Concentration of Measure Inequalities in Information Theory, Communications, and Coding

Concentration of Measure Inequalities in Information Theory, Communications, and Coding

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

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

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Communications and Information Theory

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

M. Raginsky and I. Sason. *Concentration of Measure Inequalities in Information Theory, Communications, and Coding*. Foundations and Trends® in Communications and Information Theory, vol. 10, no. 1-2, pp. 1–246, 2013.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-60198-724-2

Foundations and Trends® in Communications and Information Theory
Volume 10, Issue 1-2, 2013

Editorial Board

Editorial Scope

Topics

Foundations and Trends® in Communications and Information Theory publishes survey and tutorial articles in the following topics:

- Coded modulation
- Coding theory and practice
- Communication complexity
- Communication system design
- Cryptology and data security
- Data compression
- Data networks
- Demodulation and Equalization
- Denoising
- Detection and estimation
- Information theory and statistics
- Information theory and computer science
- Joint source/channel coding
- Modulation and signal design
- Multiuser detection
- Multiuser information theory
- Optical communication channels
- Pattern recognition and learning
- Quantization
- Quantum information processing
- Rate-distortion theory
- Shannon theory
- Signal processing for communications
- Source coding
- Storage and recording codes
- Speech and Image Compression
- Wireless Communications

Information for Librarians

Foundations and Trends® in Communications and Information Theory, 2013, Volume 10, 4 issues. ISSN paper version 1567-2190. ISSN online version 1567-2328. Also available as a combined paper and online subscription.

Foundations and Trends® in Communications and Information Theory
Vol. 10, No. 1-2 (2013) 1–246

DOI: 10.1561/0100000064

now
the essence of knowledge

Concentration of Measure Inequalities in Information Theory, Communications, and Coding

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

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

Contents

Abstract

During the last two decades, concentration inequalities have been the subject of exciting developments in various areas, including convex geometry, functional analysis, statistical physics, high-dimensional statistics, pure and applied probability theory (e.g., concentration of measure phenomena in random graphs, random matrices, and percolation), information theory, theoretical computer science, and learning theory. This monograph focuses 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. In addition to being a survey, this monograph also includes various new recent results derived by the authors.

The first part of the monograph introduces classical concentration inequalities for martingales, as well as some recent refinements and extensions. The power and versatility of the martingale approach is exemplified in the context of codes defined on graphs and iterative decoding algorithms, as well as codes for wireless communication.

The second part of the monograph introduces the entropy method, an information-theoretic technique for deriving concentration inequalities. The basic ingredients of the entropy method are discussed first in the context of logarithmic Sobolev inequalities, which underlie the so-called functional approach to concentration of measure, and then from a complementary information-theoretic viewpoint based on transportation-cost inequalities and probability in metric spaces. Some representative results on concentration for dependent random variables are briefly summarized, with emphasis on their connections to the entropy method. Finally, we discuss several applications of the entropy method to problems in communications and coding, including strong converses, empirical distributions of good channel codes, and an information-theoretic converse for concentration of measure.

M. Raginsky and I. Sason. *Concentration of Measure Inequalities in Information Theory, Communications, and Coding*. Foundations and Trends® in Communications and Information Theory, vol. 10, no. 1-2, pp. 1–246, 2013.

DOI: 10.1561/0100000064.

1

Introduction

1.1 An overview and a brief history

Concentration-of-measure inequalities provide bounds on the probability that a random variable X deviates from its mean, median or other typical value $\overline{x}$ by a given amount. These inequalities have been studied for several decades, with some fundamental and substantial contributions 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 often means that such a random variable X concentrates around $\overline{x}$ in a way that the probability of the event $\{|X - \overline{x}| \geq t\}$, for a given $t > 0$, decays exponentially in t. Detailed treatments of the concentration of measure phenomenon, including historical accounts, can be found, e.g., in [2], [3], [4], [5], [6] and [7].

In recent years, concentration inequalities have been intensively studied and used as a powerful tool in various areas. These include convex geometry, functional analysis, statistical physics, statistics, dynam-

ical systems, pure and applied probability (random matrices, Markov processes, random graphs, percolation), information theory, coding theory, learning theory, and theoretical computer science. Several techniques have been developed so far to establish concentration of measure. These include:

- The martingale approach (see, e.g., [6, 8, 9], [10, Chapter 7], [11, 12]), and its information-theoretic applications (see, e.g., [13] and references therein, [14]). This methodology will be covered in Chapter 2, which is focused on concentration inequalities for discrete-time martingales with bounded jumps, as well as on some of their potential applications in information theory, coding and communications. 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., [15] and [16]).

- The entropy method and logarithmic Sobolev inequalities (see, e.g., [3, Chapter 5], [4] and references therein). This methodology and its many remarkable links to information theory will be considered in Chapter 3.

- Transportation-cost inequalities that originated from information theory (see, e.g., [3, Chapter 6], [17], and references therein). This methodology, which is closely related to the entropy method and log-Sobolev inequalities, will be considered in Chapter 3.

- Talagrand's inequalities for product measures (see, e.g., [1], [6, Chapter 4], [7] and [18, Chapter 6]) and their links to information theory [19]. These inequalities proved to be very useful in combinatorial applications (such as the study of common and/or increasing subsequences), in statistical physics, and in functional analysis. We do not discuss Talagrand's inequalities in detail.

- Stein's method (or the method of exchangeable pairs) was recently used to prove concentration inequalities (see, e.g., [20], [21], [22] and [23]).

- Concentration inequalities that follow from rigorous methods in statistical physics (see, e.g., [24, 25, 26, 27, 28, 29, 30, 31]).
- The so-called reverse Lyapunov inequalities were recently used to derive concentration inequalities for multi-dimensional log-concave distributions [32] (see also a related work in [33]). The concentration inequalities in [32] imply an extension of the Shannon–McMillan–Breiman strong ergodic theorem to the class of discrete-time processes with log-concave marginals.

We do not address the last three items in this tutorial.

We now give a synopsis of some of the main ideas underlying the martingale approach (Chapter 2) and the entropy method (Chapter 3). Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a function that has *bounded differences*, i.e., the value of f changes at most by a bounded amount whenever any of its n input variables is changed arbitrarily while others are held fixed. A common method for proving concentration of such a function of n independent RVs around its expected value $\mathbb{E}[f]$ revolves around so-called McDiarmid's inequality or the "independent bounded-differences inequality" [6]. This inequality was originally proved via the martingale approach. Although the proof of this inequality has some similarity to the proof of the well-known Azuma–Hoeffding inequality, the bounded-difference assumption on f yields an improvement by a factor of 4 in the exponent. Some of its nice applications to algorithmic discrete mathematics were discussed in, e.g., [6, Section 3].

The Azuma–Hoeffding inequality is by now a well-known tool that has been often used to establish concentration results for discrete-time martingales whose jumps are bounded almost surely. It is due to Hoeffding [9], who proved this inequality for a sum of independent and bounded random variables, and to Azuma [8], who later extended it to bounded-difference martingales. This inequality was introduced to the computer science literature by Shamir and Spencer [34], who used it to prove concentration of the chromatic number for random graphs around its expected value. The chromatic number of a graph is defined as the minimal number of colors required to color all the vertices of this graph, so that no two adjacent vertices have the same color. Shamir and Spencer [34] established concentration of the chromatic number for

the so-called *Erdös–Rényi* ensemble of random graphs, where any pair of vertices is connected by an edge with probability $p \in (0, 1)$, independently of all other edges. It is worth noting that the concentration result in [34] was established without knowing the expected value of the chromatic number over this ensemble. This technique was later imported into coding theory in [35], [36] and [37], especially for exploring concentration phenomena pertaining to codes defined on graphs and iterative message-passing decoding algorithms. The last decade has seen an ever-expanding use of Azuma–Hoeffding inequality for proving concentration of measure in coding theory (see, e.g., [13] and references therein). In general, all these concentration inequalities serve to justify theoretically the ensemble approach to codes defined on graphs. However, much stronger concentration phenomena are observed in practice. The Azuma–Hoeffding inequality was also recently used in [38] 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 both the alphabet size and the source distribution vary the block length (so the empirical distribution does not converge to the true distribution as the block length tends to infinity). We also note that the Azuma–Hoeffding inequality was extended from martingales to "centering sequences" with bounded differences [39]; this extension provides sharper concentration results for, e.g., sequences that are related to sampling without replacement. In [40], [41] and [42], the martingale approach was also used to derive achievable rates and random coding error exponents for linear and nonlinear 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 [43], who found an alternative route to a class

of concentration inequalities for product measures originally derived by Talagrand [7] using an ingenious inductive technique. Specifically, Ledoux noticed that the well-known Chernoff bounding trick, which is discussed in detail in Section 2.2.1 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 [44], 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. By 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 $f \colon \mathbb{R}^n \to \mathbb{R}$ is any sufficiently smooth function with $\|\nabla f\| \leq 1$, can be bounded as $\Lambda(\lambda) \leq \lambda^2/2$. This bound then yields the optimal Gaussian concentration inequality $\mathbb{P}\left(|f(X^n) - \mathbb{E}[f(X^n)]| > t\right) \leq 2\exp\left(-t^2/2\right)$ 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 [45] in the context of information theory, and by Federbush [46] in the context of mathematical quantum field theory. Through the work of Stam [45], the Gaussian log-Sobolev inequality has been linked to several other information-theoretic notions, such as concavity of entropy power [47, 48, 49].)

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 can use the Herbst argument to pass from the log-Sobolev inequality to the bound $\ln \mathbb{E}[\exp(\lambda U)] \leq c\lambda^2/(2C)$, where $c > 0$ depends only on the distribution P, while $C > 0$ is determined by the energy content of f. While there is no general 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 [3, 5, 50, 51, 52], as well as in recent papers [53, 54, 55, 56, 57] and references therein.

Around the same time that Ledoux first introduced the entropy method in [43], Katalin Marton showed in a breakthrough paper [58] that 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, 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 — [59] offers a concise and fairly elementary introduction, while a more recent monograph [60] is a lot more detailed and encyclopedic. Multiple connections between optimal transportation, concentration of measure, and information theory are also explored in [17, 19, 61, 62, 63, 64, 65]. We also note that Wasserstein distances have been used in information theory in the context of lossy source coding [66, 67, 68].

The first explicit invocation of concentration inequalities in an information-theoretic context appears in the work of Ahlswede et

al. [69, 70]. These authors have shown that a certain delicate probabilistic inequality, which was referred to as the "blowing up lemma," and which we now (thanks to the contributions by Marton [58, 71]) 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 (see, e.g., [72, 73, 74, 75, 76, 77, 78, 79]). Thus, it is timely to revisit the use of concentration-of-measure ideas in information theory from a 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 researchers in information theory and related fields.

1.2 A reader's guide

This tutorial is mainly focused on the interplay between concentration of measure and information theory, as well as applications to problems related to information theory, communications and coding. For this reason, it is primarily aimed at researchers and graduate students working in these fields. The necessary mathematical background is real analysis, elementary functional analysis, and a first graduate course in probability theory and stochastic processes. As a refresher textbook for this mathematical background, the reader is referred, e.g., to [80].

Chapter 2 on the martingale approach is structured as follows: Section 2.1 lists key definitions pertaining to discrete-time martingales, while Section 2.2 presents several basic inequalities (including the celebrated Azuma–Hoeffding and McDiarmid inequalities) that form the basis of the martingale approach to concentration of measure. Section 2.3 focuses on several refined versions of the Azuma–Hoeffding inequality under additional moment conditions. Section 2.4 discusses the connections of the concentration inequalities introduced in Section 2.3 to classical limit theorems of probability theory, including central limit theorem for martingales and moderate deviations principle for i.i.d.

real-valued random variables. Section 2.5 forms the second part of the chapter, applying the concentration inequalities from Section 2.3 to information theory and some related topics. Section 2.6 gives a brief summary of the chapter.

Several very nice surveys on concentration inequalities via the martingale approach are available, including [6], [10, Chapter 11], [11, Chapter 2] and [12]. The main focus of Chapter 2 is on the presentation of some old and new concentration inequalities that form the basis of 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 the 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 recur 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 [71, 58] 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. Section 3.7 concludes with a brief summary.

2

Concentration Inequalities via the Martingale Approach

This chapter introduces concentration inequalities for discrete-time martingales with bounded increments, and shows several of their potential applications to information theory and related topics. We start by introducing the basic concentration inequalities of Azuma–Hoeffding and McDiarmid, as well as various refinements under additional moment conditions. We then move to applications, which include concentration for codes defined on graphs, concentration for OFDM signals, and the derivation of achievable rates and lower bounds on the error exponents for random codes with ML decoding over linear or non-linear communication channels.

2.1 Discrete-time martingales

We start with a brief review of martingales to set definitions and notation.

Definition 2.1 (Discrete-time martingales)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A sequence $\{X_i, \mathcal{F}_i\}_{i=0}^{n}$, $n \in \mathbb{N}$, 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. The $\mathcal{F}_i$'s form a *filtration*, i.e., $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}_n$; usually, $\mathcal{F}_0$ is the trivial σ-algebra $\{\emptyset, \Omega\}$ and $\mathcal{F}_n$ is the full σ-algebra $\mathcal{F}$.
2. $X_i \in \mathbb{L}^1(\Omega, \mathcal{F}_i, \mathbb{P})$ for every $i \in \{0, \ldots, n\}$; this means that each X_i is defined on the same sample space 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 $i \in \{1, \ldots, n\}$, the equality $X_{i-1} = \mathbb{E}[X_i|\mathcal{F}_{i-1}]$ holds almost surely (a.s.).

Here are some useful facts about martingales.

Fact 1. Since $\{\mathcal{F}_i\}_{i=0}^n$ is a filtration, it follows from the tower property for conditional expectations that, almost surely,

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

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 expectations of every term X_i of a martingale sequence are all equal to $\mathbb{E}[X_0]$.

Fact 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 $\{\mathcal{F}_i\}_{i=0}^n$, let

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

Then, the sequence $X_0, X_1, \ldots, X_n$ forms a martingale (with respect to the above filtration) since

1. The RV $X_i = \mathbb{E}[X|\mathcal{F}_i]$ is $\mathcal{F}_i$-measurable, and $\mathbb{E}[|X_i|] \leq \mathbb{E}[|X|] < \infty$.
2. By construction $\{\mathcal{F}_i\}_{i=0}^n$ is a filtration.
3. For every $i \in \{1, \ldots, n\}$

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

In the particular case where $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_n = \mathcal{F}$, we see that $X_0, X_1, \ldots, X_n$ is a martingale sequence with

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

That is, we get 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, we don't know anything about X, so we estimate it 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 2.1. Let $\{U_k\}_{k=1}^n$ be independent random variables on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and assume that $\mathbb{E}[U_k] = 0$ and $\mathbb{E}[|U_k|] < \infty$ for every k. Let us define

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

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

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

Note that $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$ denotes the minimal σ-algebra that includes all the sets of the form $\{\omega \in \Omega : (X_1(\omega) \leq \alpha_1, \ldots, X_k(\omega) \leq \alpha_k)\}$ where $\alpha_j \in \mathbb{R} \cup \{-\infty, +\infty\}$ for $j \in \{1, \ldots, k\}$. It is easy to verify that $\{X_k, \mathcal{F}_k\}_{k=0}^n$ is a martingale sequence; this 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.

If we relax the equality in the definition of a martingale to an inequality, we obtain sub- and super-martingales. More precisely, to define sub- and super-martingales, we keep the first two conditions in Definition 2.1, but replace the equality in the third condition by one of the following:

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

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

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

- If $\{X_i, \mathcal{F}_i\}$ is a martingale, h is a convex (concave) function and $\mathbb{E}\big[|h(X_i)|\big] < \infty$, 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 2.2. 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

We now turn to the main topic of the chapter, namely the martingale approach to proving concentration inequalities, i.e., sharp bounds on the *deviation probabilities* $\mathbb{P}\left(|U - \mathbb{E}U| \geq r\right)$ for all $r \geq 0$, where $U \in \mathbb{R}$ is a random variable with some additional "structure" — for instance, U may be a function of a large number n of independent or weakly dependent random variables $X_1, \ldots, X_n$. In a nutshell, the martingale approach has two basic ingredients:

1. **The martingale decomposition** — we first construct a suitable filtration $\{\mathcal{F}_i\}_{i=0}^n$ on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ that carries U, where $\mathcal{F}_0$ is the trivial σ-algebra $\{\emptyset, \Omega\}$ and $\mathcal{F}_n = \mathcal{F}$. Then

we decompose the difference $U - \mathbb{E}U$ as

$$\begin{aligned} U - \mathbb{E}U &= \mathbb{E}[U|\mathcal{F}_n] - \mathbb{E}[U|\mathcal{F}_0] \\ &= \sum_{i=1}^{n} \left(\mathbb{E}[U|\mathcal{F}_i] - \mathbb{E}[U|\mathcal{F}_{i-1}]\right). \end{aligned} \tag{2.2.1}$$

The idea is to choose the σ-algebras $\mathcal{F}_i$ in such a way that the increments $\xi_i = \mathbb{E}[U|\mathcal{F}_i] - \mathbb{E}[U|\mathcal{F}_{i-1}]$ are bounded in some sense, e.g., almost surely.

2. **The Chernoff bounding technique** — 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* $\Lambda(t) \triangleq \ln \mathbb{E}[\exp(tU)]$, $t \in \mathbb{R}$. Moreover, exploiting the martingale decomposition (2.2.1), we may write

$$\Lambda(t) = \ln \mathbb{E}\left[\prod_{i=1}^{n} \exp(t\xi_i)\right],$$

which allows us to focus on the behavior of individual terms $\exp(t\xi_i)$, $i = 1, \ldots, n$. Now, the logarithmic moment-generating function plays a key role in the theory of large deviations [81], which can be thought of as a (mainly) asymptotic analysis of the concentration of measure phenomenon. Thus, its prominent appearance here is not entirely unexpected.

There are more sophisticated variants of the martingale approach, some of which we will have occasion to see later on, but the above two ingredients are a good starting point. In the remainder of this section, we will elaborate on these ideas and examine their basic consequences.

2.2.1 The Chernoff bounding technique and the Hoeffding lemma

The first ingredient of the martingale method is the well-known Chernoff bounding technique[1]: Using Markov's inequality, for any $t > 0$ we

[1]The name of H. Chernoff is associated with this technique because of his 1952 paper [82]; however, its roots go back to S.N. Bernstein's 1927 textbook on the theory of probability [83].

have

$$\begin{aligned}\mathbb{P}(U \geq r) &= \mathbb{P}\big(\exp(tU) \geq \exp(tr)\big) \\ &\leq \exp(-tr)\mathbb{E}[\exp(tU)].\end{aligned}$$

Equivalently, if we define the *logarithmic moment generating function* $\Lambda(t) \triangleq \ln \mathbb{E}[\exp(tU)]$, $t \in \mathbb{R}$, we can write

$$\mathbb{P}(U \geq r) \leq \exp\big(\Lambda(t) - rt\big), \qquad \forall t > 0. \tag{2.2.2}$$

To bound the probability of the lower tail, $\mathbb{P}(U \leq -r)$, we follow the same steps, but with $-U$ instead of U. Now the success of the whole enterprize hinges on our ability to obtain tight upper bounds on $\Lambda(t)$. One of the basic tools available for that purpose is the following lemma due to Hoeffding [9]:

Lemma 2.2.1 (Hoeffding)**.** Let $U \in \mathbb{R}$ be a random variable, such that $U \in [a, b]$ a.s. for some finite $a < b$. Then, for any $t \geq 0$,

$$\mathbb{E}\left[\exp\big(t(U - \mathbb{E}U)\big)\right] \leq \exp\left(\frac{t^2(b-a)^2}{8}\right). \tag{2.2.3}$$

Proof. For any $p \in [0, 1]$ and any $\lambda \in \mathbb{R}$, let us define the function

$$H_p(\lambda) \triangleq \ln\left(pe^{\lambda(1-p)} + (1-p)e^{-\lambda p}\right). \tag{2.2.4}$$

Let $\xi = U - \mathbb{E}U$, where $\xi \in [a - \mathbb{E}U, b - \mathbb{E}U]$. Using the convexity of the exponential function, we can write

$$\begin{aligned}\exp(t\xi) &= \exp\left(\frac{U-a}{b-a} \cdot t(b - \mathbb{E}U) + \frac{b-U}{b-a} \cdot t(a - \mathbb{E}U)\right) \\ &\leq \left(\frac{U-a}{b-a}\right)\exp\big(t(b - \mathbb{E}U)\big) + \left(\frac{b-U}{b-a}\right)\exp\big(t(a - \mathbb{E}U)\big).\end{aligned}$$

Taking expectations of both sides, we get

$$\begin{aligned}\mathbb{E}[\exp(t\xi)] &\leq \left(\frac{\mathbb{E}U-a}{b-a}\right)\exp\big(t(b - \mathbb{E}U)\big) + \left(\frac{b-\mathbb{E}U}{b-a}\right)\exp\big(t(a - \mathbb{E}U)\big) \\ &= \exp\big(H_p(\lambda)\big) \qquad (2.2.5)\end{aligned}$$

where we have let

$$p = \frac{\mathbb{E}U - a}{b-a} \qquad \text{and} \qquad \lambda = t(b-a).$$

Using a Taylor's series expansion, we get the bound $H_p(\lambda) \leq \frac{\lambda^2}{8}$ for all $p \in [0,1]$ and all $\lambda \in \mathbb{R}$. Substituting this bound into (2.2.5) and using the above definitions of p and λ, we get (2.2.3). □

2.2.2 The Azuma–Hoeffding inequality

The Azuma–Hoeffding inequality[2], stated in Theorem 2.2.2 below, is a useful concentration inequality for bounded-difference martingales. It was proved by Hoeffding [9] for independent bounded random variables, followed by a discussion on sums of dependent random variables; and was later generalized by Azuma [8] to the more general setting of bounded-difference martingales. The proof of the Azuma–Hoeffding inequality that we present below is a nice concrete illustration of the general approach outlined in the beginning of this section. Moreover, we will have many occasions to revisit this proof in order to obtain various refinements of the original Azuma–Hoeffding bound.

Theorem 2.2.2 (Azuma–Hoeffding inequality)**.** Let $\{X_k, \mathcal{F}_k\}_{k=0}^n$ be a real-valued martingale sequence. Suppose that there exist nonnegative reals $d_1, \ldots, d_n$, such that $|X_k - X_{k-1}| \leq d_k$ a.s. for all $k \in \{1, \ldots, n\}$. Then, for every $r > 0$,

$$\mathbb{P}(|X_n - X_0| \geq r) \leq 2 \exp\left(-\frac{r^2}{2\sum_{k=1}^n d_k^2}\right). \tag{2.2.6}$$

Proof. For an arbitrary $r > 0$,

$$\mathbb{P}(|X_n - X_0| \geq r) = \mathbb{P}(X_n - X_0 \geq r) + \mathbb{P}(X_n - X_0 \leq -r). \tag{2.2.7}$$

Let $\xi_i \triangleq X_i - X_{i-1}$ for $i = 1, \ldots, n$ denote the jumps of the martingale sequence. By hypothesis, $|\xi_k| \leq d_k$ and $\mathbb{E}[\xi_k \,|\, \mathcal{F}_{k-1}] = 0$ a.s. for every $k \in \{1, \ldots, n\}$.

[2]The Azuma–Hoeffding inequality is also known as Azuma's inequality. Since we refer to it numerous times in this chapter, we will just call it Azuma's inequality for the sake of brevity.

We now apply the Chernoff technique:

$$\begin{aligned}
&\mathbb{P}(X_n - X_0 \geq r) \\
&= \mathbb{P}\left(\sum_{i=1}^{n} \xi_i \geq r\right) \\
&\leq e^{-rt}\, \mathbb{E}\left[\exp\left(t \sum_{i=1}^{n} \xi_i\right)\right], \quad \forall\, t \geq 0.
\end{aligned} \tag{2.2.8}$$

Furthermore,

$$\begin{aligned}
&\mathbb{E}\Big[\exp\Big(t \sum_{i=1}^{n} \xi_i\Big)\Big] \\
&= \mathbb{E}\left[\mathbb{E}\left[\exp\Big(t \sum_{i=1}^{n} \xi_i\Big) \,\Big|\, \mathcal{F}_{n-1}\right]\right] \\
&= \mathbb{E}\left[\exp\Big(t \sum_{i=1}^{n-1} \xi_i\Big)\, \mathbb{E}[\exp(t\xi_n) \,|\, \mathcal{F}_{n-1}]\right]
\end{aligned} \tag{2.2.9}$$

where the last equality holds since $Y \triangleq \exp(t \sum_{k=1}^{n-1} \xi_k)$ is $\mathcal{F}_{n-1}$-measurable. We now apply the Hoeffding lemma to ξ_n conditioned on $\mathcal{F}_{n-1}$. Indeed, we know that $\mathbb{E}[\xi_n|\mathcal{F}_{n-1}] = 0$ and that $\xi_n \in [-d_n, d_n]$ a.s., so Lemma 2.2.1 gives that a.s.

$$\mathbb{E}[\exp(t\xi_n) \,|\, \mathcal{F}_{n-1}] \leq \exp\left(\frac{t^2\, d_n^2}{2}\right). \tag{2.2.10}$$

Continuing recursively in a similar manner, we can bound the quantity in (2.2.9) by

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

Substituting this bound into (2.2.8), we obtain

$$\mathbb{P}(X_n - X_0 \geq r) \leq \exp\left(-rt + \frac{t^2}{2} \sum_{k=1}^{n} d_k^2\right), \quad \forall\, t \geq 0. \tag{2.2.11}$$

Finally, choosing $t = r \left(\sum_{k=1}^{n} d_k^2\right)^{-1}$ to minimize the right-hand side of (2.2.11), we get

$$\mathbb{P}(X_n - X_0 \geq r) \leq \exp\left(-\frac{r^2}{2 \sum_{k=1}^{n} d_k^2}\right). \tag{2.2.12}$$

Since $\{X_k, \mathcal{F}_k\}$ 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 $\mathbb{P}(X_n - X_0 \leq -r)$. Using these bounds in (2.2.7), we complete the proof of Theorem 2.2.2.

□

Remark 2.1. In [6, 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 $r \geq 0$,

$$\mathbb{P}\left(\left|\sum_{k=1}^n Y_k\right| \geq r\right) \leq 2\exp\left(-\frac{2r^2}{\sum_{k=1}^n (b_k - a_k)^2}\right). \tag{2.2.13}$$

Consider a 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 $k \in \{1, \ldots, n\}$. Then it is easy to see that $\{Y_k, \mathcal{F}_k\}_{k=0}^n$ is a martingale difference sequence. Since $\sum_{k=1}^n Y_k = X_n - X_0$, it follows from (2.2.13) that

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

Example 2.3. Let $\{Y_i\}_{i=0}^\infty$ be i.i.d. binary random variables which take values $\pm d$ with equal probability, where $d > 0$ is some constant. 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 generated by $Y_0, \ldots, Y_k$. 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

$$\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) \leq 2\exp\left(-\frac{\alpha^2}{2d^2}\right) \tag{2.2.14}$$

for every $\alpha \geq 0$ and $n \in \mathbb{N}$. Since the RVs $\{Y_i\}_{i=0}^\infty$ are i.i.d. with zero mean and variance d^2, the Central Limit Theorem (CLT) says

that $\frac{1}{\sqrt{n}}(X_n - X_0) = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} Y_k$ converges in distribution to $\mathcal{N}(0, d^2)$. Therefore, for every $\alpha \geq 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) = 2\,Q\Big(\frac{\alpha}{d}\Big) \tag{2.2.15}$$

where

$$Q(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp\Big(-\frac{t^2}{2}\Big)\mathrm{d}t, \quad \forall\, x \in \mathbb{R} \tag{2.2.16}$$

is the complementary standard Gaussian CDF (also known as the Q-function), for which we have the following exponential upper and lower bounds (see, e.g., [84, Section 3.3]):

$$\frac{1}{\sqrt{2\pi}} \frac{x}{1+x^2} \cdot \exp\left(-\frac{x^2}{2}\right) < Q(x) < \frac{1}{\sqrt{2\pi}\,x} \cdot \exp\left(-\frac{x^2}{2}\right), \quad \forall\, x > 0 \tag{2.2.17}$$

From (2.2.15) and (2.2.17) it follows that the exponent on the right-hand side of (2.2.14) is exact.

