U_n - U_0 \geq \varepsilon_{2,1} n) \leq \exp \left( -n D \left( \frac{\delta_{2,1} + \gamma_2}{1 + \gamma_2} \right) \right)
\]  
where the equality in (144) holds due to (140) and (134), and (145) follows from Theorem 5 with

\[
\gamma_2 \triangleq \frac{\sigma_2^2}{d_2^2}, \quad \delta_{2,1} \triangleq \frac{\varepsilon_{2,1} d_2}{d_2^2}.
\]  
and \( d_2, \sigma_2 \) are introduced, respectively, in (142) and (143).

From (116), (137) and (145), the exponent of the probability of either having an error or an erasure is lower bounded by

\[
\lim_{n \to \infty} - \frac{\ln P_{e,n}^{(1)}}{n} \geq \min_{i=1,2} D \left( \frac{\delta_{i,1} + \gamma_i}{1 + \gamma_i} \right). \quad (147)
\]

Similarly to the above analysis, one gets from (117) and (135) that the error exponent is lower bounded by

\[
\lim_{n \to \infty} - \frac{\ln P_{e,n}^{(2)}}{n} \geq \min_{i=1,2} D \left( \frac{\delta_{i,2} + \gamma_i}{1 + \gamma_i} \right) \quad (148)
\]

where

\[
\delta_{1,2} \triangleq \frac{\varepsilon_{1,2}}{d_1}, \quad \delta_{2,2} \triangleq \frac{\varepsilon_{2,2}}{d_2}.
\]  
For the case of a single threshold (i.e., \( \lambda = \lambda \triangleq \lambda \)) then (147) and (148) coincide, and one obtains that the error exponent satisfies

\[
\lim_{n \to \infty} - \frac{\ln P_{e,n}}{n} \geq \min_{i=1,2} D \left( \frac{\delta_i + \gamma_i}{1 + \gamma_i} \right) \quad (150)
\]

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., [17, p. 93]), which then yields that (150) is satisfied with

\[
\delta_1 = \frac{D(P_1 || P_2)}{d_1}, \quad \delta_2 = \frac{D(P_2 || P_1)}{d_2} \quad (151)
\]
with \( d_1 \) and \( d_2 \) from (132) and (142), respectively. The right-hand side of (150) 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 subsection via Theorem 5, 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 \( \alpha^{(1)}_n \), \( \alpha^{(2)}_n \), \( \beta^{(1)}_n \) and \( \beta^{(2)}_n \) via Azuma’s inequality, and then use them to derive lower bounds on the exponents of \( P^{(1)}_{e,n} \) and \( P^{(2)}_{e,n} \).

From (131), (132), (136), (138), and Azuma’s inequality
\[
\alpha^{(1)}_n \leq \exp\left(-\frac{\delta^2_{1,1} n}{2}\right),
\]
and, similarly, from (141), (142), (144), (146), and Azuma’s inequality
\[
\beta^{(1)}_n \leq \exp\left(-\frac{\delta^2_{1,1} n}{2}\right).
\]

From (113), (115), (135), (149) and Azuma’s inequality
\[
\alpha^{(2)}_n \leq \exp\left(-\frac{\delta^2_{1,2} n}{2}\right),
\]
\[
\beta^{(2)}_n \leq \exp\left(-\frac{\delta^2_{1,2} n}{2}\right).
\]

Therefore, it follows from (116), (117) and (152)–(155) that the resulting lower bounds on the exponents of \( P^{(1)}_{e,n} \) and \( P^{(2)}_{e,n} \) are
\[
\lim_{n \to \infty} -\frac{\ln P^{(j)}_{e,n}}{n} \geq \min_{i=1,2} \frac{\delta^2_{i,j}}{2}, \quad j = 1, 2
\]
as compared to (147) and (148) which give, for \( j = 1, 2, \)
\[
\lim_{n \to \infty} -\frac{\ln P^{(j)}_{e,n}}{n} \geq \min_{i=1,2} D\left(\frac{\delta_{i,j} + \gamma_i}{1 + \gamma_i}\right).
\]

For the specific case of a zero threshold, the lower bound on the error exponent which follows from Azuma’s inequality is given by
\[
\lim_{n \to \infty} -\frac{\ln P^{(j)}_{e,n}}{n} \geq \min_{i=1,2} \frac{\delta^2_{i,j}}{2}
\]
with the values of \( \delta_1 \) and \( \delta_2 \) in (151).

The lower bounds on the exponents in (156) and (157) are compared in the following. Note that the lower bounds in (156) are loosened as compared to those in (157) since they follow, respectively, from Azuma’s inequality and its improvement in Theorem 5.

The divergence in the exponent of (157) is equal to
\[
D\left(\frac{\delta_{i,j} + \gamma_i}{1 + \gamma_i}\right)
= \frac{\delta_{i,j} + \gamma_i}{1 + \gamma_i} \ln\left(1 + \frac{\delta_{i,j}}{\gamma_i}\right) + \frac{1 - \delta_{i,j}}{1 + \gamma_i} \ln(1 - \delta_{i,j})
= \frac{\gamma_i}{1 + \gamma_i} \left[\left(1 + \frac{\delta_{i,j}}{\gamma_i}\right) \ln\left(1 + \frac{\delta_{i,j}}{\gamma_i}\right) + \frac{(1 - \delta_{i,j}) \ln(1 - \delta_{i,j})}{\gamma_i}\right].
\]

Lemma 7:
\[
(1 + u) \ln(1 + u) \geq \begin{cases} \frac{u}{2}, & u \in [-1, 0] \\ \frac{u^2}{2} - \frac{u^3}{6}, & u \geq 0 \end{cases}
\]
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 relies on some elementary calculus. □

Since \( \delta_{i,j} \in [0, 1] \), then (159) and Lemma 7 imply that

\[
D\left( \frac{\delta_{i,j} + \gamma_i}{1 + \gamma_i} \mid \frac{\gamma_i}{1 + \gamma_i} \right) \geq \frac{\delta_{i,j}^2}{2\gamma_i} - \frac{\delta_{i,j}^3}{6\gamma_i^2(1 + \gamma_i)}. \tag{161}
\]

Hence, by comparing (156) with the combination of (157) and (161), then it follows that (up to a second-order approximation) the lower bounds on the exponents that were derived via Theorem 5 are improved by at least a factor of \( \left( \max_i \gamma_i \right)^{-1} \) as compared to those that follow from Azuma’s inequality.

**Example 11:** 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 (150) and (151) is equal to \( 1.77 \cdot 10^{-2} \), whereas the loosened lower bound in (158) is equal to \( 1.39 \cdot 10^{-2} \). In this case \( \gamma_1 = \frac{2}{3} \) and \( \gamma_2 = \frac{7}{9} \), so the improvement in the lower bound on the error exponent is indeed by a factor of approximately

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

Note that, from (137), (145) and (152)–(155), 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 (127)), 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

\[
J(\theta) \triangleq \mathbb{E}_\theta \left[ \frac{\partial}{\partial \theta} \ln P_\theta(x) \right]^2 \tag{162}
\]

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

\[
\lim_{\theta \to \theta^0} \frac{D(P_\theta \mid \mid P_{\theta^0})}{(\theta - \theta^0)^2} = \frac{J(\theta)}{2} \tag{163}
\]

(note that if it was a relative entropy to base 2 then the right-hand side of (163) would have been divided by \( \ln 2 \), and be equal to \( \frac{J(\theta)}{\ln 4} \) as in [15, Eq. (12.364)].)

**Proposition 2:** Under the above assumptions,

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

\[
\lim_{\theta \to \theta^0} \frac{C(P_\theta, P_{\theta^0})}{(\theta - \theta^0)^2} = \frac{J(\theta)}{8}. \tag{164}
\]
Let
\[ E_L(P_0, P_{\theta}) \triangleq \min_{i=1,2} D\left( \frac{\delta_i + \gamma_i}{1 + \gamma_i} \bigg| \frac{\gamma_i}{1 + \gamma_i} \right) \]  
be the lower bound on the error exponent in (150) which corresponds to \( P_1 \triangleq P_0 \) and \( P_2 \triangleq P_{\theta'} \), then also
\[ \lim_{\theta' \to \theta} \frac{E_L(P_0, P_{\theta'})}{(\theta - \theta')^2} = \frac{J(\theta)}{8}. \]  

Let
\[ \tilde{E}_L(P_0, P_{\theta'}) \triangleq \min_{i=1,2} \frac{\delta_i^2}{2} \]  
be the loosened lower bound on the error exponent in (158) which refers to \( P_1 \triangleq P_0 \) and \( P_2 \triangleq P_{\theta'} \). Then,
\[ \lim_{\theta' \to \theta} \frac{\tilde{E}_L(P_0, P_{\theta'})}{(\theta - \theta')^2} = \frac{a(\theta) J(\theta)}{8} \]  
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 C.

Proposition 2 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 behavior of the error exponent for a binary hypothesis testing when the relative entropy between the pair of probability mass functions that characterizes the two hypotheses tends to zero. This stays in contrast to the loosened error exponent for a binary hypothesis testing when the relative entropy between the pair of probability mass functions that characterizes the two hypotheses tends to zero.

