IRWIN AND JOAN JACOBS
CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGIES

Concentration of Measure Inequalities in Information Theory, Communications and Coding

Maxim Raginsky and Igal Sason

CCIT Report #823
December 2012

Electronics
Computers
Communications

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

CCIT Report #823 December 2012

Concentration of Measure Inequalities in Information Theory, Communications and Coding

TUTORIAL

Submitted to the **Foundations and Trends in Communications and Information Theory**
December 2012

Maxim Raginsky

Department of Electrical and Computer Engineering,
Coordinated Science Laboratory,
University of Illinois at Urbana-Champaign,
Urbana, IL 61801, USA.
E-mail: maxim@illinois.edu

and

Igal Sason

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

Abstract

Concentration inequalities have been the subject of exciting developments during the last two decades, and they have been intensively studied and used as a powerful tool in various areas. These include convex geometry, functional analysis, statistical physics, statistics, pure and applied probability theory (e.g., concentration of measure phenomena in random graphs, random matrices and percolation), information theory, learning theory, dynamical systems and randomized algorithms.

This tutorial article is focused on some of the key modern mathematical tools that are used for the derivation of concentration inequalities, on their links to information theory, and on their various applications to communications and coding.

The first part of this article introduces some classical concentration inequalities for martingales, and it also derives some recent refinements of these inequalities. The power and versatility of the martingale approach is exemplified in the context of binary hypothesis testing, codes defined on graphs and iterative decoding algorithms, and some other aspects that are related to wireless communications and coding.

The second part of this article introduces the entropy method for deriving concentration inequalities for functions of many independent random variables, and it also exhibits its multiple connections to information theory. The basic ingredients of the entropy method are discussed first in conjunction with the closely related topic of logarithmic Sobolev inequalities, which are typical of the so-called functional approach to studying concentration of measure phenomena. The discussion on logarithmic Sobolev inequalities is complemented by a related viewpoint based on probability in metric spaces. This viewpoint centers around the so-called transportation-cost inequalities, whose roots are in information theory. Some representative results on concentration for dependent random variables are briefly summarized, with emphasis on their connections to the entropy method. Finally, the tutorial addresses several applications of the entropy method and related information-theoretic tools to problems in communications and coding. These include strong converses for several source and channel coding problems, empirical distributions of good channel codes with non-vanishing error probability, and an information-theoretic converse for concentration of measure.

Contents

Chapter 1

Introduction

Inequalities providing upper bounds on probabilities of the type $\mathbb{P}(|X - \overline{x}| \geq t)$ (or $\mathbb{P}(X - \overline{x} \geq t)$ for a random variable (RV) X, where $\overline{x}$ denotes the expectation or median of X, have been among the main tools of probability theory. These inequalities are known as concentration inequalities, and they have been the subject of 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" [1]. The exact meaning of such a statement clearly needs to be clarified rigorously, but it will often mean that such a random variable X concentrates around $\overline{x}$ in a way that the probability of the event $\{|X - \overline{x}| > t\}$ decays exponentially in t (for $t \geq 0$). Detailed treatments of concentration of measure, including historical accounts, can be found, e.g., in [2], [3], [4], [5] and [6].

In recent years, concentration inequalities have been intensively studied and used as a powerful tool in various areas such as convex geometry, functional analysis, statistical physics, dynamical systems, pure and applied 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 measure phenomena. These include:

- The martingale approach (see, e.g., [5], [7], [8], [9, Chapter 7], [10] and [11]) with its various information-theoretic aspects (see, e.g., [12]). This methodology is covered in Chapter 2, which focuses on concentration inequalities for discrete-time martingales with bounded jumps, and on some of their potential applications in information theory, coding and communications.

- The entropy method and logarithmic Sobolev inequalities (see, e.g., [2, Chapter 5], [3] and references therein), and their information-theoretic aspects. This methodology and its remarkable information-theoretic links will be considered in Chapter 3.

- Transportation-cost inequalities that originated from information theory (see, e.g., [2, Chapter 6], [13], and references therein). This methodology and its information-theoretic aspects will be considered in Chapter 3, with a discussion of the relation between transportation-cost inequalities to the entropy method and logarithmic Sobolev inequalities.

- Talagrand's inequalities for product measures (see, e.g., [1], [5, Chapter 4], [6] and [14, Chapter 6]) and their information-theoretic applications (see, e.g., [15] and [16]). We do not discuss Talagrand's inequalities in detail.

- Stein's method is used to prove concentration inequalities, a.k.a. concentration inequalities with exchangeable pairs (see, e.g., [17], [18], [19] and [20]). This relatively recent framework is not addressed in this work.

- Concentration inequalities that follow from rigorous methods in statistical physics (see, e.g., [21, 22, 23, 24]). These methods are not addresses either in this work.

We now give a synopsis of some of the main ideas underlying the martingale approach (Chapter 2) and the entropy method (Chapter 3). Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function that is characterized by bounded differences whenever the n-dimensional vectors differ in only one coordinate. A common method for proving 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 [5, Theorem 3.1]). This inequality was proved (with some possible extensions) via the martingale approach (see [5, 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., [5, 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 [8] who proved this inequality for a sum of independent and bounded random variables, and Azuma [7] later extended it to bounded-difference martingales. It is noted that the Azuma-Hoeffding inequality for a bounded martingale-difference sequence was extended to centering sequences with bounded differences [25]; 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 [26] in order to prove concentration, around the expected value, of the chromatic number for random graphs. The chromatic number of a graph is defined to be the minimal number of colors that is required to color all the vertices of this graph so that no two vertices which are connected by an edge have the same color, and the ensemble for which concentration was demonstrated in [26] was the ensemble of random graphs with n vertices such that any ordered pair of vertices in the graph is connected by an edge with a fixed probability p for some $p \in (0, 1)$. It is noted that the concentration result in [26] 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 [27], [28] and [29]. During the last decade, the Azuma-Hoeffding inequality has been extensively used for proving concentration of measures in coding theory (see, e.g., [12] 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 [30] 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 [31], [32] and [33], 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).

However, as pointed out by Talagrand [1], "for all its qualities, the martingale method has a great drawback: it does not seem to yield results of optimal order in several key situations. In particular, it seems unable to obtain even a weak version of concentration of measure phenomenon in Gaussian space." In Chapter 3 of this tutorial, we focus on another set of techniques, fundamentally rooted in information theory, that provide very strong concentration inequalities. These techniques, commonly referred to as the *entropy method*, have originated in the work of Michel Ledoux [34], who found an alternative route to a class of concentration inequalities for product measures originally derived by Talagrand [6] using an ingenious inductive technique. Specifically, Ledoux noticed that the well-known Chernoff bounding trick, which is discussed in detail in Section 3.1 and which expresses the deviation probability of the form $\mathbb{P}(|X - \bar{x}| > t)$ (for an arbitrary $t > 0$) in terms of the moment-generating function (MGF) $\mathbb{E}[\exp(\lambda X)]$, can be combined with the so-called *logarithmic Sobolev inequalities*, which can be used to control the MGF in terms of the relative entropy.

Perhaps the best-known log-Sobolev inequality, first explicitly referred to as such by Leonard Gross

[35], pertains to the standard Gaussian distribution in Euclidean space $\mathbb{R}^n$, and bounds the relative entropy $D(P\|G_n)$ between an arbitrary probability distribution P on $\mathbb{R}^n$ and the standard Gaussian measure G_n by an "energy-like" quantity related to the squared norm of the gradient of the density of P w.r.t. G_n (here, it can be assumed without loss of generality that P is absolutely continuous w.r.t. G_n, for otherwise both sides of the log-Sobolev inequality are equal to $+\infty$). Using a clever analytic argument which he attributed to an unpublished note by Ira Herbst, Gross has used his log-Sobolev inequality to show that the logarithmic MGF $\Lambda(\lambda) = \ln \mathbb{E}[\exp(\lambda U)]$ of $U = f(X^n)$, where $X^n \sim G_n$ and $f : \mathbb{R}^n \to \mathbb{R}$ is any sufficiently smooth function with $\|\nabla f\| \le 1$, can be bounded as $\Lambda(\lambda) \le \lambda^2/2$. This bound then yields the optimal Gaussian concentration inequality $\mathbb{P}\left(|f(X^n) - \mathbb{E}[f(X^n)]| > t\right) \le 2\exp\left(-t^2/2\right)$ for $X^n \sim G_n$. (It should be pointed out that the Gaussian log-Sobolev inequality has a curious history, and seems to have been discovered independently in various equivalent forms by several people, e.g., by Stam [36] in the context of information theory, and by Federbush [37] in the context of mathematical quantum field theory. Through the work of Stam [36], the Gaussian log-Sobolev inequality has been linked to several other information-theoretic notions, such as concavity of entropy power [38, 39, 40].)

In a nutshell, the entropy method takes this idea and applies it beyond the Gaussian case. In abstract terms, log-Sobolev inequalities are functional inequalities that relate the relative entropy between an arbitrary distribution Q w.r.t. the distribution P of interest to some "energy functional" of the density $f = \mathrm{d}Q/\mathrm{d}P$. If one is interested in studying concentration properties of some function $U = f(Z)$ with $Z \sim P$, the core of the entropy method consists in applying an appropriate log-Sobolev inequality to the *tilted distributions* $P^{(\lambda f)}$ with $\mathrm{d}P^{(\lambda f)}/\mathrm{d}P \propto \exp(\lambda f)$. Provided the function f is well-behaved in the sense of having bounded "energy," one uses the "Herbst argument" to pass from the log-Sobolev inequality to the bound $\ln \mathbb{E}[\exp(\lambda U)] \le c\lambda^2/(2C)$, where $c > 0$ depends only on the distribution P, while $C > 0$ is determined by the energy content of f. While there is no universal technique for deriving log-Sobolev inequalities, there are nevertheless some underlying principles that can be exploited for that purpose. We discuss some of these principles in Chapter 3. More information on log-Sobolev inequalities can be found in several excellent monographs and lecture notes [2, 4, 41, 42, 43], as well as in [44, 45, 46, 47, 48] and references therein.

Around the same time as Ledoux first introduced the entropy method in [34], Katalin Marton has shown in a breakthrough paper [49] that to prove concentration bounds one can bypass functional inequalities and work directly on the level of probability measures. More specifically, Marton has shown that Gaussian concentration bounds can be deduced from so-called *transportation-cost inequalities*. These inequalities, discussed in detail in Section 3.4, relate information-theoretic quantities, such as the relative entropy, to a certain class of distances between probability measures on the metric space where the random variables of interest are defined. These so-called *Wasserstein distances* have been the subject of intense research activity that touches upon probability theory, functional analysis, dynamical systems and partial differential equations, statistical physics, and differential geometry. A great deal of information on this field of *optimal transportation* can be found in two books by Cédric Villani — [50] offers a concise and fairly elementary introduction, while a more recent monograph [51] is a lot more detailed and encyclopedic. Multiple connections between optimal transportation, concentration of measure, and information theory are also explored in [13, 52, 53, 54, 55, 56, 57]. (We also note that Wasserstein distances have been used in information theory in the context of lossy source coding [58, 59].)

The first explicit invocation of concentration inequalities in an information-theoretic context appears in the work of Ahlswede et al. [60, 61]. These authors have shown that a certain delicate probabilistic inequality, which they have referred to as the "blowing up lemma," and which we now (thanks to the contributions by Marton [49, 62]) recognize as a Gaussian concentration bound in Hamming space, can be used to derive strong converses for a wide variety of information-theoretic problems, including some multiterminal scenarios. The importance of sharp concentration inequalities for characterizing fundamental limits of coding schemes in information theory is evident from the recent flurry of activity on *finite-blocklength* analysis of source and channel codes [63, 64]. Thus, it is timely to revisit the use of concentration-of-measure ideas in information theory from a more modern perspective. We hope that our treatment, which above all aims to distill the core information-theoretic ideas underlying the study of concentration of measure, will be helpful to information theorists and researchers in related fields.

1.1 A reader's guide

Chapter 2 on the martingale approach is structured as follows: Section 2.1 presents briefly discrete-time (sub/ super) martingales, Section 2.2 presents some basic inequalities that are widely used for proving concentration inequalities via the martingale approach. Section 2.3 derives some refined versions of the Azuma-Hoeffding inequality, and it considers interconnections between these concentration inequalities. Section 2.4 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 2.5 considers some connections between the concentration inequalities that are introduced in Section 2.3 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 2.6 forms the second part of this work, applying the concentration inequalities from Section 2.3 to information theory and some related topics. Chapter 2 is summarized briefly in Section 2.7.

There have been so far very nice surveys on concentration inequalities via the martingale approach that include [5], [9, Chapter 11], [10, Chapter 2] and [11]. The main focus of Chapter 2 is on the presentation of some old and new concentration inequalities that are based on the martingale approach, with an emphasis on some of their potential applications in information and communication-theoretic aspects. This makes the presentation in this chapter different from these aforementioned surveys.

Chapter 3 on the entropy method is structured as follows: Section 3.1 introduces the main ingredients of the entropy method and sets up the major themes that reappears throughout the chapter. Section 3.2 focuses on the logarithmic Sobolev inequality for Gaussian measures, as well as on its numerous links to information-theoretic ideas. The general scheme of logarithmic Sobolev inequalities is introduced in Section 3.3, and then applied to a variety of continuous and discrete examples, including an alternative derivation of McDiarmid's inequality that does not rely on martingale methods and recovers the correct constant in the exponent. Thus, Sections 3.2 and 3.3 present an approach to deriving concentration bounds based on *functional* inequalities. In Section 3.4, concentration is examined through the lens of geometry in probability spaces equipped with a metric. This viewpoint centers around intrinsic properties of probability measures, and has received a great deal of attention since the pioneering work of Marton [62, 49] on transportation-cost inequalities. Although the focus in Chapter 3 is mainly on concentration for product measures, Section 3.5 contains a brief summary of a few results on concentration for functions of dependent random variables, and discusses the connection between these results and the information-theoretic machinery that has been the subject of the chapter. Several applications of concentration to problems in information theory are surveyed in Section 3.6.

Chapter 2

Concentration Inequalities via the Martingale Approach and their Applications in Information Theory, Communications and Coding

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.

2.1 Discrete-time martingales

2.1.1 Martingales

This subsection provides a brief review of 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 σ-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 σ-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. $X_i \in \mathbb{L}^1(\Omega, \mathcal{F}_i, \mathbb{P})$ for every $i \in \{0, \dots, n\}$; this means that each X_i is defined on the same sample space Ω, it is $\mathcal{F}_i$-measurable, and $\mathbb{E}[|X_i|] = \int_\Omega |X_i(\omega)| \mathbb{P}(\mathrm{d}\omega) < \infty$.

3. For all $i \in \{1, \dots, n\}$, 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 $i \in \mathbb{N}$, $\mathbb{E}[X_i] = \mathbb{E}\big[\mathbb{E}[X_i | \mathcal{F}_{i-1}]\big] = \mathbb{E}[X_{i-1}]$, so the expectation of a martingale sequence is fixed.

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

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

Then, the sequence $X_0, X_1, \dots, X_n$ 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 $i \in \{1, \dots, n\}$

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

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

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

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 $\{U_k\}_{k=1}^n$ 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

$$X_k = \sum_{j=1}^k U_j, \quad \forall\, k \in \{1, \dots, n\}$$

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

$$\begin{aligned} \mathcal{F}_k &= \sigma(X_1, \dots, X_k) \\ &= \sigma(U_1, \dots, U_k), \quad \forall\, k \in \{1, \dots, n\}. \end{aligned}$$

Note that $\mathcal{F}_k = \sigma(X_1, \dots, X_k)$ denotes the minimal σ-algebra that includes all the sets of the form $\big\{\omega \in \Omega : (X_1(\omega) \leq \alpha_1, \dots, X_k(\omega) \leq \alpha_k)\big\}$ where $\alpha_j \in \mathbb{R} \cup \{-\infty, +\infty\}$ for $j \in \{1, \dots, k\}$. 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.

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

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

2.2 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.

2.2.1 The Azuma-Hoeffding inequality

The Azuma-Hoeffding inequality[1] is a useful concentration inequality for bounded-difference martingales. It was proved in [8] for independent bounded random variables, followed by a discussion on sums of dependent random variables; this inequality was later derived in [7] for the more general setting of bounded-difference martingales. In the following, this inequality is introduced.

Theorem 1. [**Azuma-Hoeffding inequality**] Let $\{X_k, \mathcal{F}_k\}_{k=0}^n$ 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 $\alpha > 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha) \leq 2 \exp\left(-\frac{\alpha^2}{2\sum_{k=1}^n d_k^2}\right). \tag{2.1}$$

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. Therefore, we present in the following the proof of this inequality.

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

Proof. For an arbitrary $\alpha > 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha) = \mathbb{P}(X_n - X_0 \geq \alpha) + \mathbb{P}(X_n - X_0 \leq -\alpha). \tag{2.2}$$

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,

$$\begin{aligned} &\mathbb{P}(X_n - X_0 \geq \alpha) \\ &= \mathbb{P}\left(\sum_{i=1}^n \xi_i \geq \alpha\right) \\ &\leq e^{-\alpha t}\, \mathbb{E}\left[\exp\left(t \sum_{i=1}^n \xi_i\right)\right], \quad \forall\, t \geq 0. \end{aligned} \tag{2.3}$$

Furthermore,

$$\begin{aligned} &\mathbb{E}\biggl[\exp\biggl(t \sum_{k=1}^n \xi_k\biggr)\biggr] \\ &= \mathbb{E}\biggl[\mathbb{E}\biggl[\exp\biggl(t \sum_{k=1}^n \xi_k\biggr) \,|\, \mathcal{F}_{n-1}\biggr]\biggr] \\ &= \mathbb{E}\left[\exp\biggl(t \sum_{k=1}^{n-1} \xi_k\biggr)\, \mathbb{E}\bigl[\exp(t\xi_n) \,|\, \mathcal{F}_{n-1}\bigr]\right] \end{aligned} \tag{2.4}$$

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

Due to the convexity of the exponential function, and since $|\xi_k| \leq d_k$, then the straight line connecting the end points of the exponential function is below this function over the interval $[-d_k, d_k]$. Hence, for every k (note that $\mathbb{E}[\xi_k \,|\, \mathcal{F}_{k-1}] = 0$),

$$\begin{aligned} &\mathbb{E}\bigl[e^{t\xi_k} \,|\, \mathcal{F}_{k-1}\bigr] \\ &\leq \mathbb{E}\Bigl[\frac{(d_k + \xi_k)e^{td_k} + (d_k - \xi_k)e^{-td_k}}{2d_k} \,|\, \mathcal{F}_{k-1}\Bigr] \\ &= \frac{1}{2}\left(e^{td_k} + e^{-td_k}\right) \\ &= \cosh(td_k). \end{aligned} \tag{2.5}$$

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

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

Consequently, by repeatedly using the recursion in (2.4), it follows that

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

which then gives (see (2.3)) 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 $t = \alpha\left(\sum_{k=1}^{n}d_k^2\right)^{-1}$, and

$$\mathbb{P}(X_n - X_0 \geq \alpha) \leq \exp\left(-\frac{\alpha^2}{2\sum_{k=1}^{n}d_k^2}\right). \tag{2.6}$$

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 (2.2) 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 [5, 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 $r \geq 0$,

$$\mathbb{P}\left(\left|\sum_{k=1}^{n}Y_k\right| \geq r\right) \leq 2\exp\left(-\frac{2r^2}{\sum_{k=1}^{n}(b_k - a_k)^2}\right). \tag{2.7}$$

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

$$\mathbb{P}\left(|X_n - X_0| \geq r\right) \leq 2\exp\left(-\frac{2r^2}{\sum_{k=1}^{n}(b_k - a_k)^2}\right), \quad \forall\, r > 0. \tag{2.8}$$

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 $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \ldots$ where

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

is the σ-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

$$\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) \leq 2\exp\left(-\frac{\alpha^2}{2d^2}\right). \tag{2.9}$$

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 $\alpha \geq 0$,

$$\lim_{n\to\infty}\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) = 2\,Q\Bigl(\frac{\alpha}{d}\Bigr) \tag{2.10}$$

where

$$Q(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp\Bigl(-\frac{t^2}{2}\Bigr) \mathrm{d}t, \quad \forall \, x \in \mathbb{R} \tag{2.11}$$

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

$$\frac{1}{\sqrt{2\pi}} \, \frac{x}{1+x^2} \cdot e^{-\frac{x^2}{2}} < Q(x) < \frac{1}{\sqrt{2\pi}\, x} \cdot e^{-\frac{x^2}{2}}, \ \ \forall \, x > 0 \tag{2.12}$$

then it follows from (2.10) that the exponent on the right-hand side of (2.9) 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 $\alpha \geq 0$

$$\lim_{n \to \infty} \mathbb{P}(|X_n - X_0| \geq \alpha \sqrt{n}) = 2 \, Q\biggl(\frac{\alpha}{\sqrt{\gamma}\, d}\biggr). \tag{2.13}$$

From the exponential upper and lower bounds of the Q-function in (2.12), the right-hand side of (2.13) scales exponentially like $e^{-\frac{\alpha^2}{2\gamma d^2}}$. Hence, the exponent in this example is improved by a factor $\frac{1}{\gamma}$ as compared Azuma's inequality (that is the same as in Example 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 2.3.

2.2.2 McDiarmid's inequality

The following useful inequality is due to McDiarmid ([25, Theorem 3.1] or [65]), 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 $X_i : \Omega_i \to \mathbb{R}$ 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 $k \in \{1, \dots, n\}$,

$$\bigl|g(X_1, \dots, X_{k-1}, X_k, X_{k+1}, \dots, X_n) - g(X_1, \dots, X_{k-1}, \hat{X}_k, X_{k+1}, \dots, X_n)\bigr| \leq d_k \tag{2.14}$$

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

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

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

$$\mathbb{P}(\bigl|g(X_1, \dots, X_n) - \mathbb{E}\bigl[g(X_1, \dots, X_n)\bigr]\bigr| \geq \alpha) \leq 2 \exp\left(-\frac{2\alpha^2}{\sum_{k=1}^n d_k^2}\right). \tag{2.15}$$

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 $k \in \{1, \dots, n\}$, let $\mathcal{F}_k = \sigma(X_1, \dots, X_k)$ be the σ-algebra that is generated by $X_1, \dots, X_k$ with $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Define

$$\xi_k \triangleq \mathbb{E}\big[g(X_1, \dots, X_n) \,|\, \mathcal{F}_k\big] - \mathbb{E}\big[g(X_1, \dots, X_n) \,|\, \mathcal{F}_{k-1}\big], \quad \forall\, k \in \{1, \dots, n\}. \tag{2.16}$$

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

$$\begin{aligned} \mathbb{E}\big[g(X_1, \dots, X_n) \,|\, \mathcal{F}_0\big] &= \mathbb{E}\big[g(X_1, \dots, X_n)\big] \\ \mathbb{E}\big[g(X_1, \dots, X_n) \,|\, \mathcal{F}_n\big] &= g(X_1, \dots, X_n). \end{aligned} \tag{2.17}$$

Hence, it follows from the last three equalities that

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

In the following, we need a lemma:

Lemma 1. For every $k \in \{1, \dots, n\}$, 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 ξ_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 ξ_k is $\mathcal{F}_k$-measurable since $\mathcal{F}_{k-1} \subseteq \mathcal{F}_k$, and ξ_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

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

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

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

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

$$\begin{aligned} &|\xi_k - \hat{\xi}_k| \\ &= \big|\mathbb{E}\big[g(X_1, \dots, X_{k-1}, X_k, X_{k+1}, \dots, X_n) \,|\, \mathcal{F}_k\big] - \mathbb{E}\big[g(X_1, \dots, X_{k-1}, \hat{X}_k, X_{k+1}, \dots, X_n) \,|\, \hat{\mathcal{F}}_k\big]\big| \\ &= \big|\mathbb{E}\big[g(X_1, \dots, X_{k-1}, X_k, X_{k+1}, \dots, X_n) - g(X_1, \dots, X_{k-1}, \hat{X}_k, X_{k+1}, \dots, X_n) \,|\, \sigma(X_1, \dots, X_{k-1}, X_k, \hat{X}_k)\big]\big| \\ &\leq \mathbb{E}\big[|g(X_1, \dots, X_{k-1}, X_k, X_{k+1}, \dots, X_n) - g(X_1, \dots, X_{k-1}, \hat{X}_k, X_{k+1}, \dots, X_n)| \,|\, \sigma(X_1, \dots, X_{k-1}, X_k, \hat{X}_k)\big] \\ &\leq d_k. \end{aligned} \tag{2.18}$$

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

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

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

$$\begin{aligned} &\mathbb{E}[e^{t\xi_k} \,|\, \mathcal{F}_{k-1}] \\ &\leq \mathbb{E}\left[\frac{(\xi_k - a_k)e^{tb_k} + (\xi_k + b_k)e^{ta_k}}{d_k} \,\Big|\, \mathcal{F}_{k-1}\right] \\ &= \frac{b_k e^{ta_k} - a_k e^{tb_k}}{d_k}. \end{aligned}$$

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

$$\begin{aligned} &\mathbb{E}[e^{t\xi_k} \,|\, \mathcal{F}_{k-1}] \\ &\leq p_k e^{tb_k} + (1-p_k)e^{ta_k} \\ &= e^{ta_k}\big(1 - p_k + p_k e^{td_k}\big) \\ &= e^{f_k(t)} \end{aligned} \tag{2.19}$$

where

$$f_k(t) \triangleq t a_k + \ln\big(1 - p_k + p_k e^{td_k}\big), \qquad \forall\, t \in \mathbb{R}. \tag{2.20}$$

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

$$f_k(t) \leq \frac{t^2 d_k^2}{8} \tag{2.21}$$

so, from (2.19),

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

Similarly to the proof of the Azuma-Hoeffding inequality, by repeatedly using the recursion in (2.4), the last inequality implies that

$$\mathbb{E}\left[\exp\left(t \sum_{k=1}^{n} \xi_k\right)\right] \leq \exp\left(\frac{t^2}{8} \sum_{k=1}^{n} d_k^2\right) \tag{2.22}$$

which then gives from (2.3) that, for every $t \geq 0$,

$$\begin{aligned} & \mathbb{P}(g(X_1, \ldots, X_n) - \mathbb{E}[g(X_1, \ldots, X_n)] \geq \alpha) \\ & = \mathbb{P}\left(\sum_{k=1}^{n} \xi_k \geq \alpha\right) \\ & \leq \exp\left(-\alpha t + \frac{t^2}{8} \sum_{k=1}^{n} d_k^2\right). \end{aligned} \tag{2.23}$$

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

$$\mathbb{P}(g(X_1, \ldots, X_n) - \mathbb{E}[g(X_1, \ldots, X_n)] \geq \alpha) \leq \exp\left(-\frac{2\alpha^2}{\sum_{k=1}^{n} d_k^2}\right). \tag{2.24}$$

By replacing g with $-g$, it follows that this bound is also valid for the probability

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

which therefore gives the bound in (2.15). This completes the proof of Theorem 2. □

2.2.3 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 [8, Theorem 2]). An improved version of Hoeffding's inequality, due to Kearns and Saul [66], 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,

$$\mathbb{P}\left(\left|\sum_{k=1}^{n} U_k - \mu_n\right| \geq \alpha\sqrt{n}\right) \leq 2\exp\left(-\frac{2\alpha^2\, n}{\sum_{k=1}^{n}(b_k - a_k)^2}\right), \quad \forall\, \alpha \geq 0. \tag{2.25}$$

Proof. Let $g(\underline{x}) \triangleq \sum_{k=1}^{n} x_k$ for every $\underline{x} \in \mathbb{R}^n$. 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

$$\big|g(X_1, \ldots, X_{k-1}, X_k, X_{k+1}, \ldots, X_n) - g(X_1, \ldots, X_{k-1}, X_k', X_{k+1}, \ldots, X_n)\big| = |X_k - X_k'| \leq b_k - a_k$$

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

$$\mathbb{P}\big(|g(X_1, \ldots, X_n) - \mathbb{E}[g(X_1, \ldots, X_n)]| \geq \alpha\sqrt{n}\big) \leq 2\exp\left(-\frac{2\alpha^2 n}{\sum_{k=1}^{n}(b_k - a_k)^2}\right), \quad \forall\, \alpha \geq 0.$$

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 of $(U_1, \ldots, U_n)$ (note that the entries of each of these vectors are independent, and X_k is distributed like U_k), then

$$\mathbb{P}\big(|g(U_1, \ldots, U_n) - \mu_n| \geq \alpha\sqrt{n}\big) \leq 2\exp\left(-\frac{2\alpha^2 n}{\sum_{k=1}^{n}(b_k - a_k)^2}\right), \quad \forall\, \alpha \geq 0$$

which is equivalent to (2.25). □

An improved version of Hoeffding's inequality, due to Kearns and Saul [66] is introduced in the following. It is noted that a certain gap in the original proof of the improved inequality in [66] was recently solved in [67] 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 (2.19) gives

$$\begin{aligned}\mathbb{E}\big[\exp(t\xi_k)\big] &\leq (1-p_k)\,\exp\big(t(a_k - \mathbb{E}[U_k])\big) + p_k\,\exp\big(t(b_k - \mathbb{E}[U_k])\big) \\ &\triangleq \exp\big(f_k(t)\big) \end{aligned} \tag{2.26}$$

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

$$p_k \triangleq \frac{\mathbb{E}[U_k] - a_k}{b_k - a_k}, \quad \forall\, k \in \{1, \ldots, n\}. \tag{2.27}$$

The derivation of McDiarmid's inequality (see (2.21)) gives that for all $t \in \mathbb{R}$

$$f_k(t) \leq \frac{t^2(b_k - a_k)^2}{8}. \tag{2.28}$$

The improvement of this bound (see [67, Theorem 4]) gives that for all $t \in \mathbb{R}$

$$f_k(t) \leq \begin{cases} \frac{(1-2p_k)(b_k-a_k)^2t^2}{4\ln\left(\frac{1-p_k}{p_k}\right)} & \text{if } p_k \neq \frac{1}{2} \\ \frac{(b_k-a_k)^2t^2}{8} & \text{if } p_k = \frac{1}{2}. \end{cases} \tag{2.29}$$

Note that since

$$\lim_{p \to \frac{1}{2}} \frac{1-2p}{\ln\left(\frac{1-p}{p}\right)} = \frac{1}{2}$$

so the upper bound in (2.29) is continuous in p_k, and it also improves the bound on $f_k(t)$ in (2.28) unless $p_k = \frac{1}{2}$ (where both bounds coincide in this case). From (2.29), we have $f_k(t) \leq c_k t^2$, for every $k \in \{1, \ldots, n\}$ and $t \in \mathbb{R}$, where

$$c_k \triangleq \begin{cases} \frac{(1-2p_k)(b_k-a_k)^2}{4\ln\left(\frac{1-p_k}{p_k}\right)} & \text{if } p_k \neq \frac{1}{2} \\ \frac{(b_k-a_k)^2}{8} & \text{if } p_k = \frac{1}{2}. \end{cases} \tag{2.30}$$

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

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{k=1}^n U_k - \mu_n\right| \geq \alpha\sqrt{n}\right) &\leq 2\exp(-\alpha\sqrt{n}t) \prod_{k=1}^n \mathbb{E}[\exp(t\xi_k)] \\ &\leq 2\exp(-\alpha t\sqrt{n}) \cdot \exp\left(\sum_{k=1}^n c_k t^2\right), \quad \forall\, t \geq 0. \end{aligned} \tag{2.31}$$

Finally, an optimization over the non-negative free parameter t leads to the following improved version of Hoeffding's inequality in [66] (with the recent follow-up in [67]).

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,

$$\mathbb{P}\left(\left|\sum_{k=1}^n U_k - \mu_n\right| \geq \alpha\sqrt{n}\right) \leq 2\exp\left(-\frac{\alpha^2\, n}{4\sum_{k=1}^n c_k}\right), \quad \forall\, \alpha \geq 0. \tag{2.32}$$

where $\{c_k\}_{k=1}^n$ is introduced in (2.30) with the p_k's that are given in (2.27). Moreover, the exponential bound (2.32) 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 [68], followed by some numerical comparisons.

2.3 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

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

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

2.3.1 A refinement of the Azuma-Hoeffding inequality for discrete-time martingales with bounded jumps

The following theorem appears in [65] (see also [69, 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.

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

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

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) \leq 2\exp\left(-n\, D\left(\frac{\delta + \gamma}{1+\gamma} \Big|\Big| \frac{\gamma}{1+\gamma}\right)\right) \tag{2.33}$$

where

$$\gamma \triangleq \frac{\sigma^2}{d^2}, \quad \delta \triangleq \frac{\alpha}{d} \tag{2.34}$$

and

$$D(p||q) \triangleq p\ln\Big(\frac{p}{q}\Big) + (1-p)\ln\Big(\frac{1-p}{1-q}\Big), \quad \forall\, p, q \in [0, 1] \tag{2.35}$$

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 (2.33) is equal to zero.

Proof. The proof of this bound starts similarly to the proof of the Azuma-Hoeffding inequality, up to (2.4). 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., [69, 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 $\lambda \geq 0$,

$$\mathbb{E}\big[e^{\lambda X}\big] \leq \frac{e^{\lambda \overline{x}} \left[(b-\overline{x})^2 e^{-\frac{\lambda\sigma^2}{b-\overline{x}}} + \sigma^2 e^{\lambda(b-\overline{x})}\right]}{(b-\overline{x})^2 + \sigma^2}. \tag{2.36}$$

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

$$\mathbb{E}[e^Y] \leq \left(\frac{b_Y^2}{b_Y^2 + \sigma_Y^2}\right) e^{-\frac{\sigma_Y^2}{b_Y}} + \left(\frac{\sigma_Y^2}{b_Y^2 + \sigma_Y^2}\right) e^{b_Y}. \tag{2.37}$$

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

$$\mathbb{P}\left(Y_0 = -\frac{\sigma_Y^2}{b_Y}\right) = \frac{b_Y^2}{b_Y^2 + \sigma_Y^2}, \qquad \mathbb{P}(Y_0 = b_Y) = \frac{\sigma_Y^2}{b_Y^2 + \sigma_Y^2} \tag{2.38}$$

so inequality (2.37) is equivalent to showing that

$$\mathbb{E}[e^Y] \leq \mathbb{E}[e^{Y_0}]. \tag{2.39}$$

To that end, let ϕ 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 ϕ'' is constant then $f''(y) = 0$ 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 $y_1 \in \big(-\frac{\sigma_Y^2}{b_Y}, b_Y\big)$. By the same argument, applied to f' on $\big[-\frac{\sigma_Y^2}{b_Y}, y_1\big]$, it follows that $y_0 \in \big(-\frac{\sigma_Y^2}{b_Y}, y_1\big)$. The function f is convex on $(-\infty, y_0]$ (since, on this interval, $f''(y) = \phi''(y) - e^y > \phi''(y) - e^{y_0} = \phi''(y_0) - e^{y_0} = f''(y_0) = 0$), 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 ϕ'' is constant since ϕ 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 (2.38) is of zero mean and variance $\mathbb{E}[Y_0^2] = \sigma_Y^2$, so

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

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 (2.39), which therefore completes the proof of the lemma. □

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

$$\mathbb{E}\left[\exp(t\xi_k) \,|\, \mathcal{F}_{k-1}\right] \leq \frac{\sigma^2 \exp(td) + d^2 \exp\left(-\frac{t\sigma^2}{d}\right)}{d^2 + \sigma^2}. \tag{2.40}$$

Hence, it follows from (2.4) and (2.40) that, for every $t \geq 0$,

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

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

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

From the definition of γ in (2.34), this inequality is rewritten as

$$\mathbb{E}\left[\exp\left(t\sum_{k=1}^{n}\xi_k\right)\right] \leq \left(\frac{\gamma \exp(td) + \exp(-\gamma td)}{1+\gamma}\right)^n, \quad \forall\, t \geq 0. \tag{2.41}$$

Let $x \triangleq td$ (so $x \geq 0$). Combining Chernoff's inequality with (2.41) gives that, for every $\alpha \geq 0$ (where from the definition of δ in (2.34), $\alpha t = \delta x$),

$$\begin{aligned}
&\mathbb{P}(X_n - X_0 \geq \alpha n) \\
&\leq \exp(-\alpha n t)\, \mathbb{E}\left[\exp\left(t\sum_{k=1}^{n}\xi_k\right)\right] \\
&\leq \left(\frac{\gamma \exp\big((1-\delta)x\big) + \exp\big(-(\gamma+\delta)x\big)}{1+\gamma}\right)^n, \quad \forall\, x \geq 0.
\end{aligned} \tag{2.42}$$

Consider first the case where $\delta = 1$ (i.e., $\alpha = d$), then (2.42) is particularized to

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

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

$$\mathbb{P}(X_n - X_0 \geq dn) \leq \left(\frac{\gamma}{1+\gamma}\right)^n. \tag{2.43}$$

Otherwise, if $\delta \in [0, 1)$, the minimization of the base of the exponent on the right-hand side of (2.42) w.r.t. the free non-negative parameter x yields that the optimized value is

$$x = \left(\frac{1}{1+\gamma}\right) \ln\left(\frac{\gamma+\delta}{\gamma(1-\delta)}\right) \tag{2.44}$$

and its substitution into the right-hand side of (2.42) gives that, for every $\alpha \geq 0$,

$$\begin{aligned} &\mathbb{P}(X_n - X_0 \geq \alpha n) \\ &\leq \left[\left(\frac{\gamma+\delta}{\gamma}\right)^{-\frac{\gamma+\delta}{1+\gamma}} (1-\delta)^{-\frac{1-\delta}{1+\gamma}}\right]^n \\ &= \exp\left\{-n\left[\left(\frac{\gamma+\delta}{1+\gamma}\right)\ln\left(\frac{\gamma+\delta}{\gamma}\right) + \left(\frac{1-\delta}{1+\gamma}\right)\ln(1-\delta)\right]\right\} \\ &= \exp\left(-n\, D\left(\frac{\delta+\gamma}{1+\gamma}\Big|\Big|\frac{\gamma}{1+\gamma}\right)\right) \end{aligned} \tag{2.45}$$

and the exponent is equal to $+\infty$ if $\delta > 1$ (i.e., if $\alpha > d$). Applying inequality (2.45) to the martingale $\{-X_k, \mathcal{F}_k\}_{k=0}^{\infty}$ gives the same upper bound to the other tail-probability $\mathbb{P}(X_n - X_0 \leq -\alpha n)$. The probability of the union of the two disjoint events $\{X_n - X_0 \geq \alpha n\}$ and $\{X_n - X_0 \leq -\alpha n\}$, that is equal to the sum of their probabilities, therefore satisfies the upper bound in (2.33). 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 $m \in \mathbb{N}$

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

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

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

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

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

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

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

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

$$\mathbb{P}(X_k \geq kx) \leq \exp\left(-\frac{kx^2}{2d^2}\right) \tag{2.46}$$

independently of the value of ε (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 $m \in \mathbb{N}$,

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

then from (2.34)

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

and from (2.45), for every $x \geq 0$,

$$\mathbb{P}(X_k \geq kx) \leq \exp\left(-k\, D\Big(\frac{x(1-\varepsilon)}{d} + \varepsilon \,||\, \varepsilon\Big)\right). \tag{2.47}$$

Consider the case where $\varepsilon \to 0$. Then, for arbitrary $x > 0$ and $k \in \mathbb{N}$, Azuma's inequality in (2.46) provides an upper bound that is strictly positive independently of ε, whereas the one-sided concentration inequality of Theorem 5 implies a bound in (2.47) 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 [5, 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 2.1.2), so it implies that $\bigl\{\exp(t(X_k - X_0)), \mathcal{F}_k\bigr\}_{k=0}^{\infty}$ is a sub-martingale for every $t \geq 0$. Hence, by applying Doob's maximal inequality for sub-martingales, it follows that for every $\alpha \geq 0$

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

which coincides with the proof of Theorem 5 with the starting point in (2.3). 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 $\alpha \geq 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) \leq 2 \exp\left(-n f(\delta)\right) \tag{2.48}$$

where

$$f(\delta) = \begin{cases} \ln(2)\Bigl[1 - h_2\left(\frac{1-\delta}{2}\right)\Bigr], & 0 \leq \delta \leq 1 \\ +\infty, & \delta > 1 \end{cases} \tag{2.49}$$

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

Proof. By substituting $\gamma = 1$ in Theorem 5 (i.e., since there is no constraint on the conditional variance, then one can take $\sigma^2 = d^2$), the corresponding exponent in (2.33) is equal to

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

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

Remark 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 2.1 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 γ is decreased.

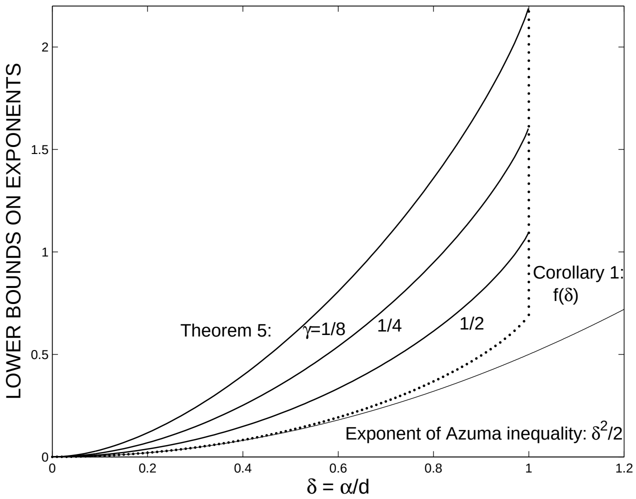

Figure 2.1: Plot of the lower bounds on the exponents from Azuma's inequality and the improved bounds in Theorem 5 and Corollary 1 (where f is defined in (2.49)). The pointed line refers to the exponent in Corollary 1, and the three solid lines for $\gamma = \frac{1}{8}, \frac{1}{4}$ and $\frac{1}{2}$ refer to the exponents in Theorem 5.

2.3.2 Geometric interpretation

A common ingredient in proving Azuma's inequality, and Theorem 5 is a derivation of an upper bound on the conditional expectation $\mathbb{E}\big[e^{t\xi_k} \,|\, \mathcal{F}_{k-1}\big]$ for $t \geq 0$ where $\mathbb{E}\big[\xi_k \,|\, \mathcal{F}_{k-1}\big] = 0$, $\text{Var}\big[\xi_k | \mathcal{F}_{k-1}\big] \leq \sigma^2$, and $|\xi_k| \leq d$ a.s. for some $\sigma, d > 0$ and for every $k \in \mathbb{N}$. The derivation of Azuma's inequality and Corollary 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 $(-\gamma d, e^{-t\gamma d})$. 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 ϕ is the parabola that satisfies the conditions

$$\phi(d) = y(d) = e^{td}, \qquad \phi(-\gamma d) = y(-\gamma d) = e^{-t\gamma d}, \qquad \phi'(-\gamma d) = y'(-\gamma d) = te^{-t\gamma d}.$$

Calculation shows that this parabola admits the form

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

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

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

which 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 $\frac{\gamma}{1+\gamma}$ and $\frac{1}{1+\gamma}$, respectively. This bound also holds when $\mathbb{E}[X^2] \leq \sigma^2$ since the right-hand side of the inequality is a monotonic non-decreasing function of $\mathbb{E}[X^2]$ (as it was verified in the proof Lemma 2). Applying Bennett's inequality to the conditional law of ξ_k given $\mathcal{F}_{k-1}$ gives (2.40) (with γ in (2.34)).

2.3.3 Improving the refined version of the Azuma-Hoeffding inequality for subclasses 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 ξ_k is symmetrically distributed around zero.

Our goal in this subsection is to demonstrate how the assumption of the conditional symmetry improves the existing the deviation inequality in Section 2.3.1 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 [70, Section 4] and [71, 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 [72], [73] and [74].

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

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 $\{\mathcal{F}_k\}_{k \geq 0}$ be the natural filtration of sub σ-algebras of $\mathcal{F}$, where $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_k = \sigma(U_1, \ldots, U_k)$ for $k \geq 1$. Furthermore, for $k \in \mathbb{N}$, let $A_k \in L^\infty(\Omega, \mathcal{F}_{k-1}, \mathbb{P})$ be an $\mathcal{F}_{k-1}$-measurable random variable with a finite essential supremum. Define a new sequence of random variables in $L^1(\Omega, \mathcal{F}, \mathbb{P})$ where

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

and $X_0 = 0$. Then, $\{X_n, \mathcal{F}_n\}_{n \in \mathbb{N}_0}$ is a martingale. Lets 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 σ-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.

$$|X_k - X_{k-1}| \leq d, \qquad \mathrm{Var}(X_k|\mathcal{F}_{k-1}) = \mathbb{E}\big[(X_k - X_{k-1})^2 \,|\, \mathcal{F}_{k-1}\big] \leq \sigma^2 \tag{2.50}$$

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

$$\mathbb{P}\left(\max_{1\leq k\leq n} |X_k - X_0| \geq \alpha n\right) \leq 2\exp\big(-nE(\gamma,\delta)\big) \tag{2.51}$$

where γ and δ are introduced in (2.34), and for $\gamma \in (0,1]$ and $\delta \in [0,1)$

$$E(\gamma,\delta) \triangleq \delta x - \ln\Big(1 + \gamma\big[\cosh(x) - 1\big]\Big) \tag{2.52}$$

$$x \triangleq \ln\left(\frac{\delta(1-\gamma) + \sqrt{\delta^2(1-\gamma)^2 + \gamma^2(1-\delta^2)}}{\gamma(1-\delta)}\right). \tag{2.53}$$

If $\delta > 1$, then the probability on the left-hand side of (2.51) is zero (so $E(\gamma, \delta) \triangleq +\infty$), and $E(\gamma, 1) = \ln\bigl(\frac{2}{\gamma}\bigr)$. Furthermore, the exponent $E(\gamma, \delta)$ is asymptotically optimal in the sense that there exists a conditionally symmetric martingale, satisfying the conditions in (2.50) 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 γ and δ.

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 $\alpha \geq 0$,

$$\mathbb{P}\left(\max_{1 \leq k \leq n} \bigl|S_k - k\,\mathbb{E}(U_1)\bigr| \geq \alpha n\right) \leq 2\exp\bigl(-nE(\gamma, \delta)\bigr), \quad \forall\, n \in \mathbb{N} \tag{2.54}$$

where $\delta \triangleq \frac{\alpha}{d}$, and $E(\gamma, \delta)$ is introduced in (2.52) and (2.53).

Remark 9. Theorem 6 should be compared to Theorem 5 (see [65, Theorem 6.1] or [69, 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 $\mathbb{P}(|X_n - X_0| \geq nd\delta)$ is equal to zero if $\delta > 1$.

If $\delta \to 1^-$ then, from (2.52) and (2.53), for every $\gamma \in (0, 1]$,

$$\lim_{\delta \to 1^-} E(\gamma, \delta) = \lim_{x \to \infty} \bigl[x - \ln\bigl(1 + \gamma(\cosh(x) - 1)\bigr)\bigr] = \ln\left(\frac{2}{\gamma}\right). \tag{2.55}$$

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

$$\lim_{\delta \to 1^-} D\left(\frac{\delta + \gamma}{1 + \gamma} \Big|\Big| \frac{\gamma}{1 + \gamma}\right) = \ln\left(1 + \frac{1}{\gamma}\right) \tag{2.56}$$

where the last equality follows from (2.35). A comparison of the limits in (2.55) and (2.56) 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 $f(\delta)$ as is defined in (2.49).

We prove in the following the new deviation inequality in Theorem 6. In order to prove Theorem 6 for a discrete-time, real-valued and conditionally symmetric martingale with bounded jumps, we deviate from the proof of Theorem 5. This is done by a replacement of Bennett's inequality for the conditional expectation in (2.40) with a tightened bound under the conditional symmetry assumption. To this end, we need a lemma to proceed.

Lemma 3. Let X be a real-valued RV with a symmetric distribution around zero, a support $[-d, d]$, and assume that $\mathbb{E}[X^2] = \text{Var}(X) \leq \gamma d^2$ for some $d > 0$ and $\gamma \in [0, 1]$. Let h be a real-valued convex function, and assume that $h(d^2) \geq h(0)$. Then

$$\mathbb{E}[h(X^2)] \leq (1 - \gamma)h(0) + \gamma h(d^2) \tag{2.57}$$

where equality holds for the symmetric distribution

$$\mathbb{P}(X = d) = \mathbb{P}(X = -d) = \frac{\gamma}{2}, \quad \mathbb{P}(X = 0) = 1 - \gamma. \tag{2.58}$$

Proof. Since h is convex and $\text{supp}(X) = [-d, d]$, then a.s. $h(X^2) \leq h(0) + \left(\frac{X}{d}\right)^2 \big(h(d^2) - h(0)\big)$. Taking expectations on both sides gives (2.57), which holds with equality for the symmetric distribution in (2.58). □

Corollary 3. If X is a random variable that satisfies the three requirements in Lemma 3 then, for every $\lambda \in \mathbb{R}$,

$$\mathbb{E}\big[\exp(\lambda X)\big] \leq 1 + \gamma\big[\cosh(\lambda d) - 1\big] \tag{2.59}$$

and (2.59) holds with equality for the symmetric distribution in Lemma 3, independently of the value of λ.

Proof. For every $\lambda \in \mathbb{R}$, due to the symmetric distribution of X, $\mathbb{E}\big[\exp(\lambda X)\big] = \mathbb{E}\big[\cosh(\lambda X)\big]$. The claim now follows from Lemma 3 since, for every $x \in \mathbb{R}$, $\cosh(\lambda x) = h(x^2)$ where $h(x) \triangleq \sum_{n=0}^{\infty} \frac{\lambda^{2n}|x|^n}{(2n)!}$ 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 ξ_k given $\mathcal{F}_{k-1}$, it follows that for every $k \in \mathbb{N}$ and $t \in \mathbb{R}$

$$\mathbb{E}\left[\exp(t\xi_k) \,|\, \mathcal{F}_{k-1}\right] \leq 1 + \gamma\big[\cosh(td) - 1\big] \tag{2.60}$$

holds a.s., and therefore it follows from (2.4) and (2.60) that for every $t \in \mathbb{R}$

$$\mathbb{E}\left[\exp\left(t \sum_{k=1}^{n} \xi_k\right)\right] \leq \Big(1 + \gamma\big[\cosh(td) - 1\big]\Big)^n. \tag{2.61}$$

By applying the maximal inequality for submartingales, then for every $\alpha \geq 0$ and $n \in \mathbb{N}$

$$\begin{aligned}
&\mathbb{P}\Big(\max_{1 \leq k \leq n} (X_k - X_0) \geq \alpha n\Big) \\
&= \mathbb{P}\Big(\max_{1 \leq k \leq n} \exp\left(t(X_k - X_0)\right) \geq \exp(\alpha n t)\Big) \qquad \forall\, t \geq 0 \\
&\leq \exp(-\alpha n t)\, \mathbb{E}\Big[\exp\big(t(X_n - X_0)\big)\Big] \\
&= \exp(-\alpha n t)\, \mathbb{E}\left[\exp\left(t \sum_{k=1}^{n} \xi_k\right)\right]
\end{aligned} \tag{2.62}$$

Therefore, from (2.62), for every $t \geq 0$,

$$\mathbb{P}\Big(\max_{1 \leq k \leq n} (X_k - X_0) \geq \alpha n\Big) \leq \exp(-\alpha n t)\Big(1 + \gamma\big[\cosh(td) - 1\big]\Big)^n. \tag{2.63}$$

From (2.34) and a replacement of td with x, then for an arbitrary $\alpha \geq 0$ and $n \in \mathbb{N}$

$$\mathbb{P}\Big(\max_{1 \leq k \leq n} (X_k - X_0) \geq \alpha n\Big) \leq \inf_{x \geq 0} \Big\{\exp\Big(-n\Big[\delta x - \ln\big(1 + \gamma\big[\cosh(x) - 1\big]\big)\Big]\Big)\Big\}. \tag{2.64}$$

Applying (2.64) to the martingale $\{-X_k, \mathcal{F}_k\}_{k \in \mathbb{N}_0}$ 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

$$\mathbb{P}\left(\max_{1 \leq k \leq n} |X_k - X_0| \geq \alpha n\right) \leq \mathbb{P}\left(\max_{1 \leq k \leq n} (X_k - X_0) \geq \alpha n\right) + \mathbb{P}\left(\min_{1 \leq k \leq n} (X_k - X_0) \leq -\alpha n\right). \tag{2.65}$$

This doubles the bound on the right-hand side of (2.64), 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 (2.52) and (2.53) 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

$$\mathbb{P}(U_i = d) = \mathbb{P}(U_i = -d) = \frac{\gamma}{2}, \quad \mathbb{P}(U_i = 0) = 1 - \gamma, \quad \forall\, i \in \mathbb{N}. \tag{2.66}$$

Consider the particular case of the conditionally symmetric martingale $\{X_n, \mathcal{F}_n\}_{n \in \mathbb{N}_0}$ in Example 6 (see Section 2.3.3) 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 $\alpha \geq \mathbb{E}[U_1] = 0$,

$$\begin{aligned} &\lim_{n \to \infty} \frac{1}{n} \ln \mathbb{P}(X_n - X_0 \geq \alpha n) \\ &= \lim_{n \to \infty} \frac{1}{n} \ln \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n U_i \geq \alpha\right) \\ &= -I(\alpha) \end{aligned} \tag{2.67}$$

where the rate function is given by

$$I(\alpha) = \sup_{t \geq 0} \{t\alpha - \ln \mathbb{E}[\exp(tU_1)]\} \tag{2.68}$$

(see, e.g., [69, Theorem 2.2.3] and [69, Lemma 2.2.5(b)] for the restriction of the supermum to the interval $[0, \infty)$). From (2.66) and (2.68), for every $\alpha \geq 0$,

$$I(\alpha) = \sup_{t \geq 0} \big\{t\alpha - \ln\big(1 + \gamma[\cosh(td) - 1]\big)\big\}$$

but it is equivalent to the optimized exponent on the right-hand side of (2.63), giving the exponent of the bound in Theorem 6. Hence, $I(\alpha) = E(\gamma, \delta)$ in (2.52) and (2.53). 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 (2.50) 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 (2.33)) 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:

$$\mathbb{P}(U_i = d) = \frac{\gamma}{1 + \gamma}, \quad \mathbb{P}(U_i = -\gamma d) = \frac{1}{1 + \gamma}, \quad \forall\, i \in \mathbb{N}$$

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 5 and 6 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.

Theorem 7. Let $\{X_k, \mathcal{F}_k\}_{k \in \mathbb{N}_0}$ be a discrete-time and real-valued conditionally symmetric martingale. Let $m \in \mathbb{N}$ be an even number, and assume that the following conditions hold a.s. for every $k \in \mathbb{N}$

$$|X_k - X_{k-1}| \leq d, \qquad \mathbb{E}\big[(X_k - X_{k-1})^l \,|\, \mathcal{F}_{k-1}\big] \leq \mu_l, \;\; \forall\, l \in \{2, 4, \ldots, m\}$$

for some $d > 0$ and non-negative numbers $\{\mu_2, \mu_4, \ldots, \mu_m\}$. Then, for every $\alpha \geq 0$ and $n \in \mathbb{N}$,

$$\mathbb{P}\left(\max_{1 \leq k \leq n} |X_k - X_0| \geq \alpha n\right) \leq 2 \left\{ \min_{x \geq 0} e^{-\delta x} \left[1 + \sum_{l=1}^{\frac{m}{2}-1} \frac{\gamma_{2l}\, x^{2l}}{(2l)!} + \gamma_m \big(\cosh(x) - 1\big)\Big)\right]\right\}^n \tag{2.69}$$

where

$$\delta \triangleq \frac{\alpha}{d}, \quad \gamma_{2l} \triangleq \frac{\mu_{2l}}{d^{2l}}, \ \ \forall\, l \in \Big\{1, \ldots, \frac{m}{2}\Big\}. \tag{2.70}$$

Proof. The starting point of the proof of Theorem 7 relies on (2.62) and (2.4). For every $k \in \mathbb{N}$ and $t \in \mathbb{R}$, since $\mathbb{E}\big[\xi_k^{2l-1} \,|\, \mathcal{F}_{k-1}\big] = 0$ for every $l \in \mathbb{N}$ (due to the conditionally symmetry property of the martingale),

$$\begin{aligned}
&\mathbb{E}\big[\exp(t\xi_k)|\mathcal{F}_{k-1}\big] \\
&= 1 + \sum_{l=1}^{\frac{m}{2}-1} \frac{t^{2l}\, \mathbb{E}\big[\xi_k^{2l} \,|\, \mathcal{F}_{k-1}\big]}{(2l)!} + \sum_{l=\frac{m}{2}}^{\infty} \frac{t^{2l}\, \mathbb{E}\big[\xi_k^{2l} \,|\, \mathcal{F}_{k-1}\big]}{(2l)!} \\
&= 1 + \sum_{l=1}^{\frac{m}{2}-1} \frac{(td)^{2l}\, \mathbb{E}\big[\big(\frac{\xi_k}{d}\big)^{2l} \,|\, \mathcal{F}_{k-1}\big]}{(2l)!} + \sum_{l=\frac{m}{2}}^{\infty} \frac{(td)^{2l}\, \mathbb{E}\big[\big(\frac{\xi_k}{d}\big)^{2l} \,|\, \mathcal{F}_{k-1}\big]}{(2l)!} \\
&\leq 1 + \sum_{l=1}^{\frac{m}{2}-1} \frac{(td)^{2l}\, \gamma_{2l}}{(2l)!} + \sum_{l=\frac{m}{2}}^{\infty} \frac{(td)^{2l}\, \gamma_m}{(2l)!} \\
&= 1 + \sum_{l=1}^{\frac{m}{2}-1} \frac{(td)^{2l}\, \big(\gamma_{2l} - \gamma_m\big)}{(2l)!} + \gamma_m \big(\cosh(td) - 1\big)
\end{aligned} \tag{2.71}$$

where the inequality above holds since $|\frac{\xi_k}{d}| \leq 1$ 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 (2.71) holds since $\cosh(x) = \sum_{n=0}^{\infty} \frac{x^{2n}}{(2n)!}$ for every $x \in \mathbb{R}$. Therefore, from (2.4),

$$\mathbb{E}\left[\exp\Big(t \sum_{k=1}^{n} \xi_k\Big)\right] \leq \left(1 + \sum_{l=1}^{\frac{m}{2}-1} \frac{(td)^{2l}\, \big(\gamma_{2l} - \gamma_m\big)}{(2l)!} + \gamma_m \big[\cosh(td) - 1\big]\right)^n \tag{2.72}$$

for an arbitrary $t \in \mathbb{R}$. The inequality then follows from (2.62). This completes the proof of Theorem 7. □

2.3.4 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 $\alpha \geq 0$

$$\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) \leq 2 \exp\Big(-\frac{\delta^2}{2\gamma}\Big) \Big(1 + O\big(n^{-\frac{1}{2}}\big)\Big). \tag{2.73}$$

Proof. See Appendix 2.A. □

Remark 10. From Proposition 1, the upper bound on $\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n})$ (for an arbitrary $\alpha \geq 0$) improves the exponent of Azuma's inequality by a factor of $\frac{1}{\gamma}$.

2.3.5 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., [10, Chapter 2] or [11, 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.

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

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

$$\mathbb{P}(X_n - X_0 \geq \alpha n) \leq \exp\left(-n\, D\left(\frac{\delta+\gamma}{1+\gamma} \Big|\Big| \frac{\gamma}{1+\gamma}\right)\right) \tag{2.74}$$

where γ and δ are defined as in (2.34), and the divergence $D(p||q)$ is introduced in (2.35). Alternatively, if $\{X_k, \mathcal{F}_k\}_{k=0}^\infty$ is a sub-martingale, the same upper bound in (2.74) 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 2.1.2,

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

a.s., where $\xi_k \triangleq X_k - \mathbb{E}[X_k \,|\, \mathcal{F}_{k-1}]$ is $\mathcal{F}_k$-measurable. Hence $\mathbb{P}((X_n - X_0 \geq \alpha n) \leq \mathbb{P}\big(\sum_{k=1}^n \xi_k \geq \alpha n\big)$ where a.s. $\xi_k \leq d$, $\mathbb{E}[\xi_k \,|\, \mathcal{F}_{k-1}] = 0$, and $\mathrm{Var}(\xi_k \,|\, \mathcal{F}_{k-1}) \leq \sigma^2$. The continuation of the proof coincides with the proof of Theorem 5 (starting from (2.3)). 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. □

2.4 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 [75]. Freedman's inequality is refined in the following to conditionally symmetric martingales with bounded jumps (see [76]). 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

$$Q_n \triangleq \sum_{k=1}^n \mathbb{E}[\xi_k^2 \,|\, \mathcal{F}_{k-1}] \tag{2.75}$$

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

$$\mathbb{P}\left(\max_{1 \leq k \leq n} (X_k - X_0) \geq z,\, Q_n \leq r \;\text{ for some } n \in \mathbb{N}\right) \leq \exp\left(-\frac{z^2}{2r} \cdot C\left(\frac{zd}{r}\right)\right) \tag{2.76}$$

where

$$C(u) \triangleq \frac{2[u \sinh^{-1}(u) - \sqrt{1+u^2} + 1]}{u^2}, \quad \forall \, u > 0. \tag{2.77}$$

Theorem 8 should be compared to Freedman's inequality in [75, Theorem 1.6] (see also [69, 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 $\xi_k \triangleq X_k - X_{k-1} \leq d$ a.s. for every $k \in \mathbb{N}$. Then, for every $z, r > 0$,

$$\mathbb{P}\left(\max_{1 \leq k \leq n} (X_k - X_0) \geq z, \; Q_n \leq r \; \text{ for some } n \in \mathbb{N} \right) \leq \exp\left(-\frac{z^2}{2r} \cdot B\left(\frac{zd}{r}\right) \right) \tag{2.78}$$

where

$$B(u) \triangleq \frac{2[(1+u)\ln(1+u) - u]}{u^2}, \quad \forall \, u > 0. \tag{2.79}$$

The proof of [75, 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 [75, Theorem 1.6], and also [69, Exercise 2.4.21(b)]). The original proof of Theorem 9 (see [75, 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 (2.60) 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, lets 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 $S_n \triangleq X_n - X_0$ 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

$$U_n \triangleq \exp(\lambda S_n - \theta Q_n), \quad \forall \, n \in \{0, 1, \ldots\} \tag{2.80}$$

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 (2.80) is $\mathcal{F}_n$-measurable (since Q_n in (2.75) 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 λ and θ in (2.80).

$$\begin{aligned}
&\mathbb{E}[U_n|\mathcal{F}_{n-1}] \\
&\overset{\text{(a)}}{=} \exp(-\theta Q_n) \, \exp(\lambda S_{n-1}) \, \mathbb{E}\big[\exp(\lambda \xi_n) \,|\, \mathcal{F}_{n-1}\big] \\
&\overset{\text{(b)}}{=} \exp(\lambda S_{n-1}) \, \exp\big(-\theta(Q_{n-1} + \mathbb{E}[\xi_n^2|\mathcal{F}_{n-1}])\big) \, \mathbb{E}\big[\exp(\lambda \xi_n) \,|\, \mathcal{F}_{n-1}\big] \\
&\overset{\text{(c)}}{=} U_{n-1} \left(\frac{\mathbb{E}\big[\exp(\lambda \xi_n) \,|\, \mathcal{F}_{n-1}\big]}{\exp(\theta \mathbb{E}[\xi_n^2 \,|\, \mathcal{F}_{n-1}])} \right)
\end{aligned} \tag{2.81}$$

where (a) follows from (2.80) 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 (2.75), and (c) follows from (2.80).

A modification of the original proof of Lemma 4 (see [75, 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 (2.40) to the conditional expectation of $e^{\lambda \xi_n}$ given $\mathcal{F}_{n-1}$ (recall that $\lambda \geq 0$) gives

$$\mathbb{E}\big[\exp(\lambda \xi_n) \,|\, \mathcal{F}_{n-1}\big] \leq \frac{\exp\big(-\lambda \mathbb{E}[\xi_n^2 \,|\, \mathcal{F}_{n-1}]\big) + \mathbb{E}[\xi_n^2 \,|\, \mathcal{F}_{n-1}] \exp(\lambda)}{1 + \mathbb{E}\big[\xi_n^2 \,|\, \mathcal{F}_{n-1}\big]}$$

which therefore implies from (2.81) and the last inequality that

$$\mathbb{E}[U_n | \mathcal{F}_{n-1}] \leq U_{n-1} \left(\frac{\exp\big(-(\lambda + \theta)\, \mathbb{E}[\xi_n^2 \,|\, \mathcal{F}_{n-1}]\big)}{1 + \mathbb{E}\big[\xi_n^2 \,|\, \mathcal{F}_{n-1}\big]} + \frac{\mathbb{E}[\xi_n^2 \,|\, \mathcal{F}_{n-1}] \exp\big(\lambda - \theta \mathbb{E}[\xi_n^2 \,|\, \mathcal{F}_{n-1}]\big)}{1 + \mathbb{E}[\xi_n^2 \,|\, \mathcal{F}_{n-1}]} \right). \tag{2.82}$$

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 (2.82) 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 $\alpha \geq 0$

$$\left(\frac{1}{1+\alpha}\right) \exp\big(-\alpha(\lambda + \theta)\big) + \left(\frac{\alpha}{1+\alpha}\right) \exp(\lambda - \alpha\theta) \leq 1 \tag{2.83}$$

which then assures that the second term on the right-hand side of (2.82) 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 (2.83) 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 (2.82) 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 (2.80) 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 ξ_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 $\lambda \geq 0$,

$$\mathbb{E}\big[\exp(\lambda \xi_n) \,|\, \mathcal{F}_{n-1}\big] \leq 1 + \mathbb{E}[\xi_n^2 \,|\, \mathcal{F}_{n-1}] \,\big(\cosh(\lambda) - 1\big). \tag{2.84}$$

Hence, combining (2.81) and (2.84) gives

$$\mathbb{E}[U_n | \mathcal{F}_{n-1}] \leq U_{n-1} \left(\frac{1 + \mathbb{E}[\xi_n^2 \,|\, \mathcal{F}_{n-1}] \,\big(\cosh(\lambda) - 1\big)}{\exp\big(\theta \mathbb{E}[\xi_n^2 | \mathcal{F}_{n-1}]\big)} \right). \tag{2.85}$$

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 (2.85)) that the following condition holds:

$$\frac{1 + \alpha\bigl(\cosh(\lambda) - 1\bigr)}{\exp(\theta\alpha)} \leq 1, \quad \forall\, \alpha \geq 0. \tag{2.86}$$

Calculus shows that, for $\lambda \geq 0$, the condition in (2.86) is satisfied if and only if

$$\theta \geq \cosh(\lambda) - 1 \triangleq \theta_{\min}(\lambda). \tag{2.87}$$

From (2.85), $\{U_n, \mathcal{F}_n\}_{n \in \mathbb{N}_0}$ 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 β 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 λ and θ, and from Doob's sampling theorem

$$\mathbb{E}[U_\beta] \leq \mathbb{E}[U_0] = 1 \tag{2.88}$$

($S_0 = Q_0 = 0$, so from (2.80), $U_0 = 1$ a.s.). Hence, it implies the following chain of inequalities:

$$\begin{aligned}
&\mathbb{P}(\exists\, n \leq M : \, S_n \geq z, Q_n \leq r) \\
&\overset{\text{(a)}}{=} \mathbb{P}(S_\beta \geq z, Q_\beta \leq r) \\
&\overset{\text{(b)}}{\leq} \mathbb{P}(\lambda S_\beta - \theta Q_\beta \geq \lambda z - \theta r) \\
&\overset{\text{(c)}}{\leq} \frac{\mathbb{E}[\exp(\lambda S_\beta - \theta Q_\beta)]}{\exp(\lambda z - \theta r)} \\
&\overset{\text{(d)}}{=} \frac{\mathbb{E}[U_\beta]}{\exp(\lambda z - \theta r)} \\
&\overset{\text{(e)}}{\leq} \exp\bigl(-(\lambda z - \theta r)\bigr)
\end{aligned} \tag{2.89}$$

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 (2.80), and finally (e) follows from (2.88). Since (2.89) holds for every $M \in \mathbb{N}$, then from the continuity theorem for non-decreasing events and (2.89)

$$\begin{aligned}
&\mathbb{P}(\exists\, n \in \mathbb{N} : \, S_n \geq z, Q_n \leq r) \\
&= \lim_{M \to \infty} \mathbb{P}(\exists\, n \leq M : \, S_n \geq z, Q_n \leq r) \\
&\leq \exp\bigl(-(\lambda z - \theta r)\bigr).
\end{aligned} \tag{2.90}$$

The choice of the non-negative parameter θ as the minimal value for which (2.90) 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 $z, r > 0$,

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

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

$$\mathbb{P}(\exists\, n \in \mathbb{N}: \, S_n \geq z, Q_n \leq r) \leq \exp\left(-\frac{z^2}{2r} \cdot B\left(\frac{z}{r}\right)\right) \tag{2.91}$$

where the function B is introduced in (2.79).

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 (2.90) for a fixed $\lambda \geq 0$. This gives the inequality

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

The optimized λ is equal to $\lambda = \sinh^{-1}\left(\frac{z}{r}\right)$. Its substitution in (2.87) gives that $\theta_{\min}(\lambda) = \sqrt{1+\frac{z^2}{r^2}} - 1$, and

$$\mathbb{P}(\exists\, n \in \mathbb{N}: \, S_n \geq z, Q_n \leq r) \leq \exp\left(-\frac{z^2}{2r} \cdot C\left(\frac{z}{r}\right)\right) \tag{2.92}$$

where the function C is introduced in (2.77).

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

$$\begin{aligned} A &\triangleq \{\exists\, n \in \mathbb{N}: \, S_n \geq z, Q_n \leq r\} \\ &= \{\exists\, n \in \mathbb{N}: \, \max_{1 \leq k \leq n} S_k \geq z, Q_n \leq r\} \triangleq B. \end{aligned} \tag{2.93}$$

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 (2.75) 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 (2.92) and (2.93) completes the proof of Theorem 8, and respectively the combination of (2.91) and (2.93) completes the proof of Theorem 9. □

Freedman's inequality can be easily specialized to a concentration inequality for a sum of centered (zero-mean) independent and bounded random variables (see Example 1). This specialization reduces to a concentration inequality of Bennett (see [77]), which can be loosened to get Bernstein's inequality (as is explained below). Furthermore, the refined inequality in Theorem 8 for conditionally symmetric martingales with bounded jumps can be specialized (again, via Example 1) to an improved concentration inequality for a sum of i.i.d. and bounded random variables 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 $\mathbb{E}[U_1] = 0$, $\mathbb{E}[U_1^2] = \sigma^2$, and $|U_1| \leq d$ a.s. for some constant $d > 0$. Then, the following inequality holds:

$$\mathbb{P}\left(\left|\sum_{i=1}^n U_i\right| \geq \alpha\right) \leq 2\exp\left(-\frac{n\sigma^2}{d^2} \cdot \phi_1\left(\frac{\alpha d}{n\sigma^2}\right)\right), \quad \forall\, \alpha > 0 \tag{2.94}$$

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 (2.94) can be improved to

$$\mathbb{P}\left(\left|\sum_{i=1}^n U_i\right| \geq \alpha\right) \leq 2\exp\left(-\frac{n\sigma^2}{d^2} \cdot \phi_2\left(\frac{\alpha d}{n\sigma^2}\right)\right), \quad \forall\, \alpha > 0 \tag{2.95}$$

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

Proof. Inequality (2.94) follows from Freedman's inequality in Theorem 9, and inequality (2.95) 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 $X_k = \sum_{i=1}^n U_i$ 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 (2.76) and (2.78) (with the related functions that are introduced in (2.79) and (2.77), respectively). Note that the same bound holds for the two one-sided tail inequalities, giving the factor 2 on the right-hand sides of (2.94) and (2.95). □

Remark 11. Bennett's concentration inequality in (2.94) can be loosened to obtain Bernstein's inequality. To this end, the following lower bound on ϕ_1 is used:

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

This gives the inequality

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

2.5 Relations of the refined inequalities to some classical results in probability theory

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

Let

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

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

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

exists in probability and is positive. The martingale CLT asserts that, under the above conditions, $\frac{S_n}{\sqrt{n}}$ converges in distribution (i.e., weakly converges) to the Gaussian distribution $\mathcal{N}(0, \sigma^2)$. It is denoted by $\frac{S_n}{\sqrt{n}} \Rightarrow \mathcal{N}(0, \sigma^2)$. We note that there exist more general versions of this statement (see, e.g., [78, pp. 475–478]).

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

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

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

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

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

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

Under the assumptions in Theorem 5 and its subsequences, for every $k \in \mathbb{N}$, one gets a.s. that

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

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

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

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

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

where the Q function is introduced in (2.11).