Example 2.4. Fix some $\gamma \in (0, 1]$. Let us generalize Example 2.3 above by considering the case where the i.i.d. binary RVs $\{Y_i\}_{i=0}^{\infty}$ have the probability law

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

Therefore, each Y_i has zero mean and variance $\sigma^2 = \gamma d^2$. Define the martingale sequence $\{X_k, \mathcal{F}_k\}_{k=0}^{\infty}$ as in Example 2.3. By 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 $\alpha \geq 0$

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) = 2\,Q\bigg(\frac{\alpha}{\sqrt{\gamma}\,d}\bigg). \tag{2.2.18}$$

From the bounds on the Q-function given in (2.2.17), it follows that the right-hand side of (2.2.18) scales exponentially like $e^{-\frac{\alpha^2}{2\gamma d^2}}$. Hence, the exponent in this example is improved by a factor $\frac{1}{\gamma}$ compared to Azuma's inequality (which gives the same bound as in Example 2.3 since $|X_k - X_{k-1}| \leq d$ for every $k \in \mathbb{N}$). This indicates that a refinement of Azuma's inequality is possible if additional information on the variance is available. Refinements of this sort were studied extensively in the probability literature, and they are the focus of Section 2.3.

2.2.3 McDiarmid's inequality

A prominent application of the martingale approach is the derivation of a powerful inequality due to McDiarmid (see [39, Theorem 3.1] or [85]), also known as the *bounded difference inequality*. Let $\mathcal{X}$ be some set, and let $f\colon \mathcal{X}^n \to \mathbb{R}$ be a function that satisfies the *bounded difference assumption*

$$\sup_{x_1,\ldots,x_n,x_i' \in \mathcal{X}} \Big| f(x_1,\ldots,x_i,\ldots,x_n) - f(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n) \Big| \le d_i \tag{2.2.19}$$

for every $1 \le i \le n$, where $d_1,\ldots,d_n$ are some nonnegative real constants. This is equivalent to saying that, for any given i, the variation of the function f with respect to its ith coordinate is upper bounded by d_i. (We assume that each argument of f takes values in the same set $\mathcal{X}$ mainly for simplicity of presentation; an extension to different domains for each variable is not difficult.)

Theorem 2.2.3 (McDiarmid's inequality)**.** Let $\{X_k\}_{i=1}^n$ be independent (not necessarily identically distributed) random variables taking values in some measurable space $\mathcal{X}$. Consider a random variable $U = f(X^n)$, where $f\colon \mathcal{X}^n \to \mathbb{R}$ is a measurable function satisfying the bounded difference assumption (2.2.19). Then, for every $r \ge 0$,

$$\mathbb{P}\left(\left|U - \mathbb{E}U\right| \ge r\right) \le 2\exp\left(-\frac{2r^2}{\sum_{k=1}^n d_k^2}\right). \tag{2.2.20}$$

Remark 2.2. One can use the Azuma–Hoeffding inequality for a derivation of a concentration inequality in the considered setting. However, the following proof provides an improvement by a factor of 4 in the exponent of the bound.

Proof. Let $\mathcal{F}_0$ be the trivial σ-algebra, and for $k \in \{1,\ldots,n\}$ let $\mathcal{F}_k = \sigma(X_1,\ldots,X_k)$ be the σ-algebra generated by $X_1,\ldots,X_k$. For every $k \in \{1,\ldots,n\}$, define

$$\xi_k \triangleq \mathbb{E}\big[f(X_1,\ldots,X_n)\,|\,\mathcal{F}_k\big] - \mathbb{E}\big[f(X_1,\ldots,X_n)\,|\,\mathcal{F}_{k-1}\big]. \tag{2.2.21}$$

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

$$\begin{aligned}\mathbb{E}\big[f(X_1,\ldots,X_n)\,|\,\mathcal{F}_0\big] &= \mathbb{E}\big[f(X_1,\ldots,X_n)\big]\\ \mathbb{E}\big[f(X_1,\ldots,X_n)\,|\,\mathcal{F}_n\big] &= f(X_1,\ldots,X_n).\end{aligned} \tag{2.2.22}$$

Hence, it follows from the last three equalities that

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

In the following, we need a lemma:

Lemma 2.2.4. For every $k \in \{1,\ldots,n\}$, the following properties hold a.s.:

1. $\mathbb{E}[\xi_k\,|\,\mathcal{F}_{k-1}] = 0$, so $\{\xi_k, \mathcal{F}_k\}$ is a martingale-difference and 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 and $\mathcal{F}_{k-1}$-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 proves the first item. The second item follows from the first and third items. To prove the third item, note that $\xi_k = f_k(X_1,\ldots,X_k)$ holds a.s. for some $\mathcal{F}_k$-measurable function $f_k\colon \mathcal{X}^k \to \mathbb{R}$. Let us define, for every $k \in \{1,\ldots,n\}$,

$$\begin{aligned}A_k &\triangleq \inf_{x\in\mathcal{X}} f_k(X_1,\ldots,X_{k-1},x),\\ B_k &\triangleq \sup_{x\in\mathcal{X}} f_k(X_1,\ldots,X_{k-1},x)\end{aligned}$$

which are $\mathcal{F}_{k-1}$-measurable[3], and by definition $\xi_k \in [A_k, B_k]$ holds almost surely. Furthermore, for every point $(x_1,\ldots,x_{k-1}) \in \mathcal{X}^{k-1}$, we

[3]This is certainly the case if $\mathcal{X}$ is countably infinite. For uncountable spaces, one needs to introduce some regularity conditions to guarantee measurability of infima and suprema. We choose not to dwell on these technicalities here to keep things simple; the book by van der Vaart and Wellner [86] contains a thorough treatment of these issues.

obtain that a.s.

$$
\begin{aligned}
&\sup_{x\in\mathcal{X}} f_k(x_1,\ldots,x_{k-1},x) - \inf_{x'\in\mathcal{X}} f_k(x_1,\ldots,x_{k-1},x') \\
&= \sup_{x,x'\in\mathcal{X}} \bigl\{f_k(x_1,\ldots,x_{k-1},x) - f_k(x_1,\ldots,x_{k-1},x')\bigr\} \\
&= \sup_{x,x'\in\mathcal{X}} \Bigl\{\mathbb{E}\bigl[f(X_1,\ldots,X_n)\,\big|\,X_1=x_1,\ldots,X_{k-1}=x_{k-1},X_k=x\bigr] \\
&\qquad\qquad - \mathbb{E}\bigl[f(X_1,\ldots,X_n)\,\big|\,X_1=x_1,\ldots,X_{k-1}=x_{k-1},X_k=x'\bigr]\Bigr\} \\
&= \sup_{x,x'\in\mathcal{X}} \Bigl\{\mathbb{E}\bigl[f(x_1,\ldots,x_{k-1},x,X_{k+1},\ldots,X_n)\bigr] \\
&\qquad\qquad - \mathbb{E}\bigl[f(x_1,\ldots,x_{k-1},x',X_{k+1},\ldots,X_n)\bigr]\Bigr\} \qquad (2.2.23) \\
&= \sup_{x,x'\in\mathcal{X}} \Bigl\{\mathbb{E}\bigl[f(x_1,\ldots,x_{k-1},x,X_{k+1},\ldots,X_n) \\
&\qquad\qquad - f(x_1,\ldots,x_{k-1},x',X_{k+1},\ldots,X_n)\bigr]\Bigr\} \\
&\le d_k \qquad (2.2.24)
\end{aligned}
$$

where (2.2.23) follows from the independence of the random variables $\{X_k\}_{k=1}^n$, and (2.2.24) follows from the condition in (2.2.19). Hence, it follows that $B_k - A_k \le d_k$ a.s., which then implies that $\xi_k \in [A_k, A_k + d_k]$. Since $\mathbb{E}[\xi_k \,|\, \mathcal{F}_{k-1}] = 0$ then a.s. the $\mathcal{F}_{k-1}$-measurable function A_k is non-positive. Note that the third item of the lemma gives better control on the range of ξ_k than what we had in the proof of the Azuma–Hoeffding inequality (there we showed that ξ_k was contained in the interval $[-d_k, d_k]$, which is twice as long as $[A_k, A_k + d_k]$.) □

We now proceed in the same manner as in the proof of the Azuma–Hoeffding inequality. Specifically, for any fixed $k \in \{1,\ldots,n\}$, $\xi_k \in [A_k, A_k + d_k]$ a.s., where A_k is $\mathcal{F}_{k-1}$-measurable, and $\mathbb{E}[\xi_k|\mathcal{F}_{k-1}] = 0$. Thus, we may apply the Hoeffding lemma to ξ_k conditionally on $\mathcal{F}_{k-1}$ to get

$$\mathbb{E}\Bigl[e^{t\xi_k}\Big|\mathcal{F}_{k-1}\Bigr] \le \exp\left(\frac{t^2 d_k^2}{8}\right).$$

Similarly to the proof of the Azuma–Hoeffding inequality, by repeatedly

using the recursion in (2.2.9), the last inequality implies that

$$\mathbb{E}\Big[\exp\Big(t\sum_{k=1}^{n}\xi_k\Big)\Big] \le \exp\left(\frac{t^2}{8}\sum_{k=1}^{n}d_k^2\right) \tag{2.2.25}$$

which then gives from (2.2.8) that, for every $t \ge 0$,

$$\begin{aligned}
&\mathbb{P}(f(X_1,\ldots,X_n) - \mathbb{E}[f(X_1,\ldots,X_n)] \ge r) \\
&\quad = \mathbb{P}\left(\sum_{i=1}^{n}\xi_i \ge r\right) \\
&\quad \le \exp\left(-rt + \frac{t^2}{8}\sum_{k=1}^{n}d_k^2\right).
\end{aligned} \tag{2.2.26}$$

The choice $t = 4r\left(\sum_{k=1}^{n} d_k^2\right)^{-1}$ minimizes the expression in (2.2.26), so

$$\mathbb{P}\Big(f(X_1,\ldots,X_n) - \mathbb{E}[f(X_1,\ldots,X_n)] \ge r\Big) \le \exp\left(-\frac{2r^2}{\sum_{k=1}^{n} d_k^2}\right). \tag{2.2.27}$$

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

$$\mathbb{P}(f(X_1,\ldots,X_n) - \mathbb{E}[f(X_1,\ldots,X_n)] \le -r),$$

so by the union bound we get (2.2.20). This completes the proof of Theorem 2.2.3. □

2.2.4 Hoeffding's inequality and its improved version (the Kearns-Saul inequality)

The following concentration inequality for sums of independent bounded random variables, originally due to Hoeffding (see [9, Theorem 2]), can be viewed as a special case of McDiarmid's inequality:

Theorem 2.2.5 (Hoeffding's 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,

$$\mathbb{P}\left(\left|\sum_{k=1}^{n}U_k - \mu_n\right| \ge r\right) \le 2\exp\left(-\frac{2r^2}{\sum_{k=1}^{n}(b_k-a_k)^2}\right), \quad \forall\, \alpha \ge 0. \tag{2.2.28}$$

Proof. Apply Theorem 2.2.3 to $f(u^n) \triangleq \sum_{k=1}^n u_k$, where $u^n \in \prod_{k=1}^n [a_k, b_k]$. □

Recall that a key step in the proof of McDiarmid's inequality is to invoke Hoeffding's lemma (Lemma 2.2.1). However, a careful look at the proof of Lemma 2.2.1 reveals a potential source of slack in the bound

$$\ln \mathbb{E}\left[\exp(tU)\right] \leq t\,\mathbb{E}U + \frac{t^2(b-a)^2}{8}$$

— namely, that this bound is the same regardless of the *location* of the mean $\mathbb{E}U$ relative to the endpoints of the interval $[a, b]$. As it turns out, one does indeed obtain an improved version of Hoeffding's inequality by making use of this information. This improved version was first stated by Kearns and Saul [87]. We note that Berend and Kontorovich [88] recently identified a gap in the original proof in [87] and were able to close this gap with some tedious calculus. A shorter information-theoretic proof of the basic inequality that underlies the Kearns–Saul inequality is given in Section 3.4.3 of Chapter 3). Here, we only state this basic inequality without proof, and then use it to derive the Kearns–Saul improvement of Hoeffding's inequality.

As before, consider a random variable $U \in [a, b]$ (a.s.) and let $\xi = U - \mathbb{E}U$. Recall the inequality (2.2.5):

$$\mathbb{E}[\exp(t\xi)] \leq \exp\left(H_p(\lambda)\right),$$

where $H_p(\lambda)$ is defined in (2.2.4), and

$$p = \frac{\mathbb{E}U - a}{b-a} \qquad \text{and} \qquad \lambda = t(b-a).$$

Earlier, we have used the bound $H_p(\lambda) \leq \lambda^2/8$ valid for all $p \in [0, 1]$ and all $\lambda \in \mathbb{R}$. However, a more careful analysis of the function $H_p(\lambda)$ [88, Theorem 4] shows that, for any $\lambda \in \mathbb{R}$,

$$H_p(\lambda) \leq \begin{cases} \dfrac{(1-2p)\lambda^2}{4\ln(\frac{1-p}{p})} & \text{if } p \neq \frac{1}{2} \\ \dfrac{\lambda^2}{8}, & \text{if } p = \frac{1}{2} \end{cases}. \tag{2.2.29}$$

Note that

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

which means that the upper bound in (2.2.29) is continuous in p, and it also improves upon the earlier bound $H_p(t) \le t^2/8$. Thus, consider a real-valued random variable U almost surely bounded between a and b, and let

$$p \triangleq \frac{\mathbb{E}U - a}{b-a}.$$

Then

$$\ln \mathbb{E}\left[\exp\left(t(U-\mathbb{E}U)\right)\right] \le ct^2(b-a)^2, \tag{2.2.30}$$

where

$$c = \begin{cases} \dfrac{1-2p}{4\ln\left(\frac{1-p}{p}\right)}, & \text{if } p \neq \dfrac{1}{2} \\ \dfrac{1}{8}, & \text{if } p = \dfrac{1}{2} \end{cases}$$

Thus, we see that the original Hoeffding bound of Lemma 2.2.1 is tight only when the mean $\mathbb{E}U$ of $U \in [a,b]$ is located exactly at the midpoint of the support interval $[a,b]$; when $\mathbb{E}U$ is closer to one of the endpoints, the new bound (2.2.30) provides a strict improvement.

Applying this improved bound to each $\xi_k = U_k - \mathbb{E}U_k$, we get the Kearns–Saul inequality [87]:

Theorem 2.2.6 (Kearns–Saul inequality)**.** Let $\{U_k\}_{k=1}^n$ be a sequence of independent and bounded random variables such that, for every $k \in \{1,\dots,n\}$, $U_k \in [a_k, b_k]$ 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,

$$\mathbb{P}\left(\left|\sum_{k=1}^n U_k - \mu_n\right| \ge r\right) \le 2\exp\left(-\frac{r^2}{4\sum_{k=1}^n c_k\,(b_k-a_k)^2}\right), \quad \forall\, r \ge 0 \tag{2.2.31}$$

where

$$c_k = \begin{cases} \dfrac{1-2p_k}{4\ln\left(\frac{1-p_k}{p_k}\right)}, & \text{if } p_k \neq \dfrac{1}{2} \\ \dfrac{1}{8}, & \text{if } p_k = \dfrac{1}{2} \end{cases} \qquad \text{with } p_k = \frac{\mathbb{E}U_k - a_k}{b_k - a_k}. \tag{2.2.32}$$

Moreover, the exponential bound (2.2.31) improves Hoeffding's inequality, unless $p_k = \frac{1}{2}$ for every $k \in \{1, \ldots, n\}$.

We also refer the reader to another recent refinement of Hoeffding's inequality in [89], followed by some numerical comparisons.

2.3 Refined versions of the Azuma–Hoeffding inequality

Example 2.4 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 a.s.

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

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

2.3.1 Martingales with bounded jumps

One of the disadvantages of Azuma–Hoeffding inequality (Theorem 2.2.2) and McDiarmid's inequality (Theorem 2.2.3) is their insensitivity to the variance, which leads to suboptimal exponents compared to the central limit theorem (CLT) and moderate deviation principle (MDP). The following theorem, which appears in [85] (see also [81, Corollary 2.4.7]), makes use of the variance:

Theorem 2.3.1. Let $\{X_k, \mathcal{F}_k\}_{k=0}^n$ be a discrete-time real-valued martingale. Assume that, for some constants $d, \sigma > 0$, the following two requirements are satisfied a.s. for every $k \in \{1, \ldots, n\}$:

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

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

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) \leq 2 \exp\left(-n\, d\Big(\frac{\delta+\gamma}{1+\gamma}\Big\|\frac{\gamma}{1+\gamma}\Big)\right) \tag{2.3.1}$$

where

$$\gamma \triangleq \frac{\sigma^2}{d^2}, \quad \delta \triangleq \frac{\alpha}{d} \tag{2.3.2}$$

and

$$d(p\|q) \triangleq p\ln\Big(\frac{p}{q}\Big) + (1-p)\ln\Big(\frac{1-p}{1-q}\Big), \quad \forall\, p, q \in [0,1] \tag{2.3.3}$$

is the divergence between the Bernoulli(p) and Bernoulli(q) measures. If $\delta > 1$, then the probability on the left-hand side of (2.3.1) is equal to zero.

Proof. The proof of this bound goes along the same lines as the proof of the Azuma–Hoeffding inequality, up to (2.2.9). The new ingredient in this proof is the use of the so-called Bennett's inequality (see, e.g., [81, Lemma 2.4.1]), which improves upon Lemma 2.2.1 by incorporating a bound on the variance: 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 $\lambda \geq 0$, Bennett's inequality states that

$$\mathbb{E}\big[e^{\lambda X}\big] \leq \frac{e^{\lambda\overline{x}}\left[(b-\overline{x})^2 e^{-\frac{\lambda\sigma^2}{b-\overline{x}}} + \sigma^2 e^{\lambda(b-\overline{x})}\right]}{(b-\overline{x})^2 + \sigma^2}. \tag{2.3.4}$$

We give the proof of (2.3.4) in Appendix 2.A for completeness.

We now apply Bennett's inequality (2.3.4) to the conditional law of ξ_k given the σ-algebra $\mathcal{F}_{k-1}$. Since $\mathbb{E}[\xi_k|\mathcal{F}_{k-1}] = 0$, $\mathsf{var}[\xi_k|\mathcal{F}_{k-1}] \leq \sigma^2$ and $\xi_k \leq d$ a.s. for $k \in \mathbb{N}$, we have

$$\mathbb{E}\left[\exp(t\xi_k)\,|\,\mathcal{F}_{k-1}\right] \leq \frac{\sigma^2\exp(td) + d^2\exp\left(-\frac{t\sigma^2}{d}\right)}{d^2+\sigma^2}, \qquad \text{a.s..} \tag{2.3.5}$$

From (2.2.9) and (2.3.5) it follows that, for every $t \geq 0$,

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

Repeating this argument inductively, we conclude that, for every $t \geq 0$,

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

Using the definition of γ in (2.3.2), we can rewrite this inequality as

$$\mathbb{E}\Big[\exp\Big(t\sum_{k=1}^{n}\xi_k\Big)\Big] \le \Big(\frac{\gamma\exp(td)+\exp(-\gamma td)}{1+\gamma}\Big)^n, \quad \forall\, t \ge 0. \tag{2.3.6}$$

Let $x \triangleq td$ (so $x \ge 0$). We can now use (2.3.6) with the Chernoff bounding technique to get, for every $\alpha \ge 0$ (where from the definition of δ in (2.3.2), $\alpha t = \delta x$),

$$\begin{aligned}
&\mathbb{P}(X_n - X_0 \ge \alpha n)\\
&\le \exp(-\alpha n t)\,\mathbb{E}\Big[\exp\Big(t\sum_{k=1}^{n}\xi_k\Big)\Big]\\
&\le \left(\frac{\gamma\exp\big((1-\delta)x\big)+\exp\big(-(\gamma+\delta)x\big)}{1+\gamma}\right)^n, \quad \forall\, x \ge 0. \qquad (2.3.7)
\end{aligned}$$

Consider first the case where $\delta = 1$ (i.e., $\alpha = d$). Then (2.3.7) becomes

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

and the expression on the right-hand side is minimized in the limit as $x \to \infty$. This gives the inequality

$$\mathbb{P}(X_n - X_0 \ge dn) \le \Big(\frac{\gamma}{1+\gamma}\Big)^n. \tag{2.3.8}$$

Otherwise, if $\delta \in [0,1)$, we minimize the base of the exponent on the right-hand side of (2.3.7) with respect to the free non-negative parameter x, and the optimal choice is

$$x = \Big(\frac{1}{1+\gamma}\Big)\ln\Big(\frac{\gamma+\delta}{\gamma(1-\delta)}\Big). \tag{2.3.9}$$

Substituting (2.3.9) into the right-hand side of (2.3.7) gives

$$\begin{aligned}
\mathbb{P}(X_n - X_0 \ge \alpha n) &\le \left[\Big(\frac{\gamma+\delta}{\gamma}\Big)^{-\frac{\gamma+\delta}{1+\gamma}}(1-\delta)^{-\frac{1-\delta}{1+\gamma}}\right]^n\\
&= \exp\left(-n\,d\Big(\frac{\delta+\gamma}{1+\gamma}\Big\|\frac{\gamma}{1+\gamma}\Big)\right), \qquad \forall \alpha \ge 0.
\end{aligned} \tag{2.3.10}$$

Finally, if $\delta > 1$ (i.e., if $\alpha > d$), the exponent is equal to $+\infty$. Applying inequality (2.3.10) to the martingale $\{-X_k, \mathcal{F}_k\}_{k=0}^{\infty}$ gives the same upper bound to the other tail probability $\mathbb{P}(X_n - X_0 \leq -\alpha n)$. Overall, we get the bound (2.3.1), which completes the proof of Theorem 2.3.1. □

Here is an illustration of how one can use Theorem 2.3.1 to get better bounds compared to Azuma's inequality:

Example 2.5. 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 $m \in \mathbb{N}$

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

This implies that $\mathbb{E}[X_m - X_{m-1} \,|\, \mathcal{F}_{m-1}] = 0$ a.s. for every $m \in \mathbb{N}$, and, since X_{m-1} is $\mathcal{F}_{m-1}$-measurable, we have $\mathbb{E}[X_m \,|\, \mathcal{F}_{m-1}] = X_{m-1}$ almost surely. Moreover, since $\varepsilon \in (0, \frac{1}{2}]$,

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

so Azuma's inequality gives

$$\mathbb{P}(X_k \geq kx) \leq \exp\left(-\frac{kx^2}{2d^2}\right), \qquad \forall x \geq 0 \tag{2.3.11}$$

independently of the value of ε (note that $X_0 = 0$ a.s.). However, we can use Theorem 2.3.1 to get a better bound: Because, for every $m \in \mathbb{N}$,

$$\mathbb{E}\big[(X_m - X_{m-1})^2 \,|\, \mathcal{F}_{m-1}\big] = \frac{d^2\varepsilon}{1-\varepsilon}, \qquad \text{a.s.}$$

it follows from (2.3.10) that

$$\mathbb{P}(X_k \geq kx) \leq \exp\left(-k\, d\Big(\frac{x(1-\varepsilon)}{d} + \varepsilon \,\Big\|\, \varepsilon\Big)\right), \qquad \forall x \geq 0. \tag{2.3.12}$$

Consider the case where $\varepsilon \to 0$. Then, for arbitrary $x > 0$ and $k \in \mathbb{N}$, Azuma's inequality in (2.3.11) provides an upper bound that is strictly positive independently of ε, whereas the one-sided concentration inequality of Theorem 2.3.1 implies a bound in (2.3.12) that tends to zero.

Remark 2.3. As pointed out by McDiarmid [6, Section 2], all the concentration inequalities for martingales that rely on the Chernoff bounding technique can be converted into analogous inequalities for 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 with respect to the same filtration (see Section 2.1), so it implies that $\{\exp(t(X_k - X_0)), \mathcal{F}_k\}_{k=0}^{\infty}$ is a sub-martingale for every $t \geq 0$. Hence, by applying Doob's maximal inequality for sub-martingales, it follows that, for every $\alpha \geq 0$ and $t \geq 0$,

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

which parallels the proof of Theorem 2.3.1 starting from (2.2.8). This idea applies to all the concentration inequalities derived in this chapter.

Corollary 2.3.2. Let $\{X_k, \mathcal{F}_k\}_{k=0}^{n}$ be a discrete-time 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 $\alpha \geq 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) \leq 2 \exp\left(-n f(\delta)\right) \tag{2.3.13}$$

where $\delta \triangleq \frac{\alpha}{d}$,

$$f(\delta) = \begin{cases} \ln(2)\Big[1 - h_2\left(\frac{1-\delta}{2}\right)\Big], & 0 \leq \delta \leq 1 \\ +\infty, & \delta > 1 \end{cases} \tag{2.3.14}$$

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

Proof. By substituting $\gamma = 1$ in Theorem 2.3.1 (since there is no constraint on the conditional variance, one can take $\sigma^2 = d^2$), the corresponding exponent in (2.3.1) is equal to

$$d\Big(\frac{1+\delta}{2}\Big\|\frac{1}{2}\Big) = f(\delta),$$

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

Remark 2.4. Corollary 2.3.2, which is a special case of Theorem 2.3.1 with $\gamma = 1$, is a tightening of the Azuma–Hoeffding inequality in the case 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 2.3.1 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.

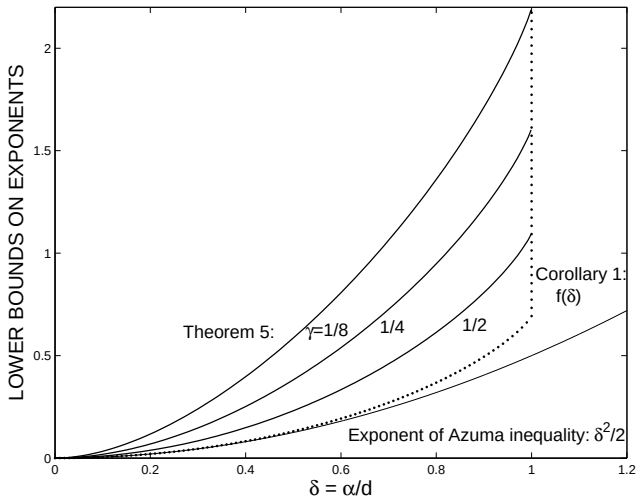

Figure 2.1: Plot of the lower bounds on the exponents from Azuma's inequality and the improved bounds in Theorem 2.3.1 and Corollary 2.3.2 (where f is defined in (2.3.14)). The pointed line refers to the exponent in Corollary 2.3.2, and the three solid lines for $\gamma = \frac{1}{8}, \frac{1}{4}$ and $\frac{1}{2}$ refer to the exponents in Theorem 2.3.1.

Theorem 2.3.1 can also be used to analyze the probabilities of *small deviations*, i.e., events of the form $\{|X_n - X_0| \geq \alpha\sqrt{n}\}$ for an arbitrary $\alpha \geq 0$ (this is in contrast to *large-deviation* events of the form $\{|X_n - X_0| \geq \alpha n\}$):

Proposition 2.1. Let $\{X_k, \mathcal{F}_k\}$ be a discrete-time real-valued martingale that satisfies the conditions of Theorem 2.3.1. Then, for every $\alpha \geq 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) \leq 2\exp\Big(-\frac{\delta^2}{2\gamma}\Big)\Big(1 + O(n^{-\frac{1}{2}})\Big). \tag{2.3.15}$$

Remark 2.5. From Proposition 2.1, 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 $\frac{1}{\gamma}$.