**Example 12:** Consider the index family of probability mass functions defined over the binary alphabet \( \mathcal{X} = \{0, 1\} \):
\[ P_0(0) = 1 - \theta, \ P_0(1) = \theta, \ \forall \theta \in (0, 1). \]

From (162), 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 (164) and (166)
\[ C(P_{\theta_1}, P_{\theta_2}), E_L(P_{\theta_1}, P_{\theta_2}) \approx \frac{J(\theta)(\theta_1 - \theta_2)^2}{8} = 2.00 \cdot 10^{-4}. \]

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

**B. 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
\[ \mathbb{E}[d_{\text{min}}(C)] = h_2^{-1}(1 - R) \]
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 [4]).

Let \( H \) designate an \( n(1 - R) \times n \) parity-check matrix of a linear block code \( C \) 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

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

This leads to the following theorem:

**Theorem 10:** [*The minimum distance of binary linear block codes*] Let \( C \) 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 \( C \) is in the interval

\[
[n h^{-1}_2(1-R) - \alpha \sqrt{n}, \ n h^{-1}_2(1-R) + \alpha \sqrt{n}]
\]

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 [9].

### C. 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 [26]. 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., [63]). 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., [66, 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 [21]. 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 [66, Section II-E]. In [66], 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 [66, 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 \( \epsilon \) designates the gap in rate to capacity, then the number of fundamental cycles should grow at least like \( \log \frac{1}{\epsilon} \). Hence, this lower bound remains unbounded as the gap to capacity tends to zero. Consistently with the study in [21] on cycle-free codes, the lower bound on the cardinality of the fundamental system of cycles in [66, 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 \( m = n(1 - R_d) \) parity-check nodes, so there are in total \( |V(\mathcal{G})| = n(2 - R_d) \) nodes. Let \( a_R \) designate the average right degree (i.e., the average degree of the parity-check nodes), then the number of edges in \( \mathcal{G} \) is given by \( |E(\mathcal{G})| = ma_R \). Therefore, for a code from the \( (n, \lambda, \rho) \) LDPC code ensemble, the cardinality of the fundamental system of cycles satisfies the equality

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

where

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

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

Let

\[
E \triangleq |E(\mathcal{G})| = n(1 - R_d)a_R
\]

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, let us 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 \( \mathcal{F}_i \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_E \) 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 1, it follows that for every \( \alpha \geq 0 \)

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

where the last transition follows from (169), and the function \( f \) was defined in (58). Hence, for \( \alpha > 1 \), this probability is zero (since \( f(\alpha) = +\infty \) for \( \alpha > 1 \)). Note that, from (169), \( \mathbb{E}_{\text{LDPC}(n, \lambda, \rho)}[\beta(\mathcal{G})] \) scales linearly with \( n \).

The combination of Eqs. (58), (170), (171) gives the following statement:

**Theorem 11:** [Concentration inequality for the cardinality of the fundamental system of cycles] Let \( \text{LDPC}(n, \lambda, \rho) \) be the LDPC code ensemble that is characterized by a block length \( n \), and a pair of degree distributions (from the edge perspective) of \( \lambda \) 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:

\[
\mathbb{P} \left( |\beta(\mathcal{G}) - \mathbb{E}_{\text{LDPC}(n, \lambda, \rho)}[\beta(\mathcal{G})]| \geq \alpha n \right) \leq 2 \cdot 2^{-\left[1 - h_2(\frac{2}{\alpha^2})\right]n}
\]
where \( h_2 \) designates the binary entropy function to the base 2, \( \eta \equiv \frac{\alpha}{(1-R_d)\beta} \), and \( R_d \) and \( a_R \) designate, respectively, the design rate and average right degree of the ensemble. Consequently, if \( \eta > 1 \), this probability is zero.

**Remark 12:** The loosened version of Theorem 11, which follows from Azuma’s inequality, gets the form

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

for every \( \alpha \geq 0 \), and \( \eta \) as defined in Theorem 11. 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 1 in Figure 1).

**Remark 13:** 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., \( R_d = (1-\varepsilon)C \) [43]. Therefore, for small values of \( \alpha \), the exponential decay rate in the inequality of Theorem 11 scales like \( (\log \frac{1}{\varepsilon})^{-2} \). This large-deviations result complements the result in [66, 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 14:** Consider small deviations from the expected value that scale like \( \sqrt{n} \). Note that Corollary 1 is a special case of Theorem 5 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 1 that Corollary 1 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.

### D. Performance of LDPC Codes under Iterative Message-Passing Decoding for Memoryless Symmetric Channels

In the following, ensembles of binary LDPC codes are considered with the same notation as of the previous subsection. The following theorem was proved by Richardson and Urbanke in [63, pp. 487–490], based on the Azuma-Hoeffding inequality, and it is a central result in the theory of codes defined on graphs and iterative decoding algorithms.

**Theorem 12:** [Concentration of the bit error probability around the ensemble average] Let \( \mathcal{C} \), a code chosen uniformly at random from the ensemble \( \text{LDPC}(n,\lambda,\rho) \), 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 \( P_b(\mathcal{C}, a_{\text{MBIOS}}, l) \) 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

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

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).

### E. On the Concentration of the Conditional Entropy for LDPC Code Ensembles

A large deviations analysis of the conditional entropy for random ensembles of LDPC codes was introduced in [55, Theorem 4] and [57, Theorem 1]. The following theorem is proved in [55, Appendix I], based on the Azuma-Hoeffding inequality, and it is rephrased in the following to consider small deviations of order \( \sqrt{n} \) (instead of large deviations of order \( n \)):

**Theorem 13:** [Concentration of the conditional entropy] Let \( \mathcal{C} \) be chosen uniformly at random from the ensemble \( \text{LDPC}(n,\lambda,\rho) \). Assume that the transmission of the code \( \mathcal{C} \) takes place over a memoryless binary-input output-symmetric (MBIOS) channel. Let \( H(X|Y) \) designate the conditional entropy of the transmitted codeword \( X \) given the received sequence \( Y \) from the channel. Then, for any \( \xi > 0 \),

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

where \( B \equiv \frac{1}{2(\max_c a_c^\text{max} + 1)^2(1-R_d)} \), \( a_c^\text{max} \) is the maximal check-node degree, and \( R_d \) 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 13 in [55, 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 \( \text{LDPC}(n, \lambda, \rho) \). Define the RV
\[
Z = H_\mathcal{G}(X|Y)
\]
which forms the conditional entropy when the transmission takes place over an MBIOS channel whose transition probability is given by \( P_{Y|X}(y|x) = \prod_{i=1}^n p_{Y_i|X}(y_i|x_i) \) where \( p_{Y_i|X}(y|1) = p_{Y_i|X}(-y|0) \). Fix an arbitrary order for the \( m = n(1-R_d) \) parity-check nodes where \( R_d \) forms the design rate of the LDPC code ensemble. Let \( \{ \mathcal{F}_t \}_{t \in \{0,1,...,m\}} \) 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, \( Z_0 = \mathbb{E}[H_\mathcal{G}(X|Y)] \) 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 [55, 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 [55, Appendix I] by increasing degree. Let \( r = (r_1, r_2, \ldots) \) 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 \( m_1 = n(1-R_d)\Gamma_{r_1} \) parity-check nodes are of degree \( r_1 \), the successive \( m_2 = n(1-R_d)\Gamma_{r_2} \) 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 [55, Appendix I]
\[
|Z_{t+1} - Z_t| \leq (r + 1) H_\mathcal{G}(\bar{X}|Y) \tag{172}
\]
where \( H_\mathcal{G}(\bar{X}|Y) \) is a RV which designates the conditional entropy of a parity-bit \( \bar{X} = X_{i_1} \oplus \ldots \oplus X_{i_t} \) (i.e., \( \bar{X} \) is equal to the modulo-2 sum of some \( r \) bits in the codeword \( X \)) given the received sequence \( Y \) at the channel output. The proof in [55, Appendix I] was then completed by upper bounding the parity-check degree \( r \) by the maximal parity-check degree \( d_c^{\text{max}} \), and also by upper bounding the conditional entropy of the parity-bit \( \bar{X} \) by 1. This gives
\[
|Z_{t+1} - Z_t| \leq d_c^{\text{max}} + 1 \quad t = 0,1,\ldots,m - 1. \tag{173}
\]
which then proves Theorem 13 from Azuma’s inequality. Note that the \( d_i \)’s in Theorem 1 are equal to \( d_c^{\text{max}} + 1 \), and \( n \) in Theorem 1 is replaced with the length \( m = n(1-R_d) \) 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 [55, Appendix I] in two respects:

- The first difference is related to the upper bound on the conditional entropy \( H_\mathcal{G}(\bar{X}|Y) \) in (172) where \( \bar{X} \) is the modulo-2 sum of some \( r \) bits of the transmitted codeword \( X \) given the channel output \( Y \). Instead of taking the most trivial upper bound that is equal to 1, as was done in [55, 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 3).

- The second difference is minor, but it proves to be helpful for tightening the concentration 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 (172) with \( d_c^{\text{max}} + 1 \), 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 \( n(1-R_d)\Gamma_r \). The effect of this simple modification will be shown in Example 14.

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

**Proposition 3:** Let \( \mathcal{G} \) be a bipartite graph which corresponds to a binary linear block code whose transmission takes place over an MBIOS channel. Let \( X \) and \( Y \) 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 $X$. Then, the conditional entropy of $\tilde{X}$ given $Y$ satisfies

$$H_G(\tilde{X}|Y) \leq h_2\left(\frac{1 - C_r^2}{2}\right).$$

(174)

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

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

(175)

and

$$1 - C^r$$

(176)

respectively, where $h_2^{-1}$ in (175) 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 (174) holds with equality (where both sides of (174) are zero), whereas the trivial upper bound is 1.