Based on the notation in (2.34), the equality $\frac{\alpha}{\sigma} = \frac{\delta}{\sqrt{\gamma}}$ holds, and

$$\lim_{n \to \infty} \mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) = 2\,Q\Big(\frac{\delta}{\sqrt{\gamma}}\Big). \tag{2.96}$$

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 (2.73) 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 2.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 (2.12), then it follows from (2.96) that the exponent in the concentration inequality (2.73) (i.e., $\frac{\delta^2}{2\gamma}$) cannot be improved under the above assumptions (unless some more information is available).

2.5.2 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., [78, Theorem 9.5]) if $\{X_k\}_{k=1}^{\infty}$ are i.i.d. real-valued RVs with zero mean and unit variance, and $S_n \triangleq \sum_{i=1}^{n} X_i$ for every $n \in \mathbb{N}$, then

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n \ln \ln n}} = 1 \quad \text{a.s.} \tag{2.97}$$

and

$$\liminf_{n\to\infty} \frac{S_n}{\sqrt{2n \ln \ln n}} = -1 \quad \text{a.s.} \tag{2.98}$$

Eqs. (2.97) and (2.98) 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. (2.97) and (2.98) 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 σ-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 $\alpha \geq 0$

$$\mathbb{P}\left(S_n \geq \alpha\sqrt{2n \ln \ln n}\right) \leq \exp\left(-nD\Big(\frac{\delta_n + \gamma}{1+\gamma} \Big|\Big| \frac{\gamma}{1+\gamma}\Big)\right) \tag{2.99}$$

where

$$\delta_n \triangleq \frac{\alpha}{c}\sqrt{\frac{2 \ln \ln n}{n}}, \quad \gamma \triangleq \frac{1}{c^2}. \tag{2.100}$$

Straightforward calculation shows that

$$\begin{aligned}
& nD\Big(\frac{\delta_n+\gamma}{1+\gamma}\Big|\Big|\frac{\gamma}{1+\gamma}\Big) \\
& = \frac{n\gamma}{1+\gamma}\left[\Big(1+\frac{\delta_n}{\gamma}\Big)\,\ln\Big(1+\frac{\delta_n}{\gamma}\Big)+\frac{1}{\gamma}\,(1-\delta_n)\ln(1-\delta_n)\right] \\
& \overset{\text{(a)}}{=} \frac{n\gamma}{1+\gamma}\left[\frac{\delta_n^2}{2}\Big(\frac{1}{\gamma^2}+\frac{1}{\gamma}\Big)+\frac{\delta_n^3}{6}\Big(\frac{1}{\gamma}-\frac{1}{\gamma^3}\Big)+\dots\right] \\
& = \frac{n\delta_n^2}{2\gamma}-\frac{n\delta_n^3(1-\gamma)}{6\gamma^2}+\dots \\
& \overset{\text{(b)}}{=} \alpha^2\ln\ln n\left[1-\frac{\alpha(c^2-1)}{6c}\sqrt{\frac{\ln\ln n}{n}}+\dots\right] \qquad (2.101)
\end{aligned}$$

where equality (a) follows from the power series expansion

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

and equality (b) follows from (2.100). A substitution of (2.101) into (2.99) gives that, for every $\alpha \geq 0$,

$$\mathbb{P}\left(S_n \geq \alpha\sqrt{2n\ln\ln n}\right) \leq \big(\ln n\big)^{-\alpha^2\left[1+O\left(\sqrt{\frac{\ln\ln n}{n}}\right)\right]} \qquad (2.102)$$

and the same bound also applies to $\mathbb{P}\big(S_n \leq -\alpha\sqrt{2n\ln\ln n}\big)$ for $\alpha \geq 0$. This provides complementary information to the limits in (2.97) and (2.98) that are provided by the LIL. From Remark 6, which follows from Doob's maximal inequality for sub-martingales, the inequality in (2.102) can be strengthened to

$$\mathbb{P}\left(\max_{1\le k\le n} S_k \geq \alpha\sqrt{2n\ln\ln n}\right) \leq \big(\ln n\big)^{-\alpha^2\left[1+O\left(\sqrt{\frac{\ln\ln n}{n}}\right)\right]}. \qquad (2.103)$$

It is shown in the following that (2.103) and the first Borel-Cantelli lemma can serve to prove one part of (2.97). 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

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

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

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

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

$$\begin{aligned}
\mathbb{P}(A_n) &\leq \mathbb{P}\left(\max_{\theta^{n-1}\le k\le\theta^n} S_k \geq \alpha\sqrt{2\theta^{n-1}\ln\ln(\theta^{n-1})}\right) \\
&= \mathbb{P}\left(\max_{\theta^{n-1}\le k\le\theta^n} S_k \geq \frac{\alpha}{\sqrt{\theta}}\,\sqrt{2\theta^{n}\ln\ln(\theta^{n-1})}\right) \\
&\leq \mathbb{P}\left(\max_{1\le k\le\theta^n} S_k \geq \frac{\alpha}{\sqrt{\theta}}\,\sqrt{2\theta^{n}\ln\ln(\theta^{n-1})}\right) \\
&\leq (n\ln\theta)^{-\frac{\alpha^2}{\theta}\left(1+\beta_n\right)} \qquad (2.104)
\end{aligned}$$

where the last inequality follows from (2.103) with $\beta_n \to 0$ as $n \to \infty$. Since

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

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

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

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

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

Theorem 5 therefore gives inequality (2.103), and it implies one side in each of the two equalities for the LIL in (2.97) and (2.98).

2.5.3 Relation of Theorem 5 with the moderate deviations principle

According to the moderate deviations theorem (see, e.g., [69, Theorem 3.7.1]) in $\mathbb{R}$, let $\{X_i\}_{i=1}^n$ be a sequence of real-valued i.i.d. RVs such that $\Lambda_X(\lambda) = \mathbb{E}[e^{\lambda X_i}] < \infty$ in some neighborhood of zero, and also assume that $\mathbb{E}[X_i] = 0$ and $\sigma^2 = \text{Var}(X_i) > 0$. Let $\{a_n\}_{n=1}^\infty$ be a non-negative sequence such that $a_n \to 0$ and $na_n \to \infty$ as $n \to \infty$, and let

$$Z_n \triangleq \sqrt{\frac{a_n}{n}} \sum_{i=1}^n X_i, \quad \forall\, n \in \mathbb{N}. \tag{2.105}$$

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

$$\begin{aligned}
&-\frac{1}{2\sigma^2} \inf_{x \in \Gamma^0} x^2 \\
&\leq \liminf_{n\to\infty} a_n \ln \mathbb{P}(Z_n \in \Gamma) \\
&\leq \limsup_{n\to\infty} a_n \ln \mathbb{P}(Z_n \in \Gamma) \\
&\leq -\frac{1}{2\sigma^2} \inf_{x \in \overline{\Gamma}} x^2
\end{aligned} \tag{2.106}$$

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

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

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

so that $a_n \to 0$ 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 (2.105),

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

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

$$\lim_{n\to\infty} n^{1-2\eta} \ln \mathbb{P}\left(\Big|\sum_{i=1}^n X_i\Big| \geq \alpha n^\eta\right) = -\frac{\alpha^2}{2\sigma^2}. \tag{2.107}$$

It is demonstrated in Appendix 2.B that, in contrast to Azuma's inequality, Theorem 5 provides an upper bound on the probability

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

which coincides with the asymptotic limit in (2.107). The analysis in Appendix 2.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.

2.5.4 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., [79, p. 473]): Let $\{X_n\}_{n \in \mathbb{N}_0}$ ($\mathbb{N}_0 \triangleq \mathbb{N} \cup \{0\}$) be a discrete-time Markov chain taking values in a countable state space $\mathcal{S}$ with transition matrix $\mathbf{P}$, and let the function $\psi : \mathcal{S} \to \mathcal{S}$ be harmonic (i.e., $\sum_{j \in \mathcal{S}} p_{i,j} \psi(j) = \psi(i), \quad \forall\, i \in \mathcal{S}$), and assume that $E[|\psi(X_n)|] < \infty$ for every n. Then, $\{Y_n, \mathcal{F}_n\}_{n \in \mathbb{N}_0}$ is a martingale where $Y_n \triangleq \psi(X_n)$ and $\{\mathcal{F}_n\}_{n \in \mathbb{N}_0}$ is the natural filtration. This relation, which follows directly from the Markov property, enables to apply the concentration inequalities in Section 2.3 for harmonic functions of Markov chains when the function ψ is bounded (so that the jumps of the martingale sequence are uniformly bounded).

Exponential deviation bounds for an important class of Markov chains, called Doeblin chains (they are characterized by an exponentially fast convergence to the equilibrium, uniformly in the initial condition) were derived in [80]. These bounds were also shown to be essentially identical to the Hoeffding inequality in the special case of i.i.d. RVs (see [80, Remark 1]).

2.6 Applications in information theory and related topics

2.6.1 Binary hypothesis testing

Binary hypothesis testing for finite alphabet models was analyzed via the method of types, e.g., in [81, Chapter 11] and [82]. 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

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

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

$$\lim_{n \to \infty} \frac{L(X_1, \ldots, X_n)}{n} = D(P_1 || P_2) \tag{2.108}$$

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

$$\lim_{n \to \infty} \frac{L(X_1, \ldots, X_n)}{n} = -D(P_2 || P_1) \tag{2.109}$$

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

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

one decides on hypothesis H_1 if

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

and on hypothesis H_2 if

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

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

Let

$$\alpha_n^{(1)} \triangleq P_1^n\Big(L(X_1, \ldots, X_n) \leq n\overline{\lambda}\Big) \tag{2.110}$$

$$\alpha_n^{(2)} \triangleq P_1^n\Big(L(X_1, \ldots, X_n) \leq n\underline{\lambda}\Big) \tag{2.111}$$

and

$$\beta_n^{(1)} \triangleq P_2^n\Big(L(X_1, \ldots, X_n) \geq n\underline{\lambda}\Big) \tag{2.112}$$

$$\beta_n^{(2)} \triangleq P_2^n\Big(L(X_1, \ldots, X_n) \geq n\overline{\lambda}\Big) \tag{2.113}$$

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

$$P_{\mathrm{e},n}^{(1)} = \pi_1 \alpha_n^{(1)} + \pi_2 \beta_n^{(1)} \tag{2.114}$$

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

$$P_{\mathrm{e},n}^{(2)} = \pi_1 \alpha_n^{(2)} + \pi_2 \beta_n^{(2)} \tag{2.115}$$

is the probability of error.

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., [69, Theorem 3.4.3] and [83, Theorem 6.4] that addresses the case where the decision is made based on a single threshold of the log-likelihood ratio. In this particular case where $\overline{\lambda} = \underline{\lambda} \triangleq \lambda$, the option of erasures does not exist, and $P_{\mathrm{e},n}^{(1)} = P_{\mathrm{e},n}^{(2)} \triangleq P_{\mathrm{e},n}$ is the error probability.

In the considered general case with erasures, let

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

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

$$\lim_{n\to\infty} -\frac{\ln \alpha_n^{(1)}}{n} = I(\lambda_1) \tag{2.116}$$

$$\lim_{n\to\infty} -\frac{\ln \alpha_n^{(2)}}{n} = I(\lambda_2) \tag{2.117}$$

$$\lim_{n\to\infty} -\frac{\ln \beta_n^{(1)}}{n} = I(\lambda_2) - \lambda_2 \tag{2.118}$$

$$\lim_{n\to\infty} -\frac{\ln \beta_n^{(2)}}{n} = I(\lambda_1) - \lambda_1 \tag{2.119}$$

where the rate function I is given by

$$I(r) \triangleq \sup_{t\in\mathbb{R}} \bigl(tr - H(t)\bigr) \tag{2.120}$$

and

$$H(t) = \ln\left(\sum_{x\in\mathcal{X}} P_1(x)^{1-t} P_2(x)^t\right), \quad \forall\, t \in \mathbb{R}. \tag{2.121}$$

The rate function I is convex, lower semi-continuous (l.s.c.) and non-negative (see, e.g., [69] and [83]). Note that

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

where $D_t(P||Q)$ designates Réyni's information divergence of order t [84, Eq. (3.3)], and I in (2.120) is the Fenchel-Legendre transform of H (see, e.g., [69, Definition 2.2.2]).

From (2.114)– (2.119), the exact exponents of $P_{\mathrm{e},n}^{(1)}$ and $P_{\mathrm{e},n}^{(2)}$ are equal to

$$\lim_{n\to\infty} -\frac{\ln P_{\mathrm{e},n}^{(1)}}{n} = \min\Bigl\{I(\lambda_1), I(\lambda_2) - \lambda_2\Bigr\} \tag{2.122}$$

and

$$\lim_{n\to\infty} -\frac{\ln P_{\mathrm{e},n}^{(2)}}{n} = \min\Bigl\{I(\lambda_2), I(\lambda_1) - \lambda_1\Bigr\}. \tag{2.123}$$

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

$$\lim_{n\to\infty} -\frac{\ln P_{\mathrm{e},n}}{n} = \min\Bigl\{I(\lambda), I(\lambda) - \lambda\Bigr\} \tag{2.124}$$

which coincides with the error exponent in [69, Theorem 3.4.3] (or [83, Theorem 6.4]). The optimal threshold for obtaining the best error exponent of the error probability $P_{\mathrm{e},n}$ is equal to zero (i.e., $\lambda = 0$); in this case, the exact error exponent is equal to

$$\begin{aligned} I(0) &= -\min_{0\le t\le 1} \ln\left(\sum_{x\in\mathcal{X}} P_1(x)^{1-t} P_2(x)^t\right) \\ &\triangleq C(P_1, P_2) \end{aligned} \tag{2.125}$$

which is the Chernoff information of the probability measures P_1 and P_2 (see [81, Eq. (11.239)]), and it is symmetric (i.e., $C(P_1, P_2) = C(P_2, P_1)$). Note that, from (2.120), $I(0) = \sup_{t\in\mathbb{R}}\bigl(-H(t)\bigr) = -\inf_{t\in\mathbb{R}}\bigl(H(t)\bigr)$; the minimization in (2.125) 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 (2.121) 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.

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

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

and

$$U_k = \mathbb{E}_{P_1^n}\big[L(X_1, \ldots, X_n) \mid \mathcal{F}_k\big]. \tag{2.126}$$

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

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

In particular

$$U_0 = n D(P_1 || P_2), \tag{2.127}$$

$$U_n = \sum_{i=1}^n \ln \frac{P_1(X_i)}{P_2(X_i)} = L(X_1, \ldots, X_n) \tag{2.128}$$

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

$$U_k - U_{k-1} = \ln \frac{P_1(X_k)}{P_2(X_k)} - D(P_1 || P_2). \tag{2.129}$$

Let

$$d_1 \triangleq \max_{x \in \mathcal{X}} \left| \ln \frac{P_1(x)}{P_2(x)} - D(P_1 || P_2) \right| \tag{2.130}$$

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 (2.129) and (2.130)

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

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

$$\begin{aligned} &\mathbb{E}_{P_1^n}\big[(U_k - U_{k-1})^2 \,|\, \mathcal{F}_{k-1}\big] \\ &= \mathbb{E}_{P_1}\left[\left(\ln \frac{P_1(X_k)}{P_2(X_k)} - D(P_1 || P_2)\right)^2\right] \\ &= \sum_{x \in \mathcal{X}} \left\{ P_1(x) \left(\ln \frac{P_1(x)}{P_2(x)} - D(P_1 || P_2) \right)^2 \right\} \\ &\triangleq \sigma_1^2. \end{aligned} \tag{2.131}$$

Let

$$\varepsilon_{1,1} = D(P_1||P_2) - \overline{\lambda}, \quad \varepsilon_{2,1} = D(P_2||P_1) + \underline{\lambda} \tag{2.132}$$

$$\varepsilon_{1,2} = D(P_1||P_2) - \underline{\lambda}, \quad \varepsilon_{2,2} = D(P_2||P_1) + \overline{\lambda} \tag{2.133}$$

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

$$\begin{aligned} \alpha_n^{(1)} &\triangleq P_1^n\big(L(X_1, \ldots, X_n) \leq n\overline{\lambda}\big) \\ &\overset{\text{(a)}}{=} P_1^n(U_n - U_0 \leq -\varepsilon_{1,1}\, n) && (2.134) \\ &\overset{\text{(b)}}{\leq} \exp\left(-n\, D\Big(\frac{\delta_{1,1} + \gamma_1}{1+\gamma_1} \Big|\Big| \frac{\gamma_1}{1+\gamma_1}\Big)\right) && (2.135) \end{aligned}$$

where equality (a) follows from (2.127), (2.128) and (2.132), and inequality (b) follows from Theorem 5 with

$$\gamma_1 \triangleq \frac{\sigma_1^2}{d_1^2}, \quad \delta_{1,1} \triangleq \frac{\varepsilon_{1,1}}{d_1}. \tag{2.136}$$

Note that if $\varepsilon_{1,1} > d_1$ then it follows from (2.129) and (2.130) that $\alpha_n^{(1)}$ is zero; in this case $\delta_{1,1} > 1$, so the divergence in (2.135) 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 (2.126), under hypothesis H_2, let us define the martingale sequence $\{U_k, \mathcal{F}_k\}_{k=0}^n$ with the same filtration and

$$U_k = \mathbb{E}_{P_2^n}\big[L(X_1, \ldots, X_n) \mid \mathcal{F}_k\big], \quad \forall\, k \in \{0, \ldots, n\}. \tag{2.137}$$

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

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

and in particular

$$U_0 = -nD(P_2||P_1), \quad U_n = L(X_1, \ldots, X_n). \tag{2.138}$$

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

$$U_k - U_{k-1} = \ln \frac{P_1(X_k)}{P_2(X_k)} + D(P_2||P_1). \tag{2.139}$$

Let

$$d_2 \triangleq \max_{x \in \mathcal{X}} \left| \ln \frac{P_2(x)}{P_1(x)} - D(P_2||P_1) \right| \tag{2.140}$$

then, the jumps of the latter martingale sequence are uniformly bounded by d_2 and, similarly to (2.131), for every $k \in \{1, \ldots, n\}$

$$\begin{aligned} &\mathbb{E}_{P_2^n}\big[(U_k - U_{k-1})^2 \,|\, \mathcal{F}_{k-1}\big] \\ &= \sum_{x \in \mathcal{X}} \left\{ P_2(x) \left(\ln \frac{P_2(x)}{P_1(x)} - D(P_2||P_1) \right)^2 \right\} \\ &\triangleq \sigma_2^2. && (2.141) \end{aligned}$$

Hence, it follows from Theorem 5 that

$$\begin{aligned} \beta_n^{(1)} &\triangleq P_2^n\big(L(X_1, \ldots, X_n) \geq n\underline{\lambda}\big) \\ &= P_2^n(U_n - U_0 \geq \varepsilon_{2,1}\, n) && (2.142) \\ &\leq \exp\left(-n\, D\Big(\frac{\delta_{2,1} + \gamma_2}{1+\gamma_2} \Big|\Big| \frac{\gamma_2}{1+\gamma_2}\Big)\right) && (2.143) \end{aligned}$$

where the equality in (2.142) holds due to (2.138) and (2.132), and (2.143) follows from Theorem 5 with

$$\gamma_2 \triangleq \frac{\sigma_2^2}{d_2^2}, \quad \delta_{2,1} \triangleq \frac{\varepsilon_{2,1}}{d_2} \tag{2.144}$$

and d_2, σ_2 are introduced, respectively, in (2.140) and (2.141).

From (2.114), (2.135) and (2.143), the exponent of the probability of either having an error or an erasure is lower bounded by

$$\lim_{n\to\infty} -\frac{\ln P_{\mathrm{e},n}^{(1)}}{n} \geq \min_{i=1,2} D\Big(\frac{\delta_{i,1}+\gamma_i}{1+\gamma_i}\Big|\Big|\frac{\gamma_i}{1+\gamma_i}\Big). \tag{2.145}$$

Similarly to the above analysis, one gets from (2.115) and (2.133) that the error exponent is lower bounded by

$$\lim_{n\to\infty} -\frac{\ln P_{\mathrm{e},n}^{(2)}}{n} \geq \min_{i=1,2} D\Big(\frac{\delta_{i,2}+\gamma_i}{1+\gamma_i}\Big|\Big|\frac{\gamma_i}{1+\gamma_i}\Big) \tag{2.146}$$

where

$$\delta_{1,2} \triangleq \frac{\varepsilon_{1,2}}{d_1}, \quad \delta_{2,2} \triangleq \frac{\varepsilon_{2,2}}{d_2}. \tag{2.147}$$

For the case of a single threshold (i.e., $\overline{\lambda} = \underline{\lambda} \triangleq \lambda$) then (2.145) and (2.146) coincide, and one obtains that the error exponent satisfies

$$\lim_{n\to\infty} -\frac{\ln P_{\mathrm{e},n}}{n} \geq \min_{i=1,2} D\Big(\frac{\delta_i+\gamma_i}{1+\gamma_i}\Big|\Big|\frac{\gamma_i}{1+\gamma_i}\Big) \tag{2.148}$$

where δ_i is the common value of $\delta_{i,1}$ and $\delta_{i,2}$ (for $i = 1, 2$). In this special case, the zero threshold is optimal (see, e.g., [69, p. 93]), which then yields that (2.148) is satisfied with

$$\delta_1 = \frac{D(P_1||P_2)}{d_1}, \quad \delta_2 = \frac{D(P_2||P_1)}{d_2} \tag{2.149}$$

with d_1 and d_2 from (2.130) and (2.140), respectively. The right-hand side of (2.148) forms a lower bound on Chernoff information which is the exact error exponent for this special case.

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

From (2.129), (2.130), (2.134), (2.136), and Azuma's inequality

$$\alpha_n^{(1)} \leq \exp\Big(-\frac{\delta_{1,1}^2 n}{2}\Big) \tag{2.150}$$

and, similarly, from (2.139), (2.140), (2.142), (2.144), and Azuma's inequality

$$\beta_n^{(1)} \leq \exp\Big(-\frac{\delta_{2,1}^2 n}{2}\Big). \tag{2.151}$$

From (2.111), (2.113), (2.133), (2.147) and Azuma's inequality

$$\alpha_n^{(2)} \leq \exp\biggl(-\frac{\delta_{1,2}^2 n}{2}\biggr) \tag{2.152}$$

$$\beta_n^{(2)} \leq \exp\biggl(-\frac{\delta_{2,2}^2 n}{2}\biggr). \tag{2.153}$$

Therefore, it follows from (2.114), (2.115) and (2.150)–(2.153) that the resulting lower bounds on the exponents of $P_{\mathrm{e},n}^{(1)}$ and $P_{\mathrm{e},n}^{(2)}$ are

$$\lim_{n\to\infty} -\frac{\ln P_{\mathrm{e},n}^{(j)}}{n} \geq \min_{i=1,2} \frac{\delta_{i,j}^2}{2}, \quad j = 1,2 \tag{2.154}$$

as compared to (2.145) and (2.146) which give, for $j = 1,2$,

$$\lim_{n\to\infty} -\frac{\ln P_{\mathrm{e},n}^{(j)}}{n} \geq \min_{i=1,2} D\Bigl(\frac{\delta_{i,j}+\gamma_i}{1+\gamma_i}\Big|\Big|\frac{\gamma_i}{1+\gamma_i}\Bigr). \tag{2.155}$$

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

$$\lim_{n\to\infty} -\frac{\ln P_{\mathrm{e},n}^{(j)}}{n} \geq \min_{i=1,2} \frac{\delta_i^2}{2} \tag{2.156}$$

with the values of δ_1 and δ_2 in (2.149).

The lower bounds on the exponents in (2.154) and (2.155) are compared in the following. Note that the lower bounds in (2.154) are loosened as compared to those in (2.155) since they follow, respectively, from Azuma's inequality and its improvement in Theorem 5.

The divergence in the exponent of (2.155) is equal to

$$\begin{aligned} & D\Bigl(\frac{\delta_{i,j}+\gamma_i}{1+\gamma_i}\Big|\Big|\frac{\gamma_i}{1+\gamma_i}\Bigr) \\ &= \left(\frac{\delta_{i,j}+\gamma_i}{1+\gamma_i}\right)\ln\left(1+\frac{\delta_{i,j}}{\gamma_i}\right) + \left(\frac{1-\delta_{i,j}}{1+\gamma_i}\right)\ln(1-\delta_{i,j}) \\ &= \frac{\gamma_i}{1+\gamma_i}\left[\left(1+\frac{\delta_{i,j}}{\gamma_i}\right)\ln\Bigl(1+\frac{\delta_{i,j}}{\gamma_i}\Bigr) + \frac{(1-\delta_{i,j})\ln(1-\delta_{i,j})}{\gamma_i}\right]. \end{aligned} \tag{2.157}$$

Lemma 7.

$$(1+u)\ln(1+u) \geq \begin{cases} u + \frac{u^2}{2}, & u \in [-1,0] \\ u + \frac{u^2}{2} - \frac{u^3}{6}, & u \geq 0 \end{cases} \tag{2.158}$$

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 (2.157) and Lemma 7 imply that

$$D\Bigl(\frac{\delta_{i,j}+\gamma_i}{1+\gamma_i}\Big|\Big|\frac{\gamma_i}{1+\gamma_i}\Bigr) \geq \frac{\delta_{i,j}^2}{2\gamma_i} - \frac{\delta_{i,j}^3}{6\gamma_i^2(1+\gamma_i)}. \tag{2.159}$$

Hence, by comparing (2.154) with the combination of (2.155) and (2.159), 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 $\bigl(\max\gamma_i\bigr)^{-1}$ 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 (2.148) and (2.149) is equal to $1.77 \cdot 10^{-2}$, whereas the loosened lower bound in (2.156) 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 (2.135), (2.143) and (2.150)–(2.153), 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.

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 (2.125)), and it will be compared to the two lower bounds on the error exponents that were derived in the previous two subsections.

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

$$J(\theta) \triangleq \mathbb{E}_\theta \left[\frac{\partial}{\partial \theta} \ln P_\theta(x) \right]^2 \tag{2.160}$$

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

$$\lim_{\theta' \to \theta} \frac{D(P_\theta || P_{\theta'})}{(\theta - \theta')^2} = \frac{J(\theta)}{2} \tag{2.161}$$

(note that if it was a relative entropy to base 2 then the right-hand side of (2.161) would have been divided by $\ln 2$, and be equal to $\frac{J(\theta)}{\ln 4}$ as in [81, Eq. (12.364)]).

Proposition 2. Under the above assumptions,

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

$$\lim_{\theta' \to \theta} \frac{C(P_\theta, P_{\theta'})}{(\theta - \theta')^2} = \frac{J(\theta)}{8}. \tag{2.162}$$

- Let
$$E_{\mathrm{L}}(P_\theta, P_{\theta'}) \triangleq \min_{i=1,2} D\Big(\frac{\delta_i+\gamma_i}{1+\gamma_i}\Big|\Big|\frac{\gamma_i}{1+\gamma_i}\Big) \tag{2.163}$$
be the lower bound on the error exponent in (2.148) which corresponds to $P_1 \triangleq P_\theta$ and $P_2 \triangleq P_{\theta'}$, then also
$$\lim_{\theta' \to \theta} \frac{E_{\mathrm{L}}(P_\theta, P_{\theta'})}{(\theta-\theta')^2} = \frac{J(\theta)}{8}. \tag{2.164}$$

- Let
$$\widetilde{E}_{\mathrm{L}}(P_\theta, P_{\theta'}) \triangleq \min_{i=1,2} \frac{\delta_i^2}{2} \tag{2.165}$$
be the loosened lower bound on the error exponent in (2.156) which refers to $P_1 \triangleq P_\theta$ and $P_2 \triangleq P_{\theta'}$. Then,
$$\lim_{\theta' \to \theta} \frac{\widetilde{E}_{\mathrm{L}}(P_\theta, P_{\theta'})}{(\theta-\theta')^2} = \frac{a(\theta)\, J(\theta)}{8} \tag{2.166}$$
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 2.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 exponent (for a concrete example, see the last part of the proof in Appendix 2.C).

Example 12. Consider the index family of of probability mass functions defined over the binary alphabet $\mathcal{X} = \{0,1\}$:
$$P_\theta(0) = 1-\theta, \;\; P_\theta(1) = \theta, \quad \forall\, \theta \in (0,1).$$
From (2.160), 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 (2.162) and (2.164)
$$C(P_{\theta_1}, P_{\theta_2}), E_{\mathrm{L}}(P_{\theta_1}, P_{\theta_2}) \approx \frac{J(\theta)(\theta_1-\theta_2)^2}{8} = 2.00 \cdot 10^{-4}.$$
Indeed, the exact values of $C(P_{\theta_1}, P_{\theta_2})$ and $E_{\mathrm{L}}(P_{\theta_1}, P_{\theta_2})$ are $2.000 \cdot 10^{-4}$ and $1.997 \cdot 10^{-4}$, respectively.

2.6.2 Minimum distance of binary linear block codes

Consider the ensemble of binary linear block codes of length n and rate R. The average value of the normalized minimum distance is equal to
$$\frac{\mathbb{E}[d_{\min}(\mathcal{C})]}{n} = h_2^{-1}(1-R)$$
where h_2^{-1} designates the inverse of the binary entropy function to the base 2, and the expectation is with respect to the ensemble where the codes are chosen uniformly at random (see [85]).

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

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

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

This leads to the following theorem:

Theorem 10. [The minimum distance of binary linear block codes] Let $\mathcal{C}$ be chosen uniformly at random from the ensemble of binary linear block codes of length n and rate R. Then for every $\alpha > 0$, with probability at least $1 - 2\exp\left(-\frac{\alpha^2}{2}\right)$, the minimum distance of $\mathcal{C}$ is in the interval

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

and it therefore concentrates around its expected value.

Note, however, that some well-known capacity-approaching families of binary linear block codes possess a minimum Hamming distance which grows sub-linearly with the block length n. For example, the class of parallel concatenated convolutional (turbo) codes was proved to have a minimum distance which grows at most like the logarithm of the interleaver length [86].

2.6.3 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 [87]. 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., [12]). As a result of their remarkable performance under practical decoding algorithms, these coding techniques have revolutionized the field of channel coding and they have been incorporated in various digital communication standards during the last decade.

In the following, we consider ensembles of binary LDPC codes. The codes are represented by bipartite graphs where the variable nodes are located on the left side of the graph, and the parity-check nodes are on the right. The parity-check equations that define the linear code are represented by edges connecting each check node with the variable nodes that are involved in the corresponding parity-check equation. The bipartite graphs representing these codes are sparse in the sense that the number of edges in the graph scales linearly with the block length n of the code. Following standard notation, let λ_i and ρ_i denote the fraction of edges attached, respectively, to variable and parity-check nodes of degree i. The

LDPC code ensemble is denoted by LDPC(n, λ, ρ) 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., [88, 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 [89]. 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 [88, Section II-E]. In [88], 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 [88, Corollary 1]. This bound was expressed in terms of the achievable gap to capacity (even under ML decoding) when the communication takes place over a memoryless binary-input output-symmetric channel. More explicitly, it was shown that if ε designates the gap in rate to capacity, then the number of fundamental cycles should grow at least like $\log \frac{1}{\varepsilon}$. Hence, this lower bound remains unbounded as the gap to capacity tends to zero. Consistently with the study in [89] on cycle-free codes, the lower bound on the cardinality of the fundamental system of cycles in [88, Corollary 1] shows quantitatively the necessity of cycles in bipartite graphs which represent good LDPC code ensembles. As a continuation to this work, we present in the following a large-deviations analysis with respect to the cardinality of the fundamental system of cycles for LDPC code ensembles.

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

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

where $E(\mathcal{G})$, $V(\mathcal{G})$ and $c(\mathcal{G})$ denote the edges, vertices and components of $\mathcal{G}$, respectively, and $|A|$ denotes the number of elements of a (finite) set A. Note that for such a bipartite graph $\mathcal{G}$, there are n variable nodes and $m = n(1 - R_{\text{d}})$ parity-check nodes, so there are in total $|V(\mathcal{G})| = n(2 - R_{\text{d}})$ 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 $|E(\mathcal{G})| = m a_{\text{R}}$. Therefore, for a code from the (n, λ, ρ) LDPC code ensemble, the cardinality of the fundamental system of cycles satisfies the equality

$$\beta(\mathcal{G}) = n\big[(1 - R_{\text{d}})a_{\text{R}} - (2 - R_{\text{d}})\big] + c(\mathcal{G}) \tag{2.167}$$

where

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

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

Let

$$E \triangleq |E(\mathcal{G})| = n(1 - R_{\text{d}})a_{\text{R}} \tag{2.168}$$

denote the number of edges of an arbitrary bipartite graph $\mathcal{G}$ from the ensemble (where we refer interchangeably to codes and to the bipartite graphs that represent these codes from the considered ensemble). Let us arbitrarily assign numbers $1, \ldots, E$ to the E edges of $\mathcal{G}$. Based on Remarks 2 and 3, lets construct a martingale sequence $X_0, \ldots, X_E$ where X_i (for $i = 0, 1, \ldots, E$) is a RV that denotes the conditional expected number of components of a bipartite graph $\mathcal{G}$, chosen uniformly at random from the ensemble, given that the first i edges of the graph $\mathcal{G}$ are revealed. Note that the corresponding filtration

$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}_E$ in this case is defined so that $\mathcal{F}_i$ is the σ-algebra that is generated by all the sets of bipartite graphs from the considered ensemble whose first i edges are fixed. For this martingale sequence

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

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 $\alpha \geq 0$

$$\begin{aligned} & \mathbb{P}\left(|c(\mathcal{G}) - \mathbb{E}_{\mathrm{LDPC}(n,\lambda,\rho)}[c(\mathcal{G})]| \geq \alpha E\right) \leq 2e^{-f(\alpha)E} \\ \Rightarrow\ & \mathbb{P}\left(|\beta(\mathcal{G}) - \mathbb{E}_{\mathrm{LDPC}(n,\lambda,\rho)}[\beta(\mathcal{G})]| \geq \alpha E\right) \leq 2e^{-f(\alpha)E} \end{aligned} \tag{2.169}$$

where the last transition follows from (2.167), and the function f was defined in (2.49). Hence, for $\alpha > 1$, this probability is zero (since $f(\alpha) = +\infty$ for $\alpha > 1$). Note that, from (2.167), $\mathbb{E}_{\mathrm{LDPC}(n,\lambda,\rho)}[\beta(\mathcal{G})]$ scales linearly with n. The combination of Eqs. (2.49), (2.168), (2.169) gives the following statement:

Theorem 11. [Concentration result for the cardinality of the fundamental system of cycles] Let $\mathrm{LDPC}(n, \lambda, \rho)$ be the LDPC code ensemble that is characterized by a block length n, and a pair of degree distributions (from the edge perspective) of λ and ρ. Let $\mathcal{G}$ be a bipartite graph chosen uniformly at random from this ensemble. Then, for every $\alpha \geq 0$, the cardinality of the fundamental system of cycles of $\mathcal{G}$, denoted by $\beta(\mathcal{G})$, satisfies the following inequality:

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

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

Remark 12. The loosened version of Theorem 11, which follows from Azuma's inequality, gets the form

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

for every $\alpha \geq 0$, and η as defined in Theorem 11. Note, however, that the exponential decay of the two bounds is similar for values of α close to zero (see the exponents in Azuma's inequality and Corollary 1 in Figure 2.1).

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 ε denotes the fractional gap to capacity under belief-propagation decoding (i.e., $R_{\mathrm{d}} = (1-\varepsilon)C$) [27]. Therefore, for small values of α, the exponential decay rate in the inequality of Theorem 11 scales like $\left(\log \frac{1}{\varepsilon}\right)^{-2}$. This large-deviations result complements the result in [88, 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.

2.6.4 Concentration Theorems for LDPC Code Ensembles over ISI channels

Concentration analysis on the number of erroneous variable-to-check messages for random ensembles of LDPC codes was introduced in [28] and [90] for memoryless channels. It was shown that the performance of an individual code from the ensemble concentrates around the expected (average) value over this

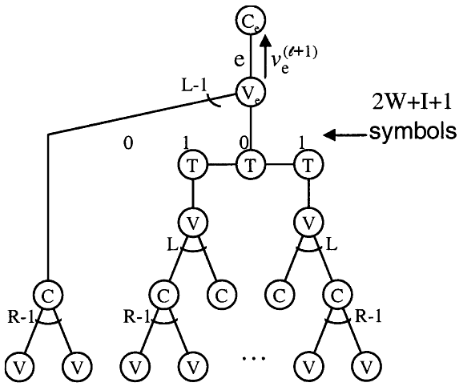

Figure 2.2: Message flow neighborhood of depth 1. In this figure $(I, W, d_{\mathrm{v}} = L, d_{\mathrm{c}} = R) = (1, 1, 2, 3)$

ensemble when the length of the block length of the code grows and that this average behavior converges to the behavior of the cycle-free case. These results were later generalized in [91] for the case of intersymbol-interference (ISI) channels. The proofs of [91, Theorems 1 and 2], which refer to regular LDPC code ensembles, are revisited in the following in order to derive an explicit expression for the exponential rate of the concentration inequality. It is then shown that particularizing the expression for memoryless channels provides a tightened concentration inequality as compared to [28] and [90]. The presentation in this subsection is based on a recent work by Ronen Eshel [92].

The ISI Channel and its message-passing decoding

In the following, we briefly describe the ISI channel and the graph used for its message-passing decoding. For a detailed description, the reader is referred to [91]. Consider a binary discrete-time ISI channel with a finite memory length, denoted by I . The channel output Y_j at time instant j is given by

$$Y_j = \sum_{i=0}^{I} h_i X_{j-i} + N_j, \quad \forall \, j \in \mathbb{Z}$$

where $\{X_j\}$ is the binary input sequence ($X_j \in \{+1, -1\}$), $\{h_i\}_{i=0}^{I}$ refers to the input response of the ISI channel, and $\{N_j\} \sim N(0, \sigma^2)$ is a sequence of i.i.d. Gaussian random variables with zero mean. It is assumed that an information block of length k is encoded by using a regular $(n, d_{\mathrm{v}}, d_{\mathrm{c}})$ LDPC code, and the resulting n coded bits are converted to the channel input sequence before its transmission over the channel. For decoding, we consider the windowed version of the sum-product algorithm when applied to ISI channels (for specific details about this decoding algorithm, the reader is referred to [91] and [93]; in general, it is an iterative message-passing decoding algorithm). The variable-to-check and check-to-variable messages are computed as in the sum-product algorithm for the memoryless case with the difference that a variable node's message from the channel is not only a function of the channel output that corresponds to the considered symbol but also a function of $2W$ neighboring channel outputs and $2W$ neighboring variables nodes as illustrated in Fig. 2.2.

Concentration

It is proved in this sub-section that for a large n, a neighborhood of depth ℓ of a variable-to-check node message is tree-like with high probability. Using the Azuma-Hoeffding inequality and the later result, it is shown that for most graphs and channel realizations, if $\underline{s}$ is the transmitted codeword, then the probability of a variable-to-check message being erroneous after ℓ rounds of message-passing decoding is highly concentrated around its expected value. This expected value is shown to converge to the value of $p^{(\ell)}(\underline{s})$ which corresponds to the cycle-free case.

In the following theorems, we consider an ISI channel and windowed message-passing decoding algorithm, when the code graph is chosen uniformly at random from the ensemble of the graphs with variable and check node degree d_{v} and d_{c}, respectively. Let $\mathcal{N}_{\vec{e}}^{(\ell)}$ denote the neighborhood of depth ℓ of an edge $\vec{e} = (\text{v}, \text{c})$ between a variable-to-check node. Let $N_{\text{c}}^{(\ell)}$, $N_{\text{v}}^{(\ell)}$ and $N_{e}^{(\ell)}$ denote, respectively, the total number of check nodes, variable nodes and code related edges in this neighborhood. Similarly, let $N_{Y}^{(\ell)}$ denote the number of variable-to-check node messages in the directed neighborhood of depth ℓ of a received symbol of the channel.

Theorem 12. [Probability of a neighborhood of depth ℓ of a variable-to-check node message to be tree-like for channels with ISI] Let $P_{\bar{t}}^{(\ell)} \equiv \Pr\left\{\mathcal{N}_{\vec{e}}^{(\ell)} \text{ not a tree}\right\}$ denote the probability that the sub-graph $\mathcal{N}_{\vec{e}}^{(\ell)}$ is not a tree (i.e., it does not contain cycles). Then, there exists a positive constant $\gamma \triangleq \gamma(d_{\text{v}}, d_{\text{c}}, \ell)$ that does not depend on the block-length n such that $P_{\bar{t}}^{(\ell)} \leq \frac{\gamma}{n}$. More explicitly, one can choose $\gamma(d_{\text{v}}, d_{\text{c}}, \ell) \triangleq \left(N_{\text{v}}^{(\ell)}\right)^2 + \left(\frac{d_{\text{c}}}{d_{\text{v}}} \cdot N_{\text{c}}^{(\ell)}\right)^2$.

Proof. This proof forms a straightforward generalization of the proof in [28] (for binary-input output-symmetric memoryless channels) to binary-input ISI channels. A detailed proof is available in [92]. □

The following concentration inequalities follow from Theorem 12 and the Azuma-Hoeffding inequality:

Theorem 13. [Concentration of the number of erroneous variable-to-check messages for channels with ISI] Let $\underline{s}$ be the transmitted codeword. Let $Z^{(\ell)}(\underline{s})$ be the number of erroneous variable-to-check messages after ℓ rounds of the windowed message-passing decoding algorithm when the code graph is chosen uniformly at random from the ensemble of the graphs with variable and check node degrees d_{v} and d_{c}, respectively. Let $p^{(\ell)}(\underline{s})$ be the expected fraction of incorrect messages passed through an edge with a tree-like directed neighborhood of depth ℓ. Then, there exist some positive constants β and γ that do not depend on the block-length n such that

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

$$\mathbb{P}\left(\left|\frac{Z^{(\ell)}(\underline{s})}{nd_v} - \frac{\mathbb{E}[Z^{(\ell)}(\underline{s})]}{nd_v}\right| > \epsilon/2\right) \leq 2e^{-\beta\epsilon^2 n}. \tag{2.170}$$

[**Convergence of expectation to the cycle-free case**] For any $\epsilon > 0$ and $n > \frac{2\gamma}{\epsilon}$, we have a.s.

$$\left|\frac{\mathbb{E}[Z^{(\ell)}(\underline{s})]}{nd_v} - p^{(\ell)}(\underline{s})\right| \leq \epsilon/2. \tag{2.171}$$

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

$$\mathbb{P}\left(\left|\frac{Z^{(\ell)}(\underline{s})}{nd_v} - p^{(\ell)}(\underline{s})\right| > \epsilon\right) \leq 2e^{-\beta\epsilon^2 n}. \tag{2.172}$$

More explicitly, it holds for

$$\beta \triangleq \beta(d_{\text{v}}, d_{\text{c}}, \ell) = \frac{d_{\text{v}}^2}{8\left(4d_{\text{v}}(N_{e}^{(\ell)})^2 + (N_{Y}^{(\ell)})^2\right)},$$

and

$$\gamma \triangleq \gamma(d_{\rm v}, d_{\rm c}, \ell) = \big(N_{\rm v}^{(\ell)}\big)^2 + \big(\frac{d_{\rm c}}{d_{\rm v}} \cdot N_{\rm c}^{(\ell)}\big)^2.$$

Proof. From the triangle inequality, we have

$$\begin{aligned} &\mathbb{P}\left(\left|\frac{Z^{(\ell)}(\underline{s})}{nd_{\rm v}} - p^{(\ell)}(\underline{s})\right| > \epsilon\right) \\ &\leq \mathbb{P}\left(\left|\frac{Z^{(\ell)}(\underline{s})}{nd_{\rm v}} - \frac{\mathbb{E}[Z^{(\ell)}(\underline{s})]}{nd_{\rm v}}\right| > \epsilon/2\right) + \mathbb{P}\left(\left|\frac{\mathbb{E}[Z^{(\ell)}(\underline{s})]}{nd_{\rm v}} - p^{(\ell)}(\underline{s})\right| > \epsilon/2\right). \end{aligned} \tag{2.173}$$

If inequality (2.171) holds a.s., then $\mathbb{P}\left(\left|\frac{Z^{(\ell)}(\underline{s})}{nd_{\rm v}} - p^{(\ell)}(\underline{s})\right| > \epsilon/2\right) = 0$; therefore, using (2.173), we deduce that (2.172) follows from (2.170) and (2.171) for any $\epsilon > 0$ and $n > \frac{2\gamma}{\epsilon}$. We start by proving (2.170). For an arbitrary sequence $\underline{s}$, the random variable $Z^{(\ell)}(\underline{s})$ denotes the number of incorrect variable-to-check node messages among all $nd_{\rm v}$ variable-to-check node messages passed in the ℓth iteration for a particular graph $\mathcal{G}$ and decoder-input $\underline{Y}$. Let us form a martingale by first exposing the $nd_{\rm v}$ edges of the graph one by one, and then exposing the n received symbols Y_i one by one. Let $\underline{a}$ denote the sequence of the $nd_{\rm v}$ variable-to-check node edges of the graph, followed by the sequence of the n received symbols at the channel output. For $i = 0, \ldots n(d_{\rm v}+1)$, let the RV $\widetilde{Z}_i \triangleq \mathbb{E}[Z^{(\ell)}(\underline{s})|a_1, \ldots a_i]$ be defined as the conditional expectation of $Z^{(\ell)}(\underline{s})$ given the first i elements of the sequence $\underline{a}$. Note that it forms a martingale sequence (see Remark 2) where $\widetilde{Z}_0 = \mathbb{E}[Z^{(\ell)}(\underline{s})]$ and $\widetilde{Z}_{n(d_{\rm v}+1)} = Z^{(\ell)}(\underline{s})$. Hence, getting an upper bound on the sequence of differences $|\widetilde{Z}_{i+1} - \widetilde{Z}_i|$ enables to apply the Azuma-Hoeffding inequality to prove concentration around the expected value $\widetilde{Z}_0$. To this end, lets consider the effect of exposing an edge of the graph. Consider two graphs $\mathcal{G}$ and $\widetilde{\mathcal{G}}$ whose edges are identical except for an exchange of an endpoint of two edges. A variable-to-check message is affected by this change if at least one of these edges is included in its directed neighborhood of depth ℓ.

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

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

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

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

variable-to-check node edges. Therefore, a change of a received output symbol can affect up to $N_Y^{(\ell)}$ variable-to-check node messages. We conclude that $|\widetilde{Z}_{i+1} - \widetilde{Z}_i| \leq 2N_e^{(\ell)}$ for the first $nd_{\rm v}$ exposures, and $|\widetilde{Z}_{i+1} - \widetilde{Z}_i| \leq N_Y^{(\ell)}$ for the last n exposures. By applying the Azuma-Hoeffding inequality, it follows that

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

and a comparison of this concentration inequality to (2.170) gives that

$$\frac{1}{\beta} = \frac{8\left(4d_{\text{v}}(N_e^{(\ell)})^2 + (N_Y^{(\ell)})^2\right)}{d_{\text{v}}^2}. \tag{2.174}$$

Next, proving inequality (2.171) relies on concepts from [28] and [91]. Let $\mathbb{E}[Z_i^{(\ell)}(\underline{s})]$ ($i \in \{1, \ldots, nd_{\text{v}}\}$) be the expected number of incorrect messages passed along edge $\overrightarrow{e_i}$ after ℓ rounds, where the average is w.r.t. all realizations of graphs and all output symbols from the channel. Then, by the symmetry in the graph construction and by the linearity of the expectation, it follows that

$$\mathbb{E}[Z^{(\ell)}(\underline{s})] = \sum_{i \in [nd_{\text{v}}]} \mathbb{E}[Z_i^{(\ell)}(\underline{s})] = nd_{\text{v}}\mathbb{E}[Z_1^{(\ell)}(\underline{s})]. \tag{2.175}$$

From Bayes rule

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

As shown in Theorem 12, $P_{\overline{t}}^{(\ell)} \leq \frac{\gamma}{n}$ where γ is a positive constant independent of n. Furthermore, we have $\mathbb{E}[Z_1^{(\ell)}(\underline{s}) \,|\, \text{neighborhood is tree}] = p^{(\ell)}(\underline{s})$, so

$$\begin{aligned}
\mathbb{E}[Z_1^{(\ell)}(\underline{s})] &\leq (1 - P_{\overline{t}}^{(\ell)})p^{(\ell)}(\underline{s}) + P_{\overline{t}}^{(\ell)} \leq p^{(\ell)}(\underline{s}) + P_{\overline{t}}^{(\ell)} \\
\mathbb{E}[Z_1^{(\ell)}(\underline{s})] &\geq (1 - P_{\overline{t}}^{(\ell)})p^{(\ell)}(\underline{s}) \geq p^{(\ell)}(\underline{s}) - P_{\overline{t}}^{(\ell)}.
\end{aligned} \tag{2.176}$$

Using (2.175), (2.176) and $P_{\overline{t}}^{(\ell)} \leq \frac{\gamma}{n}$ gives that

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

Hence, if $n > \frac{2\gamma}{\epsilon}$, then (2.171) holds. □

The concentration result proved above is a generalization of the results given in [28] for a binary-input output-symmetric memoryless channel. One can degenerate the expression of $\frac{1}{\beta}$ in (2.174) to the memoryless case by setting $W = 0$ and $I = 0$. Since we exact expressions for $N_e^{(\ell)}$ and $N_Y^{(\ell)}$ are used in the above proof, one can expect a tighter bound as compared to the earlier result $\frac{1}{\beta_{\text{old}}} = 544 d_{\text{v}}^{2\ell-1} d_{\text{c}}^{2\ell}$ given in [28]. For example for $(d_{\text{v}}, d_{\text{c}}, \ell) = (3, 4, 10)$, one gets an improvement by a factor of about 1 million. However, even with this improved expression, the required size of n according to our proof can be absurdly large. This is because the proof is very pessimistic in the sense that it assumes that any change in an edge or the decoder's input introduces an error in every message it affects. This is especially pessimistic if a large ℓ is considered, since as ℓ is increased, each message is a function of many edges and received output symbols from the channel (since the neighborhood grows with ℓ).

The same phenomena of concentration of measures that are proved above for regular LDPC code ensembles can be extended to irregular LDPC code ensembles. In the special case of memoryless binary-input output-symmetric channels, the following theorem was proved by Richardson and Urbanke in [12, pp. 487–490], based on the Azuma-Hoeffding inequality (we use here the same notation for LDPC code ensembles as in the preceding subsection).

Theorem 14. [Concentration of the bit error probability around the ensemble average] Let $\mathcal{C}$, a code chosen uniformly at random from the ensemble LDPC(n, λ, ρ), be used for transmission over a memoryless binary-input output-symmetric (MBIOS) channel characterized by its L-density a_{MBIOS}. Assume that the decoder performs l iterations of message-passing decoding, and let $P_{\text{b}}(\mathcal{C}, a_{\text{MBIOS}}, l)$

denote the resulting bit error probability. Then, for every $\delta > 0$, there exists an $\alpha > 0$ where $\alpha = \alpha(\lambda, \rho, \delta, l)$ (*independent of the block length* n) such that

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

This theorem asserts that all except an exponentially (in the block length) small fraction of codes behave within an arbitrary small δ from the ensemble average (where δ is a positive number that can be chosen arbitrarily small). Therefore, assuming a sufficiently large block length, the ensemble average is a good indicator for the performance of individual codes, and it is therefore reasonable to focus on the design and analysis of capacity-approaching ensembles (via the density evolution technique). This forms a central result in the theory of codes defined on graphs and iterative decoding algorithms.

2.6.5 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 [94, Theorem 4] and [24, Theorem 1]. The following theorem is proved in [94, 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 15. [Concentration of the conditional entropy] Let $\mathcal{C}$ be chosen uniformly at random from the ensemble $\mathrm{LDPC}(n, \lambda, \rho)$. Assume that the transmission of the code $\mathcal{C}$ takes place over a memoryless binary-input output-symmetric (MBIOS) channel. Let $H(\mathbf{X}|\mathbf{Y})$ designate the conditional entropy of the transmitted codeword $\mathbf{X}$ given the received sequence $\mathbf{Y}$ from the channel. Then, for any $\xi > 0$,

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

where $B \triangleq \frac{1}{2(d_{\mathrm{c}}^{\max}+1)^2(1-R_d)}$, $d_{\mathrm{c}}^{\max}$ 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 15 in [94, 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 $\mathrm{LDPC}(n, \lambda, \rho)$. Define the RV

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

which forms the conditional entropy when the transmission takes place over an MBIOS channel whose transition probability is given by $P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{n} p_{Y|X}(y_i|x_i)$ where $p_{Y|X}(y|1) = p_{Y|X}(-y|0)$. Fix an arbitrary order for the $m = n(1 - R_{\mathrm{d}})$ parity-check nodes where R_{d} forms the design rate of the LDPC code ensemble. Let $\{\mathcal{F}_t\}_{t\in\{0,1,\ldots,m\}}$ form a filtration of σ-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 σ-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 (λ, ρ) 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 Ω is the whole sample space of $m \times n$ binary parity-check matrices that are characterized by the pair of degree distributions (λ, ρ)). Accordingly, based on Remarks 2 and 3, let us define the following martingale sequence

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

By construction, $Z_0 = \mathbb{E}[H_{\mathcal{G}}(\mathbf{X}|\mathbf{Y})]$ is the expected value of the conditional entropy for the LDPC code ensemble, and Z_m is the RV that is equal (a.s.) to the conditional entropy of the particular code from the ensemble (see Remark 3). Similarly to [94, 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 [94, Appendix I] by increasing degree. Let $\mathbf{r} = (r_1, r_2, \ldots)$ be the set of parity-check degrees in ascending order, and Γ_i be the fraction of parity-check nodes of degree i. Hence, the first $m_1 = n(1 - R_{\text{d}})\Gamma_{r_1}$ parity-check nodes are of degree r_1, the successive $m_2 = n(1 - R_{\text{d}})\Gamma_{r_2}$ parity-check nodes are of degree r_2, and so on. The $(t+1)$th parity-check will therefore have a well defined degree, to be denoted by r. From the proof in [94, Appendix I]

$$|Z_{t+1} - Z_t| \leq (r+1)\, H_{\mathcal{G}}(\tilde{X}|\mathbf{Y}) \tag{2.177}$$

where $H_{\mathcal{G}}(\tilde{X}|\mathbf{Y})$ is a RV which designates the conditional entropy of a parity-bit $\tilde{X} = X_{i_1} \oplus \ldots \oplus X_{i_r}$ (i.e., $\tilde{X}$ is equal to the modulo-2 sum of some r bits in the codeword $\mathbf{X}$) given the received sequence $\mathbf{Y}$ at the channel output. The proof in [94, Appendix I] was then completed by upper bounding the parity-check degree r by the maximal parity-check degree $d_{\text{c}}^{\max}$, and also by upper bounding the conditional entropy of the parity-bit $\tilde{X}$ by 1. This gives

$$|Z_{t+1} - Z_t| \leq d_{\text{c}}^{\max} + 1 \quad t = 0, 1, \ldots, m-1. \tag{2.178}$$

which then proves Theorem 15 from Azuma's inequality. Note that the d_i's in Theorem 1 are equal to $d_{\text{c}}^{\max} + 1$, and n in Theorem 1 is replaced with the length $m = n(1 - R_{\text{d}})$ of the martingale sequence $\{Z_t\}$ (that is equal to the number of the parity-check nodes in the graph).

In the continuation, we deviate from the proof in [94, Appendix I] in two respects:

- The first difference is related to the upper bound on the conditional entropy $H_{\mathcal{G}}(\tilde{X}|\mathbf{Y})$ in (2.177) where $\tilde{X}$ is the modulo-2 sum of some r bits of the transmitted codeword $\mathbf{X}$ given the channel output $\mathbf{Y}$. Instead of taking the most trivial upper bound that is equal to 1, as was done in [94, 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 (2.177) with $d_{\text{c}}^{\max} + 1$, it is suggested to leave it as is since Azuma's inequality applies to the case where the bounded differences of the martingale sequence are not fixed (see Theorem 1), and since the number of the parity-check nodes of degree r is equal to $n(1 - R_{\text{d}})\Gamma_r$. The effect of this simple modification will be shown in Example 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 $\mathbf{X}$ and $\mathbf{Y}$ designate the transmitted codeword and received sequence at the channel output. Let $\tilde{X} = X_{i_1} \oplus \ldots \oplus X_{i_r}$ be a parity-bit of some r code bits of $\mathbf{X}$. Then, the conditional entropy of $\tilde{X}$ given $\mathbf{Y}$ satisfies

$$H_{\mathcal{G}}(\tilde{X}|\mathbf{Y}) \leq h_2\left(\frac{1 - C^{\frac{r}{2}}}{2}\right). \tag{2.179}$$

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

$$h_2\left(\frac{1 - \left[1 - 2h_2^{-1}(1-C)\right]^r}{2}\right) \tag{2.180}$$

and

$$1 - C^r \tag{2.181}$$

respectively, where h_2^{-1} in (2.180) 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 (2.179) holds with equality (where both sides of (2.179) are zero), whereas the trivial upper bound is 1.

Proof. Since conditioning reduces the entropy, we have $H(\tilde{X}|\mathbf{Y}) \leq H(\tilde{X}|Y_{i_1}, \ldots, Y_{i_r})$. Note that $Y_{i_1}, \ldots, Y_{i_r}$ are the corresponding channel outputs to the channel inputs $X_{i_1}, \ldots X_{i_r}$, where these r bits are used to calculate the parity-bit $\tilde{X}$. Hence, by combining the last inequality with [88, Eq. (17) and Appendix I], it follows that

$$H(\tilde{X}|\mathbf{Y}) \leq 1 - \frac{1}{2\ln 2} \sum_{p=1}^{\infty} \frac{(g_p)^r}{p(2p-1)} \tag{2.182}$$

where (see [88, Eq. (19)])

$$g_p \triangleq \int_0^{\infty} a(l)(1+e^{-l}) \tanh^{2p}\left(\frac{l}{2}\right) dl, \quad \forall\, p \in \mathbb{N} \tag{2.183}$$

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

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

Substituting this inequality in (2.182) gives that

$$\begin{aligned} H(\tilde{X}|\mathbf{Y}) &\leq 1 - \frac{1}{2\ln 2} \sum_{p=1}^{\infty} \frac{C^{pr}}{p(2p-1)} \\ &= h_2\left(\frac{1 - C^{\frac{r}{2}}}{2}\right) \end{aligned} \tag{2.184}$$

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

$$h_2(x) = 1 - \frac{1}{2\ln 2} \sum_{p=1}^{\infty} \frac{(1-2x)^{2p}}{p(2p-1)}, \quad 0 \leq x \leq 1. \tag{2.185}$$

This proves the result in (2.179).

The tightened bound on the conditional entropy for the BSC is obtained from (2.182) and the equality

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

which holds for the BSC (see [88, Eq. (97)]). This replaces C on the right-hand side of (2.184) with $\bigl(1 - 2h_2^{-1}(1-C)\bigr)^2$, thus leading to the tightened bound in (2.180).

The tightened result for the BEC follows from (2.182) where, from (2.183),

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

(see [88, Appendix II]). Substituting g_p into the right-hand side of (2.182) gives (2.180) (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 (2.177)

$$|Z_{t+1} - Z_t| \leq (r+1)\, h_2\left(\frac{1 - C^{\frac{r}{2}}}{2}\right) \tag{2.186}$$

with the corresponding two improvements for the BSC and BEC (where the second term on the right-hand side of (2.186) is replaced by (2.180) and (2.181), respectively). This improves the loosened bound of $(d_{\text{c}}^{\max} + 1)$ in [94, Appendix I]. From (2.186) and Theorem 1, we obtain the following tightened version of the concentration inequality in Theorem 15.

Theorem 16. [**A tightened concentration inequality for 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 $H(\mathbf{X}|\mathbf{Y})$ designate the conditional entropy of the transmitted codeword $\mathbf{X}$ given the received sequence $\mathbf{Y}$ at the channel output. Then, for every $\xi > 0$,

$$\mathbb{P}\big(\big|H(\mathbf{X}|\mathbf{Y}) - \mathbb{E}_{\text{LDPC}(n,\lambda,\rho)}[H(\mathbf{X}|\mathbf{Y})]\big| \geq \xi\sqrt{n}\,\big) \leq 2\exp(-B\xi^2) \tag{2.187}$$

where

$$B \triangleq \frac{1}{2(1-R_{\text{d}})\sum_{i=1}^{d_{\text{c}}^{\max}} \left\{(i+1)^2\, \Gamma_i \left[h_2\left(\frac{1-C^{\frac{i}{2}}}{2}\right)\right]^2\right\}} \tag{2.188}$$

and $d_{\text{c}}^{\max}$ 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 (2.187) can be improved (i.e., increased), respectively, to

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

and

$$B \triangleq \frac{1}{2(1-R_{\text{d}})\sum_{i=1}^{d_{\text{c}}^{\max}} \left\{(i+1)^2\, \Gamma_i\, (1-C^i)^2\right\}}\,. \tag{2.189}$$

Remark 15. From (2.188), Theorem 16 indeed yields a stronger concentration inequality than Theorem 15.

Remark 16. In the limit where $C \to 1$ bit per channel use, it follows from (2.188) that if $d_{\text{c}}^{\max} < \infty$ then $B \to \infty$. This is in contrast to the value of B in Theorem 15 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 16 is tight in this case.

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

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

Proof. See Appendix 2.D. □

This is in contrast to the value of B in Theorem 15 which vanishes when $d_{\text{c}}^{\max} = \infty$, and therefore Theorem 15 is not informative in this case (see Example 14).

Example 13. [Comparison of Theorems 15 and 16 for right-regular LDPC code ensembles] In the following, we exemplify the improvement in the tightness of Theorem 16 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 [12, Example 4.38] which expresses the capacity of the BIAWGNC in terms of the standard deviation σ 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 $d_{\mathrm{c}} = 20$), then B in Theorem 16 is improved by a factor of

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

This implies that the inequality in Theorem 16 is satisfied with a block length that is 5.134 times shorter than the block length which corresponds to Theorem 15. For the BEC, the result is improved by a factor of

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

due to the tightened value of B in (2.189) as compared to Theorem 15.

Example 14. [Comparison of Theorems 15 and 16 for a heavy-tail Poisson distribution (Tornado codes)] In the following, we compare Theorems 15 and 16 for Tornado LDPC code ensembles. This capacity-achieving sequence for the BEC refers to the heavy-tail Poisson distribution, and it was introduced in [27, Section IV], [95] (see also [12, Problem 3.20]). We rely in the following on the analysis in [88, 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 ε 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 15, and this theorem therefore is useless for the considered code ensemble. On the other hand, from Theorem 16

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

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

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

where Γ_i and ρ_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

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

where $d_{\mathrm{c}}^{\mathrm{avg}}$ denotes the average parity-check node degree, equality (d) holds since $\lambda'(0) = \lambda_2$, inequality (e) is due to the stability condition for the BEC (where $p\lambda'(0)\rho'(1) < 1$ is a necessary condition for reliable communication on the BEC under BP decoding), and finally equality (f) follows from the analysis in [88, Appendix VI] (an upper bound on λ_2 is derived in [88, Eq. (120)], and the average parity-check node degree scales like $\log \frac{1}{\varepsilon}$). Hence, from the above chain of inequalities and (2.188), it follows that for a small gap to capacity, the parameter B in Theorem 16 scales (at least) like

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

Theorem 16 is therefore useful for the analysis of this LDPC code ensemble. As is shown above, the parameter B in (2.188) tends to zero rather slowly as we let the fractional gap ε tend to zero (which therefore demonstrates a rather fast concentration in Theorem 16).

Example 15. This Example forms a direct continuation of Example 13 for the $(n, d_{\mathrm{v}}, d_{\mathrm{c}})$ regular LDPC code ensembles where $d_{\mathrm{v}} = 2$ and $d_{\mathrm{c}} = 20$. With the settings in this example, Theorem 15 gives that

$$\mathbb{P}\left(\left|H(\mathbf{X}|\mathbf{Y}) - \mathbb{E}_{\mathrm{LDPC}(n,\lambda,\rho)}[H(\mathbf{X}|\mathbf{Y})]\right| \geq \xi\sqrt{n}\right) \leq 2\exp(-0.0113\,\xi^2), \quad \forall\, \xi > 0. \tag{2.190}$$

As was mentioned already in Example 13, the exponential inequalities in Theorem 16 achieve an improvement in the exponent of Theorem 15 by factors 5.134 and 9.051 for the BIAWGNC and BEC, respectively. One therefore obtains from the concentration inequalities in Theorem 16 that, for every $\xi > 0$,

$$\mathbb{P}\left(\left|H(\mathbf{X}|\mathbf{Y}) - \mathbb{E}_{\mathrm{LDPC}(n,\lambda,\rho)}[H(\mathbf{X}|\mathbf{Y})]\right| \geq \xi\sqrt{n}\right) \leq \begin{cases} 2\exp(-0.0580\,\xi^2), & \text{(BIAWGNC)} \\ 2\exp(-0.1023\,\xi^2), & \text{(BEC)} \end{cases}. \tag{2.191}$$

2.6.6 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 [29, 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 17. [Expansion of random regular bipartite graphs] Let $\mathcal{G}$ be chosen uniformly at random from the regular ensemble $\mathrm{LDPC}(n, x^{l-1}, x^{r-1})$. Let $\alpha \in (0, 1)$ and $\delta > 0$ be fixed. Then, with probability at least $1 - \exp(-\delta n)$, all sets of αn variables in $\mathcal{G}$ have a number of neighbors that is at least

$$n\left[\frac{l\big(1 - (1-\alpha)^r\big)}{r} - \sqrt{2l\alpha\left(h(\alpha) + \delta\right)}\right] \tag{2.192}$$

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 αn variable nodes is equal to

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

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

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

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

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 2.2.1,

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

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

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

2.6.7 Concentration of the crest-factor for OFDM signals

Orthogonal-frequency-division-multiplexing (OFDM) is a modulation that converts a high-rate data stream into a number of low-rate steams that are transmitted over parallel narrow-band channels. OFDM is widely used in several international standards for digital audio and video broadcasting, and for wireless local area networks. For a textbook providing a survey on OFDM, see e.g. [96, 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; for a recent comprehensive tutorial that considers the problem of the high peak to average power ratio (PAPR) of OFDM signals and some related issues, the reader is referred to [97]. The high PAPR of OFDM signals makes their transmission 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.

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

$$s(t) = \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} X_i \exp\Bigl(\frac{j\,2\pi i t}{T}\Bigr), \quad 0 \leq t \leq T. \tag{2.193}$$

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

$$\frac{1}{T}\int_0^T |s(t)|^2 dt = 1. \tag{2.194}$$

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

$$\mathrm{CF}_n(s) \triangleq \max_{0 \le t \le T} |s(t)|. \tag{2.195}$$

Commonly, the impact of nonlinearities is described by the distribution of the crest-factor (CF) of the transmitted signal [98], but its calculation involves time-consuming simulations even for a small number of sub-carriers. From [99, Section 4] and [100], it follows that the CF scales with high probability like $\sqrt{\ln n}$ for large n. In [98, Theorem 3 and Corollary 5], a concentration inequality was derived for the CF of OFDM signals. It states that for an arbitrary $c \geq 2.5$

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

Remark 17. The analysis used to derive this rather strong concentration inequality (see [98, Appendix C]) requires some assumptions on the distribution of the X_i's (see the two conditions in [98, Theorem 3] followed by [98, 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 [98, 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.

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

$$Y_i = \mathbb{E}[\, \mathrm{CF}_n(s) \,|\, X_0, \ldots, X_{i-1}], \quad i = 0, \ldots, n \tag{2.196}$$

Based on a standard construction of martingales, $\{Y_i, \mathcal{F}_i\}_{i=0}^n$ is a martingale where $\mathcal{F}_i$ is the σ-algebra that is generated by the first i symbols $(X_0, \ldots, X_{i-1})$ in (2.193). Hence, $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}_n$ is a filtration. This martingale has also bounded jumps, and

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

for $i \in \{1, \ldots, n\}$ since revealing the additional i-th coordinate X_i affects the CF, as is defined in (2.195), by at most $\frac{2}{\sqrt{n}}$ (see the first part of Appendix 2.E). It therefore follows from Azuma's inequality that, for every $\alpha > 0$,

$$\mathbb{P}(|\mathrm{CF}_n(s) - \mathbb{E}[\mathrm{CF}_n(s)]| \geq \alpha) \leq 2 \exp\left(-\frac{\alpha^2}{8}\right) \tag{2.197}$$

which demonstrates concentration around the expected value.

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 (2.196), Appendix 2.E gives that a.s.

$$|Y_i - Y_{i-1}| \leq \frac{2}{\sqrt{n}}, \quad \mathbb{E}\big[(Y_i - Y_{i-1})^2 | \mathcal{F}_{i-1}\big] \leq \frac{2}{n} \tag{2.198}$$

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

$$\begin{aligned}
&\mathbb{P}(|\mathrm{CF}_n(s) - \mathbb{E}[\mathrm{CF}_n(s)]| \geq \alpha) \\
&= \mathbb{P}(|Y_n - Y_0| \geq \alpha) \\
&= \mathbb{P}(|Z_n - Z_0| \geq \alpha\sqrt{n}) \\
&\leq 2 \exp\left(-\frac{\alpha^2}{4}\left(1 + O\left(\frac{1}{\sqrt{n}}\right)\right)\right) \qquad (2.199)
\end{aligned}$$

(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 (2.197) via Azuma's inequality, and the term which scales like $O\left(\frac{1}{\sqrt{n}}\right)$ on the right-hand side of (2.199) is expressed explicitly for finite n (see Appendix 2.A).

A concentration inequality via Talagrand's method

In his seminal paper [6], 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., [69, Section 2.4.2], [5, Section 4] and [6]). 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 $\mathbf{x}, \mathbf{y}$ be two n-length vectors. The Hamming distance between $\mathbf{x}$ and $\mathbf{y}$ is the number of coordinates where $\mathbf{x}$ and $\mathbf{y}$ disagree, i.e.,

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

where I stands for the indicator function.

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

Definition 4. Let $a = (a_1, \dots, a_n) \in \mathbb{R}_+^n$ (i.e., a is a non-negative vector) satisfy $||a||^2 = \sum_{i=1}^{n} (a_i)^2 = 1$. Then, define

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

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

The following is a special form of Talagrand's inequalities ([1], [5, Chapter 4] and [6]).

Theorem 18 (Talagrand's inequality)**.** Let the random vector $\mathbf{X} = (X_1, \dots, X_n)$ be a vector of independent random variables with X_k taking values in a set A_k, and let $A \triangleq \prod_{k=1}^{n} A_k$. Let $f : A \to \mathbb{R}$ satisfy the condition that, for every $\mathbf{x} \in A$, there exists a non-negative, normalized n-length vector $a = a(x)$ such that

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \sigma d_a(\mathbf{x}, \mathbf{y}), \quad \forall\, \mathbf{y} \in A \qquad (2.200)$$

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

$$\mathbb{P}(|f(X) - m| \geq \alpha) \leq 4 \exp\left(-\frac{\alpha^2}{4\sigma^2}\right) \tag{2.201}$$

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 (2.201) holds if the condition in (2.200) is replaced by

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \sigma d_a(\mathbf{x}, \mathbf{y}), \quad \forall \, \mathbf{y} \in A. \tag{2.202}$$

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

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

where

$$a \triangleq \Bigl(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\Bigr) \tag{2.203}$$

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

$$\mathbb{P}(|\mathrm{CF}_n(s) - m_n| \geq \alpha) \leq 4 \exp\Bigl(-\frac{\alpha^2}{16}\Bigr) \tag{2.204}$$

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 (2.204), one obtains the following result.

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 (2.204)

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

□

Remark 18. This result applies in general to an arbitrary function f satisfying the condition in (2.200), where Talagrand's inequality in (2.201) implies that (see, e.g., [5, Lemma 4.6])

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

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

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

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

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

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

$$\mathbb{P}(|\mathrm{CF}_n(s) - \mathbb{E}[\mathrm{CF}_n(s)]| \ge \alpha) \le 2\exp\Big(-\frac{\alpha^2}{2}\Big) \tag{2.205}$$

which demonstrates concentration of this measure around its expected value. By comparing (2.204) with (2.205), 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 [98] and [100]), 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.

2.6.8 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 [31], [32] and [33] (and mainly on the latest improvements of these achievable rates in [33]).

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 [101, 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.