Proof. Let $\{X_k, \mathcal{F}_k\}_{k=0}^{\infty}$ be a discrete-time martingale that satisfies the conditions in Theorem 2.3.1. From (2.3.1), for every $\alpha \geq 0$ and $n \in \mathbb{N}$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) \leq 2\exp\left(-n\, d\Big(\frac{\delta_n + \gamma}{1+\gamma}\Big\|\frac{\gamma}{1+\gamma}\Big)\right) \tag{2.3.16}$$

where, following (2.3.2),

$$\gamma \triangleq \frac{\sigma^2}{d^2}, \quad \delta_n \triangleq \frac{\frac{\alpha}{\sqrt{n}}}{d} = \frac{\delta}{\sqrt{n}}\,. \tag{2.3.17}$$

With these definitions, we have

$$d\Big(\frac{\delta_n + \gamma}{1+\gamma}\Big\|\frac{\gamma}{1+\gamma}\Big) = \frac{\gamma}{1+\gamma}\Bigg[\left(1 + \frac{\delta}{\gamma\sqrt{n}}\right)\ln\left(1 + \frac{\delta}{\gamma\sqrt{n}}\right) + \frac{1}{\gamma}\left(1 - \frac{\delta}{\sqrt{n}}\right)\ln\left(1 - \frac{\delta}{\sqrt{n}}\right)\Bigg]. \tag{2.3.18}$$

Using the series expansion

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

in (2.3.18), we see that, for every $n > \frac{\delta^2}{\gamma^2}$,

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

Substituting this into the exponent on the right-hand side of (2.3.16) gives (2.3.15). □

2.3.2 Inequalities for sub- and super-martingales

Upper bounds on the probability $\mathbb{P}(X_n - X_0 \geq r)$ for $r \geq 0$, derived earlier in this section for martingales, can be adapted to super-martingales (similarly to, e.g., [11, Chapter 2] or [12, Section 2.7]). Alternatively, by replacing $\{X_k, \mathcal{F}_k\}_{k=0}^n$ with $\{-X_k, \mathcal{F}_k\}_{k=0}^n$, we may obtain upper bounds on the probability $\mathbb{P}(X_n - X_0 \leq -r)$ for sub-martingales. For example, the adaptation of Theorem 2.3.1 to sub- and super-martingales gives the following inequality:

Corollary 2.3.3. Let $\{X_k, \mathcal{F}_k\}_{k=0}^\infty$ be a discrete-time real-valued super-martingale. Assume that, for some constants $d, \sigma > 0$, the following two requirements are satisfied a.s.:

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

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

$$\mathbb{P}(X_n - X_0 \geq \alpha n) \leq \exp\left(-n\, d\Big(\frac{\delta+\gamma}{1+\gamma}\Big\|\frac{\gamma}{1+\gamma}\Big)\right) \tag{2.3.19}$$

where γ and δ are defined as in (2.3.2), and the binary divergence $d(p\|q)$ is introduced in (2.3.3). Alternatively, if $\{X_k, \mathcal{F}_k\}_{k=0}^\infty$ is a sub-martingale, the same upper bound in (2.3.19) 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 2.3.1. The only difference is that, for a super-martingale,

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

a.s., where $\xi_k \triangleq X_k - \mathbb{E}[X_k \,|\, \mathcal{F}_{k-1}]$ is $\mathcal{F}_k$-measurable (recall the discussion of the properties of super-martingales in Section 2.1). Hence, $\mathbb{P}\big((X_n - X_0 \geq \alpha n) \leq \mathbb{P}\big(\sum_{k=1}^n \xi_k \geq \alpha n\big)$ where, a.s., $\xi_k \leq d$, $\mathbb{E}[\xi_k \,|\, \mathcal{F}_{k-1}] = 0$, and $\mathsf{var}(\xi_k \,|\, \mathcal{F}_{k-1}) \leq \sigma^2$. The rest of the proof coincides with the proof of Theorem 2.3.1 (starting from (2.2.8)). 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 Relation of the refined inequalities to classical results

2.4.1 The martingale central limit theorem

In this section, we comment on the relationship between the martingale small-deviation bound of Proposition 2.1 and the martingale central limit theorem (CLT).

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Given a filtration $\{\mathcal{F}_k\}$, we say that $\{Y_k, \mathcal{F}_k\}_{k=0}^{\infty}$ is 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 $|Y_k| \le d$ 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 (or weakly) to the Gaussian distribution $\mathcal{N}(0, \sigma^2)$; we denote this convergence by $\frac{S_n}{\sqrt{n}} \Rightarrow \mathcal{N}(0, \sigma^2)$. (There exist more general versions of this statement — see, e.g., [90, pp. 475–478]).

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

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

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

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

and $Y_0 \triangleq 0$. Then $\{Y_k, \mathcal{F}_k\}_{k=0}^{\infty}$ is a martingale-difference sequence, and $|Y_k| \leq d$ a.s. for every $k \in \mathbb{N} \cup \{0\}$. Assume also that there exists a constant $\sigma > 0$, such that, for all k,

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

Let's first assume that this inequality holds a.s. with equality. Then it follows from the martingale CLT that

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

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

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

where the Q-function is defined in (2.2.16). In terms of the notation in (2.3.2), we have $\frac{\alpha}{\sigma} = \frac{\delta}{\sqrt{\gamma}}$, so that

$$\lim_{n\to\infty} \mathbb{P}\Big(|X_n - X_0| \geq \alpha\sqrt{n}\Big) = 2\,Q\Big(\frac{\delta}{\sqrt{\gamma}}\Big). \tag{2.4.1}$$

From the fact that

$$Q(x) \leq \frac{1}{2}\exp\left(-\frac{x^2}{2}\right), \qquad \forall x \geq 0$$

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 large-n limit of the inequality in Proposition 2.1, except for the additional factor of 2. Note also that the proof of Proposition 2.1 is applicable for finite n, and not only in the asymptotic regime $n \to \infty$. Furthermore, from the exponential upper and lower bounds on the Q-function in (2.2.17) and from (2.4.1), it follows that the exponent $\frac{\delta^2}{2\gamma}$ in the concentration inequality (2.3.15) cannot be improved without additional assumptions.

2.4.2 The moderate deviations principle

The moderate deviations principle (MDP) on the real line (see, e.g., [81, Theorem 3.7.1]) states the following: 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 $\sigma^2 = \mathsf{var}(X_i) > 0$. 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

$$Z_n \triangleq \sqrt{\frac{a_n}{n}} \sum_{i=1}^n X_i, \quad \forall\, n \in \mathbb{N}. \tag{2.4.2}$$

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

$$\begin{aligned} -\frac{1}{2\sigma^2} \inf_{x \in \Gamma^0} x^2 &\le \liminf_{n\to\infty} a_n \ln \mathbb{P}(Z_n \in \Gamma) \\ &\le \limsup_{n\to\infty} a_n \ln \mathbb{P}(Z_n \in \Gamma) \le -\frac{1}{2\sigma^2} \inf_{x \in \overline{\Gamma}} x^2 \end{aligned} \tag{2.4.3}$$

where Γ^0 and $\overline{\Gamma}$ denote, respectively, the interior and the closure 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 (2.4.2),

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

so, by the MDP,

$$\lim_{n\to\infty} n^{1-2\eta} \ln \mathbb{P}\left(\Big| \sum_{i=1}^n X_i \Big| \ge \alpha n^\eta \right) = -\frac{\alpha^2}{2\sigma^2}, \qquad \forall \alpha \ge 0 \tag{2.4.4}$$

We show in Appendix 2.B that, in contrast to Azuma's inequality, Theorem 2.3.1 provides an upper bound on the probability

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

which coincides with the asymptotic limit in (2.4.4). The analysis in Appendix 2.B provides another interesting link between Theorem 2.3.1 and a classical result in probability theory, and thus emphasizes the significance of the refinements of Azuma's inequality.

2.4.3 Functions of discrete-time Markov chains

A striking well-known relation between discrete-time Markov chains and martingales is the following (see, e.g., [91, p. 473]): Let $\{X_n\}_{n=0}^{\infty}$ be a discrete-time Markov chain taking values in a countable state space $\mathcal{S}$ with transition matrix $\mathbf{P}$. Let $\psi : \mathcal{S} \to \mathbb{R}$ be a *harmonic function* of the Markov chain, i.e.,

$$\sum_{j \in \mathcal{S}} p_{i,j} \psi(j) = \psi(i), \quad \forall\, i \in \mathcal{S}$$

and assume also that $\mathbb{E}[|\psi(X_n)|] < \infty$ for every n. Let $Y_n \triangleq \psi(X_n)$ for every n. Then $\{Y_n, \mathcal{F}_n\}_{n=0}^{\infty}$ is a martingale where $\{\mathcal{F}_n\}_{n=0}^{\infty}$ is the natural filtration. This relation, which follows directly from the Markov property, allows us to apply the concentration inequalities in Section 2.3 to bounded harmonic functions of Markov chains (the boundedness is needed to guarantee that the jumps of the resulting martingale sequence are uniformly bounded).

Exponential deviation bounds for an important class of Markov chains, so-called Doeblin chains (they are characterized by an exponentially fast convergence to equilibrium, uniformly in the initial condition) were derived by Kontoyiannis in [92]. Moreover, these bounds are essentially identical to the Hoeffding inequality in the special case of i.i.d. RVs (see [92, Remark 1]).

2.5 Applications of the martingale approach in information theory

2.5.1 Minimum distance of binary linear block codes

Consider the ensemble of binary linear block codes of length n and rate R, where the codes are chosen uniformly at random (see [93]). The asymptotic average value of the normalized minimum distance is

equal to

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

where h_2^{-1} denotes the inverse of the binary entropy function to the base 2.

Let $\mathbf{H}$ denote 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 $\mathbf{H}$ 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 sequence $X_0, \ldots, X_n$, where X_i (for $i = 0, 1, \ldots, n$) is an integer-valued RV equal to the minimal number of linearly dependent columns of a parity-check matrix chosen uniformly at random from the ensemble, given that we already revealed its first i columns. Recalling Fact 2 from Section 2.1, we see that this is a martingale sequence, where the associated filtration $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}_n$ is such that $\mathcal{F}_i$ (for $i = 0, 1, \ldots, n$) is the σ-algebra generated by all subsets of $n(1-R)\times n$ binary parity-check matrices whose first i columns have been fixed. This martingale sequence satisfies $|X_i - X_{i-1}| \leq 1$ for $i = 1, \ldots, n$ (indeed, if we observe a new column of $\mathbf{H}$, 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 n columns of $\mathbf{H}$, the code is known exactly). Hence, by Azuma's inequality,

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

This leads to the following theorem:

Theorem 2.5.1. 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 $\mathcal{C}$ lies in the interval

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

so it concentrates around its expected value.

Note, however, that some well-known capacity-approaching families of binary linear block codes have minimum Hamming distance that grows sublinearly with the block length n. For example, the class of parallel concatenated convolutional (turbo) codes was proved to have minimum distance that grows at most as the logarithm of the interleaver length [94].

2.5.2 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 [95]. A sparse parity-check matrix allows one 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., [13]). As a result of their remarkable performance under practical decoding algorithms, these coding techniques have revolutionized the field of channel coding, and 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 dis-

tributions of the ensemble from the edge perspective. It is well-known that linear block codes that can be represented by cycle-free bipartite (Tanner) graphs have poor performance even under ML decoding [96]. The bipartite graphs of capacity-approaching LDPC codes should therefore have cycles. Thus, we need to examine the cardinality of the *fundamental system of cycles* of a bipartite graph. For the required preliminary material, the reader is referred to Sections II-A and II-E of [97]. In [97], we address the following 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?

An information-theoretic lower bound on the average cardinality of the fundamental system of cycles was derived in [97, 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 the number of fundamental cycles should grow at least like $\log \frac{1}{\varepsilon}$, where ε denotes the gap in rate to capacity. This lower bound diverges as the gap to capacity tends to zero, which is consistent with the findings in in [96] on cycle-free codes, and expresses quantitatively the necessity of cycles in bipartite graphs that represent good LDPC code ensembles. As a continuation of this work, we will now give a large-deviations analysis of the cardinality of the fundamental system of cycles for LDPC code ensembles.

Let the triplet (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 $E(\mathcal{G})$ and $V(\mathcal{G})$ are the edge and the vertex sets of $\mathcal{G}$, and $c(\mathcal{G})$ denotes the number of connected components of $\mathcal{G}$, and $|A|$ denotes the cardinality of a set A. Let $R_{\mathrm{d}} \in [0,1)$ denote the *design rate* of the ensemble. Then, in any bipartite graph $\mathcal{G}$ drawn from the ensemble, there are n variable nodes and $m = n(1-R_{\mathrm{d}})$ parity-check nodes, for a total of $|V(\mathcal{G})| = n(2-R_{\mathrm{d}})$ nodes. If we let a_{R} designate the average right degree (i.e., the average degree of the parity-check nodes), then the number of edges in $\mathcal{G}$ is given by $|E(\mathcal{G})| = ma_{\mathrm{R}}$. Therefore, for a code from the (n,λ,ρ) LDPC code ensemble, the cardinality of the fundamental system of cycles satisfies the equality

$$\beta(\mathcal{G}) = n\big[(1-R_{\mathrm{d}})a_{\mathrm{R}} - (2-R_{\mathrm{d}})\big] + c(\mathcal{G}) \tag{2.5.1}$$

where the design rate and the average right degree can be computed from the degree distributions λ and ρ as

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

Let

$$E \triangleq |E(\mathcal{G})| = n(1-R_{\mathrm{d}})a_{\mathrm{R}} \tag{2.5.2}$$

denote the number of edges of an arbitrary bipartite graph $\mathcal{G}$ from the ensemble (for a fixed ensemble, we will use the terms “code" and “bipartite graph" interchangeably). Let us arbitrarily assign numbers $1,\ldots,E$ to the E edges of $\mathcal{G}$. Based on Fact 2, let us construct a martingale sequence $X_0,\ldots,X_E$, where X_i (for $i = 0,1,\ldots,E$) is a RV that denotes the conditional expected number of components of a bipartite graph $\mathcal{G}$ chosen uniformly at random from the ensemble, given that the first i edges of the graph $\mathcal{G}$ have been 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 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 revealing a new edge of $\mathcal{G}$ can change the number of components in the graph by at most 1).

By Corollary 2.3.2, it follows that for every $\alpha \geq 0$

$$\begin{aligned} &\mathbb{P}\left(|c(\mathcal{G}) - \mathbb{E}_{\mathrm{LDPC}(n,\lambda,\rho)}[c(\mathcal{G})]| \geq \alpha E\right) \leq 2e^{-f(\alpha)E} \\ \Rightarrow\ &\mathbb{P}\left(|\beta(\mathcal{G}) - \mathbb{E}_{\mathrm{LDPC}(n,\lambda,\rho)}[\beta(\mathcal{G})]| \geq \alpha E\right) \leq 2e^{-f(\alpha)E} \end{aligned} \tag{2.5.3}$$

where the implication is a consequence of (2.5.1), and the function f was defined in (2.3.14). Hence, for $\alpha > 1$, this probability is zero (since $f(\alpha) = +\infty$ for $\alpha > 1$). Note that, from (2.5.1), $\mathbb{E}_{\mathrm{LDPC}(n,\lambda,\rho)}[\beta(\mathcal{G})]$ scales linearly with n. The combination of Eqs. (2.3.14), (2.5.2), (2.5.3) gives the following statement:

Theorem 2.5.2. Let $\mathrm{LDPC}(n, \lambda, \rho)$ be the LDPC code ensemble with block length n and a pair (λ, ρ) of left and right degree distributions (from the edge perspective). 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:

$$\mathbb{P}\left(|\beta(\mathcal{G}) - \mathbb{E}_{\mathrm{LDPC}(n,\lambda,\rho)}[\beta(\mathcal{G})]| \geq \alpha n\right) \leq 2 \cdot 2^{-\left[1-h_2\left(\frac{1-\eta}{2}\right)\right]n} \tag{2.5.4}$$

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} are, respectively, the design rate and average right degree of the ensemble. Consequently, if $\eta > 1$, this probability is zero.

Remark 2.6. We can obtain the following weakened version of (2.5.4) from Azuma's inequality: for every $\alpha \geq 0$,

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

where η is defined in Theorem 2.5.2. Note, however, that the exponential decay of the two bounds is similar for values of α close to zero (see the exponents in Azuma's inequality and Corollary 2.3.2 in Figure 2.1).

Remark 2.7. 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., $R_{\mathrm{d}} = (1-\varepsilon)C$) [35]. Therefore, for small values of α, the exponential decay rate in the inequality of Theorem 2.5.2

scales like $\left(\log \frac{1}{\varepsilon}\right)^{-2}$. This large-deviations result complements the result in [97, 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 2.8. Consider small deviations from the expected value that scale like $\sqrt{n}$. Note that Corollary 2.3.2 is a special case of Theorem 2.3.1 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 2.1 that, in this case, Corollary 2.3.2 does not provide any improvement in the exponent of the concentration inequality (as compared to Azuma's inequality) when small deviations are considered.

2.5.3 Concentration theorems for LDPC code ensembles over ISI channels

Concentration analysis of the number of erroneous variable-to-check messages for random ensembles of LDPC codes was introduced in [36] and [98] for memoryless channels. It was shown that the performance of an individual code from the ensemble concentrates around the expected (average) value over this ensemble when the length of the block length of the code grows, and that this average behavior converges to the behavior of the cycle-free case. These results were later generalized in [99] for the case of intersymbol-interference (ISI) channels. The proofs of [99, 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 [36] and [98]. The presentation in this subsection is based on [100].

The ISI channel and its message-passing decoding

We start by briefly describing the ISI channel and the graph used for its message-passing decoding. For a detailed description, the reader is referred to [99]. 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 a sequence of $\{-1, +1\}$-valued binary inputs, $\{h_i\}_{i=0}^{I}$ is 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 into a 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 [99] and [101]; 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

In this section, we will prove that, for a large n, a neighborhood of depth ℓ of a variable-to-check node message is tree-like with high probability. Using this result in conjunction with the Azuma–Hoeffding inequality, we will then show 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 $p^{(\ell)}(\underline{s})$ that corresponds to the cycle-free case.

In the following theorems, we consider an ISI channel and windowed message-passing decoding algorithm, where the code graph is chosen uniformly at random from the ensemble of graphs with variable and check node degrees d_{v} and d_{c}, respectively. Let $\mathcal{N}_{\vec{e}}^{(\ell)}$ denote the neigh-

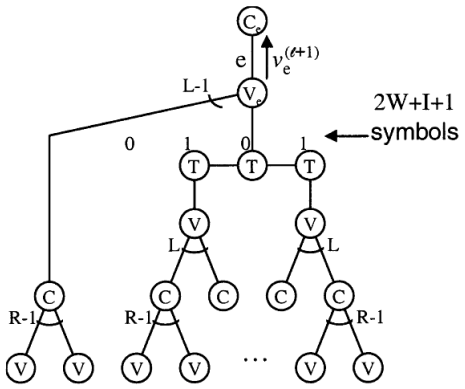

Figure 2.2: Message flow neighborhood of depth 1. In this figure $(I, W, d_{\mathrm{v}} = L, d_{\mathrm{c}} = R) = (1, 1, 2, 3)$

borhood of depth ℓ of an edge $\vec{e} = (\mathrm{v}, \mathrm{c})$ between a variable-to-check node. Let $N_{\mathrm{c}}^{(\ell)}$, $N_{\mathrm{v}}^{(\ell)}$ and $N_e^{(\ell)}$ 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 2.5.3. Let $P_{\overline{t}}^{(\ell)} \equiv \Pr\left\{\mathcal{N}_{\vec{e}}^{(\ell)} \text{ not a tree}\right\}$ denote the probability that the sub-graph $\mathcal{N}_{\vec{e}}^{(\ell)}$ is not a tree (i.e., it contains cycles). Then, there exists some positive constant $\gamma \triangleq \gamma(d_{\mathrm{v}}, d_{\mathrm{c}}, \ell)$ that does not depend on the block-length n, such that $P_{\overline{t}}^{(\ell)} \leq \frac{\gamma}{n}$. More explicitly, one can choose $\gamma(d_{\mathrm{v}}, d_{\mathrm{c}}, \ell) \triangleq \left(N_{\mathrm{v}}^{(\ell)}\right)^2 + \left(\frac{d_{\mathrm{c}}}{d_{\mathrm{v}}} \cdot N_{\mathrm{c}}^{(\ell)}\right)^2$.

Proof. This proof is a straightforward generalization of the proof in [36] (for binary-input output-symmetric memoryless channels) to binary-input ISI channels. A detailed proof is available in [100]. □

The following concentration inequalities follow from Theorem 2.5.3 and the Azuma–Hoeffding inequality:

Theorem 2.5.4. Let $\underline{s}$ be the transmitted codeword, and let $Z^{(\ell)}(\underline{s})$ be the number of erroneous variable-to-check messages after ℓ rounds of the windowed message-passing decoding algorithm. 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 the following statements hold:

Concentration around the expected value. For any $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{Z^{(\ell)}(\underline{s})}{nd_v} - \frac{\mathbb{E}[Z^{(\ell)}(\underline{s})]}{nd_v}\right| > \varepsilon/2\right) \leq 2e^{-\beta\varepsilon^2 n}. \tag{2.5.5}$$

Convergence of the expected value to the cycle-free case. For any $\varepsilon > 0$ and $n > \frac{2\gamma}{\varepsilon}$, we have a.s.

$$\left|\frac{\mathbb{E}[Z^{(\ell)}(\underline{s})]}{nd_v} - p^{(\ell)}(\underline{s})\right| \leq \varepsilon/2. \tag{2.5.6}$$

Concentration around the cycle-free case. For any $\varepsilon > 0$ and $n > \frac{2\gamma}{\varepsilon}$,

$$\mathbb{P}\left(\left|\frac{Z^{(\ell)}(\underline{s})}{nd_v} - p^{(\ell)}(\underline{s})\right| > \varepsilon\right) \leq 2e^{-\beta\varepsilon^2 n}. \tag{2.5.7}$$

More explicitly, the above statements hold for

$$\beta \triangleq \beta(d_{\mathrm{v}}, d_{\mathrm{c}}, \ell) = \frac{d_{\mathrm{v}}^2}{8\left(4d_{\mathrm{v}}(N_e^{(\ell)})^2 + (N_Y^{(\ell)})^2\right)},$$

and

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

Proof. See Appendix 2.C. □

The concentration results in Theorem 2.5.4 extend the results in [36] from binary-input output-symmetric memoryless channels to ISI channels. One can particularize the above expression for β to memoryless

channels by setting $W = 0$ and $I = 0$. Since the proof of Theorem 2.5.4 uses exact expressions for $N_e^{(\ell)}$ and $N_Y^{(\ell)}$, one would expect a tighter bound as compared to $\frac{1}{\beta_{\text{old}}} = 544 d_{\text{v}}^{2\ell-1} d_{\text{c}}^{2\ell}$ given in [36]. As an 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 improvement, 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, because the neighborhood grows with ℓ, so each message is a function of many edges and received output symbols from the channel.

The same concentration phenomena that are established above for regular LDPC code ensembles can be extended to irregular LDPC code ensembles as well. In the special case of memoryless binary-input output-symmetric (MBIOS) channels, the following theorem was proved by Richardson and Urbanke in [13, 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 2.5.5. Let $\mathcal{C}$, a code chosen uniformly at random from the ensemble LDPC(n, λ, ρ), be used for transmission over an 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 $\alpha = \alpha(\lambda, \rho, \delta, l) > 0$ where $\alpha = \alpha(\lambda, \rho, \delta, l)$ is *independent of the block length* n, such that

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

This theorem asserts that the performance of all codes, except for a fraction which is exponentially small in the block length n, is, with high probability, arbitrarily close to the ensemble average. 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.

2.5.4 On the concentration of the conditional entropy for LDPC code ensembles

A large-deviation analysis of the conditional entropy for random ensembles of LDPC codes was introduced by Méasson, Montanari and Urbanke in [102, Theorem 4] and [30, Theorem 1]. The following theorem is proved in [102, Appendix I] based on the Azuma–Hoeffding inequality (although here we rephrase it to consider small deviations of order $\sqrt{n}$ instead of large deviations of order n):

Theorem 2.5.6. Let $\mathcal{C}$ be chosen uniformly at random from the ensemble LDPC(n, λ, ρ). Assume that the transmission of the code $\mathcal{C}$ takes place over an MBIOS channel. Let $H(\mathbf{X}|\mathbf{Y})$ denote 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.

In this section, we revisit the proof of Theorem 2.5.6, originally given in [102, Appendix I], in order to derive a tightened version of this bound. To that end, let $\mathcal{G}$ be a bipartite graph that represents a code chosen uniformly at random from the ensemble LDPC(n, λ, ρ). Define the RV

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

i.e., the conditional entropy when the transmission takes place over an MBIOS channel with transition probabilities $P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{n} p_{Y|X}(y_i|x_i)$, where (by output symmetry) $p_{Y|X}(y|1) = p_{Y|X}(-y|0)$. Fix an arbitrary order for the $m = n(1 - R_{\mathrm{d}})$ parity-check nodes, where R_{d} is 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 generated by all the subsets 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 Fact 2 in Section 2.1, 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 with respect to 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. Similarly to [102, Appendix I], we obtain upper bounds on the differences $|Z_{t+1} - Z_t|$ and then rely on Azuma's inequality in Theorem 2.2.2.

Without loss of generality, we can order the parity-check nodes by increasing degree, as done in [102, Appendix I]. 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 $m_1 = n(1 - R_{\mathrm{d}})\Gamma_{r_1}$ parity-check nodes are of degree r_1, the successive $m_2 = n(1 - R_{\mathrm{d}})\Gamma_{r_2}$ parity-check nodes are of degree r_2, and so on. The $(t+1)$th parity-check will therefore have a well-defined degree, which we denote by r. From the proof in [102, Appendix I],

$$|Z_{t+1} - Z_t| \leq (r+1)\, H_{\mathcal{G}}(\tilde{X}|\mathbf{Y}) \tag{2.5.8}$$

where $H_{\mathcal{G}}(\tilde{X}|\mathbf{Y})$ is a RV equal to 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 [102, Appendix I] was then completed by upper-bounding the parity-check degree r by the maximal parity-check degree $d_{\mathrm{c}}^{\max}$, and also by upper-bounding the conditional entropy of the parity-bit $\tilde{X}$ by 1. This gives

$$|Z_{t+1} - Z_t| \leq d_{\mathrm{c}}^{\max} + 1 \quad t = 0, 1, \ldots, m-1 \tag{2.5.9}$$

which, together with Azuma's inequality, completes the proof of Theorem 2.5.6. Note that the d_i's in Theorem 2.2.2 are equal to $d_{\mathrm{c}}^{\max} + 1$, and n in Theorem 2.2.2 is replaced with the length $m = n(1 - R_{\mathrm{d}})$

of the martingale sequence $\{Z_t\}$ (that is equal to the number of the parity-check nodes in the graph).

We will now present a refined proof that departs from the one in [102, 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 (2.5.8), 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 [102, Appendix I], we derive a simple upper bound that depends on the parity-check degree r and the channel capacity C (see Proposition 2.2).

- 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 (2.5.8) with $d_{\mathrm{c}}^{\max}+1$, we propose to leave it as is, since Azuma's inequality applies to the case when the bounded differences of the martingale sequence are not fixed (see Theorem 2.2.2), and since the number of the parity-check nodes of degree r is equal to $n(1-R_{\mathrm{d}})\Gamma_r$. The effect of this simple modification will be shown in Example 2.7.

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

Proposition 2.2. Let $\mathcal{G}$ be a bipartite graph which corresponds to a binary linear block code used for transmission 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

$$H_{\mathcal{G}}(\tilde{X}|\mathbf{Y}) \leq h_2\left(\frac{1-C^{\frac{r}{2}}}{2}\right). \tag{2.5.10}$$

Furthermore, for a binary symmetric channel (BSC) or a binary erasure

channel (BEC), this bound can be improved to

$$H_{\mathcal{G}}(\tilde{X}|\mathbf{Y}) \leq h_2\left(\frac{1-\left[1-2h_2^{-1}(1-C)\right]^r}{2}\right) \tag{2.5.11}$$

and

$$H_{\mathcal{G}}(\tilde{X}|\mathbf{Y}) \leq 1 - C^r \tag{2.5.12}$$

respectively, where h_2^{-1} in (2.5.11) denotes the inverse of the binary entropy function to base 2.

Note that if the MBIOS channel is perfect (i.e., its capacity is $C = 1$ bit per channel use), then (2.5.10) holds with equality (where both sides of (2.5.10) are zero), whereas the trivial upper bound is 1.