Proof: Let us upper bound the conditional entropy $H(\tilde{X}|Y)$ with $H(\tilde{X}|Y_1, \ldots, Y_n)$, 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 [66, Eq. (17) and Appendix I], the conditional entropy of the bit $\tilde{X}$ given the $n$-length received sequence $Y$ satisfies the inequality

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

(177)

where (see [66, Eq. (19)])

$$g_p \triangleq \int_{0}^{\infty} a(l)(1 + e^{-l}) \tanh^2 \left(\frac{l}{2}\right) dl, \quad \forall \, p \in \mathbb{N}$$

(178)

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 [66, Lemmas 4 and 5], it follows that

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

Substituting this inequality in (177) gives that

$$H(\tilde{X}|Y) \leq 1 - \frac{1}{2 \ln 2} \sum_{p=1}^{\infty} \frac{C^p}{p(2p-1)}$$

$$= h_2\left(\frac{1 - C_r^2}{2}\right)$$

(179)

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

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

(180)

The tightened bound on the conditional entropy for the BSC is obtained from (177) and the equality

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

which holds for the BSC (see [66, Eq. (97)]). This replaces $C$ on the right-hand side of (179) with $(1 - 2h_2^{-1}(1-C))^2$, thus leading to the tightened bound in (175).

The tightened result for the BEC holds since from (178)

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

(see [66, Appendix II]), and a substitution of this equality in (177) gives (175) (note that $\sum_{p=1}^{\infty} \frac{1}{p(2p-1)} = 2 \ln 2$). This completes the proof of Proposition 3.

From Proposition 3 and (172)

$$|Z_{l+1} - Z_{l}| \leq (r + 1)h_2\left(\frac{1 - C_r^2}{2}\right)$$

(181)
with the corresponding two improvements for the BSC and BEC (where the second term on the right-hand side of (181) is replaced by (175) and (176), respectively). This improves the loosened bound of \((d_c^{\text{max}} + 1)\) in [55, Appendix I]. From (181) and Theorem 1, we obtain the following tightened version of the concentration inequality in Theorem 13.

**Theorem 14:** [A tightened concentration inequality for the conditional entropy] Let \(C\) be chosen uniformly at random from the ensemble \(\text{LDPC}(n, \lambda, \rho)\). Assume that the transmission of the code \(C\) takes place over a memoryless binary-input output-symmetric (MBIOS) channel. Let \(H(X|Y)\) designate the conditional entropy of the transmitted codeword \(X\) given the received sequence \(Y\) at the channel output. Then, for every \(\xi > 0\),

\[
P\left( |H(X|Y) - \mathbb{E}_{\text{LDPC}(n, \lambda, \rho)}[H(X|Y)]| \geq \xi \sqrt{n} \right) \leq 2 \exp(-B\xi^2) \tag{182}
\]

where

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

and \(d_{\text{max}}^{\text{max}}\) is the maximal check-node degree, \(R_d\) is the design rate of the ensemble, and \(C\) is the channel capacity (in bits per channel use). Furthermore, for a binary symmetric channel (BSC) or a binary erasure channel (BEC), the parameter \(B\) on the right-hand side of (182) can be improved (i.e., increased), respectively, to

\[
B \triangleq \frac{1}{2(1 - R_d) \sum_{i=1}^{d_{\text{max}}^{\text{max}}} \left\{ (i + 1)^2 \left[ \Gamma_i \left( h_2 \left( \frac{1 - C^2}{2} \right) \right) \right]^2 \right\}} \tag{184}
\]

**Remark 15:** From (183), Theorem 14 indeed yields a stronger concentration inequality than Theorem 13.

**Remark 16:** In the limit where \(C \rightarrow 1\) bit per channel use, it follows from (183) that if \(d_{\text{max}}^{\text{max}} < \infty\) then \(B \rightarrow \infty\). This is in contrast to the value of \(B\) in Theorem 13 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 14 is tight in this case.

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

**Lemma 8:** If \(d_{\text{max}}^{\text{max}} = \infty\) and \(\rho'(1) < \infty\) then \(B \rightarrow \infty\) in the limit where \(C \rightarrow 1\).

**Proof:** See Appendix D. \(\blacksquare\)

This is in contrast to the value of \(B\) in Theorem 13 which vanishes when \(d_{\text{max}}^{\text{max}} = \infty\), and therefore Theorem 13 is not informative in this case (see Example 14).

**Example 13:** [Comparison of Theorems 13 and 14 for right-regular LDPC code ensembles] In the following, we exemplify the improvement in the tightness of Theorem 14 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 [63, Example 4.38] which expresses the capacity of the BIAWNGC 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 $d_c = 20$), then $B$ in Theorem 14 is improved by a factor of

$$\frac{1}{h_2 \left( \frac{1 - C^d_2}{2} \right)^2} = 5.134.$$ 

This implies that the inequality in Theorem 14 is satisfied with a block length that is 5.134 times shorter than the block length which corresponds to Theorem 13. For the BEC, the result is improved by a factor of

$$\frac{1}{(1 - C^d)^2} = 9.051$$

due to the tightened value of $B$ in (184) as compared to Theorem 13.

**Example 14:** [Comparison of Theorems 13 and 14 for a heavy-tail Poisson distribution (Tornado codes)] In the following, we compare Theorems 13 and 14 for Tornado LDPC code ensembles. This capacity-achieving sequence for the BEC refers to the heavy-tail Poisson distribution, and it was introduced in [43, Section IV], [72] (see also [63, Problem 3.20]). We rely in the following on the analysis in [66, 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 13, and this theorem therefore is useless for the considered code ensemble. On the other hand, from Theorem 14

$$\sum_i (i + 1)^2 \Gamma_i \left[ h_2 \left( \frac{1 - C^d_2}{2} \right)^2 \right] \leq \sum_i (i + 1)^2 \Gamma_i$$

(a)

$$\leq \sum_i \rho_i (i + 2) \int_0^1 \rho(x) \, dx + 1$$

(b)

$$= (\rho'(1) + 3) d^\text{avg}_c + 1$$

(c)

$$= \left( \frac{\lambda'(0) \rho'(1)}{\lambda_2} + 3 \right) d^\text{avg}_c + 1$$

(d)

$$\leq \left( \frac{1}{p \lambda_2} + 3 \right) d^\text{avg}_c + 1$$

(e)

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

(f)

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

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

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

$$d^\text{avg}_c = \frac{1}{\int_0^1 \rho(x) \, dx}$$

where $d^\text{avg}_c$ 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 [66, Appendix VI] (an upper
bound on $\lambda_2$ is derived in [66, 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 14 scales (at least) like

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

Theorem 14 is therefore useful for the analysis of this LDPC code ensemble. As is 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 14).

Example 15: This Example forms a direct continuation of Example 13 for the $(n, d_v, d_c)$ regular LDPC code ensembles where $d_v = 2$ and $d_c = 20$. With the settings in this example, Theorem 13 gives that

$$\mathbb{P}\left(\left|H(X|Y) - \mathbb{E}_{LDPC(n, \lambda, \rho)}[H(X|Y)]\right| \geq \xi \sqrt{n}\right) \leq 2 \exp(-0.0113 \xi^2), \quad \forall \xi > 0. \quad (185)$$

As was mentioned already in Example 13, the exponential inequalities in Theorem 14 achieve an improvement in the exponent of Theorem 13 by factors 5.134 and 9.051 for the BIAWGNC and BEC, respectively. One therefore obtains from the concentration inequalities in Theorem 14 that, for every $\xi > 0$,

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

F. 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 [73, 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 15: [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

$$n \left[\frac{\left(l\left(1 - (1-\alpha)^r\right)ight)}{r} - \sqrt{2l\alpha \left(h(\alpha) + \delta\right)}\right] \quad (187)$$

where $h$ is 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

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

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 $V$ 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 $i \in \{0, \ldots, n\alpha l\}$

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

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 in Section III-A,

\[ P(\mathbb{E}[X(G)] - X(G) \geq \lambda \sqrt{\ln n}) \leq \exp\left(-\frac{\lambda^2}{2}\right), \quad \forall \lambda > 0. \]

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

Since \( \binom{n}{n\alpha} \leq e^{nh(\alpha)} \), then we get the loosened bound \( \exp(nh(\alpha) - \frac{\lambda^2}{2}) \). Finally, choosing \( \lambda = \sqrt{2n(h(\alpha) + \delta)} \) gives the required result. 

\[ \square \]

G. Concentration of the Crest-Factor for OFDM Signals

Orthogonal-frequency-division-multiplexing (OFDM) is a modulation that converts a high-rate data stream into a number of low-rate streams that are transmitted over parallel narrow-band channels. OFDM is widely used in several international standards for digital audio and video broadcasting, and for wireless local area networks. For a textbook providing a survey on OFDM, see e.g. [56, 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 [42], 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 [42], [64, Section 4] and [81]).

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

\[ s(t) = \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} X_i \exp\left(\frac{j2\pi t}{T}\right), \quad 0 \leq t \leq T. \]  

(188)

(188)

Let's 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.,

\[ \frac{1}{T} \int_0^T |s(t)|^2 dt = 1. \]  

(189)

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

\[ \text{CF}_n(s) \triangleq \max_{0 \leq t \leq T} |s(t)|. \]  

(190)

From [64, Section 4] and [81], it follows that the CF scales with high probability like \( \sqrt{\ln n} \) for large \( n \). In [42, Theorem 3 and Corollary 5], a concentration inequality was derived for the CF of OFDM signals. It states that for an arbitrary \( c \geq 2.5 \)

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

**Remark 17:** The analysis used to derive this rather strong concentration inequality (see [42, Appendix C]) requires some assumptions on the distribution of the \( X_i \)'s (see the two conditions in [42, Theorem 3] followed by [42, 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 [42, Theorem 3].