Martingale inequalities

- The first martingale inequality that will be used in the following is given in (2.41). It was used earlier in this chapter to prove the refinement of the Azuma-Hoeffding inequality in Theorem 5, and it is stated in the following as a theorem:

Theorem 19. 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

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

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

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

- The second martingale inequality that will be used in the following is similar to (2.72) (while removing the assumption that the martingale is conditionally symmetric). It leads to the following theorem:

Theorem 20. 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

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

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

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

Then, for every $t \geq 0$,

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

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 [32] 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 [32, 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 $\mathbf{C}$ of length N and rate R. Let $\mathcal{C} \in \mathbf{C}$ be a codebook in the ensemble. The number of codewords in $\mathcal{C}$ is $M = \lceil \exp(NR) \rceil$. The codewords of a codebook $\mathcal{C}$ are assumed to be independent, and the symbols in each codeword are assumed to be i.i.d. with an arbitrary probability distribution P. An ML decoding error occurs if, given the transmitted message m and the received vector $\mathbf{y}$, there exists another message $m' \neq m$ such that

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

The union bound for an AWGN channel implies that

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

where the function Q is the complementary Gaussian cumulative distribution function (see (2.11)). By using the inequality $Q(x) \leq \frac{1}{2}\exp\big(-\frac{x^2}{2}\big)$ for $x \geq 0$, it gives the loosened bound (by also ignoring the factor of one-half in the bound of Q)

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

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

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

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

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

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

$$\overline{P}_{\text{e}} \leq (M-1)\, \mathbb{E}\left[\exp\left(-\frac{\rho\, \|D\mathbf{u} - D\widetilde{\mathbf{u}}\|_2^2}{8\sigma_\nu^2}\right)\right] \tag{2.206}$$

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

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

$$\mathcal{F}_i \triangleq \sigma(U_1, \widetilde{U}_1, \ldots, U_i, \widetilde{U}_i), \quad \forall\, i \in \{1, \ldots, N\} \tag{2.207}$$

for two randomly selected codewords $\mathbf{u} = (u_1, \ldots, u_N)$, and $\widetilde{\mathbf{u}} = (\tilde{u}_1, \ldots, \tilde{u}_N)$ from the codebook; $\mathcal{F}_i$ is the minimal σ-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 σ-algebra. Furthermore, define the discrete-time martingale $\{X_k, \mathcal{F}_k\}_{k=0}^N$ by

$$X_k = \mathbb{E}[||D\mathbf{u} - D\widetilde{\mathbf{u}}||_2^2 \,|\, \mathcal{F}_k] \tag{2.208}$$

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

$$X_0 = \mathbb{E}\,[||D\mathbf{u} - D\widetilde{\mathbf{u}}||_2^2], \quad X_N = ||D\mathbf{u} - D\widetilde{\mathbf{u}}||_2^2. \tag{2.209}$$

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

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

and the substitution of the last equality into (2.206) gives that

$$\overline{P}_{\text{e}} \leq \exp(NR)\, \exp\left(-\frac{\rho\, \mathbb{E}\big[||D\mathbf{u} - D\widetilde{\mathbf{u}}||_2^2\big]}{8\sigma_\nu^2}\right) \mathbb{E}\left[\exp\left(-\frac{\rho}{8\sigma^2} \cdot \sum_{k=1}^N \xi_k\right)\right]. \tag{2.210}$$

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

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

Due to the channel model (see Eq. (2.227)) and the assumption that the symbols $\{u_i\}$ are i.i.d., it follows that $\text{Var}\,\big([D\mathbf{u}]_k\big)$ is fixed for $k = q, \ldots, N$. Let $D_{\text{v}}(P)$ designate this common value of the variance (i.e., $D_{\text{v}}(P) = \text{Var}\,\big([D\mathbf{u}]_k\big)$ for $k \geq q$), then

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

Let

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

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

$$\overline{P}_{\text{e}} \leq C_\rho(P)\ \exp\left\{-N\left(\frac{\rho\, D_{\text{v}}(P)}{4\sigma_\nu^2} - R\right)\right\}\ \mathbb{E}\left[\exp\left(\frac{\rho}{8\sigma_\nu^2} \cdot \sum_{k=1}^{N} Z_k\right)\right], \quad \forall\, \rho \in [0,1] \tag{2.211}$$

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

$$\begin{aligned} Z_k &= X_{k-1} - X_k \\ &= \mathbb{E}[||D\mathbf{u} - D\widetilde{\mathbf{u}}||_2^2 \,|\, \mathcal{F}_{k-1}] - \mathbb{E}[||D\mathbf{u} - D\widetilde{\mathbf{u}}||_2^2 \,|\, \mathcal{F}_k]. \end{aligned} \tag{2.212}$$

For the derivation of improved achievable rates and error exponents (as compared to [32]), 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 (2.211).

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 (2.211), combined with the exponential martingale inequalities that are introduced in Theorems 19 and 20, one obtains the following bounds:

1. *First Bounding Technique*: From Theorem 19, if

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

holds a.s. for every $k \geq 1$, and $\gamma_2 \triangleq \frac{\sigma^2}{d^2}$, then it follows from (2.211) that for every $\rho \in [0,1]$

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

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

$$R_1(\sigma_\nu^2) \triangleq \max_P\ \max_{\rho \in [0,1]} \left\{\frac{\rho\, D_{\text{v}}(P)}{4\sigma_\nu^2} - \ln\left(\frac{\exp\left(-\frac{\rho\,\gamma_2\, d}{8\sigma_\nu^2}\right) + \gamma_2 \exp\left(\frac{\rho\, d}{8\sigma_\nu^2}\right)}{1+\gamma_2}\right)\right\} \tag{2.213}$$

where the double maximization is performed over the input distribution P and the parameter $\rho \in [0,1]$. The inner maximization in (2.213) can be expressed in closed form, leading to the following simplified expression:

$$R_1(\sigma_\nu^2) = \max_P \begin{cases} D\left(\left(\frac{\gamma_2}{1+\gamma_2} + \frac{2\mathcal{D}_{\text{v}}(P)}{d(1+\gamma_2)}\right) \,||\, \frac{\gamma_2}{1+\gamma_2}\right), & \text{if } D_{\text{v}}(P) < \dfrac{\gamma_2\, d\left(\exp\left(\frac{d(1+\gamma_2)}{8\sigma_\nu^2}\right) - 1\right)}{2\left(1+\gamma_2 \exp\left(\frac{d(1+\gamma_2)}{8\sigma_\nu^2}\right)\right)} \\ \frac{D_{\text{v}}(P)}{4\sigma_\nu^2} - \ln\left(\frac{\exp\left(-\frac{\gamma_2\, d}{8\sigma_\nu^2}\right) + \gamma_2 \exp\left(\frac{d}{8\sigma_\nu^2}\right)}{1+\gamma_2}\right), & \text{otherwise} \end{cases} \tag{2.214}$$

where

$$D(p||q) \triangleq p \ln\left(\frac{p}{q}\right) + (1-p)\ln\left(\frac{1-p}{1-q}\right), \quad \forall\, p,q \in (0,1) \tag{2.215}$$

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 20 and Eq. (2.211), we derive in the following a second achievable rate for random coding under ML decoding. Referring to the martingale-difference sequence $\{Z_k, \mathcal{F}_k\}_{k=1}^N$ in Eqs. (2.207) and (2.212), one obtains from Eq. (2.211) that if for some even number $m \in \mathbb{N}$

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

hold a.s. for some positive constant $d > 0$ and a sequence $\{\mu_l\}_{l=2}^m$, and

$$\gamma_l \triangleq \frac{\mu_l}{d^l} \quad \forall\, l \in \{2, \ldots, m\},$$

then the average error probability satisfies, for every $\rho \in [0,1]$,

$$\overline{P}_{\mathrm{e}} \leq C_\rho(P)\, \exp\left\{-N\left(\frac{\rho\, D_{\mathrm{v}}(P)}{4\sigma_\nu^2} - R\right)\right\} \left[1 + \sum_{l=2}^{m-1} \frac{\gamma_l - \gamma_m}{l!} \left(\frac{\rho d}{8\sigma_\nu^2}\right)^l + \gamma_m \left(\exp\Big(\frac{\rho\, d}{8\,\sigma_\nu^2}\Big) - 1 - \frac{\rho\, d}{8\,\sigma_\nu^2}\right)\right]^N.$$

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

$$R_2(\sigma_\nu^2) \triangleq \max_P \max_{\rho \in [0,1]} \left\{\frac{\rho\, D_{\mathrm{v}}(P)}{4\sigma_\nu^2} - \ln\left(1 + \sum_{l=2}^{m-1} \frac{\gamma_l - \gamma_m}{l!} \left(\frac{\rho d}{8\sigma_\nu^2}\right)^l + \gamma_m \left(\exp\Big(\frac{\rho\, d}{8\,\sigma_\nu^2}\Big) - 1 - \frac{\rho\, d}{8\,\sigma_\nu^2}\right)\right)\right\} \tag{2.216}$$

where, similarly to (2.213), the double maximization in (2.216) is performed over the input distribution P and the parameter $\rho \in [0,1]$.

Achievable rates for random coding

In the following, the achievable rates for random coding over various linear and non-linear channels (with and without memory) are exemplified. In order to assess the tightness of the bounds, we start with a simple example where the mutual information for the given input distribution is known, so that its gap can be estimated (since we use here the union bound, it would have been in place also to compare the achievable rate with the cutoff rate).

1. *Binary-Input AWGN Channel*: Consider the case of a binary-input AWGN channel where

$$Y_k = U_k + \nu_k$$

where $U_i = \pm A$ for some constant $A > 0$ is a binary input, and $\nu_i \sim \mathcal{N}(0, \sigma_\nu^2)$ is an additive Gaussian noise with zero mean and variance σ_ν^2. Since the codewords $\mathbf{U} = (U_1, \ldots, U_N)$ and $\widetilde{\mathbf{U}} = (\widetilde{U}_1, \ldots, \widetilde{U}_N)$ are independent and their symbols are i.i.d., let

$$P(U_k = A) = P(\widetilde{U}_k = A) = \alpha, \quad P(U_k = -A) = P(\widetilde{U}_k = -A) = 1 - \alpha$$

for some $\alpha \in [0,1]$. Since the channel is memoryless and the all the symbols are i.i.d. then one gets from (2.207) and (2.212) that

$$\begin{aligned} Z_k &= \mathbb{E}[||\mathbf{U} - \widetilde{\mathbf{U}}||_2^2 \,|\, \mathcal{F}_{k-1}] - \mathbb{E}[||\mathbf{U} - \widetilde{\mathbf{U}}||_2^2 \,|\, \mathcal{F}_k] \\ &= \left[\sum_{j=1}^{k-1} (U_j - \widetilde{U}_j)^2 + \sum_{j=k}^{N} \mathbb{E}\big[(U_j - \widetilde{U}_j)^2\big]\right] - \left[\sum_{j=1}^{k} (U_j - \widetilde{U}_j)^2 + \sum_{j=k+1}^{N} \mathbb{E}\big[(U_j - \widetilde{U}_j)^2\big]\right] \\ &= \mathbb{E}[(U_k - \widetilde{U}_k)^2] - (U_k - \widetilde{U}_k)^2 \\ &= \alpha(1-\alpha)(-2A)^2 + \alpha(1-\alpha)(2A)^2 - (U_k - \widetilde{U}_k)^2 \\ &= 8\alpha(1-\alpha)A^2 - (U_k - \widetilde{U}_k)^2. \end{aligned}$$

Hence, for every k,

$$Z_k \leq 8\alpha(1-\alpha)A^2 \triangleq d. \tag{2.217}$$

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

$$\begin{aligned} &\mathbb{E}\big[(Z_k)^l \,|\, \mathcal{F}_{k-1}\big] \\ &= \mathbb{E}\big[(Z_k)^l\big] \\ &= \mathbb{E}\Big[\big(8\alpha(1-\alpha)A^2 - (U_k - \widetilde{U}_k)^2\big)^l\Big] \\ &= \big[1 - 2\alpha(1-\alpha)\big]\big(8\alpha(1-\alpha)A^2\big)^l + 2\alpha(1-\alpha)\,\big(8\alpha(1-\alpha)A^2 - 4A^2\big)^l \triangleq \mu_l \end{aligned} \tag{2.218}$$

and therefore, from (2.217) and (2.218), for every $l \in \mathbb{N}$

$$\gamma_l \triangleq \frac{\mu_l}{d^l} = \big[1 - 2\alpha(1-\alpha)\big]\left[1 + (-1)^l \left(\frac{1 - 2\alpha(1-\alpha)}{2\alpha(1-\alpha)}\right)^{l-1}\right]. \tag{2.219}$$

Let us now rely on the two achievable rates for random coding in Eqs. (2.214) and (2.216), 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 (2.217) and (2.219) that

$$D_{\text{v}}(P) = \text{Var}(U_k) = A^2, \quad d = 2A^2, \qquad \gamma_l = \frac{1 + (-1)^l}{2}, \ \ \forall\, l \in \mathbb{N}. \tag{2.220}$$

Based on the first bounding technique that leads to the achievable rate in Eq. (2.214), since the first condition in this equation cannot hold for the set of parameters in (2.220) then the achievable rate in this equation is equal to

$$R_1(\sigma_\nu^2) = \frac{A^2}{4\sigma_\nu^2} - \ln\cosh\Big(\frac{A^2}{4\sigma_\nu^2}\Big)$$

in units of nats per channel use. Let $\text{SNR} \triangleq \frac{A^2}{\sigma_\nu^2}$ designate the signal to noise ratio, then the first achievable rate gets the form

$$R_1'(\text{SNR}) = \frac{\text{SNR}}{4} - \ln\cosh\left(\frac{\text{SNR}}{4}\right). \tag{2.221}$$

It is observed here that the optimal value of ρ in (2.214) is equal to 1 (i.e., $\rho^\star = 1$).

Let us compare it in the following with the achievable rate that follows from (2.216). Let $m \in \mathbb{N}$ be an even number. Since, from (2.220), $\gamma_l = 1$ for all even values of $l \in \mathbb{N}$ and $\gamma_l = 0$ for all odd values of

$l \in \mathbb{N}$, then

$$\begin{aligned} & 1+\sum_{l=2}^{m-1} \frac{\gamma_l-\gamma_m}{l!}\left(\frac{\rho d}{8\sigma_\nu^2}\right)^l + \gamma_m\left(\exp\Bigl(\frac{\rho\, d}{8\,\sigma_\nu^2}\Bigr)-1-\frac{\rho\, d}{8\,\sigma_\nu^2}\right) \\ & = 1-\sum_{l=1}^{\frac{m}{2}-1} \frac{1}{(2l+1)!}\left(\frac{\rho d}{8\sigma_\nu^2}\right)^{2l+1} + \left(\exp\Bigl(\frac{\rho\, d}{8\,\sigma_\nu^2}\Bigr)-1-\frac{\rho\, d}{8\,\sigma_\nu^2}\right) \end{aligned} \tag{2.222}$$

Since the infinite sum $\sum_{l=1}^{\frac{m}{2}-1} \frac{1}{(2l+1)!}\left(\frac{\rho d}{8\sigma_\nu^2}\right)^{2l+1}$ is monotonically increasing with m (where m is even and $\rho \in [0,1]$), then from (2.216), the best achievable rate within this form is obtained in the limit where m is even and $m \to \infty$. In this asymptotic case one gets

$$\begin{aligned} & \lim_{m\to\infty}\left(1+\sum_{l=2}^{m-1} \frac{\gamma_l-\gamma_m}{l!}\left(\frac{\rho d}{8\sigma_\nu^2}\right)^l + \gamma_m\left(\exp\Bigl(\frac{\rho\, d}{8\,\sigma_\nu^2}\Bigr)-1-\frac{\rho\, d}{8\,\sigma_\nu^2}\right)\right) \\ & \overset{(a)}{=} 1-\sum_{l=1}^{\infty} \frac{1}{(2l+1)!}\left(\frac{\rho d}{8\sigma_\nu^2}\right)^{2l+1} + \left(\exp\Bigl(\frac{\rho\, d}{8\,\sigma_\nu^2}\Bigr)-1-\frac{\rho\, d}{8\,\sigma_\nu^2}\right) \\ & \overset{(b)}{=} 1-\left(\sinh\Bigl(\frac{\rho\, d}{8\,\sigma_\nu^2}\Bigr)-\frac{\rho\, d}{8\,\sigma_\nu^2}\right) + \left(\exp\Bigl(\frac{\rho\, d}{8\,\sigma_\nu^2}\Bigr)-1-\frac{\rho\, d}{8\,\sigma_\nu^2}\right) \\ & \overset{(c)}{=} \cosh\Bigl(\frac{\rho\, d}{8\,\sigma_\nu^2}\Bigr) \end{aligned} \tag{2.223}$$

where equality (a) follows from (2.222), 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 (2.216) gives (from (2.220), $\frac{d}{8\sigma_\nu^2} = \frac{A^2}{4\sigma_\nu^2}$)

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

Since the function $f(x) \triangleq x - \ln\cosh(x)$ for $x \in \mathbb{R}$ is monotonic increasing (note that $f'(x) = 1 - \tanh(x) \geq 0$), then the optimal value of $\rho \in [0,1]$ is equal to 1, and therefore the best achievable rate that follows from the second bounding technique in Eq. (2.216) is equal to

$$R_2(\sigma_\nu^2) = \frac{A^2}{4\sigma_\nu^2} - \ln\cosh\Bigl(\frac{A^2}{4\sigma_\nu^2}\Bigr)$$

in units of nats per channel use, and it is obtained in the asymptotic case where we let the even number m tend to infinity. Finally, setting $\text{SNR} = \frac{A^2}{\sigma_\nu^2}$, gives the achievable rate in (2.221), so the first and second achievable rates for the binary-input AWGN channel coincide, i.e.,

$$R_1'(\text{SNR}) = R_2'(\text{SNR}) = \frac{\text{SNR}}{4} - \ln\cosh\left(\frac{\text{SNR}}{4}\right). \tag{2.224}$$

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 (2.224), 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., [12, Example 4.38 on p. 194])

$$\begin{aligned} C(\text{SNR}) = & \ln 2 + (2\,\text{SNR}-1)\,Q(\sqrt{\text{SNR}}) - \sqrt{\frac{2\,\text{SNR}}{\pi}}\,\exp\Bigl(-\frac{\text{SNR}}{2}\Bigr) \\ & + \sum_{i=1}^{\infty}\left\{\frac{(-1)^i}{i(i+1)}\cdot\exp(2i(i+1)\,\text{SNR})\,Q\bigl((1+2i)\sqrt{\text{SNR}}\bigr)\right\} \end{aligned} \tag{2.225}$$

where the Q-function that appears in the infinite series on the right-hand side of (2.225) is the complementary Gaussian cumulative distribution function in (2.11). 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 (2.224) that follows from the martingale approach is shown in Figure 2.3.

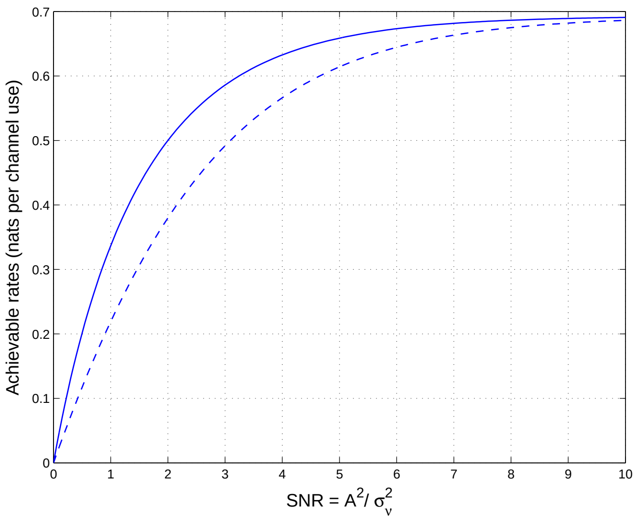

Figure 2.3: A comparison between the symmetric i.i.d. mutual information of the binary-input AWGN channel (solid line) and the common achievable rate in (2.224) (dashed line) that follows from the martingale approach in this subsection.

From the discussion in this subsection, the first and second bounding techniques in Section 2.6.8 lead to the same achievable rate (see (2.224)) 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 (2.220), 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 (2.222) and (2.223)). 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 2.4). We refer in the following to a discrete-time channel model of nonlinear Volterra channels where the input-output channel model is given by

$$y_i = [D\mathbf{u}]_i + \nu_i \tag{2.226}$$

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

$$[D\mathbf{u}]_i = h_0 + \sum_{j=1}^{L} \sum_{i_1=0}^{q} \ldots \sum_{i_j=0}^{q} h_j(i_1, \ldots, i_j) u_{i-i_1} \ldots u_{i-i_j}. \tag{2.227}$$

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

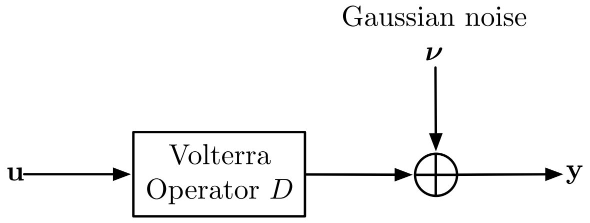

Figure 2.4: The discrete-time Volterra non-linear channel model in Eqs. (2.226) and (2.227) where the channel input and output are $\{U_i\}$ and $\{Y_i\}$, respectively, and the additive noise samples $\{\nu_i\}$, which are added to the distorted input, are i.i.d. with zero mean and variance σ_ν^2.

Table 2.1: Kernels of the 3rd order Volterra system D_1 with memory 2

kernel	$h_1(0)$	$h_1(1)$	$h_1(2)$	$h_2(0,0)$	$h_2(1,1)$	$h_2(0,1)$
value	1.0	0.5	−0.8	1.0	−0.3	0.6

kernel	$h_3(0,0,0)$	$h_3(1,1,1)$	$h_3(0,0,1)$	$h_3(0,1,1)$
value	1.0	−0.5	1.2	0.8

kernel	$h_3(0,1,2)$
value	0.6

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 2.1. 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 σ^2 and thus of γ_2 for the martingale $\{Z_i, \mathcal{F}_i\}_{i=0}^N$. In order to achieve this goal, we have to calculate $|Z_i - Z_{i-1}|$ and $\text{Var}(Z_i|\mathcal{F}_{i-1})$ for all possible combinations of the input samples which contribute to the aforementioned expressions. Thus, the analytic calculation of d and γ_l increases as the system's memory q increases. Numerical results are provided in Figure 2.5 for the case where $\sigma_\nu^2 = 1$. The new achievable rates $R_1^{(2)}(D_1, A, \sigma_\nu^2)$ and $R_2(D_1, A, \sigma_\nu^2)$, which depend on the channel input parameter A, are compared to the achievable rate provided in [32, Fig. 2] and are shown to be larger than the latter.

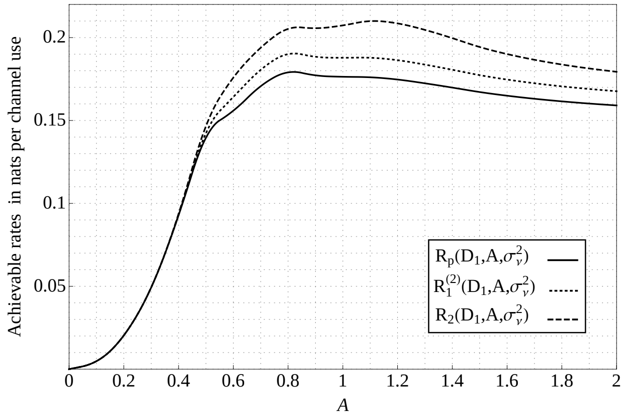

Figure 2.5: Comparison of the achievable rates in this subsection $R_1(D_1, A, \sigma_\nu^2)$ and $R_2^{(2)}(D_1, A, \sigma_\nu^2)$ (where $m = 2$) with the bound $R_p(D_1, A, \sigma_\nu^2)$ of [32, Fig.2] for the nonlinear channel with kernels depicted in Table 2.1 and noise variance $\sigma_\nu^2 = 1$. Rates are expressed in nats per channel use.

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 [102] and [103]), combined with a possible tightening of the union bound under ML decoding (see, e.g., [104]).

2.7 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. A recent interesting avenue that follows from the martingale-based inequalities that are introduced in this chapter is their generalization to random matrices (see, e.g., [105] and [106]).

2.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 (2.73).

Let $\{X_k, \mathcal{F}_k\}_{k=0}^{\infty}$ be a discrete-parameter martingale that satisfies the conditions in Theorem 5. From (2.33)

$$\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) \leq 2\exp\left(-n\, D\left(\frac{\delta'+\gamma}{1+\gamma}\Big|\Big|\frac{\gamma}{1+\gamma}\right)\right) \tag{2.228}$$

where from (2.34)

$$\delta' \triangleq \frac{\frac{\alpha}{\sqrt{n}}}{d} = \frac{\delta}{\sqrt{n}}\,. \tag{2.229}$$

From the right-hand side of (2.228)

$$\begin{aligned}&D\left(\frac{\delta'+\gamma}{1+\gamma}\Big|\Big|\frac{\gamma}{1+\gamma}\right)\\ &= \frac{\gamma}{1+\gamma}\left[\left(1+\frac{\delta}{\gamma\sqrt{n}}\right)\ln\left(1+\frac{\delta}{\gamma\sqrt{n}}\right)+\frac{1}{\gamma}\left(1-\frac{\delta}{\sqrt{n}}\right)\ln\left(1-\frac{\delta}{\sqrt{n}}\right)\right].\end{aligned} \tag{2.230}$$

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

$$\begin{aligned}nD\left(\frac{\delta'+\gamma}{1+\gamma}\Big|\Big|\frac{\gamma}{1+\gamma}\right) &= \frac{\delta^2}{2\gamma} - \frac{\delta^3(1-\gamma)}{6\gamma^2}\,\frac{1}{\sqrt{n}} + \ldots\\ &= \frac{\delta^2}{2\gamma} + O\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

Substituting this into the exponent on the right-hand side of (2.228) gives (2.73).

2.B Analysis related to the moderate deviations principle in Section 2.5.3

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

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

which coincides with the exact asymptotic limit in (2.107). 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 $\mathcal{F}_0 = \{\emptyset, \mathcal{F}\}$.

Analysis related to Azuma's inequality

The martingale sequence $\{S_k, \mathcal{F}_k\}_{k=0}^{n}$ has uniformly bounded jumps, where $|S_k - S_{k-1}| = |X_k| \leq d$ a.s. for every $k \in \{1, \ldots, n\}$. Hence it follows from Azuma's inequality that, for every $\alpha \geq 0$,

$$\mathbb{P}\left(|S_n| \geq \alpha n^{\eta}\right) \leq 2 \exp\left(-\frac{\alpha^2 n^{2\eta-1}}{2d^2}\right)$$

and therefore

$$\lim_{n\to\infty} n^{1-2\eta} \, \ln \mathbb{P}\big(|S_n| \geq \alpha n^{\eta}\big) \leq -\frac{\alpha^2}{2d^2}. \tag{2.231}$$

This differs from the limit in (2.107) where σ^2 is replaced by d^2, so Azuma's inequality does not provide the asymptotic limit in (2.107) (unless $\sigma^2 = d^2$, i.e., $|X_k| = d$ a.s. for every k).

Analysis related to Theorem 5

The analysis here is a slight modification of the analysis in Appendix 2.A with the required adaptation of the calculations for $\eta \in (\frac{1}{2}, 1)$. It follows from Theorem 5 that, for every $\alpha \geq 0$,

$$\mathbb{P}(|S_n| \geq \alpha n^{\eta}) \leq 2 \exp\left(-n\, D\Big(\frac{\delta' + \gamma}{1+\gamma} \Big|\Big| \frac{\gamma}{1+\gamma}\Big)\right)$$

where γ is introduced in (2.34), and δ' in (2.229) is replaced with

$$\delta' \triangleq \frac{\frac{\alpha}{n^{1-\eta}}}{d} = \delta n^{-(1-\eta)} \tag{2.232}$$

due to the definition of δ in (2.34). Following the same analysis as in Appendix 2.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 (2.34), $\frac{\delta^2}{\gamma} = \frac{\alpha^2}{\sigma^2}$)

$$\lim_{n\to\infty} n^{1-2\eta} \, \ln \mathbb{P}\big(|S_n| \geq \alpha n^{\eta}\big) \leq -\frac{\alpha^2}{2\sigma^2}. \tag{2.233}$$

Hence, this upper bound coincides with the exact asymptotic result in (2.107).

2.C Proof of Proposition 2

The proof of (2.162) is based on calculus, and it is similar to the proof of the limit in (2.161) that relates the divergence and Fisher information. For the proof of (2.164), note that

$$C(P_\theta, P_{\theta'}) \geq E_{\mathrm{L}}(P_\theta, P_{\theta'}) \geq \min_{i=1,2} \left\{ \frac{\delta_i^2}{2\gamma_i} - \frac{\delta_i^3}{6\gamma_i^2(1+\gamma_i)} \right\}. \tag{2.234}$$

The left-hand side of (2.234) 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 (2.234) follows from Lemma 7 (see (2.159)) and the definition of E_{L} in (2.163). By definition $\gamma_i \triangleq \frac{\sigma_i^2}{d_i^2}$ and $\delta_i \triangleq \frac{\varepsilon_i}{d_i}$ where, based on (2.149),

$$\varepsilon_1 \triangleq D(P_\theta || P_{\theta'}), \quad \varepsilon_2 \triangleq D(P'_\theta || P_\theta). \tag{2.235}$$

The term on the left-hand side of (2.234) therefore satisfies

$$\begin{aligned}
&\frac{\delta_i^2}{2\gamma_i} - \frac{\delta_i^3}{6\gamma_i^2(1+\gamma_i)} \\
&= \frac{\varepsilon_i^2}{2\sigma_i^2} - \frac{\varepsilon_i^3 d_i^3}{6\sigma_i^2(\sigma_i^2 + d_i^2)} \\
&\geq \frac{\varepsilon_i^2}{2\sigma_i^2}\left(1 - \frac{\varepsilon_i d_i}{3}\right)
\end{aligned}$$

so it follows from (2.234) and the last inequality that

$$C(P_\theta, P_{\theta'}) \geq E_{\mathrm{L}}(P_\theta, P_{\theta'}) \geq \min_{i=1,2} \left\{ \frac{\varepsilon_i^2}{2\sigma_i^2}\left(1 - \frac{\varepsilon_i d_i}{3}\right) \right\}. \tag{2.236}$$

Based on the continuity assumption of the indexed family $\{P_\theta\}_{\theta \in \Theta}$, then it follows from (2.235) that

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

and also, from (2.130) and (2.140) with P_1 and P_2 replaced by P_θ and P'_θ respectively, then

$$\lim_{\theta' \to \theta} d_i = 0, \quad \forall\, i \in \{1, 2\}.$$

It therefore follows from (2.162) and (2.236) that

$$\frac{J(\theta)}{8} \geq \lim_{\theta' \to \theta} \frac{E_{\mathrm{L}}(P_\theta, P_{\theta'})}{(\theta - \theta')^2} \geq \lim_{\theta' \to \theta} \min_{i=1,2} \left\{ \frac{\varepsilon_i^2}{2\sigma_i^2(\theta - \theta')^2} \right\}. \tag{2.237}$$

The idea is to show that the limit on the right-hand side of this inequality is $\frac{J(\theta)}{8}$ (same as the left-hand side), and hence, the limit of the middle term is also $\frac{J(\theta)}{8}$.

$$\begin{aligned}
&\lim_{\theta' \to \theta} \frac{\varepsilon_1^2}{2\sigma_1^2(\theta - \theta')^2} \\
&\overset{\text{(a)}}{=} \lim_{\theta' \to \theta} \frac{D(P_\theta || P_{\theta'})^2}{2\sigma_1^2(\theta - \theta')^2} \\
&\overset{\text{(b)}}{=} \frac{J(\theta)}{4} \lim_{\theta' \to \theta} \frac{D(P_\theta || P_{\theta'})}{\sigma_1^2}
\end{aligned}$$

$$\begin{aligned}
&\overset{\text{(c)}}{=} \frac{J(\theta)}{4} \lim_{\theta' \to \theta} \frac{D(P_\theta || P_{\theta'})}{\sum_{x \in \mathcal{X}} P_\theta(x) \left(\ln \frac{P_\theta(x)}{P_{\theta'}(x)} - D(P_\theta || P_{\theta'})\right)^2} \\
&\overset{\text{(d)}}{=} \frac{J(\theta)}{4} \lim_{\theta' \to \theta} \frac{D(P_\theta || P_{\theta'})}{\sum_{x \in \mathcal{X}} P_\theta(x) \left(\ln \frac{P_\theta(x)}{P_{\theta'}(x)}\right)^2 - D(P_\theta || P_{\theta'})^2} \\
&\overset{\text{(e)}}{=} \frac{J(\theta)^2}{8} \lim_{\theta' \to \theta} \frac{(\theta - \theta')^2}{\sum_{x \in \mathcal{X}} P_\theta(x) \left(\ln \frac{P_\theta(x)}{P_{\theta'}(x)}\right)^2 - D(P_\theta || P_{\theta'})^2} \\
&\overset{\text{(f)}}{=} \frac{J(\theta)^2}{8} \lim_{\theta' \to \theta} \frac{(\theta - \theta')^2}{\sum_{x \in \mathcal{X}} P_\theta(x) \left(\ln \frac{P_\theta(x)}{P_{\theta'}(x)}\right)^2} \\
&\overset{\text{(g)}}{=} \frac{J(\theta)}{8}
\end{aligned} \tag{2.238}$$

where equality (a) follows from (2.235), equalities (b), (e) and (f) follow from (2.161), equality (c) follows from (2.131) 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 (2.160). Similarly, also

$$\lim_{\theta' \to \theta} \frac{\varepsilon_2^2}{2\sigma_2^2 (\theta - \theta')^2} = \frac{J(\theta)}{8}$$

so

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

Hence, it follows from (2.237) that $\lim_{\theta' \to \theta} \frac{E_{\text{L}}(P_\theta, P_{\theta'})}{(\theta - \theta')^2} = \frac{J(\theta)}{8}$. This completes the proof of (2.164).

We prove now equation (2.166). From (2.130), (2.140), (2.149) and (2.165) then

$$\widetilde{E}_{\text{L}}(P_\theta, P_{\theta'}) = \min_{i=1,2} \frac{\varepsilon_i^2}{2d_i^2}$$

with ε_1 and ε_2 in (2.235). Hence,

$$\lim_{\theta' \to \theta} \frac{\widetilde{E}_{\text{L}}(P_\theta, P_{\theta'})}{(\theta' - \theta)^2} \leq \lim_{\theta' \to \theta} \frac{\varepsilon_1^2}{2d_1^2 (\theta' - \theta)^2}$$

and from (2.238) and the last inequality, it follows that

$$\begin{aligned}
&\lim_{\theta' \to \theta} \frac{\widetilde{E}_{\text{L}}(P_\theta, P_{\theta'})}{(\theta' - \theta)^2} \\
&\leq \frac{J(\theta)}{8} \lim_{\theta' \to \theta} \frac{\sigma_1^2}{d_1^2} \\
&\overset{\text{(a)}}{=} \frac{J(\theta)}{8} \lim_{\theta' \to \theta} \frac{\sum_{x \in \mathcal{X}} P_\theta(x) \left(\ln \frac{P_\theta(x)}{P_{\theta'}(x)} - D(P_\theta || P_{\theta'})\right)^2}{\left(\max_{x \in \mathcal{X}} \left|\ln \frac{P_\theta(x)}{P_{\theta'}(x)} - D(P_\theta || P_{\theta'})\right|\right)^2}.
\end{aligned} \tag{2.239}$$

It is clear that the second term on the right-hand side of (2.239) 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 (2.239) can be made arbitrarily close to zero. For a concrete example, let $\alpha \in (0,1)$ be fixed, and $\theta \in \mathbb{R}^+$ be a parameter that defines the following indexed family of probability mass functions over the ternary alphabet $\mathcal{X} = \{0,1,2\}$:

$$P_\theta(0) = \frac{\theta(1-\alpha)}{1+\theta}, \quad P_\theta(1) = \alpha, \quad P_\theta(2) = \frac{1-\alpha}{1+\theta}.$$

Then, it follows by calculus that for this indexed family

$$\lim_{\theta' \to \theta} \frac{\sum_{x \in \mathcal{X}} P_\theta(x) \left(\ln \frac{P_\theta(x)}{P_{\theta'}(x)} - D(P_\theta || P_{\theta'}) \right)^2}{\left(\max_{x \in \mathcal{X}} \left| \ln \frac{P_\theta(x)}{P_{\theta'}(x)} - D(P_\theta || P_{\theta'}) \right| \right)^2} = (1-\alpha)\theta$$

so, for any $\theta \in \mathbb{R}^+$, the above limit can be made arbitrarily close to zero by choosing α close enough to 1. This completes the proof of (2.166), and also the proof of Proposition 2.

2.D Proof of Lemma 8

In order to prove Lemma 8, one needs to show that if $\rho'(1) < \infty$ then

$$\lim_{C \to 1} \sum_{i=1}^{\infty} (i+1)^2 \Gamma_i \left[h_2 \left(\frac{1 - C^{\frac{i}{2}}}{2} \right) \right]^2 = 0 \tag{2.240}$$

which then yields from (2.188) 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

$$d_{\mathrm{c}}^{\mathrm{avg}} = \frac{1}{\sum_{i=1}^{\infty} \frac{\rho_i}{i}} < \infty.$$

The infinite sum $\sum_{i=1}^{\infty} (i+1)^2 \Gamma_i$ converges under the above assumption since

$$\begin{aligned}
&\sum_{i=1}^{\infty} (i+1)^2 \Gamma_i \\
&= \sum_{i=1}^{\infty} i^2 \Gamma_i + 2 \sum_{i=1}^{\infty} i\Gamma_i + \sum_i \Gamma_i \\
&= d_{\mathrm{c}}^{\mathrm{avg}} \left(\sum_{i=1}^{\infty} i\rho_i + 2 \right) + 1 < \infty.
\end{aligned}$$

where the last equality holds since

$$\begin{aligned}
\Gamma_i &= \frac{\frac{\rho_i}{i}}{\int_0^1 \rho(x)\, \mathrm{d}x} \\
&= d_{\mathrm{c}}^{\mathrm{avg}} \left(\frac{\rho_i}{i} \right), \quad \forall\, i \in \mathbb{N}.
\end{aligned}$$

The infinite series in (2.240) 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 (2.240) converges to zero in the limit where $C \to 1$, hence the limit in (2.240) is zero. This completes the proof of Lemma 8.

2.E Proof of the properties in (2.198) for OFDM signals

Consider an OFDM signal from Section 2.6.7. The sequence in (2.196) is a martingale due to basic properties of martingales. From (2.195), for every $i \in \{0, \dots, n\}$

$$Y_i = \mathbb{E}\Big[\max_{0 \le t \le T} \big|s(t; X_0, \dots, X_{n-1})\big| \,\Big|\, X_0, \dots, X_{i-1}\Big].$$

The conditional expectation for the RV Y_{i-1} refers to the case where only $X_0, \dots, X_{i-2}$ are revealed. Let X'_{i-1} and X_{i-1} be independent copies, which are also independent of $X_0, \dots, X_{i-2}, X_i, \dots, X_{n-1}$. Then, for every $1 \le i \le n$,

$$\begin{aligned} Y_{i-1} &= \mathbb{E}\Big[\max_{0 \le t \le T} \big|s(t; X_0, \dots, X'_{i-1}, X_i, \dots, X_{n-1})\big| \,\Big|\, X_0, \dots, X_{i-2}\Big] \\ &= \mathbb{E}\Big[\max_{0 \le t \le T} \big|s(t; X_0, \dots, X'_{i-1}, X_i, \dots, X_{n-1})\big| \,\Big|\, X_0, \dots, X_{i-2}, X_{i-1}\Big]. \end{aligned}$$

Since $|\mathbb{E}(Z)| \le \mathbb{E}(|Z|)$, then for $i \in \{1, \dots, n\}$

$$|Y_i - Y_{i-1}| \le \mathbb{E}_{X'_{i-1}, X_i, \dots, X_{n-1}}\Big[|U - V| \,\Big|\, X_0, \dots, X_{i-1}\Big] \tag{2.241}$$

where

$$\begin{aligned} U &\triangleq \max_{0 \le t \le T} \big|s(t; X_0, \dots, X_{i-1}, X_i, \dots, X_{n-1})\big| \\ V &\triangleq \max_{0 \le t \le T} \big|s(t; X_0, \dots, X'_{i-1}, X_i, \dots, X_{n-1})\big|. \end{aligned}$$

From (2.193)

$$\begin{aligned} |U - V| &\le \max_{0 \le t \le T} \big|s(t; X_0, \dots, X_{i-1}, X_i, \dots, X_{n-1}) - s(t; X_0, \dots, X'_{i-1}, X_i, \dots, X_{n-1})\big| \\ &= \max_{0 \le t \le T} \frac{1}{\sqrt{n}} \left|\big(X_{i-1} - X'_{i-1}\big) \exp\Big(\frac{j\, 2\pi i t}{T}\Big)\right| \\ &= \frac{|X_{i-1} - X'_{i-1}|}{\sqrt{n}}. \end{aligned} \tag{2.242}$$

By assumption, $|X_{i-1}| = |X'_{i-1}| = 1$, and therefore a.s.

$$|X_{i-1} - X'_{i-1}| \le 2 \Longrightarrow |Y_i - Y_{i-1}| \le \frac{2}{\sqrt{n}}.$$

In the following, an upper bound on the conditional variance $\text{Var}(Y_i \,|\, \mathcal{F}_{i-1}) = \mathbb{E}\big[(Y_i - Y_{i-1})^2 \,|\, \mathcal{F}_{i-1}\big]$ is obtained. Since $\big(\mathbb{E}(Z)\big)^2 \le \mathbb{E}(Z^2)$ for a real-valued RV Z, then from (2.241) and (2.242)

$$\mathbb{E}\big[(Y_i - Y_{i-1})^2 \,|\mathcal{F}_{i-1}\big] \le \frac{1}{n} \cdot \mathbb{E}_{X'_{i-1}}\big[|X_{i-1} - X'_{i-1}|^2 \,|\, \mathcal{F}_i\big]$$

where $\mathcal{F}_i$ is the σ-algebra that is generated by $X_0, \ldots, X_{i-1}$. Due to symmetry of the PSK constellation, then

$$\begin{aligned}
&\mathbb{E}\big[(Y_i - Y_{i-1})^2 \,|\, \mathcal{F}_{i-1}\big] \\
&\leq \frac{1}{n}\, \mathbb{E}_{X'_{i-1}}\big[|X_{i-1} - X'_{i-1}|^2 \,|\, \mathcal{F}_i\big] \\
&= \frac{1}{n}\, \mathbb{E}\big[|X_{i-1} - X'_{i-1}|^2 \,|\, X_0, \ldots, X_{i-1}\big] \\
&= \frac{1}{n}\, \mathbb{E}\big[|X_{i-1} - X'_{i-1}|^2 \,|\, X_{i-1}\big] \\
&= \frac{1}{n}\, \mathbb{E}\Big[|X_{i-1} - X'_{i-1}|^2 \,|\, X_{i-1} = e^{\frac{j\pi}{M}}\Big] \\
&= \frac{1}{nM} \sum_{l=0}^{M-1} \Big| e^{\frac{j\pi}{M}} - e^{\frac{j(2l+1)\pi}{M}} \Big|^2 \\
&= \frac{4}{nM} \sum_{l=1}^{M-1} \sin^2\Big(\frac{\pi l}{M}\Big) = \frac{2}{n}
\end{aligned}$$

where the last equality holds since

$$\begin{aligned}
&\sum_{l=1}^{M-1} \sin^2\Big(\frac{\pi l}{M}\Big) = \frac{1}{2} \sum_{l=0}^{M-1} \Big(1 - \cos\Big(\frac{2\pi l}{M}\Big)\Big) \\
&= \frac{M}{2} - \frac{1}{2}\, \mathrm{Re}\bigg\{ \sum_{l=0}^{M-1} e^{j2l\pi/M} \bigg\} \\
&= \frac{M}{2} - \frac{1}{2}\, \mathrm{Re}\bigg\{ \frac{1 - e^{2j\pi}}{1 - e^{j2\pi/M}} \bigg\} = \frac{M}{2}.
\end{aligned}$$

Chapter 3

The Entropy Method, Log-Sobolev and Transportation-Cost Inequalities: Links and Applications in Information Theory

This chapter introduces the entropy method for deriving concentration inequalities for functions of many independent random variables, and exhibits its multiple connections to information theory. The chapter is divided into four parts. The first part of the chapter introduces the basic ingredients of the entropy method and closely related topics, such as the logarithmic-Sobolev inequalities. These topics underlie the so-called functional approach to deriving concentration inequalities. The second part is devoted to a related viewpoint based on probability in metric spaces. This viewpoint centers around the so-called transportation-cost inequalities, which have been introduced into the study of concentration by Marton. The third part gives a brief summary of some results on concentration for dependent random variables, emphasizing the connections to information-theoretic ideas. The fourth part lists several applications of concentration inequalities and the entropy method to problems in information theory. The considered applications include strong converses for several source and channel coding problems, empirical distributions of good channel codes with non-vanishing error probability, and an information-theoretic converse for concentration of measures.

3.1 The main ingredients of the entropy method

As a reminder, we are interested in the following question. Let $X_1, \ldots, X_n$ be n independent random variables, each taking values in a set $\mathcal{X}$. Given a function $f : \mathcal{X}^n \to \mathbb{R}$, we would like to find tight upper bounds on the *deviation probabilities* for the random variable $U = f(X^n)$, i.e., we wish to bound from above the probability $\mathbb{P}(|U - \mathbb{E}U| \geq r)$ for each $r > 0$. Of course, if U has finite variance, then Chebyshev's inequality already gives

$$\mathbb{P}(|U - \mathbb{E}U| \geq r) \leq \frac{\mathsf{var}(U)}{r^2}, \quad \forall\, r > 0. \tag{3.1}$$

However, in many instances a bound like (3.1) is not nearly as tight as one would like, so ideally we aim for Gaussian-type bounds

$$\mathbb{P}(|U - \mathbb{E}U| \geq r) \leq K \exp\left(-\kappa r^2\right), \quad \forall\, r > 0 \tag{3.2}$$

for some constants $K, \kappa > 0$. Whenever such a bound is available, K is a small constant (usually, $K = 2$), while κ depends on the sensitivity of the function f to variations in its arguments.

In the preceding chapter, we have demonstrated the martingale method for deriving Gaussian concentration bounds of the form (3.2). In this chapter, our focus is on the so-called "entropy method," an information-theoretic technique that has become increasingly popular starting with the work of Ledoux [34] (see also [2]). In the following, we will always assume (unless specified otherwise) that the function $f : \mathcal{X}^n \to \mathbb{R}$ and the probability distribution P of X^n are such that

- $U = f(X^n)$ has zero mean: $\mathbb{E}U = \mathbb{E}f(X^n) = 0$
- U is *exponentially integrable*:

$$\mathbb{E}[\exp(\lambda U)] = \mathbb{E}\left[\exp\left(\lambda f(X^n)\right)\right] < \infty, \qquad \forall \lambda \in \mathbb{R} \tag{3.3}$$

[another way of writing this is $\exp(\lambda f) \in L^1(P)$ for all $\lambda \in \mathbb{R}$].

In a nutshell, the entropy method has three basic ingredients:

1. **The Chernoff bounding trick** — using Markov's inequality, the problem of bounding the deviation probability $\mathbb{P}(|U - \mathbb{E}U| \geq r)$ is reduced to the analysis of the *logarithmic moment-generating function* $\Lambda(\lambda) \triangleq \ln \mathbb{E}[\exp(\lambda U)]$, $\lambda \in \mathbb{R}$.

2. **The Herbst argument** — the function $\Lambda(\lambda)$ is related through a simple first-order differential equation to the relative entropy (information divergence) $D(P^{(\lambda f)} \| P)$, where $P = P_{X^n}$ is the probability distribution of X^n and $P^{(\lambda f)}$ is the *tilted probability distribution* defined by

$$\frac{\mathrm{d}P^{(\lambda f)}}{\mathrm{d}P} = \frac{\exp(\lambda f)}{\mathbb{E}[\exp(\lambda f)]} = \exp\left(\lambda f - \Lambda(\lambda)\right). \tag{3.4}$$

If the function f and the probability distribution P are such that

$$D(P^{(\lambda f)} \| P) \leq \frac{c\lambda^2}{2} \tag{3.5}$$

for some $c > 0$, then the Gaussian bound (3.2) holds with $K = 2$ and $\kappa = \frac{1}{2c}$. The standard way to establish (3.5) is through the so-called *logarithmic Sobolev inequalities.*

3. **Tensorization of the entropy** — with few exceptions, it is rather difficult to derive a bound like (3.5) directly. Instead, one typically takes a divide-and-conquer approach: Using the fact that P_{X^n} is a product distribution (by the assumed independence of the X_i's), the divergence $D(P^{(\lambda f)} \| P)$ is bounded from above by a sum of "one-dimensional" (or "local") conditional divergence terms

$$D\Big(P^{(\lambda f)}_{X_i|\bar{X}^i} \Big\| P_{X_i} \Big| P^{(\lambda f)}_{\bar{X}^i}\Big), \qquad i = 1, \dots, n \tag{3.6}$$

where, for each i, $\bar{X}^i \in \mathcal{X}^{n-1}$ denotes the $(n-1)$-tuple obtained from X^n by removing the ith coordinate, i.e., $\bar{X}^i = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$. Despite their formidable appearance, the conditional divergences in (3.6) are easier to handle because, for each given realization $\bar{X}^i = \bar{x}^i$, the ith such term involves a single-variable function $f_i(\cdot|\bar{x}^i) : \mathcal{X} \to \mathbb{R}$ defined by $f_i(y|\bar{x}^i) \triangleq f(x_1, \dots, x_{i-1}, y, x_{i+1}, \dots, x_n)$ and the corresponding tilted distribution $P^{(\lambda f)}_{X_i|\bar{X}^i = \bar{x}^i}$, where

$$\frac{\mathrm{d}P^{(\lambda f)}_{X_i|\bar{X}^i = \bar{x}^i}}{\mathrm{d}P_{X_i}} = \frac{\exp\left(\lambda f_i(\cdot|\bar{x}^i)\right)}{\mathbb{E}\left[\exp\left(\lambda f_i(X_i|\bar{x}^i)\right)\right]}, \qquad \forall \bar{x}^i \in \mathcal{X}^{n-1}. \tag{3.7}$$

In fact, from (3.4) and (3.7), it is easy to see that the conditional distribution $P^{(\lambda f)}_{X_i|\bar{X}^i=\bar{x}^i}$ is nothing but the tilted distribution $P^{(\lambda f_i(\cdot|\bar{x}^i))}_{X_i}$. This simple observation translates into the following: If the function f and the probability distribution $P = P_{X^n}$ are such that there exist constants $c_1, \ldots, c_n > 0$ so that

$$D\Big(P^{(\lambda f_i(\cdot|\bar{x}^i))}_{X_i} \Big\| P_{X_i}\Big) \le \frac{c_i \lambda^2}{2}, \qquad \forall i \in \{1, \ldots, n\}, \bar{x}^i \in \mathcal{X}^{n-1}, \tag{3.8}$$

then (3.5) holds with $c = \sum_{i=1}^n c_i$ (to be shown explicitly later), which in turn gives that

$$\mathbb{P}\Big(\, |f(X^n) - \mathbb{E}f(X^n)| \ge r\Big) \le 2\exp\left(-\frac{r^2}{2\sum_{i=1}^n c_i}\right), \quad r > 0. \tag{3.9}$$

Again, one would typically use logarithmic Sobolev inequalities to verify (3.8).

In the remainder of this section, we shall elaborate on these three ingredients. Logarithimic Sobolev inequalities and their applications to concentration bounds are described in detail in Sections 3.2 and 3.3.

3.1.1 The Chernoff bounding trick

The first ingredient of the entropy method is the well-known Chernoff bounding trick[1]: Using Markov's inequality, for any $\lambda > 0$ we have

$$\begin{aligned} \mathbb{P}(U \ge r) &= \mathbb{P}\big(\exp(\lambda U) \ge \exp(\lambda r)\big) \\ &\le \exp(-\lambda r)\mathbb{E}[\exp(\lambda U)]. \end{aligned}$$

Equivalently, if we define the *logarithmic moment generating function* $\Lambda(\lambda) \triangleq \ln \mathbb{E}[\exp(\lambda U)]$, $\lambda \in \mathbb{R}$, we can write

$$\mathbb{P}(U \ge r) \le \exp\big(\Lambda(\lambda) - \lambda r\big), \qquad \forall \lambda > 0. \tag{3.10}$$

To bound the probability of the lower tail, $\mathbb{P}(U \le -r)$, we follow the same steps, but with $-U$ instead of U. From now on, we will focus on the deviation probability $\mathbb{P}(U \ge r)$.

By means of the Chernoff bounding trick, we have reduced the problem of bounding the deviation probability $\mathbb{P}(U \ge r)$ to the analysis of the logarithmic moment-generating function $\Lambda(\lambda)$. The following properties of $\Lambda(\lambda)$ will be useful later on:

- $\Lambda(0) = 0$

- Because of the exponential integrability of U [cf. (3.3)], $\Lambda(\lambda)$ is infinitely differentiable, and one can interchange derivative and expectation. In particular,

$$\Lambda'(\lambda) = \frac{\mathbb{E}[U\exp(\lambda U)]}{\mathbb{E}[\exp(\lambda U)]} \qquad \text{and} \qquad \Lambda''(\lambda) = \frac{\mathbb{E}[U^2\exp(\lambda U)]}{\mathbb{E}[\exp(\lambda U)]} - \left(\frac{\mathbb{E}[U\exp(\lambda U)]}{\mathbb{E}[\exp(\lambda U)]}\right)^2 \tag{3.11}$$

  Since we have assumed that $\mathbb{E}U = 0$, we have $\Lambda'(0) = 0$ and $\Lambda''(0) = \mathsf{var}(U)$.

- Since $\Lambda(0) = \Lambda'(0) = 0$, we get

$$\lim_{\lambda \to 0} \frac{\Lambda(\lambda)}{\lambda} = 0. \tag{3.12}$$

[1] The name of H. Chernoff is associated with this technique because of his 1952 paper [107]; however, its roots go back to S.N. Bernstein's 1927 textbook on the theory of probability [108].

3.1.2 The Herbst argument

The second ingredient of the entropy method consists in relating this function to a certain relative entropy, and is often referred to as the *Herbst argument* because the basic idea underlying it had been described in an unpublished note by I. Herbst.

Given any function $g : \mathcal{X}^n \to \mathbb{R}$ which is exponentially integrable w.r.t. P, i.e., $\mathbb{E}[\exp(g(X^n))] < \infty$, let us denote by $P^{(g)}$ the *g-tilting* of P:

$$\frac{\mathrm{d}P^{(g)}}{\mathrm{d}P} = \frac{\exp(g)}{\mathbb{E}[\exp(g)]}.$$

Then

$$\begin{aligned}
D\big(P^{(g)} \| P\big) &= \int_{\mathcal{X}^n} \ln\left(\frac{\mathrm{d}P^{(g)}}{\mathrm{d}P}\right) \mathrm{d}P^{(g)} \\
&= \int_{\mathcal{X}^n} \frac{\mathrm{d}P^{(g)}}{\mathrm{d}P} \ln\left(\frac{\mathrm{d}P^{(g)}}{\mathrm{d}P}\right) \mathrm{d}P \\
&= \int_{\mathcal{X}^n} \frac{\exp(g)}{\mathbb{E}[\exp(g)]} \cdot \big(g - \ln \mathbb{E}[\exp(g)]\big)\, \mathrm{d}P \\
&= \frac{1}{\mathbb{E}[\exp(g)]} \int_{\mathcal{X}^n} g \exp(g)\, \mathrm{d}P - \ln \mathbb{E}[\exp(g)] \\
&= \frac{\mathbb{E}[g \exp(g)]}{\mathbb{E}[\exp(g)]} - \ln \mathbb{E}[\exp(g)].
\end{aligned}$$

In particular, if we let $g = tf$ for some $t \neq 0$, then

$$\begin{aligned}
D\big(P^{(tf)} \| P\big) &= \frac{t \cdot \mathbb{E}[f \exp(tf)]}{\mathbb{E}[\exp(tf)]} - \ln \mathbb{E}[\exp(tf)] \\
&= t\Lambda'(t) - \Lambda(t) \\
&= t^2 \left(\frac{\Lambda'(t)}{t} - \frac{\Lambda(t)}{t^2}\right) \\
&= t^2 \frac{\mathrm{d}}{\mathrm{d}t} \left(\frac{\Lambda(t)}{t}\right), \qquad (3.13)
\end{aligned}$$

where in the second line we have used (3.11). Integrating from $t = 0$ to $t = \lambda$ and using (3.12), we get

$$\Lambda(\lambda) = \lambda \int_0^\lambda \frac{D\big(P^{(tf)} \| P\big)}{t^2} \mathrm{d}t. \qquad (3.14)$$

Combining (3.14) with (3.10), we have proved the following:

Proposition 4. Let $U = f(X^n)$ be a zero-mean random variable that is exponentially integrable. Then, for any $r \geq 0$,

$$\mathbb{P}\big(U \geq r\big) \leq \exp\left(\lambda \int_0^\lambda \frac{D(P^{(tf)} \| P)}{t^2} \mathrm{d}t - \lambda r\right), \qquad \forall \lambda > 0. \qquad (3.15)$$

Thus, we have reduced the problem of bounding the deviation probabilities $\mathbb{P}(U \geq r)$ to the problem of bounding the relative entropies $D(P^{(tf)} \| P)$. In particular, we have

Corollary 7. Suppose that the function f and the probability distribution P of X^n are such that

$$D\big(P^{(tf)} \| P\big) \leq \frac{ct^2}{2}, \qquad \forall t > 0 \tag{3.16}$$

for some constant $c > 0$. Then,

$$\mathbb{P}\big(U \geq r\big) \leq \exp\left(-\frac{r^2}{2c}\right), \quad \forall\, r \geq 0. \tag{3.17}$$

Proof. Using (3.16) to upper-bound the integrand on the right-hand side of (3.16), we get

$$\mathbb{P}\big(U \geq r\big) \leq \exp\left(\frac{c\lambda^2}{2} - \lambda r\right), \qquad \forall \lambda > 0. \tag{3.18}$$

Optimizing over $\lambda > 0$ to get the tightest bound gives $\lambda = \frac{r}{c}$, and its substitution in (3.18) gives the bound in (3.17). □

3.1.3 Tensorization of the (relative) entropy

The relative entropy $D(P^{(tf)} \| P)$ involves two probability measures on the Cartesian product space $\mathcal{X}^n$, so bounding this quantity directly is generally very difficult. This is where the third ingredient of the entropy method, the so-called *tensorization step*, comes in. The name "tensorization" reflects the fact that this step involves bounding $D(P^{(tf)} \| P)$ by a sum of "one-dimensional" relative entropy terms, each involving the conditional distributions of one of the variables given the rest. The tensorization step hinges on the following simple bound:

Proposition 5. Let P and Q be two probability measures on the product space $\mathcal{X}^n$, where P is a product measure. For any $i \in \{1, \ldots, n\}$, let $\bar{X}^i$ denote the $(n-1)$-tuple $(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ obtained by removing X_i from X^n. Then

$$D(Q \| P) \leq \sum_{i=1}^{n} D\big(Q_{X_i|\bar{X}^i} \big\| P_{X_i} \big| Q_{\bar{X}^i}\big). \tag{3.19}$$

Proof. From the relative entropy chain rule

$$\begin{aligned} D(Q \| P) &= \sum_{i=1}^{n} D\big(Q_{X_i \,|\, X^{i-1}} \,||\, P_{X_i|X^{i-1}} \,|\, Q_{X^{i-1}}\big) \\ &= \sum_{i=1}^{n} D\big(Q_{X_i \,|\, X^{i-1}} \,||\, P_{X_i} \,|\, Q_{X^{i-1}}\big) \end{aligned} \tag{3.20}$$

where the last equality holds since $X_1, \ldots, X_n$ are independent random variables under P (which implies that $P_{X_i|X^{i-1}} = P_{X_i|\bar{X}^i} = P_{X_i}$). Furthermore, for every $i \in \{1, \ldots, n\}$,

$$\begin{aligned} & D\big(Q_{X_i|\bar{X}^i} \big\| P_{X_i} \big| Q_{\bar{X}^i}\big) - D\big(Q_{X_i|X^{i-1}} \big\| P_{X_i} \big| Q_{X^{i-1}}\big) \\ &= \mathbb{E}_Q\left[\ln \frac{\mathrm{d}Q_{X_i|\bar{X}^i}}{\mathrm{d}P_{X_i}}\right] - \mathbb{E}_Q\left[\ln \frac{\mathrm{d}Q_{X_i|X^{i-1}}}{\mathrm{d}P_{X_i}}\right] \\ &= \mathbb{E}_Q\left[\ln \frac{\mathrm{d}Q_{X_i|\bar{X}^i}}{\mathrm{d}Q_{X_i|X^{i-1}}}\right] \\ &= D\big(Q_{X_i|\bar{X}^i} \big\| Q_{X_i|X^{i-1}} \big| Q_{\bar{X}^i}\big) \geq 0. \end{aligned} \tag{3.21}$$

Hence, by combining (3.20) and (3.21), we get the inequality in (3.19). □

Remark 19. The quantity on the right-hand side of (3.19) is actually the so-called *erasure divergence* $D^-(Q\|P)$ between Q and P (see [109, Definition 4]), which in the case of arbitrary Q and P is defined by

$$D^-(Q\|P) \triangleq \sum_{i=1}^n D(Q_{X_i|\bar{X}^i} \| P_{X_i|\bar{X}^i} | Q_{\bar{X}^i}). \tag{3.22}$$

Because in the inequality (3.19) P is assumed to be a product measure, we can replace $P_{X_i|\bar{X}^i}$ by P_{X_i}. For a general (non-product) measure P, the erasure divergence $D^-(Q\|P)$ may be strictly larger or smaller than the ordinary divergence $D(Q\|P)$. For example, if $n = 2$, $P_{X_1} = Q_{X_1}$, $P_{X_2} = Q_{X_2}$, then

$$\frac{\mathrm{d}Q_{X_1|X_2}}{\mathrm{d}P_{X_1|X_2}} = \frac{\mathrm{d}Q_{X_2|X_1}}{\mathrm{d}P_{X_2|X_1}} = \frac{\mathrm{d}Q_{X_1,X_2}}{\mathrm{d}P_{X_1,X_2}},$$

so, from (3.22),

$$D^-(Q_{X_1,X_2}\|P_{X_1,X_2}) = D(Q_{X_1|X_2}\|P_{X_1|X_2}|Q_{X_2}) + D(Q_{X_2|X_1}\|P_{X_2|X_1}|Q_{X_1}) = 2D(Q_{X_1,X_2}\|P_{X_1,X_2}).$$

On the other hand, if $X_1 = X_2$ under both P and Q, then $D^-(Q\|P) = 0$, but $D(Q\|P) > 0$ whenever $P \neq Q$, so $D(Q\|P) > D^-(Q\|P)$ in this case.

Applying Proposition 5 with $Q = P^{(tf)}$ to bound the divergence in the integrand in (3.15), we obtain from Corollary 7 the following:

Proposition 6. For any $r \geq 0$, we have

$$\mathbb{P}\big(U \geq r\big) \leq \exp\left(\lambda \sum_{i=1}^n \int_0^\lambda \frac{D\big(P^{(tf)}_{X_i|\bar{X}^i} \| P_{X_i} | P^{(tf)}_{\bar{X}^i}\big)}{t^2} \mathrm{d}t - \lambda r\right), \qquad \forall \lambda > 0 \tag{3.23}$$

The conditional divergences in the integrand in (3.23) may look formidable, but the remarkable thing is that, for each i and a given $\bar{X}^i = \bar{x}^i$, the corresponding term involves a tilting of the marginal distribution P_{X_i}. Indeed, let us fix some $i \in \{1, \ldots, n\}$, and for each choice of $\bar{x}^i \in \mathcal{X}^{n-1}$ let us define a function $f_i(\cdot|\bar{x}^i) : \mathcal{X} \to \mathbb{R}$ by setting

$$f_i(y|\bar{x}^i) \triangleq f(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n), \qquad \forall y \in \mathcal{X}. \tag{3.24}$$

Then

$$\frac{\mathrm{d}P^{(f)}_{X_i|\bar{X}^i=\bar{x}^i}}{\mathrm{d}P_{X_i}} = \frac{\exp\big(f_i(\cdot|\bar{x}^i)\big)}{\mathbb{E}\left[\exp\big(f_i(X_i|\bar{x}^i)\big)\right]}. \tag{3.25}$$

In other words, $P^{(f)}_{X_i|\bar{X}^i=\bar{x}^i}$ is the $f_i(\cdot|\bar{x}^i)$-tilting of P_{X_i}. This is the essence of tensorization: we have effectively decomposed the n-dimensional problem of bounding $D(P^{(tf)}\|P)$ into n one-dimensional problems, where the ith problem involves the tilting of the marginal distribution P_{X_i} by functions of the form $f_i(\cdot|\bar{x}^i), \forall \bar{x}^i$. In particular, we get the following:

Corollary 8. Suppose that the function f and the probability distribution P of X^n are such that there exist some constants $c_1, \ldots, c_n > 0$, so that, for any $t > 0$,

$$D\Big(P^{(tf_i(\cdot|\bar{x}^i))}_{X_i} \| P_{X_i}\Big) \leq \frac{c_i t^2}{2}, \qquad \forall i \in \{1, \ldots, n\},\ \bar{x}^i \in \mathcal{X}^{n-1}. \tag{3.26}$$

Then

$$\mathbb{P}\Big(f(X^n) - \mathbb{E}f(X^n) \geq r\Big) \leq \exp\left(-\frac{r^2}{2\sum_{i=1}^n c_i}\right), \quad \forall r > 0. \tag{3.27}$$

Proof. For any $t > 0$

$$
\begin{aligned}
& D(P^{(tf)} || P) \\
& \leq \sum_{i=1}^n D\Big(P^{(tf)}_{X_i|\bar{X}^i} \big\| P_{X_i} \, | \, P^{(tf)}_{\bar{X}^i}\Big) && (3.28) \\
& = \sum_{i=1}^n \int_{\mathcal{X}^{n-1}} D\Big(P^{(tf)}_{X_i|\bar{X}^i=\bar{x}^i} \big\| P_{X_i}\Big) \, P^{(tf)}_{\bar{X}^i}(\mathrm{d}\bar{x}^i) && (3.29) \\
& = \sum_{i=1}^n \int_{\mathcal{X}^{n-1}} D\Big(P^{(tf_i(\cdot|\bar{x}^i))}_{X_i} \big\| P_{X_i}\Big) \, P^{(tf)}_{\bar{X}^i}(\mathrm{d}\bar{x}^i) && (3.30) \\
& \leq \sum_{i=1}^n \int_{\mathcal{X}^{n-1}} \frac{c_i t^2}{2} \, P^{(tf)}_{\bar{X}^i}(\mathrm{d}\bar{x}^i) && (3.31) \\
& = \frac{t^2}{2} \cdot \sum_{i=1}^n c_i && (3.32)
\end{aligned}
$$

where (3.28) follows from the tensorization of the relative entropy, (3.29) holds since P is a product measure (so $P_{X_i} = P_{X_i|\bar{X}^i}$) and by the definition of the conditional relative entropy, (3.30) follows from (3.24) and (3.25) which implies that $P^{(tf)}_{X_i|\bar{X}^i=\bar{x}^i} = P^{(tf_i(\cdot|\bar{x}^i))}_{X_i}$, and inequality (3.31) holds by the assumption in (3.26). Finally, the inequality in (3.27) follows from (3.32) and Corollary 7. □

3.1.4 Preview: logarithmic Sobolev inequalities

Ultimately, the success of the entropy method hinges on demonstrating that the bounds in (3.26) hold for the function $f : \mathcal{X}^n \to \mathbb{R}$ and the probability distribution $P = P_{X^n}$ of interest. In the next two sections, we will show how to derive such bounds using the so-called *logarithmic Sobolev inequalities*. Here, we will give a quick preview of this technique.

Let μ be a probability measure on $\mathcal{X}$, and let $\mathcal{A}$ be a family of real-valued functions $g : \mathcal{X} \to \mathbb{R}$, such that for any $a \geq 0$ and $g \in \mathcal{A}$, also $ag \in \mathcal{A}$. Let $E : \mathcal{A} \to \mathbb{R}^+$ be a non-negative functional that is homogeneous of degree 2, i.e., for any $a \geq 0$ and $g \in \mathcal{A}$, we have $E(ag) = a^2 E(g)$. Suppose further that there exists a constant $c > 0$, such that the inequality

$$
D(\mu^{(g)} \| \mu) \leq \frac{cE(g)}{2} \tag{3.33}
$$

holds for any $g \in \mathcal{A}$. Now, suppose that, for each $i \in \{1, \ldots, n\}$, inequality (3.33) holds with $\mu = P_{X_i}$ holds and some constant $c_i > 0$ where $\mathcal{A}$ is a suitable family of functions f such that, for any $\bar{x}^i \in \mathcal{X}^{n-1}$ and $i \in \{1, \ldots, n\}$,

1. $f_i(\cdot|\bar{x}^i) \in \mathcal{A}$
2. $E\big(f_i(\cdot|\bar{x}^i)\big) \leq 1$

where f_i is defined in (3.24). Then, the bounds in (3.26) hold since from (3.33) and the above properties of the functional E, it follows that for every $t > 0$ and $\bar{x}^i \in \mathcal{X}^{n-1}$

$$
\begin{aligned}
& D\Big(P^{(tf)}_{X_i|\bar{X}^i=\bar{x}^i} \big\| P_{X_i}\Big) \\
& \leq \frac{c_i \, E\big(t \, f_i(\cdot|\bar{x}^i)\big)}{2} \\
& = \frac{c_i t^2 \, E\big(f_i(\cdot|\bar{x}^i)\big)}{2} \\
& \leq \frac{c_i t^2}{2}, \quad \forall \, i \in \{1, \ldots, n\}.
\end{aligned}
$$

Consequently, the Gaussian concentration inequality in (3.27) follows from Corollary 8.

3.2 The Gaussian logarithmic Sobolev inequality (LSI)

Before turning to the general scheme of logarithmic Sobolev inequalities in the next section, we will illustrate the basic ideas in the particular case when $X_1, \ldots, X_n$ are i.i.d. standard Gaussian random variables. The relevant log-Sobolev inequality in this instance comes from a seminal paper of Gross [35], and it connects two key information-theoretic measures, namely the relative entropy and the relative Fisher information. In addition, there are deep links between Gross's log-Sobolev inequality and other fundamental information-theoretic inequalities, such as Stam's inequality and the entropy power inequality. Some of these fundamental links are considered in this section.

For any $n \in \mathbb{N}$ and any positive-semidefinite matrix $K \in \mathbb{R}^{n \times n}$, we will denote by G^n_K the Gaussian distribution with zero mean and covariance matrix K. When $K = sI_n$ for some $s \geq 0$ (where I_n denotes the $n \times n$ identity matrix), we will write G^n_s. We will also write G^n for G^n_1 when $n \geq 2$, and G for G^1_1. We will denote by γ^n_K, γ^n_s, γ_s, and γ the corresponding densities.

We first state Gross's inequality in its (more or less) original form:

Theorem 21. For $Z \sim G^n$ and for any smooth function $\phi : \mathbb{R}^n \to \mathbb{R}$, we have

$$\mathbb{E}[\phi^2(Z) \ln \phi^2(Z)] - \mathbb{E}[\phi^2(Z)] \ln \mathbb{E}[\phi^2(Z)] \leq 2\, \mathbb{E}\left[\|\nabla \phi(Z)\|^2\right]. \tag{3.34}$$

Remark 20. As shown by Carlen [110], equality in (3.34) holds if and only if ϕ is of the form $\phi(z) = \exp\langle a, z\rangle$ for some $a \in \mathbb{R}^n$, where $\langle \cdot, \cdot \rangle$ denotes the standard Euclidean inner product.

Remark 21. There is no loss of generality in assuming that $\mathbb{E}[\phi^2(Z)] = 1$. Then (3.34) can be rewritten as

$$\mathbb{E}[\phi^2(Z) \ln \phi^2(Z)] \leq 2\, \mathbb{E}\left[\|\nabla \phi(Z)\|^2\right], \qquad \text{if } \mathbb{E}[\phi^2(Z)] = 1,\ Z \sim G^n. \tag{3.35}$$

Moreover, a simple rescaling argument shows that, for $Z \sim G^n_s$ and an arbitrary smooth function ϕ with $\mathbb{E}[\phi^2(Z)] = 1$,

$$\mathbb{E}[\phi^2(Z) \ln \phi^2(Z)] \leq 2s\, \mathbb{E}\left[\|\nabla \phi(Z)\|^2\right]. \tag{3.36}$$

An information-theoretic proof of the Gaussian LSI (Theorem 21) is provided in the continuation to this section. The reader is also referred to [111] for another proof that is not information-theoretic.