Proof. Since conditioning reduces 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 channel outputs corresponding 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 [97, Eq. (17) and Appendix I], we can show that

$$H(\tilde{X}|\mathbf{Y}) \leq 1 - \frac{1}{2\ln 2}\sum_{k=1}^{\infty}\frac{(g_k)^r}{k(2k-1)} \tag{2.5.13}$$

where (see [97, Eq. (19)])

$$g_k \triangleq \int_0^{\infty} a(l)(1+e^{-l})\tanh^{2k}\left(\frac{l}{2}\right)\mathrm{d}l, \quad \forall\, k \in \mathbb{N} \tag{2.5.14}$$

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

$$g_k \geq C^k, \quad \forall\, k \in \mathbb{N}.$$

Substituting this inequality in (2.5.13) gives

$$\begin{aligned} H(\tilde{X}|\mathbf{Y}) &\leq 1 - \frac{1}{2\ln 2}\sum_{k=1}^{\infty}\frac{C^{kr}}{k(2k-1)} \\ &= h_2\left(\frac{1-C^{\frac{r}{2}}}{2}\right) \end{aligned} \tag{2.5.15}$$

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

$$h_2(x) = 1 - \frac{1}{2\ln 2} \sum_{k=1}^{\infty} \frac{(1-2x)^{2k}}{k(2k-1)}, \quad 0 \leq x \leq 1. \tag{2.5.16}$$

This proves the result in (2.5.10).

The tightened bound on the conditional entropy for the BSC is obtained from (2.5.13) and the equality

$$g_k = \left(1 - 2h_2^{-1}(1-C)\right)^{2k}, \quad \forall\, k \in \mathbb{N}$$

that holds for the BSC (see [97, Eq. (97)]). This replaces C on the right-hand side of (2.5.15) with $\left(1-2h_2^{-1}(1-C)\right)^2$, thus leading to the tightened bound in (2.5.11).

The tightened result for the BEC follows from (2.5.13) where, from (2.5.14),

$$g_k = C, \quad \forall\, k \in \mathbb{N}$$

(see [97, Appendix II]). Substituting g_k into the right-hand side of (2.5.13) gives (2.5.11) (note that $\sum_{k=1}^{\infty} \frac{1}{k(2k-1)} = 2\ln 2$). This completes the proof of Proposition 2.2. □

From Proposition 2.2 and (2.5.8), we get

$$|Z_{t+1} - Z_t| \leq (r+1)\, h_2\left(\frac{1 - C^{\frac{r}{2}}}{2}\right), \tag{2.5.17}$$

where the two improvements for the BSC and BEC are obtained by replacing the second term, $h_2(\cdot)$, on the right-hand side of (2.5.17) by (2.5.11) and (2.5.12), respectively. This improves upon the earlier bound of $(d_{\mathrm{c}}^{\max} + 1)$ in [102, Appendix I]. From (2.5.17) and Theorem 2.2.2, we obtain the following tightened version of the concentration inequality in Theorem 2.5.6:

Theorem 2.5.7. Let $\mathcal{C}$ be chosen uniformly at random from the ensemble LDPC(n, λ, ρ). 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 $\xi > 0$,

$$\mathbb{P}\big(\big|H(\mathbf{X}|\mathbf{Y}) - \mathbb{E}_{\mathrm{LDPC}(n,\lambda,\rho)}[H(\mathbf{X}|\mathbf{Y})]\big| \geq \xi\sqrt{n}\,\big) \leq 2\exp(-B\xi^2), \tag{2.5.18}$$

where

$$B \triangleq \frac{1}{2(1-R_{\mathrm{d}})\displaystyle\sum_{i=1}^{d_{\mathrm{c}}^{\max}} \left\{(i+1)^2\,\Gamma_i\,\left[h_2\left(\frac{1-C^{\frac{i}{2}}}{2}\right)\right]^2\right\}}, \tag{2.5.19}$$

$d_{\mathrm{c}}^{\max}$ 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 (2.5.18) can be improved (i.e., increased), respectively, to

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

and

$$B_{\mathrm{BEC}} \triangleq \frac{1}{2(1-R_{\mathrm{d}})\displaystyle\sum_{i=1}^{d_{\mathrm{c}}^{\max}} \Big\{(i+1)^2\,\Gamma_i\,(1-C^i)^2\Big\}}. \tag{2.5.20}$$

Remark 2.9. By inspection of (2.5.19), we see that Theorem 2.5.7 indeed yields a stronger concentration inequality than the one in Theorem 2.5.6.

Remark 2.10. In the limit where $C \to 1$ bit per channel use, it follows from (2.5.19) that, if $d_{\mathrm{c}}^{\max} < \infty$, then $B \to \infty$. This is in contrast to the value of B in Theorem 2.5.6, which does not depend on the channel capacity and is finite. Note that B should indeed be infinity for a perfect channel, and therefore Theorem 2.5.7 is tight in this case. Moreover, in the case where $d_{\mathrm{c}}^{\max}$ is not finite, we prove the following:

Lemma 2.5.8. If $d_{\mathrm{c}}^{\max} = \infty$ 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 2.5.6, which vanishes when $d_{\mathrm{c}}^{\max} = \infty$, and therefore Theorem 2.5.6 is not informative in this case (see Example 2.7).

Example 2.6 (Comparison of Theorems 2.5.6 and 2.5.7 for right-regular LDPC code ensembles)**.** Let us examine the improvement resulting from the tighter bounds in Theorem 2.5.7 for right-regular LDPC code ensembles. Consider the case where the communication 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

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

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 $\sigma_{\mathrm{BP}} = 0.4156590$. From [13, Example 4.38] that 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 over both channels, the capacity of the BEC and BIAWGNC is set to 0.98 bits per channel use.

Since the considered code ensemble is right-regular (i.e., the parity-check degree is fixed to $d_{\mathrm{c}} = 20$), the value of B in Theorem 2.5.7 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 2.5.7 is satisfied with a block length that is 5.134 times shorter than the block length which

corresponds to Theorem 2.5.6. 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 (2.5.20) as compared to Theorem 2.5.6.

Example 2.7 (Comparison of Theorems 2.5.6 and 2.5.7 for a heavy-tail Poisson distribution (Tornado codes))**.** In this example, we compare Theorems 2.5.6 and 2.5.7 for Tornado codes. This capacity-achieving sequence for the BEC refers to the heavy-tail Poisson distribution, and it was introduced in [35, Section IV], [103] (see also [13, Problem 3.20]). We rely in the following on the analysis in [97, 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 denote 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 2.5.6, which renders this theorem useless for the considered code ensemble. On the other hand, from Theorem 2.5.7,

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

where:

- inequality (a) holds since the binary entropy function to 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 [97, Appendix VI] (an upper bound on λ_2 is derived in [97, Eq. (120)], and the average parity-check node degree scales like $\log \frac{1}{\varepsilon}$).

Hence, from the above chain of inequalities and (2.5.19), it follows that, for a small gap to capacity, the parameter B in Theorem 2.5.7 scales (at least) like

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

Theorem 2.5.7 is therefore useful for the analysis of this LDPC code ensemble. As is shown above, the parameter B in (2.5.19) tends to zero rather slowly as we let the fractional gap ε tend to zero (which therefore demonstrates a rather fast concentration in Theorem 2.5.7).

Example 2.8. Here, we continue with the setting of Example 2.6 on the $(n, d_{\mathrm{v}}, d_{\mathrm{c}})$ regular LDPC code ensemble, where $d_{\mathrm{v}} = 2$ and $d_{\mathrm{c}} = 20$. With the setting of this example, Theorem 2.5.6 gives

$$\begin{aligned}\mathbb{P}\big(\big|H(\mathbf{X}|\mathbf{Y}) - \mathbb{E}_{\mathrm{LDPC}(n,\lambda,\rho)}[H(\mathbf{X}|\mathbf{Y})]\big| \geq \xi\sqrt{n}\,\big) \\ \leq 2\exp(-0.0113\,\xi^2), \quad \forall\,\xi > 0. \end{aligned} \tag{2.5.21}$$

As was mentioned already in Example 2.6, the exponential inequalities in Theorem 2.5.7 achieve an improvement in the exponent of Theorem 2.5.6 by factors 5.134 and 9.051 for the BIAWGNC and BEC, respectively. One therefore obtains from the concentration inequalities in Theorem 2.5.7 that, for every $\xi > 0$,

$$\begin{aligned}\mathbb{P}\big(\big|H(\mathbf{X}|\mathbf{Y}) - \mathbb{E}_{\mathrm{LDPC}(n,\lambda,\rho)}[H(\mathbf{X}|\mathbf{Y})]\big| \geq \xi\sqrt{n}\,\big) \\ \leq \begin{cases} 2\exp(-0.0580\,\xi^2), & \text{(BIAWGNC)} \\ 2\exp(-0.1023\,\xi^2), & \text{(BEC)} \end{cases}. \end{aligned} \tag{2.5.22}$$

2.5.5 Expansion in random regular bipartite graphs

Azuma's inequality is useful for analyzing the expansion properties of random bipartite graphs. The following theorem was originally proved by Sipser and Spielman [37, Theorem 25], but we state it here in a more precise form that captures the relation between the deviation from the expected value and the exponential convergence rate of the resulting probability:

Theorem 2.5.9. Let $\mathcal{G}$ be chosen uniformly at random from the regular ensemble LDPC(n, x^{l-1}, x^{r-1}). Let $\alpha \in (0,1)$ and $\delta > 0$ be fixed. Then, with probability at least $1 - \exp(-\delta n)$, all sets of αn variables in $\mathcal{G}$ have a number of neighbors that is at least

$$n\left[\frac{l(1-(1-\alpha)^r)}{r} - \sqrt{2l\alpha\,(h(\alpha)+\delta)}\right] \tag{2.5.23}$$

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 expected number of neighbors of αn variable nodes is equal to

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

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

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

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

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 $|X_i - X_{i-1}| \le 1$ 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 2.2.2,

$$\mathbb{P}\Big(\mathbb{E}[X(\mathcal{G})] - X(\mathcal{G}) \ge \lambda\sqrt{l\alpha n}\Big) \le \exp\left(-\frac{\lambda^2}{2}\right), \quad \forall\,\lambda > 0.$$

Since there are $\binom{n}{n\alpha}$ choices for the set $\mathcal{V}$, 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 at most $\binom{n}{n\alpha}\exp(-\frac{\lambda^2}{2})$, by the union bound. Since $\binom{n}{n\alpha} \le e^{nh(\alpha)}$, we get the loosened bound that is equal to $\exp\big(nh(\alpha) - \frac{\lambda^2}{2}\big)$. Finally, choosing $\lambda = \sqrt{2n\big(h(\alpha)+\delta\big)}$ gives the bound in (2.5.23). □

2.5.6 Concentration of the crest factor for OFDM signals

Orthogonal-frequency-division-multiplexing (OFDM) is a modulation scheme that converts a high-rate data stream into a number of 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 treatment of OFDM, see, e.g., [104, 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 [105]. 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, the symbol error rate the power efficiency of OFDM signals are reduced 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

$$s(t) = \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} X_i \exp\Big(\frac{j\,2\pi i t}{T}\Big), \quad 0 \le t \le T. \tag{2.5.24}$$

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

$$\frac{1}{T} \int_0^T |s(t)|^2 dt = 1. \tag{2.5.25}$$

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

$$\mathrm{CF}_n(s) \triangleq \max_{0 \le t \le T} |s(t)|. \tag{2.5.26}$$

Commonly, the impact of nonlinearities is described by the distribution of the CF of the transmitted signal [106], but its calculation involves time-consuming simulations even for a small number of sub-carriers. From [107, Section 4] and [108], it follows that the CF scales with high probability like $\sqrt{\ln n}$ for large n. In [106, Theorem 3 and Corollary 5], a concentration inequality was derived for the CF of OFDM signals. It

states that, for an arbitrary $c \geq 2.5$,

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

Remark 2.11. The analysis used to derive this rather strong concentration inequality (see [106, Appendix C]) requires some assumptions on the distribution of the X_i's (see the two conditions in [106, Theorem 3] followed by [106, Corollary 5]). These requirements are not needed in the following analysis, and the derivation of concentration inequalities that are introduced in this subsection is much simpler and provides some insight into the problem, although the resulting concentration result is weaker than the one in [106, 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. The following comparison of concentration inequalities for the crest factor of OFDM signals is based on the work in [109].

Concentration of the crest factor via Azuma's inequality

Let us define the RVs

$$Y_i = \mathbb{E}[\,\mathrm{CF}_n(s) \,|\, X_0, \ldots, X_{i-1}], \quad i = 0, \ldots, n. \tag{2.5.27}$$

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 generated by the first i symbols $(X_0, \ldots, X_{i-1})$ in (2.5.24). Hence, $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}_n$ is a filtration. This martingale also has bounded jumps:

$$|Y_i - Y_{i-1}| \leq \frac{2}{\sqrt{n}}, \qquad i \in \{1, \ldots, n\}$$

since revealing the additional ith coordinate X_i affects the CF, as defined in (2.5.26), 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 $\alpha > 0$,

$$\mathbb{P}(|\mathrm{CF}_n(s) - \mathbb{E}[\mathrm{CF}_n(s)]| \geq \alpha) \leq 2 \exp\left(-\frac{\alpha^2}{8}\right), \tag{2.5.28}$$

which demonstrates concentration around the expected value.

Concentration of the crest-factor via Proposition 2.1

We will now use Proposition 2.1 to derive an improved concentration result. For the martingale sequence $\{Y_i\}_{i=0}^n$ in (2.5.27), Appendix 2.E gives that a.s.

$$|Y_i - Y_{i-1}| \leq \frac{2}{\sqrt{n}}, \quad \mathbb{E}\big[(Y_i - Y_{i-1})^2 | \mathcal{F}_{i-1}\big] \leq \frac{2}{n} \tag{2.5.29}$$

for every $i \in \{1, \ldots, n\}$. Note that the conditioning on the σ-algebra $\mathcal{F}_{i-1}$ is equivalent to conditioning on the symbols $X_0, \ldots, X_{i-2}$, and there is no conditioning for $i = 1$. Further, let $Z_i = \sqrt{n} Y_i$ for $0 \leq i \leq n$. Proposition 2.1 therefore implies that, for an arbitrary $\alpha > 0$,

$$\begin{aligned} \mathbb{P}(|\mathrm{CF}_n(s) - \mathbb{E}[\mathrm{CF}_n(s)]| \geq \alpha) &= \mathbb{P}(|Y_n - Y_0| \geq \alpha) \\ &= \mathbb{P}(|Z_n - Z_0| \geq \alpha \sqrt{n}) \\ &\leq 2 \exp\left(-\frac{\alpha^2}{4}\left(1 + O\Big(\frac{1}{\sqrt{n}}\Big)\right)\right) \end{aligned} \tag{2.5.30}$$

(since $\delta = \frac{\alpha}{2}$ and $\gamma = \frac{1}{2}$ in the setting of Proposition 2.1). Note that the exponent in the last inequality is doubled as compared to the bound that was obtained in (2.5.28) via Azuma's inequality, and the term that scales like $O\Big(\frac{1}{\sqrt{n}}\Big)$ on the right-hand side of (2.5.30) is expressed explicitly for finite n (see the proof of Proposition 2.1.

A concentration inequality via Talagrand's method

In his seminal paper [7], Talagrand introduced another 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., [81, Section 2.4.2], [6, Section 4] and [7]). We will now demonstrate the use of this approach by deriving a concentration result for the crest factor around its median and an upper bound on the distance between the median and the expected value. We start by listing the

definitions needed to introduce a special form of Talagrand's inequalities:

Definition 2.2 (Hamming distance)**.** Let $\mathbf{x}, \mathbf{y}$ be two n-length vectors. The Hamming distance between $\mathbf{x}$ and $\mathbf{y}$ is the number of coordinates where $\mathbf{x}$ and $\mathbf{y}$ disagree, i.e.,

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

where $1_{\{\cdot\}}$ stands for the indicator function.

Definition 2.3 (Weighted Hamming distance)**.** Let $a = (a_1, \ldots, a_n) \in \mathbb{R}_+^n$ (i.e., a is a non-negative vector) satisfy the normalization condition $||a||^2 = \sum_{i=1}^n (a_i)^2 = 1$. Then, define

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

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

The following is a special form of Talagrand's inequalities ([1], [6, Chapter 4] and [7]):

Theorem 2.5.10 (Talagrand's inequality)**.** Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a vector of independent random variables, where, for each $k \in \{1, \ldots, n\}$, X_k takes values in a set A_k, and let $A \triangleq \prod_{k=1}^n A_k$. Let $f \colon A \to \mathbb{R}$ satisfy the condition that, for every $\mathbf{x} \in A$, there exists a non-negative, normalized n-length vector $a = a(\mathbf{x})$, such that

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \sigma d_a(\mathbf{x}, \mathbf{y}), \quad \forall\, \mathbf{y} \in A \tag{2.5.31}$$

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

$$\mathbb{P}(|f(X) - m| \geq \alpha) \leq 4 \exp\left(-\frac{\alpha^2}{4\sigma^2}\right) \tag{2.5.32}$$

where m is any *median*[4] 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 (2.5.32) holds if the condition in (2.5.31) is replaced by

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \sigma d_a(\mathbf{x}, \mathbf{y}), \quad \forall\, \mathbf{y} \in A. \tag{2.5.33}$$

[4]Unlike the mean, a median is not necessarily unique.

We are now in a position to use 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 $|X_i| = |Y_i| = 1$. Note that

$$\begin{aligned}
&\max_{0\le t\le T} \big| s(t; X_0, \ldots, X_{n-1})\big| - \max_{0\le t\le T} \big| s(t; Y_0, \ldots, Y_{n-1})\big| \\
&\quad \le \max_{0\le t\le T} \big| s(t; X_0, \ldots, X_{n-1}) - s(t; Y_0, \ldots, Y_{n-1})\big| \\
&\quad = \frac{1}{\sqrt{n}} \max_{0\le t\le T} \left| \sum_{i=0}^{n-1} (X_i - Y_i) \exp\Big(\frac{j\,2\pi i t}{T}\Big)\right| \\
&\quad \le \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} |X_i - Y_i| \\
&\quad \le \frac{2}{\sqrt{n}} \sum_{i=0}^{n-1} I_{\{X_i \ne Y_i\}} \\
&\quad = 2 d_a(X, Y)
\end{aligned}$$

where

$$a \triangleq \left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right) \tag{2.5.34}$$

is a nonnegative unit vector in $\mathbb{R}^n$ (note that a in this case is independent of $\mathbf{x}$). Hence, Talagrand's inequality in Theorem 2.5.10 implies that, for every $\alpha \ge 0$,

$$\mathbb{P}(|\mathrm{CF}_n(s) - m_n| \ge \alpha) \le 4 \exp\left(-\frac{\alpha^2}{16}\right), \tag{2.5.35}$$

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 the CF around its median. As a simple consequence of (2.5.35), we obtain the following result:

Corollary 2.5.11. 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 (2.5.35),

$$\begin{aligned}\big|\mathbb{E}[\mathrm{CF}_n(s)] - m_n\big| &\le \mathbb{E}\,|\mathrm{CF}_n(s) - m_n| \\ &= \int_0^\infty \mathbb{P}(|\mathrm{CF}_n(s) - m_n| \ge \alpha)\, d\alpha \\ &\le \int_0^\infty 4\exp\Big(-\frac{\alpha^2}{16}\Big)\, d\alpha \\ &= 8\sqrt{\pi}.\end{aligned}$$

□

Remark 2.12. This result applies in general to an arbitrary function f satisfying the condition in (2.5.31), where Talagrand's inequality in (2.5.32) implies that (see, e.g., [6, Lemma 4.6])

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

Establishing concentration via McDiarmid's inequality

Finally, we use McDiarmid's inequality (Theorem 2.2.3) in order 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

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

where the last inequality holds since $|X_{i-1}| = |X'_{i-1}| = 1$. Hence, McDiarmid's inequality in Theorem 2.2.3 implies that, for every $\alpha \geq 0$,

$$\mathbb{P}(|\mathrm{CF}_n(s) - \mathbb{E}[\mathrm{CF}_n(s)]| \geq \alpha) \leq 2\exp\Big(-\frac{\alpha^2}{2}\Big) \tag{2.5.36}$$

which demonstrates concentration of the CF around its expected value. By comparing (2.5.35) with (2.5.36), 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 2.1. As we will see later in Chapter 3, there are deep connections between McDiarmid's inequality and the concentration of Lipschitz functions of many independent random variables around their means (or medians), where the Lipschitz property is with respect to a weighted Hamming distance of the type introduced earlier in Definition 2.3.

To conclude, four concentration inequalities for the crest factor (CF) of OFDM signals have been derived above under the assumption that the symbols are independent. The first two concentration inequalities rely on variants of Azuma's inequality, while the last two bounds are based on Talagrand's and McDiarmid's inequalities. Although these concentration results are weaker than some existing results from the literature (see [106] and [108]), they establish concentration in a rather simple way and provide some additional 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 2.1 by a factor of 2. Note, however, that Proposition 2.1 may, in general, be tighter than McDiarmid's inequality (if $\gamma < \frac{1}{4}$ in the setting of Proposition 2.1). 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.

2.5.7 Random coding theorems via martingale inequalities

In the following section, we establish new error exponents and achievable rates of random coding, for channels with and without memory,

under maximum-likelihood (ML) decoding. The analysis relies on 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 section is based on [40], [41] and [42].

Random coding theorems are concerned with 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, possibly subject to some constraints, and the codebook is known to the encoder and decoder.

Nonlinear effects, which degrade the quality of the information transmission, are typically encountered in wireless communication systems and optical fibers. In satellite communication systems, the amplifiers located onboard 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 [110, 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.

Channel Model

Before we proceed to the performance analysis, the channel model is presented in the following (see Figure 2.3). We refer in the following to a discrete-time channel model of nonlinear Volterra channels where the input-output channel model is given by

$$y_i = [D\mathbf{u}]_i + \nu_i \tag{2.5.37}$$

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

$$[D\mathbf{u}]_i = h_0 + \sum_{j=1}^{L} \sum_{i_1=0}^{q} \cdots \sum_{i_j=0}^{q} h_j(i_1, \ldots, i_j) u_{i-i_1} \ldots u_{i-i_j}. \qquad (2.5.38)$$

and $\boldsymbol{\nu}$ is an additive Gaussian noise vector with i.i.d. components $\nu_i \sim \mathcal{N}(0, \sigma_\nu^2)$. The real-valued functions h_j in (2.5.38) form the kernels of the system. Such system models are used in the baseband representation of nonlinear narrowband communication channels. Note that a linear time-invariant (LTI) channel is a special case of (2.5.37) and (2.5.38) with $L = 1$ and $q = 0$, and in this case, the kernel is the impulse response of the channel.

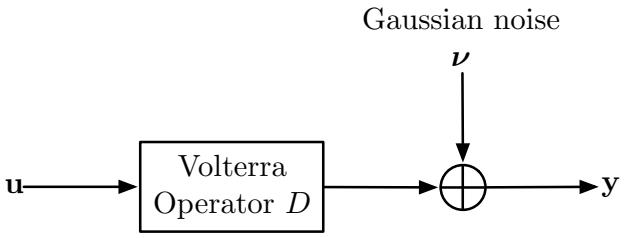

Figure 2.3: The discrete-time Volterra non-linear channel model in Eqs. (2.5.37) and (2.5.38) where the channel input and output are $\{U_i\}$ and $\{Y_i\}$, respectively, and the additive noise samples $\{\nu_i\}$, which are added to the distorted signal, are i.i.d. with zero mean and variance σ_ν^2.

Martingale inequalities

We start by writing down explicitly some martingale inequalities that will be needed later on:

The first martingale inequality that will be used in the following is given in (2.3.6). It was used earlier in this chapter to prove the refinement of the Azuma–Hoeffding inequality in Theorem 2.3.1; here, we state it as a standalone theorem:

Theorem 2.5.12. Let $\{X_k, \mathcal{F}_k\}_{k=0}^n$, for some $n \in \mathbb{N}$, be a 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 conditions

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

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

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

The second martingale inequality that we will need is similar to [111, Theorem 3], except for removing the assumption that the martingale is conditionally symmetric. It leads to the following theorem:

Theorem 2.5.13. 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 $t \geq 0$,

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

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 [41] for the derivation of the upper bound on the ML decoding error probability. 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 with noise variance σ_ν^2 gives

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

where Q is the complementary Gaussian cdf (see (2.2.16)). Using the inequality $Q(x) \leq \frac{1}{2}\exp(-\frac{x^2}{2})$ for $x \geq 0$ and also ignoring the factor of one-half in the bound of Q, we obtain the loosened bound

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

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

$$\overline{P}_{\mathrm{e}} \leq (M-1)\,\mathbb{E}\left[\exp\left(-\frac{\rho\,\|D\mathbf{u}-D\tilde{\mathbf{u}}\|_2^2}{8\sigma_\nu^2}\right)\right], \tag{2.5.39}$$

where the expectation is taken over two randomly chosen codewords $\mathbf{u}$ and $\tilde{\mathbf{u}}$ that are independent and whose 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 σ-algebra $\mathcal{F}_i$ is given by

$$\mathcal{F}_i \triangleq \sigma(U_1, \widetilde{U}_1, \ldots, U_i, \widetilde{U}_i), \quad \forall\, i \in \{1, \ldots, N\} \tag{2.5.40}$$

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

$$X_k = \mathbb{E}\Big[||D\mathbf{u}-D\tilde{\mathbf{u}}||_2^2 \mid \mathcal{F}_k\Big] \tag{2.5.41}$$

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

$$X_0 = \mathbb{E}\Big[||D\mathbf{u}-D\tilde{\mathbf{u}}||_2^2\Big], \quad X_N = ||D\mathbf{u}-D\tilde{\mathbf{u}}||_2^2. \tag{2.5.42}$$

Following earlier notation, let $\xi_k \triangleq X_k - X_{k-1}$ be the jumps of the martingale. Then, we have

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

and the substitution of the last equality into (2.5.39) gives that

$$\overline{P}_{\mathrm{e}} \leq \exp(NR)\,\exp\left(-\frac{\rho\,\mathbb{E}\big[||D\mathbf{u}-D\tilde{\mathbf{u}}||_2^2\big]}{8\sigma_\nu^2}\right)\mathbb{E}\left[\exp\left(-\frac{\rho}{8\sigma^2}\cdot\sum_{k=1}^N \xi_k\right)\right]. \tag{2.5.43}$$

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

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

Due to the channel model (see Eq. (2.5.38)) and the assumption that the symbols $\{u_i\}$ are i.i.d., it follows that $\mathsf{var}\left([D\mathbf{u}]_k\right)$ is fixed for $k=q,\ldots,N$. If we let $D_{\mathrm{v}}(P)$ designate this common value of the variance (i.e., $D_{\mathrm{v}}(P)=\mathsf{var}\left([D\mathbf{u}]_k\right)$ for $k\geq q$), then

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

Furthermore, define the quantity

$$C_\rho(P)\triangleq\exp\left\{-\frac{\rho}{4\sigma_\nu^2}\ \left(\sum_{k=1}^{q-1}\mathsf{var}\left([D\mathbf{u}]_k\right)-(q-1)D_{\mathrm{v}}(P)\right)\right\},$$

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

$$\begin{aligned}\overline{P}_{\mathrm{e}} &\leq C_\rho(P)\ \exp\left\{-N\left(\frac{\rho\, D_{\mathrm{v}}(P)}{4\sigma_\nu^2}-R\right)\right\}\\ &\qquad\times\mathbb{E}\left[\exp\left(\frac{\rho}{8\sigma_\nu^2}\cdot\sum_{k=1}^{N}Z_k\right)\right],\quad\forall\,\rho\in[0,1]\end{aligned}\tag{2.5.44}$$

where $Z_k\triangleq-\xi_k$, so $\{Z_k,\mathcal{F}_k\}$ is a martingale difference sequence that corresponds to the jumps of the martingale $\{-X_k,\mathcal{F}_k\}$. From (2.5.41),

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

$$\begin{aligned} Z_k &= X_{k-1} - X_k \\ &= \mathbb{E}[||D\mathbf{u} - D\widetilde{\mathbf{u}}||_2^2 \,|\, \mathcal{F}_{k-1}] - \mathbb{E}[||D\mathbf{u} - D\widetilde{\mathbf{u}}||_2^2 \,|\, \mathcal{F}_k]. \end{aligned} \tag{2.5.45}$$

We will now improve upon the bounds on achievable rates and error exponents from [41] by appealing to the two martingale inequalities presented earlier. More specifically, we obtain two exponential upper bounds (in terms of N) on the last term on the right-hand side of (2.5.44).