In the following, Azuma's inequality and a refined version of this inequality are considered under the assumption that \( \{X_j\}_{j=0}^{n-1} \) are independent complex-valued random variables with magnitude 1, attaining the \( M \) points of an \( M \)-ary PSK constellation with equal probability.
1) Establishing Concentration of the Crest-Factor via Azuma’s Inequality: In the following, Azuma’s inequality is used to derive a concentration result. Let us define

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

(191)

Based on a standard construction of martingales, \( \{Y_i, F_i\}_{i=0}^n \) is a martingale where \( F_i \) is the \( \sigma \)-algebra that is generated by the first \( i \) symbols \( (X_0, \ldots, X_{i-1}) \) in (188). Hence, \( F_0 \subseteq F_1 \subseteq \ldots \subseteq F_n \) 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 (190), by at most \( \frac{2}{\sqrt{n}} \) (see the first part of Appendix E). It therefore follows from Azuma’s inequality that, for every \( \alpha > 0 \),

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

(192)

which demonstrates concentration around the expected value.

2) Establishing Concentration of the Crest-Factor via the Refined Version of Azuma’s Inequality in Proposition 1: In the following, we rely on Proposition 1 to derive an improved concentration result. For the martingale sequence \( \{Y_i\}_{i=0}^n \) in (191), Appendix E gives that a.s.

\[ |Y_i - Y_{i-1}| \leq \frac{2}{\sqrt{n}}, \quad \mathbb{E}[(Y_i - Y_{i-1})^2 \mid F_{i-1}] \leq \frac{2}{n} \]

(193)

for every \( i \in \{1, \ldots, n\} \). Note that the conditioning on the \( \sigma \)-algebra \( F_{i-1} \) 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 1 therefore implies that for an arbitrary \( \alpha > 0 \)

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

(194)

(since \( \delta = \frac{\alpha}{2} \) and \( \gamma = \frac{1}{2} \) in the setting of Proposition 1). Note that the exponent in the last inequality is doubled as compared to the bound that was obtained in (192) via Azuma’s inequality, and the term which scales like \( O \left( \frac{1}{\sqrt{n}} \right) \) on the right-hand side of (194) is expressed explicitly for finite \( n \) (see Appendix A).

3) A Concentration Inequality via Talagrand’s Method: In his seminal paper [76], 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., [17, Section 2.4.2], [53, Section 4] and [76]). 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 3 (Hamming distance): Let \( x, y \) be two \( n \)-length vectors. The Hamming distance between \( x \) and \( y \) is the number of coordinates where \( x \) and \( y \) disagree, i.e.,

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

where \( I \) stands for the indicator function.

The following suggests a generalization and normalization of the previous distance metric.
\textbf{Definition 4:} 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
\[
d_a(x, y) \triangleq \sum_{i=1}^n a_i I_{\{x_i \neq y_i\}}.
\]
Hence, \( d_H(x, y) = \sqrt{n} d_a(x, y) \) for \( a = \left( \frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}} \right) \).

The following is a special form of Talagrand’s inequalities ([53, Chapter 4], [76], [77]).

\textbf{Theorem 16 (Talagrand’s inequality):} Let the random vector \( X = (X_1, \ldots, X_n) \) 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 \( f : A \rightarrow \mathbb{R} \) satisfy the condition that, for every \( x \in A \), there exists a non-negative, normalized \( n \)-length vector \( a = a(x) \) such that
\[
f(x) \leq f(y) + \sigma d_a(x, y), \quad \forall y \in A
\]
for some fixed value \( \sigma > 0 \). Then, for every \( \alpha \geq 0 \),
\[
\mathbb{P}(\|f(X) - m\| \geq \alpha) \leq 4 \exp \left( -\frac{\alpha^2}{4\sigma^2} \right)
\]
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 (196) holds if the condition in (195) is replaced by
\[
f(y) \leq f(x) + \sigma d_a(x, y), \quad \forall y \in A.
\]

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
\[
\max_{0 \leq t \leq T} \left| s(t; X_0, \ldots, X_{n-1}) - \max_{0 \leq t \leq T} |s(t; Y_0, \ldots, Y_{n-1})| \right|
\leq \max_{0 \leq t \leq T} \left| s(t; X_0, \ldots, X_{n-1}) - s(t; Y_0, \ldots, Y_{n-1}) \right|
\leq \frac{1}{\sqrt{n}} \left| \sum_{i=0}^{n-1} (X_i - Y_i) \exp \left( \frac{j2\pi i t}{T} \right) \right|
\leq \frac{2}{\sqrt{n}} \sum_{i=0}^{n-1} |X_i - Y_i|
\leq \frac{2}{\sqrt{n}} \sum_{i=0}^{n-1} I_{\{x_i \neq y_i\}}
= 2d_a(X, Y)
\]
where
\[
a \triangleq \left( \frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}} \right)
\]
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 16 implies that, for every \( \alpha \geq 0 \),
\[
\mathbb{P}(\|\text{CF}_n(s) - m_n\| \geq \alpha) \leq 4 \exp \left( -\frac{\alpha^2}{16} \right)
\]
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 (199), one obtains the following result.

\textbf{Corollary 6:} 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 (199)

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

Remark 18: This result applies in general to an arbitrary function \( f \) satisfying the condition in (195), where Talagrand’s inequality in (196) implies that (see, e.g., [53, Lemma 4.6])

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

4) Establishing Concentration via McDiarmid’s Inequality: McDiarmid’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

\[ U \triangleq \max_{0 \leq t \leq T} |s(t; X_0, \ldots, X_{i-1}, X_i, \ldots, X_{n-1})| \]
\[ V \triangleq \max_{0 \leq t \leq T} |s(t; X_0, \ldots, X'_i, X_i, \ldots, X_{n-1})| \]

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

\[ |U - V| \leq \max_{0 \leq t \leq T} \left| s(t; X_0, \ldots, X_{i-1}, X_i, \ldots, X_{n-1}) - s(t; X_0, \ldots, X'_i, X_i, \ldots, X_{n-1}) \right| \]
\[ = \max_{0 \leq t \leq T} \frac{1}{\sqrt{n}} \left| (X_{i-1} - X'_i) \exp\left(\frac{j2\pi it}{T}\right) \right| \]
\[ = \frac{|X_{i-1} - X'_i|}{\sqrt{n}} \leq \frac{2}{\sqrt{n}} \]

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

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

(200)

which demonstrates concentration of this measure around its expected value. By comparing (199) with (200), 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 1.

To conclude, 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 [42] and [81]), they establish concentration in a rather simple way and provide some insight to the problem. 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 1 by a factor of 2. Note however that Proposition 1 may be in general tighter than McDiarmid’s inequality (if \( \gamma < \frac{1}{4} \) in the setting of Proposition 1). 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.
H. Random Coding Theorems via Martingale Inequalities

The following subsection 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 subsection is based on [82], [83] and [84] (and mainly on the latest improvements of these achievable rates in [84]).

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 [5, 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 [17, Corollary 2.4.7] and [52]) that will be useful later in this paper.

**Theorem 17:** 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

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

 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 $k \in \{1, \ldots, n\}$. Let $\gamma \triangleq \frac{\sigma^2}{4d}$. Then, for every $t \geq 0$,

$$\mathbb{E}\left[\exp\left(t \sum_{k=1}^n \xi_k\right)\right] \leq \left(\frac{e^{-\gamma td} + \gamma e^{td}}{1 + \gamma}\right)^n. \quad (201)$$

The proof of this theorem relies on Lemma 2.

- Second inequality: The following theorem presents a new martingale inequality that will be useful later in this subsection (the methodology of its proof is similar to the analysis used to prove Theorem 6).

**Theorem 18:** 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

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

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

$$\xi_k \leq d, \quad (202)$$

$$\mathbb{E}[|\xi_k|^l | \mathcal{F}_{k-1}] \leq \mu_l, \quad \forall l \in \{2, \ldots, m\} \quad (203)$$

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

$$\gamma_l \triangleq \frac{\mu_l}{d^l}, \quad \forall l \in \{2, \ldots, m\}. \quad (204)$$
Then, for every \( t \geq 0 \),
\[
\mathbb{E} \left[ \exp \left( t \sum_{k=1}^{n} \xi_k \right) \right] \leq \left( 1 + \sum_{l=2}^{m-1} \frac{(\gamma_l - \gamma_m)(td)^l}{l!} + \gamma_m(e^{td} - 1 - td) \right)^n.
\]
\[\text{(205)}\]

2) Achievable Rates under ML Decoding: The goal of this subsection is to derive achievable rates in the random coding setting under ML decoding. We first review briefly the analysis in [83] 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 [83, 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 subsection.