From an information-theoretic point of view, the Gaussian LSI (3.34) relates two measures of (dis)similarity between probability measures — the relative entropy (or divergence) and the *relative Fisher information* (or *Fisher information distance*). The latter is defined as follows. Let P_1 and P_2 be two Borel probability measures on $\mathbb{R}^n$ with differentiable densities p_1 and p_2. Then the *relative Fisher information* (or *Fisher information distance*) between P_1 and P_2 is defined as (see [112, Eq. (6.4.12)])

$$I(P_1 \| P_2) \triangleq \int_{\mathbb{R}^n} \left\| \nabla \ln \frac{p_1(z)}{p_2(z)} \right\|^2 p_1(z) \mathrm{d}z = \mathbb{E}_{P_1}\left[\left\| \nabla \ln \frac{\mathrm{d}P_1}{\mathrm{d}P_2} \right\|^2 \right], \tag{3.37}$$

whenever the above integral converges. Under suitable regularity conditions, $I(P_1 \| P_2)$ admits the equivalent form (see [113, Eq. (1.108)])

$$I(P_1 \| P_2) = 4 \int_{\mathbb{R}^n} p_2(z) \left\| \nabla \sqrt{\frac{p_1(z)}{p_2(z)}} \right\|^2 \mathrm{d}z = 4\, \mathbb{E}_{P_2}\left[\left\| \nabla \sqrt{\frac{\mathrm{d}P_1}{\mathrm{d}P_2}} \right\|^2 \right]. \tag{3.38}$$

Remark 22. One condition under which (3.38) holds is as follows. Let $\xi : \mathbb{R}^n \to \mathbb{R}^n$ be the *distributional* (or *weak*) *gradient* of $\sqrt{\mathrm{d}P_1/\mathrm{d}P_2} = \sqrt{p_1/p_2}$, i.e., the equality

$$\int_{-\infty}^{\infty} \sqrt{\frac{p_1(z)}{p_2(z)}} \partial_i \psi(z) \mathrm{d}z = - \int_{-\infty}^{\infty} \xi_i(z) \psi(z) \mathrm{d}z$$

holds for all $i = 1, \dots, n$ and all test functions $\psi \in C_c^\infty(\mathbb{R}^n)$ [114, Sec. 6.6]. Then (3.38) holds, provided $\xi \in L^2(P_2)$.

Now let us fix a smooth function $\phi : \mathbb{R}^n \to \mathbb{R}$ satisfying the normalization condition $\int_{\mathbb{R}^n} \phi^2 \, \mathrm{d}G^n = 1$; we can assume w.l.o.g. that $\phi \geq 0$. Let Z be a standard n-dimensional Gaussian random variable, i.e., $P_Z = G^n$, and let $Y \in \mathbb{R}^n$ be a random vector with distribution P_Y satisfying

$$\frac{\mathrm{d}P_Y}{\mathrm{d}P_Z} = \frac{\mathrm{d}P_Y}{\mathrm{d}G^n} = \phi^2.$$

Then, on the one hand, we have

$$\mathbb{E}\left[\phi^2(Z) \ln \phi^2(Z)\right] = \mathbb{E}\left[\left(\frac{\mathrm{d}P_Y}{\mathrm{d}P_Z}(Z)\right) \ln \left(\frac{\mathrm{d}P_Y}{\mathrm{d}P_Z}(Z)\right)\right] = D(P_Y \| P_Z), \tag{3.39}$$

and on the other, from (3.38),

$$\mathbb{E}\left[\|\nabla\phi(Z)\|^2\right] = \mathbb{E}\left[\left\|\nabla \sqrt{\frac{\mathrm{d}P_Y}{\mathrm{d}P_Z}(Z)}\right\|^2\right] = \frac{1}{4} I(P_Y \| P_Z). \tag{3.40}$$

Substituting (3.39) and (3.40) into (3.35), we obtain the inequality

$$D(P_Y \| P_Z) \leq \frac{1}{2} I(P_Y \| P_Z), \qquad P_Z = G^n \tag{3.41}$$

which holds for any $P_Y \ll G^n$ with $\nabla \sqrt{\mathrm{d}P_Y/\mathrm{d}G^n} \in L^2(G^n)$. Conversely, for any $P_Y \ll G^n$ satisfying (3.41), we can derive (3.35) by letting $\phi = \sqrt{\mathrm{d}P_Y/\mathrm{d}G^n}$, provided $\nabla\phi$ exists (e.g., in the distributional sense). Similarly, for any $s > 0$, (3.36) can be written as

$$D(P_Y \| P_Z) \leq \frac{s}{2} I(P_Y \| P_Z), \qquad P_Z = G_s^n. \tag{3.42}$$

Now let us apply the Gaussian LSI (3.34) to functions of the form $\phi = \exp(g/2)$ for all suitably well-behaved $g : \mathbb{R}^n \to \mathbb{R}$. Doing this, we obtain

$$\mathbb{E}\left[\exp(g) \ln \frac{\exp(g)}{\mathbb{E}[\exp(g)]}\right] \leq \frac{1}{2} \mathbb{E}\left[\|\nabla g\|^2 \exp(g)\right], \tag{3.43}$$

where the expectation is w.r.t. G^n. If we let $P = G^n$, then we can recognize the left-hand side of (3.43) as $\mathbb{E}[\exp(g)] \cdot D(P^{(g)} \| P)$, where $P^{(g)}$ denotes, as usual, the g-tilting of P. Moreover, the right-hand side is equal to $\mathbb{E}[\exp(g)] \cdot \mathbb{E}_P^{(g)}[\|\nabla g\|^2]$ with $\mathbb{E}_P^{(g)}[\cdot]$ denoting expectation w.r.t. $P^{(g)}$. We therefore obtain the so-called *modified log-Sobolev inequality* for the standard Gaussian measure:

$$D(P^{(g)} \| P) \leq \frac{1}{2} \mathbb{E}_P^{(g)}\left[\|\nabla g\|^2\right], \qquad P = G^n \tag{3.44}$$

which holds for all smooth functions $g : \mathbb{R}^n \to \mathbb{R}$ that are exponentially integrable w.r.t. G^n. Observe that (3.44) implies (3.33) with $\mu = G^n$, $c = 1$, and $E(g) = \|\nabla g\|_\infty^2$.

In the remainder of this section, we first present a proof of Theorem 21, and then discuss several applications of the modified log-Sobolev inequality (3.44) to derivation of Gaussian concentration inequalities via the Herbst argument.

3.2.1 An information-theoretic proof of Gross's log-Sobolev inequality

In accordance with our general theme, we will prove Theorem 21 via tensorization: We first scale up to general n using suitable (sub)additivity properties, and then establish the $n = 1$ case. Indeed, suppose that (3.34) holds in dimension 1. For $n \geq 2$, let $X = (X_1, \ldots, X_n)$ be an n-tuple of i.i.d. $\mathcal{N}(0, 1)$ variables and consider a smooth function $\phi : \mathbb{R}^n \to \mathbb{R}$, such that $\mathbb{E}_P[\phi^2(X)] = 1$, where $P = P_X = G^n$ is the product of n copies of the standard Gaussian distribution G. If we define a probability measure $Q = Q_X$ with $\mathrm{d}Q_X/\mathrm{d}P_X = \phi^2$, then using Proposition 5 we can write

$$\begin{aligned}
\mathbb{E}_P\left[\phi^2(X)\ln\phi^2(X)\right] &= \mathbb{E}_P\left[\frac{\mathrm{d}Q}{\mathrm{d}P}\ln\frac{\mathrm{d}Q}{\mathrm{d}P}\right] \\
&= D(Q\|P) \\
&\leq \sum_{i=1}^n D\big(Q_{X_i|\bar{X}^i}\big\|P_{X_i}\big|Q_{\bar{X}^i}\big). \qquad (3.45)
\end{aligned}$$

Following the same steps as the ones that led to (3.24), we can define for each $i = 1, \ldots, n$ and each $\bar{x}^i = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \in \mathbb{R}^{n-1}$ the function $\phi_i(\cdot|\bar{x}^i) : \mathbb{R} \to \mathbb{R}$ via

$$\phi_i(y|\bar{x}^i) \triangleq \phi(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n), \qquad \forall \bar{x}^i \in \mathbb{R}^{n-1},\ y \in \mathbb{R}.$$

Then

$$\frac{\mathrm{d}Q_{X_i|\bar{X}^i=\bar{x}^i}}{\mathrm{d}P_{X_i}} = \frac{\phi_i^2(\cdot|\bar{x}^i)}{\mathbb{E}_P[\phi_i^2(X_i|\bar{x}^i)]}$$

for all $i \in \{1, \ldots, n\}, \bar{x}^i \in \mathbb{R}^{n-1}$. With this, we can write

$$\begin{aligned}
D\big(Q_{X_i|\bar{X}^i}\big\|P_{X_i}\big|Q_{\bar{X}^i}\big) &= \mathbb{E}_Q\left[\ln\frac{\mathrm{d}Q_{X_i|\bar{X}^i}}{\mathrm{d}P_{X_i}}\right] \\
&= \mathbb{E}_P\left[\frac{\mathrm{d}Q}{\mathrm{d}P}\ln\frac{\mathrm{d}Q_{X_i|\bar{X}^i}}{\mathrm{d}P_{X_i}}\right] \\
&= \mathbb{E}_P\left[\phi^2(X)\ln\frac{\phi_i^2(X_i|\bar{X}^i)}{\mathbb{E}_P[\phi_i^2(X_i|\bar{X}^i)|\bar{X}^i]}\right] \\
&= \mathbb{E}_P\left[\phi_i^2(X_i|\bar{X}^i)\ln\frac{\phi_i^2(X_i|\bar{X}^i)}{\mathbb{E}_P[\phi_i^2(X_i|\bar{X}^i)|\bar{X}^i]}\right] \\
&= \int_{\mathbb{R}^{n-1}}\mathbb{E}_P\left[\phi_i^2(X_i|\bar{x}^i)\ln\frac{\phi_i^2(X_i|\bar{x}^i)}{\mathbb{E}_P[\phi_i^2(X_i|\bar{x}^i)]}\right]P_{\bar{X}^i}(\mathrm{d}\bar{x}^i). \qquad (3.46)
\end{aligned}$$

Since each $X_i \sim G$, we can apply the Gaussian LSI (3.34) to the univariate functions $\phi_i(\cdot|\bar{x}^i)$ to get

$$\mathbb{E}_P\left[\phi_i^2(X_i|\bar{x}^i)\ln\frac{\phi_i^2(X_i|\bar{x}^i)}{\mathbb{E}_P[\phi_i^2(X_i|\bar{x}^i)]}\right] \leq 2\,\mathbb{E}_P\left[\big(\phi_i'(X_i|\bar{x}^i)\big)^2\right], \qquad \forall i = 1, \ldots, n; \bar{x}^i \in \mathbb{R}^{n-1} \qquad (3.47)$$

where

$$\phi_i'(y|\bar{x}^i) = \frac{\mathrm{d}\phi_i(y|\bar{x}^i)}{\mathrm{d}y} = \frac{\partial\phi(x)}{\partial x_i}\Big|_{x_i=y}.$$

Since $X_1, \ldots, X_n$ are i.i.d. under P, we can express (3.47) as

$$\mathbb{E}_P\left[\phi_i^2(X_i|\bar{x}^i)\ln\frac{\phi_i^2(X_i|\bar{x}^i)}{\mathbb{E}_P[\phi_i^2(X_i|\bar{x}^i)]}\right] \leq 2\,\mathbb{E}_P\left[\big(\partial_i\phi(X)\big)^2\Big|\bar{X}^i = \bar{x}^i\right].$$

Substituting this bound into (3.46), we have

$$D\big(Q_{X_i|\bar{X}^i} \| P_{X_i} | Q_{\bar{X}^i}\big) \le 2\,\mathbb{E}_P\left[\big(\partial_i \phi(X)\big)^2\right].$$

In turn, using this to bound each term in the summation on the right-hand side of (3.45) together with the fact that $\sum_{i=1}^n \big(\partial_i \phi(x)\big)^2 = \|\nabla \phi(x)\|^2$, we get

$$\mathbb{E}_P\left[\phi^2(X)\ln\phi^2(X)\right] \le 2\,\mathbb{E}_P\left[\big\|\nabla\phi^2(X)\big\|^2\right], \tag{3.48}$$

which is precisely the n-dimensional Gaussian LSI (3.35) for general $n \ge 2$ provided that it holds for $n = 1$.

Based on the above argument, we will now focus on proving the Gaussian LSI for $n = 1$. To that end, it will be convenient to express it in a different but equivalent form that relates the Fisher information and the entropy power of a real-valued random variable with a sufficiently regular density. In this form, the Gaussian LSI was first derived by Stam [36], and the equivalence between Stam's inequality and (3.34) was only noted much later by Carlen [110]. We will first establish this equivalence following Carlen's argument, and then give a new information-theoretic proof of Stam's inequality that, unlike existing proofs [115, 38], does not require de Bruijn's identity or the entropy-power inequality.

First, some definitions. Let Y be a real-valued random variable with density p_Y. The *differential entropy* of Y (in nats) is given by

$$h(Y) = h(p_Y) \triangleq -\int_{-\infty}^{\infty} p_Y(y)\ln p_Y(y)\mathrm{d}y, \tag{3.49}$$

provided the integral exists. If it does, then the *entropy power* of Y is given by

$$N(Y) \triangleq \frac{\exp(2h(Y))}{2\pi e}. \tag{3.50}$$

Moreover, if the density p_Y is differentiable, then the *Fisher information* (w.r.t. a location parameter) is given by

$$J(Y) = J(p_Y) = \int_{-\infty}^{\infty}\left(\frac{\mathrm{d}}{\mathrm{d}y}\ln p_Y(y)\right)^2 p_Y(y)\mathrm{d}y = \mathbb{E}[\rho_Y^2(Y)], \tag{3.51}$$

where $\rho_Y(y) \triangleq (\mathrm{d}/\mathrm{d}y)\ln p_Y(y)$ is known as the *score function.*

Remark 23. In theoretical statistics, an alternative definition of the Fisher information (w.r.t. a location parameter) in a real-valued random variable Y is (see [116, Definition 4.1])

$$J(Y) \triangleq \sup\left\{\big|\mathbb{E}\psi'(Y)\big|^2 : \psi \in C^1, \mathbb{E}\psi^2(Y) = 1\right\}. \tag{3.52}$$

Note that this definition does not involve derivatives of any functions of the density of Y (nor assumes that such a density even exists). It can be shown that the quantity defined in (3.52) exists and is finite if and only if Y has an absolutely continuous density p_Y, in which case $J(Y)$ is equal to (3.51) (see [116, Theorem 4.2]).

We will need the following facts:

1. If $D(P_Y \| G_s) < \infty$, then

$$D(P_Y\|G_s) = \frac{1}{2}\ln\frac{1}{N(Y)} + \frac{1}{2}\ln s - \frac{1}{2} + \frac{1}{2s}\mathbb{E}Y^2. \tag{3.53}$$

This is proved by direct calculation: Since $D(P_Y \| G_s) < \infty$, we have $P_Y \ll G_s$ and $\mathrm{d}P_Y/\mathrm{d}G_s = p_Y/\gamma_s$. Then

$$\begin{aligned} D(P_Y \| G_s) &= \int_{-\infty}^{\infty} p_Y(y) \ln \frac{p_Y(y)}{\gamma_s(y)} \mathrm{d}y \\ &= -h(Y) + \frac{1}{2}\ln(2\pi s) + \frac{1}{2s}\mathbb{E}Y^2 \\ &= -\frac{1}{2}\left(2h(Y) - \ln(2\pi e)\right) + \frac{1}{2}\ln s - \frac{1}{2} + \frac{1}{2s}\mathbb{E}Y^2 \\ &= \frac{1}{2}\ln\frac{1}{N(Y)} + \frac{1}{2}\ln s - \frac{1}{2} + \frac{1}{2s}\mathbb{E}Y^2, \end{aligned}$$

which is (3.53).

2. If $J(Y) < \infty$ and $\mathbb{E}Y^2 < \infty$, then for any $s > 0$

$$I(P_Y \| G_s) = J(Y) + \frac{1}{s^2}\mathbb{E}Y^2 - \frac{2}{s} < \infty, \tag{3.54}$$

where $I(\cdot\|\cdot)$ is the relative Fisher information, cf. (3.37). Indeed:

$$\begin{aligned} I(P_Y \| G_s) &= \int_{-\infty}^{\infty} p_Y(y) \left(\frac{\mathrm{d}}{\mathrm{d}y}\ln p_Y(y) - \frac{\mathrm{d}}{\mathrm{d}y}\ln\gamma_s(y)\right)^2 \mathrm{d}y \\ &= \int_{-\infty}^{\infty} p_Y(y)\left(\rho_Y(y) + \frac{y}{s}\right)^2 \mathrm{d}y \\ &= \mathbb{E}[\rho_Y^2(Y)] + \frac{2}{s}\mathbb{E}[Y\rho_Y(Y)] + \frac{1}{s^2}\mathbb{E}Y^2 \\ &= J(Y) + \frac{2}{s}\mathbb{E}[Y\rho_Y(Y)] + \frac{1}{s^2}\mathbb{E}Y^2. \end{aligned}$$

Because $J(Y) < \infty$ and $\mathbb{E}Y^2 < \infty$, then $\mathbb{E}[Y\rho_Y(Y)] = -1$ (see [117, Lemma A1]), and we get (3.54).

We are now in a position to prove the following:

Proposition 7 (Carlen [110])**.** *Let Y be a real-valued random variable with a smooth density p_Y, such that $J(Y) < \infty$ and $\mathbb{E}Y^2 < \infty$. Then the following statements are equivalent:*

1. *Gaussian log-Sobolev inequality, $D(P_Y \| G) \le (1/2) I(P_Y \| G)$.*
2. *Stam's inequality, $N(Y)J(Y) \ge 1$.*

Remark 24. Carlen's original derivation in [110] requires p_Y to be in the Schwartz space $\mathcal{S}(\mathbb{R})$ of infinitely differentiable functions, all of whose derivatives vanish sufficiently rapidly at infinity. In comparison, the regularity conditions of the above proposition are much weaker, requiring only that P_Y has a differentiable and absolutely continuous density, as well as a finite second moment.

Proof. We first show the implication 1) ⇒ 2). If 1) holds, then

$$D(P_Y \| G_s) \le \frac{s}{2} I(P_Y \| G_s), \qquad \forall s > 0. \tag{3.55}$$

Since $J(Y)$ and $\mathbb{E}Y^2$ are finite by assumption, the right-hand side of (3.55) is finite and equal to (3.54). Therefore, $D(P_Y \| G_s)$ is also finite, and it is equal to (3.53). Hence, we can rewrite (3.55) as

$$\frac{1}{2}\ln\frac{1}{N(Y)} + \frac{1}{2}\ln s - \frac{1}{2} + \frac{1}{2s}\mathbb{E}Y^2 \le \frac{s}{2}J(Y) + \frac{1}{2s}\mathbb{E}Y^2 - 1.$$

Because $\mathbb{E}Y^2 < \infty$, we can cancel the corresponding term from both sides and, upon rearranging, obtain

$$\ln \frac{1}{N(Y)} \le sJ(Y) - \ln s - 1.$$

Importantly, this bound holds for *every* $s > 0$. Therefore, using the fact that, for any $a > 0$,

$$1 + \ln a = \inf_{s>0} (as - \ln s),$$

we obtain Stam's inequality $N(Y)J(Y) \ge 1$.

To establish the converse implication 2) ⇒ 1), we simply run the above proof backwards. □

We now turn to the proof of Stam's inequality. Without loss of generality, we may assume that $\mathbb{E}Y = 0$ and $\mathbb{E}Y^2 = 1$. Our proof will exploit the formula, due to Verdú [118], that expresses the divergence in terms of an integral of the excess mean squared error (MSE) in a certain estimation problem with additive Gaussian noise. Specifically, consider the problem of estimating a real-valued random variable Y on the basis of a noisy observation $\sqrt{s}Y + Z$, where $s > 0$ is the signal-to-noise ratio (SNR) and the additive standard Gaussian noise $Z \sim G$ is independent of Y. If Y has distribution P, then the minimum MSE (MMSE) at SNR s is defined as

$$\mathsf{mmse}(Y, s) \triangleq \inf_{\varphi} \mathbb{E}[(Y - \varphi(\sqrt{s}Y + Z))^2], \tag{3.56}$$

where the infimum is over all measurable functions (estimators) $\varphi : \mathbb{R} \to \mathbb{R}$. It is well-known that the infimum in (3.56) is achieved by the conditional expectation $u \mapsto \mathbb{E}[Y|\sqrt{s}Y + Z = u]$, so

$$\mathsf{mmse}(Y, s) = \mathbb{E}\left[\left(Y - \mathbb{E}[Y|\sqrt{s}Y + Z]\right)^2\right].$$

On the other hand, suppose we instead assume that Y has distribution Q and therefore use the *mismatched estimator* $u \mapsto \mathbb{E}_Q[Y|\sqrt{s}Y+Z]$, where the conditional expectation is now computed assuming that $Y \sim Q$. Then the resulting *mismatched* MSE is given by

$$\mathsf{mse}_Q(Y, s) = \mathbb{E}\left[\left(Y - \mathbb{E}_Q[Y|\sqrt{s}Y + Z]\right)^2\right],$$

where the expectation on the outside is computed using the correct distribution P of Y. Then the following relation holds for the divergence between P and Q (see [118, Theorem 1]):

$$D(P\|Q) = \frac{1}{2}\int_0^\infty [\mathsf{mse}_Q(Y, s) - \mathsf{mmse}(Y, s)]\, \mathrm{d}s. \tag{3.57}$$

We will apply the formula (3.57) to $P = P_Y$ and $Q = G$, where P_Y satisfies $\mathbb{E}Y = 0$ and $\mathbb{E}Y^2 = 1$. Then it can be shown that, for any γ,

$$\mathsf{mse}_Q(Y, s) = \mathsf{mse}_G(Y, s) = \mathsf{lmmse}(Y, s),$$

where $\mathsf{lmmse}(Y, s)$ is the *linear* MMSE, i.e., the MMSE attainable by any *affine* estimator $u \mapsto au + b$, $a, b \in \mathbb{R}$:

$$\mathsf{lmmse}(Y, s) = \inf_{a,b\in\mathbb{R}} \mathbb{E}\left[\left(Y - a(\sqrt{s}Y + Z) - b\right)^2\right]. \tag{3.58}$$

The infimum in (3.58) is achieved by $a^* = \sqrt{s}/(1 + s)$ and $b = 0$, giving

$$\mathsf{lmmse}(Y, s) = \frac{1}{1 + s}. \tag{3.59}$$

Moreover, $\mathsf{mmse}(\gamma)$ can be bounded from below using the so-called *van Trees inequality* [119] (cf. also Appendix 3.A):

$$\mathsf{mmse}(Y, s) \geq \frac{1}{J(Y) + s}. \tag{3.60}$$

Then

$$\begin{aligned} D(P_Y \| G) &= \frac{1}{2} \int_0^\infty \big(\mathsf{lmmse}(Y, s) - \mathsf{mmse}(Y, s)\big)\, \mathrm{d}s \\ &\leq \frac{1}{2} \int_0^\infty \left(\frac{1}{1+s} - \frac{1}{J(Y)+s} \right) \mathrm{d}s \\ &= \frac{1}{2} \lim_{\lambda \to \infty} \int_0^\lambda \left(\frac{1}{1+s} - \frac{1}{J(Y)+s} \right) \mathrm{d}s \\ &= \frac{1}{2} \lim_{\lambda \to \infty} \ln \left(\frac{J(Y)\,(1+\lambda)}{J(Y)+\lambda} \right) \\ &= \frac{1}{2} \ln J(Y), \end{aligned} \tag{3.61}$$

where the second step uses (3.59) and (3.60). On the other hand, using (3.53) with $s = \mathbb{E}Y^2 = 1$, we get $D(P_Y \| G) = \frac{1}{2} \ln(1/N(Y))$. Combining this with (3.61), we recover Stam's inequality $N(Y)J(Y) \geq 1$. Moreover, the van Trees bound (3.60) is achieved with equality if and only if Y is a standard Gaussian random variable.

3.2.2 From Gaussian log-Sobolev inequality to Gaussian concentration inequalities

We are now ready to apply the log-Sobolev machinery to establish Gaussian concentration for random variables of the form $U = f(X^n)$, where $X_1, \ldots, X_n$ are i.i.d. standard normal random variables and $f : \mathbb{R}^n \to \mathbb{R}$ is any Lipschitz function. We start by considering the special case when f is also differentiable.

Proposition 8. Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(0, 1)$ random variables. Then, for every differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ such that $\|\nabla f(X^n)\| \leq 1$ almost surely, we have

$$\mathbb{P}\Big(f(X^n) \geq \mathbb{E} f(X^n) + r \Big) \leq \exp\left(-\frac{r^2}{2} \right), \quad \forall\, r \geq 0 \tag{3.62}$$

Proof. Let $P = G^n$ denote the distribution of X^n. If Q is any probability measure such that $Q \ll P$ and $P \ll Q$ (i.e., P and Q are mutually absolutely continuous), then any event that has P-probability 1 will also have Q-probability 1 and vice versa. Since the function f is differentiable, it is everywhere finite, so $P^{(f)}$ and P are mutually absolutely continuous. Hence, any event that occurs P-a.s. also occurs $P^{(tf)}$-a.s. for all $t \in \mathbb{R}$. In particular, $\|\nabla f(X^n)\| \leq 1$ $P^{(tf)}$-a.s. for all t. Therefore, applying the modified log-Sobolev inequality (3.44) to $g = tf$ for some $t > 0$, we get

$$D(P^{(tf)} \| P) \leq \frac{t^2}{2} \mathbb{E}_P^{(tf)} \left[\|\nabla f(X^n)\|^2 \right] \leq \frac{t^2}{2}. \tag{3.63}$$

Using Corollary 7 with $U = f(X^n) - \mathbb{E} f(X^n)$, we get (3.62). □

Remark 25. Corollary 7 and inequality (3.44) with $g = tf$ imply that, for any smooth f with $\|\nabla f(X^n)\|^2 \leq L$ a.s.,

$$\mathbb{P}\Big(f(X^n) \geq \mathbb{E} f(X^n) + r \Big) \leq \exp\left(-\frac{r^2}{2L} \right), \quad \forall\, r \geq 0. \tag{3.64}$$

Thus, the constant κ in the corresponding Gaussian concentration bound (3.2) is controlled by the sensitivity of f to modifications of its coordinates.

Having established concentration for smooth f, we can now proceed to the general case:

Theorem 22. Let X^n be as before, and let $f : \mathbb{R}^n \to \mathbb{R}$ be a Lipschitz function with Lipschitz constant 1, i.e.,

$$|f(x^n) - f(y^n)| \le \|x^n - y^n\|, \qquad \forall x^n, y^n \in \mathbb{R}^n.$$

Then

$$\mathbb{P}\Big(f(X^n) \ge \mathbb{E}f(X^n) + r\Big) \le \exp\left(-\frac{r^2}{2}\right), \quad \forall\, r \ge 0. \tag{3.65}$$

Proof. The trick is to slightly perturb f to get a *differentiable* function with the norm of its gradient bounded by the Lipschitz constant of f. Then we can apply Proposition 8, and consider the limit of vanishing perturbation.

We construct the perturbation as follows. Let $Z_1, \dots, Z_n$ be n i.i.d. $\mathcal{N}(0,1)$ random variables, independent of X^n. For any $\delta > 0$, define the function

$$\begin{aligned} f_\delta(x^n) \triangleq \mathbb{E}\left[f\left(x^n + \sqrt{\delta}Z^n\right)\right] &= \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} f(x^n + \sqrt{\delta}z^n) \exp\left(-\frac{\|z^n\|^2}{2}\right) \mathrm{d}z^n \\ &= \frac{1}{(2\pi\delta)^{n/2}} \int_{\mathbb{R}^n} f(z^n) \exp\left(-\frac{\|z^n - x^n\|^2}{2\delta}\right) \mathrm{d}z^n. \end{aligned}$$

It is easy to see that f_δ is differentiable (in fact, it is in C^∞; this is known as the smoothing property of the Gaussian convolution kernel). Moreover, using Jensen's inequality and the fact that f is 1-Lipschitz, we have

$$\begin{aligned} |f_\delta(x^n) - f(x^n)| &= \Big|\mathbb{E}[f(x^n + \sqrt{\delta}Z^n)] - f(x^n)\Big| \\ &\le \mathbb{E}\Big|f(x^n + \sqrt{\delta}Z^n) - f(x^n)\Big| \\ &\le \sqrt{\delta}\mathbb{E}\|Z^n\|. \end{aligned}$$

Therefore, $\lim_{\delta\to 0} f_\delta(x^n) = f(x^n)$ for every $x^n \in \mathbb{R}^n$. Moreover, because f is 1-Lipschitz, it is differentiable almost everywhere by Rademacher's theorem [120, Section 3.1.2], and $\|\nabla f\| \le 1$ almost everywhere. Consequently, since $\nabla f_\delta(x^n) = \mathbb{E}\big[\nabla f\big(x^n + \sqrt{\delta}Z^n\big)\big]$, Jensen's inequality gives $\|\nabla f_\delta(x^n)\| \le \mathbb{E}\big\|\nabla f\big(x^n + \sqrt{\delta}Z^n\big)\big\| \le 1$ for *every* $x^n \in \mathbb{R}^n$. Therefore, we can apply Proposition 8 to get, for all $\delta > 0$ and $r > 0$,

$$\mathbb{P}\Big(f_\delta(X^n) \ge \mathbb{E}f_\delta(X^n) + r\Big) \le \exp\left(-\frac{r^2}{2}\right).$$

Using the fact that $f_\delta(x^n)$ converges to $f(x^n)$ everywhere as $\delta \to 0$, we obtain (3.65):

$$\begin{aligned} \mathbb{P}\Big(f(X^n) \ge \mathbb{E}f(X^n) + r\Big) &= \mathbb{E}\big[1_{\{f(X^n)\ge \mathbb{E}f(X^n)+r\}}\big] \\ &\le \lim_{\delta\to 0} \mathbb{E}\big[1_{\{f_\delta(X^n)\ge \mathbb{E}f_\delta(X^n)+r\}}\big] \\ &= \lim_{\delta\to 0} \mathbb{P}\Big(f_\delta(X^n) \ge \mathbb{E}f_\delta(X^n) + r\Big) \\ &\le \exp\left(-\frac{r^2}{2}\right) \end{aligned}$$

where the first inequality is by Fatou's lemma. □

3.2.3 Hypercontractivity, Gaussian log-Sobolev inequality, and Rényi divergence

We close our treatment of the Gaussian log-Sobolev inequality with a striking result, proved by Gross in his original paper [35], that this inequality is equivalent to a very strong contraction property (dubbed *hypercontractivity*) of a certain class of stochastic transformations. The original motivation behind the work of Gross [35] came from problems in quantum field theory. However, we will take an information-theoretic point of view and relate it to data processing inequalities for a certain class of channels with additive Gaussian noise, as well as to the rate of convergence in the second law of thermodynamics for Markov processes [121].

Consider a pair (X, Y) of real-valued random variables that are related through the stochastic transformation

$$Y = e^{-t}X + \sqrt{1 - e^{-2t}}\,Z \tag{3.66}$$

for some $t \ge 0$, where the additive noise $Z \sim G$ is independent of X. For reasons that will become clear shortly, we will refer to the channel that implements the transformation (3.66) for a given $t \ge 0$ as the *Ornstein–Uhlenbeck channel with noise parameter* t and denote it by $\mathrm{OU}(t)$. Similarly, we will refer to the collection of channels $\{\mathrm{OU}(t)\}_{t=0}^{\infty}$ indexed by all $t \ge 0$ as the *Ornstein–Uhlenbeck channel family.* We immediately note the following properties:

1. $\mathrm{OU}(0)$ is the ideal channel, $Y = X$.

2. If $X \sim G$, then $Y \sim G$ as well, for any t.

3. Using the terminology of [12, Chapter 4], the channel family $\{\mathrm{OU}(t)\}_{t=0}^{\infty}$ is *ordered by degradation*: for any $t_1, t_2 \ge 0$ we have

$$\mathrm{OU}(t_1 + t_2) = \mathrm{OU}(t_2) \circ \mathrm{OU}(t_1) = \mathrm{OU}(t_1) \circ \mathrm{OU}(t_2), \tag{3.67}$$

which is shorthand for the following statement: for any input random variable X, any standard Gaussian Z independent of X, and any $t_1, t_2 \ge 0$, we can always find independent standard Gaussian random variables Z_1, Z_2 that are also independent of X, such that

$$\begin{aligned} e^{-(t_1+t_2)}X + \sqrt{1 - e^{-2(t_1+t_2)}}\,Z &\overset{\mathrm{d}}{=} e^{-t_2}\left[e^{-t_1}X + \sqrt{1 - e^{-2t_1}}\,Z_1\right] + \sqrt{1 - e^{-2t_2}}\,Z_2 \\ &\overset{\mathrm{d}}{=} e^{-t_1}\left[e^{-t_2}X + \sqrt{1 - e^{-2t_2}}\,Z_1\right] + \sqrt{1 - e^{-2t_1}}\,Z_2 \end{aligned} \tag{3.68}$$

where $\overset{\mathrm{d}}{=}$ denotes equality of distributions. In other words, we can always define real-valued random variables X, Y_1, Y_2, Z_1, Z_2 on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, such that $Z_1, Z_2 \sim G$, (X, Z_1, Z_2) are mutually independent,

$$\begin{aligned} Y_1 &\overset{\mathrm{d}}{=} e^{-t_1}X + \sqrt{1 - e^{-2t_1}}\,Z_1 \\ Y_2 &\overset{\mathrm{d}}{=} e^{-(t_1+t_2)}X + \sqrt{1 - e^{-2(t_1+t_2)}}\,Z_2 \end{aligned}$$

and $X \longrightarrow Y_1 \longrightarrow Y_2$ is a Markov chain. Even more generally, given any real-valued random variable X, we can construct a continuous-time Markov process $\{Y_t\}_{t=0}^{\infty}$ with $Y_0 \overset{\mathrm{d}}{=} X$ and $Y_t \overset{\mathrm{d}}{=} e^{-t}X + \sqrt{1 - e^{-2t}}\,\mathcal{N}(0,1)$ for all $t \ge 0$. One way to do this is to let $\{Y_t\}_{t=0}^{\infty}$ be governed by the Itô stochastic differential equation (SDE)

$$\mathrm{d}Y_t = -Y_t\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \qquad t \ge 0 \tag{3.69}$$

with the initial condition $Y_0 \overset{\mathrm{d}}{=} X$, where $\{B_t\}$ denotes the standard one-dimensional Wiener process (a.k.a. Brownian motion). The SDE (3.69) is known as the *Langevin equation* [122, p. 75], and the random process $\{Y_t\}$ that solves it is called the *Ornstein–Uhlenbeck process*; the solution of (3.69) is given by (see, e.g., [123, p. 358] or [124, p. 127]) is given by

$$Y_t = Xe^{-t} + \sqrt{2}\int_0^t e^{-(t-s)}\,\mathrm{d}B_s, \qquad t \geq 0$$

where by the Itô isometry the variance of the (zero-mean) additive Gaussian noise is indeed

$$\mathbb{E}\left[\left(\sqrt{2}\int_0^t e^{-(t-s)}\,\mathrm{d}B_s\right)^2\right] = 2\int_0^t e^{-2(t-s)}\mathrm{d}s = 2e^{-2t}\int_0^s e^{2s}\mathrm{d}s = 1 - e^{-2t}, \quad \forall\, t \geq 0.$$

This explains our choice of the name "Ornstein–Uhlenbeck channel" for the random transformation (3.66).

In order to state the main result to be proved in this section, we need the following definition: the *Rényi divergence* of order $\alpha \in \mathbb{R}^+\backslash\{0,1\}$ between two probability measures, P and Q, is defined as

$$D_\alpha(P\|Q) \triangleq \begin{cases} \frac{1}{\alpha-1}\ln \mathbb{E}_Q\left[\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)^\alpha\right], & \text{if } P \ll Q \\ +\infty, & \text{otherwise.} \end{cases} \tag{3.70}$$

We recall several key properties of the Rényi divergence (see, for example, [125]):

1. The ordinary divergence $D(P\|Q)$ is the limit of $D_\alpha(P\|Q)$ as $\alpha \downarrow 1$.
2. If we *define* $D_1(P\|Q)$ as $D(P\|Q)$, then the function $\alpha \mapsto D_\alpha(P\|Q)$ is nondecreasing.
3. For all $\alpha > 0$, $D_\alpha(\cdot\|\cdot)$ satisfies the *data processing inequality*: if we have two possible distributions P and Q for a random variable U, then for any channel (stochastic transformation) T that takes U as input we have

$$D_\alpha(\tilde{P}\|\tilde{Q}) \leq D_\alpha(P\|Q), \qquad \forall \alpha > 0 \tag{3.71}$$

where $\tilde{P}$ (resp., $\tilde{Q}$) is the distribution of the output of T when the input has distribution P (resp., Q).

Now consider the following set-up. Let X be a real-valued random variable with a sufficiently well-behaved distribution P (at the very least, we assume $P \ll G$). For any $t \geq 0$, let P_t denote the output distribution of the OU(t) channel with input $X \sim G$. Then, using the fact that the standard Gaussian distribution G is left invariant by the Ornstein–Uhlenbeck channel family together with the data processing inequality (3.71), we have

$$D_\alpha(P_t\|G) \leq D_\alpha(P\|G), \qquad \forall\, t \geq 0,\ \alpha > 0. \tag{3.72}$$

In other words, as we increase the noise parameter t, the output distribution P_t starts to resemble the invariant distribution G more and more, where the measure of resemblance is given by any of the Rényi divergences. This is, of course, nothing but the second law of thermodynamics for Markov chains (see, e.g., [81, Section 4.4] or [121]) applied to the continuous-time Markov process governed by the Langevin equation (3.69). We will now show, however, that the Gaussian log-Sobolev inequality of Gross (see Theorem 21) implies a stronger statement: For any $\alpha > 1$ and any $\varepsilon \in (0,1)$, there exists a positive constant $\tau = \tau(\alpha, \varepsilon)$, such that

$$D_\alpha(P_t\|G) \leq \varepsilon D_\alpha(P\|G), \qquad \forall t \geq \tau. \tag{3.73}$$

Here is the precise result:

Theorem 23 (Hypercontractive estimate for the Ornstein–Uhlenbeck channel)**.** The Gaussian log-Sobolev inequality of Theorem 21 is equivalent to the following statement: For any $1 < \beta < \alpha < \infty$

$$D_\alpha(P_t \| G) \le \left(\frac{\alpha(\beta-1)}{\beta(\alpha-1)}\right) D_\beta(P\|G), \qquad \forall\, t \ge \frac{1}{2}\ln\left(\frac{\alpha-1}{\beta-1}\right). \tag{3.74}$$

Remark 26. To see that Theorem 23 implies (3.73), fix $\alpha > 1$ and $\varepsilon \in (0,1)$. Let

$$\beta = \beta(\varepsilon, \alpha) \triangleq \frac{\alpha}{\alpha - \varepsilon(\alpha-1)}.$$

It is easy to verify that $1 < \beta < \alpha$ and that $\frac{\alpha(\beta-1)}{\beta(\alpha-1)} = \varepsilon$. Hence, Theorem 23 implies that

$$D_\alpha(P_t\|P) \le \varepsilon D_\beta(P\|G), \quad \forall\, t \ge \frac{1}{2}\ln\left(1 + \frac{\alpha(1-\varepsilon)}{\varepsilon}\right) \triangleq \tau(\alpha, \varepsilon).$$

Since the Rényi divergence $D_\alpha(\cdot\|\cdot)$ is monotonic non-decreasing in the parameter α, and $1 < \beta < \alpha$, then it follows that $D_\beta(P\|G) \le D_\alpha(P\|G)$. It therefore follows from the last inequality that

$$D_\alpha(P_t\|P) \le \varepsilon D_\alpha(P\|G), \quad \forall\, t \ge \tau(\alpha, \varepsilon).$$

We now turn to the proof of Theorem 23.

Proof. As a reminder, the L^p norm of a real-valued random variable U is defined by $\|U\|_p \triangleq (\mathbb{E}[|U|^p])^{1/p}$. It will be convenient to work with the following equivalent form of the Rényi divergence in (3.70): For any two random variables U and V such that $P_U \ll P_V$, we have

$$D_\alpha(P_U\|P_V) = \frac{\alpha}{\alpha-1}\ln\left\|\frac{\mathrm{d}P_U}{\mathrm{d}P_V}(V)\right\|_\alpha, \qquad \alpha > 1. \tag{3.75}$$

Let us denote by g the Radon–Nikodym derivative $\mathrm{d}P/\mathrm{d}G$. It is easy to show that $P_t \ll G$ for all t, so the Radon–Nikodym derivative $g_t \triangleq \mathrm{d}P_t/\mathrm{d}G$ exists. Moreover, $g_0 = g$. Also, let us define the function $\alpha : [0,\infty) \to [\beta, \infty)$ by $\alpha(t) = 1 + (\beta-1)e^{2t}$ for some $\beta > 1$. Let $Z \sim G$. Using (3.75), it is easy to verify that the desired bound (3.74) is equivalent to the statement that the function $F : [0,\infty) \to \mathbb{R}$, defined by

$$F(t) \triangleq \ln\left\|\frac{\mathrm{d}P_t}{\mathrm{d}G}(Z)\right\|_{\alpha(t)} \equiv \ln\|g_t(Z)\|_{\alpha(t)},$$

is non-increasing. From now on, we will adhere to the following notational convention: we will use either the dot or $\mathrm{d}/\mathrm{d}t$ to denote derivatives w.r.t. the "time" t, and the prime to denote derivatives w.r.t. the "space" variable z. We start by computing the derivative of F w.r.t. t, which gives

$$\begin{aligned}\dot{F}(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\left\{\frac{1}{\alpha(t)}\ln\mathbb{E}\left[\left(g_t(Z)\right)^{\alpha(t)}\right]\right\} \\ &= -\frac{\dot{\alpha}(t)}{\alpha^2(t)}\ln\mathbb{E}\left[\left(g_t(Z)\right)^{\alpha(t)}\right] + \frac{1}{\alpha(t)}\frac{\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\left(g_t(Z)\right)^{\alpha(t)}\right]}{\mathbb{E}\left[\left(g_t(Z)\right)^{\alpha(t)}\right]}.\end{aligned} \tag{3.76}$$

To handle the derivative w.r.t. t in the second term in (3.76), we need to delve a bit into the theory of the so-called *Ornstein–Uhlenbeck semigroup*, which is an alternative representation of the Ornstein–Uhlenbeck channel (3.66).

For any $t \ge 0$, let us define a linear operator K_t acting on any sufficiently regular (e.g., $L^1(G)$) function h as

$$K_t h(x) \triangleq \mathbb{E}\left[h\left(e^{-t}x + \sqrt{1-e^{-2t}}Z\right)\right], \tag{3.77}$$

where $Z \sim G$, as before. The family of operators $\{K_t\}_{t=0}^\infty$ has the following properties:

1. K_0 is the identity operator, $K_0 h = h$ for any h.

2. For any $t \geq 0$, if we consider the OU(t) channel, given by the random transformation (3.66), then for any measurable function F such that $\mathbb{E}[F(Y)] < \infty$ with Y in (3.66), we can write

$$K_t F(x) = \mathbb{E}[F(Y)|X = x], \qquad \forall x \in \mathbb{R} \tag{3.78}$$

and

$$\mathbb{E}[F(Y)] = \mathbb{E}[K_t F(X)]. \tag{3.79}$$

Here, (3.78) easily follows from (3.66), and (3.79) is immediate from (3.78).

3. A particularly useful special case of the above is as follows. Let X have distribution P with $P \ll G$, and let P_t denote the output distribution of the OU(t) channel. Then, as we have seen before, $P_t \ll G$, and the corresponding densities satisfy

$$g_t(x) = K_t g(x). \tag{3.80}$$

To prove (3.80), we can either use (3.78) and the fact that $g_t(x) = \mathbb{E}[g(Y)|X = x]$, or proceed directly from (3.66):

$$\begin{aligned} g_t(x) &= \frac{1}{\sqrt{2\pi(1-e^{-2t})}} \int_{\mathbb{R}} \exp\left(-\frac{(u - e^{-t}x)^2}{2(1-e^{-2t})}\right) g(u) \mathrm{d}u \\ &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} g\left(e^{-t}x + \sqrt{1-e^{-2t}}z\right) \exp\left(-\frac{z^2}{2}\right) \mathrm{d}z \\ &\equiv \mathbb{E}\left[g\left(e^{-t}x + \sqrt{1-e^{-t}}Z\right)\right] \end{aligned} \tag{3.81}$$

where in the second line we have made the change of variables $z = \frac{u - e^{-t}x}{\sqrt{1-e^{-2t}}}$, and in the third line $Z \sim G$.

4. The family of operators $\{K_t\}_{t=0}^{\infty}$ forms a semigroup, i.e., for any $t_1, t_2 \geq 0$ we have

$$K_{t_1+t_2} = K_{t_1} \circ K_{t_2} = K_{t_2} \circ K_{t_1},$$

which is shorthand for saying that $K_{t_1+t_2} h = K_{t_2}(K_{t_1} h) = K_{t_1}(K_{t_2} h)$ for any sufficiently regular h. This follows from (3.78) and (3.79) and from the fact that the channel family $\{\mathrm{OU}(t)\}_{t=0}^{\infty}$ is ordered by degradation. For this reason, $\{K_t\}_{t=0}^{\infty}$ is referred to as the *Ornstein–Uhlenbeck semigroup*. In particular, if $\{Y_t\}_{t=0}^{\infty}$ is the Ornstein–Uhlenbeck process, then for any sufficiently regular function $F : \mathbb{R} \to \mathbb{R}$ we have

$$K_t F(x) = \mathbb{E}[F(Y_t)|Y_0 = x], \qquad \forall x \in \mathbb{R}.$$

Two deeper results concerning the Ornstein–Uhlenbeck semigroup, which we will need, are as follows: Define the second-order differential operator $\mathcal{L}$ by

$$\mathcal{L}h(x) \triangleq h''(x) - xh'(x)$$

for all sufficiently smooth functions $h : \mathbb{R} \to \mathbb{R}$. Then:

1. The *Ornstein–Uhlenbeck flow* $\{h_t\}_{t=0}^{\infty}$, where $h_t = K_t h$ with sufficiently smooth initial condition $h_0 = h$, satisfies the partial differential equation (PDE)

$$\dot{h}_t = \mathcal{L}h_t. \tag{3.82}$$

2. For $Z \sim G$ and all sufficiently smooth functions $g, h : \mathbb{R} \to \mathbb{R}$ we have the *integration-by-parts formula*

$$\mathbb{E}[g(Z)\mathcal{L}h(Z)] = \mathbb{E}[h(Z)\mathcal{L}g(Z)] = -\mathbb{E}[g'(Z)h'(Z)]. \tag{3.83}$$

We provide the details in Appendix 3.B.

We are now ready to tackle the second term in (3.76). Noting that the family of densities $\{g_t\}_{t=0}^{\infty}$ forms an Ornstein–Uhlenbeck flow with initial condition $g_0 = g$, we have (assuming enough regularity conditions to permit interchanges of derivatives and expectations)

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\big(g_t(Z)\big)^{\alpha(t)}\right] &= \mathbb{E}\left[\big(g_t(Z)\big)^{\alpha(t)}\frac{\mathrm{d}}{\mathrm{d}t}\ln\big(g_t(Z)\big)^{\alpha(t)}\right]\\ &= \dot{\alpha}(t)\cdot\mathbb{E}\left[\big(g_t(Z)\big)^{\alpha(t)}\ln g_t(Z)\right] + \alpha(t)\,\mathbb{E}\left[\big(g_t(Z)\big)^{\alpha(t)-1}\frac{\mathrm{d}}{\mathrm{d}t}g_t(Z)\right]\\ &= \dot{\alpha}(t)\cdot\mathbb{E}\left[\big(g_t(Z)\big)^{\alpha(t)}\ln g_t(Z)\right] + \alpha(t)\,\mathbb{E}\left[\big(g_t(Z)\big)^{\alpha(t)-1}\mathcal{L}g_t(Z)\right] && (3.84)\\ &= \dot{\alpha}(t)\cdot\mathbb{E}\left[\big(g_t(Z)\big)^{\alpha(t)}\ln g_t(Z)\right] - \alpha(t)\,\mathbb{E}\left[\Big(\big(g_t(Z)\big)^{\alpha(t)-1}\Big)'\,\big(g_t(Z)\big)'\right] && (3.85)\\ &= \dot{\alpha}(t)\cdot\mathbb{E}\left[\big(g_t(Z)\big)^{\alpha(t)}\ln g_t(Z)\right] - \alpha(t)\big(\alpha(t)-1\big)\cdot\mathbb{E}\left[\big(g_t(Z)\big)^{\alpha(t)-2}\left|\big(g_t(Z)\big)'\right|^2\right] && (3.86)\end{aligned}$$

where we have used (3.82) to get (3.84), and (3.83) to get (3.85). If we define the function $\phi_t = g_t^{\alpha(t)/2}$, then we can rewrite (3.86) as

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\big(g_t(Z)\big)^{\alpha(t)}\right] = \frac{\dot{\alpha}(t)}{\alpha(t)}\mathbb{E}\left[\phi_t^2(Z)\ln\phi_t^2(Z)\right] - \frac{4\big(\alpha(t)-1\big)}{\alpha(t)}\mathbb{E}\left[\left|\phi_t'(Z)\right|^2\right]. \tag{3.87}$$

Using the definition of ϕ_t and a substitution of (3.87) into the right-hand side of (3.76) gives that

$$\alpha^2(t)\,\mathbb{E}[\phi_t^2(Z)]\,\dot{F}(t) = \dot{\alpha}(t)\cdot\big(\mathbb{E}[\phi_t^2(Z)\ln\phi_t^2(Z)] - \mathbb{E}[\phi_t^2(Z)]\ln\mathbb{E}[\phi_t^2(Z)]\big) - 4(\alpha(t)-1)\mathbb{E}\left[\left|\phi_t'(Z)\right|^2\right]. \tag{3.88}$$

If we now apply the Gaussian log-Sobolev inequality (3.34) to ϕ_t, then from (3.88) we get

$$\alpha^2(t)\,\mathbb{E}[\phi_t^2(Z)]\,\dot{F}(t) \le 2\left(\dot{\alpha}(t) - 2(\alpha(t)-1)\right)\mathbb{E}\left[\left|\phi_t'(Z)\right|^2\right]. \tag{3.89}$$

Since $\alpha(t) = 1 + (\beta-1)e^{2t}$, then $\dot{\alpha}(t) - 2(\alpha(t)-1) = 0$ and the right-hand side of (3.89) is equal to zero. Moreover, because $\alpha(t) > 0$ and $\phi_t^2(Z) > 0$ a.s. (note that $\phi_t^2 > 0$ if and only if $g_t > 0$, but the latter follows from (3.81) where g is a probability density function) then we conclude that $\dot{F}(t) \le 0$.

What we have proved so far is that, for any $\beta > 1$ and any $t \ge 0$,

$$D_{\alpha(t)}(P_t\|G) \le \left(\frac{\alpha(t)(\beta-1)}{\beta(\alpha(t)-1)}\right)D_\beta(P\|G) \tag{3.90}$$

where $\alpha(t) = 1 + (\beta-1)e^{2t}$. By the monotonicity property of the Rényi divergence, the left-hand side of (3.90) is greater than or equal to $D_\alpha(P_t\|G)$ as soon as $\alpha \le \alpha(t)$. By the same token, because the function $u \in (1,\infty) \mapsto u/(u-1)$ is strictly decreasing, the right-hand side of (3.90) can be upper-bounded by $\frac{\alpha(\beta-1)}{\beta(\alpha-1)}D_\beta(P\|G)$ for all $\alpha \ge \alpha(t)$. Putting all these facts together, we conclude that the Gaussian log-Sobolev inequality (3.34) implies (3.74).

We now show that (3.74) implies the log-Sobolev inequality of Theorem 21. To that end, we recall that (3.74) is equivalent to the right-hand side of (3.88) being less than or equal to zero for all $t \ge 0$ and all $\beta > 1$. Let us choose $t = 0$ and $\beta = 2$, in which case

$$\alpha(0) = \dot{\alpha}(0) = 2, \qquad \phi_0 = g.$$

Using this in (3.88) for $t = 0$, we get

$$2\left(\mathbb{E}\left[g^2(Z)\ln g^2(Z)\right] - \mathbb{E}[g^2(Z)]\ln \mathbb{E}[g^2(Z)]\right) - 4\,\mathbb{E}\left[\left|g'(Z)\right|^2\right] \le 0$$

which is precisely the log-Sobolev inequality (3.34). □

As a consequence, we can establish a strong version of the data processing inequality for the ordinary divergence:

Corollary 9. In the notation of Theorem 23, we have for any $t \ge 0$

$$D(P_t \| G) \le e^{-2t} D(P\|G). \tag{3.91}$$

Proof. Let $\alpha = 1 + \varepsilon e^{2t}$ and $\beta = 1 + \varepsilon$ for some $\varepsilon > 0$. Then using Theorem 23, we have

$$D_{1+\varepsilon e^{2t}}(P_t\|G) \le \left(\frac{e^{-2t}+\varepsilon}{1+\varepsilon}\right) D_{1+\varepsilon}(P\|G), \qquad \forall t \ge 0 \tag{3.92}$$

Taking the limit of both sides of (3.92) as $\varepsilon \downarrow 0$ and using $D(P\|Q) = \lim_{\alpha\downarrow 1} D_\alpha(P\|Q)$, we get (3.91). □

3.3 Logarithmic Sobolev inequalities: the general scheme

Now that we have seen the basic idea behind log-Sobolev inequalities in the concrete case of i.i.d. Gaussian random variables, we are ready to take a more general viewpoint. To that end, we adopt the framework of Bobkov and Götze [44] and consider a probability space $(\Omega, \mathcal{F}, \mu)$ together with a pair $(\mathcal{A}, \Gamma)$ that satisfies the following requirements:

- **(LSI-1)** $\mathcal{A}$ is a family of bounded measurable functions on Ω, such that if $f \in \mathcal{A}$, then $af + b \in \mathcal{A}$ as well for any $a \ge 0$ and $b \in \mathbb{R}$.
- **(LSI-2)** Γ is an operator that maps functions in $\mathcal{A}$ to nonnegative measurable functions on Ω.
- **(LSI-3)** For any $f \in \mathcal{A}$, $a \ge 0$, and $b \in \mathbb{R}$, $\Gamma(af+b) = a\,\Gamma f$.

Then we say that μ satisfies a *logarithmic Sobolev inequality* with constant $c \ge 0$, or LSI(c) for short, if

$$D(\mu^{(f)}\|\mu) \le \frac{c}{2}\,\mathbb{E}_\mu^{(f)}\left[(\Gamma f)^2\right], \qquad \forall f \in \mathcal{A}. \tag{3.93}$$

Here, as before, $\mu^{(f)}$ denotes the f-tilting of μ, i.e.,

$$\frac{\mathrm{d}\mu^{(f)}}{\mathrm{d}\mu} = \frac{\exp(f)}{\mathbb{E}_\mu[\exp(f)]},$$

and $\mathbb{E}_\mu^{(f)}[\cdot]$ denotes expectation w.r.t. $\mu^{(f)}$.

Remark 27. We have expressed the log-Sobolev inequality using standard information-theoretic notation. Most of the mathematics literature dealing with the subject, however, uses a different notation, which we briefly summarize for the reader's benefit. Given a probability measure μ on Ω and a nonnegative function $g : \Omega \to \mathbb{R}$, define the *entropy functional*

$$\begin{aligned}\mathrm{Ent}_\mu(g) &\triangleq \int g \ln g \,\mathrm{d}\mu - \int g\,\mathrm{d}\mu \cdot \ln\left(\int g\,\mathrm{d}\mu\right)\\ &\equiv \mathbb{E}_\mu[g\ln g] - \mathbb{E}_\mu[g]\,\ln \mathbb{E}_\mu[g].\end{aligned}$$

Then the LSI(c) condition can be equivalently written as (cf. [44, p. 2])

$$\mathrm{Ent}_\mu\big(\exp(f)\big) \le \frac{c}{2} \int (\Gamma f)^2 \exp(f)\, \mathrm{d}\mu \tag{3.94}$$

with the convention that $0 \ln 0 \triangleq 0$. To see the equivalence of (3.93) and (3.94), note that

$$\begin{aligned} &\mathrm{Ent}_\mu\big(\exp(f)\big) \\ &= \int \exp(f) \ln\left(\frac{\exp(f)}{\int \exp(f) \mathrm{d}\mu}\right) \mathrm{d}\mu \\ &= \mathbb{E}_\mu[\exp(f)] \int \Big(\frac{\mathrm{d}\mu^{(f)}}{\mathrm{d}\mu}\Big) \ln\Big(\frac{\mathrm{d}\mu^{(f)}}{\mathrm{d}\mu}\Big)\, \mathrm{d}\mu \\ &= \mathbb{E}_\mu[\exp(f)] \cdot D(\mu^{(f)} \| \mu) \end{aligned} \tag{3.95}$$

and

$$\begin{aligned} &\int (\Gamma f)^2 \exp(f)\, \mathrm{d}\mu \\ &= \mathbb{E}_\mu[\exp(f)] \int (\Gamma f)^2\, \mathrm{d}\mu^{(f)} \\ &= \mathbb{E}_\mu[\exp(f)] \cdot \mathbb{E}_\mu^{(f)} \left[(\Gamma f)^2\right]. \end{aligned} \tag{3.96}$$

Substituting (3.95) and (3.96) into (3.94), we obtain (3.93). We note that the entropy functional Ent is homogeneous: for any g such that $\mathrm{Ent}_\mu(g) < \infty$ and any $c > 0$, we have

$$\mathrm{Ent}_\mu(cg) = c\, \mathbb{E}_\mu\left[g \ln \frac{g}{\mathbb{E}_\mu[g]}\right] = c\, \mathrm{Ent}_\mu(g).$$

Remark 28. Strictly speaking, (3.93) should be called a modified (or exponential) logarithmic Sobolev inequality. The ordinary log-Sobolev inequality takes the form

$$\mathrm{Ent}_\mu(g^2) \le 2c \int (\Gamma g)^2\, \mathrm{d}\mu \tag{3.97}$$

for all strictly positive $g \in \mathcal{A}$. If the pair $(\mathcal{A}, \Gamma)$ is such that $\psi \circ g \in \mathcal{A}$ for any $g \in \mathcal{A}$ and any C^∞ function $\psi : \mathbb{R} \to \mathbb{R}$, and Γ obeys the chain rule

$$\Gamma(\psi \circ g) = |\psi' \circ g|\, \Gamma g, \qquad \forall g \in \mathcal{A}, \psi \in C^\infty \tag{3.98}$$

then (3.93) and (3.97) are equivalent. Indeed, if (3.97) holds, then using it with $g = \exp(f/2)$ gives

$$\begin{aligned} \mathrm{Ent}_\mu\big(\exp(f)\big) &\le 2c \int \Big(\Gamma\big(\exp(f/2)\big)\Big)^2\, \mathrm{d}\mu \\ &= \frac{c}{2} \int (\Gamma f)^2 \exp(f)\, \mathrm{d}\mu \end{aligned}$$

which is (3.94). Note that the last equality follows from (3.98) which implies that

$$\Gamma\big(\exp(f/2)\big) = \frac{1}{2}\, \exp(f/2) \cdot \Gamma f.$$

Conversely, using (3.94) with $f = 2\ln g$, we get (note that if follows from (3.98) that $\Gamma(2\ln g) = \frac{2\Gamma g}{g}$ where $g \geq 0$)

$$\begin{aligned} \mathrm{Ent}_\mu(g^2) &\leq \frac{c}{2}\int |\Gamma(2\ln g)|^2\, g^2\, \mathrm{d}\mu \\ &= 2c\int (\Gamma g)^2 \mathrm{d}\mu, \end{aligned}$$

which is (3.97). In fact, the Gaussian log-Sobolev inequality we have looked at in Section 3.2 is an instance, in which this equivalence holds with $\Gamma f = ||\nabla f||$ clearly satisfying the product rule (3.98).

Recalling the discussion of Section 3.1.4, we now show how we can pass from a log-Sobolev inequality to a concentration inequality via the Herbst argument. Indeed, let $\Omega = \mathcal{X}^n$ and $\mu = P$, and suppose that P satisfies LSI(c) on an appropriate pair $(\mathcal{A}, \Gamma)$. Suppose, furthermore, that the function of interest f is an element of $\mathcal{A}$ and that $\|\Gamma(f)\|_\infty < \infty$ (otherwise, LSI(c) is vacuously true for any c). Then $tf \in \mathcal{A}$ for any $t \geq 0$, so applying (3.93) to $g = tf$ we get

$$\begin{aligned} D\big(P^{(tf)} \| P\big) &\leq \frac{c}{2}\mathbb{E}_P^{(f)}\left[(\Gamma(tf))^2\right] \\ &= \frac{ct^2}{2}\mathbb{E}_P^{(tf)}\left[(\Gamma f)^2\right] \\ &\leq \frac{c\|\Gamma f\|_\infty^2 t^2}{2}, \end{aligned} \tag{3.99}$$

where the second step uses the fact that $\Gamma(tf) = t\Gamma f$ for any $f \in \mathcal{A}$ and any $t \geq 0$. In other words, P satisfies the bound (3.33) for every $g \in \mathcal{A}$ with $E(g) = \|\Gamma g\|_\infty^2$. Therefore, using the bound (3.99) together with Corollary 7, we arrive at

$$\mathbb{P}\big(f(X^n) \geq \mathbb{E}f(X^n) + r\big) \leq \exp\left(-\frac{r^2}{2c\|\Gamma f\|_\infty^2}\right), \qquad \forall r \geq 0. \tag{3.100}$$

3.3.1 Tensorization of the logarithmic Sobolev inequality

In the above demonstration, we have capitalized on an appropriate log-Sobolev inequality in order to derive a concentration inequality. Showing that a log-Sobolev inequality actually holds can be very difficult for reasons discussed in Section 3.1.3. However, when the probability measure P is a product measure, i.e., the random variables $X_1, \ldots, X_n \in \mathcal{X}$ are independent under P, we can, once again, use the "divide-and-conquer" tensorization strategy: we break the original n-dimensional problem into n one-dimensional subproblems, then establish that each marginal distribution P_{X_i}, $i = 1, \ldots, n$, satisfies a log-Sobolev inequality for a suitable class of real-valued functions on $\mathcal{X}$, and finally appeal to the tensorization bound for the relative entropy.

Let us provide the abstract scheme first. Suppose that for each $i \in \{1, \ldots, n\}$ we have a pair $(\mathcal{A}_i, \Gamma_i)$ defined on $\mathcal{X}$ that satisfies the requirements (LSI-1)–(LSI-3) listed at the beginning of Section 3.3. Recall that for any function $f : \mathcal{X}^n \to \mathbb{R}$, for any $i \in \{1, \ldots, n\}$, and any $(n-1)$-tuple $\bar{x}^i = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$, we have defined a function $f_i(\cdot|\bar{x}^i) : \mathcal{X} \to \mathbb{R}$ via $f_i(x_i|\bar{x}^i) \triangleq f(x^n)$. Then, we have the following:

Theorem 24. Let $X_1, \ldots, X_n \in \mathcal{X}$ be n independent random variables, and let $P = P_{X_1} \otimes \ldots \otimes P_{X_n}$ be their joint distribution. Let $\mathcal{A}$ consist of all functions $f : \mathcal{X}^n \to \mathbb{R}$ such that, for every $i \in \{1, \ldots, n\}$,

$$f_i(\cdot|\bar{x}^i) \in \mathcal{A}_i, \qquad \forall\, \bar{x}^i \in \mathcal{X}^{n-1}. \tag{3.101}$$

Define the operator Γ that maps each $f \in \mathcal{A}$ to

$$\Gamma f = \sqrt{\sum_{i=1}^n (\Gamma_i f_i)^2}, \tag{3.102}$$

which is shorthand for

$$\Gamma f(x^n) = \sqrt{\sum_{i=1}^n \Big(\Gamma_i f_i(x_i|\bar{x}^i)\Big)^2}, \qquad \forall\, x^n \in \mathcal{X}^n. \tag{3.103}$$

Then the following statements hold:

1. If there exists a constant $c \ge 0$ such that, for every i, P_{X_i} satisfies LSI(c) with respect to $(\mathcal{A}_i, \Gamma_i)$, then P satisfies LSI(c) with respect to $(\mathcal{A}, \Gamma)$.
2. For any $f \in \mathcal{A}$ with $\mathbb{E}[f(X^n)] = 0$, and any $r \ge 0$,

$$\mathbb{P}\big(f(X^n) \ge r\big) \le \exp\left(-\frac{r^2}{2c\|\Gamma f\|_\infty^2}\right). \tag{3.104}$$

Proof. We first check that the pair $(\mathcal{A}, \Gamma)$, defined in the statement of the theorem, satisfies the requirements (LSI-1)–(LSI-3). Thus, consider some $f \in \mathcal{A}$, choose some $a \ge 0$ and $b \in \mathbb{R}$, and let $g = af + b$. Then, for any i and any $\bar{x}^i$,

$$\begin{aligned} g_i(\cdot|\bar{x}^i) &= g(x_1, \dots, x_{i-1}, \cdot, x_{i+1}, \dots, x_n) \\ &= af(x_1, \dots, x_{i-1}, \cdot, x_{i+1}, \dots, x_n) + b \\ &= af_i(\cdot|\bar{x}^i) + b \in \mathcal{A}_i, \end{aligned}$$

where the last step uses (3.101). Hence, $f \in \mathcal{A}$ implies that $g = af + b \in \mathcal{A}$ for any $a \ge 0, b \in \mathbb{R}$, so (LSI-1) holds. From the definitions of Γ in (3.102) and (3.103) it is readily seen that (LSI-2) and (LSI-3) hold as well.

Next, for any $f \in \mathcal{A}$ and any $t \ge 0$, we have

$$\begin{aligned} D\big(P^{(tf)} \| P\big) &\le \sum_{i=1}^n D\Big(P^{(tf)}_{X_i|\bar{X}^i} \Big\| P_{X_i} \Big| P^{(tf)}_{\bar{X}^i}\Big) \\ &= \sum_{i=1}^n \int P^{(tf)}_{\bar{X}^i}(\mathrm{d}\bar{x}^i) D\Big(P^{(tf)}_{X_i|\bar{X}^i=\bar{x}^i} \Big\| P_{X_i}\Big) \\ &= \sum_{i=1}^n \int P^{(tf)}_{\bar{X}^i}(\mathrm{d}\bar{x}^i) D\Big(P^{(tf_i(\cdot|\bar{x}^i))}_{X_i} \Big\| P_{X_i}\Big) \\ &\le \frac{ct^2}{2} \sum_{i=1}^n \int P^{(tf)}_{\bar{X}^i}(\mathrm{d}\bar{x}^i)\, \mathbb{E}^{(tf_i(\cdot|\bar{x}^i))}_{X_i} \Big[\big(\Gamma_i f_i(X_i|\bar{x}^i)\big)^2\Big] \\ &= \frac{ct^2}{2} \sum_{i=1}^n \mathbb{E}^{(tf)}_{P_{\bar{X}^i}} \Big\{ \mathbb{E}^{(tf)}_{P_{X_i}} \Big[\big(\Gamma_i f_i(X_i|\bar{X}^i)\big)^2 \Big| \bar{X}^i \Big] \Big\} \\ &= \frac{ct^2}{2} \cdot \mathbb{E}^{(tf)}_P \big[(\Gamma f)^2\big], \end{aligned} \tag{3.105}$$

where the first step uses Proposition 5 with $Q = P^{(tf)}$, the second is by the definition of conditional divergence where $P_{X_i} = P_{X_i|\bar{X}^i}$, the third is due to (3.25), the fourth uses the fact that (a) $f_i(\cdot|\bar{x}^i) \in \mathcal{A}_i$

for all $\bar{x}^i$ and (b) P_{X_i} satisfies LSI(c) w.r.t. $(\mathcal{A}_i, \Gamma_i)$, and the last step uses the tower property of the conditional expectation, as well as (3.102). We have thus proved the first part of the proposition, i.e., that P satisfies LSI(c) w.r.t. the pair $(\mathcal{A}, \Gamma)$. The second part follows from the same argument that was used to prove (3.100). □

3.3.2 Maurer's thermodynamic method

With Theorem 24 at our disposal, we can now establish concentration inequalities in product spaces whenever an appropriate log-Sobolev inequality can be shown to hold for each individual variable. Thus, the bulk of the effort is in showing that this is, indeed, the case for a given probability measure P and a given class of functions. Ordinarily, this is done on a case-by-case basis. However, as shown recently by A. Maurer in an insightful paper [126], it is possible to derive log-Sobolev inequalities in a wide variety of settings by means of a single unified method. This method has two basic ingredients:

1. A certain "thermodynamic" representation of the divergence $D(\mu^{(f)} \| \mu)$, $f \in \mathcal{A}$, as an integral of the *variances* of f w.r.t. the tilted measures $\mu^{(tf)}$ for all $t \in (0, 1)$.

2. Derivation of upper bounds on these variances in terms of an appropriately chosen operator Γ acting on $\mathcal{A}$, where $\mathcal{A}$ and Γ are the objects satisfying the conditions (LSI-1)–(LSI-3).

In this section, we will state two lemmas that underlie these two ingredients and then describe the overall method in broad strokes. Several detailed demonstrations of the method in action will be given in the sections that follow.

Once again, consider a probability space $(\Omega, \mathcal{F}, \mu)$ and recall the definition of the g-tilting of μ:

$$\frac{\mathrm{d}\mu^{(g)}}{\mathrm{d}\mu} = \frac{\exp(g)}{\mathbb{E}_\mu[\exp(g)]}.$$

The variance of any $h : \Omega \to \mathbb{R}$ w.r.t. $\mu^{(g)}$ is then given by

$$\mathsf{var}_\mu^{(g)}[h] \triangleq \mathbb{E}_\mu^{(g)}[h^2] - \left(\mathbb{E}_\mu^{(g)}[h]\right)^2.$$

The first ingredient of Maurer's method is encapsulated in the following (see [126, Theorem 3]):

Lemma 9 (Representation of the divergence in terms of thermal fluctuations)**.** Consider a function $f : \Omega \to \mathbb{R}$, such that $\mathbb{E}_\mu[\exp(\lambda f)] < \infty$ for all $\lambda > 0$. Then

$$D\big(\mu^{(\lambda f)} \big\| \mu\big) = \int_0^\lambda \int_t^\lambda \mathsf{var}_\mu^{(sf)}[f] \, \mathrm{d}s \, \mathrm{d}t. \tag{3.106}$$

Remark 29. The "thermodynamic" interpretation of the above result stems from the fact that the tilted measures $\mu^{(tf)}$ can be viewed as the *Gibbs measures* that are used in statistical mechanics as a probabilistic description of physical systems in thermal equilibrium. In this interpretation, the underlying space Ω is the state (or configuration) space of some physical system Σ, the elements $x \in \Omega$ are the states (or configurations) of Σ, μ is some base (or reference) measure, and f is the energy function. We can view μ as some initial distribution of the system state. According to the postulates of statistical physics, the thermal equilibrium of Σ at absolute temperature θ corresponds to that distribution ν on Ω that will globally minimize the *free energy functional*

$$\Psi_\theta(\nu) \triangleq \mathbb{E}_\nu[f] + \theta D(\nu \| \mu). \tag{3.107}$$

It is claimed that $\Psi_\theta(\nu)$ is uniquely minimized by $\nu^* = \mu^{(-tf)}$, where $t = 1/\theta$ is the *inverse temperature*. To see this, consider an arbitrary ν, where we may assume, without loss of generality, that $\nu \ll \mu$. Let $\psi \triangleq \mathrm{d}\nu/\mathrm{d}\mu$. Then

$$\frac{\mathrm{d}\nu}{\mathrm{d}\mu^{(-tf)}} = \frac{\frac{\mathrm{d}\nu}{\mathrm{d}\mu}}{\frac{\mathrm{d}\mu^{(-tf)}}{\mathrm{d}\mu}} = \frac{\psi}{\frac{\exp(-tf)}{\mathbb{E}_\mu[\exp(-tf)]}} = \psi\,\exp(tf)\,\mathbb{E}_\mu[\exp(-tf)]$$

and

$$\begin{aligned}\Psi_\theta(\nu) &= \frac{1}{t}\,\mathbb{E}_\nu[tf + \ln\psi] \\ &= \frac{1}{t}\,\mathbb{E}_\nu\left[\ln\bigl(\psi\exp(tf)\bigr)\right] \\ &= \frac{1}{t}\,\mathbb{E}_\nu\left[\ln\frac{\mathrm{d}\nu}{\mathrm{d}\mu^{(-tf)}} - \Lambda(-t)\right] \\ &= \frac{1}{t}\left[D(\nu\|\mu^{(-tf)}) - \Lambda(-t)\right],\end{aligned}$$

where, as before, $\Lambda(-t) \triangleq \ln\bigl(\mathbb{E}_\mu[\exp(-tf)]\bigr)$ is the logarithmic moment generating function of f w.r.t. μ. Therefore, $\Psi_\theta(\nu) = \Psi_{1/t}(\nu) \ge -\Lambda(-t)/t$, with equality if and only if $\nu = \mu^{(-tf)}$.