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 (2.5.44), combined with Theorems 2.5.12 and 2.5.13, we obtain the following bounds:

1. *First Bounding Technique*: From Theorem 2.5.12, if the inequalities
$$Z_k \leq d, \quad \mathsf{var}(Z_k \,|\, \mathcal{F}_{k-1}) \leq \sigma^2$$
hold a.s. for every $k \geq 1$, and $\gamma_2 \triangleq \frac{\sigma^2}{d^2}$, then it follows from (2.5.44) that, for every $\rho \in [0,1]$,
$$\begin{aligned} \overline{P}_{\mathrm{e}} \leq C_\rho(P) \, \exp\left\{-N\left(\frac{\rho\, D_{\mathrm{v}}(P)}{4\sigma_\nu^2} - R\right)\right\} \\ \times \left(\frac{\exp\left(-\frac{\rho\gamma_2\, d}{8\sigma_\nu^2}\right) + \gamma_2 \exp\left(\frac{\rho\, d}{8\sigma_\nu^2}\right)}{1+\gamma_2}\right)^N. \end{aligned}$$
Therefore, the maximal achievable rate that follows from this bound is given by
$$\begin{aligned} R_1(\sigma_\nu^2) \triangleq \max_P \max_{\rho\in[0,1]} \Bigg\{ &\frac{\rho\, D_{\mathrm{v}}(P)}{4\sigma_\nu^2} \\ &- \ln\left(\frac{\exp\left(-\frac{\rho\gamma_2\, d}{8\sigma_\nu^2}\right) + \gamma_2 \exp\left(\frac{\rho\, d}{8\sigma_\nu^2}\right)}{1+\gamma_2}\right)\Bigg\} \end{aligned} \tag{2.5.46}$$
where the double maximization is performed over the input distribution P and the parameter $\rho \in [0,1]$. The inner maximization in

(2.5.46) can be expressed in closed form, leading to the following simplified expression:

$$R_1(\sigma_\nu^2) = \max_P \begin{cases} D\left(\left(\frac{\gamma_2}{1+\gamma_2} + \frac{2\mathcal{D}_{\mathrm{v}}(P)}{d(1+\gamma_2)}\right) \middle\| \frac{\gamma_2}{1+\gamma_2}\right), \\ \qquad\qquad \text{if } \; D_{\mathrm{v}}(P) < \frac{\gamma_2\, d\left(\exp\left(\frac{d(1+\gamma_2)}{8\sigma_\nu^2}\right)-1\right)}{2\left(1+\gamma_2 \exp\left(\frac{d(1+\gamma_2)}{8\sigma_\nu^2}\right)\right)} \\ \frac{D_{\mathrm{v}}(P)}{4\sigma_\nu^2} - \ln\left(\dfrac{\exp\left(-\frac{\gamma_2\, d}{8\sigma_\nu^2}\right) + \gamma_2 \exp\left(\frac{d}{8\sigma_\nu^2}\right)}{1+\gamma_2}\right), \\ \qquad\qquad \text{otherwise} \end{cases} \tag{2.5.47}$$

where

$$D(p||q) \triangleq p \ln\left(\frac{p}{q}\right) + (1-p)\ln\left(\frac{1-p}{1-q}\right), \quad \forall\, p, q \in (0,1) \tag{2.5.48}$$

denotes the relative entropy (or information divergence) between Bernoulli(p) and Bernoulli(q) measures.

2. *Second Bounding Technique:* Based on the combination of Theorem 2.5.13 and Eq. (2.5.44), we can derive another bound on the achievable rate for random coding under ML decoding. Referring to the martingale-difference sequence $\{Z_k, \mathcal{F}_k\}_{k=1}^N$ in Eqs. (2.5.40) and (2.5.45), we can see from Eq. (2.5.44) that, if for some even number $m \in \mathbb{N}$, the inequalities

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

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

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

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

$$R_2(\sigma_\nu^2) \triangleq \max_P \max_{\rho\in[0,1]} \left\{ \frac{\rho\, D_{\mathrm{v}}(P)}{4\sigma_\nu^2} - \ln\left(1 + \sum_{l=2}^{m-1} \frac{\gamma_l - \gamma_m}{l!} \left(\frac{\rho d}{8\sigma_\nu^2}\right)^l \right.\right.$$
$$\left.\left. + \gamma_m \left(e^{\frac{\rho\, d}{8\,\sigma_\nu^2}} - 1 - \frac{\rho\, d}{8\,\sigma_\nu^2}\right)\right)\right\} \tag{2.5.49}$$

where, similarly to (2.5.46), the double maximization in (2.5.49) is performed over the input distribution P and the parameter $\rho \in [0,1]$.

Achievable rates for random coding under ML decoding

In the following, we particularize the above results on the achievable rates for random coding to various linear and non-linear channels (with and without memory). 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 the union bound, it would have been worth comparing the achievable rate with the cutoff rate).

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

$$Y_k = U_k + \nu_k,$$

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

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

for some $\alpha \in [0, 1]$. Since the channel is memoryless and the all the symbols are i.i.d., from (2.5.40) and (2.5.45) one gets

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

Hence, for every k,

$$Z_k \le 8\alpha(1-\alpha)A^2 \triangleq d. \tag{2.5.50}$$

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

$$\begin{aligned}
&\mathbb{E}\big[(Z_k)^l \,|\, \mathcal{F}_{k-1}\big] \\
&= \mathbb{E}\big[(Z_k)^l\big] \\
&= \mathbb{E}\Big[\big(8\alpha(1-\alpha)A^2 - (U_k - \widetilde{U}_k)^2\big)^l\Big] \\
&= \big[1 - 2\alpha(1-\alpha)\big]\big(8\alpha(1-\alpha)A^2\big)^l \\
&\qquad + 2\alpha(1-\alpha)\left(8\alpha(1-\alpha)A^2 - 4A^2\right)^l \\
&\triangleq \mu_l
\end{aligned} \tag{2.5.51}$$

and therefore, from (2.5.50) and (2.5.51), for every $l \in \mathbb{N}$

$$\gamma_l \triangleq \frac{\mu_l}{d^l} = \big[1 - 2\alpha(1-\alpha)\big]\left[1 + (-1)^l\left(\frac{1 - 2\alpha(1-\alpha)}{2\alpha(1-\alpha)}\right)^{l-1}\right]. \tag{2.5.52}$$

Let us now apply the two bounds on the achievable rates for random coding in Eqs. (2.5.47) and (2.5.49) 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 (2.5.50) and (2.5.52) that

$$D_{\mathrm{v}}(P) = \mathsf{var}(U_k) = A^2, \quad d = 2A^2, \qquad \gamma_l = \frac{1 + (-1)^l}{2}, \ \ \forall\, l \in \mathbb{N}. \tag{2.5.53}$$

We start with the first bounding technique that leads to Eq. (2.5.47). Since the first condition in this equation cannot hold for the set of parameters in (2.5.53), 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 $\mathrm{SNR} \triangleq \frac{A^2}{\sigma_\nu^2}$ designate the signal-to-noise ratio; then the first achievable rate takes the form

$$R_1'(\mathrm{SNR}) = \frac{\mathrm{SNR}}{4} - \ln\cosh\left(\frac{\mathrm{SNR}}{4}\right). \tag{2.5.54}$$

It should be pointed out that the optimal value of ρ in (2.5.47) is equal to 1 (i.e., $\rho^\star = 1$).

Now let us compare it with the achievable rate that follows from (2.5.49). Let $m \in \mathbb{N}$ be an even number. Since, from (2.5.53), $\gamma_l = 1$ for all even values of $l \in \mathbb{N}$ and $\gamma_l = 0$ for all odd values of $l \in \mathbb{N}$, it follows that

$$\begin{aligned} &1 + \sum_{l=2}^{m-1} \frac{\gamma_l - \gamma_m}{l!} \left(\frac{\rho d}{8\sigma_\nu^2}\right)^l + \gamma_m \left(\exp\Big(\frac{\rho\, d}{8\,\sigma_\nu^2}\Big) - 1 - \frac{\rho\, d}{8\,\sigma_\nu^2}\right) \\ &\quad = 1 - \sum_{l=1}^{\frac{m}{2}-1} \frac{1}{(2l+1)!} \left(\frac{\rho d}{8\sigma_\nu^2}\right)^{2l+1} \\ &\qquad\qquad + \left(\exp\Big(\frac{\rho\, d}{8\,\sigma_\nu^2}\Big) - 1 - \frac{\rho\, d}{8\,\sigma_\nu^2}\right). \end{aligned} \tag{2.5.55}$$

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 $\rho \in [0, 1]$), from (2.5.49)

we see that 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

$$\begin{aligned}
&\lim_{m\to\infty} \left(1 + \sum_{l=2}^{m-1} \frac{\gamma_l - \gamma_m}{l!} \left(\frac{\rho d}{8\sigma_\nu^2}\right)^l \right. \\
&\qquad\qquad \left. + \gamma_m \left(\exp\Big(\frac{\rho\, d}{8\,\sigma_\nu^2}\Big) - 1 - \frac{\rho\, d}{8\,\sigma_\nu^2}\right)\right) \\
&\overset{\text{(a)}}{=} 1 - \sum_{l=1}^{\infty} \frac{1}{(2l+1)!} \left(\frac{\rho d}{8\sigma_\nu^2}\right)^{2l+1} + \left(\exp\Big(\frac{\rho\, d}{8\,\sigma_\nu^2}\Big) - 1 - \frac{\rho\, d}{8\,\sigma_\nu^2}\right) \\
&\overset{\text{(b)}}{=} 1 - \left(\sinh\Big(\frac{\rho\, d}{8\,\sigma_\nu^2}\Big) - \frac{\rho\, d}{8\,\sigma_\nu^2}\right) + \left(\exp\Big(\frac{\rho\, d}{8\,\sigma_\nu^2}\Big) - 1 - \frac{\rho\, d}{8\,\sigma_\nu^2}\right) \\
&\overset{\text{(c)}}{=} \cosh\Big(\frac{\rho\, d}{8\,\sigma_\nu^2}\Big) \qquad\qquad (2.5.56)
\end{aligned}$$

where equality (a) follows from (2.5.55), 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 (2.5.49) gives (from (2.5.53), $\frac{d}{8\sigma_\nu^2} = \frac{A^2}{4\sigma_\nu^2}$)

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

Since the function $f(x) \triangleq x - \ln\cosh(x)$ for $x \in \mathbb{R}$ is monotonically increasing (note that $f'(x) = 1 - \tanh(x) \geq 0$), 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. (2.5.49) is equal to

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

in units of nats per channel use, and it is obtained in the asymptotic case where we let the even number m tend to infinity. Finally, setting $\mathrm{SNR} = \frac{A^2}{\sigma_\nu^2}$, gives the achievable rate in (2.5.54), so the first and second achievable rates for the binary-input AWGN

channel coincide, i.e.,

$$R_1'(\mathrm{SNR}) = R_2'(\mathrm{SNR}) = \frac{\mathrm{SNR}}{4} - \ln\cosh\left(\frac{\mathrm{SNR}}{4}\right). \tag{2.5.57}$$

Note that this common rate tends to zero as we let the SNR tend to zero, and it tends to $\ln 2$ nats per channel use (i.e., 1 bit per channel use) as we let the SNR tend to infinity.

In order to exemplify the tightness of the achievable random-coding rate in (2.5.57), we compare it 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., [13, Example 4.38 on p. 194])

$$\begin{aligned} &C(\mathrm{SNR}) \\ &= \ln 2 + (2\ \mathrm{SNR} - 1)\, Q(\sqrt{\mathrm{SNR}}) - \sqrt{\frac{2\,\mathrm{SNR}}{\pi}}\ \exp\Big(-\frac{\mathrm{SNR}}{2}\Big) \\ &\quad + \sum_{i=1}^{\infty} \left\{ \frac{(-1)^i}{i(i+1)} \cdot \exp(2i(i+1)\ \mathrm{SNR})\ Q\big((1+2i)\sqrt{\mathrm{SNR}}\big) \right\} \end{aligned} \tag{2.5.58}$$

where $Q(\cdot)$ is, again, the complementary Gaussian cdf, see Eq. (2.2.16). Furthermore, this infinite series exhibits fast convergence — the absolute value of its nth remainder is bounded by the $(n+1)$st 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 nth 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 (2.5.57) that follows from the martingale approach is shown in Figure 2.4.

From this discussion, we see that the first and second bounding techniques in Section 2.5.7 lead to the same achievable rate (see (2.5.57)) in the setting of random coding and ML decoding where

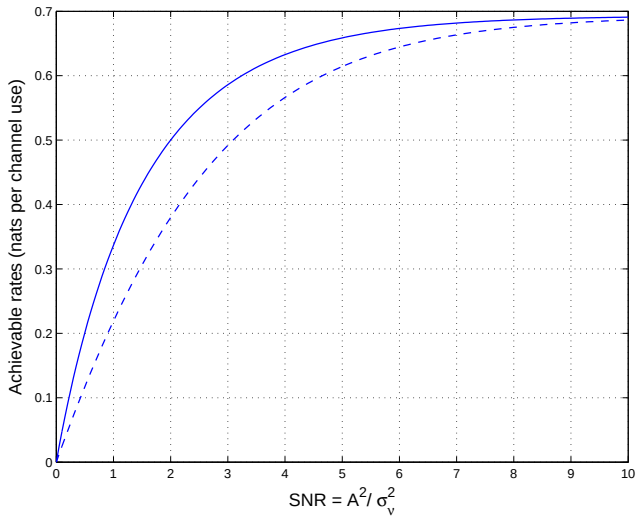

Figure 2.4: A comparison between the symmetric i.i.d. mutual information of the binary-input AWGN channel (solid line) and the common achievable rate in (2.5.57) (dashed line) that follows from the martingale approach in this subsection.

we assume a symmetric input distribution (i.e., $P(\pm A) = \frac{1}{2}$). But this is due to the fact that, from (2.5.53), the sequence $\{\gamma_l\}_{l \geq 2}$ is equal to zero for odd values of l and it is equal to 1 for even values of l (see the derivation of (2.5.55) and (2.5.56)). 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*: Consider the following input-output channel model shown in Figure 2.3: Consider the channel model in (2.5.37) and (2.5.38), and let us assume that the transmission of information takes place over a Volterra system D_1 of order $L = 3$ and memory $q = 2$, whose kernels are depicted in Table 2.1. Due to complexity of the channel model, the calculation of the achievable rates provided earlier in this subsection requires numerical calculation of the parameters d and σ^2, and thus of γ_2, for the martingale $\{Z_i, \mathcal{F}_i\}_{i=0}^{N}$.

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

In order to achieve this goal, we have to calculate $|Z_i - Z_{i-1}|$ and $\mathsf{var}(Z_i|\mathcal{F}_{i-1})$ for all possible combinations of the input samples which contribute to the aforementioned expressions. Thus, the complexity of calculating 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 $R_1(D_1, A, \sigma_\nu^2)$ and $R_2^{(2)}(D_1, A, \sigma_\nu^2)$, which depend on the channel input parameter A, are compared to the achievable rate provided in [41, Fig. 2] and are shown to be larger than the latter.

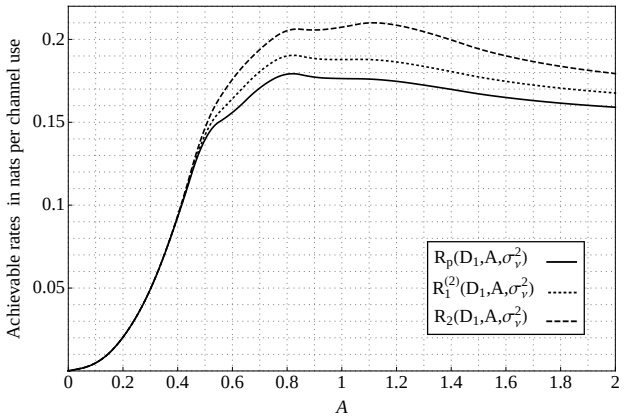

Figure 2.5: Comparison of the achievable rates in this subsection $R_1(D_1, A, \sigma_\nu^2)$ and $R_2^{(2)}(D_1, A, \sigma_\nu^2)$ (where $m = 2$) with the bound $R_p(D_1, A, \sigma_\nu^2)$ of [41, Fig.2] for the nonlinear channel with kernels depicted in Table 2.1 and noise variance $\sigma_\nu^2 = 1$. 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 refinements of Bennett's

inequality (see [112] and [113]), combined with a possible tightening of the union bound under ML decoding (see, e.g., [114]).

2.6 Summary

This chapter deals with several 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 was focused on the derivation of these inequalities, followed by a short discussion of their relations to some selected classical results in probability theory. The second part of this work focused on applications of these martingale inequalities in the context of information theory, communication, and coding theory. 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., [15] and [16]).

2.A Proof of Bennett's inequality

The inequality is trivial if $\lambda = 0$, so we only prove it 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 $\mathsf{var}(Y) \leq \lambda^2\sigma^2 \triangleq \sigma_Y^2$. It is therefore required to show that, if $\mathbb{E}[Y] = 0$, $Y \leq b_Y$, and $\mathsf{var}(Y) \leq \sigma_Y^2$, then

$$\mathbb{E}[e^Y] \leq \left(\frac{b_Y^2}{b_Y^2 + \sigma_Y^2}\right) e^{-\frac{\sigma_Y^2}{b_Y}} + \left(\frac{\sigma_Y^2}{b_Y^2 + \sigma_Y^2}\right) e^{b_Y}. \tag{2.A.1}$$

Let Y_0 be a random variable that takes two possible values $-\frac{\sigma_Y^2}{b_Y}$ and b_Y with probabilities

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

Then inequality (2.A.1) is equivalent to

$$\mathbb{E}[e^Y] \leq \mathbb{E}[e^{Y_0}], \tag{2.A.3}$$

which is what we will prove.

To that end, let ϕ be the unique parabola so that the function

$$f(y) \triangleq \phi(y) - e^y, \quad \forall\, y \in \mathbb{R}$$

is zero at $y = b_Y$, and has $f(y) = f'(y) = 0$ at $y = -\frac{\sigma_Y^2}{b_Y}$. Since ϕ'' is constant, $f''(y) = 0$ at exactly one value of y, say, y_0. Furthermore, since $f(-\frac{\sigma_Y^2}{b_Y}) = f(b_Y)$ (both are equal to zero), we must have $f'(y) = 0$ for some $y_1 \in (-\frac{\sigma_Y^2}{b_Y}, b_Y)$. By the same argument applied to f' on $[-\frac{\sigma_Y^2}{b_Y}, y_1]$, it follows that $y_0 \in (-\frac{\sigma_Y^2}{b_Y}, y_1)$. The function f is convex on $(-\infty, y_0]$ (since, on this interval, $f''(y) = \phi''(y) - e^y \geq \phi''(y) - e^{y_0} = \phi''(y_0) - e^{y_0} = f''(y_0) = 0$), and its minimal value on this interval is attained 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 attains its maximal value on this interval at $y = y_1$. This 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

$$\mathbb{E}[e^Y] \leq \mathbb{E}[\phi(Y)],$$

with equality if $\mathbb{P}(Y \in \{-\frac{\sigma_Y^2}{b_Y}, b_Y\}) = 1$. Since $f''(y) \geq 0$ for $y < y_0$, it must be the case that $\phi''(y) - e^y = f''(y) \geq 0$, 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.A.2), 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.A.3), which therefore completes the proof of Bennett's inequality.

2.B On the moderate deviations principle in Section 2.4.2

Here we show that, in contrast to Azuma's inequality, Theorem 2.3.1 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 (2.4.4) under an extra 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 $\mathcal{F}_0 = \{\emptyset, \mathcal{F}\}$. This martingale sequence has uniformly bounded jumps: $|S_k - S_{k-1}| = |X_k| \leq d$ 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

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

This differs from the limit in (2.4.4) where σ^2 is replaced by d^2, so Azuma's inequality does not provide the asymptotic limit in (2.4.4) (unless $\sigma^2 = d^2$, i.e., $|X_k| = d$ a.s. for every k).

2.C Proof of Theorem 2.5.4

From the triangle inequality, we have

$$\begin{aligned}\mathbb{P}\left(\left|\frac{Z^{(\ell)}(\underline{s})}{nd_{\mathrm{v}}} - p^{(\ell)}(\underline{s})\right| > \varepsilon\right) &\leq \mathbb{P}\left(\left|\frac{Z^{(\ell)}(\underline{s})}{nd_{\mathrm{v}}} - \frac{\mathbb{E}[Z^{(\ell)}(\underline{s})]}{nd_{\mathrm{v}}}\right| > \varepsilon/2\right) \\ &\quad + \mathbb{P}\left(\left|\frac{\mathbb{E}[Z^{(\ell)}(\underline{s})]}{nd_{\mathrm{v}}} - p^{(\ell)}(\underline{s})\right| > \varepsilon/2\right). \end{aligned} \tag{2.C.1}$$

If inequality (2.5.6) holds a.s., then $\mathbb{P}\left(\left|\frac{Z^{(\ell)}(\underline{s})}{nd_{\mathrm{v}}} - p^{(\ell)}(\underline{s})\right| > \varepsilon/2\right) = 0$; therefore, using (2.C.1), we deduce that (2.5.7) follows from (2.5.5) and (2.5.6) for any $\varepsilon > 0$ and $n > \frac{2\gamma}{\varepsilon}$. We start by proving (2.5.5). 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_{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_{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_{v} variable-to-check node edges of the graph, followed by the sequence of the n received symbols at the channel output. For $i = 0, ...n(d_{\mathrm{v}} + 1)$, let the RV $\widetilde{Z}_i \triangleq \mathbb{E}[Z^{(\ell)}(\underline{s})|a_1, ...a_i]$ 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 Fact 2 in Section 2.1), where $\widetilde{Z}_0 = \mathbb{E}[Z^{(\ell)}(\underline{s})]$ and $\widetilde{Z}_{n(d_{\mathrm{v}}+1)} = Z^{(\ell)}(\underline{s})$. Hence, getting an upper bound on the sequence of differences $|\widetilde{Z}_{i+1} - \widetilde{Z}_i|$ will enable us to apply the Azuma–Hoeffding inequality to prove concentration around the expected value $\widetilde{Z}_0$. To this end, let's 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 $\alpha \equiv (d_{\mathrm{v}} - 1 + 2Wd_{\mathrm{v}})(d_{\mathrm{c}} - 1)$, then there are a total of

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

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 includes 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

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

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_{v} exposures, and $|\widetilde{Z}_{i+1} - \widetilde{Z}_i| \leq N_Y^{(\ell)}$ for the last n exposures. Applying the Azuma–Hoeffding inequality, we get

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{Z^{(\ell)}(\underline{s})}{nd_{\mathrm{v}}} - \frac{\mathbb{E}[Z^{(\ell)}(\underline{s})]}{nd_{\mathrm{v}}}\right| > \frac{\varepsilon}{2}\right) \\
&\qquad \leq 2\exp\left(-\frac{(nd_{\mathrm{v}}\varepsilon/2)^2}{2\left(nd_{\mathrm{v}}\big(2N_e^{(\ell)}\big)^2 + n\big(N_Y^{(\ell)}\big)^2\right)}\right)
\end{aligned}$$

and a comparison of this concentration inequality with (2.5.5) gives that

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

Next, proving inequality (2.5.6) relies on concepts from [36] and [99]. Let $\mathbb{E}[Z_i^{(\ell)}(\underline{s})]$ ($i \in \{1, \ldots, nd_{\mathrm{v}}\}$) be the expected number of incorrect messages passed along edge $\overrightarrow{e_i}$ after ℓ rounds, where the average is with respect to all realizations of graphs and all output symbols from the channel. Then, by symmetry in the graph construction and by linearity of expectation, it follows that

$$\mathbb{E}[Z^{(\ell)}(\underline{s})] = \sum_{i \in [nd_{\mathrm{v}}]} \mathbb{E}[Z_i^{(\ell)}(\underline{s})] = nd_{\mathrm{v}}\mathbb{E}[Z_1^{(\ell)}(\underline{s})]. \tag{2.C.3}$$

By the law of total probability,

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

As shown in Theorem 2.5.3, $P_{\overline{t}}^{(\ell)} \leq \frac{\gamma}{n}$, where γ is a positive constant independent of n. Furthermore, we have $\mathbb{E}[Z_1^{(\ell)}(\underline{s}) \,|\, \text{neighborhood is a tree}] = p^{(\ell)}(\underline{s})$, so

$$\begin{aligned} \mathbb{E}[Z_1^{(\ell)}(\underline{s})] &\leq (1 - P_{\overline{t}}^{(\ell)}) p^{(\ell)}(\underline{s}) + P_{\overline{t}}^{(\ell)} \leq p^{(\ell)}(\underline{s}) + P_{\overline{t}}^{(\ell)} \\ \mathbb{E}[Z_1^{(\ell)}(\underline{s})] &\geq (1 - P_{\overline{t}}^{(\ell)}) p^{(\ell)}(\underline{s}) \geq p^{(\ell)}(\underline{s}) - P_{\overline{t}}^{(\ell)}. \end{aligned} \tag{2.C.4}$$

Using (2.C.3), (2.C.4) and $P_{\overline{t}}^{(\ell)} \leq \frac{\gamma}{n}$ gives that

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

Hence, if $n > \frac{2\gamma}{\varepsilon}$, then (2.5.6) holds.

Analysis related to Theorem 2.3.1

The analysis here is a slight modification of the analysis in the proof of Proposition 2.1, with the required adaptation of the calculations for $\eta \in (\frac{1}{2}, 1)$. It follows from Theorem 2.3.1 that, for every $\alpha \geq 0$,

$$\mathbb{P}(|S_n| \geq \alpha n^{\eta}) \leq 2 \exp\left(-n \, D\Big(\frac{\delta_n + \gamma}{1+\gamma} \Big\| \frac{\gamma}{1+\gamma} \Big) \right)$$

where γ is introduced in (2.3.2), and δ_n in (2.3.17) is replaced with

$$\delta_n \triangleq \frac{\frac{\alpha}{n^{1-\eta}}}{d} = \delta n^{-(1-\eta)} \tag{2.C.5}$$

due to the definition of δ in (2.3.2). Following the same analysis as in the proof of Proposition 2.1, 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 (2.3.2), $\frac{\delta^2}{\gamma} = \frac{\alpha^2}{\sigma^2}$)

$$\lim_{n \to \infty} n^{1-2\eta} \, \ln \mathbb{P}(|S_n| \geq \alpha n^{\eta}) \leq -\frac{\alpha^2}{2\sigma^2}. \tag{2.C.6}$$

Hence, this upper bound coincides with the exact asymptotic result in (2.4.4).

2.D Proof of Lemma 2.5.8

In order to prove Lemma 2.5.8, one needs to show that, if $\rho'(1) < \infty$, then

$$\lim_{C \to 1} \sum_{i=1}^{\infty} (i+1)^2 \Gamma_i \left[h_2 \left(\frac{1 - C^{\frac{i}{2}}}{2} \right) \right]^2 = 0 \tag{2.D.1}$$

which then yields from (2.5.19) that $B \to \infty$ in the limit where $C \to 1$.

By the assumption in Lemma 2.5.8, if $\rho'(1) < \infty$, 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

$$\Gamma_i = \frac{\frac{\rho_i}{i}}{\displaystyle\int_0^1 \rho(x)\,\mathrm{d}x} = d_{\mathrm{c}}^{\mathrm{avg}} \left(\frac{\rho_i}{i} \right), \quad \forall\, i \in \mathbb{N}.$$

The infinite series in (2.D.1) therefore converges uniformly for $C \in [0, 1]$, so the order of the limit and the infinite sum can be exchanged. Every term of the infinite series in (2.D.1) converges to zero in the limit $C \to 1$, hence the limit in (2.D.1) is zero. This completes the proof of Lemma 2.5.8.