Consider an ensemble of block codes \( \mathcal{C} \) of length \( N \) and rate \( R \). Let \( \mathcal{C} \subset \mathcal{C} \) be a codebook in the ensemble. The number of codewords in \( \mathcal{C} \) is \( M = \lceil \exp(\nu R) \rceil \). 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 \( y \), there exists another message \( m' \neq m \) such that
\[
|| y - Du_m ||_2 < || y - Du_{m'} ||_2.
\]
The union bound for an AWGN channel implies that
\[
Pe_m(C) \leq \sum_{m' \neq m} Q \left( \frac{|| Du_m - Du_{m'} ||_2}{2\sigma^2} \right)
\]
where
\[
Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp \left( -\frac{t^2}{2} \right) dt, \quad \forall x \in \mathbb{R}
\]
\[\text{(206)}\]
is the complementary Gaussian cumulative distribution function. By using the inequality \( Q(x) \leq \frac{1}{2} \exp \left( -\frac{x^2}{2} \right) \) for \( x \geq 0 \), it gives the loosened bound (by also ignoring the factor of one-half in the bound of \( Q \))
\[
Pe_m(C) \leq \sum_{m' \neq m} \exp \left( -\frac{|| Du_m - Du_{m'} ||_2^2}{8\sigma^2} \right).
\]
At this stage, let us introduce a new parameter \( \rho \in [0, 1] \), and write
\[
Pe_m(C) \leq \sum_{m' \neq m} \exp \left( -\rho \frac{|| Du_m - Du_{m'} ||_2^2}{8\sigma^2} \right).
\]
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
\[
\mathbb{P}_{e|m} \leq \mathbb{E} \left[ \sum_{m' \neq m} \exp \left( -\rho \frac{|| Du_m - Du_{m'} ||_2^2}{8\sigma^2} \right) \right]
\]
and the average ML decoding error probability over the code ensemble and the transmitted message satisfies
\[
\mathbb{P}_{e} \leq (M - 1) \mathbb{E} \left[ \exp \left( -\rho \frac{|| Du - D\tilde{u} ||_2^2}{8\sigma^2} \right) \right]
\]
\[\text{(207)}\]
where the expectation is taken over two randomly chosen codewords \( u \) and \( \tilde{u} \) 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
\[
\mathcal{F}_i = \sigma(U_1, \tilde{U}_1, \ldots, U_i, \tilde{U}_i), \quad \forall i \in \{1, \ldots, N\}
\]
\[\text{(208)}\]
for two randomly selected codewords \( u = (u_1, \ldots, u_N) \), and \( \tilde{u} = (\tilde{u}_1, \ldots, \tilde{u}_N) \) 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

\[
X_k = \mathbb{E}[||D u - D \tilde{u}||_2^2 | \mathcal{F}_k]
\]

(209)
designates the conditional expectation of the squared Euclidean distance between the distorted codewords \( D u \) and \( D \tilde{u} \) given the first \( i \) coordinates of the two codewords \( u \) and \( \tilde{u} \). The first and last entries of this martingale sequence are, respectively, equal to

\[
X_0 = \mathbb{E}[||D u - D \tilde{u}||_2^2], \quad X_N = ||D u - D \tilde{u}||_2^2.
\]

(210)

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

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

and the substitution of the last equality into (207) gives that

\[
\overline{P}_e \leq \exp(NR) \exp \left( -\frac{\rho \mathbb{E}[||D u - D \tilde{u}||_2^2]}{8\sigma^2} \right) \mathbb{E} \left[ \exp \left( -\frac{\rho}{8\sigma^2} \sum_{k=1}^N \xi_k \right) \right].
\]

(211)

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

\[
\mathbb{E}[||D u - D \tilde{u}||_2^2]
\]

\[
= \sum_{k=1}^N \mathbb{E} \left[ (|D u|_k - |D \tilde{u}|_k)^2 \right]
\]

\[
= \sum_{k=1}^N \text{Var} \left( |D u|_k - |D \tilde{u}|_k \right)
\]

\[
= 2 \sum_{k=1}^N \text{Var} \left( |D u|_k \right)
\]

\[
= 2 \left( \sum_{k=1}^{q-1} \text{Var} \left( |D u|_k \right) + \sum_{k=q}^N \text{Var} \left( |D u|_k \right) \right).
\]

Due to the channel model (see Eq. (228)) and the assumption that the symbols \( \{u_i\} \) are i.i.d., it follows that \( \text{Var} \left( |D u|_k \right) \) is fixed for \( k = q, \ldots, N \). Let \( D_v(P) \) designate this common value of the variance (i.e., \( D_v(P) = \text{Var} \left( |D u|_k \right) \) for \( k \geq q \)), then

\[
\mathbb{E}[||D u - D \tilde{u}||_2^2] = 2 \left( \sum_{k=1}^{q-1} \text{Var} \left( |D u|_k \right) + (N - q + 1)D_v(P) \right).
\]

Let

\[
C_\rho(P) \triangleq \exp \left\{ -\frac{\rho}{8\sigma^2} \left( \sum_{k=1}^{q-1} \text{Var} \left( |D u|_k \right) - (q-1)D_v(P) \right) \right\}
\]

which is a bounded constant, under the assumption that \( ||u||_\infty \leq K < +\infty \) 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

\[
\overline{P}_e \leq C_\rho(P) \exp \left\{ -N \left( \frac{\rho D_v(P)}{4\sigma^2} - R \right) \right\} \mathbb{E} \left[ \exp \left( \frac{\rho}{8\sigma^2} \sum_{k=1}^N Z_k \right) \right], \quad \forall \rho \in [0, 1]
\]

(212)

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 (209), it follows that the martingale-difference sequence \( \{Z_k, \mathcal{F}_k\} \) is given by

\[
Z_k = X_{k-1} - X_k
\]

\[
= \mathbb{E}[||D u - D \tilde{u}||_2^2 | \mathcal{F}_{k-1}] - \mathbb{E}[||D u - D \tilde{u}||_2^2 | \mathcal{F}_k].
\]

(213)
For the derivation of improved achievable rates and error exponents (as compared to [83]), the two martingale inequalities presented earlier in this subsection are applied to the obtain two possible exponential upper bounds (in terms of $N$) on the last term on the right-hand side of (212).

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 (212), combined with the exponential martingale inequalities that are introduced in Theorems 17 and 18, one obtains the following bounds:

1) **First Bounding Technique:** From Theorem 17, if

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

holds a.s. for every $k \geq 1$, and $\gamma_2 \triangleq \frac{\sigma^2}{d}$, then it follows from (212) that for every $\rho \in [0,1]$

$$\mathcal{P}_e \leq C_\rho(P) \exp \left\{ -N \left( \frac{\rho D_N(P)}{4\sigma^2} - R \right) \right\} \left( \frac{\exp \left( -\frac{\rho \gamma_2 d}{8\sigma^2} \right) + \gamma_2 \exp \left( \frac{\rho d}{8\sigma^2} \right)}{1 + \gamma_2} \right)^N.$$

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

$$R_1(\sigma^2_\rho) \triangleq \max_P \max_{\rho \in [0,1]} \left\{ \frac{\rho D_N(P)}{4\sigma^2} - \ln \left( \frac{\exp \left( -\frac{\rho \gamma_2 d}{8\sigma^2} \right) + \gamma_2 \exp \left( \frac{\rho d}{8\sigma^2} \right)}{1 + \gamma_2} \right) \right\}$$

where the double maximization is performed over the input distribution $P$ and the parameter $\rho \in [0,1]$. The inner maximization in (214) can be expressed in closed form, leading to the following simplified expression:

$$R_1(\sigma^2_\rho) = \max_P \left\{ D \left( \left. \gamma_2 \frac{d}{1+\gamma_2} + \frac{2D_N(P)}{d(1+\gamma_2)} \right\| \gamma_2 \right) \right\}, \quad \text{if} \quad D_N(P) < \frac{\gamma_2 d \exp \left( \frac{(1+\gamma_2)}{8\sigma^2} \right) - 1}{2 \left( 1 + \gamma_2 \exp \left( \frac{d(1+\gamma_2)}{8\sigma^2} \right) \right)}$$

(215)

where

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

(216)
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 18 and Eq. (212), we derive in the following a second achievable rate for random coding under ML decoding. Referring to the martingale-difference sequence $\{Z_k, \mathcal{F}_k\}$ in Eqs. (208) and (213), one obtains from Eq. (212) that if for some even number $m \in \mathbb{N}$

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

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 $\rho \in [0,1]$,

$$\mathcal{P}_e \leq C_\rho(P) \exp \left\{ -N \left( \frac{\rho D_N(P)}{4\sigma^2} - R \right) \right\} \left[ 1 + \sum_{l=2}^{m-1} \frac{\gamma_l - \gamma_m}{l!} \left( \frac{\rho d}{8\sigma^2} \right)^l + \gamma_m \left( \exp \left( \frac{\rho d}{8\sigma^2} \right) - 1 - \frac{\rho d}{8\sigma^2} \right) \right]^N.$$

This gives the following achievable rate, for an arbitrary even number $m \in \mathbb{N}$,

$$R_2(\sigma^2_\rho) \triangleq \max_P \max_{\rho \in [0,1]} \left\{ \frac{\rho D_N(P)}{4\sigma^2} - \ln \left( 1 + \sum_{l=2}^{m-1} \frac{\gamma_l - \gamma_m}{l!} \left( \frac{\rho d}{8\sigma^2} \right)^l + \gamma_m \left( \exp \left( \frac{\rho d}{8\sigma^2} \right) - 1 - \frac{\rho d}{8\sigma^2} \right) \right) \right\}$$