Now we give the proof of Lemma 9:

Proof. We start by noting that (see (3.11))

$$\Lambda'(t) = \mathbb{E}_\mu^{(tf)}[f] \qquad \text{and} \qquad \Lambda''(t) = \mathsf{var}_\mu^{(tf)}[f], \tag{3.108}$$

and, in particular, $\Lambda'(0) = \mathbb{E}_\mu[f]$. Moreover, from (3.13), we get

$$D\bigl(\mu^{(\lambda f)}\big\|\mu\bigr) = \lambda^2\,\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{\Lambda(\lambda)}{\lambda}\right) = \lambda\Lambda'(\lambda) - \Lambda(\lambda). \tag{3.109}$$

Now, using (3.108), we get

$$\begin{aligned}\lambda\Lambda'(\lambda) &= \int_0^\lambda \Lambda'(\lambda)\mathrm{d}t \\ &= \int_0^\lambda \left(\int_0^\lambda \Lambda''(s)\mathrm{d}s + \Lambda'(0)\right)\mathrm{d}t \\ &= \int_0^\lambda \left(\int_0^\lambda \mathsf{var}_\mu^{(sf)}[f]\,\mathrm{d}s + \mathbb{E}_\mu[f]\right)\mathrm{d}t\end{aligned} \tag{3.110}$$

and

$$\begin{aligned}\Lambda(\lambda) &= \int_0^\lambda \Lambda'(t)\,\mathrm{d}t \\ &= \int_0^\lambda \left(\int_0^t \Lambda''(s)\,\mathrm{d}s + \Lambda'(0)\right)\mathrm{d}t \\ &= \int_0^\lambda \left(\int_0^t \mathsf{var}_\mu^{(sf)}[f]\,\mathrm{d}s + \mathbb{E}_\mu[f]\right)\mathrm{d}t.\end{aligned} \tag{3.111}$$

Substituting (3.110) and (3.111) into (3.109), we get (3.106). □

Now the whole affair hinges on the second step, which involves bounding the variances $\mathsf{var}_\mu^{(tf)}[f]$, for $t > 0$, from above in terms of expectations $\mathbb{E}_\mu^{(tf)}\left[(\Gamma f)^2\right]$ for an appropriately chosen Γ. The following is sufficiently general for our needs:

Theorem 25. Let the objects $(\mathcal{A}, \Gamma)$ and $\{(\mathcal{A}_i, \Gamma_i)\}_{i=1}^n$ be constructed as in the statement of Theorem 24. Suppose, furthermore, that, for each i, the operator Γ_i maps each $g \in \mathcal{A}_i$ to a constant (which may depend on g), and there exists a constant $c > 0$ such that the bound

$$\mathsf{var}_i^{(sg)}[g(X_i)|\bar{X}^i = \bar{x}^i] \le c\,(\Gamma_i g)^2\,, \qquad \forall \bar{x}^i \in \mathcal{X}^{n-1} \tag{3.112}$$

holds for all $i \in \{1, \dots, n\}$, $s > 0$, and $g \in \mathcal{A}_i$, where $\mathsf{var}_i^{(g)}[\cdot|\bar{X}^i = \bar{x}^i]$ denotes the (conditional) variance w.r.t. $P^{(g)}_{X_i|\bar{X}^i=\bar{x}^i}$. Then, the pair $(\mathcal{A}, \Gamma)$ satisfies LSI(c) w.r.t. P_{X^n}.

Proof. Given a function $g : \mathcal{X}_i \to \mathbb{R}$ in $\mathcal{A}_i$, let $\mathsf{var}_i^{(g)}[\cdot|\bar{X}^i]$ denote the conditional variance w.r.t. $P^{(g)}_{X_i|\bar{X}^i}$. Then we can write

$$\begin{aligned}
D\Big(P^{(f)}_{X_i|\bar{X}^i=\bar{x}^i}\Big\|P_{X_i}\Big) &= D\Big(P^{(f_i(\cdot|\bar{x}^i))}_{X_i}\Big\|P_{X_i}\Big) \\
&= \int_0^1 \int_t^1 \mathsf{var}_i^{(sf_i(\cdot|\bar{x}^i))}[f_i(X_i|\bar{X}^i)|\bar{X}^i = \bar{x}^i]\,\mathrm{d}s\,\mathrm{d}t \\
&\le c\,(\Gamma_i f_i)^2 \int_0^\lambda \int_t^\lambda \mathrm{d}s\,\mathrm{d}t \\
&= \frac{c(\Gamma_i f_i)^2\,\lambda^2}{2}.
\end{aligned}$$

where the first step uses the fact that $P^{(f)}_{X_i|\bar{X}^i=\bar{x}^i}$ is equal to the $f_i(\cdot|\bar{x}^i)$-tilting of P_{X_i}, the second step uses Lemma 9, and the third step uses (3.112) with $g = f_i(\cdot|\bar{x}^i)$. We have therefore established that, for each i, the pair $(\mathcal{A}_i, \Gamma_i)$ satisfies LSI(c). Therefore, the pair $(\mathcal{A}, \Gamma)$ satisfies LSI(c) by Theorem 24. □

The following two lemmas will be useful for establishing bounds like (3.112):

Lemma 10. Let $U \in \mathbb{R}$ be a random variable such that $U \in [a, b]$ a.s. for some $-\infty < a \le b < +\infty$. Then

$$\mathsf{var}[U] \le (b - \mathbb{E}U)(\mathbb{E}U - a) \le \frac{(b-a)^2}{4}. \tag{3.113}$$

Proof. The first inequality in (3.113) follows by direct calculation:

$$\begin{aligned}
\mathsf{var}[U] &= \mathbb{E}[(U - \mathbb{E}U)^2] \\
&\le (b - \mathbb{E}U)(\mathbb{E}U - a).
\end{aligned}$$

The second line is due to the fact that the function $u \mapsto (b - u)(u - a)$ takes its maximum value of $(b - a)^2/4$ at $u = (a + b)/2$. □

Lemma 11. [126, Lemma 9] Let $f : \Omega \to \mathbb{R}$ be such that $f - \mathbb{E}_\mu[f] \le C$ for some $C \in \mathbb{R}$. Then for any $t > 0$ we have

$$\mathsf{var}_\mu^{(tf)}[f] \le \exp(tC)\,\mathsf{var}_\mu[f]$$

Proof. Because $\mathsf{var}_\mu[f] = \mathsf{var}_\mu[f + c]$ for any constant $c \in \mathbb{R}$, we have

$$\begin{aligned}
\mathsf{var}_\mu^{(tf)}[f] &= \mathsf{var}_\mu^{(tf)}\Big\{f - \mathbb{E}_\mu\,[f]\Big\} \\
&\le \mathbb{E}_\mu^{(tf)}\Big[(f - \mathbb{E}_\mu[f])^2\Big] && (3.114) \\
&= \mathbb{E}_\mu\left[\frac{\exp(tf)\,(f - \mathbb{E}_\mu[f])^2}{\mathbb{E}_\mu[\exp(tf)]}\right] && (3.115) \\
&\le \mathbb{E}_\mu\Big\{(f - \mathbb{E}_\mu[f])^2 \exp\left[t\,(f - \mathbb{E}_\mu[f])\right]\Big\} && (3.116) \\
&\le \exp(tC)\,\mathbb{E}_\mu\Big[(f - \mathbb{E}_\mu\,[f])^2\Big]\,, && (3.117)
\end{aligned}$$

where:

- (3.114) uses the bound $\mathsf{var}[U] \le \mathbb{E}U^2$;
- (3.115) is by definition of the tilted distribution $\mu^{(tf)}$;
- (3.116) follows from applying Jensen's inequality to the denominator; and
- (3.117) uses the assumption that $f - \mathbb{E}_\mu[f] \le C$ and the monotonicity of $\exp(\cdot)$.

This completes the proof of Lemma 11. □

3.3.3 Discrete logarithmic Sobolev inequalities on the Hamming cube

We now use Maurer's method to derive log-Sobolev inequalities for functions of n i.i.d. Bernoulli random variables. Let $\mathcal{X}$ be the two-point set $\{0,1\}$, and let $e_i \in \mathcal{X}^n$ denote the binary string that has 1 in the ith position and zeros elsewhere. Finally, for any $f : \mathcal{X}^n \to \mathbb{R}$ define

$$\Gamma f(x^n) \triangleq \sqrt{\sum_{i=1}^n \big(f(x^n \oplus e_i) - f(x^n)\big)^2}, \qquad \forall x^n \in \mathcal{X}^n, \tag{3.118}$$

where the modulo-2 addition $\oplus$ is defined componentwise. In other words, Γf measures the sensitivity of f to local bit flips. We consider the symmetric, i.e., Bernoulli(1/2), case first:

Theorem 26 (Discrete log-Sobolev inequality for the symmetric Bernoulli measure)**.** Let $\mathcal{A}$ be the set of all the functions $f : \mathcal{X}^n \to \mathbb{R}$. Then, the pair $(\mathcal{A}, \Gamma)$ with Γ defined in (3.118) satisfies the conditions (LSI-1)–(LSI-3). Let $X_1, \ldots, X_n$ be n i.i.d. Bernoulli(1/2) random variables, and let P denote their distribution. Then, P satisfies LSI(1/4) w.r.t. $(\mathcal{A}, \Gamma)$. In other words, for any $f : \mathcal{X}^n \to \mathbb{R}$,

$$D\big(P^{(f)} \| P\big) \le \frac{1}{8}\, \mathbb{E}_P^{(f)} \Big[(\Gamma f)^2\Big]. \tag{3.119}$$

Proof. Let $\mathcal{A}_0$ be the set of all functions $g : \{0,1\} \to \mathbb{R}$, and let Γ_0 be the operator that maps every $g \in \mathcal{A}_0$ to

$$\Gamma g \triangleq |g(0) - g(1)| = |g(x) - g(x \oplus 1)|, \quad \forall\, x \in \{0,1\}. \tag{3.120}$$

For each $i \in \{1, \ldots, n\}$, let $(\mathcal{A}_i, \Gamma_i)$ be a copy of $(\mathcal{A}_0, \Gamma_0)$. Then, each Γ_i maps every function $g \in \mathcal{A}_i$ to the constant $|g(0) - g(1)|$. Moreover, for any $g \in \mathcal{A}_i$, the random variable $U_i = g(X_i)$ is bounded between $g(0)$ and $g(1)$, where we can assume without loss of generality that $g(0) \le g(1)$. Hence, by Lemma 10, we have

$$\mathsf{var}_{P_i}^{(sg)}[g(X_i)|\bar{X}^i = \bar{x}^i] \le \frac{\big(g(0) - g(1)\big)^2}{4} = \frac{(\Gamma_i g)^2}{4}, \qquad \forall g \in \mathcal{A}_i,\ \bar{x}^i \in \mathcal{X}^{n-1}. \tag{3.121}$$

In other words, the condition (3.112) of Theorem 25 holds with $c = 1/4$. In addition, it is easy to see that the operator Γ constructed from $\Gamma_1, \ldots, \Gamma_n$ according to (3.102) is precisely the one in (3.118). Therefore, by Theorem 25, the pair $(\mathcal{A}, \Gamma)$ satisfies LSI(1/4) w.r.t. P, which proves (3.119). This completes the proof of Theorem 26. □

Remark 30. The log-Sobolev inequality in (3.119) is an exponential form of the original log-Sobolev inequality for the Bernoulli(1/2) measure derived by Gross [35], which reads:

$$\mathrm{Ent}_P[g^2] \le \frac{(g(0) - g(1))^2}{2}. \tag{3.122}$$

To see this, define f by $e^f = g^2$, where we may assume without loss of generality that $0 < g(0) \le g(1)$. To show that (3.122) implies (3.119), note that

$$\begin{aligned}
\big(g(0)-g(1)\big)^2 &= \big(\exp\big(f(0)/2\big)-\exp\big(f(1)/2\big)\big)^2 \\
&\le \frac{1}{8}\left[\exp\big(f(0)\big)+\exp\big(f(1)\big)\right]\big(f(0)-f(1)\big)^2 \\
&= \frac{1}{4}\mathbb{E}_P\left[\exp(f)(\Gamma f)^2\right] \qquad (3.123)
\end{aligned}$$

with $\Gamma f = |f(0)-f(1)|$, where the inequality follows from the easily verified fact that $(1-x)^2 \le \frac{(1+x^2)(\ln x)^2}{2}$ for all $x \ge 0$, which we apply to $x \triangleq g(1)/g(0)$. Therefore, the inequality in (3.122) implies the following:

$$\begin{aligned}
& D(P^{(f)}\|P) \\
&= \frac{\mathrm{Ent}_P[\exp(f)]}{\mathbb{E}_P[\exp(f)]} \qquad (3.124) \\
&= \frac{\mathrm{Ent}_P[g^2]}{\mathbb{E}_P[\exp(f)]} \qquad (3.125) \\
&\le \frac{\big(g(0)-g(1)\big)^2}{2\,\mathbb{E}_P[\exp(f)]} \qquad (3.126) \\
&\le \frac{\mathbb{E}_P[\exp(f)\,(\Gamma f)^2]}{8\,\mathbb{E}_P[\exp(f)]} \qquad (3.127) \\
&= \frac{1}{8}\,\mathbb{E}_P^{(f)}\left[(\Gamma f)^2\right] \qquad (3.128)
\end{aligned}$$

where equality (3.124) follows from (3.95), equality (3.125) holds due to the equality $e^f = g^2$, inequality (3.126) holds due to (3.122), inequality (3.127) follows from (3.123), and equality (3.128) follows by definition of the expectation w.r.t. the tilted probability measure $P^{(f)}$. Therefore, it is concluded that indeed (3.122) implies (3.119).

Gross used (3.122) and the central limit theorem to establish his Gaussian log-Sobolev inequality (see Theorem 21). We can follow the same steps and arrive at (3.34) from (3.119). To that end, let $g : \mathbb{R} \to \mathbb{R}$ be a sufficiently smooth function (to guarantee, at least, that both $g\exp(g)$ and the derivative of g are continuous and bounded), and define the function $f : \{0,1\}^n \to \mathbb{R}$ by

$$f(x_1,\ldots,x_n) \triangleq g\left(\frac{x_1+x_2+\ldots+x_n-n/2}{\sqrt{n/4}}\right).$$

If $X_1,\ldots,X_n$ are i.i.d. Bernoulli(1/2) random variables, then, by the central limit theorem, the sequence of probability measures $\{P_{Z^n}\}_{n=1}^{\infty}$ with

$$Z_n \triangleq \frac{X_1+\ldots+X_n-n/2}{\sqrt{n/4}}$$

converges weakly to the standard Gaussian distribution G as $n \to \infty$. Therefore, by the assumed smoothness properties of g we have

$$\begin{aligned}
\mathbb{E}\left[\exp\big(f(X^n)\big)\right]\cdot D\big(P_{X^n}^{(f)}\big\|P_{X^n}\big) &= \mathbb{E}\left[f(X^n)\exp\big(f(X^n)\big)\right] - \mathbb{E}[\exp\big(f(X^n)\big)]\ln\mathbb{E}[\exp\big(f(X^n)\big)] \\
&= \mathbb{E}\left[g(Z_n)\exp\big(g(Z_n)\big)\right] - \mathbb{E}[\exp\big(g(Z_n)\big)]\ln\mathbb{E}[\exp\big(g(Z_n)\big)] \\
&\xrightarrow{n\to\infty} \mathbb{E}\left[g(Z)\exp\big(g(Z)\big)\right] - \mathbb{E}[\exp\big(g(Z)\big)]\ln\mathbb{E}[\exp\big(g(Z)\big)] \\
&= \mathbb{E}\left[\exp\left(g(Z)\right)\right] D\big(P_Z^{(g)}\big\|P_Z\big) \qquad (3.129)
\end{aligned}$$

where $Z \sim G$ is a standard Gaussian random variable. Moreover, using the definition (3.118) of Γ and the smoothness of g, for any $i \in \{1, \ldots, n\}$ and $x^n \in \{0,1\}^n$ we have

$$\begin{aligned} |f(x^n \oplus e_i) - f(x^n)|^2 &= \left| g\left(\frac{x_1 + \ldots + x_n - n/2}{\sqrt{n/4}} + \frac{(-1)^{x_i}}{\sqrt{n/4}} \right) - g\left(\frac{x_1 + \ldots + x_n - n/2}{\sqrt{n/4}} \right) \right|^2 \\ &= \frac{4}{n} \left(g'\left(\frac{x_1 + \ldots + x_n - n/2}{\sqrt{n/4}} \right) \right)^2 + o\left(\frac{1}{n}\right), \end{aligned}$$

which implies that

$$\begin{aligned} |\Gamma f(x^n)|^2 &= \sum_{i=1}^n \left(f(x^n \oplus e_i) - f(x^n) \right)^2 \\ &= 4 \left(g'\left(\frac{x_1 + \ldots + x_n - n/2}{\sqrt{n/4}} \right) \right)^2 + o\left(1\right). \end{aligned}$$

Consequently,

$$\begin{aligned} \mathbb{E}\left[\exp\left(f(X^n)\right)\right] \cdot \mathbb{E}^{(f)}\left[\left(\Gamma f(X^n)\right)^2 \right] &= \mathbb{E}\left[\exp\left(f(X^n)\right) \left(\Gamma f(X^n)\right)^2 \right] \\ &= 4\,\mathbb{E}\left[\exp\left(g(Z_n)\right) \left(\left(g'(Z_n)\right)^2 + o(1) \right) \right] \\ &\xrightarrow{n\to\infty} 4\,\mathbb{E}\left[\exp\left(g(Z)\right) \left(g'(Z)\right)^2 \right] \\ &= 4\,\mathbb{E}\left[\exp\left(g(Z)\right)\right] \cdot \mathbb{E}^{(g)}\left[\left(g'(Z)\right)^2 \right]. \end{aligned} \tag{3.130}$$

Taking the limit of both sides of (3.119) as $n \to \infty$ and then using (3.129) and (3.130), we obtain

$$D\big(P_Z^{(g)} \| P_Z\big) \le \frac{1}{2}\,\mathbb{E}^{(g)}\left[\left(g'(Z)\right)^2 \right],$$

which is (3.44).

Now let us consider the case when $X_1, \ldots, X_n$ are i.i.d. Bernoulli(p) random variables with some $p \neq 1/2$. We will use Maurer's method to give an alternative, simpler proof of the following result of Ledoux [42, Corollary 5.9]:

Theorem 27. Consider any function $f : \{0,1\}^n \to \mathbb{R}$ with the property that

$$\max_{i \in \{1,\ldots,n\}} |f(x^n \oplus e_i) - f(x^n)| \le c \tag{3.131}$$

for all $x^n \in \{0,1\}^n$. Let $X_1, \ldots, X_n$ be n i.i.d. Bernoulli(p) random variables, and let P be their joint distribution. Then

$$D\big(P^{(f)} \| P\big) \le pq \left(\frac{(c-1)\exp(c) + 1}{c^2} \right) \mathbb{E}^{(f)}\left[(\Gamma f)^2 \right], \tag{3.132}$$

where $q = 1 - p$.

Proof. Following the usual route, we will establish the $n = 1$ case first, and then scale up to arbitrary n by tensorization.

Let $a = |\Gamma(f)| = |f(0) - f(1)|$, where Γ is defined as in (3.120). Without loss of generality, we may assume that $f(0) = 0$ and $f(1) = a$. Then

$$\mathbb{E}[f] = pa \qquad \text{and} \qquad \mathsf{var}[f] = pqa^2. \tag{3.133}$$

Using (3.133) and Lemma 11, we can write for any $t > 0$

$$\mathsf{var}^{(tf)}[f] \le pqa^2 \exp(tqa).$$

Therefore, by Lemma 9 we have

$$\begin{aligned} D\big(P^{(f)} \big\| P\big) &\le pqa^2 \int_0^1 \int_t^1 \exp(sqa) \, \mathrm{d}s \, \mathrm{d}t \\ &= pqa^2 \left(\frac{(qa-1)\exp(qa)+1}{(qa)^2} \right) \\ &\le pqa^2 \left(\frac{(c-1)\exp(c)+1}{c^2} \right), \end{aligned}$$

where the last step follows from the fact that the function $u \mapsto u^{-2}[(u-1)\exp(u)+1]$ is nondecreasing in $u \ge 0$, and $0 \le qa \le a \le c$. Since $a^2 = (\Gamma f)^2$, we can write

$$D\big(P^{(f)} \big\| P\big) \le pq \left(\frac{(c-1)\exp(c)+1}{c^2} \right) \mathbb{E}^{(f)} \left[(\Gamma f)^2 \right],$$

so we have established (3.132) for $n = 1$.

Now consider an arbitrary $n \in \mathbb{N}$. Since the condition in (3.131) can be expressed as

$$\big| f_i(0|\bar{x}^i) - f_i(1|\bar{x}^i) \big| \le c, \qquad \forall\, i \in \{1, \dots, n\}, \ \bar{x}^i \in \{0,1\}^{n-1},$$

we can use (3.132) to write

$$D\Big(P_{X_i}^{(tf_i(\cdot|\bar{x}^i))} \Big\| P_{X_i} \Big) \le pq \left(\frac{(c-1)\exp c + 1}{c^2} \right) \mathbb{E}^{(f_i(\cdot|\bar{x}^i))} \Big[\big(\Gamma_i f_i(X_i|\bar{X}^i) \big)^2 \Big| \bar{X}^i = \bar{x}^i \Big]$$

for every $i = 1, \dots, n$ and all $\bar{x}^i \in \{0,1\}^{n-1}$. With this, the same sequence of steps that led to (3.105) in the proof of Theorem 24 can be used to complete the proof of (3.132) for arbitrary n. □

Remark 31. In order to capture the correct dependence on the Bernoulli parameter p, we had to use a more refined, distribution-dependent variance bound of Lemma 11, as opposed to a cruder bound of Lemma 10 that does not depend on the underlying distribution. Maurer's paper [126] has other examples.

Remark 32. The same technique based on the central limit theorem that was used to arrive at the Gaussian log-Sobolev inequality (3.44) can be utilized here as well: given a sufficiently smooth function $g : \mathbb{R} \to \mathbb{R}$, define $f : \{0,1\}^n \to \mathbb{R}$ by

$$f(x^n) \triangleq g\left(\frac{x_1 + \ldots + x_n - np}{\sqrt{npq}} \right).$$

and then apply (3.132) to it.

3.3.4 The method of bounded differences revisited

As our second illustration of the use of Maurer's method, we will give an information-theoretic proof of McDiarmid's inequality with the correct constant in the exponent (recall that the original proof in [25, 5] used the martingale method; the reader is referred to the derivation of McDiarmid's inequality via the martingale approach in Theorem 2 of the preceding chapter). Following the exposition in [126, Section 4.1], we have:

Theorem 28. Let $X_1, \dots, X_n \in \mathcal{X}$ be independent random variables. Consider a function $f : \mathcal{X}^n \to \mathbb{R}$ with $\mathbb{E}[f(X^n)] = 0$, and also suppose that there exist some constants $0 \le c_1, \dots, c_n < +\infty$ such that, for each $i \in \{1, \dots, n\}$,

$$\left|f_i(x|\bar{x}^i) - f_i(y|\bar{x}^i)\right| \le c_i, \qquad \forall x, y \in \mathcal{X}, \bar{x}^i \in \mathcal{X}^{n-1}. \tag{3.134}$$

Then, for any $r \ge 0$,

$$\mathbb{P}\Big(f(X^n) \ge r\Big) \le \exp\left(-\frac{2r^2}{\sum_{i=1}^n c_i^2}\right). \tag{3.135}$$

Proof. Let $\mathcal{A}_0$ be the set of all bounded measurable functions $g : \mathcal{X} \to \mathbb{R}$, and let Γ_0 be the operator that maps every $g \in \mathcal{A}_0$ to

$$\Gamma_0 g \triangleq \sup_{x \in \mathcal{X}} g(x) - \inf_{x \in \mathcal{X}} g(x).$$

Clearly, $\Gamma_0(ag+b) = a\Gamma_0 g$ for any $a \ge 0$ and $b \in \mathbb{R}$. Now, for each $i \in \{1, \dots, n\}$, let $(\mathcal{A}_i, \Gamma_i)$ be a copy of $(\mathcal{A}_0, \Gamma_0)$. Then, each Γ_i maps every function $g \in \mathcal{A}_i$ to a non-negative constant. Moreover, for any $g \in \mathcal{A}_i$, the random variable $U_i = g(X_i)$ is bounded between $\inf_{x \in \mathcal{X}} g(x)$ and $\sup_{x \in \mathcal{X}} g(x) \equiv \inf_{x \in \mathcal{X}} g(x) + \Gamma_i g$. Therefore, Lemma 10 gives

$$\mathsf{var}_i^{(sg)}[g(X_i)|\bar{X}^i = \bar{x}^i] \le \frac{(\Gamma_i g)^2}{4}, \qquad \forall\, g \in \mathcal{A}_i,\ \bar{x}^i \in \mathcal{X}^{n-1}.$$

Hence, the condition (3.112) of Theorem 25 holds with $c = 1/4$. Now let $\mathcal{A}$ be the set of all bounded measurable functions $f : \mathcal{X}^n \to \mathbb{R}$. Then for any $f \in \mathcal{A}$, $i \in \{1, \dots, n\}$, and $x^n \in \mathcal{X}^n$ we have

$$\begin{aligned}
&\sup_{x_i \in \mathcal{X}_i} f(x_1, \dots, x_i, \dots, x_n) - \inf_{x_i \in \mathcal{X}_i} f(x_1, \dots, x_i, \dots, x_n) \\
&\quad = \sup_{x_i \in \mathcal{X}_i} f_i(x_i|\bar{x}^i) - \inf_{x_i \in \mathcal{X}_i} f_i(x_i|\bar{x}^i) \\
&\quad = \Gamma_i f_i(\cdot|\bar{x}^i).
\end{aligned}$$

Thus, if we construct an operator Γ on $\mathcal{A}$ from $\Gamma_1, \dots, \Gamma_n$ according to (3.102), the pair $(\mathcal{A}, \Gamma)$ will satisfy the conditions of Theorem 24. Therefore, by Theorem 25, it follows that the pair $(\mathcal{A}, \Gamma)$ satisfies LSI(1/4) for *any* product probability measure on $\mathcal{X}^n$, i.e., the inequality

$$\mathbb{P}\Big(f(X^n) \ge r\Big) \le \exp\left(-\frac{2r^2}{\|\Gamma f\|_\infty^2}\right) \tag{3.136}$$

holds for any $r \ge 0$ and bounded f with $\mathbb{E}[f] = 0$. Now, if f satisfies (3.134), then

$$\begin{aligned}
\|\Gamma f\|_\infty^2 &= \sup_{x^n \in \mathcal{X}^n} \sum_{i=1}^n \big(\Gamma_i f_i(x_i|\bar{x}^i)\big)^2 \\
&\le \sum_{i=1}^n \sup_{x^n \in \mathcal{X}^n} \big(\Gamma_i f_i(x_i|\bar{x}^i)\big)^2 \\
&= \sum_{i=1}^n \sup_{x^n \in \mathcal{X}^n, y \in \mathcal{X}} |f_i(x_i|\bar{x}^i) - f(y|\bar{x}^i)|^2 \\
&\le \sum_{i=1}^n c_i^2.
\end{aligned}$$

Substituting this bound into the right-hand side of (3.136), we get (3.135). □

It is instructive to compare the strategy used to prove Theorem 28 with an earlier approach by Boucheron, Lugosi and Massart [127] using the entropy method. Their starting point is the following lemma:

Lemma 12. Define the function $\psi : \mathbb{R} \to \mathbb{R}$ by $\psi(u) = \exp(u) - u - 1$. Consider a probability space $(\Omega, \mathcal{F}, \mu)$ and a measurable function $f : \Omega \to \mathbb{R}$ such that tf is exponentially integrable w.r.t. μ for all $t \in \mathbb{R}$. Then, the following inequality holds for any $c \in \mathbb{R}$:

$$D\big(\mu^{(tf)} \| \mu\big) \le \mathbb{E}_{\mu}^{(tf)} \left[\psi\big(- t(f - c)\big)\right]. \tag{3.137}$$

Proof. Recall that

$$\begin{aligned} D\big(\mu^{(tf)} \| \mu\big) &= t\mathbb{E}_{\mu}^{(tf)} [f] + \ln \frac{1}{\mathbb{E}_{\mu}[\exp(tf)]} \\ &= t\mathbb{E}_{\mu}^{(tf)}[f] - tc + \ln \frac{\exp(tc)}{\mathbb{E}_{\mu}[\exp(tf)]} \end{aligned}$$

Using this together with the inequality $\ln u \le u - 1$ for every $u > 0$, we can write

$$\begin{aligned} D\big(\mu^{(tf)} \| \mu\big) &\le t\mathbb{E}_{\mu}^{(tf)}[f] - tc + \frac{\exp(tc)}{\mathbb{E}_{\mu}[\exp(tf)]} - 1 \\ &= t\mathbb{E}_{\mu}^{(tf)}[f] - tc + \mathbb{E}_{\mu}\left[\frac{\exp(t(f+c))\exp(-tf)}{\mathbb{E}_{\mu}[\exp(tf)]}\right] - 1 \\ &= t\mathbb{E}_{\mu}^{(tf)}[f] + \exp(tc)\,\mathbb{E}_{\mu}^{(tf)}\left[\exp(-tf)\right] - tc - 1, \end{aligned}$$

and we get (3.137). This completes the proof of Lemma 12. □

Notice that, while (3.137) is only an upper bound on $D\big(\mu^{(tf)} \| \mu\big)$, the thermal fluctuation representation (3.106) of Lemma 9 is an *exact* expression. Lemma 12 leads to the following inequality of log-Sobolev type:

Theorem 29. Let $X_1, \ldots, X_n$ be n independent random variables taking values in a set $\mathcal{X}$, and let $U = f(X^n)$ for a function $f : \mathcal{X}^n \to \mathbb{R}$. Let $P = P_{X^n} = P_{X_1} \otimes \ldots \otimes P_{X_n}$ be the product probability distribution of X^n. Also, let $X_1', \ldots, X_n'$ be independent copies of the X_i's, and define for each $i \in \{1, \ldots, n\}$

$$U^{(i)} \triangleq f(X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n).$$

Then,

$$D\big(P^{(tf)} \| P\big) \le \exp\big(- \Lambda(t)\big) \sum_{i=1}^n \mathbb{E}\left[\exp(tU)\, \psi\left(-t(U - U^{(i)})\right)\right], \tag{3.138}$$

where $\psi(u) \triangleq \exp(u) - u - 1$ for $u \in \mathbb{R}$, $\Lambda(t) \triangleq \ln \mathbb{E}[\exp(tU)]$ is the logarithmic moment-generating function, and the expectation on the right-hand side is w.r.t. X^n and $(X')^n$. Moreover, if we define the function $\tau : \mathbb{R} \to \mathbb{R}$ by $\tau(u) = u\big(\exp(u) - 1\big)$, then

$$D\big(P^{(tf)} \| P\big) \le \exp\big(- \Lambda(t)\big) \sum_{i=1}^n \mathbb{E}\left[\exp(tU)\, \tau\left(-t(U - U^{(i)})\right) 1_{\{U > U^{(i)}\}}\right] \tag{3.139}$$

and

$$D\big(P^{(tf)} \| P\big) \le \exp\big(- \Lambda(t)\big) \sum_{i=1}^n \mathbb{E}\left[\exp(tU)\, \tau\left(-t(U - U^{(i)})\right) 1_{\{U < U^{(i)}\}}\right]. \tag{3.140}$$

Proof. Applying Proposition 5 to P and $Q = P^{(tf)}$, we have

$$\begin{aligned} D\big(P^{(tf)}\big\|P\big) &\le \sum_{i=1}^n D\Big(P^{(tf)}_{X_i|\bar{X}^i}\Big\|P_{X_i}\Big|P^{(tf)}_{\bar{X}^i}\Big) \\ &= \sum_{i=1}^n \int P^{(tf)}_{\bar{X}^i}(\mathrm{d}\bar{x}^i)\, D\Big(P^{(tf)}_{X_i|\bar{X}^i=\bar{x}^i}\Big\|P_{X_i}\Big) \\ &= \sum_{i=1}^n \int P^{(tf)}_{\bar{X}^i}(\mathrm{d}\bar{x}^i)\, D\Big(P^{(tf_i(\cdot|\bar{x}^i))}_{X_i}\Big\|P_{X_i}\Big). \end{aligned} \tag{3.141}$$

Fix some $i \in \{1, \dots, n\}$, $\bar{x}^i \in \mathcal{X}^{n-1}$, and $x'_i \in \mathcal{X}$. Let us apply Lemma 12 to the ith term of the summation in (3.141) with $\mu = P_{X_i}$, $f = f_i(\cdot|\bar{x}^i)$, and $c = f(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_n) = f_i(x'_i|\bar{x}^i)$ to get

$$D\Big(P^{(tf_i(\cdot|\bar{x}^i))}_{X_i}\Big\|P_{X_i}\Big) \le \mathbb{E}^{(tf_i(\cdot|\bar{x}^i))}_{P_{X_i}}\Big[\psi\Big(-t\big(f_i(X_i|\bar{x}^i) - f_i(x'_i|\bar{x}^i)\big)\Big)\Big].$$

Substituting this into (3.141), and then taking expectation w.r.t. both X^n and $(X')^n$, we get (3.138).

To prove (3.139) and (3.140), let us write (note that $\psi(0) = 0$)

$$\begin{aligned} &\exp(tU)\,\psi\left(-t(U - U^{(i)})\right) \\ &\quad = \exp(tU)\,\psi\left(-t(U - U^{(i)})\right)1_{\{U>U^{(i)}\}} + \exp(tU)\,\psi\left(t(U^{(i)} - U)\right)1_{\{U<U^{(i)}\}}. \end{aligned} \tag{3.142}$$

Since X_i and X'_i are i.i.d. and independent of $\bar{X}^i$, we have (due to a symmetry consideration which follows from the identical distribution of U and $U^{(i)}$)

$$\begin{aligned} &\mathbb{E}\Big[\exp(tU)\,\psi\left(t(U^{(i)} - U)\right)1_{\{U<U^{(i)}\}}\Big|\bar{X}^i\Big] \\ &\quad = \mathbb{E}\Big[\exp(tU^{(i)})\,\psi\left(t(U - U^{(i)})\right)1_{\{U>U^{(i)}\}}\Big|\bar{X}^i\Big] \\ &\quad = \mathbb{E}\Big[\exp(tU)\exp\left(t(U^{(i)} - U)\right)\psi\left(t(U - U^{(i)})\right)1_{\{U>U^{(i)}\}}\Big|\bar{X}^i\Big]. \end{aligned}$$

Using this and (3.142), we can write

$$\begin{aligned} &\mathbb{E}\left[\exp(tU)\psi\left(-t(U - U^{(i)})\right)\right] \\ &\quad = \mathbb{E}\Big\{\exp(tU)\left[\psi\left(-t(U - U^{(i)})\right) + \exp\left(t(U^{(i)} - U)\right)\psi\left(t(U - U^{(i)})\right)\right]1_{\{U>U^{(i)}\}}\Big\}. \end{aligned}$$

Using the equality $\psi(u) + \exp(u)\psi(-u) = \tau(u)$ for every $u \in \mathbb{R}$, we get (3.139). The proof of (3.140) is similar. □

Now suppose that f satisfies the bounded difference condition in (3.134). Using this together with the fact that $\tau(-u) = u\big(1 - \exp(-u)\big) \le u^2$ for every $u > 0$, then for every $t > 0$ we can write

$$\begin{aligned} D\big(P^{(tf)}\big\|P\big) &\le \exp\big(-\Lambda(t)\big)\sum_{i=1}^n \mathbb{E}\left[\exp(tU)\,\tau\big(-t(U - U^{(i)})\big)\,1_{\{U>U^{(i)}\}}\right] \\ &\le t^2\exp\big(-\Lambda(t)\big)\sum_{i=1}^n \mathbb{E}\left[\exp(tU)\,\big(U - U^{(i)}\big)^2\,1_{\{U>U^{(i)}\}}\right] \\ &\le t^2\exp\big(-\Lambda(t)\big)\sum_{i=1}^n c_i^2\,\mathbb{E}\left[\exp(tU)\,1_{\{U>U^{(i)}\}}\right] \end{aligned}$$

$$\begin{aligned}
&\leq t^2 \exp\big(-\Lambda(t)\big) \left(\sum_{i=1}^n c_i^2\right) \mathbb{E}\left[\exp(tU)\right] \\
&= \left(\sum_{i=1}^n c_i^2\right) t^2.
\end{aligned}$$

Applying Corollary 7, we get

$$\mathbb{P}\Big(f(X^n) \geq \mathbb{E}f(X^n) + r\Big) \leq \exp\left(-\frac{r^2}{4\sum_{i=1}^n c_i^2}\right), \quad \forall\, r > 0$$

which has the same dependence on r and the c_i's as McDiarmid's inequality (3.135), but has a worse constant in the exponent by a factor of 8.

3.3.5 Log-Sobolev inequalities for Poission and compound Poisson measures

Let P_λ denote the Poisson(λ) measure. Bobkov and Ledoux [45] have established the following log-Sobolev inequality: for any function $f : \mathbb{Z}_+ \to \mathbb{R}$,

$$D\Big(\mathsf{P}_\lambda^{(f)} \| \mathsf{P}_\lambda\Big) \leq \lambda\, \mathbb{E}_{\mathsf{P}_\lambda}^{(f)} \Big[(\Gamma f)\, e^{\Gamma f} - e^{\Gamma f} + 1\Big], \tag{3.143}$$

where Γ is the modulus of the discrete gradient:

$$\Gamma f(x) \triangleq |f(x) - f(x+1)|, \qquad \forall x \in \mathbb{Z}_+. \tag{3.144}$$

Using tensorization of (3.143), Kontoyiannis and Madiman [128] gave a simple proof of a log-Sobolev inequality for a *compound Poisson distribution*. We recall that a compound Poisson distribution is defined as follows: given $\lambda > 0$ and a probability measure μ on $\mathbb{N}$, the compound Poisson distribution $\mathsf{CP}_{\lambda,\mu}$ is the distribution of the random sum $Z = \sum_{i=1}^N Y_i$, where $N \sim \mathsf{P}_\lambda$ and $Y_1, Y_2, \ldots$ are i.i.d. random variables with distribution μ, independent of N.

Theorem 30 (Log-Sobolev inequality for compound Poisson measures [128])**.** For any $\lambda > 0$, any probability measure μ on $\mathbb{N}$, and any bounded function $f : \mathbb{Z}_+ \to \mathbb{R}$,

$$D\Big(\mathsf{CP}_{\lambda,\mu}^{(f)} \| \mathsf{CP}_{\lambda,\mu}\Big) \leq \lambda \sum_{k=1}^\infty \mu(k)\, \mathbb{E}_{\mathsf{CP}_{\lambda,\mu}}^{(f)} \Big[(\Gamma_k f)\, e^{\Gamma_k f} - \Gamma_k f + 1\Big], \tag{3.145}$$

where $\Gamma_k f(x) \triangleq |f(x) - f(x+k)|$ for each $k, x \in \mathbb{Z}_+$.

Proof. The proof relies on the following alternative representation of the $\mathsf{CP}_{\lambda,\mu}$ probability measure: if $Z \sim \mathsf{CP}_{\lambda,\mu}$, then

$$Z \overset{\mathrm{d}}{=} \sum_{k=1}^\infty k Y_k, \qquad Y_k \sim \mathsf{P}_{\lambda\mu(k)},\ k \in \mathbb{Z}_+ \tag{3.146}$$

where $\{Y_k\}_{k=1}^\infty$ are independent random variables (this equivalence can be verified by showing, e.g., that these two representations yield the same characteristic function). For each n, let P_n denote the product distribution of $Y_1, \ldots, Y_n$. Consider a function f from the statement of Theorem 30, and define the function $g : \mathbb{Z}_+^n \to \mathbb{R}$ by

$$g(y_1, \ldots, y_n) \triangleq f\left(\sum_{k=1}^n k y_k\right), \qquad \forall y_1, \ldots, y_n \in \mathbb{Z}_+.$$

If we now denote by $\bar{P}_n$ the distribution of the sum $S_n = \sum_{k=1}^n kY_k$, then

$$\begin{aligned} D\Big(\bar{P}_n^{(f)} \big\| \bar{P}_n\Big) &= \mathbb{E}_{\bar{P}_n}\left[\frac{\exp\big(f(S_n)\big)}{\mathbb{E}_{\bar{P}_n}[\exp\big(f(S_n)\big)]} \ln \frac{\exp\big(f(S_n)\big)}{\mathbb{E}_{\bar{P}_n}[\exp\big(f(S_n)\big)]}\right] \\ &= \mathbb{E}_{P_n}\left[\frac{\exp\big(g(Y^n)\big)}{\mathbb{E}_{P_n}[\exp\big(g(Y^n)\big)]} \ln \frac{\exp\big(g(Y^n)\big)}{\mathbb{E}_{P_n}[\exp\big(g(Y^n)\big)]}\right] \\ &= D\big(P_n^{(g)} \big\| P_n\big) \\ &\le \sum_{k=1}^n D\Big(P^{(g)}_{Y_k|\bar{Y}^k} \Big\| P_{Y_k} \Big| P^{(g)}_{\bar{Y}^k}\Big), \end{aligned} \tag{3.147}$$

where the last line uses Proposition 5 and the fact that P_n is a product distribution. Using the fact that

$$\frac{\mathrm{d}P^{(g)}_{Y_k|\bar{Y}^k=\bar{y}^k}}{\mathrm{d}P_{Y_k}} = \frac{\exp\big(g_k(\cdot|\bar{y}^k)\big)}{\mathbb{E}_{\mathsf{P}_{\lambda\mu(k)}}[\exp\big(g_k(Y_k|\bar{y}^k)\big)]}, \qquad P_{Y_k} = \mathsf{P}_{\lambda\mu(k)}$$

and applying the Bobkov–Ledoux inequality in (3.143) to P_{Y_k} and all functions of the form $g_k(\cdot|\bar{y}^k)$, we can write

$$D\Big(P^{(g)}_{Y_k|\bar{Y}^k} \big\| P_{Y_k} \,\big|\, P^{(g)}_{\bar{Y}^k}\Big) \le \lambda\mu(k)\, \mathbb{E}^{(g)}_{P_n}\Big[\big(\Gamma g_k(Y_k|\bar{Y}^k)\big)\, e^{\Gamma g_k(Y_k|\bar{Y}^k)} - e^{\Gamma g_k(Y_k|\bar{Y}^k)} + 1\Big] \tag{3.148}$$

where Γ is the absolute value of the "one-dimensional" discrete gradient in (3.144). Now, for any $y^n \in \mathbb{Z}_+^n$, we have

$$\begin{aligned} \Gamma g_k(y_k|\bar{y}^k) &= \Big| g_k(y_k|\bar{y}^k) - g_k(y_k+1|\bar{y}^k) \Big| \\ &= \left| f\left(ky_k + \sum_{j\in\{1,\dots,n\}\setminus\{k\}} jy_j\right) - f\left(k(y_k+1) + \sum_{j\in\{1,\dots,n\}\setminus\{k\}} jy_j\right)\right| \\ &= \left| f\left(\sum_{j=1}^n jy_j\right) - f\left(\sum_{j=1}^n jy_j + k\right)\right| \\ &= \Gamma_k f\left(\sum_{j=1}^n jy_j\right). \end{aligned}$$

Using this in (3.148) and performing the reverse change of measure from P_n to $\bar{P}_n$, we can write

$$D\Big(P^{(g)}_{Y_k|\bar{Y}^k} \big\| P_{Y_k} \,\big|\, P^{(g)}_{\bar{Y}^k}\Big) \le \lambda\mu(k)\, \mathbb{E}^{(f)}_{\bar{P}_n}\Big[\big(\Gamma_k f(S_n)\big) e^{\Gamma_k f(S_n)} - e^{\Gamma_k f(S_n)} + 1\Big]. \tag{3.149}$$

Therefore, the combination of (3.147) and (3.149) gives

$$\begin{aligned} D\big(\bar{P}_n^{(f)} \big\| \bar{P}_n\big) &\le \lambda \sum_{k=1}^n \mu(k)\, \mathbb{E}^{(f)}_{\bar{P}_n}\Big[(\Gamma_k f)\, e^{\Gamma_k f} - e^{\Gamma_k f} + 1\Big] \\ &\le \lambda \sum_{k=1}^\infty \mu(k)\, \mathbb{E}^{(f)}_{\bar{P}_n}\Big[(\Gamma_k f)\, e^{\Gamma_k f} - e^{\Gamma_k f} + 1\Big] \end{aligned} \tag{3.150}$$

where the second line follows from the inequality $xe^x - e^x + 1 \ge 0$ that holds for all $x \ge 0$.

Now we will take the limit as $n \to \infty$ of both sides of (3.150). For the left-hand side, we use the fact that, by (3.146), $\bar{P}_n$ converges weakly (or in distribution) to $\mathsf{CP}_{\lambda,\mu}$ as $n \to \infty$. Since f is bounded, $\bar{P}_n^{(f)} \to \mathsf{CP}_{\lambda,\mu}^{(f)}$ in distribution. Therefore, by the bounded convergence theorem we have

$$\lim_{n\to\infty} D\big(\bar{P}_n^{(f)} \| \bar{P}_n\big) = D\Big(\mathsf{CP}_{\lambda,\mu}^{(f)} \,\|\, \mathsf{CP}_{\lambda,\mu}\Big). \tag{3.151}$$

For the right-hand side, we have

$$\begin{aligned}\sum_{k=1}^{\infty} \mu(k)\, \mathbb{E}_{\bar{P}_n}^{(f)} \Big[(\Gamma_k f)\, e^{\Gamma_k f} - e^{\Gamma_k f} + 1\Big] &= \mathbb{E}_{\bar{P}_n}^{(f)} \left\{ \sum_{k=1}^{\infty} \mu(k) \Big[(\Gamma_k f)\, e^{\Gamma_k f} - e^{\Gamma_k f} + 1\Big] \right\} \\ &\xrightarrow{n\to\infty} \mathbb{E}_{\mathsf{CP}_{\lambda,\mu}}^{(f)} \left[\sum_{k=1}^{\infty} \mu(k) \Big((\Gamma_k f)\, e^{\Gamma_k f} - e^{\Gamma_k f} + 1\Big) \right] \\ &= \sum_{k=1}^{\infty} \mu(k)\, \mathbb{E}_{\mathsf{CP}_{\lambda,\mu}}^{(f)} \Big[(\Gamma_k f)\, e^{\Gamma_k f} - e^{\Gamma_k f} + 1\Big] \end{aligned} \tag{3.152}$$

where the first and the last steps follow from Fubini's theorem, and the second step follows from the bounded convergence theorem. Putting (3.150)–(3.152) together, we get the inequality in (3.145). This completes the proof of Theorem 30. □

3.3.6 Bounds on the variance: Efron–Stein–Steele and Poincaré inequalities

As we have seen, tight bounds on the *variance* of a function $f(X^n)$ of independent random variables $X_1, \ldots, X_n$ are key to obtaining tight bounds on the deviation probabilities $\mathbb{P}\big(f(X^n) \ge \mathbb{E}f(X^n) + r\big)$ for $r \ge 0$. It turns out that the reverse is also true: assuming that f has Gaussian-like concentration behavior,

$$\mathbb{P}\big(f(X^n) \ge \mathbb{E}f(X^n) + r\big) \le K \exp\big(-\kappa r^2\big), \qquad \forall r \ge 0$$

it is possible to derive tight bounds on the variance of $f(X^n)$.

We start by deriving a version of a well-known inequality due to Efron and Stein [129], with subsequent refinements by Steele [130].

In the following, we say that a function f is "sufficiently regular" if the functions tf are exponentially integrable for all sufficiently small $t > 0$.

Theorem 31. Let $X_1, \ldots, X_n$ be independent $\mathcal{X}$-valued random variables. Then, for any sufficiently regular $f : \mathcal{X}^n \to \mathbb{R}$ we have

$$\mathsf{var}[f(X^n)] \le \sum_{i=1}^{n} \mathbb{E}\left\{\mathsf{var}\big[f(X^n) \big| \bar{X}^i\big]\right\} \tag{3.153}$$

Proof. By Proposition 5, for any $t > 0$, we have

$$D\big(P^{(tf)} \| P\big) \le \sum_{i=1}^{n} D\big(P_{X_i|\bar{X}^i}^{(tf)} \| P_{X_i} | P_{\bar{X}^i}\big).$$

Using Lemma 9, we can rewrite this inequality as

$$\int_0^t \int_s^t \mathsf{var}^{(\tau f)}[f]\, \mathrm{d}\tau\, \mathrm{d}s \le \sum_{i=1}^{n} \mathbb{E}\left[\int_0^t \int_s^t \mathsf{var}^{(\tau f_i(\cdot|\bar{X}^i))}[f_i(X_i|\bar{X}^i)]\, \mathrm{d}\tau\, \mathrm{d}s\right]$$

Dividing both sides by t^2, passing to the limit of $t \to 0$, and using the fact that

$$\lim_{t\to 0} \frac{1}{t^2} \int_0^t \int_s^t \mathsf{var}^{(\tau f)}[f] \, \mathrm{d}\tau \, \mathrm{d}s = \frac{\mathsf{var}[f]}{2},$$

we get (3.153). □

Next, we discuss the connection between log-Sobolev inequalities and another class of functional inequalities: the *Poincaré inequalities*. Consider, as before, a probability space $(\Omega, \mathcal{F}, \mu)$ and a pair $(\mathcal{A}, \Gamma)$ satisfying the conditions (LSI-1)–(LSI-3). Then we say that μ satisfies a *Poincaré inequality* with constant $c \geq 0$ if

$$\mathsf{var}_\mu[f] \leq c \, \mathbb{E}_\mu \left[|\Gamma f|^2 \right], \qquad \forall f \in \mathcal{A}. \tag{3.154}$$

Theorem 32. Suppose that μ satisfies LSI(c) w.r.t. $(\mathcal{A}, \Gamma)$. Then μ also satisfies a Poincaré inequality with constant c.

Proof. For any $f \in \mathcal{A}$ and any $t > 0$, we can use Lemma 9 to express the corresponding LSI(c) for the function tf as

$$\int_0^t \int_s^t \mathsf{var}_\mu^{(\tau f)}[f] \, \mathrm{d}\tau \, \mathrm{d}s \leq \frac{ct^2}{2} \cdot \mathbb{E}_\mu^{(tf)} \left[(\Gamma f)^2 \right]. \tag{3.155}$$

Proceeding exactly as in the proof of Theorem 31 above (i.e., by dividing both sides of the above inequality by t^2 and taking the limit where $t \to 0$), we obtain

$$\frac{1}{2} \mathsf{var}_\mu[f] \leq \frac{c}{2} \cdot \mathbb{E}_\mu \left[(\Gamma f)^2 \right].$$

Multiplying both sides by 2, we see that μ indeed satisfies (3.154). □

Moreover, Poincaré inequalities tensorize, as the following analogue of Theorem 24 shows:

Theorem 33. Let $X_1, \ldots, X_n \in \mathcal{X}$ be n independent random variables, and let $P = P_{X_1} \otimes \ldots P_{X_n}$ be their joint distribution. Let $\mathcal{A}$ consist of all functions $f : \mathcal{X}^n \to \mathbb{R}$, such that, for every i,

$$f_i(\cdot|\bar{x}^i) \in \mathcal{A}_i, \qquad \forall \bar{x}^i \in \mathcal{X}^{n-1} \tag{3.156}$$

Define the operator Γ that maps each $f \in \mathcal{A}$ to

$$\Gamma f = \sqrt{\sum_{i=1}^n (\Gamma_i f_i)^2}, \tag{3.157}$$

which is shorthand for

$$\Gamma f(x^n) = \sqrt{\sum_{i=1}^n \big(\Gamma_i f_i(x_i|\bar{x}^i)\big)^2}, \qquad \forall x^n \in \mathcal{X}^n. \tag{3.158}$$

Suppose that, for every $i \in \{1, \ldots, n\}$, P_{X_i} satisfies a Poincare inequality with constant c with respect to $(\mathcal{A}_i, \Gamma_i)$. Then P satisfies a Poincare inequality with constant c with respect to $(\mathcal{A}, \Gamma)$.

Proof. The proof is conceptually similar to the proof of Theorem 24 (which refers to the tensorization of the logarithmic Sobolev inequality), except that now we use the Efron–Stein–Steele inequality of Theorem 31 to tensorize the variance of f. □

3.4 Transportation-cost inequalities

So far, we have been discussing concentration of measure through the lens of various *functional* inequalities, primarily log-Sobolev inequalities. In a nutshell, if we are interested in the concentration properties of a given function $f(X^n)$ of a random n-tuple $X^n \in \mathcal{X}^n$, we seek to control the divergence $D(P^{(f)} \| P)$, where P is the distribution of X^n and $P^{(f)}$ is its f-tilting, $\mathrm{d}P^{(f)}/\mathrm{d}P \propto \exp(f)$, by some quantity related to the sensitivity of f to modifications of its arguments (e.g., the squared norm of the gradient of f, as in the Gaussian log-Sobolev inequality of Gross [35]). The common theme underlying these functional inequalities is that any such measure of sensitivity is tied to a particular *metric structure* on the underlying product space $\mathcal{X}^n$. To see this, suppose that $\mathcal{X}^n$ is equipped with some metric $d(\cdot,\cdot)$, and consider the following generalized definition of the modulus of the gradient of any function $f : \mathcal{X}^n \to \mathbb{R}$:

$$|\nabla f|(x^n) \triangleq \limsup_{y^n : d(x^n,y^n) \downarrow 0} \frac{|f(x^n) - f(y^n)|}{d(x^n,y^n)}. \tag{3.159}$$

If we also define the Lipschitz constant of f by

$$\|f\|_{\mathrm{Lip}} \triangleq \sup_{x^n \neq y^n} \frac{|f(x^n) - f(y^n)|}{d(x^n,y^n)}$$

and consider the class $\mathcal{A}$ of all functions f with $\|f\|_{\mathrm{Lip}} < \infty$, then it is easy to see that the pair $(\mathcal{A}, \Gamma)$ with $\Gamma f(x^n) \triangleq |\nabla f|(x^n)$ satisfies the conditions (LSI-1)–(LSI-3) listed in Section 3.3. Consequently, if a given probability distribution P for a random n-tuple $X^n \in \mathcal{X}^n$ satisfies LSI(c) w.r.t. the pair $(\mathcal{A}, \Gamma)$, we can use the Herbst argument to obtain the concentration inequality

$$\mathbb{P}\Big(f(X^n) \geq \mathbb{E}f(X^n) + r\Big) \leq \exp\left(-\frac{r^2}{2c\|f\|_{\mathrm{Lip}}^2}\right), \quad \forall\, r \geq 0. \tag{3.160}$$

All the examples of concentration we have discussed so far can be seen to fit this theme. Consider, for instance, the following cases:

1. **Euclidean metric:** for $\mathcal{X} = \mathbb{R}$, equip the product space $\mathcal{X}^n = \mathbb{R}^n$ with the ordinary Euclidean metric:

$$d(x^n, y^n) = \|x^n - y^n\| = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}.$$

Then the Lipschitz constant $\|f\|_{\mathrm{Lip}}$ of any function $f : \mathcal{X}^n \to \mathbb{R}$ is given by

$$\|f\|_{\mathrm{Lip}} \triangleq \sup_{x^n \neq y^n} \frac{|f(x^n) - f(y^n)|}{d(x^n,y^n)} = \sup_{x^n \neq y^n} \frac{|f(x^n) - f(y^n)|}{\|x^n - y^n\|}, \tag{3.161}$$

and for any probability measure P on $\mathbb{R}^n$ that satisfies LSI(c) we have the bound (3.160). We have already seen in (3.44) a particular instance of this with $P = G^n$, which satisfies LSI(1).

2. **Weighted Hamming metric:** for any n constants $c_1, \ldots, c_n > 0$ and any measurable space $\mathcal{X}$, let us equip the product space $\mathcal{X}^n$ with the metric

$$d_{c^n}(x^n, y^n) \triangleq \sum_{i=1}^n c_i 1_{\{x_i \neq y_i\}}.$$

The corresponding Lipschitz constant $\|f\|_{\mathrm{Lip}}$, which we also denote by $\|f\|_{\mathrm{Lip},\, c^n}$ to emphasize the role of the weights $\{c_i\}_{i=1}^n$, is given by

$$\|f\|_{\mathrm{Lip},c^n} \triangleq \sup_{x^n \neq y^n} \frac{|f(x^n) - f(y^n)|}{d_{c^n}(x^n,y^n)}.$$

Then it is easy to see that the condition $\|f\|_{\mathrm{Lip},\, c^n} \le 1$ is equivalent to (3.134). As we have shown in Section 3.3.4, *any* product probability measure P on $\mathcal{X}^n$ equipped with the metric d_{c^n} satisfies LSI(1/4) w.r.t.

$$\mathcal{A} = \Big\{ f : \|f\|_{\mathrm{Lip},\, c^n} < \infty \Big\}$$

and $\Gamma f(\cdot) = |\nabla f|(\cdot)$ with $|\nabla f|$ given by (3.159) with $d = d_{c^n}$. In this case, the concentration inequality (3.160) (with $c = 1/4$) is precisely McDiarmid's inequality (3.135).

The above two examples suggest that the metric structure plays the primary role, while the functional concentration inequalities like (3.160) are simply a consequence. In this section, we describe an alternative approach to concentration that works directly on the level of *probability measures*, rather than functions, and that makes this intuition precise. The key tool underlying this approach is the notion of *transportation cost*, which can be used to define a metric on probability distributions over the space of interest in terms of a given base metric on this space. This metric on distributions is then related to the divergence via so-called *transporation cost inequalities*. The pioneering work by K. Marton in [62] and [49] has shown that one can use these inequalities to deduce concentration.

3.4.1 Concentration and isoperimetry

We start by giving rigorous meaning to the notion that the concentration of measure phenomenon is fundamentally geometric in nature. In order to talk about concentration, we need the notion of a *metric probability space* in the sense of M. Gromov [131]. Specifically, we say that a triple $(\mathcal{X}, d, \mu)$ is a metric probability space if $(\mathcal{X}, d)$ is a Polish space (i.e., a complete and separable metric space) and μ is a probability measure on the Borel sets of $(\mathcal{X}, d)$.

For any set $A \subseteq \mathcal{X}$ and any $r > 0$, define the *r-blowup of A* by

$$A_r \triangleq \{x \in \mathcal{X} : d(x, A) < r\}, \tag{3.162}$$

where $d(x, A) \triangleq \inf_{y \in A} d(x, y)$ is the distance from the point x to the set A. We then say that the probability measure μ has *normal* (or *Gaussian*) *concentration* on $(\mathcal{X}, d)$ if there exist some constants $K, \kappa > 0$, such that

$$\mu(A) \ge 1/2 \qquad \Longrightarrow \qquad \mu(A_r) \ge 1 - K e^{-\kappa r^2}, \ \forall\, r > 0. \tag{3.163}$$

Remark 33. Of the two constants K and κ in (3.163), it is κ that is more important. For that reason, sometimes we will say that μ has normal concentration with constant $\kappa > 0$ to mean that (3.163) holds with that value of κ and some $K > 0$.

Here are a few standard examples (see [2, Section 1.1]):

1. **Standard Gaussian distribution** — if $\mathcal{X} = \mathbb{R}^n$, $d(x, y) = \|x - y\|$ is the standard Euclidean metric, and $\mu = G^n$, the standard Gaussian distribution, then for any Borel set $A \subseteq \mathbb{R}^n$ with $G^n(A) \ge 1/2$ we have

$$\begin{aligned} G^n(A_r) &\ge \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{r} \exp\left(-\frac{t^2}{2}\right) dt \\ &\ge 1 - \frac{1}{2} \exp\left(-\frac{r^2}{2}\right), \qquad \forall\, r \ge 0 \end{aligned} \tag{3.164}$$

i.e., (3.163) holds with $K = \frac{1}{2}$ and $\kappa = \frac{1}{2}$.

2. **Uniform distribution on the unit sphere** — if $\mathcal{X} = \mathbb{S}^n \equiv \big\{x \in \mathbb{R}^{n+1} : \|x\| = 1\big\}$, d is given by the geodesic distance on $\mathbb{S}^n$, and $\mu = \sigma^n$ (the uniform distribution on $\mathbb{S}^n$), then for any Borel set $A \subseteq \mathbb{S}^n$ with $\sigma^n(A) \ge 1/2$ we have

$$\sigma^n(A_r) \ge 1 - \exp\left(-\frac{(n-1)r^2}{2}\right), \qquad \forall\, r \ge 0. \tag{3.165}$$

   In this instance, (3.163) holds with $K = 1$ and $\kappa = (n-1)/2$. Notice that κ is actually increasing with the ambient dimension n.

3. **Uniform distribution on the Hamming cube** — if $\mathcal{X} = \{0,1\}^n$, d is the normalized Hamming metric

$$d(x,y) = \frac{1}{n}\sum_{i=1}^n 1_{\{x_i \neq y_i\}}$$

   for all $x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n) \in \{0,1\}^n$, and $\mu = B^n$ is the uniform distribution on $\{0,1\}^n$ (which is equal to the product of n copies of a Bernoulli(1/2) measure on $\{0,1\}$), then for any $A \subseteq \{0,1\}^n$ we have

$$B^n(A_r) \ge 1 - \exp\left(-2nr^2\right), \qquad \forall\, r \ge 0 \tag{3.166}$$

   so (3.163) holds with $K = 1$ and $\kappa = 2n$.

Remark 34. Gaussian concentration of the form (3.163) is often discussed in the context of the so-called *isoperimetric inequalities*, which relate the full measure of a set to the measure of its boundary. To be more specific, consider a metric probability space $(\mathcal{X}, d, \mu)$, and for any Borel set $A \subseteq \mathcal{X}$ define its *surface measure* as (see [2, Section 2.1])

$$\mu^+(A) \triangleq \liminf_{r \to 0} \frac{\mu(A_r \setminus A)}{r} = \liminf_{r \to 0} \frac{\mu(A_r) - \mu(A)}{r}. \tag{3.167}$$

Then the classical Gaussian isoperimetric inequality can be stated as follows: If H is a half-space in $\mathbb{R}^n$, i.e., $H = \{x \in \mathbb{R}^n : \langle x, u\rangle < c\}$ for some $u \in \mathbb{R}^n$ with $\|u\| = 1$ and some $c \in [-\infty, +\infty]$, and if $A \subseteq \mathbb{R}^n$ is a Borel set with $G^n(A) = G^n(H)$, then

$$(G^n)^+(A) \ge (G^n)^+(H), \tag{3.168}$$

with equality if and only if A is a half-space. In other words, the Gaussian isoperimetric inequality (3.168) says that, among all Borel subsets of $\mathbb{R}^n$ with a given Gaussian volume, the half-spaces have the smallest surface measure. An equivalent integrated version of (3.168) says the following (see, e.g., [132]): Consider a Borel set A in $\mathbb{R}^n$ and a half-space $H = \{x : \langle x, u\rangle < c\}$ with $\|u\| = 1$, $c \ge 0$ and $G^n(A) = G^n(H)$. Then for any $r \ge 0$ we have

$$G^n(A_r) \ge G^n(H_r),$$

with equality if and only if A is itself a half-space. Moreover, an easy calculation shows that

$$G^n(H_r) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{c+r} \exp\left(-\frac{\xi^2}{2}\right) \mathrm{d}\xi \ge 1 - \frac{1}{2}\exp\left(-\frac{(r+c)^2}{2}\right), \quad \forall\, r \ge 0.$$

So, if $G(A) \ge 1/2$, we can always choose $c = 0$ and get (3.164).

Intuitively, what (3.163) says is that, if μ has normal concentration on $(\mathcal{X}, d)$, then most of the probability mass in $\mathcal{X}$ is concentrated around any set with probability at least 1/2. At first glance, this seems to have nothing to do with what we have been looking at all this time, namely the concentration of Lipschitz functions on $\mathcal{X}$ around their mean. However, as we will now show, the geometric and the functional pictures of the concentration of measure phenomenon are, in fact, equivalent. To that end, let us define the *median* of a function $f : \mathcal{X} \to \mathbb{R}$: we say that a real number m_f is a median of f w.r.t. μ (or a μ*-median* of f) if

$$\mathbb{P}_\mu\big(f(X) \ge m_f\big) \ge \frac{1}{2} \qquad \text{and} \qquad \mathbb{P}_\mu\big(f(X) \le m_f\big) \ge \frac{1}{2} \tag{3.169}$$

(note that a median of f may not be unique). The precise result is as follows:

Theorem 34. Let $(\mathcal{X}, d, \mu)$ be a metric probability space. Then μ has the normal concentration property (3.163) (with arbitrary constants $K, \kappa > 0$) if and only if for every Lipschitz function $f : \mathcal{X} \to \mathbb{R}$ (where the Lipschitz property is defined w.r.t. the metric d) we have

$$\mathbb{P}_\mu\Big(f(X) \ge m_f + r\Big) \le K \exp\bigg(-\frac{\kappa r^2}{\|f\|_{\mathrm{Lip}}^2}\bigg), \qquad \forall\, r \ge 0 \tag{3.170}$$

where m_f is any μ-median of f.

Proof. Suppose that μ satisfies (3.163). Fix any Lipschitz function f where, without loss of generality, we may assume that $\|f\|_{\mathrm{Lip}} = 1$, let m_f be any median of f, and define the set $A^f \triangleq \Big\{x \in \mathcal{X} : f(x) \le m_f\Big\}$. By definition of the median in (3.169), $\mu(A^f) \ge 1/2$. Consequently, by (3.163), we have

$$\mu(A^f_r) \equiv \mathbb{P}_\mu\Big(d(X, A^f) < r\Big) \ge 1 - K \exp(-\kappa r^2), \qquad \forall r \ge 0. \tag{3.171}$$

By the Lipschitz property of f, for any $y \in A^f$ we have $f(X) - m_f \le f(X) - f(y) \le d(X, y)$, so $f(X) - m_f \le d(X, A^f)$. This, together with (3.171), implies that

$$\mathbb{P}_\mu\Big(f(X) - m_f < r\Big) \ge \mathbb{P}_\mu\Big(d(X, A^f) < r\Big) \ge 1 - K \exp(-\kappa r^2), \qquad \forall\, r \ge 0$$

which is (3.170).

Conversely, suppose (3.170) holds for every Lipschitz f. Choose any Borel set A with $\mu(A) \ge 1/2$ and define the function $f_A(x) \triangleq d(x, A)$ for every $x \in \mathcal{X}$. Then f_A is 1-Lipschitz, since

$$\begin{aligned}
|f_A(x) - f_A(y)| &= \Big|\inf_{u \in A} d(x,u) - \inf_{u \in A} d(y,u)\Big| \\
&\le \sup_{u \in A} |d(x,u) - d(y,u)| \\
&\le d(x,y),
\end{aligned}$$

where the last step is by the triangle inequality. Moreover, zero is a median of f_A, since

$$\mathbb{P}_\mu\big(f_A(X) \le 0\big) = \mathbb{P}_\mu\big(X \in A\big) \ge \frac{1}{2} \qquad \text{and} \qquad \mathbb{P}_\mu\big(f_A(X) \ge 0\big) \ge \frac{1}{2},$$

where the second bound is vacuously true since $f_A \ge 0$ everywhere. Consequently, with $m_f = 0$, we get

$$\begin{aligned}
1 - \mu(A_r) &= \mathbb{P}_\mu\big(d(X, A) \ge r\big) \\
&= \mathbb{P}_\mu\big(f_A(X) \ge m_f + r\big) \\
&\le K \exp\left(-\kappa r^2\right), \qquad \forall\, r \ge 0
\end{aligned}$$

which gives (3.163). □

It is shown in the following that for Lipschitz functions, concentration around the mean also implies concentration around any median, but possibly with worse constants [2, Proposition 1.7]:

Theorem 35. Let $(\mathcal{X}, d, \mu)$ be a metric probability space, such that for any 1-Lipschitz function $f : \mathcal{X} \to \mathbb{R}$ we have

$$\mathbb{P}_\mu\Big(f(X) \geq \mathbb{E}_\mu[f(X)] + r\Big) \leq K_0 \exp\big(-\kappa_0 r^2\big), \qquad \forall\, r \geq 0 \tag{3.172}$$

with some constants $K_0, \kappa_0 > 0$. Then, μ has the normal concentration property (3.163) with $K = K_0$ and $\kappa = \kappa_0/4$. Consequently, the concentration inequality in (3.170) around any median m_f is satisfied with the same constants of κ and K.

Proof. Let $A \subseteq \mathcal{X}$ be an arbitrary Borel set with $\mu(A) > 0$, and fix some $r > 0$. Define the function $f_{A,r}(x) \triangleq \min\{d(x,A), r\}$. Then, from the triangle inequality, $\|f_{A,r}\|_{\text{Lip}} \leq 1$ and

$$\begin{aligned}
\mathbb{E}_\mu[f_{A,r}(x)] &= \int_{\mathcal{X}} \min\{d(x,A), r\}\, \mu(\mathrm{d}x) \\
&= \underbrace{\int_A \min\{d(x,A), r\}\, \mu(\mathrm{d}x)}_{=0} + \int_{A^c} \min\{d(x,A), r\}\, \mu(\mathrm{d}x) \\
&\leq r\mu(A^c) \\
&= (1 - \mu(A))r.
\end{aligned} \tag{3.173}$$

Then

$$\begin{aligned}
1 - \mu(A_r) &= \mathbb{P}_\mu\Big(d(X,A) \geq r\Big) \\
&= \mathbb{P}_\mu\Big(f_{A,r}(X) \geq r\Big) \\
&\leq \mathbb{P}_\mu\Big(f_{A,r}(X) \geq \mathbb{E}_\mu[f_{A,r}(X)] + r\mu(A)\Big) \\
&\leq K_0 \exp\Big(-\kappa\,(\mu(A)r)^2\Big), \qquad \forall r \geq 0,
\end{aligned}$$

where the first two steps use the definition of $f_{A,r}$, the third step uses (3.173), and the last step uses (3.172). Consequently, if $\mu(A) \geq 1/2$, we get (3.163) with $K = K_0$ and $\kappa = \kappa_0/4$. Consequently, from Theorem 34, also the concentration inequality in (3.170) holds for any median m_f with the same constants of κ and K. □

Remark 35. Let $(\mathcal{X}, d, \mu)$ be a metric probability space, and suppose that μ has the normal concentration property (3.163) (with arbitrary constants $K, \kappa > 0$). Let $f : \mathcal{X} \to \mathbb{R}$ be an arbitrary Lipschitz function (where the Lipschitz property is defined w.r.t. the metric d), and let $\mathbb{E}_\mu[f(X)]$ and m_f be, respectively, the mean and any median of f w.r.t. μ. Theorem 3.172 considers concentration of f around the mean and the median. In the following, we provide an upper bound on the distance between the mean and any median of f in terms of the parameters κ and K of (3.163), and the Lipschitz constant of f. From Theorem 34, it follows that

$$\begin{aligned}
&\big|\mathbb{E}_\mu[f(X)] - m_f\big| \\
&\leq \mathbb{E}_\mu\big[|f(X) - m_f|\big] \\
&= \int_0^\infty \mathbb{P}_\mu(|f(X) - m_f| \geq r)\, \mathrm{d}r \\
&\leq \int_0^\infty 2K \exp\Big(-\frac{\kappa r^2}{\|f\|_{\text{Lip}}^2}\Big)\, \mathrm{d}r \\
&= \sqrt{\frac{\pi}{\kappa}}\, K \|f\|_{\text{Lip}}
\end{aligned} \tag{3.174}$$

where the last inequality follows from the (one-sided) concentration inequality in (3.170) and since f and $-f$ are both Lipschitz functions with the same constant. This shows that the larger is κ and also the smaller is K (so that the concentration inequality in (3.163) is more pronounced), then the mean and any median of f get closer to each other, so the concentration of f around both the mean and median becomes more well expected. Indeed, Theorem 34 provides a better concentration inequality around the median when this situation takes place.