2.E Proof of the properties in (2.5.29) for OFDM signals

Consider an OFDM signal from Section 2.5.6. The sequence in (2.5.27) is a martingale. From (2.5.26), for every $i \in \{0, \ldots, n\}$,

$$Y_i = \mathbb{E}\Big[\max_{0 \le t \le T} |s(t; X_0, \ldots, X_{n-1})| \Big| X_0, \ldots, X_{i-1} \Big].$$

The conditional expectation for the RV Y_{i-1} refers to the case where only $X_0, \ldots, X_{i-2}$ are revealed. Let X'_{i-1} and X_{i-1} be independent copies, which are also independent of $X_0, \ldots, X_{i-2}, X_i, \ldots, X_{n-1}$. Then, for every $1 \le i \le n$,

$$\begin{aligned} Y_{i-1} &= \mathbb{E}\Big[\max_{0 \le t \le T} |s(t; X_0, \ldots, X'_{i-1}, X_i, \ldots, X_{n-1})| \,\Big|\, X_0, \ldots, X_{i-2} \Big] \\ &= \mathbb{E}\Big[\max_{0 \le t \le T} |s(t; X_0, \ldots, X'_{i-1}, X_i, \ldots, X_{n-1})| \,\Big|\, X_0, \ldots, X_{i-2}, X_{i-1} \Big]. \end{aligned}$$

Since $|\mathbb{E}(Z)| \le \mathbb{E}(|Z|)$, then for $i \in \{1, \ldots, n\}$

$$|Y_i - Y_{i-1}| \le \mathbb{E}_{X'_{i-1}, X_i, \ldots, X_{n-1}} \Big[|U - V| \,\Big|\, X_0, \ldots, X_{i-1} \Big] \tag{2.E.1}$$

where

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

From (2.5.24)

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

By assumption, $|X_{i-1}| = |X'_{i-1}| = 1$, and therefore a.s.

$$|X_{i-1} - X'_{i-1}| \le 2 \Longrightarrow |Y_i - Y_{i-1}| \le \frac{2}{\sqrt{n}}.$$

We now obtain an upper bound on the conditional variance $\mathsf{var}(Y_i \,|\, \mathcal{F}_{i-1}) = \mathbb{E}\big[(Y_i - Y_{i-1})^2 \,|\, \mathcal{F}_{i-1}\big]$. Since $\big(\mathbb{E}(Z)\big)^2 \le \mathbb{E}(Z^2)$ for a real-valued RV Z, from (2.E.1), (2.E.2) and the tower property for conditional expectations it follows that

$$\mathbb{E}\big[(Y_i - Y_{i-1})^2 \,|\mathcal{F}_{i-1}\big] \le \frac{1}{n} \cdot \mathbb{E}_{X'_{i-1}}\big[|X_{i-1} - X'_{i-1}|^2 \,|\, \mathcal{F}_{i-1}\big]$$

where $\mathcal{F}_{i-1}$ is the σ-algebra generated by $X_0, \ldots, X_{i-2}$. Due to symmetry in the PSK constellation, and also due to the independence of X_{i-1}, X'_{i-1} in $X_0, \ldots, X_{i-2}$, we have

$$\begin{aligned}
\mathbb{E}\big[(Y_i - Y_{i-1})^2 \,|\, \mathcal{F}_{i-1}\big] &\le \frac{1}{n}\, \mathbb{E}\big[|X_{i-1} - X'_{i-1}|^2 \,|\, X_0, \ldots, X_{i-2}\big] \\
&= \frac{1}{n}\, \mathbb{E}\big[|X_{i-1} - X'_{i-1}|^2\big] \\
&= \frac{1}{n}\, \mathbb{E}\Big[|X_{i-1} - X'_{i-1}|^2 \,|\, X_{i-1} = e^{\frac{j\pi}{M}}\Big] \\
&= \frac{1}{nM} \sum_{l=0}^{M-1} \Big| e^{\frac{j\pi}{M}} - e^{\frac{j(2l+1)\pi}{M}} \Big|^2 \\
&= \frac{4}{nM} \sum_{l=1}^{M-1} \sin^2\Big(\frac{\pi l}{M}\Big) = \frac{2}{n}.
\end{aligned}$$

The last equality holds since

$$\begin{aligned}
\sum_{l=1}^{M-1} \sin^2\Big(\frac{\pi l}{M}\Big) &= \frac{1}{2} \sum_{l=0}^{M-1} \Big(1 - \cos\Big(\frac{2\pi l}{M}\Big)\Big) \\
&= \frac{M}{2} - \frac{1}{2}\, \mathrm{Re}\Big\{ \sum_{l=0}^{M-1} e^{j2l\pi/M} \Big\} \\
&= \frac{M}{2} - \frac{1}{2}\, \mathrm{Re}\Big\{ \frac{1 - e^{2j\pi}}{1 - e^{j2\pi/M}} \Big\} = \frac{M}{2}.
\end{aligned}$$

3

The Entropy Method, Logarithmic Sobolev Inequalities, and Transportation-Cost Inequalities

This chapter introduces the entropy method for deriving concentration inequalities for functions of a large number of independent random variables, and exhibits its multiple connections to information theory. The chapter is divided into four parts. Sections 3.1–3.3 introduce the basic ingredients of the entropy method and closely related topics, such as logarithmic Sobolev inequalities. These topics underlie the so-called functional approach to deriving concentration inequalities. Section 3.4 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. Section 3.5 gives a brief summary of some results on concentration for dependent random variables, emphasizing the connections to information-theoretic ideas. Section 3.6 lists several applications of concentration inequalities and the entropy method to problems in information theory, including 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.

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 $f\colon \mathcal{X}^n \to \mathbb{R}$, 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

$$\mathbb{P}(|U - \mathbb{E}U| \geq r) \leq \frac{\mathsf{var}(U)}{r^2}, \quad \forall\, r > 0. \tag{3.1.1}$$

However, in many instances a bound like (3.1.1) is not nearly as tight as one would like, so ideally we aim for Gaussian-type bounds

$$\mathbb{P}(|U - \mathbb{E}U| \geq r) \leq K \exp\left(-\kappa r^2\right), \quad \forall\, r > 0 \tag{3.1.2}$$

for some constants $K, \kappa > 0$. Whenever such a bound is available, K is typically 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 (3.1.2), such as the inequalities of Azuma (Theorem 2.2.2) and McDiarmid (Theorem 2.2.3). 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 [43] (see also [3]). In the following, we will always assume (unless specified otherwise) that the function $f\colon \mathcal{X}^n \to \mathbb{R}$ 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*:

$$\mathbb{E}[\exp(\lambda U)] = \mathbb{E}\left[\exp\left(\lambda f(X^n)\right)\right] < \infty, \qquad \forall \lambda \in \mathbb{R} \tag{3.1.3}$$

  [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 technique** — 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 $\Lambda(\lambda) \triangleq \ln \mathbb{E}[\exp(\lambda U)]$, $\lambda \in \mathbb{R}$. (This is also the starting point of the martingale approach, see Chapter 2.)

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

$$\frac{\mathrm{d}P^{(\lambda f)}}{\mathrm{d}P} = \frac{\exp(\lambda f)}{\mathbb{E}[\exp(\lambda f)]} = \exp\big(\lambda f - \Lambda(\lambda)\big). \tag{3.1.4}$$

   If the function f and the probability distribution P are such that

$$D(P^{(\lambda f)}\|P) \leq \frac{c\lambda^2}{2} \tag{3.1.5}$$

   for some $c > 0$, then the Gaussian bound (3.1.2) holds with $K = 2$ and $\kappa = \frac{1}{2c}$. The standard way to establish (3.1.5) 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 (3.1.5) 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[1] terms

$$D\Big(P^{(\lambda f)}_{X_i|\bar{X}^i}\Big\|P_{X_i}\Big|P^{(\lambda f)}_{\bar{X}^i}\Big), \qquad i = 1, \ldots, n \tag{3.1.6}$$

[1]Recall the usual definition of the conditional divergence:

$$D(P_{V|U}\|Q_{V|U}|P_U) \triangleq \int P_U(\mathrm{d}u) D(P_{V|U=u}\|Q_{V|U=u}).$$

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., $\bar{X}^i = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$. Despite their formidable appearance, the conditional divergences in (3.1.6) are easier to handle because, for each given realization $\bar{X}^i = \bar{x}^i$, the ith such term involves a single-variable function $f_i(\cdot|\bar{x}^i)\colon \mathcal{X} \to \mathbb{R}$ defined by $f_i(y|\bar{x}^i) \triangleq f(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n)$ and the corresponding tilted distribution $P^{(\lambda f)}_{X_i|\bar{X}^i=\bar{x}^i}$, where

$$\frac{\mathrm{d}P^{(\lambda f)}_{X_i|\bar{X}^i=\bar{x}^i}}{\mathrm{d}P_{X_i}} = \frac{\exp\left(\lambda f_i(\cdot|\bar{x}^i)\right)}{\mathbb{E}\left[\exp\left(\lambda f_i(X_i|\bar{x}^i)\right)\right]}, \qquad \forall \bar{x}^i \in \mathcal{X}^{n-1}. \quad (3.1.7)$$

In fact, from (3.1.4) and (3.1.7), 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

$$D\Big(P^{(\lambda f_i(\cdot|\bar{x}^i))}_{X_i}\Big\|P_{X_i}\Big) \le \frac{c_i\lambda^2}{2}, \ \forall i \in \{1, \ldots, n\}, \bar{x}^i \in \mathcal{X}^{n-1} \quad (3.1.8)$$

then (3.1.5) holds with $c = \sum_{i=1}^n c_i$ (to be shown explicitly later), which in turn gives the bound

$$\mathbb{P}\Big(|f(X^n) - \mathbb{E}f(X^n)| \ge r\Big) \le 2\exp\left(-\frac{r^2}{2\sum_{i=1}^n c_i}\right), \quad (3.1.9)$$

for all $r > 0$. Again, one would typically use logarithmic Sobolev inequalities to verify (3.1.8). Conceptually, the tensorization step is similar to "single-letter" techniques common in information theory.

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 technique revisited

We start by recalling the Chernoff technique (see Section 2.2.1), which reduces the problem of bounding the deviation probability $\mathbb{P}(U \ge r)$

to the analysis of the logarithmic moment-generating function $\Lambda(\lambda) = \ln \mathbb{E}[\exp(\lambda U)]$:

$$\mathbb{P}(U \geq r) \leq \exp\Big(\Lambda(\lambda) - \lambda r\Big), \qquad \forall \lambda > 0.$$

The following properties of $\Lambda(\lambda)$ will be useful later on:

- $\Lambda(0) = 0$
- Because of the exponential integrability of U [cf. (3.1.3)], $\Lambda(\lambda)$ is infinitely differentiable, and one can interchange derivative and expectation. In particular,

$$\Lambda'(\lambda) = \frac{\mathbb{E}[U \exp(\lambda U)]}{\mathbb{E}[\exp(\lambda U)]},$$
$$\Lambda''(\lambda) = \frac{\mathbb{E}[U^2 \exp(\lambda U)]}{\mathbb{E}[\exp(\lambda U)]} - \left(\frac{\mathbb{E}[U \exp(\lambda U)]}{\mathbb{E}[\exp(\lambda U)]}\right)^2. \tag{3.1.10}$$

  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

$$\lim_{\lambda \to 0} \frac{\Lambda(\lambda)}{\lambda} = 0. \tag{3.1.11}$$

3.1.2 The Herbst argument

The second ingredient of the entropy method consists in relating the logarithmic moment-generating function to a certain relative entropy. The underlying technique is often referred to as the *Herbst argument* because its basic idea had been described in an unpublished 1975 letter from I. Herbst to L. Gross (the first explicit mention of this letter appears in a paper by Davies and Simon [115]).

Given any function $g\colon \mathcal{X}^n \to \mathbb{R}$ which is exponentially integrable with respect to 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

$$\begin{aligned} D(P^{(g)}\|P) &= \int_{\mathcal{X}^n} \ln\left(\frac{\mathrm{d}P^{(g)}}{\mathrm{d}P}\right) \mathrm{d}P^{(g)} \\ &= \int_{\mathcal{X}^n} \frac{\mathrm{d}P^{(g)}}{\mathrm{d}P} \ln\left(\frac{\mathrm{d}P^{(g)}}{\mathrm{d}P}\right) \mathrm{d}P \\ &= \frac{\mathbb{E}[g\exp(g)]}{\mathbb{E}[\exp(g)]} - \ln \mathbb{E}[\exp(g)]. \end{aligned}$$

In particular, if we let $g = tf$ for some $t \neq 0$, then

$$\begin{aligned} D(P^{(tf)}\|P) &= \frac{t \cdot \mathbb{E}[f\exp(tf)]}{\mathbb{E}[\exp(tf)]} - \ln \mathbb{E}[\exp(tf)] \\ &= t\Lambda'(t) - \Lambda(t) \\ &= t^2 \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\Lambda(t)}{t}\right), \end{aligned} \tag{3.1.12}$$

where in the second line we have used (3.1.10). Integrating from $t = 0$ to $t = \lambda$ and using (3.1.11), we get

$$\Lambda(\lambda) = \lambda \int_0^\lambda \frac{D(P^{(tf)}\|P)}{t^2} \mathrm{d}t. \tag{3.1.13}$$

Combining (3.1.13) with (2.2.2), we have proved the following:

Proposition 3.1. Let $U = f(X^n)$ be a zero-mean random variable that is exponentially integrable. Then, for any $r \geq 0$,

$$\mathbb{P}(U \geq r) \leq \exp\left(\lambda \int_0^\lambda \frac{D(P^{(tf)}\|P)}{t^2} \mathrm{d}t - \lambda r\right), \qquad \forall \lambda > 0. \tag{3.1.14}$$

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 3.1.1. Suppose that the function f and the probability distribution P of X^n are such that

$$D(P^{(tf)}\|P) \leq \frac{ct^2}{2}, \qquad \forall t > 0 \tag{3.1.15}$$

for some constant $c > 0$. Then

$$\mathbb{P}(U \geq r) \leq \exp\left(-\frac{r^2}{2c}\right), \quad \forall\, r \geq 0. \tag{3.1.16}$$

Proof. Using (3.1.15) to upper-bound the integrand on the right-hand side of (3.1.14), we get

$$\mathbb{P}(U \geq r) \leq \exp\left(\frac{c\lambda^2}{2} - \lambda r\right), \qquad \forall \lambda > 0. \tag{3.1.17}$$

Optimizing over $\lambda > 0$ to get the tightest bound gives $\lambda = \frac{r}{c}$, and its substitution in (3.1.17) gives the bound in (3.1.16). □

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 a conditional distribution of one of the variables given the rest. The tensorization step hinges on the following simple bound:

Proposition 3.2. 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

$$D(Q\|P) \leq \sum_{i=1}^{n} D(Q_{X_i|\bar{X}^i}\|P_{X_i}|Q_{\bar{X}^i}). \tag{3.1.18}$$

Proof. From the relative entropy chain rule

$$\begin{aligned} D(Q||P) &= \sum_{i=1}^{n} D(Q_{X_i\,|\,X^{i-1}} \,\|\, P_{X_i|X^{i-1}} \,|\, Q_{X^{i-1}}) \\ &= \sum_{i=1}^{n} D(Q_{X_i\,|\,X^{i-1}} \,\|\, P_{X_i} \,|\, Q_{X^{i-1}}) \end{aligned} \tag{3.1.19}$$

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 $i \in \{1, \ldots, n\}$,

$$\begin{aligned} & D(Q_{X_i|\bar{X}^i} \| P_{X_i} | Q_{\bar{X}^i}) - D(Q_{X_i|X^{i-1}} \| P_{X_i} | Q_{X^{i-1}}) \\ & \quad = \mathbb{E}_Q \left[\ln \frac{\mathrm{d}Q_{X_i|\bar{X}^i}}{\mathrm{d}P_{X_i}} \right] - \mathbb{E}_Q \left[\ln \frac{\mathrm{d}Q_{X_i|X^{i-1}}}{\mathrm{d}P_{X_i}} \right] \\ & \quad = \mathbb{E}_Q \left[\ln \frac{\mathrm{d}Q_{X_i|\bar{X}^i}}{\mathrm{d}Q_{X_i|X^{i-1}}} \right] \\ & \quad = D(Q_{X_i|\bar{X}^i} \| Q_{X_i|X^{i-1}} | Q_{\bar{X}^i}) \geq 0. \end{aligned} \tag{3.1.20}$$

Hence, by combining (3.1.19) and (3.1.20), we get the inequality in (3.1.18). □

Remark 3.1. The quantity on the right-hand side of (3.1.18) is actually the so-called *erasure divergence* $D^-(Q\|P)$ between Q and P (see [116, Definition 4]), which in the case of arbitrary Q and P is defined by

$$D^-(Q\|P) \triangleq \sum_{i=1}^{n} D(Q_{X_i|\bar{X}^i} \| P_{X_i|\bar{X}^i} | Q_{\bar{X}^i}). \tag{3.1.21}$$

Because in the inequality (3.1.18) 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 (3.1.21),

$$\begin{aligned} & D^-(Q_{X_1,X_2} \| P_{X_1,X_2}) \\ & \quad = D(Q_{X_1|X_2} \| P_{X_1|X_2} | Q_{X_2}) + D(Q_{X_2|X_1} \| P_{X_2|X_1} | Q_{X_1}) \\ & \quad = 2D(Q_{X_1,X_2} \| P_{X_1,X_2}). \end{aligned}$$

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 3.2 with $Q = P^{(tf)}$ to bound the divergence in the integrand in (3.1.14), we obtain from Corollary 3.1.1 the following:

Proposition 3.3. For any $r \geq 0$, we have

$$\mathbb{P}(U \geq r) \leq \exp\left(\lambda \sum_{i=1}^{n} \int_0^\lambda \frac{D(P_{X_i|\bar{X}^i}^{(tf)} \| P_{X_i} | P_{\bar{X}^i}^{(tf)})}{t^2} \mathrm{d}t - \lambda r\right), \ \forall \lambda > 0. \tag{3.1.22}$$

The conditional divergences in the integrand in (3.1.22) 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 $f_i(\cdot|\bar{x}^i) \colon \mathcal{X} \to \mathbb{R}$ by setting

$$f_i(y|\bar{x}^i) \triangleq f(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n), \qquad \forall y \in \mathcal{X}. \tag{3.1.23}$$

Then

$$\frac{\mathrm{d}P_{X_i|\bar{X}^i=\bar{x}^i}^{(f)}}{\mathrm{d}P_{X_i}} = \frac{\exp\left(f_i(\cdot|\bar{x}^i)\right)}{\mathbb{E}\left[\exp\left(f_i(X_i|\bar{x}^i)\right)\right]}. \tag{3.1.24}$$

In other words, $P_{X_i|\bar{X}^i=\bar{x}^i}^{(f)}$ is the $f_i(\cdot|\bar{x}^i)$-tilting of P_{X_i}, the marginal distribution of 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 3.1.2. 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 $t > 0$,

$$D\Big(P_{X_i}^{(tf_i(\cdot|\bar{x}^i))} \| P_{X_i}\Big) \leq \frac{c_i t^2}{2}, \qquad \forall i \in \{1, \ldots, n\}, \ \bar{x}^i \in \mathcal{X}^{n-1}. \tag{3.1.25}$$

Then

$$\mathbb{P}\Big(f(X^n) - \mathbb{E}f(X^n) \geq r\Big) \leq \exp\left(-\frac{r^2}{2\sum_{i=1}^n c_i}\right), \quad \forall\, r > 0. \tag{3.1.26}$$

Remark 3.2. Note the obvious similarity between the bound (3.1.26) and McDiarmid's inequality (2.2.20). Indeed, as we will show later on in Section 3.3.4, it is possible to derive McDiarmid's inequality using the entropy method.

Proof. For any $t > 0$

$$D(P^{(tf)}\|P) \le \sum_{i=1}^{n} D\Big(P^{(tf)}_{X_i|\bar{X}^i}\|P_{X_i} \,|\, P^{(tf)}_{\bar{X}^i}\Big) \tag{3.1.27}$$

$$= \sum_{i=1}^{n} \int_{\mathcal{X}^{n-1}} D\Big(P^{(tf)}_{X_i|\bar{X}^i=\bar{x}^i}\|P_{X_i}\Big)\, P^{(tf)}_{\bar{X}^i}(\mathrm{d}\bar{x}^i) \tag{3.1.28}$$

$$= \sum_{i=1}^{n} \int_{\mathcal{X}^{n-1}} D\Big(P^{(tf_i(\cdot|\bar{x}^i))}_{X_i}\|P_{X_i}\Big)\, P^{(tf)}_{\bar{X}^i}(\mathrm{d}\bar{x}^i) \tag{3.1.29}$$

$$\le \frac{t^2}{2} \cdot \sum_{i=1}^{n} c_i \tag{3.1.30}$$

where (3.1.27) follows from the tensorization of the relative entropy, (3.1.28) 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, (3.1.29) follows from (3.1.23) and (3.1.24) which implies that $P^{(tf)}_{X_i|\bar{X}^i=\bar{x}^i} = P^{(tf_i(\cdot|\bar{x}^i))}_{X_i}$, and inequality (3.1.30) holds by the assumption in (3.1.25). Finally, the inequality in (3.1.26) follows from (3.1.30) and Corollary 3.1.1. □

3.1.4 Preview: logarithmic Sobolev inequalities

Ultimately, the success of the entropy method hinges on demonstrating that the bounds in (3.1.25) hold for the function $f\colon \mathcal{X}^n \to \mathbb{R}$ 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 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 $g\colon \mathcal{X} \to \mathbb{R}$, such that for any $a \ge 0$ and $g \in \mathcal{A}$, we also have $ag \in \mathcal{A}$. Let $E\colon \mathcal{A} \to \mathbb{R}^+$ be a non-negative functional that is homogeneous of degree 2, i.e., for any $a \ge 0$ and $g \in \mathcal{A}$, we have $E(ag) = a^2 E(g)$. We are interested in the case when there exists

a constant $c > 0$, such that the inequality

$$D(\mu^{(g)}\|\mu) \leq \frac{cE(g)}{2} \tag{3.1.31}$$

holds for any $g \in \mathcal{A}$. Now suppose that, for each $i \in \{1, \ldots, n\}$, inequality (3.1.31) holds with $\mu = P_{X_i}$ 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 (3.1.23). Then, the bounds in (3.1.25) hold, since from (3.1.31) and the above properties of the functional E it follows that for every $t > 0$ and $\bar{x}^i \in \mathcal{X}^{n-1}$

$$\begin{aligned} D\Big(P^{(tf)}_{X_i|\bar{X}^i=\bar{x}^i}\|P_{X_i}\Big) &\leq \frac{c_i\, E\big(t\, f_i(\cdot|\bar{x}^i)\big)}{2} \\ &= \frac{c_i t^2\, E\big(f_i(\cdot|\bar{x}^i)\big)}{2} \\ &\leq \frac{c_i t^2}{2}, \quad \forall\, i \in \{1, \ldots, n\}. \end{aligned}$$

Consequently, the Gaussian concentration inequality in (3.1.26) follows from Corollary 3.1.2.

3.2 The Gaussian logarithmic Sobolev inequality

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 [44], and it connects two key information-theoretic quantities, 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_K^n 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_s^n. We will also write G^n for G_1^n when $n \geq 2$, and G for G_1^1. We will denote by γ_K^n, γ_s^n, γ_s, and γ the corresponding densities.

We first state Gross's inequality in its (more or less) original form:

Theorem 3.2.1 (Log-Sobolev inequality for the Gaussian measure)**.** For $Z \sim G^n$ and for any smooth[2] function $\phi\colon \mathbb{R}^n \to \mathbb{R}$, we have

$$\mathbb{E}[\phi^2(Z)\ln\phi^2(Z)] - \mathbb{E}[\phi^2(Z)]\ln\mathbb{E}[\phi^2(Z)] \leq 2\,\mathbb{E}\left[\|\nabla\phi(Z)\|^2\right], \quad (3.2.1)$$

where $\|\cdot\|$ denotes the usual Euclidean norm on $\mathbb{R}^n$.

Remark 3.3. As shown by Carlen [117], equality in (3.2.1) 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 3.4. There is no loss of generality in assuming that $\mathbb{E}[\phi^2(Z)] = 1$. Then (3.2.1) can be rewritten as

$$\mathbb{E}[\phi^2(Z)\ln\phi^2(Z)] \leq 2\,\mathbb{E}\left[\|\nabla\phi(Z)\|^2\right] \text{ if } \mathbb{E}[\phi^2(Z)] = 1,\ Z \sim G^n. \quad (3.2.2)$$

Moreover, a simple rescaling argument shows that, for $Z \sim G_s^n$ and an arbitrary smooth function ϕ with $\mathbb{E}[\phi^2(Z)] = 1$,

$$\mathbb{E}[\phi^2(Z)\ln\phi^2(Z)] \leq 2s\,\mathbb{E}\left[\|\nabla\phi(Z)\|^2\right]. \quad (3.2.3)$$

We give an information-theoretic proof of the Gaussian LSI (Theorem 3.2.1) later in this section; we refer the reader to [118] as an example of a typical proof using techniques from functional analysis.

From an information-theoretic point of view, the Gaussian LSI (3.2.1) relates two measures of (dis)similarity between probability mea-

[2]Here and elsewhere, we will use the term "smooth" somewhat loosely to mean "satisfying enough regularity conditions to make sure that all relevant quantities are well-defined." In the present context, smooth means that both ϕ and $\nabla\phi$ should be square-integrable with respect to the standard Gaussian measure G^n.

sures — 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, and suppose that the Radon–Nikodym derivative $\mathrm{d}P_1/\mathrm{d}P_2 \equiv p_1/p_2$ is differentiable P_2-a.e. Then the *relative Fisher information* (or *Fisher information distance*) between P_1 and P_2 is defined as (see [119, Eq. (6.4.12)])

$$I(P_1\|P_2) \triangleq \int_{\mathbb{R}^n} \left\|\nabla \ln \frac{p_1(z)}{p_2(z)}\right\|^2 p_1(z)\mathrm{d}z = \mathbb{E}_{P_1}\left[\left\|\nabla \ln \frac{\mathrm{d}P_1}{\mathrm{d}P_2}\right\|^2\right], \quad (3.2.4)$$

whenever the above integral converges. Under suitable regularity conditions, $I(P_1\|P_2)$ admits the equivalent form (see [120, Eq. (1.108)])

$$\begin{aligned} I(P_1\|P_2) &= 4\int_{\mathbb{R}^n} p_2(z)\left\|\nabla\sqrt{\frac{p_1(z)}{p_2(z)}}\right\|^2 \mathrm{d}z \\ &= 4\,\mathbb{E}_{P_2}\left[\left\|\nabla\sqrt{\frac{\mathrm{d}P_1}{\mathrm{d}P_2}}\right\|^2\right]. \end{aligned} \quad (3.2.5)$$

Remark 3.5. One condition under which (3.2.5) holds is as follows. Let $\xi\colon \mathbb{R}^n \to \mathbb{R}^n$ be the *distributional* (or *weak*) *gradient* of $\sqrt{\mathrm{d}P_1/\mathrm{d}P_2} = \sqrt{p_1/p_2}$, so that 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 $\psi \in C_c^\infty(\mathbb{R}^n)$ [121, Sec. 6.6]. (Here, $\partial_i\psi$ denotes the ith coordinate of $\nabla\psi$.) Then (3.2.5) holds, provided $\xi \in L^2(P_2)$.