(217)

where, similarly to (214), the double maximization in (217) 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 \( \nu_i \sim \mathcal{N}(0, \sigma^2) \) is an additive Gaussian noise with zero mean and variance \( \sigma^2 \). Since the codewords \( U = (U_1, \ldots, U_N) \) and \( \bar{U} = (\bar{U}_1, \ldots, \bar{U}_N) \) are independent and their symbols are i.i.d., let

\[ P(U_k = A) = P(\bar{U}_k = A) = \alpha, \quad P(U_k = -A) = P(\bar{U}_k = -A) = 1 - \alpha \]

for some \( \alpha \in [0, 1] \). Since the channel is memoryless and the all the symbols are i.i.d. then one gets from (208) and (213) that

\[
Z_k = \mathbb{E}[\|U - \bar{U}\|^2_2 \mid \mathcal{F}_{k-1}] - \mathbb{E}[\|U - \bar{U}\|^2_2 \mid \mathcal{F}_k]
\]

\[
= \sum_{j=1}^{k-1} (U_j - \bar{U}_j)^2 + \sum_{j=k}^{N} \mathbb{E}[(U_j - \bar{U}_j)^2] - \sum_{j=1}^{k} (U_j - \bar{U}_j)^2 - \sum_{j=k+1}^{N} \mathbb{E}[(U_j - \bar{U}_j)^2]
\]

\[
= \mathbb{E}[(U_k - \bar{U}_k)^2] - (U_k - \bar{U}_k)^2
\]

\[
= \alpha(1 - \alpha)(-2A)^2 + \alpha(1 - \alpha)(2A)^2 - (U_k - \bar{U}_k)^2
\]

\[
= 8\alpha(1 - \alpha)A^2 - (U_k - \bar{U}_k)^2.
\]

Hence, for every \( k \),

\[
Z_k \leq 8\alpha(1 - \alpha)A^2 \triangleq d. \tag{218}
\]

Furthermore, for every \( k, l \in \mathbb{N} \), due to the above properties

\[
\mathbb{E}[Z_k^l \mid \mathcal{F}_{k-1}]
\]

\[
= \mathbb{E}[Z_k^l]
\]

\[
= \mathbb{E}\left[\left(8\alpha(1 - \alpha)A^2 - (U_k - \bar{U}_k)^2\right)^l\right]
\]

\[
= \left[1 - 2\alpha(1 - \alpha)\right] (8\alpha(1 - \alpha)A^2)^l + 2\alpha(1 - \alpha) (8\alpha(1 - \alpha)A^2 - 4A^2)^l \triangleq \mu_l \tag{219}
\]

and therefore, from (218) and (219), for every \( l \in \mathbb{N} \)

\[
\gamma_l \triangleq \frac{\mu_l}{d^l} = \left[1 - 2\alpha(1 - \alpha)\right] \left[1 + (-1)^l \left(\frac{1 - 2\alpha(1 - \alpha)}{2\alpha(1 - \alpha)}\right)^{l-1}\right]. \tag{220}
\]

Let us now rely on the two achievable rates for random coding in Eqs. (215) and (217), 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 (218) and (220) that

\[
D_\gamma(P) = \text{Var}(U_k) = A^2, \quad d = 2A^2, \quad \gamma_l = \frac{1 + (-1)^l}{2}, \quad \forall l \in \mathbb{N}. \tag{221}
\]

Based on the first bounding technique that leads to the achievable rate in Eq. (215), since the first condition in this equation cannot hold for the set of parameters in (221) then the achievable rate in this equation is equal to

\[
R_1(\sigma^2) = \frac{A^2}{4\sigma^2} - \ln \cosh \left(\frac{A^2}{4\sigma^2}\right)
\]
in units of nats per channel use. Let \( \text{SNR} \triangleq \frac{A^2}{\sigma_v^2} \) designate the signal to noise ratio, then the first achievable rate gets the form

\[
R'_1(\text{SNR}) = \frac{\text{SNR}}{4} - \ln \left( \frac{\text{SNR}}{4} \right). \tag{222}
\]

It is observed here that the optimal value of \( \rho \) in (215) is equal to 1 (i.e., \( \rho^* = 1 \)).

Let us compare it in the following with the achievable rate that follows from (217). Let \( m \in \mathbb{N} \) be an even number. Since, from (221), \( \gamma_l = 1 \) for all even values of \( l \in \mathbb{N} \) and \( \gamma_l = 0 \) for all odd values of \( l \in \mathbb{N} \), then

\[
1 + \sum_{l=2}^{m-1} \frac{\gamma_l - \gamma_m}{l!} \left( \frac{\rho d}{8\sigma_v^2} \right)^l + \gamma_m \left( \exp \left( \frac{\rho d}{8\sigma_v^2} \right) - 1 - \frac{\rho d}{8\sigma_v^2} \right) = 1 - \sum_{l=1}^{\frac{m-1}{2}} \frac{1}{(2l+1)!} \left( \frac{\rho d}{8\sigma_v^2} \right)^{2l+1} + \left( \exp \left( \frac{\rho d}{8\sigma_v^2} \right) - 1 - \frac{\rho d}{8\sigma_v^2} \right) \tag{223}
\]

Since the infinite sum \( \sum_{l=1}^{\infty} \frac{1}{(2l+1)!} \left( \frac{\rho d}{8\sigma_v^2} \right)^{2l+1} \) is monotonically increasing with \( m \) (where \( m \) is even and \( \rho \in [0, 1] \)), then from (217), 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

\[
\lim_{m \to \infty} \left( 1 + \sum_{l=2}^{m-1} \frac{\gamma_l - \gamma_m}{l!} \left( \frac{\rho d}{8\sigma_v^2} \right)^l + \gamma_m \left( \exp \left( \frac{\rho d}{8\sigma_v^2} \right) - 1 - \frac{\rho d}{8\sigma_v^2} \right) \right) = 1 - \sum_{l=1}^{\infty} \frac{1}{(2l+1)!} \left( \frac{\rho d}{8\sigma_v^2} \right)^{2l+1} + \left( \exp \left( \frac{\rho d}{8\sigma_v^2} \right) - 1 - \frac{\rho d}{8\sigma_v^2} \right) \tag{224}
\]

where equality (a) follows from (223), 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 (217) gives from (221), \( \frac{d}{8\sigma_v^2} = \frac{A^2}{4\sigma_v^2} \)

\[
R_2(\sigma_v^2) = \max_{\rho \in [0,1]} \left( \frac{\rho A^2}{4\sigma_v^2} - \ln \cosh \left( \frac{\rho A^2}{4\sigma_v^2} \right) \right).
\]

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. (217) is equal to

\[
R_2(\sigma_v^2) = \frac{A^2}{4\sigma_v^2} - \ln \cosh \left( \frac{A^2}{4\sigma_v^2} \right)
\]

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 \( \text{SNR} = \frac{A^2}{\sigma_v^2} \), gives the achievable rate in (222), so the first and second achievable rates for the binary-input AWGN channel coincide, i.e.,

\[
R'_1(\text{SNR}) = R_2(\sigma_v^2) = \frac{\text{SNR}}{4} - \ln \cosh \left( \frac{\text{SNR}}{4} \right). \tag{225}
\]

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 (225), 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., [63, Example 4.38 on p. 194])

$$C(\text{SNR}) = \ln 2 + (2 \text{SNR} - 1) Q(\sqrt{\text{SNR}}) - \sqrt{\frac{2 \text{SNR}}{\pi}} \exp\left(-\frac{\text{SNR}}{2}\right)$$

$$+ \sum_{i=1}^{\infty} \left\{ (-1)^i \frac{1}{i(i+1)} \cdot \exp(2i(i+1) \text{SNR}) \cdot Q((1+2i) \sqrt{\text{SNR}}) \right\}$$

(226)

where the $Q$-function that appears in the infinite series on the right-hand side of (226) is the complementary Gaussian cumulative distribution function in (206). 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 (225) that follows from the martingale approach is shown in Figure 2.

![Figure 2](image-url)

**Fig. 2.** A comparison between the symmetric i.i.d. mutual information of the binary-input AWGN channel (solid line) and the common achievable rate in (225) (dashed line) that follows from the martingale approach in this subsection.

From the discussion in this subsection, the first and second bounding techniques in Section VII-H2 lead to the same achievable rate (see (225)) 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 (221), 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 (223) and (224)). 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 $l > 2$.

2) **Nonlinear Channels with Memory - Third-Order Volterra Channels:** The channel model is first presented in the following (see Figure 3). We refer in the following to a discrete-time channel model of nonlinear Volterra channels where the input-output channel model is given by

$$y_i = [D u]_i + \nu_i$$

(227)

where $i$ is the time index. Volterra’s operator $D$ of order $L$ and memory $q$ is given by

$$[D u]_i = h_0 + \sum_{j=1}^{L} \sum_{i_1=0}^{q} \cdots \sum_{i_j=0}^{q} h_j(i_1, \ldots, i_j) u_{i-i_1} \cdots u_{i-i_j}.$$

(228)
and \( \nu \) is an additive Gaussian noise vector with i.i.d. entries \( \nu_i \sim \mathcal{N}(0, \sigma^2_\nu) \).

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 I. 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 subsection requires the numerical calculation of the parameters \( d \) and \( \sigma^2 \) and thus of \( \gamma_2 \) for the martingale \( \{ Z_i, \mathcal{F}_i \}_{i=0}^\infty \). In order to achieve this goal, we have to calculate \( |Z_i - Z_{i-1}| \) and \( \operatorname{Var}(Z_i | \mathcal{F}_{i-1}) \) for all possible combinations of the input samples which contribute to the aforementioned expressions. Thus, the analytic calculation of \( d \) and \( \gamma_2 \) increases as the system’s memory \( q \) increases. Numerical results are provided in Figure 4 for the case where \( \sigma^2_\nu = 1 \). The new achievable rates \( R_{1}^{(2)}(D_1, A, \sigma^2_\nu) \) and \( R_{2}^{(2)}(D_1, A, \sigma^2_\nu) \), which depend on the channel input parameter \( A \), are compared to the achievable rate provided in [83, Fig. 2] and are shown to be larger than the latter.