3.4.2 Marton's argument: from transportation to concentration

As shown above, the phenomenon of concentration is fundamentally geometric in nature, as captured by the isoperimetric inequality (3.163). Once we have established (3.163) on a given metric probability space $(\mathcal{X}, d, \mu)$, we immediately obtain Gaussian concentration for all Lipschitz functions $f : \mathcal{X} \to \mathbb{R}$ by Theorem 34.

There is a powerful information-theoretic technique for deriving concentration inequalities like (3.163). This technique, first introduced by Marton (see [62] and [49]), hinges on a certain type of inequality that relates the divergence between two probability measures to a quantity called the *transportation cost*. Let $(\mathcal{X}, d)$ be a Polish space. Given $p \geq 1$, let $\mathcal{P}_p(\mathcal{X})$ denote the space of all Borel probability measures μ on $\mathcal{X}$, such that the moment bound

$$\mathbb{E}_\mu[d^p(X, x_0)] < \infty \tag{3.175}$$

holds for some (and hence all) $x_0 \in \mathcal{X}$.

Definition 5. Given $p \geq 1$, the *L^p Wasserstein distance* between any pair $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$ is defined as

$$W_p(\mu, \nu) \triangleq \inf_{\pi \in \Pi(\mu,\nu)} \left(\int_{\mathcal{X} \times \mathcal{X}} d^p(x, y) \pi(\mathrm{d}x, \mathrm{d}y) \right)^{1/p}, \tag{3.176}$$

where $\Pi(\mu, \nu)$ is the set of all probability measures π on the product space $\mathcal{X} \times \mathcal{X}$ with marginals μ and ν.

Remark 36. Another equivalent way of writing down the definition of $W_p(\mu, \nu)$ is

$$W_p(\mu, \nu) = \inf_{X \sim \mu, Y \sim \nu} \left\{ \mathbb{E}[d^p(X, Y)] \right\}^{1/p}, \tag{3.177}$$

where the infimum is over all pairs (X, Y) of jointly distributed random variables with values in $\mathcal{X}$, such that $P_X = \mu$ and $P_Y = \nu$.

Remark 37. The name "transportation cost" comes from the following interpretation: Let μ (resp., ν) represent the initial (resp., desired) distribution of some matter (say, sand) in space, such that the total mass in both cases is normalized to one. Thus, both μ and ν correspond to sand piles of some given shapes. The objective is to rearrange the initial sand pile with shape μ into one with shape ν with minimum cost, where the cost of transporting a grain of sand from location x to location y is given by $c(x, y)$ for some sufficiently regular function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. If we allow randomized transportation policies, i.e., those that associate with each location x in the initial sand pile a conditional probability distribution $\pi(\mathrm{d}y|x)$ for the destination in the final sand pile, then the minimum transportation cost is given by

$$C^*(\mu, \nu) \triangleq \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{X}} c(x, y) \pi(\mathrm{d}x, \mathrm{d}y) \tag{3.178}$$

When the cost function is given by $c = d^p$ for some $p \geq 1$ and d is a metric on $\mathcal{X}$, we will have $C^*(\mu, \nu) = W_p^p(\mu, \nu)$. The optimal transportation problem (3.178) has a rich history, dating back to a 1781 essay by Gaspard Monge, who has considered a particular special case of the problem

$$C_0^*(\mu, \nu) \triangleq \inf_{\varphi : \mathcal{X} \to \mathcal{X}} \left\{ \int_{\mathcal{X}} c(x, \varphi(x)) \mu(\mathrm{d}x) : \mu \circ \varphi^{-1} = \nu \right\}. \tag{3.179}$$

Here, the infimum is over all *deterministic* transportation policies, i.e., measurable mappings $\varphi : \mathcal{X} \to \mathcal{X}$, such that the desired final measure ν is the image of μ under φ, or, in other words, if $X \sim \mu$, then $Y = \varphi(X) \sim \nu$. The problem (3.179) (or the *Monge optimal transportation problem*, as it has now come to be called) does not always admit a solution (incidentally, an optimal mapping does exist in the case considered by Monge, namely $\mathcal{X} = \mathbb{R}^3$ and $c(x, y) = \|x - y\|$). A stochastic relaxation of Monge's problem, given by (3.178), was considered in 1942 by Leonid Kantorovich (and reprinted more recently [133]). We recommend the books by Villani [50, 51] for a much more detailed historical overview and rigorous treatment of optimal transportation.

Lemma 13. The Wasserstein distances have the following properties:

1. For each $p \geq 1$, $W_p(\cdot, \cdot)$ is a metric on $\mathcal{P}_p(\mathcal{X})$.

2. If $1 \leq p \leq q$, then $\mathcal{P}_p(\mathcal{X}) \supseteq \mathcal{P}_q(\mathcal{X})$, and $W_p(\mu, \nu) \leq W_q(\mu, \nu)$ for any $\mu, \nu \in \mathcal{P}_q(\mathcal{X})$.

3. W_p metrizes weak convergence plus convergence of pth-order moments: a sequence $\{\mu_n\}_{n=1}^\infty$ in $\mathcal{P}_p(\mathcal{X})$ converges to $\mu \in \mathcal{P}_p(\mathcal{X})$ in W_p, i.e., $W_p(\mu_n, \mu) \xrightarrow{n \to \infty} 0$, if and only if:

   (a) $\{\mu_n\}$ converges to μ weakly, i.e., $\mathbb{E}_{\mu_n}[\varphi] \xrightarrow{n \to \infty} \mathbb{E}_\mu[\varphi]$ for any continuous bounded function $\varphi : \mathcal{X} \to \mathbb{R}$

   (b) for some (and hence all) $x_0 \in \mathcal{X}$,

   $$\int_{\mathcal{X}} d^p(x, x_0) \mu_n(\mathrm{d}x) \xrightarrow{n \to \infty} \int_{\mathcal{X}} d^p(x, x_0) \mu(\mathrm{d}x).$$

   If the above two statements hold, then we say that $\{\mu_n\}$ converges to μ *weakly in* $\mathcal{P}_p(\mathcal{X})$.

4. The mapping $(\mu, \nu) \mapsto W_p(\mu, \nu)$ is continuous on $\mathcal{P}_p(\mathcal{X})$, i.e., if $\mu_n \to \mu$ and $\nu_n \to \nu$ weakly in $\mathcal{P}_p(\mathcal{X})$, then $W_p(\mu_n, \nu_n) \to W_p(\mu, \nu)$. However, it is only *lower semicontinuous* in the usual weak topology (without the convergence of pth-order moments): if $\mu_n \to \mu$ and $\nu_n \to \nu$ weakly, then

   $$\liminf_{n \to \infty} W_p(\mu_n, \nu_n) \geq W_p(\mu, \nu).$$

5. The infimum in (3.176) [and therefore in (3.177)] is actually a minimum; in other words, there exists an *optimal coupling* $\pi^* \in \Pi(\mu, \nu)$, such that

   $$W_p^p(\mu, \nu) = \int_{\mathcal{X} \times \mathcal{X}} d^p(x, y) \pi^*(\mathrm{d}x, \mathrm{d}y).$$

   Equivalently, there exists a pair (X^*, Y^*) of jointly distributed $\mathcal{X}$-valued random variables with $P_{X^*} = \mu$ and $P_{Y^*} = \nu$, such that

   $$W_p^p(\mu, \nu) = \mathbb{E}[d^p(X^*, Y^*)].$$

6. If $p = 2$, $\mathcal{X} = \mathbb{R}$ with $d(x, y) = |x - y|$, and μ is atomless (i.e., if $\mu(\{x\}) = 0$ for all $x \in \mathbb{R}$), then the optimal coupling between μ and any ν is given by the deterministic mapping

   $$Y = \mathsf{F}_\nu^{-1} \circ \mathsf{F}_\mu(X)$$

   for $X \sim \mu$, where F_μ denotes the cumulative distribution (cdf) function of μ, i.e., $\mathsf{F}_\mu(x) = \mathbb{P}_\mu(X \leq x)$, and F_ν^{-1} is the *quantile function* of ν, i.e., $\mathsf{F}_\nu^{-1}(x) \triangleq \inf\{\alpha : \mathsf{F}_\nu(x) \geq \alpha\}$.

Definition 6. We say that a probability measure μ on $(\mathcal{X}, d)$ satisfies an *L^p transportation cost inequality with constant* $c > 0$, or a $T_p(c)$ inequality for short, if for any probability measure $\nu \ll \mu$ we have

$$W_p(\mu, \nu) \le \sqrt{2cD(\nu\|\mu)}. \tag{3.180}$$

Example 16 (Total variation distance and Pinsker's inequality)**.** Here is a specific example illustrating this abstract machinery, which should be a familiar territory to information theorists. Let $\mathcal{X}$ be a discrete set equipped with the Hamming metric $d(x, y) = 1_{\{x \neq y\}}$. In this case, the corresponding L^1 Wasserstein distance between any two probability measures μ and ν on $\mathcal{X}$ takes the simple form

$$W_1(\mu, \nu) = \inf_{X \sim \mu, Y \sim \nu} \mathbb{P}\left(X \neq Y\right).$$

As we will now show, this turns out to be nothing but the usual total variation distance

$$\|\mu - \nu\|_{\text{TV}} = \sup_{A \subseteq \mathcal{X}} |\mu(A) - \nu(A)| = \frac{1}{2} \sum_{x \in \mathcal{X}} |\mu(x) - \nu(x)|$$

(we are abusing the notation here, writing $\mu(x)$ for the μ-probability of the singleton $\{x\}$). To see this, consider any $\pi \in \Pi(\mu, \nu)$. Then for any x we have

$$\mu(x) = \sum_{y \in \mathcal{X}} \pi(x, y) \ge \pi(x, x),$$

and the same goes for ν. Consequently, $\pi(x, x) \le \min\left\{\mu(x), \nu(x)\right\}$, and so

$$\begin{aligned} \mathbb{E}_\pi[d(X \neq Y)] &= 1 - \sum_{x \in \mathcal{X}} \pi(x, x) && (3.181) \\ &\ge 1 - \sum_{x \in \mathcal{X}} \min\left\{\mu(x), \nu(x)\right\}. && (3.182) \end{aligned}$$

On the other hand, if we define the set $A = \{x \in \mathcal{X} : \mu(x) \ge \nu(x)\}$, then

$$\begin{aligned} \|\mu - \nu\|_{\text{TV}} &= \frac{1}{2} \sum_{x \in A} |\mu(x) - \nu(x)| + \frac{1}{2} \sum_{x \in A^c} |\mu(x) - \nu(x)| \\ &= \frac{1}{2} \sum_{x \in A} [\mu(x) - \nu(x)] + \frac{1}{2} \sum_{x \in A^c} [\nu(x) - \mu(x)] \\ &= \frac{1}{2} \Big(\mu(A) - \nu(A) + \nu(A^c) - \mu(A^c)\Big) \\ &= \mu(A) - \nu(A) \end{aligned}$$

and

$$\begin{aligned} \sum_{x \in \mathcal{X}} \min\left\{\mu(x), \nu(x)\right\} &= \sum_{x \in A} \nu(x) + \sum_{x \in A^c} \mu(x) \\ &= \nu(A) + \mu(A^c) \\ &= 1 - \Big(\mu(A) - \nu(A)\Big) \\ &= 1 - \|\mu - \nu\|_{\text{TV}}. && (3.183) \end{aligned}$$

Consequently, for any $\pi \in \Pi(\mu, \nu)$ we see from (3.181)–(3.183) that

$$\mathbb{P}_\pi\big(X \neq Y\big) = \mathbb{E}_\pi[d(X, Y)] \ge \|\mu - \nu\|_{\text{TV}}. \tag{3.184}$$

Moreover, the lower bound in (3.184) is actually achieved by π^* taking

$$\pi^*(x,y) = \min\left\{\mu(x), \nu(x)\right\} 1_{\{x=y\}} + \frac{\big(\mu(x)-\nu(x)\big)1_{\{x \in A\}}\big(\nu(y)-\mu(y)\big)1_{\{y \in A^c\}}}{\mu(A)-\nu(A)} 1_{\{x \neq y\}}. \tag{3.185}$$

Now that we have expressed the total variation distance $\|\mu - \nu\|_{\text{TV}}$ as the L^1 Wasserstein distance induced by the Hamming metric on $\mathcal{X}$, we can recognize the well-known Pinsker's inequality,

$$\|\mu - \nu\|_{\text{TV}} \le \sqrt{\frac{1}{2} D(\nu \| \mu)}, \tag{3.186}$$

as a $T_1(1/4)$ inequality that holds for any probability measure μ on $\mathcal{X}$.

Remark 38. It should be pointed out that the constant $c = 1/4$ in Pinsker's inequality (3.186) is not necessarily the best possible *for a given distribution* P. Ordentlich and Weinberger [134] have obtained the following *distribution-dependent* refinement of Pinsker's inequality. Let the function $\varphi : [0, 1/2] \to \mathbb{R}^+$ be defined by

$$\varphi(p) \triangleq \begin{cases} \left(\frac{1}{1-2p}\right) \ln\left(\frac{1-p}{p}\right), & \text{if } p \in \left[0, \frac{1}{2}\right) \\ 2, & \text{if } p = 1/2 \end{cases} \tag{3.187}$$

(in fact, $\varphi(p) \to 2$ as $p \uparrow 1/2$, $\varphi(p) \to \infty$ as $p \downarrow 0$, and φ is a monotonic decreasing and convex function). Let $\mathcal{X}$ be a discrete set. For any $P \in \mathcal{P}(\mathcal{X})$, define the *balance coefficient*

$$\pi_P \triangleq \max_{A \subseteq \mathcal{X}} \min\left\{P(A), 1 - P(A)\right\} \quad \Longrightarrow \quad \pi_P \in \left[0, \frac{1}{2}\right].$$

Then (cf. Theorem 2.1 in [134]), for any $Q \in \mathcal{P}(\mathcal{X})$,

$$\|P - Q\|_{\text{TV}} \le \sqrt{\left(\frac{1}{\varphi(\pi_P)}\right) D(Q \| P)} \tag{3.188}$$

From the above properties of the function φ, it follows that the distribution-dependent refinement of Pinsker's inequality is more pronounced when the balance coefficient is small (i.e., $\pi_P \ll 1$). Moreover, this bound is optimal for a given P, in the sense that

$$\varphi(\pi_P) = \inf_{Q \in \mathcal{P}(\mathcal{X})} \frac{D(Q \| P)}{\|P - Q\|_{\text{TV}}^2}. \tag{3.189}$$

For instance, if $\mathcal{X} = \{0, 1\}$ and P is the distribution of a Bernoulli(p) random variable, then $\pi_P = \min\{p, 1-p\}$, and (since $\varphi(p)$ in (3.187) is symmetric around one-half)

$$\varphi(\pi_P) = \begin{cases} \left(\frac{1}{1-2p}\right) \ln\left(\frac{1-p}{p}\right), & \text{if } p \neq \frac{1}{2} \\ 2, & \text{if } p = \frac{1}{2} \end{cases}$$

and for any other $Q \in \mathcal{P}(\{0,1\})$ we have, from (3.188),

$$\|P - Q\|_{\text{TV}} \le \begin{cases} \sqrt{\left(\frac{1-2p}{\ln[(1-p)/p]}\right) D(Q \| P)}, & \text{if } p \neq \frac{1}{2} \\ \sqrt{\frac{1}{2}\, D(Q \| P)}, & \text{if } p = \frac{1}{2}. \end{cases} \tag{3.190}$$

The above inequality provides an upper bound on the total variation distance in terms of the divergence. In general, a bound in the reverse direction cannot be derived since the total variation distance can be arbitrarily close to zero, whereas the divergence is equal to infinity. However, consider an i.i.d. sample of size n that is generated from a probability distribution P. Sanov's theorem implies that the probability that the empirical distribution of the generated sample deviates in total variation from P by at least some $\varepsilon \in (0, 2]$ scales asymptotically like $\exp\bigl(-n\, D^*(P, \varepsilon)\bigr)$ where

$$D^*(P, \varepsilon) \triangleq \inf_{Q \colon \|P-Q\|_{\mathrm{TV}} \geq \varepsilon} D(Q\|P).$$

Although a reverse form of Pinsker's inequality (or its probability-dependent refinement in [134]) cannot be derived, it was recently proved in [135] that

$$D^*(P, \varepsilon) \leq \varphi(\pi_P)\, \varepsilon^2 + O(\varepsilon^3).$$

This inequality shows that the probability-dependent refinement of Pinsker's inequality in (3.188) is actually tight for $D^*(P, \varepsilon)$ when ε is small, since both upper and lower bounds scale like $\varphi(\pi_P)\, \varepsilon^2$ if $\varepsilon \ll 1$.

Apart of providing a refined upper bound on the total variation distance between two discrete probability distributions, the inequality in (3.188) also enables to derive a refined lower bound on the relative entropy when a lower bound on the total variation distance is available. This approach was studied in [136, Section III] in the context of the Poisson approximation where (3.188) was combined with a new lower bound on the total variation distance (using the so-called Chen-Stein method) between the distribution of a sum of independent Bernoulli random variables and the Poisson distribution with the same mean. It is noted that for a sum of i.i.d. Bernoulli random variables, the resulting lower bound on this relative entropy (see [136, Theorem 7]) scales similarly to the upper bound on this relative entropy by Kontoyiannis et al. (see [137, Theorem 1]), where the derivation of the latter upper bound relies on the logarithmic Sobolev inequality for the Poisson distribution by Bobkov and Ledoux [45] (see Section 3.3.5 here).

Marton's procedure for deriving Gaussian concentration from a transportation cost inequality [62, 49] can be distilled in the following:

Proposition 9. Suppose μ satisfies a $T_1(c)$ inequality. Then, the Gaussian concentration inequality in (3.163) holds with $\kappa = 1/(2c)$ and $K = 1$ for all $r \geq \sqrt{2c \ln 2}$.

Proof. Fix two Borel sets $A, B \subset \mathcal{X}$ with $\mu(A), \mu(B) > 0$. Define the conditional probability measures

$$\mu_A(C) \triangleq \frac{\mu(C \cap A)}{\mu(A)} \qquad \text{and} \qquad \mu_B(C) \triangleq \frac{\mu(C \cap B)}{\mu(B)},$$

where C is an arbitrary Borel set in $\mathcal{X}$. Then $\mu_A, \mu_B \ll \mu$, and

$$W_1(\mu_A, \mu_B) \leq W_1(\mu, \mu_A) + W_1(\mu, \mu_B) \tag{3.191}$$

$$\leq \sqrt{2cD(\mu_A\|\mu)} + \sqrt{2cD(\mu_B\|\mu)}, \tag{3.192}$$

where (3.191) is by the triangle inequality, while (3.192) is because μ satisfies $T_1(c)$. Now, for any Borel set C, we have

$$\mu_A(C) = \int_C \frac{1_A(x)}{\mu(A)}\, \mu(\mathrm{d}x),$$

so it follows that $\mu_A \ll \mu$ with $\mathrm{d}\mu_A/\mathrm{d}\mu = 1_A/\mu(A)$, and the same holds for μ_B. Therefore,

$$D(\mu_A\|\mu) = \mathbb{E}_\mu\left[\frac{\mathrm{d}\mu_A}{\mathrm{d}\mu} \ln \frac{\mathrm{d}\mu_A}{\mathrm{d}\mu}\right] = \ln \frac{1}{\mu(A)}, \tag{3.193}$$

and an analogous formula holds for μ_B in place of μ_A. Substituting this into (3.192) gives

$$W_1(\mu_A, \mu_B) \le \sqrt{2c \ln \frac{1}{\mu(A)}} + \sqrt{2c \ln \frac{1}{\mu(B)}}. \tag{3.194}$$

We now obtain a lower bound on $W_1(\mu_A, \mu_B)$. Since μ_A (resp., μ_B) is supported on A (resp., B), any $\pi \in \Pi(\mu_A, \nu_A)$ is supported on $A \times B$. Consequently, for any such π we have

$$\begin{aligned} \int_{\mathcal{X}\times\mathcal{X}} d(x,y)\ \pi(\mathrm{d}x, \mathrm{d}y) &= \int_{A\times B} d(x,y)\ \pi(\mathrm{d}x, \mathrm{d}y) \\ &\ge \int_{A\times B} \inf_{y\in B} d(x,y)\ \pi(\mathrm{d}x, \mathrm{d}y) \\ &= \int_A d(x,B)\ \mu_A(\mathrm{d}x) \\ &\ge \inf_{x\in A} d(x,B)\ \mu_A(A) \\ &= d(A,B), \end{aligned} \tag{3.195}$$

where $\mu_A(A) = 1$, and $d(A,B) \triangleq \inf_{x\in A, y\in B} d(x,y)$ is the distance between A and B. Since (3.195) holds for every $\pi \in \Pi(\mu_A, \nu_A)$, we can take the infimum over all such π and get $W_1(\mu_A, \mu_B) \ge d(A,B)$. Combining this with (3.194) gives the inequality

$$d(A,B) \le \sqrt{2c \ln \frac{1}{\mu(A)}} + \sqrt{2c \ln \frac{1}{\mu(B)}},$$

which holds for all Borel sets A and B that have nonzero μ-probability.

Let $B = A_r^c$, then $\mu(B) = 1 - \mu(A_r)$ and $d(A,B) \ge r$. Consequently,

$$r \le \sqrt{2c \ln \frac{1}{\mu(A)}} + \sqrt{2c \ln \frac{1}{1-\mu(A_r)}}. \tag{3.196}$$

If $\mu(A) \ge 1/2$ and $r \ge \sqrt{2c\ln 2}$, then (3.196) gives

$$\mu(A_r) \ge 1 - \exp\left(-\frac{1}{2c}\left(r - \sqrt{2c\ln 2}\right)^2\right). \tag{3.197}$$

Hence, the Gaussian concentration inequality in (3.163) indeed holds with $\kappa = 1/(2c)$ and $K = 1$ for all $r \ge \sqrt{2c\ln 2}$. □

Remark 39. The formula (3.193), apparently first used explicitly by Csiszár [138, Eq. (4.13)], is actually quite remarkable: it states that the probability of any event can be expressed as an exponential of a divergence.

While the method described in the proof of Proposition 9 does not produce optimal concentration estimates (which typically have to be derived on a case-by-case basis), it hints at the potential power of the transportation cost inequalities. To make full use of this power, we first establish an important fact that, for $p \in [1,2]$, the T_p inequalities tensorize (see, for example, [51, Proposition 22.5]):

Proposition 10 (Tensorization of transportation cost inequalities)**.** For any $p \in [1,2]$, the following statement is true: If μ satisfies $T_p(c)$ on $(\mathcal{X}, d)$, then, for any $n \in \mathbb{N}$, the product measure $\mu^{\otimes n}$ satisfies $T_p(cn^{2/p-1})$ on $(\mathcal{X}^n, d_{p,n})$ with the metric

$$d_{p,n}(x^n, y^n) \triangleq \left(\sum_{i=1}^n d^p(x_i, y_i)\right)^{1/p}, \qquad \forall x^n, y^n \in \mathcal{X}^n. \tag{3.198}$$

Proof. Suppose μ satisfies $T_p(c)$. Fix some n and an arbitrary probability measure ν on $(\mathcal{X}^n, d_{p,n})$. Let $X^n, Y^n \in \mathcal{X}^n$ be two independent random n-tuples, such that

$$P_{X^n} = P_{X_1} \otimes P_{X_2|X_1} \otimes \ldots \otimes P_{X_n|X^{n-1}} = \nu \tag{3.199}$$

$$P_{Y^n} = P_{Y_1} \otimes P_{Y_2} \otimes \ldots \otimes P_{Y_n} = \mu^{\otimes n}. \tag{3.200}$$

For each $i \in \{1, \ldots, n\}$, let us define the "conditional" W_p distance

$$W_p(P_{X_i|X^{i-1}}, P_{Y_i}|P_{X^{i-1}}) \triangleq \left(\int_{\mathcal{X}^{i-1}} W_p^p(P_{X_i|X^{i-1}=x^{i-1}}, P_{Y_i}) P_{X^{i-1}}(\mathrm{d}x^{i-1}) \right)^{1/p}.$$

We will now prove that

$$W_p^p(\nu, \mu^{\otimes n}) = W_p^p(P_{X^n}, P_{Y^n}) \le \sum_{i=1}^n W_p^p(P_{X_i|X^{i-1}}, P_{Y_i}|P_{X^{i-1}}), \tag{3.201}$$

where the L^p Wasserstein distance on the left-hand side is computed w.r.t. the $d_{p,n}$ metric. By Lemma 13, there exists an optimal coupling of P_{X_1} and P_{Y_1}, i.e., a pair (X_1^*, Y_1^*) of jointly distributed $\mathcal{X}$-valued random variables, such that $P_{X_1^*} = P_{X_1}$, $P_{Y_1^*} = P_{Y_1}$, and

$$W_p^p(P_{X_1}, P_{Y_1}) = \mathbb{E}[d^p(X_1^*, Y_1^*)].$$

Now for each $i = 2, \ldots, n$ and each choice of $x^{i-1} \in \mathcal{X}^{i-1}$, again by Lemma 13, there exists an optimal coupling of $P_{X_i|X^{i-1}=x^{i-1}}$ and P_{Y_i}, i.e., a pair $(X_i^*(x^{i-1}), Y_i^*(x^{i-1}))$ of jointly distributed $\mathcal{X}$-valued random variables, such that $P_{X_i^*(x^{i-1})} = P_{X_i|X^{i-1}=x^{i-1}}$, $P_{Y_i^*(x^{i-1})} = P_{Y_i}$, and

$$W_p^p(P_{X_i|X^{i-1}=x^{i-1}}, P_{Y_i}) = \mathbb{E}[d^p(X_i^*(x^{i-1}), Y_i^*(x^{i-1}))]. \tag{3.202}$$

Moreover, because $\mathcal{X}$ is a Polish space, all couplings can be constructed in such a way that the mapping $x^{i-1} \mapsto \mathbb{P}\big((X_i^*(x^{i-1}), Y_i^*(x^{i-1})) \in C\big)$ is measurable for each Borel set $C \subseteq \mathcal{X} \times \mathcal{X}$ [51]. In other words, for each i we can define the regular conditional distributions

$$P_{X_i^* Y_i^* | X^{*i-1}=x^{i-1}} \triangleq P_{X_i^*(x^{i-1}) Y_i^*(x^{i-1})}, \qquad \forall x^{i-1} \in \mathcal{X}^{i-1}$$

such that

$$P_{X^{*n} Y^{*n}} = P_{X_1^* Y_1^*} \otimes P_{X_2^* Y_2^* | X_1^*} \otimes \ldots \otimes P_{X_n^* | X^{*n-1}}$$

is a coupling of $P_{X^n} = \nu$ and $P_{Y^n} = \mu^{\otimes n}$, and

$$W_p^p(P_{X_i|X^{i-1}}, P_{Y_i}) = \mathbb{E}[d^p(X_i^*, Y_i^*)|X^{*i-1}], \qquad i = 1, \ldots, n. \tag{3.203}$$

By definition of W_p, we then have

$$W_p^p(\nu, \mu^{\otimes n}) \le \mathbb{E}[d_{p,n}^p(X^{*n}, Y^{*n})] \tag{3.204}$$

$$= \sum_{i=1}^n \mathbb{E}[d^p(X_i^*, Y_i^*)] \tag{3.205}$$

$$= \sum_{i=1}^n \mathbb{E}\Big[\mathbb{E}[d^p(X_i^*, Y_i^*)|X^{*i-1}]\Big] \tag{3.206}$$

$$= \sum_{i=1}^n W_p^p(P_{X_i|X^{i-1}}, P_{Y_i}|P_{X^{i-1}}), \tag{3.207}$$

where:

- (3.204) is due to the fact that (X^{*n}, Y^{*n}) is a (not necessarily optimal) coupling of $P_{X^n} = \nu$ and $P_{Y^n} = \mu^{\otimes n}$;
- (3.205) is by the definition (3.198) of $d_{p,n}$;
- (3.206) is by the law of iterated expectation; and
- (3.207) is by (3.202).

We have thus proved (3.201). By hypothesis, μ satisfies $T_p(c)$ on $(\mathcal{X}, d)$. Therefore, since $P_{Y_i} = \mu$ for every i, we can write

$$\begin{aligned} W_p^p(P_{X_i|X^{i-1}}, P_{Y_i}|P_{X^{i-1}}) &= \int_{\mathcal{X}^{i-1}} W_p^p(P_{X_i|X^{i-1}=x^{i-1}}, P_{Y_i}) P_{X^{i-1}}(\mathrm{d}x^{i-1}) \\ &\le \int_{\mathcal{X}^{i-1}} \big(2cD(P_{X_i|X^{i-1}=x^{i-1}} \| P_{Y_i})\big)^{p/2} P_{X^{i-1}}(\mathrm{d}x^{i-1}) \\ &= (2c)^{p/2} \big(D(P_{X_i|X^{i-1}} \| P_{Y_i} | P_{X^{i-1}})\big)^{p/2}. \end{aligned} \tag{3.208}$$

Summing from $i = 1$ to $i = n$ and using (3.201), (3.208) and Hölder's inequality, we obtain

$$\begin{aligned} W_p^p(\nu, \mu^{\otimes n}) &\le (2c)^{p/2} \sum_{i=1}^n \big(D(P_{X_i|X^{i-1}} \| P_{Y_i} | P_{X^{i-1}})\big)^{p/2} \\ &\le (2c)^{p/2} n^{1-p/2} \left(\sum_{i=1}^n D(P_{X_i|X^{i-1}} \| P_{Y_i} | P_{X^{i-1}}) \right)^{p/2} \\ &= (2c)^{p/2} n^{1-p/2} \left(D(P_{X^n} \| P_{Y^n})\right)^{p/2} \\ &= (2c)^{p/2} n^{1-p/2} D(\nu \| \mu^{\otimes n})^{p/2}, \end{aligned}$$

where the third line is by the chain rule for the divergence, and since P_{Y^n} is a product probability measure. Taking the p-th root of both sides, we finally get

$$W_p(\nu, \mu^{\otimes n}) \le \sqrt{2cn^{2/p-1} D(\nu \| \mu^{\otimes n})},$$

i.e., $\mu^{\otimes n}$ indeed satisfies the $T_p(cn^{2/p-1})$ inequality. □

Since W_2 dominates W_1 (cf. item 2 of Lemma 13), a $T_2(c)$ inequality is stronger than a $T_1(c)$ inequality (for an arbitrary $c > 0$). Moreover, as Proposition 10 above shows, T_2 inequalities tensorize *exactly*: if μ satisfies T_2 with a constant $c > 0$, then $\mu^{\otimes n}$ also satisfies T_2 for every n with the *same* constant c. By contrast, if μ only satisfies $T_1(c)$, then the product measure $\mu^{\otimes n}$ satisfies T_1 with the much worse constant cn. As we shall shortly see, this sharp difference between the T_1 and T_2 inequalities actually has deep consequences. In a nutshell, in the two sections that follow, we will show that, for $p \in \{1, 2\}$, a given probability measure μ satisfies a $T_p(c)$ inequality on $(\mathcal{X}, d)$ if and only if it has Gaussian concentration with constant $1/(2c)$. Suppose now that we wish to show Gaussian concentration for the product measure $\mu^{\otimes n}$ on the product space $(\mathcal{X}^n, d_{1,n})$. Following our tensorization programme, we could first show that μ satisfies a transportation cost inequality for some $p \in [1, 2]$, then apply Proposition 10 and consequently also apply Proposition 9. If we go through with this approach, we will see that:

- if μ satisfies $T_1(c)$ on $(\mathcal{X}, d)$, then $\mu^{\otimes n}$ satisfies $T_1(cn)$ on $(\mathcal{X}^n, d_{1,n})$, which is equivalent to Gaussian concentration with constant $1/(2cn)$. In this case, the concentration phenomenon becomes weaker and weaker as the dimension n increases.

- if, on the other hand, μ satisfies $T_2(c)$ on $(\mathcal{X}, d)$, then $\mu^{\otimes n}$ satisfies $T_2(c)$ on $(\mathcal{X}^n, d_{2,n})$, which is equivalent to Gaussian concentration with the same constant $1/(2c)$, and this constant is *independent* of the dimension n. Of course, these two approaches give the same constants in concentration inequalities for sums of independent random variables: if f is a 1-Lipschitz function on $(\mathcal{X}, d)$, then from the fact that

$$\begin{aligned} d_{1,n}(x^n, y^n) &= \sum_{i=1}^n d(x_i, y_i) \\ &\le \sqrt{n} \left(\sum_{i=1}^n d^2(x_i, y_i) \right)^{\frac{1}{2}} \\ &= \sqrt{n}\, d_{2,n}(x^n, y^n) \end{aligned}$$

we can conclude that, for $f_n(x^n) \triangleq (1/n) \sum_{i=1}^n f(x_i)$,

$$\|f_n\|_{\mathrm{Lip},1} \triangleq \sup_{x^n \neq y^n} \frac{|f_n(x^n) - f_n(y^n)|}{d_{1,n}(x^n, y^n)} \le \frac{1}{n}$$

and

$$\|f_n\|_{\mathrm{Lip},2} \triangleq \sup_{x^n \neq y^n} \frac{|f_n(x^n) - f_n(y^n)|}{d_{2,n}(x^n, y^n)} \le \frac{1}{\sqrt{n}}$$

(the latter estimate cannot be improved). Therefore, both $T_1(c)$ and $T_2(c)$ give

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n f(X_i) \ge r \right) \le \exp\left(-\frac{nr^2}{2c\|f\|_{\mathrm{Lip}}^2} \right),$$

where $X_1, \ldots, X_n \in \mathcal{X}$ are i.i.d. random variables whose common marginal μ satisfies either $T_2(c)$ or $T_1(c)$, and f is a Lipschitz function on $\mathcal{X}$ with $\mathbb{E}[f(X_1)] = 0$. The difference between T_1 and T_2 inequalities becomes quite pronounced in the case of "nonlinear" functions of $X_1, \ldots, X_n$.

However, it is an experimental fact that T_1 inequalities are easier to work with than T_2 inequalities.

The same strategy as above can be used to prove the following generalization of Proposition 10:

Proposition 11. For any $p \in [1, 2]$, the following statement is true: Let $\mu_1, \ldots, \mu_n$ be n Borel probability measures on a Polish space $(\mathcal{X}, d)$, such that μ_i satisfies $T_p(c_i)$ for some $c_i > 0$, for each $i = 1, \ldots, n$. Let $c \triangleq \max_{1 \le i \le n} c_i$. Then $\mu = \mu_1 \otimes \ldots \mu_n$ satisfies $T_p(cn^{2/p-1})$ on $(\mathcal{X}^n, d_{p,n})$.

3.4.3 Gaussian concentration and T_1 inequalities

As we have shown above, Marton's argument can be used to deduce Gaussian concentration from a transportation cost inequality. As we will demonstrate here and in the following section, in certain cases these properties are *equivalent*. We will consider first the case when μ satisfies a T_1 inequality. The first proof of equivalence between T_1 and Gaussian concentration is due to Bobkov and Götze [44], and it relies on the following variational representations of the L^1 Wasserstein distance and the divergence:

1. **Kantorovich–Rubinstein theorem** [50, Theorem 1.14] For any two $\mu, \nu \in \mathcal{P}_1(\mathcal{X})$,

$$W_1(\mu, \nu) = \sup_{f:\, \|f\|_{\mathrm{Lip}} \le 1} \left| \mathbb{E}_\mu[f] - \mathbb{E}_\nu[f] \right|. \tag{3.209}$$

2. **Donsker–Varadhan lemma** [69, Lemma 6.2.13]: for any two Borel probability measures μ, ν,

$$D(\nu \| \mu) = \sup_{g:\, \exp(g) \in L^1(\mu)} \left\{ \mathbb{E}_\nu[g] - \ln \mathbb{E}_\mu[\exp(g)] \right\} \tag{3.210}$$

Theorem 36 (Bobkov–Götze [44])**.** A Borel probability measure $\mu \in \mathcal{P}_1(\mathcal{X})$ satisfies $T_1(c)$ if and only if the inequality

$$\mathbb{E}_\mu \left\{ \exp[tf(X)] \right\} \le \exp[ct^2/2] \tag{3.211}$$

holds for all 1-Lipschitz functions $f : \mathcal{X} \to \mathbb{R}$ with $\mathbb{E}_\mu[f(X)] = 0$, and all $t \in \mathbb{R}$.

Remark 40. The moment condition $\mathbb{E}_\mu[d(X, x_0)] < \infty$ is needed to ensure that every Lipschitz function $f : \mathcal{X} \to \mathbb{R}$ is μ-integrable:

$$\mathbb{E}_\mu\big[|f(X)|\big] \le |f(x_0)| + \mathbb{E}_\mu\big[|f(X) - f(x_0)|\big] \le |f(x_0)| + \|f\|_{\text{Lip}}\, \mathbb{E}_\mu\big[d(X, x_0)\big] < \infty.$$

Proof. Without loss of generality, we may consider (3.211) only for $t \ge 0$.

Suppose first that μ satisfies $T_1(c)$. Consider some $\nu \ll \mu$. Using the $T_1(c)$ property of μ together with the Kantorovich–Rubinstein formula (3.209), we can write

$$\int f \mathrm{d}\nu \le W_1(\nu, \mu) \le \sqrt{2cD(\nu\|\mu)}$$

for any 1-Lipschitz $f : \mathcal{X} \to \mathbb{R}$ with $\mathbb{E}_\mu[f] = 0$. Next, from the fact that

$$\inf_{t>0} \left(\frac{a}{t} + \frac{bt}{2} \right) = \sqrt{2ab} \tag{3.212}$$

for any $a, b \ge 0$, we see that any such f must satisfy

$$\int_{\mathcal{X}} f \, \mathrm{d}\nu \le \frac{1}{t} D(\nu\|\mu) + \frac{ct}{2}, \qquad \forall\, t > 0.$$

Rearranging, we obtain

$$\int_{\mathcal{X}} tf \, \mathrm{d}\nu - \frac{ct^2}{2} \le D(\nu\|\mu), \qquad \forall\, t > 0.$$

Applying this inequality to $\nu = \mu^{(g)}$ (the g-tilting of μ) where $g \triangleq tf$, and using the fact that

$$\begin{aligned} D(\mu^{(g)}\|\mu) &= \int_{\mathcal{X}} g \, \mathrm{d}\mu^{(g)} - \ln \int \exp(g) \, \mathrm{d}\mu \\ &= \int_{\mathcal{X}} tf \, \mathrm{d}\nu - \ln \int \exp(tf) \, \mathrm{d}\mu \end{aligned}$$

we deduce that

$$\ln \left(\int_{\mathcal{X}} \exp(tf) \, \mathrm{d}\mu \right) \le \frac{ct^2}{2} \quad \Longrightarrow \quad \ln \mathbb{E}_\mu \left\{ \exp\left[tf(X) - \frac{ct^2}{2} \right] \right\} \le 0 \tag{3.213}$$

for all $t \ge 0$, and all f with $\|f\|_{\text{Lip}} \le 1$ and $\mathbb{E}_\mu[f] = 0$, which is precisely (3.211).

Conversely, assume that μ satisfies (3.211). Then any function of the form tf, where $t > 0$ and f is as in (3.211), is feasible for the supremization in (3.210). Consequently, given any $\nu \ll \mu$, we can write

$$\begin{aligned} D(\nu\|\mu) &\ge \int_{\mathcal{X}} tf \, \mathrm{d}\nu - \ln \int_{\mathcal{X}} \exp(tf) \, \mathrm{d}\mu \\ &= \int_{\mathcal{X}} tf \, \mathrm{d}\nu - \int_{\mathcal{X}} tf \, \mathrm{d}\mu - \frac{ct^2}{2} \end{aligned}$$

where in the second step we have used the fact that $\int f \, \mathrm{d}\mu = 0$ by hypothesis, as well as (3.211). Rearranging gives

$$\left| \int_{\mathcal{X}} f \, \mathrm{d}\nu - \int_{\mathcal{X}} f \, \mathrm{d}\mu \right| \le \frac{1}{t} D(\nu \| \mu) + \frac{ct}{2}, \qquad \forall t > 0 \tag{3.214}$$

(the absolute value in the left-hand side is a consequence of the fact that exactly the same argument goes through with $-f$ instead of f). Applying (3.212), we see that the bound

$$\left| \int_{\mathcal{X}} f \, \mathrm{d}\nu - \int_{\mathcal{X}} f \, \mathrm{d}\mu \right| \le \sqrt{2cD(\nu \| \mu)}. \tag{3.215}$$

holds for all 1-Lipschitz f with $\mathbb{E}_\mu[f] = 0$. In fact, we may now drop the condition that $\mathbb{E}_\mu[f] = 0$ by replacing f with $f - \mathbb{E}_\mu[f]$. Thus, taking the supremum over all 1-Lipschitz f on the left-hand side of (3.215) and using the Kantorovich–Rubinstein formula (3.209), we conclude that $W_1(\mu, \nu) \le \sqrt{2cD(\nu \| \mu)}$ for every $\nu \ll \mu$, i.e., μ satisfies $T_1(c)$. This completes the proof of Theorem 36. □

The above theorem gives us an alternative way of deriving Gaussian concentration for Lipschitz functions:

Corollary 10. Let $\mathcal{A}$ be the space of all Lipschitz functions on $\mathcal{X}$, and define the operator Γ on $\mathcal{A}$ via

$$\Gamma f(x) \triangleq \limsup_{y \in \mathcal{X} \colon d(x,y) \downarrow 0} \frac{|f(x) - f(y)|}{d(x,y)}, \qquad \forall x \in \mathcal{X}.$$

Suppose that μ satisfies $T_1(c)$, then it implies the following concentration inequality for every $f \in \mathcal{A}$:

$$\mathbb{P}_\mu\Big(f(X) \ge \mathbb{E}[f(X)] + r\Big) \le \exp\left(-\frac{r^2}{2c\|f\|_{\mathrm{Lip}}^2}\right), \qquad \forall\, r \ge 0.$$

Corollary 10 shows that the method based on transportation cost inequalities gives the same (sharp) constants as the entropy method. As another illustration, we prove the following sharp estimate:

Theorem 37. Let $\mathcal{X} = \{0,1\}^n$, equipped with the metric

$$d(x^n, y^n) = \sum_{i=1}^n 1_{\{x_i \neq y_i\}}. \tag{3.216}$$

Let $X_1, \ldots, X_n \in \{0,1\}$ be i.i.d. Bernoulli(p) random variables. Then, for any Lipschitz function $f : \{0,1\}^n \to \mathbb{R}$,

$$\mathbb{P}\left(f(X^n) - \mathbb{E}[f(X^n)] \ge r\right) \le \exp\left(-\frac{\ln[(1-p)/p]\, r^2}{n\|f\|_{\mathrm{Lip}}^2 (1-2p)}\right), \qquad \forall r \ge 0. \tag{3.217}$$

Proof. Taking into account Remark 41, we may assume without loss of generality that $p \neq 1/2$. From the distribution-dependent refinement of Pinsker's inequality (3.190), it follows that the Bernoulli(p) measure satisfies $T_1(1/(2\varphi(p)))$ w.r.t. the Hamming metric, where $\varphi(p)$ is defined in (3.187). By Proposition 10, the product of n Bernoulli(p) measures satisfies $T_1(n/(2\varphi(p)))$ w.r.t. the metric (3.216). The bound (3.217) then follows from Corollary 10. □

Remark 41. In the limit as $p \to 1/2$, the right-hand side of (3.217) becomes $\exp\left(-\frac{2r^2}{n\|f\|_{\mathrm{Lip}}^2}\right)$.

Remark 42. If $\|f\|_{\text{Lip}} \le C/n$ for some $C > 0$, then (3.217) implies that

$$\mathbb{P}\bigg(f(X^n) - \mathbb{E}[f(X^n)] \ge r\bigg) \le \exp\left(-\frac{\ln[(1-p)/p]}{C^2(1-2p)} \cdot nr^2\right), \qquad \forall\, r \ge 0.$$

This will be the case, for instance, if $f(x^n) = (1/n)\sum_{i=1}^n f_i(x_i)$ for some functions $f_1, \dots, f_n : \{0,1\} \to \mathbb{R}$ satisfying $|f_i(0) - f_i(1)| \le C$ for all $i = 1, \dots, n$. More generally, any f satisfying (3.134) with $c_i = c'_i/n$, $i = 1, \dots, n$, for some constants $c'_1, \dots, c'_n \ge 0$, satisfies

$$\mathbb{P}\bigg(f(X^n) - \mathbb{E}[f(X^n)] \ge r\bigg) \le \exp\left(-\frac{\ln[(1-p)/p]}{(1-2p)\sum_{i=1}^n (c'_i)^2} \cdot nr^2\right), \qquad \forall\, r \ge 0.$$

3.4.4 Dimension-free Gaussian concentration and T_2 inequalities

So far, we have confined our discussion to the "one-dimensional" case of a probability measure μ on a Polish space $(\mathcal{X}, d)$. Recall, however, that in most applications our interest is in functions of n independent random variables taking values in $\mathcal{X}$. Proposition 10 shows that the transportation cost inequalities tensorize, so in principle this property can be used to derive concentration inequalities for such functions.

As before, let $(\mathcal{X}, d, \mu)$ be a metric probability space. We say that μ has *dimension-free Gaussian concentration* if there exist constants $K, \kappa > 0$, such that for any $k \in \mathbb{N}$

$$A \subseteq \mathcal{X}^k \text{ and } \mu^{\otimes k}(A) \ge 1/2 \qquad \Longrightarrow \qquad \mu^{\otimes k}(A_r) \ge 1 - Ke^{-\kappa r^2}, \forall r > 0 \tag{3.218}$$

where the isoperimetric enlargement A_r of a Borel set $A \subseteq \mathcal{X}^k$ is defined w.r.t. the metric $d_k \equiv d_{2,k}$ defined according to (3.198):

$$A_r \triangleq \left\{ y^k \in \mathcal{X}^k : \sum_{i=1}^k d^2(x_i, y_i) < r^2,\, \forall x^k \in A \right\}.$$

Remark 43. As before, we are mainly interested in the constant κ in the exponent. Thus, we may explicitly say that μ has dimension-free Gaussian concentration with constant $\kappa > 0$, meaning that (3.218) holds with that κ and some $K > 0$.

Theorem 38 (Talagrand [139])**.** *Let $\mathcal{X} = \mathbb{R}^n$, $d(x,y) = \|x - y\|$, and $\mu = G^n$. Then G^n satisfies a $T_2(1)$ inequality.*

Proof. The proof starts for $n = 1$: let $\mu = G$, let $\nu \in \mathcal{P}(\mathbb{R})$ have density f w.r.t. μ: $f = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}$, and let Φ denote the standard Gaussian cdf, i.e.,

$$\Phi(x) = \int_{-\infty}^x \gamma(y)\mathrm{d}y = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^x \exp\left(-\frac{y^2}{2}\right)\mathrm{d}y, \quad \forall\, x \in \mathbb{R}.$$

If $X \sim G$, then (by item 6 of Lemma 13) the optimal coupling of $\mu = G$ and ν, i.e., the one that achieves the infimum in

$$W_2(\nu, \mu) = W_2(\nu, G) = \inf_{X \sim G,\, Y \sim \nu} \left(\mathbb{E}[(X-Y)^2]\right)^{1/2}$$

is given by $Y = h(X)$ with $h = \mathsf{F}_\nu^{-1} \circ \Phi$. Consequently,

$$\begin{aligned} W_2^2(\nu, G) &= \mathbb{E}[(X - h(X))^2] \\ &= \int_{-\infty}^{\infty} \big(x - h(x)\big)^2 \gamma(x)\,\mathrm{d}x. \end{aligned} \tag{3.219}$$

Since $\mathrm{d}\nu = f \, \mathrm{d}\mu$ with $\mu = G$, and $\mathsf{F}_\nu(h(x)) = \Phi(x)$ for every $x \in \mathbb{R}$, then

$$\int_{-\infty}^{x} \gamma(y) \, \mathrm{d}y = \Phi(x) = \mathsf{F}_\nu(h(x)) = \int_{-\infty}^{h(x)} \mathrm{d}\nu = \int_{-\infty}^{h(x)} f \, \mathrm{d}\mu = \int_{-\infty}^{h(x)} f(y)\gamma(y) \, \mathrm{d}y. \tag{3.220}$$

Differentiating both sides of (3.220) w.r.t. x gives

$$h'(x)f(h(x))\gamma(h(x)) = \gamma(x), \quad \forall \, x \in \mathbb{R} \tag{3.221}$$

and, since $h = \mathsf{F}_\nu^{-1} \circ \Phi$, then h is a monotonic increasing function and

$$\lim_{x \to -\infty} h(x) = -\infty, \quad \lim_{x \to \infty} h(x) = \infty.$$

Moreover,

$$\begin{aligned} D(\nu \| G) &= D(\nu \| \mu) \\ &= \int_{\mathbb{R}} \mathrm{d}\nu \, \ln \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \\ &= \int_{-\infty}^{\infty} \ln \bigl(f(x)\bigr) \, \mathrm{d}\nu(x) \\ &= \int_{-\infty}^{\infty} f(x) \, \ln \bigl(f(x)\bigr) \, \mathrm{d}\mu(x) \\ &= \int_{-\infty}^{\infty} f(x) \, \ln \bigl(f(x)\bigr) \, \gamma(x) \, \mathrm{d}x \\ &= \int_{-\infty}^{\infty} f\bigl(h(x)\bigr) \, \ln \bigl(f\bigl(h(x)\bigr)\bigr) \, \gamma\bigl(h(x)\bigr) \, h'(x) \, \mathrm{d}x \\ &= \int_{-\infty}^{\infty} \ln \bigl(f(h(x))\bigr) \, \gamma(x) \, \mathrm{d}x \end{aligned} \tag{3.222}$$

while using above the change-of-variables formula, and also (3.221) for the last equality. From (3.221), we have

$$\ln f(h(x)) = \ln \left(\frac{\gamma(x)}{h'(x) \, \gamma\bigl(h(x)\bigr)} \right) = \frac{h^2(x) - x^2}{2} - \ln h'(x)$$

so, by substituting this into (3.222), it follows that

$$\begin{aligned} D(\nu \| \mu) &= \frac{1}{2} \int_{-\infty}^{\infty} \bigl[h^2(x) - x^2\bigr] \, \gamma(x) \, \mathrm{d}x - \int_{-\infty}^{\infty} \ln h'(x) \, \gamma(x) \, \mathrm{d}x \\ &= \frac{1}{2} \int_{-\infty}^{\infty} \bigl(x - h(x)\bigr)^2 \, \gamma(x) \, \mathrm{d}x + \int_{-\infty}^{\infty} x\bigl(h(x) - x\bigr) \, \gamma(x) \, \mathrm{d}x - \int_{-\infty}^{\infty} \ln h'(x) \, \gamma(x) \, \mathrm{d}x \\ &= \frac{1}{2} \int_{-\infty}^{\infty} \bigl(x - h(x)\bigr)^2 \, \gamma(x) \, dx + \int_{-\infty}^{\infty} (h'(x) - 1) \, \gamma(x) \, \mathrm{d}x - \int_{-\infty}^{\infty} \ln h'(x) \, \gamma(x) \, \mathrm{d}x \\ &\geq \frac{1}{2} \int_{-\infty}^{\infty} \bigl(x - h(x)\bigr)^2 \, \gamma(x) \, \mathrm{d}x \\ &= \frac{1}{2} \, W_2^2(\nu, \mu) \end{aligned}$$

where the third line relies on integration by parts, the forth line follows from the inequality $\ln t \leq t - 1$ for $t > 0$, and the last line holds due to (3.219). This shows that $\mu = G$ satisfies $T_2(1)$, so it completes the proof of Theorem 38 for $n = 1$. Finally, this theorem is generalized for an arbitrary n by tensorization via Proposition 10. □

We now get to the main result of this section, namely that dimension-free Gaussian concentration and T_2 are equivalent:

Theorem 39. Let $(\mathcal{X}, d, \mu)$ be a metric probability space. Then, the following statements are equivalent:

1. μ satisfies $T_2(c)$.
2. μ has dimension-free Gaussian concentration with constant $\kappa = 1/(2c)$.

Remark 44. As we will see, the implication 1) ⇒ 2) follows easily from the tensorization property of transportation cost inequalities (Proposition 10). The reverse implication 2) ⇒ 1) is a nontrivial result, which was proved by Gozlan [56] using an elegant probabilistic approach relying on the theory of large deviations [69].

Proof. We first prove that 1) ⇒ 2). Assume that μ satisfies $T_2(c)$ on $(\mathcal{X}, d)$. Fix some $k \in \mathbb{N}$ and consider the metric probability space $(\mathcal{X}^k, d_{2,k}, \mu^{\otimes k})$, where the metric $d_{2,k}$ is defined by (3.198) with $p = 2$. By the tensorization property of transportation cost inequalities (Proposition 10), the product measure $\mu^{\otimes k}$ satisfies $T_2(c)$ on $(\mathcal{X}^k, d_{2,k})$. Because the L^2 Wasserstein distance dominates the L^1 Wasserstein distance (by item 2 of Lemma 13), $\mu^{\otimes k}$ also satisfies $T_1(c)$ on $(\mathcal{X}^k, d_{2,k})$. Therefore, by the Bobkov–Götze theorem (Theorem 36 in the preceding section), $\mu^{\otimes k}$ has Gaussian concentration (3.163) with respect to $d_{2,k}$ with constant $\kappa = 1/(2c)$. Since this holds for every $k \in \mathbb{N}$, we conclude that μ indeed has dimension-free Gaussian concentration with constant $\kappa = 1/(2c)$.

We now prove the converse implication 2) ⇒ 1). Suppose that μ has dimension-free Gaussian concentration with constant $\kappa > 0$. Let us fix some $k \in \mathbb{N}$ and consider the metric probability space $(\mathcal{X}^k, d_{2,k}, \mu^{\otimes k})$. Given $x^k \in \mathcal{X}^k$, let P_{x^k} be the corresponding *empirical measure*, i.e.,

$$\mathsf{P}_{x^k} = \frac{1}{k} \sum_{i=1}^{k} \delta_{x_i}, \tag{3.223}$$

where δ_x denotes a Dirac measure (unit mass) concentrated at $x \in \mathcal{X}$. Now consider a probability measure ν on $\mathcal{X}$, and define the function $f_\nu : \mathcal{X}^k \to \mathbb{R}$ by

$$f_\nu(x^k) \triangleq W_2(\mathsf{P}_{x^k}, \nu), \quad \forall\, x^k \in \mathcal{X}^k.$$

We claim that this function is Lipschitz w.r.t. $d_{2,k}$ with Lipschitz constant $1/\sqrt{k}$. To verify this, note that

$$\begin{aligned}
\left|f_\nu(x^k) - f_\nu(y^k)\right| &= \left|W_2(\mathsf{P}_{x^k}, \nu) - W_2(\mathsf{P}_{y^k}, \nu)\right| \\
&\le W_2(\mathsf{P}_{x^k}, \mathsf{P}_{y^k}) && (3.224) \\
&= \inf_{\pi \in \Pi(\mathsf{P}_{x^k}, \mathsf{P}_{y^k})} \left(\int_{\mathcal{X}} d^2(x, y)\, \pi(\mathrm{d}x, \mathrm{d}y) \right)^{1/2} && (3.225) \\
&\le \left(\frac{1}{k} \sum_{i=1}^{k} d^2(x_i, y_i) \right)^{1/2} && (3.226) \\
&= \frac{1}{\sqrt{k}}\, d_{2,k}(x^k, y^k), && (3.227)
\end{aligned}$$

where

- (3.224) is by the triangle inequality;
- (3.225) is by definition of W_2;

- (3.226) uses by the fact that the measure that places mass $1/k$ on each (x_i, y_i) for $i \in \{1, \ldots, k\}$, is an element of $\Pi(\mathsf{P}_{x^k}, \mathsf{P}_{y^k})$ (due to the definition of an empirical distribution in (3.223), the marginals of the above measure are indeed P_{x^k} and P_{y^k}); and
- (3.227) uses the definition (3.198) of $d_{2,k}$.

Now let us consider the function $f_k \triangleq f_\mu \equiv W_2(\mathsf{P}_{x^k}, \mu)$, for which, as we have just seen, we have $\|f_k\|_{\text{Lip},2} = 1/\sqrt{k}$. Let $X_1, \ldots, X_k$ be i.i.d. draws from μ. Then, by the assumed dimension-free Gaussian concentration property of μ, we have

$$\begin{aligned}\mathbb{P}\Big(f_k(X^k) \geq \mathbb{E}[f_k(X^k)] + r\Big) &\leq \exp\left(-\frac{r^2}{2c\|f\|^2_{\text{Lip},2}}\right)\\ &= \exp\left(-\kappa k r^2\right), \qquad \forall\, r \geq 0 \end{aligned} \tag{3.228}$$

and this inequality holds for every $k \in \mathbb{N}$; note that the last equality holds since $c = \frac{1}{2\kappa}$ and $\|f\|^2_{\text{Lip},2} = \frac{1}{k}$.

Now, if $X_1, X_2, \ldots$ are i.i.d. draws from μ, then the sequence of empirical distributions $\{\mathsf{P}_{X^k}\}_{k=1}^\infty$ almost surely converges weakly to μ (this is known as Varadarajan's theorem [140, Theorem 11.4.1]). Since W_2 metrizes the topology of weak convergence together with the convergence of second moments (cf. Lemma 13), we have $\lim_{k\to\infty} \mathbb{E}[f_k(X^k)] = 0$. Consequently, taking logarithms of both sides of (3.228), dividing by k, and taking limit superior as $k \to \infty$, we get

$$\limsup_{k\to\infty} \frac{1}{k} \ln \mathbb{P}\Big(W_2(\mathsf{P}_{X^k}, \mu) \geq r\Big) \leq -\kappa r^2. \tag{3.229}$$

On the other hand, for a fixed μ, the mapping $\nu \mapsto W_2(\nu, \mu)$ is lower semicontinuous in the topology of weak convergence of probability measures (cf. Lemma 13). Consequently, the set $\{\mu : W_2(\mathsf{P}_{X^k}, \mu) > r\}$ is open in the weak topology, so by Sanov's theorem [69, Theorem 6.2.10]

$$\liminf_{k\to\infty} \frac{1}{k} \ln \mathbb{P}\Big(W_2(\mathsf{P}_{X^k}, \mu) \geq r\Big) \geq -\inf\left\{D(\nu\|\mu) : W_2(\mu, \nu) > r\right\}. \tag{3.230}$$

Combining (3.229) and (3.230), we get that

$$\inf\big\{D(\nu\|\mu) :\, W_2(\mu, \nu) > r\big\} \geq \kappa r^2$$

which then implies that $D(\nu\|\mu) \geq \kappa\, W_2^2(\mu, \nu)$. Upon rearranging, we obtain $W_2(\mu, \nu) \leq \sqrt{\left(\frac{1}{\kappa}\right) D(\nu\|\mu)}$, which is a $T_2(c)$ inequality with $c = \frac{1}{2\kappa}$. This completes the proof of Theorem 39. □

3.4.5 A grand unification: the HWI inequality

At this point, we have seen two perspectives on the concentration of measure phenomenon: functional (through various log-Sobolev inequalities) and probabilistic (through transportation cost inequalities). We now show that these two perspectives are, in a very deep sense, equivalent, at least in the Euclidean setting of $\mathbb{R}^n$. This equivalence is captured by a striking inequality, due to Otto and Villani [141], which relates three measures of similarity between probability measures: the divergence, L^2 Wasserstein distance, and Fisher information distance. In the literature on optimal transport, the divergence between two probability measures Q and P is often denoted by $H(Q\|P)$ or $H(Q, P)$, due to its close links to the Boltzmann H-functional of statistical physics. For this reason, the inequality we have alluded to above has been dubbed the *HWI inequality*, where H stands for the divergence, W for the Wasserstein distance, and I for the Fisher information distance.

As a warm-up, we first state a weaker version of the HWI inequality specialized to the Gaussian distribution, and give a self-contained information-theoretic proof following [142]:

Theorem 40. Let G be the standard Gaussian probability distribution on $\mathbb{R}$. Then, the inequality

$$D(P\|G) \le W_2(P,G)\sqrt{I(P\|G)}, \tag{3.231}$$

where W_2 is the L^2 Wasserstein distance w.r.t. the absolute-value metric $d(x,y)=|x-y|$, holds for any Borel probability distribution P on $\mathbb{R}$, for which the right-hand side of (3.231) is finite.

Proof. Without loss of generality, we may assume that P has zero mean and unit variance. We first show the following: Let X and Y be a pair of real-valued random variables, and let $N \sim G$ be independent of (X,Y). Then for any $t>0$

$$D(P_{X+\sqrt{t}N}\|P_{Y+\sqrt{t}N}) \le \frac{1}{2t}\, W_2^2(P_X,P_Y). \tag{3.232}$$

Using the chain rule for divergence, we can expand $D(P_{X,Y,X+\sqrt{t}N}\|P_{X,Y,Y+\sqrt{t}N})$ in two ways as

$$\begin{aligned} D(P_{X,Y,X+\sqrt{t}N}\|P_{X,Y,Y+\sqrt{t}N}) &= D(P_{X+\sqrt{t}N}\|P_{Y+\sqrt{t}N}) + D(P_{X,Y|X+\sqrt{t}N}\|P_{X,Y|Y+\sqrt{t}N}|P_{X+\sqrt{t}N}) \\ &\ge D(P_{X+\sqrt{t}N}\|P_{Y+\sqrt{t}N}) \end{aligned}$$

and since N is independent of (X,Y), then

$$\begin{aligned} D(P_{X,Y,X+\sqrt{t}N}\|P_{X,Y,Y+\sqrt{t}N}) &= D(P_{X+\sqrt{t}N}\,\|\,P_{Y+\sqrt{t}N}|P_{X,Y}) \\ &= \mathbb{E}[D(\mathcal{N}(X,t)\,\|\,\mathcal{N}(Y,t))\,|\,X,Y] \\ &= \frac{1}{2t}\,\mathbb{E}[(X-Y)^2] \end{aligned}$$

where the last equality is a special case of the equality

$$D\big(\mathcal{N}(m_1,\sigma_1^2)\,\|\,\mathcal{N}(m_2,\sigma_2^2)\big) = \frac{1}{2}\ln\left(\frac{\sigma_1^2}{\sigma_2^2}\right) + \frac{1}{2}\left(\frac{(m_1-m_2)^2}{\sigma_2^2} + \frac{\sigma_1^2}{\sigma_2^2} - 1\right)$$

where $\sigma_1^2=\sigma_2^2=t$, $m_1=X$ and $m_2=Y$ (given the values of X and Y). Therefore, for any pair (X,Y) of jointly distributed real-valued random variables, we have

$$D(P_{X+\sqrt{t}N}\|P_{Y+\sqrt{t}N}) \le \frac{1}{2t}\,\mathbb{E}[(X-Y)^2]. \tag{3.233}$$

The left-hand side of (3.233) only depends on the marginal distributions of X and Y. Hence, taking the infimum of the right-hand side of (3.233) w.r.t. all couplings of P_X and P_Y (i.e., all $\mu \in \Pi(P_X,P_Y)$), we get (3.232) (see (3.177)).

Let X have distribution P, Y have distribution G, and define the function

$$F(t) \triangleq D(P_{X+\sqrt{t}Z}\|P_{Y+\sqrt{t}Z}),$$

where $Z \sim G$ is independent of (X,Y). Then $F(0)=D(P\|G)$, and from (3.232) we have

$$F(t) \le \frac{1}{2t}\,W_2^2(P_X,P_Y) = \frac{1}{2t}\,W_2^2(P,G). \tag{3.234}$$

Moreover, the function $F(t)$ is differentiable, and it follows from [118, Eq. (32)] that

$$F'(t) = \frac{1}{2t^2}\big[\mathsf{mmse}(X,t^{-1}) - \mathsf{lmmse}(X,t^{-1})\big] \tag{3.235}$$

where $\mathsf{mmse}(X, \cdot)$ and $\mathsf{lmmse}(X, \cdot)$ have been defined in (3.56) and (3.58), respectively. Now, for any $t > 0$ we have

$$\begin{aligned} D(P\|G) &= F(0) \\ &= -\big(F(t) - F(0)\big) + F(t) \\ &= -\int_0^t F'(s)\mathrm{d}s + F(t) \\ &= \frac{1}{2}\int_0^t \frac{1}{s^2}\left(\mathsf{lmmse}(X, s^{-1}) - \mathsf{mmse}(X, s^{-1})\right)\mathrm{d}s + F(t) && (3.236) \\ &\le \frac{1}{2}\int_0^t \left(\frac{1}{s(s+1)} - \frac{1}{s(sJ(X)+1)}\right)\mathrm{d}s + \frac{1}{2t}W_2^2(P, G) && (3.237) \\ &= \frac{1}{2}\left(\ln\frac{tJ(X)+1}{t+1} + \frac{W_2^2(P, G)}{t}\right) && (3.238) \\ &\le \frac{1}{2}\left(\frac{t(J(X)-1)}{t+1} + \frac{W_2^2(P, G)}{t}\right) && (3.239) \\ &\le \frac{1}{2}\left(I(P\|G)\,t + \frac{W_2^2(P, G)}{t}\right) && (3.240) \end{aligned}$$

where

- (3.236) uses (3.235);
- (3.237) uses (3.59), the Van Trees inequality (3.60), and (3.234);
- (3.238) is an exercise in calculus;
- (3.239) uses the inequality $\ln x \le x - 1$ for $x > 0$; and
- (3.240) uses the formula (3.54) (so $I(P\|G) = J(X) - 1$ since $X \sim P$ has zero mean and unit variance, and one needs to substitute $s = 1$ in (3.54) to get $G_s = G$), and the fact that $t \ge 0$.

Optimizing the choice of t in (3.240), we get (3.231). □

Remark 45. Note that the HWI inequality (3.231) together with the T_2 inequality for the Gaussian distribution imply a weaker version of the log-Sobolev inequality (3.41) (i.e., with a larger constant). Indeed, using the T_2 inequality of Theorem 38 on the right-hand side of (3.231), we get

$$\begin{aligned} D(P\|G) &\le W_2(P, G)\sqrt{I(P\|G)} \\ &\le \sqrt{2D(P\|G)}\sqrt{I(P\|G)}, \end{aligned}$$

which gives $D(P\|G) \le 2I(P\|G)$. It is not surprising that we end up with a suboptimal constant here: the series of bounds leading up to (3.240) contributes a lot more slack than the single use of the van Trees inequality (3.60) in our proof of Stam's inequality (which is equivalent to the Gaussian log-Sobolev inequality of Gross) in Section 3.2.1.

We are now ready to state the HWI inequality in its strong form:

Theorem 41 (Otto–Villani [141]). Let P be a Borel probability measure on $\mathbb{R}^n$ that is absolutely continuous w.r.t. the Lebesgue measure, and let the corresponding pdf p be such that

$$\nabla^2 \ln\left(\frac{1}{p}\right) \succeq KI_n \tag{3.241}$$

for some $K \in \mathbb{R}$ (where ∇^2 denotes the Hessian, and the matrix inequality $A \succeq B$ means that $A - B$ is positive semidefinite). Then, any probability measure $Q \ll P$ satisfies

$$D(Q\|P) \le W_2(Q,P)\sqrt{I(Q\|P)} - \frac{K}{2}W_2^2(Q,P). \tag{3.242}$$

We omit the proof, which relies on deep structural properties of optimal transportation mappings achieving the infimum in the definition of the L^2 Wasserstein metric w.r.t. the Euclidean norm in $\mathbb{R}^n$. (An alternative, simpler proof was given later by Cordero–Erausquin [143].) We can, however, highlight a couple of key consequences (see [141]):

1. Suppose that P, in addition to satisfying the conditions of Theorem 41, also satisfies a $T_2(c)$ inequality. Using this fact in (3.242), we get

$$D(Q\|P) \le \sqrt{2cD(Q\|P)}\sqrt{I(Q\|P)} - \frac{K}{2}W_2^2(Q,P) \tag{3.243}$$

If the pdf p of P is log-concave, so that (3.241) holds with $K = 0$, then (3.243) implies the inequality

$$D(Q\|P) \le 2c\, I(Q\|P) \tag{3.244}$$

for any $Q \ll P$. This is, of course, an Euclidean log-Sobolev inequality similar to the one satisfied by $P = G^n$. Of course, the constant in front of the Fisher information distance $I(\cdot\|\cdot)$ on the right-hand side of (3.244) is suboptimal, as can be easily seen by letting $P = G^n$, which satisfies $T_2(1)$, and going through the above steps — as we know from Section 3.2 (in particular, see (3.41)), the optimal constant should be 1/2, so the one in (3.244) is off by a factor of 4. On the other hand, it is quite remarkable that, up to constants, the Euclidean log-Sobolev and T_2 inequalities are equivalent.

2. If the pdf p of P is *strongly* log-concave, i.e., if (3.241) holds with some $K > 0$, then P satisfies the Euclidean log-Sobolev inequality with constant $1/K$. Indeed, using Young's inequality $ab \le a^2/2 + b^2/2$, we can write

$$\begin{aligned} D(Q\|P) &\le \sqrt{K}W_2(Q,P)\sqrt{\frac{I(Q\|P)}{K}} - \frac{K}{2}W_2^2(Q,P) \\ &\le \frac{1}{2K}\, I(Q\|P), \end{aligned}$$

which shows that P satisfies the Euclidean LSI($1/K$) inequality. In particular, the standard Gaussian distribution $P = G^n$ satisfies (3.241) with $K = 1$, so we even get the right constants. In fact, the statement that (3.241) with $K > 0$ implies Euclidean LSI($1/K$) was first proved in 1985 by Bakry and Emery [144] using very different means.

3.5 Extension to non-product distributions

Our focus in this chapter has been mostly on functions of independent random variables. However, there is extensive literature on the concentration of measure for weakly dependent random variables. In this section, we describe (without proof) a few results along this direction that explicitly use information-theoretic methods. The examples we give are by no means exhaustive, and are only intended to show that, even in the case of dependent random variables, the underlying ideas are essentially the same as in the independent case.

The basic scenario is exactly as before: We have n random variables $X_1, \ldots, X_n$ with a given joint distribution P (which is now not necessarily of a product form, i.e., $P = P_{X^n}$ may not be equal to $P_{X_1} \otimes \ldots \otimes P_{X_n}$), and we are interested in the concentration properties of some function $f(X^n)$.

3.5.1 Samson's transporation cost inequalities for weakly dependent random variables

Samson [145] has developed a general approach for deriving transportation cost inequalities for dependent random variables that revolves around a certain L^2 measure of dependence. Given the distribution $P = P_{X^n}$ of $(X_1, \dots, X_n)$, consider an upper triangular matrix $\Delta \in \mathbb{R}^{n \times n}$, such that $\Delta_{i,j} = 0$ for $i > j$, $\Delta_{i,i} = 1$ for all i, and for $i < j$

$$\Delta_{i,j} = \sup_{x_i, x_i'} \sup_{x^{i-1}} \sqrt{\left\| P_{X_j^n | X_i = x_i, X^{i-1} = x^{i-1}} - P_{X_j^n | X_i = x_i', X^{i-1} = x^{i-1}} \right\|_{\mathrm{TV}}}. \tag{3.245}$$

Note that in the special case where P is a product measure, the matrix Δ is equal to the $n \times n$ identity matrix. Let $\|\Delta\|$ denote the operator norm of Δ in the Euclidean topology, i.e.,

$$\|\Delta\| \triangleq \sup_{v \in \mathbb{R}^n \colon v \neq 0} \frac{\|\Delta v\|}{\|v\|} = \sup_{v \in \mathbb{R}^n \colon \|v\| = 1} \|\Delta v\|.$$

Following Marton [146], Samson considers a Wasserstein-type distance on the space of probability measures on $\mathcal{X}^n$, defined by

$$d_2(P, Q) \triangleq \inf_{\pi \in \Pi(P,Q)} \sup_{\alpha} \int \sum_{i=1}^n \alpha_i(y) 1_{\{x_i \neq y_i\}} \pi(\mathrm{d}x^n, \mathrm{d}y^n),$$

where the supremum is over all vector-valued positive functions $\alpha = (\alpha_1, \dots, \alpha_n) : \mathcal{X}^n \to \mathbb{R}^n$, such that

$$\mathbb{E}_Q \left[\|\alpha(Y^n)\|^2 \right] \le 1.$$

The main result of [145] goes as follows:

Theorem 42. The probability distribution P of X^n satisfies the following transportation cost inequality:

$$d_2(Q, P) \le \|\Delta\| \sqrt{2D(Q \| P)} \tag{3.246}$$

for all $Q \ll P$.