Now let us fix a smooth function $\phi\colon \mathbb{R}^n \to \mathbb{R}$ 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

$$\begin{aligned}\mathbb{E}\left[\phi^2(Z)\ln\phi^2(Z)\right] &= \mathbb{E}\left[\left(\frac{\mathrm{d}P_Y}{\mathrm{d}P_Z}(Z)\right)\ln\left(\frac{\mathrm{d}P_Y}{\mathrm{d}P_Z}(Z)\right)\right]\\ &= D(P_Y\|P_Z),\end{aligned} \tag{3.2.6}$$

and on the other, from (3.2.5),

$$\mathbb{E}\left[\|\nabla\phi(Z)\|^2\right] = \mathbb{E}\left[\left\|\nabla\sqrt{\frac{\mathrm{d}P_Y}{\mathrm{d}P_Z}(Z)}\right\|^2\right] = \frac{1}{4}I(P_Y\|P_Z). \tag{3.2.7}$$

Substituting (3.2.6) and (3.2.7) into (3.2.2), we obtain the inequality

$$D(P_Y\|P_Z) \le \frac{1}{2}I(P_Y\|P_Z), \qquad P_Z = G^n \tag{3.2.8}$$

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 (3.2.8), we can derive (3.2.2) 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 $s > 0$, (3.2.3) can be written as

$$D(P_Y\|P_Z) \le \frac{s}{2}I(P_Y\|P_Z), \qquad P_Z = G_s^n. \tag{3.2.9}$$

Now let us apply the Gaussian LSI (3.2.1) to functions of the form $\phi = \exp(g/2)$ for all suitably well-behaved $g\colon \mathbb{R}^n \to \mathbb{R}$. Then we obtain

$$\mathbb{E}\left[\exp\big(g(Z)\big)\ln\frac{\exp\big(g(Z)\big)}{\mathbb{E}[\exp\big(g(Z)\big)]}\right] \le \frac{1}{2}\mathbb{E}\left[\|\nabla g(Z)\|^2\exp\big(g(Z)\big)\right], \tag{3.2.10}$$

where $Z \sim G^n$. If we let $P = G^n$ and denote by $P^{(g)}$ the g-tilting of P, then we can recognize the left-hand side of (3.2.10) as $\mathbb{E}[\exp\big(g(Z)\big)]\cdot D(P^{(g)}\|P)$. Similarly, 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 with respect to $P^{(g)}$. We therefore obtain the so-called *modified log-Sobolev inequality* for the standard Gaussian measure:

$$D(P^{(g)}\|P) \le \frac{1}{2}\mathbb{E}_P^{(g)}\left[\|\nabla g\|^2\right], \qquad P = G^n \tag{3.2.11}$$

which holds for all smooth functions $g\colon \mathbb{R}^n \to \mathbb{R}$ that are exponentially integrable with respect to G^n. Observe that (3.2.11) implies (3.1.31) 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 3.2.1, and then discuss several applications of the modified log-Sobolev inequality (3.2.11) 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 3.2.1 via tensorization: We first establish the $n = 1$ case and then scale up to general n using suitable (sub)additivity properties. Indeed, suppose that (3.2.1) 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 $\phi\colon \mathbb{R}^n \to \mathbb{R}$, 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 3.2 we can write

$$\begin{aligned}\mathbb{E}_P\left[\phi^2(X)\ln\phi^2(X)\right] &= \mathbb{E}_P\left[\frac{\mathrm{d}Q}{\mathrm{d}P}\ln\frac{\mathrm{d}Q}{\mathrm{d}P}\right] \\ &= D(Q\|P) \\ &\leq \sum_{i=1}^n D(Q_{X_i|\bar{X}^i}\|P_{X_i}|Q_{\bar{X}^i}). \qquad (3.2.12)\end{aligned}$$

Following the same steps as the ones that led to (3.1.23), 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 $\phi_i(\cdot|\bar{x}^i)\colon \mathbb{R} \to \mathbb{R}$ 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 $i \in \{1, \dots, n\}, \bar{x}^i \in \mathbb{R}^{n-1}$. With this, we can write

$$\begin{aligned}
&D(Q_{X_i|\bar{X}^i} \| P_{X_i} | Q_{\bar{X}^i}) \\
&\quad = \mathbb{E}_Q \left[\ln \frac{\mathrm{d}Q_{X_i|\bar{X}^i}}{\mathrm{d}P_{X_i}} \right] \\
&\quad = \mathbb{E}_P \left[\frac{\mathrm{d}Q}{\mathrm{d}P} \ln \frac{\mathrm{d}Q_{X_i|\bar{X}^i}}{\mathrm{d}P_{X_i}} \right] \\
&\quad = \mathbb{E}_P \left[\phi^2(X) \ln \frac{\phi_i^2(X_i|\bar{X}^i)}{\mathbb{E}_P[\phi_i^2(X_i|\bar{X}^i)|\bar{X}^i]} \right] \\
&\quad = \mathbb{E}_P \left[\phi_i^2(X_i|\bar{X}^i) \ln \frac{\phi_i^2(X_i|\bar{X}^i)}{\mathbb{E}_P[\phi_i^2(X_i|\bar{X}^i)|\bar{X}^i]} \right] \\
&\quad = \int_{\mathbb{R}^{n-1}} \mathbb{E}_P \left[\phi_i^2(X_i|\bar{x}^i) \ln \frac{\phi_i^2(X_i|\bar{x}^i)}{\mathbb{E}_P[\phi_i^2(X_i|\bar{x}^i)]} \right] P_{\bar{X}^i}(\mathrm{d}\bar{x}^i). \qquad (3.2.13)
\end{aligned}$$

Since each $X_i \sim G$, we can apply the Gaussian LSI (3.2.1) to the univariate functions $\phi_i(\cdot|\bar{x}^i)$ to get

$$\mathbb{E}_P \left[\phi_i^2(X_i|\bar{x}^i) \ln \frac{\phi_i^2(X_i|\bar{x}^i)}{\mathbb{E}_P[\phi_i^2(X_i|\bar{x}^i)]} \right] \leq 2\, \mathbb{E}_P \left[\left(\phi_i'(X_i|\bar{x}^i) \right)^2 \right] \qquad (3.2.14)$$

for all $i = 1, \dots, n$ and all $\bar{x}^i \in \mathbb{R}^{n-1}$, where the prime denotes the derivative of $\phi_i(y|\bar{x}^i)$ with respect to y:

$$\phi_i'(y|\bar{x}^i) = \frac{\mathrm{d}\phi_i(y|\bar{x}^i)}{\mathrm{d}y} = \frac{\partial \phi(x)}{\partial x_i}\Big|_{x_i = y}.$$

Since $X_1, \dots, X_n$ are i.i.d. under P, we can express (3.2.14) as

$$\mathbb{E}_P \left[\phi_i^2(X_i|\bar{x}^i) \ln \frac{\phi_i^2(X_i|\bar{x}^i)}{\mathbb{E}_P[\phi_i^2(X_i|\bar{x}^i)]} \right] \leq 2\, \mathbb{E}_P \left[\left(\partial_i \phi(X) \right)^2 \Big| \bar{X}^i = \bar{x}^i \right],$$

where $\partial_i \phi$ denotes the ith coordinate of the gradient $\nabla \phi$. Substituting this bound into (3.2.13), we have

$$D(Q_{X_i|\bar{X}^i} \| P_{X_i} | Q_{\bar{X}^i}) \leq 2\, \mathbb{E}_P \left[\left(\partial_i \phi(X) \right)^2 \right].$$

In turn, using this to bound each term in the summation on the right-hand side of (3.2.12) together with the fact that $\sum_{i=1}^n \left(\partial_i \phi(x) \right)^2 =$

$\|\nabla\phi(x)\|^2$, we get

$$\mathbb{E}_P\left[\phi^2(X)\ln\phi^2(X)\right] \leq 2\,\mathbb{E}_P\left[\|\nabla\phi(X)\|^2\right], \tag{3.2.15}$$

which is precisely the Gaussian LSI (3.2.2) in $\mathbb{R}^n$. Thus, if the Gaussian LSI holds for $n = 1$, then it holds for all $n \geq 2$.

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 [45], and the equivalence between Stam's inequality and (3.2.1) was only noted much later by Carlen [117]. 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 [122, 47], does not directly rely on de Bruijn's identity or on the entropy-power inequality.

First, let us start with some definitions. Let Y be a real-valued random variable with density p_Y. The *differential entropy* of Y is given by

$$h(Y) = h(p_Y) \triangleq -\int_{-\infty}^{\infty} p_Y(y)\ln p_Y(y)\mathrm{d}y, \tag{3.2.16}$$

provided the integral exists. If it does, then the *entropy power* of Y is given by

$$N(Y) \triangleq \frac{\exp(2h(Y))}{2\pi e}. \tag{3.2.17}$$

Moreover, if the density p_Y is differentiable, then the *Fisher information* (with respect to a location parameter) is given by

$$J(Y) = J(p_Y) = \int_{-\infty}^{\infty}\left(\frac{\mathrm{d}}{\mathrm{d}y}\ln p_Y(y)\right)^2 p_Y(y)\mathrm{d}y = \mathbb{E}[\rho_Y^2(Y)], \tag{3.2.18}$$

where $\rho_Y(y) \triangleq (\mathrm{d}/\mathrm{d}y)\ln p_Y(y) = \frac{p_Y'(y)}{p_Y(y)}$ is known as the *score function*.

Remark 3.6. In theoretical statistics, an alternative definition of the Fisher information (with respect to a location parameter) of a real-valued random variable Y is (see [123, Definition 4.1])

$$J(Y) \triangleq \sup\left\{\left(\mathbb{E}\psi'(Y)\right)^2 : \psi \in C^1, \mathbb{E}[\psi^2(Y)] = 1\right\} \tag{3.2.19}$$

where the supremum is taken over the set of all continuously differentiable functions ψ with compact support, such that $\mathbb{E}[\psi^2(Y)] = 1$. 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 (3.2.19) exists and is finite if and only if Y has an absolutely continuous density p_Y, in which case $J(Y)$ is equal to (3.2.18) (see [123, Theorem 4.2]).

We will need the following facts:

1. If $D(P_Y\|G_s) < \infty$, then
$$D(P_Y\|G_s) = \frac{1}{2}\ln\frac{1}{N(Y)} + \frac{1}{2}\ln s - \frac{1}{2} + \frac{1}{2s}\,\mathbb{E}Y^2. \tag{3.2.20}$$
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 (3.2.20).

2. If $J(Y) < \infty$ and $\mathbb{E}Y^2 < \infty$, then for any $s > 0$
$$I(P_Y\|G_s) = J(Y) + \frac{1}{s^2}\,\mathbb{E}Y^2 - \frac{2}{s} < \infty, \tag{3.2.21}$$
where $I(\cdot\|\cdot)$ is the relative Fisher information, cf. (3.2.4). Indeed:
$$\begin{aligned} I(P_Y\|G_s) &= \int_{-\infty}^{\infty} p_Y(y)\left(\frac{\mathrm{d}}{\mathrm{d}y}\ln p_Y(y) - \frac{\mathrm{d}}{\mathrm{d}y}\ln\gamma_s(y)\right)^2\mathrm{d}y \\ &= \int_{-\infty}^{\infty} p_Y(y)\left(\rho_Y(y) + \frac{y}{s}\right)^2\mathrm{d}y \\ &= \mathbb{E}[\rho_Y^2(Y)] + \frac{2}{s}\,\mathbb{E}[Y\rho_Y(Y)] + \frac{1}{s^2}\,\mathbb{E}Y^2 \\ &= J(Y) + \frac{2}{s}\,\mathbb{E}[Y\rho_Y(Y)] + \frac{1}{s^2}\,\mathbb{E}Y^2. \end{aligned} \tag{3.2.22}$$

Because $\mathbb{E}Y^2 < \infty$, we have $\mathbb{E}|Y| < \infty$, so $\lim_{y\to\pm\infty} y\, p_Y(y) = 0$. Furthermore, integration by parts gives

$$\begin{aligned}\mathbb{E}[Y\rho_Y(Y)] &= \int_{-\infty}^{\infty} y\, \rho_Y(y)\, p_Y(y)\, \mathrm{d}y \\ &= \int_{-\infty}^{\infty} y\, p_Y'(y)\, \mathrm{d}y \\ &= \left(\lim_{y\to\infty} y\, p_Y(y) - \lim_{y\to-\infty} y\, p_Y(y)\right) - \int_{-\infty}^{\infty} p_Y(y)\, \mathrm{d}y \\ &= -1\end{aligned}$$

(see [124, Lemma A1] for another proof). Substituting this into (3.2.22), we get (3.2.21).

We are now in a position to prove the following result of Carlen [117]:

Proposition 3.4. 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, $D(P_Y\|G) \le \frac{1}{2}\, I(P_Y\|G)$.
2. Stam's inequality, $N(Y)J(Y) \ge 1$.

Remark 3.7. Carlen's original derivation in [117] 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) $\Rightarrow$ 2). If 1) holds, then

$$D(P_Y\|G_s) \le \frac{s}{2} I(P_Y\|G_s), \qquad \forall s > 0. \tag{3.2.23}$$

Since $J(Y)$ and $\mathbb{E}Y^2$ are finite by assumption, the right-hand side of (3.2.23) is finite and equal to (3.2.21). Therefore, $D(P_Y\|G_s)$ is also finite, and it is equal to (3.2.20). Hence, we can rewrite (3.2.23) 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

$$\ln \frac{1}{N(Y)} \leq sJ(Y) - \ln s - 1.$$

Importantly, this bound holds for *every* $s > 0$. Therefore, using the fact that

$$1 + \ln a = \inf_{s>0}(as - \ln s), \qquad \forall a > 0$$

we obtain Stam's inequality $N(Y)J(Y) \geq 1$.

To establish the converse implication 2) $\Rightarrow$ 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 [125], that expresses the divergence between two probability distributions 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

$$\mathsf{mmse}(Y, s) \triangleq \inf_{\varphi} \mathbb{E}[(Y - \varphi(\sqrt{s}Y + Z))^2], \tag{3.2.24}$$

where the infimum is over all measurable functions (estimators) $\varphi\colon \mathbb{R} \to \mathbb{R}$. It is well-known that the infimum in (3.2.24) 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* $u \mapsto \mathbb{E}_Q[Y|\sqrt{s}Y + Z = u]$, 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 outer expectation on the right-hand side is computed using the correct distribution P of Y. Then, the following relation holds for the divergence between P and Q (see [125, Theorem 1]):

$$D(P\|Q) = \frac{1}{2}\int_0^\infty [\mathsf{mse}_Q(Y,s) - \mathsf{mmse}(Y,s)]\,\mathrm{d}s. \tag{3.2.25}$$

We will apply the formula (3.2.25) to $P = P_Y$ and $Q = G$, where P_Y satisfies $\mathbb{E}Y = 0$ and $\mathbb{E}Y^2 = 1$. In that case it can be shown that, for any $s > 0$,

$$\mathsf{mse}_Q(Y,s) = \mathsf{mse}_G(Y,s) = \mathsf{lmmse}(Y,s),$$

where $\mathsf{lmmse}(Y,s)$ is the *linear* MMSE, i.e., the MMSE attainable by any *affine* estimator $u \mapsto au + b$, $a, b \in \mathbb{R}$:

$$\mathsf{lmmse}(Y,s) = \inf_{a,b\in\mathbb{R}} \mathbb{E}\left[\left(Y - a(\sqrt{s}Y + Z) - b\right)^2\right]. \tag{3.2.26}$$

The infimum in (3.2.26) is achieved by $a^* = \sqrt{s}/(1+s)$ and $b = 0$, giving

$$\mathsf{lmmse}(Y,s) = \frac{1}{1+s}. \tag{3.2.27}$$

Moreover, $\mathsf{mmse}(Y,s)$ can be bounded from below using the so-called *van Trees inequality* [126] (see also Appendix 3.A):

$$\mathsf{mmse}(Y,s) \geq \frac{1}{J(Y)+s}. \tag{3.2.28}$$

Then

$$\begin{aligned} D(P_Y\|G) &= \frac{1}{2}\int_0^\infty (\mathsf{lmmse}(Y,s) - \mathsf{mmse}(Y,s))\,\mathrm{d}s \\ &\leq \frac{1}{2}\int_0^\infty \left(\frac{1}{1+s} - \frac{1}{J(Y)+s}\right)\mathrm{d}s \\ &= \frac{1}{2}\lim_{\lambda\to\infty}\int_0^\lambda \left(\frac{1}{1+s} - \frac{1}{J(Y)+s}\right)\mathrm{d}s \\ &= \frac{1}{2}\lim_{\lambda\to\infty} \ln\left(\frac{J(Y)\,(1+\lambda)}{J(Y)+\lambda}\right) \\ &= \frac{1}{2}\ln J(Y), \end{aligned} \tag{3.2.29}$$

where the second step uses (3.2.27) and (3.2.28). On the other hand, using (3.2.20) with $s = \mathbb{E}Y^2 = 1$, we get $D(P_Y\|G) = \frac{1}{2}\ln(1/N(Y))$. Combining this with (3.2.29), we recover Stam's inequality $N(Y)J(Y) \geq 1$. Moreover, the van Trees bound (3.2.28) 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 $f\colon \mathbb{R}^n \to \mathbb{R}$ is any Lipschitz function. We start by considering the special case when f is also differentiable.

Proposition 3.5. Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(0,1)$ random variables. Then, for every differentiable function $f\colon \mathbb{R}^n \to \mathbb{R}$ such that $\|\nabla f(X^n)\| \leq 1$ almost surely, we have

$$\mathbb{P}\Big(f(X^n) \geq \mathbb{E}f(X^n) + r\Big) \leq \exp\left(-\frac{r^2}{2}\right), \quad \forall\, r \geq 0 \tag{3.2.30}$$

Proof. Let $P = G^n$ denote the distribution of X^n. If Q is any probability measure such that P and Q are mutually absolutely continuous (i.e., $Q \ll P$ and $P \ll Q$), 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 $t > 0$. Therefore, applying the modified log-Sobolev inequality (3.2.11) to $g = tf$ for some $t > 0$, we get

$$D(P^{(tf)}\|P) \leq \frac{t^2}{2}\mathbb{E}_P^{(tf)}\left[\|\nabla f(X^n)\|^2\right] \leq \frac{t^2}{2}. \tag{3.2.31}$$

Now for the Herbst argument: using Corollary 3.1.1 with $U = f(X^n) - \mathbb{E}f(X^n)$, we get (3.2.30). □

Remark 3.8. Corollary 3.1.1 and inequality (3.2.11) with $g = tf$ imply

that, for any smooth function f with $\|\nabla f(X^n)\|^2 \le L$ a.s.,

$$\mathbb{P}\Big(f(X^n) \ge \mathbb{E}f(X^n) + r\Big) \le \exp\left(-\frac{r^2}{2L}\right), \quad \forall\, r \ge 0. \tag{3.2.32}$$

Thus, the constant κ in the corresponding Gaussian concentration bound (3.1.2) 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 3.2.2. Let X^n be as before, and let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a 1-Lipschitz function, i.e.,

$$|f(x^n) - f(y^n)| \le \|x^n - y^n\|, \qquad \forall x^n, y^n \in \mathbb{R}^n.$$

Then

$$\mathbb{P}\Big(f(X^n) \ge \mathbb{E}f(X^n) + r\Big) \le \exp\left(-\frac{r^2}{2}\right), \quad \forall\, r \ge 0. \tag{3.2.33}$$

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 3.5, and consider the limit of vanishing perturbation.

We construct the perturbation as follows. Let $Z_1, \ldots, Z_n$ be n i.i.d. $\mathcal{N}(0,1)$ random variables, independent of X^n. For any $\delta > 0$, define the function

$$\begin{aligned} f_\delta(x^n) &\triangleq \mathbb{E}\left[f(x^n + \sqrt{\delta}Z^n)\right] \\ &= \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} f(x^n + \sqrt{\delta}z^n) \exp\left(-\frac{\|z^n\|^2}{2}\right) \mathrm{d}z^n \\ &= \frac{1}{(2\pi\delta)^{n/2}} \int_{\mathbb{R}^n} f(z^n) \exp\left(-\frac{\|z^n - x^n\|^2}{2\delta}\right) \mathrm{d}z^n. \end{aligned}$$

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 kernel, see, e.g., [121,

p. 180]). Using Jensen's inequality and the fact that f is 1-Lipschitz,

$$\begin{aligned}|f_\delta(x^n) - f(x^n)| &= \left|\mathbb{E}[f(x^n + \sqrt{\delta}Z^n)] - f(x^n)\right| \\ &\leq \mathbb{E}\left|f(x^n + \sqrt{\delta}Z^n) - f(x^n)\right| \\ &\leq \sqrt{\delta}\,\mathbb{E}\|Z^n\|.\end{aligned}$$

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 [127, Section 3.1.2], and $\|\nabla f\| \leq 1$ almost everywhere. Since $\nabla f_\delta(x^n) = \mathbb{E}\big[\nabla f\big(x^n + \sqrt{\delta}Z^n\big)\big]$, Jensen's inequality gives

$$\|\nabla f_\delta(x^n)\| \leq \mathbb{E}\|\nabla f\big(x^n + \sqrt{\delta}Z^n\big)\| \leq 1$$

for *every* $x^n \in \mathbb{R}^n$. Therefore, we can apply Proposition 3.5 to get, for all $\delta > 0$ and $r > 0$,

$$\mathbb{P}\Big(f_\delta(X^n) \geq \mathbb{E}f_\delta(X^n) + r\Big) \leq \exp\left(-\frac{r^2}{2}\right).$$

Using the fact that f_δ converges to f pointwise as $\delta \to 0$, we obtain (3.2.33):

$$\begin{aligned}\mathbb{P}\Big(f(X^n) \geq \mathbb{E}f(X^n) + r\Big) &= \mathbb{E}\big[1_{\{f(X^n)\geq \mathbb{E}f(X^n)+r\}}\big] \\ &\leq \lim_{\delta\to 0} \mathbb{E}\big[1_{\{f_\delta(X^n)\geq \mathbb{E}f_\delta(X^n)+r\}}\big] \\ &= \lim_{\delta\to 0} \mathbb{P}\big(f_\delta(X^n) \geq \mathbb{E}f_\delta(X^n) + r\big) \\ &\leq \exp\left(-\frac{r^2}{2}\right)\end{aligned}$$

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 [44], 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 [44] 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 [128].

Consider a pair (X, Y) of real-valued random variables that are related through the stochastic transformation

$$Y = e^{-t}X + \sqrt{1 - e^{-2t}}Z \tag{3.2.34}$$

for some $t \geq 0$, 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 (3.2.34) for a given $t \geq 0$ 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 $t \geq 0$ 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 [13, Chapter 4], the channel family $\{\mathrm{OU}(t)\}_{t=0}^{\infty}$ is *ordered by degradation*: for any $t_1, t_2 \geq 0$ we have

$$\mathrm{OU}(t_1 + t_2) = \mathrm{OU}(t_2) \circ \mathrm{OU}(t_1) = \mathrm{OU}(t_1) \circ \mathrm{OU}(t_2), \tag{3.2.35}$$

which is shorthand for the following statement: for any input random variable X, any standard Gaussian Z independent of X, and any $t_1, t_2 \geq 0$, we can always find independent standard Gaussian random variables Z_1, Z_2 that are also independent of X, such that

$$\begin{aligned} & e^{-(t_1+t_2)}X + \sqrt{1 - e^{-2(t_1+t_2)}}Z \\ & \quad \stackrel{\mathrm{d}}{=} e^{-t_2}\left[e^{-t_1}X + \sqrt{1 - e^{-2t_1}}Z_1\right] + \sqrt{1 - e^{-2t_2}}Z_2 \\ & \quad \stackrel{\mathrm{d}}{=} e^{-t_1}\left[e^{-t_2}X + \sqrt{1 - e^{-2t_2}}Z_1\right] + \sqrt{1 - e^{-2t_1}}Z_2 \end{aligned} \tag{3.2.36}$$

where $\stackrel{\mathrm{d}}{=}$ 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,

$$\begin{aligned} Y_1 &\stackrel{\mathrm{d}}{=} e^{-t_1} X + \sqrt{1 - e^{-2t_1}} Z_1 \\ Y_2 &\stackrel{\mathrm{d}}{=} e^{-(t_1+t_2)} X + \sqrt{1 - e^{-2(t_1+t_2)}} Z_2 \end{aligned}$$

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 $Y_0 \stackrel{\mathrm{d}}{=} X$ and $Y_t \stackrel{\mathrm{d}}{=} e^{-t} X + \sqrt{1 - e^{-2t}} \mathcal{N}(0,1)$ for all $t \geq 0$. One way to do this is to let $\{Y_t\}_{t=0}^{\infty}$ be governed by the Itô stochastic differential equation (SDE)

$$\mathrm{d}Y_t = -Y_t \, \mathrm{d}t + \sqrt{2} \, \mathrm{d}B_t, \qquad t \geq 0 \tag{3.2.37}$$

with the initial condition $Y_0 \stackrel{\mathrm{d}}{=} X$, where $\{B_t\}$ denotes the standard one-dimensional Wiener process (Brownian motion). The solution of this SDE (which is known as the *Langevin equation* [129, p. 75]) is given by the so-called *Ornstein–Uhlenbeck process*

$$Y_t = X e^{-t} + \sqrt{2} \int_0^t e^{-(t-s)} \, \mathrm{d}B_s, \qquad t \geq 0$$

where, by the Itô isometry, the variance of the (zero-mean) additive Gaussian noise is indeed

$$\begin{aligned} \mathbb{E}\left[\left(\sqrt{2} \int_0^t e^{-(t-s)} \, \mathrm{d}B_s \right)^2 \right] &= 2 \int_0^t e^{-2(t-s)} \mathrm{d}s \\ &= 2e^{-2t} \int_0^t e^{2s} \mathrm{d}s \\ &= 1 - e^{-2t}, \quad \forall \, t \geq 0 \end{aligned}$$

(see, e.g., [130, p. 358] or [131, p. 127]). This explains our choice of the name "Ornstein–Uhlenbeck channel" for the random transformation (3.2.34).

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

$$D_\alpha(P\|Q) \triangleq \frac{1}{\alpha - 1} \ln \int \mathrm{d}\mu \left(\frac{\mathrm{d}P}{\mathrm{d}\mu}\right)^\alpha \left(\frac{\mathrm{d}Q}{\mathrm{d}\mu}\right)^{1-\alpha}, \tag{3.2.38}$$

where μ is any σ-finite measure that dominates both P and Q. If $P \ll Q$, we have the equivalent form

$$D_\alpha(P\|Q) \frac{1}{\alpha - 1} \ln \mathbb{E}_Q \left[\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)^\alpha\right]. \tag{3.2.39}$$

We recall several key properties of the Rényi divergence (see, for example, [132]):

1. The Kullback-Leibler divergence $D(P\|Q)$ is the limit of $D_\alpha(P\|Q)$ as α tends to 1 from *below*:

   $$D(P\|Q) = \lim_{\alpha \uparrow 1} D_\alpha(P\|Q)$$

   ; in addition,

   $$D(P\|Q) = \sup_{0<\alpha<1} D_\alpha(P\|Q) \leq \inf_{\alpha>1} D_\alpha(P\|Q).$$

   Moreover, if $D(P\|Q) = \infty$ or there exists some $\beta > 1$ such that $D_\beta(P\|Q) < \infty$, then also

   $$D(P\|Q) = \lim_{\alpha \downarrow 1} D_\alpha(P\|Q). \tag{3.2.40}$$

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

   $$D_\alpha(\tilde{P}\|\tilde{Q}) \leq D_\alpha(P\|Q), \qquad \forall \alpha > 0 \tag{3.2.41}$$

   where $\tilde{P}$ or $\tilde{Q}$ is the distribution of the output of T when the input has distribution P or Q, respectively.

4. The Rényi divergence is non-negative for any order $\alpha > 0$.

Now consider the following set-up. Let X be a real-valued random variable with distribution P, such that $P \ll G$. For any $t \geq 0$, let P_t denote the output distribution of the $\mathrm{OU}(t)$ channel with input $X \sim P$. 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 (3.2.41), we have

$$D_\alpha(P_t\|G) \leq D_\alpha(P\|G), \qquad \forall\, t \geq 0,\ \alpha > 0. \tag{3.2.42}$$

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., [133, Section 4.4] or [128]) applied to the continuous-time Markov process governed by the Langevin equation (3.2.37). We will now show, however, that the Gaussian log-Sobolev inequality of Gross (see Theorem 3.2.1) 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

$$D_\alpha(P_t\|G) \leq \varepsilon D_\alpha(P\|G), \qquad \forall t \geq \tau. \tag{3.2.43}$$

Here is the precise result:

Theorem 3.2.3. The Gaussian log-Sobolev inequality of Theorem 3.2.1 is equivalent to the following statement: For any $1 < \beta < \alpha < \infty$

$$D_\alpha(P_t\|G) \leq \left(\frac{\alpha(\beta-1)}{\beta(\alpha-1)}\right) D_\beta(P\|G),\ \forall\, t \geq \frac{1}{2}\ln\left(\frac{\alpha-1}{\beta-1}\right). \tag{3.2.44}$$

Remark 3.9. The original hypercontractivity result of Gross is stated as an inequality relating suitable norms of $g_t = \mathrm{d}P_t/\mathrm{d}G$ and $g = \mathrm{d}P/\mathrm{d}G$; we refer the reader to the original paper [44] or to the lecture notes of Guionnet and Zegarlinski [50] for the traditional treatment of hypercontractivity.