![Fig. 3. The discrete-time Volterra non-linear channel model in Eqs. (227) and (228) 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 \( \sigma^2_\nu \).](image)

![Fig. 4. Comparison of the achievable rates in this subsection \( R_{1}(D_1, A, \sigma^2_\nu) \) and \( R_{2}^{(2)}(D_1, A, \sigma^2_\nu) \) (where \( m = 2 \)) with the bound \( R_{p}(D_1, A, \sigma^2_\nu) \) of [83, Fig.2] for the nonlinear channel with kernels depicted in Table I and noise variance \( \sigma^2_\nu = 1 \). Rates are expressed in nats per channel use.](image)

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 [22] and [23]), combined with a possible tightening of the union bound under ML decoding (see, e.g., [65]).
VIII. 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 concentration inequalities from the literature. The second part of this work exemplifies these martingale 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 chapter.

APPENDIX A

PROOF OF PROPOSITION 1

Let \( \{X_k, \mathcal{F}_k\}_{k=0}^{\infty} \) be a discrete-parameter martingale. We prove in the following that Theorem 5 implies (75). Let \( \{X_k, \mathcal{F}_k\}_{k=0}^{\infty} \) be a discrete-parameter martingale that satisfies the conditions in Theorem 5. From (33)

\[
\mathbb{P}(|X_n - X_0| \geq \alpha \sqrt{n}) \leq 2 \exp \left( -n D \left( \frac{\delta' + \gamma}{1 + \gamma} \right) \right) \tag{229}
\]

where from (52)

\[
\delta' \triangleq \frac{\alpha}{\sqrt{n}} = \frac{\delta}{\sqrt{n}}. \tag{230}
\]

From the right-hand side of (229)

\[
D \left( \frac{\delta' + \gamma}{1 + \gamma} \right) = \frac{\gamma}{1 + \gamma} \left[ \left(1 + \frac{\delta}{\gamma \sqrt{n}}\right) \ln \left(1 + \frac{\delta}{\gamma \sqrt{n}}\right) + \frac{1}{\gamma} \left(1 - \frac{\delta}{\gamma \sqrt{n}}\right) \ln \left(1 - \frac{\delta}{\gamma \sqrt{n}}\right) \right]. \tag{231}
\]

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 (231) that for every \( n > \frac{\delta^2}{\gamma^2} \)

\[
n D \left( \frac{\delta' + \gamma}{1 + \gamma} \right) = \frac{\delta^2}{2 \gamma} - \frac{\delta^2 (1 - \gamma)}{6 \gamma^2} \frac{1}{\sqrt{n}} + \ldots
\]

\[
= \frac{\delta^2}{2 \gamma} + O \left( \frac{1}{\sqrt{n}} \right). \tag{232}
\]

Substituting this into the exponent on the right-hand side of (229) gives (75).

APPENDIX B

ANALYSIS RELATED TO THE MODERATE DEVIATIONS PRINCIPLE IN SECTION VI-C

It is demonstrated in the following that, in contrast to Azuma’s inequality, Theorem 5 provides an upper bound on

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

which coincides with the exact asymptotic limit in (109). 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 \( \alpha \geq 0 \),
\[
\mathbb{P}( |S_n| \geq \alpha n^\eta ) \leq 2 \exp \left( -\frac{\alpha^2 n^{2\eta - 1}}{2d^2} \right)
\]
and therefore
\[
\lim_{n \to \infty} n^{1-2\eta} \ln \mathbb{P}( |S_n| \geq \alpha n^\eta ) \leq -\frac{\alpha^2}{2d^2}.
\] (232)

This differs from the limit in (109) where \( \sigma^2 \) is replaced by \( d^2 \), so Azuma’s inequality does not provide the asymptotic limit in (109) (unless \( \sigma^2 = d^2 \), i.e., \( |X_k| = d \) a.s. for every \( k \)).

2) Analysis related to Theorem 5: The analysis here is a slight modification of the analysis in Appendix A with the required adaptation of the calculations for \( \eta \in (\frac{1}{2}, 1) \). It follows from Theorem 5 that, for every \( \alpha \geq 0 \),
\[
\mathbb{P}( |S_n| \geq \alpha n^\eta ) \leq 2 \exp \left( -n D\left( \frac{\delta' + \gamma}{1 + \gamma}, \frac{\gamma}{1 + \gamma} \right) \right)
\]
where \( \gamma \) is introduced in (52), and \( \delta' \) in (230) is replaced with
\[
\delta' \triangleq \frac{\alpha}{d} = \delta n^{-(1-\eta)}
\] (233)
due to the definition of \( \delta \) in (52). Following the same analysis as in Appendix A, it follows that for every \( n \in \mathbb{N} \)
\[
\mathbb{P}( |S_n| \geq \alpha n^\eta ) \leq 2 \exp \left( -\frac{\delta^2 n^{2\eta - 1}}{2\gamma} \left[ 1 + \frac{\alpha(1 - \gamma)}{3\gamma d} \cdot n^{-(1-\eta)} + \ldots \right] \right)
\]
and therefore (since, from (52), \( \frac{\delta^2}{\gamma} = \frac{\sigma^2}{\sigma^2} \))
\[
\lim_{n \to \infty} n^{1-2\eta} \ln \mathbb{P}( |S_n| \geq \alpha n^\eta ) \leq -\frac{\alpha^2}{2\sigma^2}.
\] (234)

Hence, this upper bound coincides with the exact asymptotic result in (109).

APPENDIX C

PROOF OF PROPOSITION 2

The proof of (164) is based on calculus, and it is similar to the proof of the limit in (163) that relates the divergence and Fisher information. For the proof of (166), note that
\[
C(P_0, P_{\theta'}) \geq E_L(P_0, P_{\theta'}) \geq \min_{i=1,2} \left\{ \frac{\delta_i^2}{2\gamma_i} - \frac{\delta_i^3}{6\gamma_i^2 (1 + \gamma_i)} \right\}.
\] (235)
The left-hand side of (235) holds since \( E_L \) is a lower bound on the error exponent, and the exact value of this error exponent is the Chernoff information. The right-hand side of (235) follows from Lemma 7 (see (161)) and the definition of \( E_L \) in (165). By definition \( \gamma_i \triangleq \frac{\sigma_i^2}{d_i^2} \) and \( \delta_i \triangleq \frac{\delta_i}{d_i} \), where, based on (151),
\[
\varepsilon_1 \triangleq D(P_0 || P_{\theta'}), \quad \varepsilon_2 \triangleq D(P_{\theta'} || P_\theta).
\] (236)
The term on the left-hand side of (235) therefore satisfies
\[
\frac{\delta_i^2}{2\gamma_i} - \frac{\delta_i^3}{6\gamma_i^2 (1 + \gamma_i)} = \frac{\varepsilon_1^2}{2\sigma_i^2} - \frac{\varepsilon_2^3 d_i^2}{6\sigma_i^2 (\sigma_i^2 + d_i^2)} \geq \frac{\varepsilon_1^2}{2\sigma_i^2} \left( 1 - \frac{\varepsilon_2 d_i}{3} \right)
\]
so it follows from (235) and the last inequality that
\[
C(P_0, P_{\theta'}) \geq E_L(P_0, P_{\theta'}) \geq \min_{i=1,2} \left\{ \frac{\varepsilon_1^2}{2\sigma_i^2} \left( 1 - \frac{\varepsilon_2 d_i}{3} \right) \right\}.
\] (237)
Based on the continuity assumption of the indexed family \( \{P_\theta\}_{\theta \in \Theta} \), then it follows from (236) that

\[
\lim_{\theta' \to \theta} \varepsilon_i = 0, \quad \forall i \in \{1, 2\}
\]

and also, from (132) and (142) 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 (164) and (237) that

\[
\frac{J(\theta)}{8} \geq \lim_{\theta' \to \theta} \frac{E_L(P_\theta, P_{\theta'})}{(\theta - \theta')^2} \geq \lim_{\theta' \to \theta} \min_{i=1,2} \left\{ \frac{\varepsilon_i^2}{2\sigma_i^2(\theta - \theta')^2} \right\}.
\]

(238)

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 \( \frac{J(\theta)}{8} \).

\[
\lim_{\theta' \to \theta} \frac{\varepsilon_i^2}{2\sigma_i^2(\theta - \theta')^2}
\]

\[
= \lim_{\theta' \to \theta} \frac{D(P_\theta||P_{\theta'})^2}{2\sigma_i^2(\theta - \theta')^2}
\]

\[
\lim_{\theta' \to \theta} \frac{D(P_\theta||P_{\theta'})}{4\theta' - \theta} \sum_{x \in X} P_\theta(x) \left( \ln \frac{P_\theta(x)}{P_{\theta'}(x)} \right)^2 - D(P_\theta||P_{\theta'})^2
\]

\[
= \frac{J(\theta)}{8} \lim_{\theta' \to \theta} \sum_{x \in X} P_\theta(x) \left( \ln \frac{P_\theta(x)}{P_{\theta'}(x)} \right)^2 - D(P_\theta||P_{\theta'})^2
\]

\[
= \frac{J(\theta)}{8} \lim_{\theta' \to \theta} \sum_{x \in X} P_\theta(x) \left( \ln \frac{P_\theta(x)}{P_{\theta'}(x)} \right)^2
\]

\[
= \frac{J(\theta)}{8}
\]

(239)

where equality (a) follows from (236), equalities (b), (e) and (f) follow from (163), equality (c) follows from (133) 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 (162). Similarly, also

\[
\lim_{\theta' \to \theta} \frac{\varepsilon_i^2}{2\sigma_i^2(\theta - \theta')^2} = \frac{J(\theta)}{8}
\]

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 (238) that \( \frac{E_L(P_\theta, P_{\theta'})}{(\theta - \theta')^2} = \frac{J(\theta)}{8} \). This completes the proof of (166).