Let us examine some implications:

1. Let $\mathcal{X} = [0, 1]$. Then Theorem 42 implies that any probability measure P on the unit cube $\mathcal{X}^n = [0,1]^n$ satisfies the following Euclidean log-Sobolev inequality: for any smooth convex function $f : [0,1]^n \to \mathbb{R}$,

$$D\big(P^{(f)} \| P\big) \le 2\|\Delta\|^2 \, \mathbb{E}^{(f)} \left[\|\nabla f(X^n)\|^2 \right] \tag{3.247}$$

(see [145, Corollary 1]). The same method as the one we used to prove Proposition 8 and Theorem 22 can be applied to obtain from (3.247) the following concentration inequality for any convex function $f : [0,1]^n \to \mathbb{R}$ with $\|f\|_{\mathrm{Lip}} \le 1$:

$$\mathbb{P}\Big(f(X^n) \ge \mathbb{E} f(X^n) + r \Big) \le \exp\left(-\frac{r^2}{2\|\Delta\|^2} \right), \qquad \forall r \ge 0. \tag{3.248}$$

2. While (3.246) and its corollaries, (3.247) and (3.248), hold in full generality, these bounds are nontrivial only if the operator norm $\|\Delta\|$ is independent of n. This is the case whenever the dependence between the X_i's is sufficiently weak. For instance, if $X_1, \dots, X_n$ are independent, then $\Delta = I_{n \times n}$. In this case, (3.246) becomes

$$d_2(Q, P) \le \sqrt{2D(Q \| P)},$$

and we recover the usual concentration inequalities for Lipschitz functions. To see some examples with dependent random variables, suppose that $X_1, \ldots, X_n$ is a Markov chain, i.e., for each i, X_{i+1}^n is conditionally independent of X^{i-1} given X_i. In that case, from (3.245), the upper triangular part of Δ is given by

$$\Delta_{i,j} = \sup_{x_i, x_i'} \sqrt{\left\| P_{X_j|X_i=x_i} - P_{X_j|X_i=x_i'} \right\|_{\mathrm{TV}}}, \qquad i < j$$

and $\|\Delta\|$ will be independent of n under suitable ergodicity assumptions on the Markov chain $X_1, \ldots, X_n$. For instance, suppose that the Markov chain is homogeneous, i.e., the conditional probability distribution $P_{X_i|X_{i-1}}$ $(i > 1)$ is independent of i, and that

$$\sup_{x_i, x_i'} \|P_{X_{i+1}|X_i=x_i} - P_{X_{i+1}|X_i=x_i'}\|_{\mathrm{TV}} \le 2\rho$$

for some $\rho < 1$. Then it can be shown (see [145, Eq. (2.5)]) that

$$\begin{aligned} \|\Delta\| &\le \sqrt{2} \left(1 + \sum_{k=1}^{n-1} \rho^{k/2} \right) \\ &\le \frac{\sqrt{2}}{1 - \sqrt{\rho}}. \end{aligned}$$

More generally, following Marton [146], we will say that the (not necessarily homogeneous) Markov chain $X_1, \ldots, X_n$ is *contracting* if, for every i,

$$\delta_i \triangleq \sup_{x_i, x_i'} \|P_{X_{i+1}|X_i=x_i} - P_{X_{i+1}|X_i=x_i'}\|_{\mathrm{TV}} < 1.$$

In this case, it can be shown that

$$\|\Delta\| \le \frac{1}{1 - \delta^{1/2}}, \qquad \text{where } \delta \triangleq \max_{i=1,\ldots,n} \delta_i.$$

3.5.2 Marton's transportation cost inequalities for L^2 Wasserstein distance

Another approach to obtaining concentration results for dependent random variables, due to Marton [147, 148], relies on another measure of dependence that pertains to the sensitivity of the conditional distributions of X_i given $\bar{X}^i$ to the particular realization $\bar{x}^i$ of $\bar{X}^i$. The results of [147, 148] are set in the Euclidean space $\mathbb{R}^n$, and center around a transportation cost inequality for the L^2 Wasserstein distance

$$W_2(P, Q) \triangleq \inf_{X^n \sim P, Y^n \sim Q} \sqrt{\mathbb{E}\|X^n - Y^n\|^2}, \tag{3.249}$$

where $\|\cdot\|$ denotes the usual Euclidean norm.

We will state a particular special case of Marton's results (a more general development considers conditional distributions of $(X_i : i \in S)$ given $(X_j : j \in S^c)$ for a suitable system of sets $S \subset \{1, \ldots, n\}$). Let P be a probability measure on $\mathbb{R}^n$ which is absolutely continuous w.r.t. the Lebesgue measure. For each $x^n \in \mathbb{R}^n$ and each $i \in \{1, \ldots, n\}$ we denote by $\bar{x}^i$ the vector in $\mathbb{R}^{n-1}$ obtained by deleting the ith coordinate of x^n:

$$\bar{x}^i = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n).$$

Following Marton [147], we say that P is *δ-contractive*, with $0 < \delta < 1$, if for any $y^n, z^n \in \mathbb{R}^n$

$$\sum_{i=1}^n W_2^2(P_{X_i|\bar{X}^i=\bar{y}^i}, P_{X_i|\bar{X}^i=\bar{z}^i}) \le (1 - \delta)\|y^n - z^n\|_2. \tag{3.250}$$

Remark 46. Marton's contractivity condition (3.250) is closely related to the so-called *Dobrushin–Shlosman mixing condition* from mathematical statistical physics.

Theorem 43 (Marton [147, 148])**.** Suppose that P is absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}^n$ and δ-contractive, and that the conditional distributions $P_{X_i|\bar{X}^i}$, $i \in \{1, \dots, n\}$, have the following properties:

1. for each i, the function $x^n \mapsto p_{X_i|\bar{X}^i}(x_i|\bar{x}^i)$ is continuous, where $p_{X_i|\bar{X}^{i-1}}(\cdot|\bar{x}^i)$ denotes the univariate probability density function of $P_{X_i|\bar{X}^i=\bar{x}^i}$

2. for each i and each $\bar{x}^i \in \mathbb{R}^{n-1}$, $P_{X_i|\bar{X}^i=\bar{x}^{i-1}}$ satisfies $T_2(c)$ w.r.t. the L^2 Wasserstein distance (3.249) (cf. Definition 6)

Then for any probability measure Q on $\mathbb{R}^n$ we have

$$W_2(Q,P) \le \left(\frac{K}{\sqrt{\delta}} + 1\right) \sqrt{2cD(Q\|P)}, \tag{3.251}$$

where $K > 0$ is an absolute constant. In other words, any P satisfying the conditions of the theorem admits a $T_2(c')$ inequality with $c' = (K/\sqrt{\delta} + 1)^2 c$.

The contractivity criterion (3.250) is not easy to verify in general. Let us mention one sufficient condition [147]. Let p denote the probability density of P, and suppose that it takes the form

$$p(x^n) = \frac{1}{Z} \exp\left(-\Psi(x^n)\right) \tag{3.252}$$

for some C^2 function $\Psi : \mathbb{R}^n \to \mathbb{R}$, where Z is the normalization factor. For any $x^n, y^n \in \mathbb{R}^n$, let us define a matrix $B(x^n, y^n) \in \mathbb{R}^{n \times n}$ by

$$B_{ij}(x^n, y^n) \triangleq \begin{cases} \nabla^2_{ij} \Psi(x_i \odot \bar{y}^i), & i \neq j \\ 0, & i = j \end{cases} \tag{3.253}$$

where $\nabla^2_{ij} F$ denotes the (i,j) entry of the Hessian matrix of $F \in C^2(\mathbb{R}^n)$, and $x_i \odot \bar{y}^i$ denotes the n-tuple obtained by replacing the deleted ith coordinate in $\bar{y}^i$ with x_i:

$$x_i \odot \bar{y}^i = (y_1, \dots, y_{i-1}, x_i, y_{i+1}, \dots, y_n).$$

For example, if Ψ is a sum of one-variable and two-variable terms

$$\Psi(x^n) = \sum_{i=1}^n V_i(x_i) + \sum_{i<j} b_{ij} x_i x_j$$

for some smooth functions $V_i : \mathbb{R} \to \mathbb{R}$ and some constants $b_{ij} \in \mathbb{R}$, which is often the case in statistical physics, then the matrix B is independent of x^n, y^n, and has off-diagonal entries b_{ij}, $i \neq j$. Then (see Theorem 2 in [147]) the conditions of Theorem 43 will be satisfied, provided the following holds:

1. For each i and each $\bar{x}^i \in \mathbb{R}^{n-1}$, the conditional probability distributions $P_{X_i|\bar{X}^i=\bar{x}^i}$ satisfy the Euclidean log-Sobolev inequality

$$D(Q\|P_{X_i|\bar{X}^i=\bar{x}^i}) \le \frac{c}{2} I(Q\|P_{X_i|\bar{X}^i=\bar{x}^i}),$$

where $I(\cdot\|\cdot)$ is the Fisher information distance, cf. (3.37) for the definition.

2. The operator norms of $B(x^n, y^n)$ are uniformly bounded as

$$\sup_{x^n, y^n} \|B(x^n, y^n)\|^2 \le \frac{1-\delta}{c^2}.$$

We also refer the reader to more recent follow-up work by Marton [149, 150], which further elaborates on the theme of studying the concentration properties of dependent random variables by focusing on the conditional probability distributions $P_{X_i|\bar{X}^i}$, $i = 1, \dots, n$. These papers describe sufficient conditions on the joint distribution P of $X_1, \dots, X_n$, such that, for any other distribution Q,

$$D(Q\|P) \le K(P) \cdot D^{-}(Q\|P), \tag{3.254}$$

where $D^{-}(\cdot\|\cdot)$ is the erasure divergence (cf. (3.22) for the definition), and the P-dependent constant $K(P) > 0$ is controlled by suitable contractivity properties of P. At this point, the utility of a tensorization inequality like (3.254) should be clear: each term in the erasure divergence

$$D^{-}(Q\|P) = \sum_{i=1}^{n} D(Q_{X_i|\bar{X}^i} \| P_{X_i|\bar{X}^i} | Q_{\bar{X}^i})$$

can be handled by appealing to appropriate log-Sobolev inequalities or transportation-cost inequalities for probability measures on $\mathcal{X}$ (indeed, one can just treat $P_{X_i|\bar{X}^i = \bar{x}^i}$ for each fixed $\bar{x}^i$ as a probability measure on $\mathcal{X}$, in just the same way as with P_{X_i} before), and then these "one-dimensional" bounds can be assembled together to derive concentration for the original "n-dimensional" distribution.

3.6 Applications in information theory and related topics

3.6.1 The "blowing up" lemma and strong converses

The first explicit invocation of the concentration of measure phenomenon in an information-theoretic context appears in the work of Ahlswede et al. [60, 61]. These authors have shown that the following result, now known as the "blowing up lemma" (see, e.g., [151, Lemma 1.5.4]), provides a versatile tool for proving strong converses in a variety of scenarios, including some multiterminal problems:

Lemma 14. For every two finite sets $\mathcal{X}$ and $\mathcal{Y}$ and every positive sequence $\varepsilon_n \to 0$, there exist positive sequences $\delta_n, \eta_n \to 0$, such that the following holds: For every discrete memoryless channel (DMC) with input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$, and transition probabilities $T(y|x), x \in \mathcal{X}, y \in \mathcal{Y}$, and every $n \in \mathbb{N}$, $x^n \in \mathcal{X}^n$, and $B \subseteq \mathcal{Y}^n$,

$$T^n(B|x^n) \ge \exp(-n\varepsilon_n) \qquad \Longrightarrow \qquad T^n(B_{n\delta_n}|x^n) \ge 1 - \eta_n. \tag{3.255}$$

Here, for an arbitrary $B \subseteq \mathcal{Y}^n$ and $r > 0$, the set B_r denotes the *r-blowup* of B (see the definition in (3.162)) w.r.t. the Hamming metric

$$d_n(y^n, u^n) \triangleq \sum_{i=1}^{n} 1_{\{y_i \ne u_i\}}, \qquad \forall y^n, u^n \in \mathcal{Y}^n.$$

The proof of the blowing-up lemma, given in [60], was rather technical and made use of a very delicate isoperimetric inequality for discrete probability measures on a Hamming space, due to Margulis [152]. Later, the same result was obtained by Marton [62] using purely information-theoretic methods. We will use a sharper, "nonasymptotic" version of the blowing-up lemma, which is more in the spirit of the modern viewpoint on the concentration of measure (cf. Marton's follow-up paper [49]):

Lemma 15. Let $X_1, \ldots, X_n$ be n independent random variables taking values in a finite set $\mathcal{X}$. Then, for any $A \subseteq \mathcal{X}^n$ with $P_{X^n}(A) > 0$,

$$P_{X^n}(A_r) \geq 1 - \exp\left[-\frac{2}{n}\left(r - \sqrt{\frac{n}{2}\ln\left(\frac{1}{P_{X^n}(A)}\right)}\right)^2\right], \qquad \forall\, r > \sqrt{\frac{n}{2}\ln\left(\frac{1}{P_{X^n}(A)}\right)}. \tag{3.256}$$

Proof. The proof of Lemma 15 is similar to the proof of Proposition 9, as is shown in the following: Consider the L^1 Wasserstein metric on $\mathcal{P}(\mathcal{X}^n)$ induced by the Hamming metric d_n on $\mathcal{X}^n$, i.e., for any $P_n, Q_n \in \mathcal{P}(\mathcal{X}^n)$,

$$\begin{aligned} W_1(P_n, Q_n) &\triangleq \inf_{X^n \sim P_n,\, Y^n \sim Q_n} \mathbb{E}\big[d_n(X^n, Y^n)\big] \\ &= \inf_{X^n \sim P_n,\, Y^n \sim Q_n} \mathbb{E}\left[\sum_{i=1}^n 1_{\{X_i \neq Y_i\}}\right] \\ &= \inf_{X^n \sim P_n,\, Y^n \sim Q_n} \sum_{i=1}^n \Pr(X_i \neq Y_i). \end{aligned}$$

Let P_n denote the product measure $P_{X^n} = P_{X_1} \otimes \ldots \otimes P_{X_n}$. By Pinsker's inequality, any $\mu \in \mathcal{P}(\mathcal{X})$ satisfies $T_1(1/4)$ on $(\mathcal{X}, d)$ where $d = d_1$ is the Hamming metric. By Proposition 11, the product measure P_n satisfies $T_1(n/4)$ on the product space $(\mathcal{X}^n, d_n)$, i.e., for any $\mu_n \in \mathcal{P}(\mathcal{X}^n)$,

$$W_1(\mu_n, P_n) \leq \sqrt{\frac{n}{2} D(\mu_n \| P_n)}. \tag{3.257}$$

For any set $C \subseteq \mathcal{X}^n$ with $P_n(C) > 0$, let $P_{n,C}$ denote the conditional probability measure $P_n(\cdot|C)$. Then, it follows that (see (3.193))

$$D\big(P_{n,C} \big\| P_n\big) = \ln\left(\frac{1}{P_n(C)}\right). \tag{3.258}$$

Now, given any $A \subseteq \mathcal{X}^n$ with $P_n(A) > 0$ and any $r > 0$, consider the probability measures $Q_n = P_{n,A}$ and $\bar{Q}_n = P_{n,A_r^c}$. Then

$$\begin{aligned} W_1(Q_n, \bar{Q}_n) &\leq W_1(Q_n, P_n) + W_1(\bar{Q}_n, P_n) &(3.259)\\ &\leq \sqrt{\frac{n}{2} D(Q_n \| P_n)} + \sqrt{\frac{n}{2} D(\bar{Q}_n \| P_n)} &(3.260)\\ &= \sqrt{\frac{n}{2}\ln\left(\frac{1}{P_n(A)}\right)} + \sqrt{\frac{n}{2}\ln\left(\frac{1}{1 - P_n(A_r)}\right)} &(3.261) \end{aligned}$$

where (3.259) uses the triangle inequality, (3.260) follows from (3.257), and (3.261) uses (3.258). Following the same reasoning that leads to (3.195), it follows that

$$W_1(Q_n, \bar{Q}_n) = W_1(P_{n,A}, P_{n,A_r^c}) \geq d_n(A, A_r^c) \geq r.$$

Using this to bound the left-hand side of (3.259) from below, we obtain (3.256). □

We can now easily prove the blowing-up lemma (see Lemma 14). To this end, given a positive sequence $\{\varepsilon_n\}_{n=1}^\infty$ that tends to zero, let us choose a positive sequence $\{\delta_n\}_{n=1}^\infty$ such that

$$\delta_n > \sqrt{\frac{\varepsilon_n}{2}}, \quad \delta_n \xrightarrow{n \to \infty} 0, \quad \eta_n \triangleq \exp\left(-2n\left(\delta_n - \sqrt{\frac{\varepsilon_n}{2}}\right)^2\right) \xrightarrow{n \to \infty} 0.$$

These requirements can be satisfied, e.g., by the setting

$$\delta_n \triangleq \sqrt{\frac{\varepsilon_n}{2}} + \sqrt{\frac{\alpha \ln n}{n}}, \quad \eta_n = \frac{1}{n^{2\alpha}}, \quad \forall\, n \in \mathbb{N}$$

where $\alpha > 0$ can be made arbitrarily small. Using this selection for $\{\delta_n\}_{n=1}^{\infty}$ in (3.256), we get (3.255) with the r_n-blowup of the set B where $r_n \triangleq n\delta_n$. Note that the above selection does not depend on the transition probabilities of the DMC with input $\mathcal{X}$ and output $\mathcal{Y}$ (the correspondence between Lemmas 14 and 15 is given by $P_{X^n} = T^n(\cdot|x^n)$ where $x^n \in \mathcal{X}^n$ is arbitrary).

We are now ready to demonstrate how the blowing-up lemma can be used to obtain strong converses. Following [151], from this point on, we will use the notation $T : \mathcal{U} \to \mathcal{V}$ for a DMC with input alphabet $\mathcal{U}$, output alphabet $\mathcal{V}$, and transition probabilities $T(v|u), u \in \mathcal{U}, v \in \mathcal{V}$.

We first consider the problem of characterizing the capacity region of a degraded broadcast channel (DBC). Let $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ be finite sets. A DBC is specified by a pair of DMC's $T_1 : \mathcal{X} \to \mathcal{Y}$ and $T_2 : \mathcal{X} \to \mathcal{Z}$ where there exists a DMC $T_3 : \mathcal{Y} \to \mathcal{Z}$ such that

$$T_2(z|x) = \sum_{y \in \mathcal{Y}} T_3(z|y) T_1(y|x), \qquad \forall x \in \mathcal{X}, z \in \mathcal{Z}. \tag{3.262}$$

(More precisely, this is an instance of a *stochastically degraded* broadcast channel – see, e.g., [81, Section 5.6] and [153, Chapter 5]). Given $n, M_1, M_2 \in \mathbb{N}$, an (n, M_1, M_2)*-code* $\mathcal{C}$ for the DBC (T_1, T_2) consists of the following objects:

1. an *encoding map* $f_n : \{1, \ldots, M_1\} \times \{1, \ldots, M_2\} \to \mathcal{X}^n$;
2. a collection $\mathcal{D}_1$ of M_1 disjoint *decoding sets* $D_{1,i} \subset \mathcal{Y}^n$, $1 \le i \le M_1$; and, similarly,
3. a collection $\mathcal{D}_2$ of M_2 disjoint decoding sets $D_{2,j} \subset \mathcal{Z}^n$, $1 \le j \le M_2$.

Given $0 < \varepsilon_1, \varepsilon_2 \le 1$, we say that $\mathcal{C} = (f_n, \mathcal{D}_1, \mathcal{D}_2)$ is an $(n, M_1, M_2, \varepsilon_1, \varepsilon_2)$*-code* if

$$\max_{1 \le i \le M_1} \max_{1 \le j \le M_2} T_1^n\Big(D_{1,i}^c \Big| f_n(i,j)\Big) \le \varepsilon_1$$
$$\max_{1 \le i \le M_1} \max_{1 \le j \le M_2} T_2^n\Big(D_{2,j}^c \Big| f_n(i,j)\Big) \le \varepsilon_2.$$

In other words, we are using the maximal probability of error criterion. It should be noted that, although for some multiuser channels the capacity region w.r.t. the maximal probability of error is strictly smaller than the capacity region w.r.t. the average probability of error [154], these two capacity regions are identical for broadcast channels [155]. We say that a pair of rates (R_1, R_2) (in nats per channel use) is $(\varepsilon_1, \varepsilon_2)$*-achievable* if for any $\delta > 0$ and sufficiently large n, there exists an $(n, M_1, M_2, \varepsilon_1, \varepsilon_2)$-code with

$$\frac{1}{n} \ln M_k \ge R_k - \delta, \qquad k = 1, 2.$$

Likewise, we say that (R_1, R_2) is *achievable* if it is $(\varepsilon_1, \varepsilon_2)$-achievable for all $0 < \varepsilon_1, \varepsilon_2 \le 1$. Now let $\mathcal{R}(\varepsilon_1, \varepsilon_2)$ denote the set of all $(\varepsilon_1, \varepsilon_2)$-achievable rates, and let $\mathcal{R}$ denote the set of all achievable rates. Clearly,

$$\mathcal{R} = \bigcap_{(\varepsilon_1, \varepsilon_2) \in (0,1]^2} \mathcal{R}(\varepsilon_1, \varepsilon_2).$$

The following result was proved by Ahlswede and Körner [156]:

Theorem 44. A rate pair (R_1, R_2) is achievable for the DBC (T_1, T_2) if and only if there exist random variables $U \in \mathcal{U}, X \in \mathcal{X}, Y \in \mathcal{Y}, Z \in \mathcal{Z}$ such that $U \to X \to Y \to Z$ is a Markov chain, $P_{Y|X} = T_1$, $P_{Z|Y} = T_3$ (see (3.262)), and

$$R_1 \le I(X;Y|U), \quad R_2 \le I(U;Z).$$

Moreover, the domain $\mathcal{U}$ of U can be chosen so that $|\mathcal{U}| \le \min\{|\mathcal{X}|, |\mathcal{Y}|, |\mathcal{Z}|\}$.

The *strong converse* for the DBC, due to Ahlswede, Gács and Körner [60], states that allowing for nonvanishing probabilities of error does not enlarge the achievable region:

Theorem 45 (Strong converse for the DBC)**.**

$$\mathcal{R}(\varepsilon_1.\varepsilon_2) = \mathcal{R}, \qquad \forall(\varepsilon_1, \varepsilon_2) \in (0,1]^2.$$

Before proceeding with the formal proof of this theorem, we briefly describe the way in which the blowing up lemma enters the picture. The main idea is that, given any code, one can "blow up" the decoding sets in such a way that the probability of decoding error can be as small as one desires (for large enough n). Of course, the blown-up decoding sets are no longer disjoint, so the resulting object is no longer a code according to the definition given earlier. On the other hand, the blowing-up operation transforms the original code into a *list code* with a subexponential list size, and one can use Fano's inequality to get nontrivial converse bounds.

Proof (Theorem 45). Let $\widetilde{\mathcal{C}} = (f_n, \widetilde{\mathcal{D}}_1, \widetilde{\mathcal{D}}_2)$ be an arbitrary $(n, M_1, M_2, \widetilde{\varepsilon}_1, \widetilde{\varepsilon}_2)$-code for the DBC (T_1, T_2) with

$$\widetilde{\mathcal{D}}_1 = \left\{\widetilde{D}_{1,i}\right\}_{i=1}^{M_1} \qquad \text{and} \qquad \widetilde{\mathcal{D}}_2 = \left\{\widetilde{D}_{2,j}\right\}_{j=1}^{M_2}.$$

Let $\{\delta_n\}_{n=1}^{\infty}$ be a sequence of positive reals, such that

$$\delta_n \to 0, \quad \sqrt{n}\delta_n \to \infty \qquad \text{as } n \to \infty.$$

For each $i \in \{1, \dots, M_1\}$ and $j \in \{1, \dots, M_2\}$, define the "blown-up" decoding sets

$$D_{1,i} \triangleq \left[\widetilde{D}_{1,i}\right]_{n\delta_n} \qquad \text{and} \qquad D_{2,j} \triangleq \left[\widetilde{D}_{2,j}\right]_{n\delta_n}.$$

By hypothesis, the decoding sets in $\widetilde{\mathcal{D}}_1$ and $\widetilde{\mathcal{D}}_2$ are such that

$$\min_{1\le i\le M_1} \min_{1\le j\le M_2} T_1^n\Big(\widetilde{D}_{1,i}\Big|f_n(i,j)\Big) \ge 1 - \widetilde{\varepsilon}_1$$
$$\min_{1\le i\le M_1} \min_{1\le j\le M_2} T_2^n\Big(\widetilde{D}_{2,j}\Big|f_n(i,j)\Big) \ge 1 - \widetilde{\varepsilon}_2.$$

Therefore, by Lemma 15, we can find a sequence $\varepsilon_n \to 0$, such that

$$\min_{1\le i\le M_1} \min_{1\le j\le M_2} T_1^n\Big(D_{1,i}\Big|f_n(i,j)\Big) \ge 1 - \varepsilon_n \tag{3.263a}$$
$$\min_{1\le i\le M_1} \min_{1\le j\le M_2} T_2^n\Big(D_{2,j}\Big|f_n(i,j)\Big) \ge 1 - \varepsilon_n \tag{3.263b}$$

Let $\mathcal{D}_1 = \{D_{1,i}\}_{i=1}^{M_1}$, and $\mathcal{D}_2 = \{D_{2,j}\}_{j=1}^{M_2}$. We have thus constructed a triple $(f_n, \mathcal{D}_1, \mathcal{D}_2)$ satisfying (3.263). Note, however, that this new object is not a code because the blown-up sets $D_{1,i} \subseteq \mathcal{Y}^n$ are not disjoint, and the same holds for the blow-up sets $\{D_{2,j}\}$. On the other hand, each given n-tuple $y^n \in \mathcal{Y}^n$

belongs to a small number of the $D_{1,i}$'s, and the same applies to $D_{2,j}$'s. More precisely, let us define for each $y^n \in \mathcal{Y}^n$ the set

$$\mathcal{N}_1(y^n) \triangleq \{i : y^n \in D_{1,i}\},$$

and similarly for $\mathcal{N}_2(z^n)$, $z^n \in \mathcal{Z}^n$. Then a simple combinatorial argument (see [60, Lemma 5 and Eq. (37)] for details) can be used to show that there exists a sequence $\{\eta_n\}_{n=1}^{\infty}$ of positive reals, such that $\eta_n \to 0$ and

$$|\mathcal{N}_1(y^n)| \le |\mathcal{B}_{n\delta_n}(y^n)| \le \exp(n\eta_n), \qquad \forall y^n \in \mathcal{Y}^n \tag{3.264a}$$

$$|\mathcal{N}_2(z^n)| \le |\mathcal{B}_{n\delta_n}(z^n)| \le \exp(n\eta_n), \qquad \forall z^n \in \mathcal{Z}^n \tag{3.264b}$$

where, for any $y^n \in \mathcal{Y}^n$ and any $r \ge 0$, $\mathcal{B}_r(y^n) \subseteq \mathcal{Y}^n$ denotes the ball of d_n-radius r centered at y^n:

$$\mathcal{B}_r(y^n) \triangleq \{v^n \in \mathcal{Y}^n : d_n(v^n, y^n) \le r\} \equiv \{y^n\}_r$$

(the last expression denotes the r-blowup of the singleton set $\{y^n\}$).

We are now ready to apply Fano's inequality, just as in [156]. Specifically, let U have a uniform distribution over $\{1, \dots, M_2\}$, and let $X^n \in \mathcal{X}^n$ have a uniform distribution over the set $\mathcal{T}(U)$, where for each $j \in \{1, \dots, M_2\}$ we let

$$\mathcal{T}(j) \triangleq \{f_n(i,j) : 1 \le i \le M_1\}.$$

Finally, let $Y^n \in \mathcal{Y}^n$ and $Z^n \in \mathcal{Z}^n$ be generated from X^n via the DMC's T_1^n and T_2^n, respectively. Now, for each $z^n \in \mathcal{Z}^n$, consider the error event

$$E_n(z^n) \triangleq \{U \notin \mathcal{N}_2(z^n)\}, \quad \forall\, z^n \in \mathcal{Z}^n$$

and let $\zeta_n \triangleq \mathbb{P}(E_n(Z^n))$. Then, using a modification of Fano's inequality for list decoding (see Appendix 3.C) together with (3.264), we get

$$H(U|Z^n) \le h(\zeta_n) + (1-\zeta_n)n\eta_n + \zeta_n \ln M_2. \tag{3.265}$$

On the other hand, $\ln M_2 = H(U) = I(U;Z^n) + H(U|Z^n)$, so

$$\begin{aligned}\frac{1}{n}\ln M_2 &\le \frac{1}{n}\Big[I(U;Z^n) + h(\zeta_n) + \zeta_n \ln M_2\Big] + (1-\zeta_n)\eta_n \\ &= \frac{1}{n}I(U;Z^n) + o(1),\end{aligned}$$

where the second step uses the fact that, by (3.263), $\zeta_n \le \varepsilon_n$, which converges to zero. Using a similar argument, we can also prove that

$$\frac{1}{n}\ln M_1 \le \frac{1}{n}I(X^n;Y^n|U) + o(1).$$

By the weak converse for the DBC [156], the pair (R_1, R_2) with $R_1 = \frac{1}{n}I(X^n;Y^n|U)$ and $R_2 = \frac{1}{n}I(U;Z^n)$ belongs to the achievable region $\mathcal{R}$. Since any element of $\mathcal{R}(\varepsilon_1,\varepsilon_2)$ can be expressed as a limit of rates $\left(\frac{1}{n}\ln M_1, \frac{1}{n}\ln M_2\right)$, and since the achievable region $\mathcal{R}$ is closed, we conclude that $\mathcal{C}(\varepsilon_1,\varepsilon_2) \subseteq \mathcal{C}$ for all $\varepsilon_1, \varepsilon_2 \in (0,1]$, and Theorem 45 is proved. □

Our second example of the use of the blowing-up lemma to prove a strong converse is a bit more sophisticated, and concerns the problem of lossless source coding with side information. Let $\mathcal{X}$ and $\mathcal{Y}$ be finite sets, and $\{(X_i, Y_i)\}_{i=1}^{\infty}$ be a sequence of i.i.d. samples drawn from a given joint distribution $P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. The $\mathcal{X}$-valued and the $\mathcal{Y}$-valued parts of this sequence are observed by two independent encoders. An (n, M_1, M_2)-code is a triple $\mathcal{C} = \big(f_n^{(1)}, f_n^{(2)}, g_n\big)$, where $f_n^{(1)} : \mathcal{X}^n \to \{1, \dots, M_1\}$ and $f_n^{(2)} : \mathcal{Y}^n \to \{1, \dots, M_2\}$ are the encoding maps and $g_n : \{1, \dots, M_1\} \times \{1, \dots, M_2\} \to \mathcal{Y}^n$ is the decoding map. The decoder observes

$$J_n^{(1)} = f_n^{(1)}(X^n) \qquad \text{and} \qquad J_n^{(2)} = f_n^{(2)}(Y^n)$$

and wishes to reconstruct Y^n with a small probability of error. The reconstruction is given by

$$\begin{aligned} \widehat{Y}^n &= g_n\big(J_n^{(1)}, J_n^{(2)}\big) \\ &= g_n\big(f_n^{(1)}(X^n), f_n^{(2)}(Y^n)\big). \end{aligned}$$

We say that $\mathcal{C} = \Big(f_n^{(1)}, f_n^{(2)}, g_n\Big)$ is an $(n, M_1, M_2, \varepsilon)$-*code* if

$$\mathbb{P}\left(\widehat{Y}^n \neq Y^n\right) = \mathbb{P}\left(g_n\big(f_n^{(1)}(X^n), f_n^{(2)}(Y^n)\big) \neq Y^n\right) \le \varepsilon. \tag{3.266}$$

We say that a rate pair (R_1, R_2) is ε-*achievable* if, for any $\delta > 0$ and sufficiently large $n \in \mathbb{N}$, there exists an $(n, M_1, M_2, \varepsilon)$-code $\mathcal{C}$ with

$$\frac{1}{n} \ln M_k \le R_k + \delta, \qquad k = 1, 2. \tag{3.267}$$

A rate pair (R_1, R_2) is *achievable* if it is ε-achievable for all $\varepsilon \in (0, 1]$. Again, let $\mathcal{R}(\varepsilon)$ (resp., $\mathcal{R}$) denote the set of all ε-achievable (resp., achievable) rate pairs. Clearly,

$$\mathcal{R} = \bigcap_{\varepsilon \in (0,1]} \mathcal{R}(\varepsilon).$$

The following characterization of the achievable region was obtained in [156]:

Theorem 46. A rate pair (R_1, R_2) is achievable if and only if there exist random variables $U \in \mathcal{U}$, $X \in \mathcal{X}$, $Y \in \mathcal{Y}$, such that $U \to X \to Y$ is a Markov chain, (X, Y) has the given joint distribution P_{XY}, and

$$\begin{aligned} R_1 &\ge I(X; U) \\ R_2 &\ge H(Y|U) \end{aligned}$$

Moreover, the domain $\mathcal{U}$ of U can be chosen so that $|\mathcal{U}| \le |\mathcal{X}| + 2$.

Our goal is to prove the corresponding *strong converse* (originally established in [60]), which states that allowing for a nonvanishing error probability, as in (3.266), does not asymptotically enlarge the achievable region:

Theorem 47 (Strong converse for source coding with side information)**.**

$$\mathcal{R}(\varepsilon) = \mathcal{R}, \qquad \forall \varepsilon \in (0, 1].$$

In preparation for the proof of Theorem 47, we need to introduce some additional terminology and definitions. Given two finite sets $\mathcal{U}$ and $\mathcal{V}$, a DMC $S : \mathcal{U} \to \mathcal{V}$, and a parameter $\eta \in [0, 1]$, we say, following [151], that a set $B \subseteq \mathcal{V}$ is an η-*image of* $u \in \mathcal{U}$ *under* S if $S(B|u) \ge \eta$. For any $B \subseteq \mathcal{V}$, let $\mathcal{D}_\eta(B; S) \subseteq \mathcal{U}$ denote the set of all $u \in \mathcal{U}$, such that B is an η-image of u under S:

$$\mathcal{D}_\eta(B; S) \triangleq \Big\{u \in \mathcal{U} : S(B|u) \ge \eta\Big\}.$$

Now, given $P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, let $T : \mathcal{X} \to \mathcal{Y}$ be the DMC corresponding to the conditional probability distribution $P_{Y|X}$. Finally, given a strictly positive probability measure $Q_Y \in \mathcal{P}(\mathcal{Y})$ and the parameters $c \ge 0$ and $\varepsilon \in (0, 1]$, we define

$$\widehat{\Gamma}_n(c, \varepsilon; Q_Y) \triangleq \min_{B \subseteq \mathcal{Y}^n} \left\{ \frac{1}{n} \ln Q_Y^n(B) : \frac{1}{n} \ln P_X^n \left(\mathcal{D}_{1-\varepsilon}(B; T^n) \cap \mathcal{T}^n_{[X]} \right) \ge -c \right\} \tag{3.268}$$

where $\mathcal{T}^n_{[X]} \subset \mathcal{X}^n$ denotes the typical set induced by the marginal distribution P_X.

Theorem 48. For any $c \ge 0$ and any $\varepsilon \in (0, 1]$,

$$\lim_{n \to \infty} \widehat{\Gamma}_n(c, \varepsilon; Q_Y) = \Gamma(c; Q_Y), \tag{3.269}$$

where

$$\Gamma(c; Q_Y) \triangleq - \max_{\mathcal{U}: |\mathcal{U}| \le |\mathcal{X}|+2} \max_{U \in \mathcal{U}} \left\{ D(P_{Y|U} \| Q_Y | P_U) : U \to X \to Y;\, I(X; U) \le c \right\}. \tag{3.270}$$

Moreover, the function $c \mapsto \Gamma(c; Q_Y)$ is continuous.

Proof. The proof consists of two major steps. The first is to show that (3.269) holds for $\varepsilon = 0$, and that the limit $\Gamma(c; Q_Y)$ is equal to (3.270). We omit the details of this step and instead refer the reader to the original paper by Ahlswede, Gács and Körner [60]. The second step, which actually relies on the blowing-up lemma, is to show that

$$\lim_{n \to \infty} \left[\widehat{\Gamma}_n(c, \varepsilon; Q_Y) - \widehat{\Gamma}_n(c, \varepsilon'; Q_Y) \right] = 0 \tag{3.271}$$

for any $\varepsilon, \varepsilon' \in (0, 1]$. To that end, let us fix an ε and choose a sequence of positive reals, such that

$$\delta_n \to 0 \text{ and } \sqrt{n}\delta_n \to \infty \qquad \text{as } n \to \infty. \tag{3.272}$$

For a fixed n, let us consider any set $B \subseteq \mathcal{Y}^n$. If $T^n(B|x^n) \ge 1 - \varepsilon$ for some $x^n \in \mathcal{X}^n$, then by Lemma 15

$$\begin{aligned} T^n(B_{n\delta_n}|x^n) &\ge 1 - \exp\left[-\frac{2}{n} \left(n\delta_n - \sqrt{\frac{n}{2} \ln\left(\frac{1}{1-\varepsilon} \right)} \right)^2 \right] \\ &= 1 - \exp\left[-2 \left(\sqrt{n}\,\delta_n - \sqrt{\frac{1}{2} \ln\left(\frac{1}{1-\varepsilon} \right)} \right)^2 \right] \\ &\triangleq 1 - \varepsilon_n. \end{aligned} \tag{3.273}$$

Owing to (3.272), the right-hand side of (3.273) will tend to 1 as $n \to \infty$, which implies that, for all large n,

$$\mathcal{D}_{1-\varepsilon_n}(B_{n\delta_n}; T^n) \cap \mathcal{T}^n_{[X]} \supseteq \mathcal{D}_{1-\varepsilon}(B; T^n) \cap \mathcal{T}^n_{[X]}. \tag{3.274}$$

On the other hand, since Q_Y is strictly positive,

$$\begin{aligned} Q_Y^n(B_{n\delta_n}) &= \sum_{y^n \in B_{n\delta_n}} Q_Y^n(y^n) \\ &\le \sum_{y^n \in B} Q_Y^n\left(\mathcal{B}_{n\delta_n}(y^n) \right) \\ &\le \sup_{y^n \in \mathcal{Y}^n} \frac{Q_Y^n\left(\mathcal{B}_{n\delta_n}(y^n) \right)}{Q_Y^n(y^n)} \sum_{y^n \in B} Q_Y^n(y^n) \\ &= \sup_{y^n \in \mathcal{Y}^n} \frac{Q_Y^n\left(\mathcal{B}_{n\delta_n}(y^n) \right)}{Q_Y^n(y^n)} \cdot Q_Y^n(B). \end{aligned}$$

Using this together with the fact that

$$\lim_{n\to\infty} \frac{1}{n} \ln \sup_{y^n \in \mathcal{Y}^n} \frac{Q_Y^n\left(\mathcal{B}_{n\delta_n}(y^n)\right)}{Q_Y^n(y^n)} = 0$$

(see [60, Lemma 5]), we can write

$$\lim_{n\to\infty} \sup_{B \subseteq \mathcal{Y}^n} \frac{1}{n} \ln \frac{Q_Y^n\left(B_{n\delta_n}\right)}{Q_Y^n(B)} = 0. \tag{3.275}$$

From (3.274) and (3.275), it follows that

$$\lim_{n\to\infty} \left[\widehat{\Gamma}_n(c, \varepsilon; Q_Y) - \widehat{\Gamma}_n(c, \varepsilon_n; Q_Y)\right] = 0.$$

This completes the proof of Theorem 48. □

We are now ready to prove Theorem 47. Let $\mathcal{C} = \left(f_n^{(1)}, f_n^{(2)}, g_n\right)$ be an arbitrary $(n, M_1, M_2, \varepsilon)$-code. For a given index $j \in \{1, \ldots, M_1\}$, we define the set

$$B(j) \triangleq \left\{y^n \in \mathcal{Y}^n : y^n = g_n\left(j, f_n^{(2)}(y^n)\right)\right\},$$

which consists of all $y^n \in \mathcal{Y}^n$ that are correctly decoded for any $x^n \in \mathcal{X}^n$ such that $f_n^{(1)}(x^n) = j$. Using this notation, we can write

$$\mathbb{E}\left[T^n\big(B(f_n^{(1)}(X^n))\big|X^n\big)\right] \ge 1 - \varepsilon. \tag{3.276}$$

If we define the set

$$A_n \triangleq \left\{x^n \in \mathcal{X}^n : T^n\big(B(f_n^{(1)}(x^n))\big|x^n\big) \ge 1 - \sqrt{\varepsilon}\right\},$$

then, using the so-called "reverse Markov inequality"[2] and (3.276), we see that

$$\begin{aligned} P_X^n(A_n) &= 1 - P_X^n(A_n^c) \\ &= 1 - P_X^n\Bigg(\underbrace{T^n\Big(B(f_n^{(1)}(X^n))\,|\,X^n\Big)}_{\le 1} < 1 - \sqrt{\varepsilon}\Bigg) \\ &\ge 1 - \frac{1 - \mathbb{E}\left[T^n\big(B(f_n^{(1)}(X^n))\big|X^n\big)\right]}{1 - (1 - \sqrt{\varepsilon})} \\ &\ge 1 - \frac{1 - (1 - \varepsilon)}{\sqrt{\varepsilon}} = 1 - \sqrt{\varepsilon}. \end{aligned}$$

Consequently, for all sufficiently large n, we have

$$P_X^n\left(A_n \cap \mathcal{T}_{[X]}^n\right) \ge 1 - 2\sqrt{\varepsilon}.$$

[2] The reverse Markov inequality states that if Y is a random variable such that $Y \le b$ a.s. for some constant b, then for all $a < b$

$$\mathbb{P}(Y \le a) \le \frac{b - \mathbb{E}[Y]}{b - a}.$$

This implies, in turn, that there exists some $j^* \in f_n^{(1)}(\mathcal{X}^n)$, such that

$$P_X^n \left(\mathcal{D}_{1-\sqrt{\varepsilon}}(B(j^*)) \cap \mathcal{T}_{[X]}^n \right) \geq \frac{1-2\sqrt{\varepsilon}}{M_1}. \tag{3.277}$$

On the other hand,

$$M_2 = \left| f_n^{(2)}(Y^n) \right| \geq |B(j^*)|. \tag{3.278}$$

We are now in a position to apply Theorem 48. If we choose Q_Y to be the uniform distribution on $\mathcal{Y}$, then it follows from (3.277) and (3.278) that

$$\begin{aligned} \frac{1}{n} \ln M_2 &\geq \frac{1}{n} \ln |B(j^*)| \\ &= \frac{1}{n} \ln Q_Y^n(B(j^*)) + \ln |\mathcal{Y}| \\ &\geq \widehat{\Gamma}_n \left(-\frac{1}{n} \ln(1-2\sqrt{\varepsilon}) + \frac{1}{n} \ln M_1, \sqrt{\varepsilon};\, Q_Y \right) + \ln |\mathcal{Y}|. \end{aligned}$$

Using Theorem 48, we conclude that the bound

$$\frac{1}{n} \ln M_2 \geq \Gamma \left(-\frac{1}{n} \ln(1-2\sqrt{\varepsilon}) + \frac{1}{n} \ln M_1;\, Q_Y \right) + \ln |\mathcal{Y}| + o(1) \tag{3.279}$$

holds for any $(n, M_1, M_2, \varepsilon)$-code. If $(R_1, R_2) \in \mathcal{R}(\varepsilon)$, then there exists a sequence $\{\mathcal{C}_n\}_{n=1}^{\infty}$, where each $\mathcal{C}_n = \left(f_n^{(1)}, f_n^{(2)}, g_n \right)$ is an $(n, M_{1,n}, M_{2,n}, \varepsilon)$-code, and

$$\lim_{n\to\infty} \frac{1}{n} \ln M_{k,n} = R_k, \qquad k = 1, 2.$$

Using this in (3.279), together with the continuity of the mapping $c \mapsto \Gamma(c; Q_Y)$, we get

$$R_2 \geq \Gamma(R_1; Q_Y) + \ln |\mathcal{Y}|, \qquad \forall (R_1, R_2) \in \mathcal{R}(\varepsilon). \tag{3.280}$$

By definition of Γ in (3.270), there exists a triple $U \to X \to Y$ such that $I(X; U) \leq R_1$ and

$$\Gamma(R_1; Q_Y) = -D(P_{Y|U} \| Q_Y | P_U) = -\ln |\mathcal{Y}| + H(Y|U), \tag{3.281}$$

where the second equality is due to the fact that $U \to X \to Y$ is a Markov chain and Q_Y is the uniform distribution on $\mathcal{Y}$. Therefore, (3.280) and (3.281) imply that

$$R_2 \geq H(Y|U).$$

Consequently, the triple $(U, X, Y) \in \mathcal{R}$ by Theorem 46, and hence $\mathcal{R}(\varepsilon) \subseteq \mathcal{R}$ for all $\varepsilon > 0$. Since $\mathcal{R} \subseteq \mathcal{R}(\varepsilon)$ by definition, the proof of Theorem 47 is completed.

3.6.2 Empirical distributions of good channel codes with nonvanishing error probability

A more recent application of concentration of measure to information theory has to do with characterizing stochastic behavior of output sequences of good channel codes. On a conceptual level, the random coding argument originally used by Shannon, and many times since, to show the existence of good channel codes suggests that the input (resp., output) sequence of such a code should resemble, as much as possible, a typical realization of a sequence of i.i.d. random variables sampled from a capacity-achieving input (resp., output) distribution. For capacity-achieving sequences of codes with asymptotically vanishing probability of error, this intuition has been analyzed rigorously by Shamai and Verdú [157], who have proved the following remarkable statement [157, Theorem 2]: given a DMC $T : \mathcal{X} \to \mathcal{Y}$, any capacity-achieving sequence of channel codes with asymptotically vanishing probability of error (maximal or average) has the property that

$$\lim_{n\to\infty} \frac{1}{n} D(P_{Y^n} \| P^*_{Y^n}) = 0, \tag{3.282}$$

where for each n P_{Y^n} denotes the output distribution on $\mathcal{Y}^n$ induced by the code (assuming the messages are equiprobable), while $P^*_{Y^n}$ is the product of n copies of the single-letter capacity-achieving output distribution (see below for a more detailed exposition). In fact, the convergence in (3.282) holds not just for DMC's, but for arbitrary channels satisfying the condition

$$C = \lim_{n\to\infty} \frac{1}{n} \sup_{P_{X^n} \in \mathcal{P}(\mathcal{X}^n)} I(X^n; Y^n).$$

In a recent preprint [158], Polyanskiy and Verdú have extended the results of [157] and showed that (3.282) holds for codes with *nonvanishing* probability of error, provided one uses the maximal probability of error criterion and deterministic decoders.

In this section, we will present some of the results from [158] in the context of the material covered earlier in this chapter. To keep things simple, we will only focus on channels with finite input and output alphabets. Thus, let $\mathcal{X}$ and $\mathcal{Y}$ be finite sets, and consider a DMC $T : \mathcal{X} \to \mathcal{Y}$. The capacity C is given by solving the optimization problem

$$C = \max_{P_X \in \mathcal{P}(\mathcal{X})} I(X; Y),$$

where X and Y are related via T. Let $P^*_X \in \mathcal{P}(\mathcal{X})$ be any capacity-achieving input distribution (there may be several). It can be shown ([159, 160]) that the corresponding output distribution $P^*_Y \in \mathcal{P}(\mathcal{Y})$ is unique, and that for any $n \in \mathbb{N}$, the product distribution $P^*_{Y^n} \equiv (P^*_Y)^{\otimes n}$ has the key property

$$D(P_{Y^n|X^n=x^n} \| P^*_{Y^n}) \le nC, \qquad \forall x^n \in \mathcal{X}^n \tag{3.283}$$

where $P_{Y^n|X^n=x^n}$ is shorthand for the product distribution $T^n(\cdot|x^n)$. From the bound (3.283), we see that the capacity-achieving output distribution $P^*_{Y^n}$ dominates any output distribution P_{Y^n} induced by an arbitrary input distribution $P_{X^n} \in \mathcal{P}(\mathcal{X}^n)$:

$$P_{Y^n|X^n=x^n} \ll P^*_{Y^n}, \forall x^n \in \mathcal{X}^n \qquad \Longrightarrow \qquad P_{Y^n} \ll P^*_{Y^n}, \forall P_{X^n} \in \mathcal{P}(\mathcal{X}^n).$$

This has two important consequences:

1. The information density is well-defined for any $x^n \in \mathcal{X}^n$ and $y^n \in \mathcal{Y}^n$:

$$i^*_{X^n;Y^n}(x^n; y^n) \triangleq \ln \frac{\mathrm{d}P_{Y^n|X^n=x^n}(y^n)}{\mathrm{d}P^*_{Y^n}}.$$

2. For any input distribution P_{X^n}, the corresponding output distribution P_{Y^n} satisfies

$$D(P_{Y^n} \| P^*_{Y^n}) \le nC - I(X^n; Y^n)$$

Indeed, by the chain rule for divergence for any input distribution $P_{X^n} \in \mathcal{P}(\mathcal{X}^n)$ we have

$$\begin{aligned} I(X^n; Y^n) &= D(P_{Y^n|X^n} \| P_{Y^n} | P_{X^n}) \\ &= D(P_{Y^n|X^n} \| P^*_{Y^n} | P_{X^n}) - D(P_{Y^n} \| P^*_{Y^n}) \\ &\le nC - D(P_{Y^n} \| P^*_{Y^n}). \end{aligned}$$

The claimed bound follows upon rearranging this inequality.

Now let us bring codes into the picture. Given $n, M \in \mathbb{N}$, an (n, M)-*code* for T is a pair $\mathcal{C} = (f_n, g_n)$ consisting of an *encoding map* $f_n : \{1, \dots, M\} \to \mathcal{X}^n$ and a *decoding map* $g_n : \mathcal{Y}^n \to \{1, \dots, M\}$. Given $0 < \varepsilon \le 1$, we say that $\mathcal{C}$ is an (n, M, ε)-*code* if

$$\max_{1 \le i \le M} \mathbb{P}\big(g_n(Y^n) \ne i \big| X^n = f_n(i)\big) \le \varepsilon. \tag{3.284}$$

Remark 47. Polyanskiy and Verdú [158] use a more precise nomenclature and say that any such $\mathcal{C} = (f_n, g_n)$ satisfying (3.284) is an $(n, M, \varepsilon)_{\text{max,det}}$-*code* to indicate explicitly that the decoding map g_n is deterministic and that the maximal probability of error criterion is used. Here, we will only consider codes of this type, so we will adhere to our simplified terminology.

Consider any (n, M)-code $\mathcal{C} = (f_n, g_n)$ for T, and let J be a random variable uniformly distributed on $\{1, \dots, M\}$. Hence, we can think of any $1 \le i \le M$ as one of M equiprobable messages to be transmitted over T. Let $P^{(\mathcal{C})}_{X^n}$ denote the distribution of $X^n = f_n(J)$, and let $P^{(\mathcal{C})}_{Y^n}$ denote the corresponding output distribution. The central result of [158] is that the output distribution $P^{(\mathcal{C})}_{Y^n}$ of any (n, M, ε)-code satisfies

$$D\big(P^{(\mathcal{C})}_{Y^n} \big\| P^*_{Y^n}\big) \le nC - \ln M + o(n); \tag{3.285}$$

moreover, the $o(n)$ term may be refined to $O(\sqrt{n})$ for any DMC T, except those that have zeroes in their transition matrix. For the proof of (3.285) with the $O(\sqrt{n})$ term, we will need the following strong converse for channel codes due to Augustin [161] (see also [162]):

Theorem 49 (Augustin)**.** Let $S : \mathcal{U} \to \mathcal{V}$ be a DMC with finite input and output alphabets, and let $P_{V|U}$ be the transition probability induced by S. For any $M \in \mathbb{N}$ and $0 < \varepsilon \le 1$, let $f : \{1, \dots, M\} \to \mathcal{U}$ and $g : \mathcal{V} \to \{1, \dots, M\}$ be two mappings, such that

$$\max_{1 \le i \le M} \mathbb{P}\big(g(V) \ne i \big| U = f(i)\big) \le \varepsilon.$$

Let $Q_V \in \mathcal{P}(\mathcal{V})$ be an auxiliary output distribution, and fix an arbitrary map $\gamma : \mathcal{U} \to \mathbb{R}$. Then, the following inequality holds:

$$M \le \frac{\exp\{\mathbb{E}[\gamma(U)]\}}{\inf\limits_{u \in \mathcal{U}} P_{V|U=u}\left(\ln \dfrac{\mathrm{d}P_{V|U=u}}{\mathrm{d}Q_V} < \gamma(u)\right) - \varepsilon}, \tag{3.286}$$

provided the denominator is strictly positive. The expectation in the numerator is taken w.r.t. the distribution of $U = f(J)$ with $J \sim \text{Uniform}\{1, \dots, M\}$.

We first establish the bound (3.285) for the case when the DMC T is such that

$$C_1 \triangleq \max_{x, x' \in \mathcal{X}} D(P_{Y|X=x} \| P_{Y|X=x'}) < \infty. \tag{3.287}$$

Note that $C_1 < \infty$ if and only if the transition matrix of T does not have any zeroes. Consequently,

$$c(T) \triangleq 2 \max_{x, x' \in \mathcal{X}} \max_{y, y' \in \mathcal{Y}} \left| \ln \frac{P_{Y|X}(y|x)}{P_{Y|X}(y'|x')} \right| < \infty.$$

We can now establish the following sharpened version of Theorem 5 from [158]:

Theorem 50. Let $T : \mathcal{X} \to \mathcal{Y}$ be a DMC with $C > 0$ satisfying (3.287). Then, any (n, M, ε)-code $\mathcal{C}$ for T with $0 < \varepsilon < 1/2$ satisfies

$$D\big(P^{(C)}_{Y^n} \big\| P^*_{Y^n}\big) \le nC - \ln M + \ln \frac{1}{\varepsilon} + c(T) \sqrt{\frac{n}{2} \ln \frac{1}{1 - 2\varepsilon}}. \tag{3.288}$$

Remark 48. Our sharpening of the corresponding result from [158] consists mainly in identifying an explicit form for the constant in front of $\sqrt{n}$ in (3.288).

Remark 49. As shown in [158], the restriction to codes with deterministic decoders and to the maximal probability of error criterion is necessary both for this theorem and for the next one.

Proof. Fix an input sequence $x^n \in \mathcal{X}^n$ and consider the function $h_{x^n} : \mathcal{Y}^n \to \mathbb{R}$ defined by

$$h_{x^n}(y^n) \triangleq \ln \frac{\mathrm{d}P_{Y^n|X^n=x^n}}{\mathrm{d}P^{(\mathcal{C})}_{Y^n}}(y^n).$$

Then $\mathbb{E}[h_{x^n}(Y^n)|X^n = x^n] = D(P_{Y^n|X^n=x^n} \| P^{(\mathcal{C})}_{Y^n})$. Moreover, for any $i \in \{1, \dots, n\}$, $y, y' \in \mathcal{Y}$, and $\overline{y}^i \in \mathcal{Y}^{n-1}$, we have (see the notation used in (3.24))

$$\begin{aligned}
\left|h_{i,x^n}(y|\overline{y}^i) - h_{i,x^n}(y'|\overline{y}^i)\right| &\le \left|\ln P_{Y^n|X^n=x^n}(y^{i-1}, y, y^n_{i+1}) - \ln P_{Y^n|X^n=x^n}(y^{i-1}, y', y^n_{i+1})\right| \\
&\quad + \left|\ln P^{(\mathcal{C})}_{Y^n}(y^{i-1}, y, y^n_{i+1}) - \ln P^{(\mathcal{C})}_{Y^n}(y^{i-1}, y', y^n_{i+1})\right| \\
&\le \left|\ln \frac{P_{Y_i|X_i=x_i}(y)}{P_{Y_i|X_i=x_i}(y')}\right| + \left|\ln \frac{P^{(\mathcal{C})}_{Y_i|\overline{Y}^i}(y|\overline{y}^i)}{P^{(\mathcal{C})}_{Y_i|\overline{Y}^i}(y'|\overline{y}^i)}\right| \\
&\le 2 \max_{x,x' \in \mathcal{X}} \max_{y,y' \in \mathcal{Y}} \left|\ln \frac{P_{Y|X}(y|x)}{P_{Y|X}(y'|x')}\right| && (3.289) \\
&= c(T) < \infty && (3.290)
\end{aligned}$$

(see Appendix 3.D for a detailed explanation of the inequality in (3.289)). Hence, for each fixed $x^n \in \mathcal{X}^n$, the function $h_{x^n} : \mathcal{Y}^n \to \mathbb{R}$ satisfies the bounded differences condition (3.134) with $c_1 = \ldots = c_n = c(T)$. Theorem 28 therefore implies that, for any $r \ge 0$, we have

$$P_{Y^n|X^n=x^n}\left(\ln \frac{\mathrm{d}P_{Y^n|X^n=x^n}}{\mathrm{d}P^{(\mathcal{C})}_{Y^n}}(Y^n) \ge D(P_{Y^n|X^n=x^n} \| P^{(\mathcal{C})}_{Y^n}) + r\right) \le \exp\left(-\frac{2r^2}{nc^2(T)}\right) \tag{3.291}$$

(In fact, the above derivation goes through for any possible output distribution P_{Y^n}, not necessarily one induced by a code.) This is where we have departed from the original proof by Polyanskiy and Verdú [158]: we have used McDiarmid's (or bounded differences) inequality to control the deviation probability for the "conditional" information density h_{x^n} directly, whereas they bounded the *variance* of h_{x^n} using a suitable Poincaré inequality, and then derived a bound on the derivation probability using Chebyshev's inequality. As we will see shortly, the sharp concentration inequality (3.291) allows us to explicitly identify the dependence of the constant multiplying $\sqrt{n}$ in (3.288) on the channel T and on the maximal error probability ε.

We are now in a position to apply Augustin's strong converse. To that end, we let $\mathcal{U} = \mathcal{X}^n$, $\mathcal{V} = \mathcal{Y}^n$, and consider the DMC $S = T^n$ together with an (n, M, ε)-code $(f, g) = (f_n, g_n)$. Furthermore, let

$$\zeta_n = \zeta_n(\varepsilon) \triangleq c(T)\sqrt{\frac{n}{2} \ln \frac{1}{1 - 2\varepsilon}} \tag{3.292}$$

and take $\gamma(x^n) = D(P_{Y^n|X^n=x^n} \| P^{(\mathcal{C})}_{Y^n}) + \zeta_n$. Using (3.286) with the auxiliary distribution $Q_V = P^{(\mathcal{C})}_{Y^n}$, we get

$$M \le \frac{\exp\{\mathbb{E}[\gamma(\mathcal{X}^n)]\}}{\inf_{x^n \in \mathcal{X}^n} P_{Y^n|X^n=x^n}\left(\ln \frac{\mathrm{d}P_{Y^n|X^n=x^n}}{\mathrm{d}P^{(\mathcal{C})}_{Y^n}} < \gamma(x^n)\right) - \varepsilon} \tag{3.293}$$

where $\mathbb{E}[\gamma(X^n)] = D\big(P_{Y^n|X^n} \| P_{Y^n}^{(\mathcal{C})} \,|\, P_{X^n}^{(\mathcal{C})}\big) + \zeta_n$. The concentration inequality in (3.291) with ζ_n in (3.292) therefore gives that, for every $x^n \in \mathcal{X}^n$,

$$\begin{aligned} P_{Y^n|X^n=x^n}\left(\ln \frac{\mathrm{d}P_{Y^n|X^n=x^n}}{\mathrm{d}P_{Y^n}^{(\mathcal{C})}} \geq \gamma(x^n)\right) &\leq \exp\left(-\frac{2\zeta_n^2}{nc^2(T)}\right) \\ &= 1 - 2\varepsilon \end{aligned}$$

which implies that

$$\inf_{x^n \in \mathcal{X}^n} P_{Y^n|X^n=x^n}\left(\ln \frac{\mathrm{d}P_{Y^n|X^n=x^n}}{\mathrm{d}P_{Y^n}^{(\mathcal{C})}} < \gamma(x^n)\right) \geq 2\varepsilon.$$

Hence, from (3.293) and the last inequality, it follows that

$$M \leq \frac{1}{\varepsilon} \, \exp\Big(D\big(P_{Y^n|X^n} \| P_{Y^n}^{(\mathcal{C})} \,|\, P_{X^n}^{(\mathcal{C})}\big) + \zeta_n\Big)$$

so, by taking logarithms on both sides of the last inequality and rearranging terms, we get from (3.292) that

$$\begin{aligned} D(P_{Y^n|X^n} \| P_{Y^n}^{(\mathcal{C})} \,|\, P_{X^n}^{(\mathcal{C})}) &\geq \ln M + \ln \varepsilon - \zeta_n \\ &= \ln M + \ln \varepsilon - c(T)\sqrt{\frac{n}{2} \ln \frac{1}{1-2\varepsilon}}. \end{aligned} \tag{3.294}$$

We are now ready to derive (3.288):

$$\begin{aligned} &D\big(P_{Y^n}^{(\mathcal{C})} \big\| P_{Y^n}^*\big) \\ &\quad = D\big(P_{Y^n|X^n} \big\| P_{Y^n}^* | P_{X^n}^{(\mathcal{C})}\big) - D\big(P_{Y^n|X^n} \big\| P_{Y^n}^{(\mathcal{C})} | P_{X^n}^{(\mathcal{C})}\big) && (3.295) \\ &\quad \leq nC - \ln M + \ln \frac{1}{\varepsilon} + c(T)\sqrt{\frac{n}{2} \ln \frac{1}{1-2\varepsilon}} && (3.296) \end{aligned}$$

where (3.295) uses the chain rule for divergence, while (3.296) uses (3.294) and (3.283). This completes the proof of Theorem 50. □

For an arbitrary DMC T with nonzero capacity and zeroes in its transition matrix, we have the following result from [158]:

Theorem 51. Let $T : \mathcal{X} \to \mathcal{Y}$ be a DMC with $C > 0$. Then, for any $0 < \varepsilon < 1$, any (n, M, ε)-code $\mathcal{C}$ for T satisfies

$$D\big(P_{Y^n}^{(\mathcal{C})} \big\| P_{Y^n}^*\big) \leq nC - \ln M + O\left(\sqrt{n} \ln^{3/2} n\right).$$

More precisely, for any such code we have

$$\begin{aligned} &D\big(P_{Y^n}^{(\mathcal{C})} \big\| P_{Y^n}^*\big) \\ &\quad \leq nC - \ln M + \sqrt{2n}\,(\ln n)^{3/2}\left(1 + \sqrt{\frac{1}{\ln n} \ln\left(\frac{1}{1-\varepsilon}\right)}\right)\left(1 + \frac{\ln|\mathcal{Y}|}{\ln n}\right) + 3\ln n + \ln\big(2|\mathcal{X}||\mathcal{Y}|^2\big). \end{aligned} \tag{3.297}$$

Proof. Given an (n, M, ε)-code $\mathcal{C} = (f_n, g_n)$, let $c_1, \ldots, c_M \in \mathcal{X}^n$ be its codewords, and let $\widetilde{D}_1, \ldots, \widetilde{D}_M \subset \mathcal{Y}^n$ be the corresponding decoding regions:

$$\widetilde{D}_i = g_n^{-1}\left(\mathcal{Y}^n\right) \equiv \left\{y^n \in \mathcal{Y}^n : g_n^{-1}(y^n) = i\right\}, \qquad i = 1, \ldots, M.$$

If we choose

$$\delta_n = \delta_n(\varepsilon) = \frac{1}{n}\left\lceil n\left(\sqrt{\frac{\ln n}{2n}} + \sqrt{\frac{1}{2n}\ln\frac{1}{1-\varepsilon}}\right)\right\rceil \tag{3.298}$$

(note that $n\delta_n$ is an integer), then by Lemma 15 the "blown-up" decoding regions $D_i \triangleq \left[\widetilde{D}_i\right]_{n\delta_n}$, $1 \le i \le M$, satisfy

$$\begin{aligned} P_{Y^n|X^n=c_i}\left(D_i^c\right) &\le \exp\left[-2n\left(\delta_n - \sqrt{\frac{1}{2n}\ln\frac{1}{1-\varepsilon}}\right)^2\right] \\ &\le \frac{1}{n}, \qquad\qquad \forall\, i \in \{1, \ldots, M\}. \end{aligned} \tag{3.299}$$

We now complete the proof by a random coding argument. For

$$N \triangleq \frac{M}{n\binom{n}{n\delta_n}|\mathcal{Y}|^{n\delta_n}}, \tag{3.300}$$

let $U_1, \ldots, U_N$ be independent random variables, each uniformly distributed on the set $\{1, \ldots, M\}$. For each realization $V = U^N$, let $P_{X^n(V)} \in \mathcal{P}(\mathcal{X}^n)$ denote the induced distribution of $X^n(V) = f_n(c_J)$, where J is uniformly distributed on the set $\{U_1, \ldots, U_N\}$, and let $P_{Y^n(V)}$ denote the corresponding output distribution of $Y^n(V)$:

$$P_{Y^n(V)} = \frac{1}{N}\sum_{i=1}^{N} P_{Y^n|X^n=c_{U_i}}. \tag{3.301}$$

It is easy to show that $\mathbb{E}\left[P_{Y^n}^{(V)}\right] = P_{Y^n}^{(\mathcal{C})}$, the output distribution of the original code $\mathcal{C}$, where the expectation is w.r.t. the distribution of $V = U^N$. Now, for $V = U^N$ and for every $y^n \in \mathcal{Y}^n$, let $\mathcal{N}_V(y^n)$ denote the list of all those indices in $(U_1, \ldots, U_N)$ such that $y^n \in D_{U_j}$:

$$\mathcal{N}_V(y^n) = \left\{U_j : y^n \in D_{U_j}\right\}.$$

Consider the list decoder $Y^n \mapsto \mathcal{N}_V(Y^n)$, and let $\varepsilon(V)$ denote its average decoding error probability: $\varepsilon(V) = P(J \notin \mathcal{N}_V(Y^n)|V)$. Then, for each realization of V, we have

$$\begin{aligned} & D\big(P_{Y^n(V)} \big\| P^*_{Y^n}\big) \\ &\quad = D\big(P_{Y^n|X^n} \big\| P^*_{Y^n} \big| P_{X^n(V)}\big) - I(X^n(V); Y^n(V)) && (3.302) \\ &\quad \le nC - I(X^n(V); Y^n(V)) && (3.303) \\ &\quad \le nC - I(J; Y^n(V)) && (3.304) \\ &\quad = nC - H(J) + H(J|Y^n(V)) && (3.305) \\ &\quad \le nC - \ln N + (1 - \varepsilon(V)) \ln|\mathcal{N}_V(Y^n)| + n\varepsilon(V) \ln|\mathcal{X}| + \ln 2 && (3.306) \end{aligned}$$

where:

- (3.302) is by the chain rule for divergence;

- (3.303) is by (3.283);
- (3.304) is by the data processing inequality and the fact that $J \to X^n(V) \to Y^n(V)$ is a Markov chain; and
- (3.306) is by Fano's inequality for list decoding (see Appendix 3.C), and also since (i) $N \leq |\mathcal{X}|^n$, (ii) J is uniformly distributed on $\{U_1, \ldots, U_N\}$, so $H(J|U_1, \ldots, U_N) = \ln N$ and $H(J) \geq \ln N$.

(Note that all the quantities indexed by V in the above chain of estimates are actually random variables, since they depend on the realization $V = U^N$.) Now, from (3.300) it follows that

$$\begin{aligned} \ln N &= \ln M - \ln n - \ln \binom{n}{n\delta_n} - n\delta_n \ln |\mathcal{Y}| \\ &\geq \ln M - \ln n - n\delta_n \left(\ln n + \ln |\mathcal{Y}|\right) \end{aligned} \tag{3.307}$$

where the last inequality uses the simple inequality $\binom{n}{k} \leq n^k$ for $k \leq n$ with $k \triangleq n\delta_n$ (it is noted that the gain in using instead the inequality $\binom{n}{n\delta_n} \leq \exp\bigl(n\, h(\delta_n)\bigr)$ is marginal, and it does not have any advantage asymptotically for large n). Moreover, each $y^n \in \mathcal{Y}^n$ can belong to at most $\binom{n}{n\delta_n} |\mathcal{Y}|^{n\delta_n}$ blown-up decoding sets, so

$$\begin{aligned} \ln |\mathcal{N}_V(Y^n)| &\leq \ln \binom{n}{n\delta_n} + n\delta_n \ln |\mathcal{Y}| \\ &\leq n\delta_n \left(\ln n + \ln |\mathcal{Y}|\right). \end{aligned} \tag{3.308}$$

Substituting (3.307) and (3.308) into (3.306), we get

$$D\bigl(P_{Y^n(V)} \big\| P^*_{Y^n}\bigr) \leq nC - \ln M + \ln n + 2n\delta_n \left(\ln n + \ln |\mathcal{Y}|\right) + n\varepsilon(V) \ln |\mathcal{X}| + \ln 2. \tag{3.309}$$

Using the fact that $\mathbb{E}\left[P_{Y^n(V)}\right] = P^{(\mathcal{C})}_{Y^n}$, convexity of the relative entropy, and (3.309), we get

$$D\bigl(P^{(\mathcal{C})}_{Y^n} \big\| P^*_{Y^n}\bigr) \leq nC - \ln M + \ln n + 2n\delta_n \left(\ln n + \ln |\mathcal{Y}|\right) + n\, \mathbb{E}\left[\varepsilon(V)\right] \ln |\mathcal{X}| + \ln 2. \tag{3.310}$$

To finish the proof and get (3.297), we use the fact that

$$\mathbb{E}\left[\varepsilon(V)\right] \leq \max_{1 \leq i \leq M} P_{Y^n|X^n=c_i}\left(D_i^c\right) \leq \frac{1}{n},$$

which follows from (3.299), as well as the substitution of (3.298) in (3.310) (note that, from (3.298), it follows that $\delta_n < \sqrt{\frac{\ln n}{2n}} + \sqrt{\frac{1}{2n} \ln \frac{1}{1-\varepsilon}} + \frac{1}{n}$). This completes the proof of Theorem 51. □

We are now ready to examine some consequences of Theorems 50 and 51. To start with, consider a sequence $\{\mathcal{C}_n\}_{n=1}^{\infty}$, where each $\mathcal{C}_n = (f_n, g_n)$ is an (n, M_n, ε)-code for a DMC $T : \mathcal{X} \to \mathcal{Y}$ with $C > 0$. We say that $\{\mathcal{C}_n\}_{n=1}^{\infty}$ is *capacity-achieving* if

$$\lim_{n \to \infty} \frac{1}{n} \ln M_n = C. \tag{3.311}$$

Then, from Theorems 50 and 51, it follows that any such sequence satisfies

$$\lim_{n \to \infty} \frac{1}{n} D\bigl(P^{(\mathcal{C}_n)}_{Y^n} \big\| P^*_{Y^n}\bigr) = 0. \tag{3.312}$$

Moreover, as shown in [158], if the restriction to either deterministic decoding maps or to the maximal probability of error criterion is lifted, then the convergence in (3.312) may no longer hold. This is in sharp contrast to [157, Theorem 2], which states that (3.312) holds for *any* capacity-achieving sequence of codes with vanishing probability of error (maximal or average).

Another remarkable fact that follows from the above theorems is that a broad class of functions evaluated on the output of a good code concentrate sharply around their expectations with respect to the capacity-achieving output distribution. Specifically, we have the following version of [158, Proposition 10] (again, we have streamlined the statement and the proof a bit to relate them to earlier material in this chapter):

Theorem 52. Let $T : \mathcal{X} \to \mathcal{Y}$ be a DMC with $C > 0$ and $C_1 < \infty$. Let $d : \mathcal{Y}^n \times \mathcal{Y}^n \to \mathbb{R}_+$ be a metric, and suppose that there exists a constant $c > 0$, such that the conditional probability distributions $P_{Y^n|X^n=x^n}$, $x^n \in \mathcal{X}^n$, as well as $P^*_{Y^n}$ satisfy $T_1(c)$ on the metric space $(\mathcal{Y}^n, d)$. Then, for any $\varepsilon \in (0,1)$, there exists a constant $a > 0$ that depends only on T and on ε, such that for any (n, M, ε)-code $\mathcal{C}$ for T and any function $f : \mathcal{Y}^n \to \mathbb{R}$ we have

$$P^{(\mathcal{C})}_{Y^n}\Big(|f(Y^n) - \mathbb{E}[f(Y^{*n})]| \geq r \Big) \leq 4 \exp\left(nC - \ln M + a\sqrt{n} - \frac{r^2}{8c\|f\|^2_{\mathrm{Lip}}} \right), \qquad \forall\, r \geq 0 \tag{3.313}$$

where $\mathbb{E}[f(Y^{*n})]$ designates the expected value of $f(Y^n)$ w.r.t. the capacity-achieving output distribution $P^*_{Y^n}$, and

$$\|f\|_{\mathrm{Lip}} \triangleq \sup_{y^n \neq v^n} \frac{|f(y^n) - f(v^n)|}{d(y^n, v^n)}$$

is the Lipschitz constant of f w.r.t. the metric d.