We prove now equation (168). From (132), (142), (151) and (167) then

\[
\bar{E}_L(P_\theta, P_{\theta'}) = \min_{i=1,2} \frac{\varepsilon_i^2}{2d_i^2}
\]
with $\varepsilon_1$ and $\varepsilon_2$ in (236). Hence,
\[
\lim_{\theta' \to \theta} \frac{\tilde{E}_\lambda(P_\theta, P_{\theta'})}{(\theta' - \theta)^2} \leq \lim_{\theta' \to \theta} \frac{\varepsilon_1^2}{2\sigma_1^2(\theta' - \theta)^2}
\]
and from (239) and the last inequality, it follows that
\[
\lim_{\theta' \to \theta} \tilde{E}_\lambda(P_\theta, P_{\theta'}) \leq \frac{J(\theta)}{8} \lim_{\theta' \to \theta} \frac{\sigma_1^2}{\sigma_1^2} \sum_{x \in \mathcal{X}} P_\theta(x) \left( \frac{\ln P_\theta(x) - D(P_\theta \| P_{\theta'})}{\max_{x \in \mathcal{X}} \left| \frac{\ln P_\theta(x)}{P_{\theta'}(x)} - D(P_\theta \| P_{\theta'}) \right|} \right)^2.
\]

It is clear that the second term on the right-hand side of (240) 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 (240) 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
\[
\lim_{\theta' \to \theta} \sum_{x \in \mathcal{X}} P_\theta(x) \left( \frac{\ln P_\theta(x) - D(P_\theta \| P_{\theta'})}{\max_{x \in \mathcal{X}} \left| \frac{\ln P_\theta(x)}{P_{\theta'}(x)} - D(P_\theta \| P_{\theta'}) \right|} \right)^2 = (1 - \alpha)\theta
\]
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 (168), and also the proof of Proposition 2.

**APPENDIX D**

**PROOF OF LEMMA 8**

In order to prove Lemma 8, one needs to show that if $\rho'(1) < \infty$ then
\[
\lim_{C \to 1} \sum_{i=1}^{\infty} (i + 1)^2 \Gamma_i \left[ h_2 \left( \frac{1 - C^2}{2} \right) \right]^2 = 0
\]
which then yields from (183) that $B \to \infty$ in the limit where $C \to 1$.

By the assumption in Lemma 8 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
\[
d^\text{avg}_c = \frac{1}{\sum_{i=1}^{\infty} \rho_i} < \infty.
\]
The infinite sum \( \sum_{i=1}^{\infty} (i+1)^2 \Gamma_i \) converges under the above assumption since
\[
\sum_{i=1}^{\infty} (i+1)^2 \Gamma_i = \sum_{i=1}^{\infty} i^2 \Gamma_i + 2 \sum_{i=1}^{\infty} i \Gamma_i + \sum_{i=1}^{\infty} \Gamma_i = d^\text{avg}_c \left( \sum_{i=1}^{\infty} i \rho_i + 2 \right) + 1 < \infty,
\]
where the last equality holds since
\[
\Gamma_i = \frac{\rho_i}{\int_0^1 \rho(x) \, dx} = d^\text{avg}_c \left( \frac{\rho_i}{\gamma} \right), \quad \forall \, i \in \mathbb{N}.
\]
The infinite series in (241) 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 (241) converges to zero in the limit where \( C \to 1 \), hence the limit in (241) is zero. This completes the proof of Lemma 8.

**APPENDIX E**

**PROOF OF THE PROPERTIES IN (193) FOR OFDM SIGNALS**

Consider an OFDM signal from Section VII-G. The sequence in (191) is a martingale due to basic properties of martingales. From (190), for every \( i \in \{0, \ldots, n\} \)
\[
Y_i = \mathbb{E} \left[ \max_{0 \leq t \leq T} |s(t; X_0, \ldots, X_{n-1})| \bigg| X_0, \ldots, X_{i-1} \right].
\]
The conditional expectation for the RV \( Y_{i-1} \) refers to the case where only \( X_0, \ldots, X_{i-2} \) are revealed. Let \( X'_{i-1} \) and \( X_{i-1} \) be independent copies, which are also independent of \( X_0, \ldots, X_{i-2}, X_i, \ldots, X_{n-1} \). Then, for every \( 1 \leq i \leq n \),
\[
Y_{i-1} = \mathbb{E} \left[ \max_{0 \leq t \leq T} |s(t; X_0, \ldots, X'_{i-1}, X_i, \ldots, X_{n-1})| \bigg| X_0, \ldots, X_{i-2} \right] = \mathbb{E} \left[ \max_{0 \leq t \leq T} |s(t; X_0, \ldots, X'_{i-1}, X_i, \ldots, X_{n-1})| \bigg| X_0, \ldots, X_{i-2}, X_{i-1} \right].
\]
Since \( |\mathbb{E}(Z)| \leq \mathbb{E}(|Z|) \), then for \( i \in \{1, \ldots, n\} \)
\[
|Y_i - Y_{i-1}| \leq \mathbb{E}_{X'_{i-1}, X_i, \ldots, X_{n-1}} \left[ |U - V| \bigg| X_0, \ldots, X_{i-1} \right] \tag{242}
\]
where
\[
U \triangleq \max_{0 \leq t \leq T} |s(t; X_0, \ldots, X_{i-1}, X_i, \ldots, X_{n-1})| \\
V \triangleq \max_{0 \leq t \leq T} |s(t; X_0, \ldots, X'_{i-1}, X_i, \ldots, X_{n-1})|.
\]
From (188)
\[
|U - V| \leq \max_{0 \leq t \leq T} |s(t; X_0, \ldots, X_{i-1}, X_i, \ldots, X_{n-1}) - s(t; X_0, \ldots, X'_{i-1}, X_i, \ldots, X_{n-1})| = \max_{0 \leq t \leq T} \frac{1}{\sqrt{n}} \left| (X_{i-1} - X'_{i-1}) \exp \left( \frac{j2\pi i t}{T} \right) \right| = \frac{|X_{i-1} - X'_{i-1}|}{\sqrt{n}}. \tag{243}
\]
By assumption, \(|X_{i-1}| = |X'_{i-1}| = 1\), and therefore a.s.

\[ |X_{i-1} - X'_{i-1}| \leq 2 \implies |Y_i - Y_{i-1}| \leq \frac{2}{\sqrt{n}}. \]

In the following, an upper bound on the conditional variance \(\text{Var}(Y_i | F_{i-1}) = \mathbb{E}[(Y_i - Y_{i-1})^2 | F_{i-1}]\) is obtained. Since \((\mathbb{E}(Z))^2 \leq \mathbb{E}(Z^2)\) for a real-valued RV \(Z\), then from (242) and (243)

\[
\mathbb{E}[(Y_i - Y_{i-1})^2 | F_{i-1}] \leq \frac{1}{n} \cdot \mathbb{E}_{X_{i-1}}[|X_{i-1} - X'_{i-1}|^2 | F_i]
\]

where \(F_i\) is the \(\sigma\)-algebra that is generated by \(X_0, \ldots, X_{i-1}\). Due to symmetry of the PSK constellation, then

\[
\mathbb{E}[(Y_i - Y_{i-1})^2 | F_{i-1}] \\
\leq \frac{1}{n} \cdot \mathbb{E}_{X_{i-1}}[|X_{i-1} - X'_{i-1}|^2 | F_i] \\
= \frac{1}{n} \mathbb{E}[|X_{i-1} - X'_{i-1}|^2 | X_0, \ldots, X_{i-1}] \\
= \frac{1}{n} \mathbb{E}[|X_{i-1} - X'_{i-1}|^2 | X_{i-1}] \\
= \frac{1}{n} \mathbb{E}[|X_{i-1} - X'_{i-1}|^2 | X_{i-1} = e^{j\pi}] \\
= \frac{1}{nM} \sum_{l=0}^{M-1} \left| e^{j\pi} - e^{j(2l+1)\pi/M} \right|^2 \\
= \frac{4}{nM} \sum_{l=1}^{M-1} \sin^2 \left( \frac{\pi l}{M} \right) = \frac{2}{n}
\]

where the last equality holds since

\[
\sum_{l=1}^{M-1} \sin^2 \left( \frac{\pi l}{M} \right) = \frac{M}{2} - \frac{1}{2} \Re \left\{ \sum_{l=0}^{M-1} e^{j2l\pi/M} \right\} \\
= \frac{M}{2} - \frac{1}{2} \Re \left\{ \frac{1 - e^{j2\pi}}{1 - e^{j2\pi/M}} \right\} = \frac{M}{2}.
\]

REFERENCES