Proof. For any f, define

$$\mu^*_f \triangleq \mathbb{E}[f(Y^{*n})], \qquad \phi(x^n) \triangleq \mathbb{E}[f(Y^n)|X^n = x^n],\ \forall\, x^n \in \mathcal{X}^n. \tag{3.314}$$

Since each $P_{Y^n|X^n=x^n}$ satisfies $T_1(c)$, by the Bobkov–Götze theorem (Theorem 36), we have

$$\mathbb{P}\Big(|f(Y^n) - \phi(x^n)| \geq r \Big| X^n = x^n \Big) \leq 2 \exp\left(-\frac{r^2}{2c\|f\|^2_{\mathrm{Lip}}} \right), \qquad \forall\, r \geq 0. \tag{3.315}$$

Now, given $\mathcal{C}$, consider a subcode $\mathcal{C}'$ with codewords $x^n \in \mathcal{X}^n$ satisfying $\phi(x^n) > \mu^*_f + r$ for $r > 0$. The number of codewords M' of $\mathcal{C}'$ satisfies

$$M' = M P^{(\mathcal{C})}_{X^n}\left(\phi(X^n) \geq \mu^*_f + r \right). \tag{3.316}$$

Let $Q = P^{(\mathcal{C}')}_{Y^n}$ be the output distribution induced by $\mathcal{C}'$. Then

$$\begin{aligned}
\mu^*_f + r &\leq \frac{1}{M'} \sum_{x^n \in\, \mathrm{codewords}(\mathcal{C}')} \phi(x^n) && (3.317)\\
&= \mathbb{E}_Q[f(Y^n)] && (3.318)\\
&\leq \mathbb{E}[f(Y^{*n})] + \|f\|_{\mathrm{Lip}} \sqrt{2cD(Q_{Y^n} \| P^*_{Y^n})} && (3.319)\\
&\leq \mu^*_f + \|f\|_{\mathrm{Lip}} \sqrt{2c\left(nC - \ln M' + a\sqrt{n} \right)}, && (3.320)
\end{aligned}$$

where:

- (3.317) is by definition of $\mathcal{C}'$;

- (3.318) is by definition of ϕ in (3.314);
- (3.319) follows from the fact that $P^*_{Y^n}$ satisfies $T_1(c)$ and from the Kantorovich–Rubinstein formula (3.209); and
- (3.320) holds, for an appropriate $a = a(T, \varepsilon) > 0$, by Theorem 50, because $\mathcal{C}'$ is an (n, M', ε)-code for T.

From this and (3.316), we get

$$r \le \|f\|_{\text{Lip}} \sqrt{2c \left(nC - \ln M - \ln P^{(\mathcal{C})}_{X^n}\left(\phi(X^n) \ge \mu^*_f + r\right) + a\sqrt{n}\right)}$$

so, it follows that

$$P^{(\mathcal{C})}_{X^n}\left(\phi(X^n) \ge \mu^*_f + r\right) \le \exp\left(nC - \ln M + a\sqrt{n} - \frac{r^2}{2c\|f\|^2_{\text{Lip}}}\right).$$

Following the same line of reasoning with $-f$ instead of f, we conclude that

$$P^{(\mathcal{C})}_{X^n}\left(\left|\phi(X^n) - \mu^*_f\right| \ge r\right) \le 2\exp\left(nC - \ln M + a\sqrt{n} - \frac{r^2}{2c\|f\|^2_{\text{Lip}}}\right). \tag{3.321}$$

Finally, for every $r \ge 0$,

$$\begin{aligned}
&P^{(\mathcal{C})}_{Y^n}\left(\left|f(Y^n) - \mu^*_f\right| \ge r\right) \\
&\quad \le P^{(\mathcal{C})}_{X^n,Y^n}\left(|f(Y^n) - \phi(X^n)| \ge r/2\right) + P^{(\mathcal{C})}_{X^n}\left(\left|\phi(X^n) - \mu^*_f\right| \ge r/2\right) \\
&\quad \le 2\exp\left(-\frac{r^2}{8c\|f\|^2_{\text{Lip}}}\right) + 2\exp\left(nC - \ln M + a\sqrt{n} - \frac{r^2}{8c\|f\|^2_{\text{Lip}}}\right) &(3.322)\\
&\quad = 2\exp\left(-\frac{r^2}{8c\|f\|^2_{\text{Lip}}}\right)\left(1 + \exp\left(nC - \ln M + a\sqrt{n}\right)\right) \\
&\quad \le 4\exp\left(nC - \ln M + a\sqrt{n} - \frac{r^2}{8c\|f\|^2_{\text{Lip}}}\right), &(3.323)
\end{aligned}$$

where (3.322) is by (3.315) and (3.321), while (3.323) follows from the fact that

$$nC - \ln M + a\sqrt{n} \ge D(P^{(\mathcal{C})}_{Y^n} \| P^*_{Y^n}) \ge 0$$

by Theorem 50, and the way that the constant a was selected above (see (3.320)). This proves (3.313). □

As an illustration, let us consider $\mathcal{Y}^n$ with the product metric

$$d(y^n, v^n) = \sum_{i=1}^n 1_{\{y_i \ne v_i\}} \tag{3.324}$$

(this is the metric $d_{1,n}$ induced by the Hamming metric on $\mathcal{Y}$). Then any function $f : \mathcal{Y}^n \to \mathbb{R}$ of the form

$$f(y^n) = \frac{1}{n}\sum_{i=1}^n f_i(y_i), \qquad \forall y^n \in \mathcal{Y}^n \tag{3.325}$$

where $f_1, \ldots, f_n : \mathcal{Y} \to \mathbb{R}$ are Lipschitz functions on $\mathcal{Y}$, will satisfy

$$\|f\|_{\text{Lip}} \le \frac{L}{n}, \qquad L \triangleq \max_{1 \le i \le n} \|f_i\|_{\text{Lip}}.$$

Any probability distribution P on $\mathcal{Y}$ equipped with the Hamming metric satisfies $T_1(1/4)$ (this is simply Pinsker's inequality); by Proposition 11, any product probability distribution on $\mathcal{Y}^n$ satisfies $T_1(n/4)$ w.r.t. the product metric (3.324). Consequently, for any (n, M, ε)-code for T and any function $f : \mathcal{Y}^n \to \mathbb{R}$ of the form (3.325), Theorem 52 gives the concentration inequality

$$P^{(\mathcal{C})}_{Y^n}\Big(|f(Y^n) - \mathbb{E}[f(Y^{*n})]| \ge r\Big) \le 4 \exp\left(nC - \ln M + a\sqrt{n} - \frac{2nr^2}{\|f\|^2_{\text{Lip}}}\right), \qquad \forall r \ge 0. \tag{3.326}$$

Concentration inequalities like (3.313) or its more specialized version (3.326), can be very useful in characterizing various performance characteristics of good channel codes without having to explicitly construct such codes: all one needs to do is to find the capacity-achieving output distribution P^*_Y and evaluate $\mathbb{E}[f(Y^{*n})]$ for any f of interest. Then, Theorem 52 guarantees that $f(Y^n)$ concentrates tightly around $\mathbb{E}[f(Y^{*n})]$, which is relatively easy to compute since $P^*_{Y^n}$ is a product distribution.

Remark 50. This sub-section considers the empirical output distributions of good channel codes with non-vanishing probability of error via the use of concentration inequalities. As a concluding remark, it is noted that the combined result in [163, Eqs. (A17), (A19)] provides a lower bound on the rate loss with respect to fully random block codes (with a binomial distribution) in terms of the normalized divergence between the distance spectrum of the considered code and the binomial distribution. This result refers to the empirical input distribution of good codes, and it was derived via the use of variations on the Gallager bounds.

3.6.3 An information-theoretic converse for concentration of measure

If we were to summarize the main idea behind concentration of measure, it would be this: if a subset of a metric probability space does not have a "too small" probability mass, then its isoperimetric enlargements (or blowups) will eventually take up most of the probability mass. On the other hand, it makes sense to ask whether a converse of this statement is true — given a set whose blowups eventually take up most of the probability mass, how small can this set be? This question was answered precisely by Kontoyiannis [164] using information-theoretic techniques.

The following setting is considered in [164]: Let $\mathcal{X}$ be a finite set, together with a nonnegative distortion function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ (which is not necessarily a metric) and a strictly positive mass function $M : \mathcal{X} \to (0, \infty)$ (which is not necessarily normalized to one). As before, let us extend the "single-letter" distortion d to $d_n : \mathcal{X}^n \to \mathbb{R}^+$, $n \in \mathbb{N}$, where

$$d_n(x^n, y^n) \triangleq \sum_{i=1}^n d(x_i, y_i), \qquad \forall x^n, y^n \in \mathcal{X}^n.$$

For every $n \in \mathbb{N}$ and for every set $C \subseteq \mathcal{X}^n$, let us define

$$M^n(C) \triangleq \sum_{x^n \in C} M^n(x^n)$$

where

$$M^n(x^n) \triangleq \prod_{i=1}^n M(x_i), \qquad \forall x^n \in \mathcal{X}^n.$$

As before, we define the r-blowup of any set $A \subseteq \mathcal{X}^n$ by

$$A_r \triangleq \{x^n \in \mathcal{X}^n : d_n(x^n, A) \le r\},$$

where $d_n(x^n, A) \triangleq \min_{y^n \in A} d_n(x^n, y^n)$. Fix a probability distribution $P \in \mathcal{P}(\mathcal{X})$, where we assume without loss of generality that P is strictly positive. We are interested in the following question: Given a sequence of sets $A^{(n)} \subseteq \mathcal{X}^n$, $n \in \mathbb{N}$, such that

$$P^{\otimes n}\left(A^{(n)}_{n\delta}\right) \to 1, \qquad \text{as } n \to \infty$$

for some $\delta \ge 0$, how small can their masses $M^n(A^{(n)})$ be?

In order to state and prove the main result of [164] that answers this question, we need a few preliminary definitions. For any $n \in \mathbb{N}$, any pair P_n, Q_n of probability measures on $\mathcal{X}^n$, and any $\delta \ge 0$, let us define the set

$$\Pi_n(P_n, Q_n, \delta) \triangleq \left\{\pi_n \in \Pi_n(P_n, Q_n) : \frac{1}{n}\mathbb{E}_{\pi_n}\left[d_n(X^n, Y^n)\right] \le \delta\right\} \tag{3.327}$$

of all couplings $\pi_n \in \mathcal{P}(\mathcal{X}^n \times \mathcal{X}^n)$ of P_n and Q_n, such that the per-letter expected distortion between X^n and Y^n with $(X^n, Y^n) \sim \pi_n$ is at most δ. With this, we define

$$I_n(P_n, Q_n, \delta) \triangleq \inf_{\pi_n \in \Pi_n(P_n, Q_n, \delta)} D(\pi_n \| P_n \otimes Q_n),$$

and consider the following *rate function*:

$$\begin{aligned} R_n(\delta) &\equiv R_n(\delta; P_n, M^n) \\ &\triangleq \inf_{Q_n \in \mathcal{P}(\mathcal{X}^n)} \Big\{ I_n(P_n, Q_n, \delta) + \mathbb{E}_{Q_n}[\ln M^n(Y^n)] \Big\} \\ &\equiv \inf_{P_{X^nY^n}} \left\{ I(X^n; Y^n) + \mathbb{E}[\ln M^n(Y^n)] : P_{X^n} = P_n, \frac{1}{n}\mathbb{E}[d_n(X^n, Y^n)] \le \delta \right\}. \end{aligned}$$

When $n = 1$, we will simply write $\Pi(P, Q, \delta)$, $I(P, Q, \delta)$ and $R(\delta)$. For the special case when each P_n is the product measure $P^{\otimes n}$, we have

$$R(\delta) = \lim_{n \to \infty} \frac{1}{n} R_n(\delta) = \inf_{n \ge 1} \frac{1}{n} R_n(\delta) \tag{3.328}$$

(see [164, Lemma 2]). We are now ready to state the main result of [164]:

Theorem 53. Consider an arbitrary set $A^{(n)} \subseteq \mathcal{X}^n$, and denote

$$\delta \triangleq \frac{1}{n}\mathbb{E}[d_n(X^n, A^{(n)})].$$

Then

$$\frac{1}{n}\ln M^n(A^{(n)}) \ge R(\delta; P, M). \tag{3.329}$$

Proof. Given $A_n \subseteq \mathcal{X}^n$, let $\varphi_n : \mathcal{X}^n \to A_n$ be the function that maps each $x^n \in \mathcal{X}^n$ to the closest element $y^n \in A_n$, i.e.,

$$d_n(x^n, \varphi_n(x^n)) = d_n(x^n, A_n)$$

(we assume some fixed rule for resolving ties). If $X^n \sim P^{\otimes n}$, then let $Q_n \in \mathcal{P}(\mathcal{X}^n)$ denote the distribution of $Y^n = \varphi_n(X^n)$, and let $\pi_n \in \mathcal{P}(\mathcal{X}^n \times \mathcal{X}^n)$ denote the joint distribution of X^n and Y^n:

$$Q_n(x^n, y^n) = P^{\otimes n}(x^n) 1_{\{y^n = \varphi_n(x^n)\}}.$$

Then, the two marginals of π_n are $P^{\otimes n}$ and Q_n and

$$\begin{aligned}\mathbb{E}_{\pi_n}[d_n(X^n, Y^n)] &= \mathbb{E}_{\pi_n}[d_n(X^n, \varphi_n(X^n))] \\ &= \mathbb{E}_{\pi_n}\left[d_n(X^n, A^n)\right] \\ &= n\delta,\end{aligned}$$

so $\pi_n \in \Pi_n(P^{\otimes n}, Q_n, \delta)$. Moreover,

$$\begin{aligned}\ln M^n(A_n) &= \ln \sum_{y^n \in A_n} M^n(y^n) \\ &= \ln \sum_{y^n \in A_n} Q_n(y^n) \cdot \frac{M^n(y^n)}{Q_n(y^n)} \\ &\geq \sum_{y^n \in A_n} Q_n(y^n) \ln \frac{M^n(y^n)}{Q_n(y^n)} && (3.330) \\ &= \sum_{x^n \in \mathcal{X}^n, y^n \in A_n} \pi_n(x^n, y^n) \ln \frac{\pi_n(x^n, y^n)}{P^{\otimes n}(x^n) Q_n(y^n)} + \sum_{y^n \in A_n} Q_n(y^n) \ln M^n(y^n) && (3.331) \\ &= I(X^n; Y^n) + \mathbb{E}_{Q_n}[\ln M^n(Y^n)] && (3.332) \\ &\geq R_n(\delta), && (3.333)\end{aligned}$$

where (3.330) is by Jensen's inequality, (3.331) and (3.332) use the fact that π_n is a coupling of $P^{\otimes n}$ and Q_n, and (3.333) is by definition of $R_n(\delta)$. Using (3.328), we get (3.329), and the theorem is proved. □

Remark 51. In the same paper [164], an achievability result was also proved: For any $\delta \geq 0$ and any $\varepsilon > 0$, there is a sequence of sets $A^{(n)} \subseteq \mathcal{X}^n$ such that

$$\frac{1}{n} \ln M^n(A^{(n)}) \leq R(\delta) + \varepsilon, \qquad \forall n \in \mathbb{N} \tag{3.334}$$

and

$$\frac{1}{n} d_n(X^n, A^{(n)}) \leq \delta, \qquad \text{eventually a.s.} \tag{3.335}$$

We are now ready to use Theorem 53 to answer the question posed at the beginning of this section. Specifically, we consider the case when $M = P$. Defining the *concentration exponent* $R_{\mathrm{c}}(r; P) \triangleq R(r; P, P)$, we have:

Corollary 11 (Converse concentration of measure)**.** If $A^{(n)} \subseteq \mathcal{X}^n$ is an arbitrary set, then

$$P^{\otimes n}\left(A^{(n)}\right) \geq \exp\left(n\, R_{\mathrm{c}}(\delta; P)\right), \tag{3.336}$$

where $\delta = \frac{1}{n}\mathbb{E}\left[d_n(X^n, A^{(n)})\right]$. Moreover, if the sequence of sets $\{A^{(n)}\}_{n=1}^{\infty}$ is such that, for some $\delta \geq 0$, $P^{\otimes n}\left(A^{(n)}_{n\delta}\right) \to 1$ as $n \to \infty$, then

$$\liminf_{n \to \infty} \frac{1}{n} \ln P^{\otimes n}\left(A^{(n)}\right) \geq R_{\mathrm{c}}(\delta; P). \tag{3.337}$$

Remark 52. A moment of reflection shows that the concentration exponent $R_{\mathrm{c}}(\delta; P)$ is nonpositive. Indeed, from definitions,

$$\begin{aligned}
R_{\mathrm{c}}(\delta; P) &= R(\delta; P, P) \\
&= \inf_{P_{XY}} \Big\{ I(X;Y) + \mathbb{E}[\ln P(Y)] : P_X = P,\, \mathbb{E}[d(X,Y)] \le \delta \Big\} \\
&= \inf_{P_{XY}} \Big\{ H(Y) - H(Y|X) + \mathbb{E}[\ln P(Y)] : P_X = P,\, \mathbb{E}[d(X,Y)] \le \delta \Big\} \\
&= \inf_{P_{XY}} \Big\{ - D(P_Y \| P) - H(Y|X) : P_X = P,\, \mathbb{E}[d(X,Y)] \le \delta \Big\} \\
&= - \sup_{P_{XY}} \Big\{ D(P_Y \| P) + H(Y|X) : P_X = P,\, \mathbb{E}[d(X,Y)] \le \delta \Big\}, \qquad (3.338)
\end{aligned}$$

which proves the claim, since both the divergence and the (conditional) entropy are nonnegative.

Remark 53. Using the achievability result from [164] (cf. Remark 51), one can also prove that there exists a sequence of sets $\{A^{(n)}\}_{n=1}^{\infty}$, such that

$$\lim_{n\to\infty} P^{\otimes n}\left(A^{(n)}_{n\delta}\right) = 1 \qquad \text{and} \qquad \lim_{n\to\infty} \frac{1}{n} \ln P^{\otimes n}\left(A^{(n)}\right) \le R_{\mathrm{c}}(\delta; P).$$

As an illustration, let us consider the case when $\mathcal{X} = \{0,1\}$ and d is the Hamming distortion, $d(x,y) = 1_{\{x \neq y\}}$. Then $\mathcal{X}^n = \{0,1\}^n$ is the n-dimensional binary cube. Let P be the Bernoulli(p) probability measure, which satisfies a $T_1\left(\frac{1}{2\varphi(p)}\right)$ transportation-cost inequality w.r.t. the L^1 Wasserstein distance induced by the Hamming metric, where $\varphi(p)$ is defined in (3.187). By Proposition 10, the product measure $P^{\otimes n}$ satisfies a $T_1\left(\frac{n}{2\varphi(p)}\right)$ transportation-cost inequality on the product space $(\mathcal{X}^n, d_n)$. Consequently, it follows from (3.197) that for any $\delta \ge 0$ and any $A^{(n)} \subseteq \mathcal{X}^n$,

$$\begin{aligned}
P^{\otimes n}\left(A^{(n)}_{n\delta}\right) &\ge 1 - \exp\left(-\frac{\varphi(p)}{n} \left(n\delta - \sqrt{\frac{n}{\varphi(p)} \ln \frac{1}{P^{\otimes n}\left(A^{(n)}\right)}} \right)^2 \right) \\
&= 1 - \exp\left(-n\,\varphi(p) \left(\delta - \sqrt{\frac{1}{n\,\varphi(p)} \ln \frac{1}{P^{\otimes n}\left(A^{(n)}\right)}} \right)^2 \right). \qquad (3.339)
\end{aligned}$$

Thus, if a sequence of sets $A^{(n)} \subseteq \mathcal{X}^n$, $n \in \mathbb{N}$, satisfies

$$\liminf_{n\to\infty} \frac{1}{n} \ln P^{\otimes n}\left(A^{(n)}\right) \ge -\varphi(p)\delta^2, \qquad (3.340)$$

then

$$P^{\otimes n}\left(A^{(n)}_{n\delta}\right) \xrightarrow{n\to\infty} 1. \qquad (3.341)$$

The converse result, Corollary 11, says that if a sequence of sets $A^{(n)} \subseteq \mathcal{X}^n$ satisfies (3.341), then (3.337) holds. Let us compare the concentration exponent $R_{\mathrm{c}}(\delta; P)$, where P is the Bernoulli(p) measure, with the exponent $-\varphi(p)\delta^2$ on the right-hand side of (3.340):

Theorem 54. If P is the Bernoulli(p) measure with $p \in [0, 1/2]$, then the concentration exponent $R_{\mathrm{c}}(\delta; P)$ satisfies

$$R_{\mathrm{c}}(\delta; P) \le -\varphi(p)\delta^2 - (1-p)h\left(\frac{\delta}{1-p}\right), \qquad \forall\, \delta \in [0, 1-p] \qquad (3.342)$$

and

$$R_{\rm c}(\delta;P) = \ln p, \qquad \forall\, \delta \in [1-p,1] \tag{3.343}$$

where $h(x) \triangleq -x\ln x - (1-x)\ln(1-x)$, $x \in [0,1]$, is the binary entropy function (in nats).

Proof. From (3.338), we have

$$R_{\rm c}(\delta;P) = -\sup_{P_{XY}} \Big\{ D(P_Y\|P) + H(Y|X) : P_X = P,\, \mathbb{P}(X \neq Y) \le \delta \Big\}. \tag{3.344}$$

For a given $\delta \in [0,1-p]$, let us choose P_Y so that $\|P_Y - P\|_{\rm TV} = \delta$. Then from (3.189),

$$\begin{aligned} \frac{D(P_Y\|P)}{\delta^2} &= \frac{D(P_Y\|P)}{\|P_Y-P\|^2_{\rm TV}} \\ &\ge \inf_Q \frac{D(Q\|P)}{\|Q-P\|^2_{\rm TV}} \\ &= \varphi(p). \end{aligned} \tag{3.345}$$

By the coupling representation of the total variation distance, we can choose a joint distribution $P_{\widetilde{X}\widetilde{Y}}$ with marginals $P_{\widetilde{X}} = P$ and $P_{\widetilde{Y}} = P_Y$, such that $\mathbb{P}(\widetilde{X} \neq \widetilde{Y}) = \|P_Y - P\|_{\rm TV} = \delta$. Moreover, using (3.185), we can compute

$$P_{\tilde{Y}|\tilde{X}=0} = \text{Bernoulli}\left(\frac{\delta}{1-p}\right) \qquad \text{and} \qquad P_{\tilde{Y}|\tilde{X}=1}(\tilde{y}) = \delta_1(\tilde{y}) \triangleq 1_{\{\tilde{y}=1\}}.$$

Consequently,

$$H(\widetilde{Y}|\widetilde{X}) = (1-p)H(\widetilde{Y}|\widetilde{X}=0) = (1-p)h\left(\frac{\delta}{1-p}\right). \tag{3.346}$$

From (3.344), (3.345) and (3.346), we obtain

$$\begin{aligned} R_{\rm c}(\delta;P) &\le -D(P_{\widetilde{Y}}\|P) - H(\widetilde{Y}|\widetilde{X}) \\ &\le -\varphi(p)\delta^2 - (1-p)h\left(\frac{\delta}{1-p}\right). \end{aligned}$$

To prove (3.343), it suffices to consider the case where $\delta = 1-p$. If we let Y be independent of $X \sim P$, then $I(X;Y) = 0$, so we have to minimize $\mathbb{E}_Q[\ln P(Y)]$ over all distributions Q of Y. But then

$$\min_Q \mathbb{E}_Q[\ln P(Y)] = \min_{y\in\{0,1\}} \ln P(y) = \min\{\ln p, \ln(1-p)\} = \ln p,$$

where the last equality holds since $p \le 1/2$. □

3.A Van Trees inequality

Consider the problem of estimating a random variable $Y \sim P_Y$ based on a noisy observation $U = \sqrt{s}Y + Z$, where $s > 0$ is the SNR parameter, while the additive noise $Z \sim G$ is independent of Y. We assume that P_Y has a differentiable, absolutely continuous density p_Y with $I(Y) < \infty$. Our goal is to prove the van Trees inequality (3.60) and to establish that equality in (3.60) holds if and only if Y is Gaussian.

In fact, we will prove a more general statement: Let $\varphi(U)$ be an arbitrary (Borel-measurable) estimator of Y. Then

$$\mathbb{E}\left[(Y - \varphi(U))^2\right] \ge \frac{1}{s + J(Y)}, \tag{3.347}$$

with equality if and only if Y has a standard normal distribution, and $\varphi(U)$ is the MMSE estimator of Y given U.

The strategy of the proof is, actually, very simple. Define two random variables

$$\begin{aligned}\Delta(U,Y) &\triangleq \varphi(U) - Y,\\ \Upsilon(U,Y) &\triangleq \frac{\mathrm{d}}{\mathrm{d}y}\ln\left[p_{U|Y}(U|y)p_Y(y)\right]\Bigg|_{y=Y}\\ &= \frac{\mathrm{d}}{\mathrm{d}y}\ln\left[\gamma(U-\sqrt{s}y)p_Y(y)\right]\Bigg|_{y=Y}\\ &= \sqrt{s}(U-\sqrt{s}Y) + \rho_Y(Y)\\ &= \sqrt{s}Z + \rho_Y(Y)\end{aligned}$$

where $\rho_Y(y) \triangleq \frac{\mathrm{d}}{\mathrm{d}y}\ln P_Y(y)$ for $y \in \mathbb{R}$ is the score function. We will show below that $\mathbb{E}[\Delta(U,Y)\Upsilon(U,Y)] = 1$. Then, applying the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}1 &= |\mathbb{E}[\Delta(U,Y)\Upsilon(U,Y)]|^2\\ &\le \mathbb{E}[\Delta^2(U,Y)]\cdot\mathbb{E}[\Upsilon^2(U,Y)]\\ &= \mathbb{E}[(\varphi(U)-Y)^2]\cdot\mathbb{E}[(\sqrt{s}Z+\rho_Y(Y))^2]\\ &= \mathbb{E}[(\varphi(U)-Y)^2]\cdot(s+J(Y)).\end{aligned}$$

Upon rearranging, we obtain (3.347). Now, the fact that $J(Y) < \infty$ implies that the density p_Y is bounded (see [117, Lemma A.1]). Using this together with the rapid decay of the Gaussian density γ at infinity, we have

$$\int_{-\infty}^{\infty}\frac{\mathrm{d}}{\mathrm{d}y}\left[p_{U|Y}(u|y)p_Y(y)\right]\mathrm{d}y = \gamma(u-\sqrt{s}y)p_Y(y)\Bigg|_{-\infty}^{\infty} = 0. \tag{3.348}$$

Integration by parts gives

$$\begin{aligned}\int_{-\infty}^{\infty} y\frac{\mathrm{d}}{\mathrm{d}y}\left[p_{U|Y}(u|y)p_Y(y)\right]\mathrm{d}y &= y\gamma(u-\sqrt{s}y)p_Y(y)\Bigg|_{-\infty}^{\infty} - \int_{-\infty}^{\infty}p_{U|Y}(u|y)p_Y(y)\mathrm{d}y\\ &= -\int_{-\infty}^{\infty}p_{U|Y}(u|y)p_Y(y)\mathrm{d}y\\ &= -p_U(u).\end{aligned} \tag{3.349}$$

Using (3.348) and (3.349), we have

$$\begin{aligned}&\mathbb{E}[\Delta(U,Y)\Upsilon(U,Y)]\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(\varphi(u)-y)\frac{\mathrm{d}}{\mathrm{d}y}\ln\left[p_{U|Y}(u|y)p_Y(y)\right]p_{U|Y}(u|y)p_Y(y)\mathrm{d}u\,\mathrm{d}y\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(\varphi(u)-y)\frac{\mathrm{d}}{\mathrm{d}y}\left[p_{U|Y}(u|y)p_Y(y)\right]\mathrm{d}u\,\mathrm{d}y\\ &= \int_{-\infty}^{\infty}\varphi(u)\underbrace{\left(\int_{-\infty}^{\infty}\frac{\mathrm{d}}{\mathrm{d}y}\left[p_{U|Y}(u|y)p_Y(y)\right]\mathrm{d}y\right)}_{=0}\mathrm{d}u - \int_{-\infty}^{\infty}\underbrace{\left(\int_{-\infty}^{\infty}y\frac{\mathrm{d}}{\mathrm{d}y}\left[p_{U|Y}(u|y)p_Y(y)\right]\mathrm{d}y\right)}_{=-p_U(u)}\mathrm{d}u\\ &= \int_{-\infty}^{\infty}p_U(u)\mathrm{d}u\\ &= 1,\end{aligned}$$

as was claimed. It remains to establish the necessary and sufficient condition for equality in (3.347). The Cauchy–Schwarz inequality for the product of $\Delta(U,Y)$ and $\Upsilon(U,Y)$ holds if and only if $\Delta(U,Y) = c\Upsilon(U,Y)$ for some constant $c \in \mathbb{R}$, almost surely. This is equivalent to

$$\begin{aligned}\varphi(U) &= Y + c\sqrt{s}(U - \sqrt{s}Y) + c\rho_Y(Y) \\ &= c\sqrt{s}U + (1-cs)Y + c\rho_Y(Y)\end{aligned}$$

for some $c \in \mathbb{R}$. In fact, c must be nonzero, for otherwise we will have $\varphi(U) = Y$, which is not a valid estimator. But then it must be the case that $(1-cs)Y + c\rho_Y(Y)$ is independent of Y, i.e., there exists some other constant $c' \in \mathbb{R}$, such that

$$\rho_Y(y) \triangleq \frac{p'_Y(y)}{p_Y(y)} = \frac{c'}{c} + (s - 1/c)y.$$

In other words, the score $\rho_Y(y)$ must be an affine function of y, which is the case if and only if Y is a Gaussian random variable.

3.B Details on the Ornstein–Uhlenbeck semigroup

In this appendix, we will prove the formulas (3.82) and (3.83) pertaining to the Ornstein–Uhlenbeck semigroup. We start with (3.82). Recalling that

$$h_t(x) = K_t h(x) = \mathbb{E}\left[h\left(e^{-t}x + \sqrt{1-e^{-2t}}Z\right)\right],$$

we have

$$\begin{aligned}\dot{h}_t(x) &= \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[h\left(e^{-t}x + \sqrt{1-e^{-2t}}Z\right)\right] \\ &= -e^{-t}x\,\mathbb{E}\left[h'\left(e^{-t}x + \sqrt{1-e^{-2t}}Z\right)\right] + \frac{e^{-2t}}{\sqrt{1-e^{-2t}}}\cdot\mathbb{E}\left[Zh'\left(e^{-t}x + \sqrt{1-e^{-2t}}Z\right)\right].\end{aligned}$$

For any sufficiently smooth function h and any $m, \sigma \in \mathbb{R}$,

$$\mathbb{E}[Zh'(m+\sigma Z)] = \sigma\mathbb{E}[h''(m+\sigma Z)]$$

(which is proved straightforwardly using integration by parts, provided that $\lim_{x\to\pm\infty} e^{-\frac{x^2}{2}}h'(m+\sigma x) = 0$). Using this equality, we can write

$$\mathbb{E}\left[Zh'\left(e^{-t}x + \sqrt{1-e^{-2t}}Z\right)\right] = \sqrt{1-e^{-2t}}\mathbb{E}\left[h''\left(e^{-t}x + \sqrt{1-e^{-2t}}Z\right)\right].$$

Therefore,

$$\dot{h}_t(x) = -e^{-t}x \cdot K_t h'(x) + e^{-2t}K_t h''(x). \tag{3.350}$$

On the other hand,

$$\begin{aligned}\mathcal{L}h_t(x) &= h''_t(x) - xh'_t(x) \\ &= e^{-2t}\mathbb{E}\left[h''\left(e^{-t}x + \sqrt{1-e^{-2t}}Z\right)\right] - xe^{-t}\mathbb{E}\left[h'\left(e^{-t}x + \sqrt{1-e^{-2t}}Z\right)\right] \\ &= e^{-2t}K_t h''(x) - e^{-t}xK_t h'(x).\end{aligned} \tag{3.351}$$

Comparing (3.350) and (3.351), we get (3.82).

The proof of the integration-by-parts formula (3.83) is more subtle, and relies on the fact that the Ornstein–Uhlenbeck process $\{Y_t\}_{t=0}^{\infty}$ with $Y_0 \sim G$ is stationary and *reversible* in the sense that, for any two $t, t' \ge 0$, $(Y_t, Y_{t'}) \overset{\mathrm{d}}{=} (Y_{t'}, Y_t)$. To see this, let

$$p^{(t)}(y|x) \triangleq \frac{1}{\sqrt{2\pi(1-e^{-2t})}} \exp\left(-\frac{(y-e^{-t}x)^2}{2(1-e^{-2t})}\right)$$

be the transition density of the OU(t) channel. Then it is not hard to establish that

$$p^{(t)}(y|x)\gamma(x) = p^{(t)}(x|y)\gamma(y), \qquad \forall x, y \in \mathbb{R}$$

(recall that γ denotes the standard Gaussian pdf). For $Z \sim G$ and any two smooth functions g, h, this implies that

$$\begin{aligned}\mathbb{E}[g(Z)K_t h(Z)] &= \mathbb{E}[g(Y_0)K_t h(Y_0)]\\ &= \mathbb{E}[g(Y_0)\mathbb{E}[h(Y_t)|Y_0]]\\ &= \mathbb{E}[g(Y_0)h(Y_t)]\\ &= \mathbb{E}[g(Y_t)h(Y_0)]\\ &= \mathbb{E}[K_t g(Y_0)h(Y_0)]\\ &= \mathbb{E}[K_t g(Z)h(Z)],\end{aligned}$$

where we have used (3.78) and the reversibility property of the Ornstein–Uhlenbeck process. Taking the derivative of both sides w.r.t. t, we conclude that

$$\mathbb{E}[g(Z)\mathcal{L}h(Z)] = \mathbb{E}[\mathcal{L}g(Z)h(Z)]. \tag{3.352}$$

In particular, since $\mathcal{L}1 = 0$ (where on the left-hand side 1 denotes the constant function $x \mapsto 1$), we have

$$\mathbb{E}[\mathcal{L}g(Z)] = \mathbb{E}[1\mathcal{L}g(Z)] = \mathbb{E}[g(Z)\mathcal{L}1] = 0 \tag{3.353}$$

for all smooth g.

Remark 54. If we consider the Hilbert space $L^2(G)$ of all functions $g : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}[g^2(Z)] < \infty$ with $Z \sim G$, then (3.352) expresses the fact that $\mathcal{L}$ is a self-adjoint linear operator on this space. Moreover, (3.353) shows that the constant functions are in the kernel of $\mathcal{L}$ (the closed linear subspace of $L^2(G)$ consisting of all g with $\mathcal{L}g = 0$).

We are now ready to prove (3.83). To that end, let us first define the operator Γ on pairs of functions g, h by

$$\Gamma(g,h) \triangleq \frac{1}{2}\left[\mathcal{L}(gh) - g\mathcal{L}h - h\mathcal{L}g\right]. \tag{3.354}$$

Remark 55. This operator was introduced into the study of Markov processes by Paul Meyer under the name “carré du champ” (French for “square of the field”). In the general theory, $\mathcal{L}$ can be any linear operator that serves as an infinitesimal generator of a Markov semigroup. Intuitively, Γ measures how far a given $\mathcal{L}$ is from being a derivation, where we say that an operator $\mathcal{L}$ acting on a function space is a *derivation* (or that it satisfies the *Leibniz rule*) if, for any g, h in its domain,

$$\mathcal{L}(gh) = g\mathcal{L}h + h\mathcal{L}g.$$

An example of a derivation is the first-order linear differential operator $\mathcal{L}g = g'$, in which case the Leibniz rule is simply the product rule of differential calculus.

Now, for our specific definition of $\mathcal{L}$, we have

$$\begin{aligned}\Gamma(g,h)(x) &= \frac{1}{2}\left[(gh)''(x) - x(gh)'(x) - g(x)\big(h''(x) - xh'(x)\big) - h(x)\big(g''(x) - xg'(x)\big)\right] \\ &= \frac{1}{2}\Big[g''(x)h(x) + 2g'(x)h'(x) + g(x)h''(x) \\ &\qquad - xg'(x)h(x) - xg(x)h'(x) - g(x)h''(x) + xg(x)h'(x) - g''(x)h(x) + xg'(x)h(x)\Big] \\ &= g'(x)h'(x), \qquad (3.355)\end{aligned}$$

or, more succinctly, $\Gamma(g,h) = g'h'$. Therefore,

$$\begin{aligned}\mathbb{E}[g(Z)\mathcal{L}h(Z)] &= \frac{1}{2}\Big\{\mathbb{E}[g(Z)\mathcal{L}h(Z)] + \mathbb{E}[h(Z)\mathcal{L}g(Z)]\Big\} && (3.356)\\ &= \frac{1}{2}\mathbb{E}[\mathcal{L}(gh)(Z)] - \mathbb{E}[\Gamma(g,h)(Z)] && (3.357)\\ &= -\mathbb{E}[g'(Z)h'(Z)], && (3.358)\end{aligned}$$

where (3.356) uses (3.352), (3.357) uses the definition (3.354) of Γ, and (3.358) uses (3.355) together with (3.353). This proves (3.83).

3.C Fano's inequality for list decoding

The following generalization of Fano's inequality has been used in the proof of Theorem 45: Let $\mathcal{X}$ and $\mathcal{Y}$ be finite sets, and let $(X,Y) \in \mathcal{X} \times \mathcal{Y}$ be a pair of jointly distributed random variables. Consider an arbitrary mapping $L : \mathcal{Y} \to 2^{\mathcal{X}}$ which maps any $y \in \mathcal{Y}$ to a set $L(y) \subseteq \mathcal{X}$. Let $P_{\rm e} = \mathbb{P}\left(X \notin L(Y)\right)$. Then

$$H(X|Y) \le h(P_{\rm e}) + (1-P_{\rm e})\mathbb{E}\left[\ln|L(Y)|\right] + P_{\rm e}\ln|\mathcal{X}| \qquad (3.359)$$

(see, e.g., [156] or [165, Lemma 1]).

To prove (3.359), define the indicator random variable $E \triangleq 1_{\{X \notin L(Y)\}}$. Then we can expand the conditional entropy $H(E,X|Y)$ in two ways as

$$\begin{aligned}H(E,X|Y) &= H(E|Y) + H(X|E,Y) && (3.360a)\\ &= H(X|Y) + H(E|X,Y). && (3.360b)\end{aligned}$$

Since X and Y uniquely determine E (for the given L), the quantity on the right-hand side of (3.360b) is equal to $H(X|Y)$. On the other hand, we can upper-bound the right-hand side of (3.360a) as

$$\begin{aligned}H(E|Y) + H(X|E,Y) &\le H(E) + H(X|E,Y) \\ &= h(P_{\rm e}) + \mathbb{P}(E=0)H(X|E=0,Y) + \mathbb{P}(E=1)H(X|E=1,Y) \\ &\le h(P_{\rm e}) + (1-P_{\rm e})\mathbb{E}\left[\ln|L(Y)|\right] + P_{\rm e}\ln|\mathcal{X}|,\end{aligned}$$

where the last line uses the fact that when $E=0$ (resp, $E=1$), the uncertainty about X is at most $\mathbb{E}[\ln|L(Y)|]$ (respectively, $\ln|\mathcal{X}|$). More precisely,

$$\begin{aligned}H(X|E=0,Y) &= -\sum_{y\in\mathcal{Y}} \mathbb{P}(Y=y,E=0) \sum_{x\in\mathcal{X}} \mathbb{P}(X=x|Y=y,E=0)\, \ln \mathbb{P}(X=x|Y=y,E=0) \\ &= -\sum_{y\in\mathcal{Y}} \mathbb{P}(Y=y,E=0) \sum_{x\in L(y)} \mathbb{P}(X=x|Y=y)\, \ln \mathbb{P}(X=x|Y=y) \\ &\le \sum_{y\in\mathcal{Y}} \mathbb{P}(Y=y,E=0)\, \ln|L(y)| \\ &\le \sum_{y\in\mathcal{Y}} \mathbb{P}(Y=y)\, \ln|L(y)| \\ &= \mathbb{E}\left[\ln|L(Y)|\right].\end{aligned}$$

In particular, when L is such that $L(Y) \le N$ a.s., we can apply Jensen's inequality to the second term on the right-hand side of (3.359) to get

$$H(X|Y) \le h(P_{\mathrm{e}}) + (1 - P_{\mathrm{e}}) \ln N + P_{\mathrm{e}} \ln |\mathcal{X}|.$$

This is precisely the inequality we used to derive the bound (3.265) in the proof of Theorem 45.

3.D Details for the derivation of (3.290)

Let $X^n \sim P_{X^n}$ and $Y^n \in \mathcal{Y}^n$ be the input and output sequences of a DMC with transition matrix $T : \mathcal{X} \to \mathcal{Y}$, where the DMC is used without feedback. In other words, $(X^n, Y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ is a random variable with $X^n \sim P_{X^n}$ and

$$P_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^n P_{Y|X}(y_i|x_i), \qquad \forall y^n \in \mathcal{Y}^n,\ \forall x^n \in \mathcal{X}^n \text{ s.t. } P_{X^n}(x^n) > 0.$$

Because the channel is memoryless and there is no feedback, the ith output symbol $Y_i \in \mathcal{Y}$ depends only on the ith input symbol $X_i \in \mathcal{X}$ and not on the rest of the input symbols $\overline{X}^i$. Consequently, $\overline{Y}^i \to X_i \to Y_i$ is a Markov chain for every $i = 1, \ldots, n$, so we can write

$$P_{Y_i|\overline{Y}^i}(y|\overline{y}^i) = \sum_{x \in \mathcal{X}} P_{Y_i|X_i}(y|x) P_{X_i|\overline{Y}^i}(x|\overline{y}^i) \tag{3.361}$$

$$= \sum_{x \in \mathcal{X}} P_{Y|X}(y|x) P_{X_i|\overline{Y}^i}(x|\overline{y}^i) \tag{3.362}$$

for all $y \in \mathcal{Y}$ and all $\overline{y}^i \in \mathcal{Y}^{n-1}$ such that $P_{\overline{Y}^i}(\overline{y}^i) > 0$. Therefore, for any two $y, y' \in \mathcal{Y}$ we have

$$\begin{aligned}
\ln \frac{P_{Y_i|\overline{Y}^i}(y|\overline{y}^i)}{P_{Y_i|\overline{Y}^i}(y'|\overline{y}^i)} &= \ln P_{Y_i|\overline{Y}^i}(y|\overline{y}^i) - \ln P_{Y_i|\overline{Y}^i}(y'|\overline{y}^i) \\
&= \ln \sum_{x \in \mathcal{X}} P_{Y|X}(y|x) P_{X_i|\overline{Y}^i}(x|\overline{y}^i) - \ln \sum_{x \in \mathcal{X}} P_{Y|X}(y'|x) P_{X_i|\overline{Y}^i}(x|\overline{y}^i) \\
&\le \max_{x \in \mathcal{X}} \ln P_{Y|X}(y|x) - \min_{x \in \mathcal{X}} \ln P_{Y|X}(y'|x).
\end{aligned}$$

Interchanging the roles of y and y', we get

$$\ln \frac{P_{Y_i|\overline{Y}^i}(y'|\overline{y}^i)}{P_{Y_i|\overline{Y}^i}(y|\overline{y}^i)} \le \max_{x, x' \in \mathcal{X}} \ln \frac{P_{Y|X}(y'|x)}{P_{Y|X}(y|x')}.$$

This implies, in turn, that

$$\left| \ln \frac{P_{Y_i|\overline{Y}^i}(y|\overline{y}^i)}{P_{Y_i|\overline{Y}^i}(y'|\overline{y}^i)} \right| \le \max_{x, x' \in \mathcal{X}} \max_{y, y' \in \mathcal{Y}} \left| \ln \frac{P_{Y|X}(y|x)}{P_{Y|X}(y'|x')} \right| = \frac{1}{2} c(T)$$

for all $y, y' \in \mathcal{Y}$.

Bibliography

[1] M. Talagrand, "A new look at independence," *Annals of Probability*, vol. 24, no. 1, pp. 1–34, January 1996.

[2] M. Ledoux, *The Concentration of Measure Phenomenon*, ser. Mathematical Surveys and Monographs. American Mathematical Society, 2001, vol. 89.

[3] G. Lugosi, "Concentration of measure inequalities - lecture notes," 2009. [Online]. Available: http://www.econ.upf.edu/~lugosi/anu.pdf.

[4] P. Massart, *The Concentration of Measure Phenomenon*, ser. Lecture Notes in Mathematics. Springer, 2007, vol. 1896.

[5] C. McDiarmid, "Concentration," in *Probabilistic Methods for Algorithmic Discrete Mathematics*. Springer, 1998, pp. 195–248.

[6] M. Talagrand, "Concentration of measure and isoperimteric inequalities in product space," *Publications Mathématiques de l'I.H.E.S*, vol. 81, pp. 73–205, 1995.

[7] K. Azuma, "Weighted sums of certain dependent random variables," *Tohoku Mathematical Journal*, vol. 19, pp. 357–367, 1967.

[8] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, March 1963.

[9] N. Alon and J. H. Spencer, *The Probabilistic Method*, 3rd ed. Wiley Series in Discrete Mathematics and Optimization, 2008.

[10] F. Chung and L. Lu, *Complex Graphs and Networks*, ser. Regional Conference Series in Mathematics. Wiley, 2006, vol. 107.

[11] ——, "Concentration inequalities and martingale inequalities: a survey," *Internet Mathematics*, vol. 3, no. 1, pp. 79–127, March 2006. [Online]. Available: http://www.ucsd.edu/~fan/wp/concen.pdf.

[12] T. J. Richardson and R. Urbanke, *Modern Coding Theory*. Cambridge University Press, 2008.

[13] N. Gozlan and C. Leonard, "Transport inequalities: a survey," *Markov Processes and Related Fields*, vol. 16, no. 4, pp. 635–736, 2010.

[14] J. M. Steele, *Probability Theory and Combinatorial Optimization*, ser. CBMS–NSF Regional Conference Series in Applied Mathematics. Siam, Philadelphia, PA, USA, 1997, vol. 69.

[15] S. B. Korada and N. Macris, "On the concentration of the capacity for a code division multiple access system," in *Proceedings of the 2007 IEEE International Symposium on Information Theory*, Nice, France, June 2007, pp. 2801–2805.

[16] ——, "Tight bounds on the capacity of binary input random CDMA systems," *IEEE Trans. on Information Theory*, vol. 56, no. 11, pp. 5590–5613, November 2010.

[17] S. Chatterjee, "Concentration inequalities with exchangeable pairs," Ph.D. dissertation, Stanford University, California, USA, February 2008. [Online]. Available: http://arxiv.org/abs/0507526.

[18] ——, "Stein's method for concentration inequalities," *Probability Theory and Related Fields*, vol. 138, pp. 305–321, 2007.

[19] S. Chatterjee and P. S. Dey, "Applications of Stein's method for concentration inequalities," *Annals of Probability*, vol. 38, no. 6, pp. 2443–2485, June 2010.

[20] N. Ross, "Fundamentals of Stein's method," *Probability Surveys*, vol. 8, pp. 210–293, 2011.

[21] E. Abbe and A. Montanari, "On the concentration of the number of solutions of random satisfiability formulas," 2010. [Online]. Available: http://arxiv.org/abs/1006.3786.

[22] S. Kudekar, "Statistical physics methods for sparse graph codes," Ph.D. dissertation, EPFL - Swiss Federal Institute of Technology, Lausanne, Switzeland, July 2009. [Online]. Available: http://infoscience.epfl.ch/record/138478/files/EPFL_TH4442.pdf.

[23] S. Kudekar and N. Macris, "Sharp bounds for optimal decoding of low-density parity-check codes," *IEEE Trans. on Information Theory*, vol. 55, no. 10, pp. 4635–4650, October 2009.

[24] A. Montanari, "Tight bounds for LDPC and LDGM codes under MAP decoding," *IEEE Trans. on Information Theory*, vol. 51, no. 9, pp. 3247–3261, September 2005.

[25] C. McDiarmid, "Centering sequences with bounded differences," *Combinatorics, Probability and Computing*, vol. 6, no. 1, pp. 79–86, March 1997.

[26] E. Shamir and J. Spencer, "Sharp concentration of the chromatic number on random graphs," *Combinatorica*, vol. 7, no. 1, pp. 121–129, 1987.

[27] M. G. Luby, Mitzenmacher, M. A. Shokrollahi, and D. A. Spielmann, "Efficient erasure-correcting codes," *IEEE Trans. on Information Theory*, vol. 47, no. 2, pp. 569–584, February 2001.

[28] T. J. Richardson and R. Urbanke, "The capacity of low-density parity-check codes under message-passing decoding," *IEEE Trans. on Information Theory*, vol. 47, no. 2, pp. 599–618, February 2001.

[29] M. Sipser and D. A. Spielman, "Expander codes," *IEEE Trans. on Information Theory*, vol. 42, no. 6, pp. 1710–1722, November 1996.

[30] A. B. Wagner, P. Viswanath, and S. R. Kulkarni, "Probability estimation in the rare-events regime," *IEEE Trans. on Information Theory*, vol. 57, no. 6, pp. 3207–3229, June 2011.

[31] K. Xenoulis and N. Kalouptsidis, "On the random coding exponent of nonlinear Gaussian channels," in *Proceedings of the 2009 IEEE International Workshop on Information Theory*, Volos, Greece, June 2009, pp. 32–36.

[32] ——, "Achievable rates for nonlinear Volterra channels," *IEEE Trans. on Information Theory*, vol. 57, no. 3, pp. 1237–1248, March 2011.

[33] K. Xenoulis, N. Kalouptsidis, and I. Sason, "New achievable rates for nonlinear Volterra channels via martingale inequalities," in *Proceedings of the 2012 IEEE International Workshop on Information Theory*, MIT, Boston, MA, USA, July 2012, pp. 1430–1434.

[34] M. Ledoux, "On Talagrand's deviation inequalities for product measures," *ESAIM: Probability and Statistics*, vol. 1, pp. 63–87, 1997.

[35] L. Gross, "Logarithmic Sobolev inequalities," *American Journal of Mathematics*, vol. 97, no. 4, pp. 1061–1083, 1975.

[36] A. J. Stam, "Some inequalities satisfied by the quantities of information of Fisher and Shannon," *Information and Control*, vol. 2, pp. 101–112, 1959.

[37] P. Federbush, "A partially alternate derivation of a result of Nelson," *Journal of Mathematical Physics*, vol. 10, no. 1, pp. 50–52, 1969.

[38] A. Dembo, T. M. Cover, and J. A. Thomas, "Information theoretic inequalities," *IEEE Trans. on Information Theory*, vol. 37, no. 6, pp. 1501–1518, November 1991.

[39] C. Villani, "A short proof of the 'concavity of entropy power'," *IEEE Trans. on Information Theory*, vol. 46, no. 4, pp. 1695–1696, July 2000.

[40] G. Toscani, "An information-theoretic proof of Nash's inequality," *Rendiconti Lincei: Matematica e Applicazioni*, 2012, in press.

[41] A. Guionnet and B. Zegarlinski, "Lectures on logarithmic Sobolev inequalities," *Séminaire de probabilités (Strasbourg)*, vol. 36, pp. 1–134, 2002.

[42] M. Ledoux, "Concentration of measure and logarithmic Sobolev inequalities," in *Séminaire de Probabilités XXXIII*, ser. Lecture Notes in Math. Springer, 1999, vol. 1709, pp. 120–216.

[43] G. Royer, *An Invitation to Logarithmic Sobolev Inequalities*, ser. SFM/AMS Texts and Monographs. American Mathematical Society and Société Mathématiques de France, 2007, vol. 14.

[44] S. G. Bobkov and F. Götze, "Exponential integrability and transportation cost related to logarithmic Sobolev inequalities," *Journal of Functional Analysis*, vol. 163, pp. 1–28, 1999.

[45] S. G. Bobkov and M. Ledoux, "On modified logarithmic Sobolev inequalities for Bernoulli and Poisson measures," *Journal of Functional Analysis*, vol. 156, no. 2, pp. 347–365, 1998.

[46] S. G. Bobkov and P. Tetali, "Modified logarithmic Sobolev inequalities in discrete settings," *Journal of Theoretical Probability*, vol. 19, no. 2, pp. 289–336, 2006.

[47] D. Chafaï, "Entropies, convexity, and functional inequalities: Φ-entropies and Φ-Sobolev inequalities," *J. Math. Kyoto University*, vol. 44, no. 2, pp. 325–363, 2004.

[48] C. P. Kitsos and N. K. Tavoularis, "Logarithmic Sobolev inequalities for information measures," *IEEE Trans. on Information Theory*, vol. 55, no. 6, pp. 2554–2561, June 2009.

[49] K. Marton, "Bounding $\bar{d}$-distance by informational divergence: a method to prove measure concentration," *Annals of Probability*, vol. 24, no. 2, pp. 857–866, 1996.

[50] C. Villani, *Topics in Optimal Transportation.* Providence, RI: American Mathematical Society, 2003.

[51] ——, *Optimal Transport: Old and New.* Springer, 2008.

[52] P. Cattiaux and A. Guillin, "On quadratic transportation cost inequalities," *Journal de Matématiques Pures et Appliquées*, vol. 86, pp. 342–361, 2006.

[53] A. Dembo, "Information inequalities and concentration of measure," *Annals of Probability*, vol. 25, no. 2, pp. 927–939, 1997.

[54] A. Dembo and O. Zeitouni, "Transportation approach to some concentration inequalities in product spaces," *Electronic Communications in Probability*, vol. 1, pp. 83–90, 1996.

[55] H. Djellout, A. Guillin, and L. Wu, "Transportation cost-information inequalities and applications to random dynamical systems and diffusions," *Annals of Probability*, vol. 32, no. 3B, pp. 2702–2732, 2004.

[56] N. Gozlan, "A characterization of dimension free concentration in terms of transportation inequalities," *Annals of Probability*, vol. 37, no. 6, pp. 2480–2498, 2009.

[57] E. Milman, "Properties of isoperimetric, functional and transport-entropy inequalities via concentration," *Probability Theory and Related Fields*, vol. 152, pp. 475–507, 2012.

[58] R. M. Gray, D. L. Neuhoff, and P. C. Shields, "A generalization of Ornstein's $\bar{d}$ distnace with applications to information theory," *Annals of Probability*, vol. 3, no. 2, pp. 315–328, 1975.

[59] R. M. Gray, D. L. Neuhoff, and J. K. Omura, "Process definitions of distortion-rate functions and source coding theorems," *IEEE Trans. on Information Theory*, vol. 21, no. 5, pp. 524–532, September 1975.

[60] R. Ahlswede, P. Gács, and J. Körner, "Bounds on conditional probabilities with applications in multi-user communication," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, vol. 34, pp. 157–177, 1976, see correction in vol. 39, no. 4, pp. 353–354, 1977.

[61] R. Ahlswede and G. Dueck, "Every bad code has a good subcode: a local converse to the coding theorem," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, vol. 34, pp. 179–182, 1976.

[62] K. Marton, "A simple proof of the blowing-up lemma," *IEEE Trans. on Information Theory*, vol. 32, no. 3, pp. 445–446, May 1986.

[63] V. Kostina and S. Verdú, "Fixed-length lossy compression in the finite blocklength regime," *IEEE Trans. on Information Theory*, vol. 58, no. 6, pp. 3309–3338, June 2012.

[64] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in finite blocklength regime," *IEEE Trans. on Information Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[65] C. McDiarmid, "On the method of bounded differences," in *Surveys in Combinatorics*. Cambridge University Press, 1989, vol. 141, pp. 148–188.

[66] M. J. Kearns and L. K. Saul, "Large deviation methods for approximate probabilistic inference," in *Proceedings of the 14th Conference on Uncertaintly in Artifical Intelligence*, San-Francisco, CA, USA, March 16-18 1998, pp. 311–319.

[67] D. Berend and A. Kontorovich, "On the concentration of the missing mass," 2012. [Online]. Available: http://arxiv.org/abs/1210.3248.

[68] S. G. From and A. W. Swift, "A refinement of Hoeffding's inequality," *Journal of Statistical Computation and Simulation*, pp. 1–7, December 2011.

[69] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. Springer, 1997.

[70] K. Dzhaparide and J. H. van Zanten, "On Bernstein-type inequalities for martingales," *Stochastic Processes and their Applications*, vol. 93, no. 1, pp. 109–117, May 2001.

[71] V. H. de la Pena, "A general class of exponential inequalities for martingales and ratios," *Annals of Probability*, vol. 27, no. 1, pp. 537–564, January 1999.

[72] A. Osekowski, "Weak type inequalities for conditionally symmetric martingales," *Statistics and Probability Letters*, vol. 80, no. 23-24, pp. 2009–2013, December 2010.

[73] ——, "Sharp ratio inequalities for a conditionally symmetric martingale," *Bulletin of the Polish Academy of Sciences Mathematics*, vol. 58, no. 1, pp. 65–77, 2010.

[74] G. Wang, "Sharp maximal inequalities for conditionally symmetric martingales and Brownian motion," *Proceedings of the American Mathematical Society*, vol. 112, no. 2, pp. 579–586, June 1991.

[75] D. Freedman, "On tail probabilities for martingales," *Annals of Probability*, vol. 3, no. 1, pp. 100–118, January 1975.

[76] I. Sason, "Tightened exponential bounds for discrete-time conditionally symmetric martingales with bounded increments," in *Proceedings of the 2012 International Workshop on Applied Probability*, Jerusalem, Israel, June 2012, p. 59.

[77] G. Bennett, "Probability inequalities for the sum of independent random variables," *Journal of the American Statistical Association*, vol. 57, no. 297, pp. 33–45, March 1962.

[78] P. Billingsley, *Probability and Measure*, 3rd ed. Wiley Series in Probability and Mathematical Statistics, 1995.

[79] G. Grimmett and D. Stirzaker, *Probability and Random Processes*, 3rd ed. Oxford University Press, 2001.

[80] I. Kontoyiannis, L. A. Latras-Montano, and S. P. Meyn, "Relative entropy and exponential deviation bounds for general Markov chains," in *Proceedings of the 2005 IEEE International Symposium on Information Theory*, Adelaide, Australia, September 2005, pp. 1563–1567.

[81] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley and Sons, 2006.

[82] I. Csiszár and P. C. Shields, *Information Theory and Statistics: A Tutorial*, ser. Foundations and Trends in Communications and Information Theory. Now Publishers, Delft, the Netherlands, 2004, vol. 1, no. 4.

[83] F. den Hollander, *Large Deviations*, ser. Fields Institute Monographs. American Mathematical Society, 2000.

[84] A. Réyni, "On measures of entropy and information," in *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, California, USA, 1961, pp. 547–561.

[85] A. Barg and G. D. Forney, "Random codes: minimum distances and error exponents," *IEEE Trans. on Information Theory*, vol. 48, no. 9, pp. 2568–2573, September 2002.

[86] M. Breiling, "A logarithmic upper bound on the minimum distance of turbo codes," *IEEE Trans. on Information Theory*, vol. 50, no. 8, pp. 1692–1710, August 2004.

[87] R. G. Gallager, "Low-Density Parity-Check Codes," Ph.D. dissertation, MIT, Cambridge, MA, USA, 1963.

[88] I. Sason, "On universal properties of capacity-approaching LDPC code ensembles," *IEEE Trans. on Information Theory*, vol. 55, no. 7, pp. 2956–2990, July 2009.

[89] T. Etzion, A. Trachtenberg, and A. Vardy, “Which codes have cycle-free Tanner graphs?” *IEEE Trans. on Information Theory*, vol. 45, no. 6, pp. 2173–2181, September 1999.

[90] M. G. Luby, Mitzenmacher, M. A. Shokrollahi, and D. A. Spielmann, “Improved low-density parity-check codes using irregular graphs,” *IEEE Trans. on Information Theory*, vol. 47, no. 2, pp. 585–598, February 2001.

[91] A. Kavčić, X. Ma, and M. Mitzenmacher, “Binary intersymbol interference channels: Gallager bounds, density evolution, and code performance bounds,” *IEEE Trans. on Information Theory*, vol. 49, no. 7, pp. 1636–1652, July 2003.

[92] R. Eshel, “Aspects of Convex Optimization and Concentration in Coding,” MSc thesis, Department of Electrical Engineering, Technion - Israel Institute of Technology, Haifa, Israel, February 2012.

[93] J. Douillard, M. Jezequel, C. Berrou, A. Picart, P. Didier, and A. Glavieux, “Iterative correction of intersymbol interference: turbo-equalization,” *Eurpoean Transactions on Telecommunications*, vol. 6, no. 1, pp. 507–511, September 1995.

[94] C. Méasson, A. Montanari, and R. Urbanke, “Maxwell construction: the hidden bridge between iterative and maximum apposteriori decoding,” *IEEE Trans. on Information Theory*, vol. 54, no. 12, pp. 5277–5307, December 2008.

[95] A. Shokrollahi, “Capacity-achieving sequences,” in *Volume in Mathematics and its Applications*, vol. 123, 2000, pp. 153–166.

[96] A. F. Molisch, *Wireless Communications.* John Wiley and Sons, 2005.

[97] G. Wunder, R. F. H. Fischer, H. Boche, S. Litsyn, and J. S. No, “The PAPR problem in OFDM transmission: new directions for a long-lasting problem,” accepted to the *IEEE Signal Processing Magazine*, December 2012. [Online]. Available: http://arxiv.org/abs/1212.2865.

[98] S. Litsyn and G. Wunder, “Generalized bounds on the crest-factor istribution of OFDM signals with applications to code design,” *IEEE Trans. on Information Theory*, vol. 52, no. 3, pp. 992–1006, March 2006.

[99] R. Salem and A. Zygmund, “Some properties of trigonometric series whose terms have random signs,” *Acta Mathematica*, vol. 91, no. 1, pp. 245–301, 1954.

[100] G. Wunder and H. Boche, “New results on the statistical distribution of the crest-factor of OFDM signals,” *IEEE Trans. on Information Theory*, vol. 49, no. 2, pp. 488–494, February 2003.

[101] S. Benedetto and E. Biglieri, *Principles of Digital Transmission with Wireless Applications.* Kluwer Academic/ Plenum Publishers, 1999.

[102] X. Fan, I. Grama, and Q. Liu, “Hoeffding’s inequality for supermartingales,” 2011. [Online]. Available: http://arxiv.org/abs/1109.4359.

[103] ——, “The missing factor in Bennett’s inequality,” 2012. [Online]. Available: http://arxiv.org/abs/1206.2592.

[104] I. Sason and S. Shamai, *Performance Analysis of Linear Codes under Maximum-Likelihood Decoding: A Tutorial*, ser. Foundations and Trends in Communications and Information Theory. Now Publishers, Delft, the Netherlands, July 2006, vol. 3, no. 1-2.

[105] J. A. Tropp, “User-friendly tail bounds for sums of random matrices,” *Foundations of Computational Mathematics*, vol. 12, no. 4, pp. 389–434, August 2012.

[106] ——, "Freedman's inequality for matrix martingales," *Electronic Communications in Probability*, vol. 16, pp. 262–270, March 2011.

[107] H. Chernoff, "A measure of asymptotic efficiency of tests of a hypothesis based on the sum of observations," *Annals of Mathematical Statistics*, vol. 23, no. 4, pp. 493–507, 1952.

[108] S. N. Bernstein, *The Theory of Probability.* Moscow/Leningrad: Gos. Izdat., 1927, in Russian.

[109] S. Verdú and T. Weissman, "The information lost in erasures," *IEEE Trans. on Information Theory*, vol. 54, no. 11, pp. 5030–5058, November 2008.

[110] E. A. Carlen, "Superadditivity of Fisher's information and logarithmic Sobolev inequalities," *Journal of Functional Analysis*, vol. 101, pp. 194–211, 1991.

[111] R. A. Adams and F. H. Clarke, "Gross's logarithmic Sobolev inequality: a simple proof," *American Journal of Mathematics*, vol. 101, no. 6, pp. 1265–1269, December 1979.

[112] G. Blower, *Random Matrices: High Dimensional Phenomena*, ser. London Mathematical Society Lecture Notes. Cambridge, U.K.: Cambridge University Press, 2009.

[113] O. Johnson, *Information Theory and the Central Limit Theorem.* London: Imperial College Press, 2004.

[114] E. H. Lieb and M. Loss, *Analysis*, 2nd ed. Providence, RI: American Mathematical Society, 2001.

[115] M. H. M. Costa and T. M. Cover, "On the similarity of the entropy power inequality and the Brunn–Minkowski inequality," *IEEE Trans. on Information Theory*, vol. 30, no. 6, pp. 837–839, November 1984.

[116] P. J. Huber and E. M. Ronchetti, *Robust Statistics*, 2nd ed. Wiley Series in Probability and Statistics, 2009.

[117] O. Johnson and A. Barron, "Fisher information inequalities and the central limit theorem," *Probability Theory and Related Fields*, vol. 129, pp. 391–409, 2004.

[118] S. Verdú, "Mismatched estimation and relative entropy," *IEEE Trans. on Information Theory*, vol. 56, no. 8, pp. 3712–3720, August 2010.

[119] H. L. van Trees, *Detection, Estimation and Modulation Theory, Part I.* Wiley, 1968.

[120] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions.* CRC Press, 1992.

[121] M. C. Mackey, *Time's Arrow: The Origins of Thermodynamic Behavior.* New York: Springer, 1992.

[122] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*, 5th ed. Berlin: Springer, 1998.

[123] I. Karatzas and S. Shreve, *Brownian Motion and Stochastic Calculus*, 2nd ed. Springer, 1988.

[124] F. C. Klebaner, *Introduction to Stochastic Calculus with Applications*, 2nd ed. Imperial College Press, 2005.

[125] T. van Erven and P. Harremoës, "Rényi divergence and Kullback–Leibler divergence," *IEEE Trans. on Information Theory*, 2012, submitted, 2012. [Online]. Available: http://arxiv.org/abs/1206.2459.

[126] A. Maurer, "Thermodynamics and concentration," *Bernoulli*, vol. 18, no. 2, pp. 434–454, 2012.

[127] S. Boucheron, G. Lugosi, and P. Massart, "Concentration inequalities using the entropy method," *Annals of Probability*, vol. 31, no. 3, pp. 1583–1614, 2003.

[128] I. Kontoyiannis and M. Madiman, "Measure concentration for compound Poisson distributions," *Electronic Communications in Probability*, vol. 11, pp. 45–57, 2006.

[129] B. Efron and C. Stein, "The jackknife estimate of variance," *Annals of Statistics*, vol. 9, pp. 586–596, 1981.

[130] J. M. Steele, "An Efron–Stein inequality for nonsymmetric statistics," *Annals of Statistics*, vol. 14, pp. 753–758, 1986.

[131] M. Gromov, *Metric Structures for Riemannian and Non-Riemannian Spaces.* Birkhäuser, 2001.

[132] S. Bobkov, "A functional form of the isoperimetric inequality for the Gaussian measure," *Journal of Functional Analysis*, vol. 135, pp. 39–49, 1996.

[133] L. V. Kantorovich, "On the translocation of masses," *Journal of Mathematical Sciences*, vol. 133, no. 4, pp. 1381–1382, 2006.

[134] E. Ordentlich and M. Weinberger, "A distribution dependent refinement of Pinsker's inequality," *IEEE Trans. on Information Theory*, vol. 51, no. 5, pp. 1836–1840, May 2005.

[135] D. Berend, P. Harremoës, and A. Kontorovich, "A reverse Pinsker inequality," 2012. [Online]. Available: http://arxiv.org/abs/1206.6544.

[136] I. Sason, "An information-theoretic perspective of the Poisson approximation via the Chen-Stein method," 2012. [Online]. Available: http://arxiv.org/abs/1206.6811.

[137] I. Kontoyiannis, P. Harremoës, and O. Johnson, "Entropy and the law of small numbers," *IEEE Trans. on Information Theory*, vol. 51, no. 2, pp. 466–472, February 2005.

[138] I. Csiszár, "Sanov property, generalized I-projection and a conditional limit theorem," *Annals of Probability*, vol. 12, no. 3, pp. 768–793, 1984.

[139] M. Talagrand, "Transportation cost for Gaussian and other product measures," *Geometry and Functional Analysis*, vol. 6, no. 3, pp. 587–600, 1996.

[140] R. M. Dudley, *Real Analysis and Probability.* Cambridge University Press, 2004.

[141] F. Otto and C. Villani, "Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality," *Journal of Functional Analysis*, vol. 173, pp. 361–400, 2000.

[142] Y. Wu, "On the HWI inequality," a work in progress.

[143] D. Cordero-Erausquin, "Some applications of mass transport to Gaussian-type inequalities," *Arch. Rational Mech. Anal.*, vol. 161, pp. 257–269, 2002.

[144] D. Bakry and M. Emery, "Diffusions hypercontractives," in *Séminaire de Probabilités XIX*, ser. Lecture Notes in Mathematics. Springer, 1985, vol. 1123, pp. 177–206.

[145] P.-M. Samson, "Concentration of measure inequalities for Markov chains and ϕ-mixing processes," *Annals of Probability*, vol. 28, no. 1, pp. 416–461, 2000.

[146] K. Marton, "A measure concentration inequality for contracting Markov chains," *Geometric and Functional Analysis*, vol. 6, pp. 556–571, 1996, see also erratum in *Geometric and Functional Analysis*, vol. 7, pp. 609–613, 1997.

[147] ——, "Measure concentration for Euclidean distance in the case of dependent random variables," *Annals of Probability*, vol. 32, no. 3B, pp. 2526–2544, 2004.

[148] ——, "Correction to 'Measure concentration for Euclidean distance in the case of dependent random variables'," *Annals of Probability*, vol. 38, no. 1, pp. 439–442, 2010.

[149] ——, "Bounding relative entropy by the relative entropy of local specifications in product spaces," 2009. [Online]. Available: http://arxiv.org/abs/0907.4491.

[150] ——, "An inequality for relative entropy and logarithmic Sobolev inequalities in Euclidean spaces," 2012. [Online]. Available: http://arxiv.org/abs/1206.4868.

[151] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge University Press, 2011.

[152] G. Margulis, "Probabilistic characteristics of graphs with large connectivity," *Problems of Information Transmission*, vol. 10, no. 2, pp. 174–179, 1974.

[153] A. El Gamal and Y. Kim, *Network Information Theory.* Cambridge University Press, 2011.

[154] G. Dueck, "Maximal error capacity regions are smaller than average error capacity regions for multi-user channels," *Problems of Control and Information Theory*, vol. 7, no. 1, pp. 11–19, 1978.

[155] F. M. J. Willems, "The maximal-error and average-error capacity regions for the broadcast channels are identical: a direct proof," *Problems of Control and Information Theory*, vol. 19, no. 4, pp. 339–347, 1990.

[156] R. Ahlswede and J. Körner, "Source coding with side information and a converse for degraded broadcast channels," *IEEE Trans. on Information Theory*, vol. 21, no. 6, pp. 629–637, November 1975.

[157] S. Shamai and S. Verdú, "The empirical distribution of good codes," *IEEE Trans. on Information Theory*, vol. 43, no. 3, pp. 836–846, May 1997.

[158] Y. Polyanskiy and S. Verdú, "Empirical distribution of good channel codes with non-vanishing error probability," January 2012, preprint. [Online]. Available: http://people.lids.mit.edu/yp/homepage/data/optcodes_journal.pdf.

[159] F. Topsøe, "An information theoretical identity and a problem involving capacity," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 291–292, 1967.

[160] J. H. B. Kemperman, "On the Shannon capacity of an arbitrary channel," *Indagationes Mathematicae*, vol. 36, pp. 101–115, 1974.

[161] U. Augustin, "Gedächtnisfreie Kanäle für diskrete Zeit," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, vol. 6, pp. 10–61, 1966.

[162] R. Ahlswede, "An elementary proof of the strong converse theorem for the multiple-access channel," *Journal of Combinatorics, Information and System Sciences*, vol. 7, no. 3, pp. 216–230, 1982.

[163] S. Shamai and I. Sason, "Variations on the Gallager bounds, connections and applications," *IEEE Trans. on Information Theory*, vol. 48, no. 12, pp. 3029–3051, December 2001.

[164] Y. Kontoyiannis, “Sphere-covering, measure concentration, and source coding,” *IEEE Trans. on Information Theory*, vol. 47, no. 4, pp. 1544–1552, May 2001.

[165] Y. Kim, A. Sutivong, and T. M. Cover, “State amplification,” *IEEE Trans. on Information Theory*, vol. 54, no. 5, pp. 1850–1859, May 2008.