Remark 3.10. To see that Theorem 3.2.3 implies (3.2.43), 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 $\frac{\alpha(\beta-1)}{\beta(\alpha-1)} = \varepsilon$. Hence, Theorem 3.2.3 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 non-decreasing in the parameter α, and $1 < \beta < \alpha$, it follows that $D_\beta(P||G) \le D_\alpha(P||G)$. Therefore, the last inequality implies 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 3.2.3.

Proof. As a reminder, the L^p norm of a real-valued random variable U is defined by $\|U\|_p \triangleq (\mathbb{E}[|U|^p])^{1/p}$ for $p \ge 1$. It will be convenient to work with the following equivalent form of the Rényi divergence in (3.2.39): For any two random variables U and V such that $P_U \ll P_V$, we have

$$D_\alpha(P_U\|P_V) = \frac{\alpha}{\alpha-1}\ln\left\|\frac{\mathrm{d}P_U}{\mathrm{d}P_V}(V)\right\|_\alpha, \qquad \alpha > 1. \tag{3.2.45}$$

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 $\alpha\colon [0,\infty) \to [\beta,\infty)$ by $\alpha(t) = 1 + (\beta-1)e^{2t}$ for some $\beta > 1$. Let $Z \sim G$. Using (3.2.45), it is easy to verify that the desired bound (3.2.44) is equivalent to the statement that the function $F\colon [0,\infty) \to \mathbb{R}$, 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 with respect to the "time" t, and the prime to denote derivatives with respect to the "space" variable z. We start by computing the

derivative of F with respect to t, which gives

$$\begin{aligned}\dot{F}(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\left\{\frac{1}{\alpha(t)}\ln\mathbb{E}\left[\left(g_t(Z)\right)^{\alpha(t)}\right]\right\} \\ &= -\frac{\dot{\alpha}(t)}{\alpha^2(t)}\ln\mathbb{E}\left[\left(g_t(Z)\right)^{\alpha(t)}\right] + \frac{1}{\alpha(t)}\frac{\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\left(g_t(Z)\right)^{\alpha(t)}\right]}{\mathbb{E}\left[\left(g_t(Z)\right)^{\alpha(t)}\right]}. \quad (3.2.46)\end{aligned}$$

To handle the derivative with respect to t in the second term in (3.2.46), 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 (3.2.34).

For any $t \geq 0$, let us define a linear operator K_t acting on any sufficiently regular (e.g., $L^1(G)$) function h as

$$K_t h(x) \triangleq \mathbb{E}\left[h\left(e^{-t}x + \sqrt{1-e^{-2t}}Z\right)\right], \quad (3.2.47)$$

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 (3.2.34), then for any measurable function F such that $\mathbb{E}\big|F(Y)\big| < \infty$ with Y in (3.2.34), we can write

$$K_t F(x) = \mathbb{E}[F(Y)|X = x], \qquad \forall x \in \mathbb{R} \quad (3.2.48)$$

and

$$\mathbb{E}[F(Y)] = \mathbb{E}[K_t F(X)]. \quad (3.2.49)$$

Here, (3.2.48) easily follows from (3.2.34), and (3.2.49) is immediate from (3.2.48).

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

$$g_t(x) = K_t g(x). \quad (3.2.50)$$

To prove (3.2.50), we can either use (3.2.48) and the fact that $g_t(x) = \mathbb{E}[g(Y)|X = x]$, or proceed directly from (3.2.34):

$$\begin{aligned} g_t(x) &= \frac{1}{\sqrt{2\pi(1-e^{-2t})}} \int_{\mathbb{R}} \exp\left(-\frac{(u-e^{-t}x)^2}{2(1-e^{-2t})}\right) g(u)\mathrm{d}u \\ &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} g\left(e^{-t}x + \sqrt{1-e^{-2t}}z\right) \exp\left(-\frac{z^2}{2}\right) \mathrm{d}z \\ &\equiv \mathbb{E}\left[g\left(e^{-t}x + \sqrt{1-e^{-t}}Z\right)\right] \end{aligned} \tag{3.2.51}$$

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 (3.2.48) and (3.2.49) 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 function $F \in L^1(G)$ 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 C^2 functions $h\colon \mathbb{R} \to \mathbb{R}$. Then:

1. The *Ornstein–Uhlenbeck flow* $\{h_t\}_{t=0}^{\infty}$, where $h_t = K_t h$ with a C^2 initial condition $h_0 = h$, satisfies the partial differential equation (PDE)

$$\dot{h}_t = \mathcal{L}h_t. \tag{3.2.52}$$

2. For $Z \sim G$ and all C^2 functions $g, h\colon \mathbb{R} \to \mathbb{R}$ we have the *integration-by-parts formula*

$$\mathbb{E}[g(Z)\mathcal{L}h(Z)] = \mathbb{E}[h(Z)\mathcal{L}g(Z)] = -\mathbb{E}[g'(Z)h'(Z)]. \tag{3.2.53}$$

We provide the details in Appendix 3.B.

We are now ready to tackle the second term in (3.2.46). Noting that the family of densities $\{g_t\}_{t=0}^{\infty}$ forms an Ornstein–Uhlenbeck flow with initial condition $g_0 = g$, we have

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\left(g_t(Z)\right)^{\alpha(t)}\right] \\
&= \mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}t}\left\{\left(g_t(Z)\right)^{\alpha(t)}\right\}\right] \\
&= \dot{\alpha}(t)\,\mathbb{E}\left[\left(g_t(Z)\right)^{\alpha(t)}\ln g_t(Z)\right] + \alpha(t)\,\mathbb{E}\left[\left(g_t(Z)\right)^{\alpha(t)-1}\frac{\mathrm{d}}{\mathrm{d}t}g_t(Z)\right] \\
&= \dot{\alpha}(t)\,\mathbb{E}\left[\left(g_t(Z)\right)^{\alpha(t)}\ln g_t(Z)\right] \\
&\qquad + \alpha(t)\,\mathbb{E}\left[\left(g_t(Z)\right)^{\alpha(t)-1}\mathcal{L}g_t(Z)\right] &(3.2.54)\\
&= \dot{\alpha}(t)\,\mathbb{E}\left[\left(g_t(Z)\right)^{\alpha(t)}\ln g_t(Z)\right] \\
&\qquad - \alpha(t)\,\mathbb{E}\left[\left(\left(g_t(Z)\right)^{\alpha(t)-1}\right)' g_t'(Z)\right] &(3.2.55)\\
&= \dot{\alpha}(t)\,\mathbb{E}\left[\left(g_t(Z)\right)^{\alpha(t)}\ln g_t(Z)\right] \\
&\qquad - \alpha(t)(\alpha(t)-1)\,\mathbb{E}\left[\left(g_t(Z)\right)^{\alpha(t)-2}\left(g_t'(Z)\right)^2\right] &(3.2.56)
\end{aligned}$$

where we use (3.2.52) to get (3.2.54), and (3.2.53) to get (3.2.55). (Referring back to (3.2.51), we see that the functions g_t, for all $t > 0$, are C^∞ due to the smoothing property of the Gaussian kernel, so all interchanges of expectations and derivatives in the above display are justified.) If we define the function $\phi_t(z) \triangleq \left(g_t(z)\right)^{\alpha(t)/2}$, then we can rewrite (3.2.56) as

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\left(g_t(Z)\right)^{\alpha(t)}\right] &= \frac{\dot{\alpha}(t)}{\alpha(t)}\,\mathbb{E}\left[\phi_t^2(Z)\ln\phi_t^2(Z)\right] \\
&\qquad - \frac{4(\alpha(t)-1)}{\alpha(t)}\,\mathbb{E}\left[\left(\phi_t'(Z)\right)^2\right]. &(3.2.57)
\end{aligned}$$

Using the definition of ϕ_t and substituting (3.2.57) into the right-hand side of (3.2.46), we get

$$\alpha^2(t)\,\mathbb{E}[\phi_t^2(Z)]\,\dot{F}(t) = \dot{\alpha}(t)\,\left(\mathbb{E}[\phi_t^2(Z)\ln\phi_t^2(Z)] - \mathbb{E}[\phi_t^2(Z)]\ln\mathbb{E}[\phi_t^2(Z)]\right) - 4(\alpha(t)-1)\mathbb{E}\left[\left(\phi_t'(Z)\right)^2\right]. \tag{3.2.58}$$

If we now apply the Gaussian log-Sobolev inequality (3.2.1) to ϕ_t, then from (3.2.58) we get

$$\alpha^2(t)\,\mathbb{E}[\phi_t^2(Z)]\,\dot{F}(t) \le 2\left(\dot{\alpha}(t) - 2(\alpha(t)-1)\right)\mathbb{E}\left[\left(\phi_t'(Z)\right)^2\right]. \tag{3.2.59}$$

Since $\alpha(t) = 1+(\beta-1)e^{2t}$, $\dot{\alpha}(t)-2(\alpha(t)-1) = 0$, which implies that the right-hand side of (3.2.59) 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 (3.2.51) where g is a probability density function), we conclude that $\dot{F}(t) \le 0$.

What we have proved so far is that, for any $\beta > 1$ and any $t \ge 0$,

$$D_{\alpha(t)}(P_t\|G) \le \left(\frac{\alpha(t)(\beta-1)}{\beta(\alpha(t)-1)}\right) D_\beta(P\|G) \tag{3.2.60}$$

where $\alpha(t) = 1 + (\beta-1)e^{2t}$. By the monotonicity property of the Rényi divergence, the left-hand side of (3.2.60) 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 $u \in (1,\infty) \mapsto \frac{u}{u-1}$ is strictly decreasing, the right-hand side of (3.2.60) can be upper-bounded by $\left(\frac{\alpha(\beta-1)}{\beta(\alpha-1)}\right)D_\beta(P\|G)$ for all $\alpha \ge \alpha(t)$. Putting all these facts together, we conclude that the Gaussian log-Sobolev inequality (3.2.1) implies (3.2.44).

We now show that (3.2.44) implies the log-Sobolev inequality of Theorem 3.2.1. To that end, we recall that (3.2.44) is equivalent to the right-hand side of (3.2.58) 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 (3.2.58) 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 (3.2.1) where $\mathbb{E}[g(Z)] = \mathbb{E}_G\left[\frac{\mathrm{d}P}{\mathrm{d}G}\right] = 1$. □

As a consequence, we can establish a strong version of the data processing inequality for the ordinary divergence:

Corollary 3.2.4. In the notation of Theorem 3.2.3, we have for any $t \geq 0$

$$D(P_t \| G) \leq e^{-2t} D(P \| G). \tag{3.2.61}$$

Proof. Let $\alpha = 1 + \varepsilon e^{2t}$ and $\beta = 1 + \varepsilon$ for some $\varepsilon > 0$. Then using Theorem 3.2.3, we have

$$D_{1+\varepsilon e^{2t}}(P_t \| G) \leq \left(\frac{e^{-2t} + \varepsilon}{1 + \varepsilon} \right) D_{1+\varepsilon}(P \| G), \qquad \forall\, t \geq 0. \tag{3.2.62}$$

Taking the limit of both sides of (3.2.62) as $\varepsilon \downarrow 0$ and using (3.2.40) (note that $D_\alpha(P\|G) < \infty$ for $\alpha > 1$), we get (3.2.61). □

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 [53] 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 $a \geq 0$ 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 $f \in \mathcal{A}$, $a \geq 0$, and $b \in \mathbb{R}$, $\Gamma(af + b) = a\,\Gamma f$.

Then we say that μ satisfies a *logarithmic Sobolev inequality* with constant $c \geq 0$, or LSI(c) for short, if

$$D(\mu^{(f)} \| \mu) \leq \frac{c}{2}\, \mathbb{E}_\mu^{(f)} \left[(\Gamma f)^2 \right], \qquad \forall f \in \mathcal{A}. \tag{3.3.1}$$

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 with respect to $\mu^{(f)}$.

Remark 3.11. We have expressed the log-Sobolev inequality using standard information-theoretic notation. Most of the mathematical 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 $g\colon \Omega \to \mathbb{R}$, define the *entropy functional*

$$\begin{aligned}\mathrm{Ent}_\mu(g) &\triangleq \int g \ln g \,\mathrm{d}\mu - \int g \,\mathrm{d}\mu \cdot \ln\left(\int g \,\mathrm{d}\mu\right) \\ &\equiv \mathbb{E}_\mu[g \ln g] - \mathbb{E}_\mu[g] \,\ln \mathbb{E}_\mu[g]. \end{aligned} \tag{3.3.2}$$

Then, the LSI(c) condition can be equivalently written as (cf. [53, p. 2])

$$\mathrm{Ent}_\mu\big(\exp(f)\big) \le \frac{c}{2}\int (\Gamma f)^2 \exp(f)\,\mathrm{d}\mu \tag{3.3.3}$$

with the convention that $0 \ln 0 \triangleq 0$. To see the equivalence of (3.3.1) and (3.3.3), note that

$$\begin{aligned}\mathrm{Ent}_\mu\big(\exp(f)\big) &= \int \exp(f) \ln\left(\frac{\exp(f)}{\int \exp(f)\mathrm{d}\mu}\right)\mathrm{d}\mu \\ &= \mathbb{E}_\mu[\exp(f)] \int \Big(\frac{\mathrm{d}\mu^{(f)}}{\mathrm{d}\mu}\Big)\ln\Big(\frac{\mathrm{d}\mu^{(f)}}{\mathrm{d}\mu}\Big)\,\mathrm{d}\mu \\ &= \mathbb{E}_\mu[\exp(f)] \cdot D(\mu^{(f)}\|\mu) \end{aligned} \tag{3.3.4}$$

and

$$\begin{aligned}\int (\Gamma f)^2 \exp(f)\,\mathrm{d}\mu &= \mathbb{E}_\mu[\exp(f)] \int (\Gamma f)^2 \,\mathrm{d}\mu^{(f)} \\ &= \mathbb{E}_\mu[\exp(f)] \cdot \mathbb{E}_\mu^{(f)}\left[(\Gamma f)^2\right]. \end{aligned} \tag{3.3.5}$$

Substituting (3.3.4) and (3.3.5) into (3.3.3), we obtain (3.3.1). We note that the entropy functional Ent is homogeneous of degree 1: for any g such that $\mathrm{Ent}_\mu(g) < \infty$ and any $a > 0$, we have

$$\mathrm{Ent}_\mu(ag) = a\,\mathbb{E}_\mu\left[g \ln \frac{g}{\mathbb{E}_\mu[g]}\right] = a\,\mathrm{Ent}_\mu(g).$$

Remark 3.12. Strictly speaking, (3.3.1) should be called a modified (or exponential) logarithmic Sobolev inequality. The ordinary log-Sobolev inequality takes the form

$$\mathrm{Ent}_\mu(g^2) \le 2c \int (\Gamma g)^2 \, \mathrm{d}\mu \tag{3.3.6}$$

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

$$\Gamma(\psi \circ g) = |\psi' \circ g| \, \Gamma g, \qquad \forall g \in \mathcal{A}, \psi \in C^\infty \tag{3.3.7}$$

then (3.3.1) and (3.3.6) are equivalent. Indeed, if (3.3.6) holds, then using it with $g = \exp(f/2)$ gives

$$\begin{aligned} \mathrm{Ent}_\mu(\exp(f)) &\le 2c \int \Big(\Gamma\big(\exp(f/2)\big)\Big)^2 \, \mathrm{d}\mu \\ &= \frac{c}{2} \int (\Gamma f)^2 \exp(f) \, \mathrm{d}\mu \end{aligned}$$

which is (3.3.3). The last equality in the above display follows from (3.3.7) which implies that

$$\Gamma\big(\exp(f/2)\big) = \frac{1}{2} \exp(f/2) \cdot \Gamma f.$$

Conversely, using (3.3.3) with $f = 2 \ln g$ gives

$$\begin{aligned} \mathrm{Ent}_\mu(g^2) &\le \frac{c}{2} \int \big(\Gamma(2 \ln g)\big)^2 g^2 \, \mathrm{d}\mu \\ &= 2c \int (\Gamma g)^2 \mathrm{d}\mu, \end{aligned}$$

which is (3.3.6). Again, the last equality is a consequence of (3.3.7), which gives $\Gamma(2 \ln g) = \frac{2 \Gamma g}{g}$ for all strictly positive $g \in \mathcal{A}$). 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 (3.3.7).

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 (3.3.1) to $g = tf$ we get

$$\begin{aligned} D(P^{(tf)}\|P) &\leq \frac{c}{2}\,\mathbb{E}_P^{(tf)}\left[(\Gamma(tf))^2\right] \\ &= \frac{ct^2}{2}\,\mathbb{E}_P^{(tf)}\left[(\Gamma f)^2\right] \\ &\leq \frac{c\|\Gamma f\|_\infty^2 t^2}{2}, \end{aligned} \tag{3.3.8}$$

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 (3.1.31) for every $g \in \mathcal{A}$ with $E(g) = \|\Gamma g\|_\infty^2$. Therefore, using the bound (3.3.8) together with Corollary 3.1.1, we arrive at

$$\mathbb{P}(f(X^n) \geq \mathbb{E}f(X^n) + r) \leq \exp\left(-\frac{r^2}{2c\|\Gamma f\|_\infty^2}\right), \qquad \forall r \geq 0. \tag{3.3.9}$$

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 $f\colon \mathcal{X}^n \to \mathbb{R}$, 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 $f_i(\cdot|\bar{x}^i)\colon \mathcal{X} \to \mathbb{R}$ via $f_i(x_i|\bar{x}^i) \triangleq f(x^n)$. Then, we have the following:

Theorem 3.3.1. 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 $f\colon \mathcal{X}^n \to \mathbb{R}$ such that, for every $i \in \{1, \ldots, n\}$,

$$f_i(\cdot|\bar{x}^i) \in \mathcal{A}_i, \qquad \forall\, \bar{x}^i \in \mathcal{X}^{n-1}. \tag{3.3.10}$$

Define the operator Γ that maps each $f \in \mathcal{A}$ to

$$\Gamma f = \sqrt{\sum_{i=1}^{n} (\Gamma_i f_i)^2}, \tag{3.3.11}$$

which is shorthand for

$$\Gamma f(x^n) = \sqrt{\sum_{i=1}^{n} \Big(\Gamma_i f_i(x_i|\bar{x}^i)\Big)^2}, \qquad \forall\, x^n \in \mathcal{X}^n. \tag{3.3.12}$$

Then, the following statements hold:

1. If there exists a constant $c \geq 0$ 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 $r \geq 0$,

$$\mathbb{P}(f(X^n) \geq r) \leq \exp\left(-\frac{r^2}{2c\|\Gamma f\|_\infty^2}\right). \tag{3.3.13}$$

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 $a \geq 0$ and $b \in \mathbb{R}$, and let $g = af + b$. Then, for any i and any $\bar{x}^i$,

$$\begin{aligned} g_i(\cdot|\bar{x}^i) &= g(x_1, \ldots, x_{i-1}, \cdot, x_{i+1}, \ldots, x_n) \\ &= af(x_1, \ldots, x_{i-1}, \cdot, x_{i+1}, \ldots, x_n) + b \\ &= af_i(\cdot|\bar{x}^i) + b \in \mathcal{A}_i, \end{aligned}$$

where the last step uses (3.3.10). Hence, $f \in \mathcal{A}$ implies that $g = af + b \in \mathcal{A}$ for any $a \geq 0, b \in \mathbb{R}$, so (LSI-1) holds. From the definitions of Γ in

(3.3.11) and (3.3.12) it is readily seen that (LSI-2) and (LSI-3) hold as well.

Next, for any $f \in \mathcal{A}$ and any $t \geq 0$, we have

$$\begin{aligned} D(P^{(tf)}\|P) &\leq \sum_{i=1}^n D\Big(P^{(tf)}_{X_i|\bar{X}^i}\Big\|P_{X_i}\Big|P^{(tf)}_{\bar{X}^i}\Big) \\ &= \sum_{i=1}^n \int P^{(tf)}_{\bar{X}^i}(\mathrm{d}\bar{x}^i) D\Big(P^{(tf)}_{X_i|\bar{X}^i=\bar{x}^i}\Big\|P_{X_i}\Big) \\ &= \sum_{i=1}^n \int P^{(tf)}_{\bar{X}^i}(\mathrm{d}\bar{x}^i) D\Big(P^{(tf_i(\cdot|\bar{x}^i))}_{X_i}\Big\|P_{X_i}\Big) \\ &\leq \frac{ct^2}{2}\sum_{i=1}^n \int P^{(tf)}_{\bar{X}^i}(\mathrm{d}\bar{x}^i)\, \mathbb{E}^{(tf_i(\cdot|\bar{x}^i))}_{X_i}\Big[\Big(\Gamma_i f_i(X_i|\bar{x}^i)\Big)^2\Big] \\ &= \frac{ct^2}{2}\sum_{i=1}^n \mathbb{E}^{(tf)}_{P_{\bar{X}^i}}\Big\{\mathbb{E}^{(tf)}_{P_{X_i}}\Big[\Big(\Gamma_i f_i(X_i|\bar{X}^i)\Big)^2\Big|\bar{X}^i\Big]\Big\} \\ &= \frac{ct^2}{2}\cdot \mathbb{E}^{(tf)}_P\Big[(\Gamma f)^2\Big], \end{aligned} \tag{3.3.14}$$

where the first step uses Proposition 3.2 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 (3.1.24), 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) with respect to $(\mathcal{A}_i, \Gamma_i)$, and the last step uses the tower property of the conditional expectation, as well as (3.3.11). We have thus proved the first part of the theorem, i.e., that P satisfies LSI(c) with respect to the pair $(\mathcal{A}, \Gamma)$. The second part follows from the same argument that was used to prove (3.3.9). □

3.3.2 Maurer's thermodynamic method

With Theorem 3.3.1 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 [134], it is possible to derive log-Sobolev inequal-

ities 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 with respect to 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 $h\colon \Omega \to \mathbb{R}$ with respect to $\mu^{(g)}$ 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 [134, Theorem 3]):

Lemma 3.3.2. Consider a function $f\colon \Omega \to \mathbb{R}$, such that $\mathbb{E}_\mu[\exp(\lambda f)] < \infty$ for all $\lambda > 0$. Then

$$D(\mu^{(\lambda f)}\|\mu) = \int_0^\lambda \int_t^\lambda \mathsf{var}_\mu^{(sf)}[f]\,\mathrm{d}s\,\mathrm{d}t. \tag{3.3.15}$$

Remark 3.13. 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*

$$\Psi_\theta(\nu) \triangleq \mathbb{E}_\nu[f] + \theta D(\nu\|\mu). \tag{3.3.16}$$

Then we claim 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

$$\begin{aligned}
\Psi_\theta(\nu) &= \frac{1}{t}\, \mathbb{E}_\nu[tf + \ln\psi] \\
&= \frac{1}{t}\, \mathbb{E}_\nu\left[\ln\big(\psi \exp(tf)\big)\right] \\
&= \frac{1}{t}\, \mathbb{E}_\nu\left[\ln \frac{\mathrm{d}\nu}{\mathrm{d}\mu^{(-tf)}} - \Lambda(-t)\right] \\
&= \frac{1}{t}\left[D(\nu\|\mu^{(-tf)}) - \Lambda(-t)\right],
\end{aligned}$$

where, as before, $\Lambda(-t) \triangleq \ln\big(\mathbb{E}_\mu[\exp(-tf)]\big)$ is the logarithmic moment generating function of f with respect to μ. Therefore, $\Psi_\theta(\nu) = \Psi_{1/t}(\nu) \geq -\Lambda(-t)/t$, with equality if and only if $\nu = \mu^{(-tf)}$.

We refer the reader to a recent monograph by Merhav [135] that highlights some interesting relations between information theory and statistical physics. This monograph relates thermodynamic potentials (like the thermodynamical entropy and free energy) to information measures (like the Shannon entropy and information divergence); it also provides some rigorous mathematical tools that were inspired by the physical point of view and which proved to be useful in dealing with information-theoretic problems.

Now we give the proof of Lemma 3.3.2:

Proof. We start by noting that (see (3.1.10))

$$\Lambda'(t) = \mathbb{E}_\mu^{(tf)}[f] \qquad \text{and} \qquad \Lambda''(t) = \mathsf{var}_\mu^{(tf)}[f], \tag{3.3.17}$$

and, in particular, $\Lambda'(0) = \mathbb{E}_\mu[f]$. Moreover, from (3.1.12), we get

$$D(\mu^{(\lambda f)}\|\mu) = \lambda^2 \, \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{\Lambda(\lambda)}{\lambda}\right) = \lambda\Lambda'(\lambda) - \Lambda(\lambda). \tag{3.3.18}$$

Now, using (3.3.17), we get

$$\begin{aligned}
\lambda\Lambda'(\lambda) &= \int_0^\lambda \Lambda'(\lambda)\,\mathrm{d}t \\
&= \int_0^\lambda \left(\int_0^\lambda \Lambda''(s)\,\mathrm{d}s + \Lambda'(0)\right)\mathrm{d}t \\
&= \int_0^\lambda \left(\int_0^\lambda \mathsf{var}_\mu^{(sf)}[f]\,\mathrm{d}s + \mathbb{E}_\mu[f]\right)\mathrm{d}t
\end{aligned} \tag{3.3.19}$$

and

$$\begin{aligned}
\Lambda(\lambda) &= \int_0^\lambda \Lambda'(t)\,\mathrm{d}t \\
&= \int_0^\lambda \left(\int_0^t \Lambda''(s)\,\mathrm{d}s + \Lambda'(0)\right)\mathrm{d}t \\
&= \int_0^\lambda \left(\int_0^t \mathsf{var}_\mu^{(sf)}[f]\,\mathrm{d}s + \mathbb{E}_\mu[f]\right)\mathrm{d}t.
\end{aligned} \tag{3.3.20}$$

Substituting (3.3.19) and (3.3.20) into (3.3.18), we get (3.3.15). □

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 3.3.3. Let the objects $(\mathcal{A}, \Gamma)$ and $\{(\mathcal{A}_i, \Gamma_i)\}_{i=1}^n$ be constructed as in the statement of Theorem 3.3.1. Furthermore, suppose that for each $i \in \{1, \ldots, n\}$, 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

$$\mathsf{var}_i^{(sg)}[g(X_i)|\bar{X}^i = \bar{x}^i] \le c\,(\Gamma_i g)^2\,, \qquad \forall \bar{x}^i \in \mathcal{X}^{n-1} \tag{3.3.21}$$

holds for all $i \in \{1, \ldots, n\}$, $s > 0$, and $g \in \mathcal{A}_i$, where $\mathsf{var}_i^{(g)}[\cdot|\bar{X}^i = \bar{x}^i]$ denotes the (conditional) variance with respect to $P^{(g)}_{X_i|\bar{X}^i=\bar{x}^i}$. Then, the pair $(\mathcal{A}, \Gamma)$ satisfies LSI(c) with respect to P_{X^n}.

Proof. Consider an arbitrary function $f \in \mathcal{A}$. Then, by construction, $f_i \colon \mathcal{X}_i \to \mathbb{R}$ is in $\mathcal{A}_i$ for each $i \in \{1, \ldots, n\}$. We can write

$$\begin{aligned} D\Big(P^{(f)}_{X_i|\bar{X}^i=\bar{x}^i}\Big\|P_{X_i}\Big) &= D\Big(P_{X_i}^{(f_i(\cdot|\bar{x}^i))}\Big\|P_{X_i}\Big) \\ &= \int_0^1 \int_t^1 \mathsf{var}_i^{(sf_i(\cdot|\bar{x}^i))}[f_i(X_i|\bar{X}^i)|\bar{X}^i = \bar{x}^i]\,\mathrm{d}s\,\mathrm{d}t \\ &\le c\,(\Gamma_i f_i)^2 \int_0^1 \int_t^1 \mathrm{d}s\,\mathrm{d}t \\ &= \frac{c(\Gamma_i f_i)^2}{2} \end{aligned}$$

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 3.3.2, and the third step uses (3.3.21) 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 3.3.1. □

The following two lemmas from [134] will be useful for establishing bounds like (3.3.21):

Lemma 3.3.4. Let U be a random variable such that $U \in [a, b]$ a.s. for some $-\infty < a \le b < +\infty$. Then

$$\mathsf{var}[U] \le \frac{(b-a)^2}{4}. \tag{3.3.22}$$

Proof. Since the support of U is the interval $[a, b]$, the maximal variance of U is attained when the random variable U is binary and equiprobable on the endpoints of this interval. The bound in (3.3.22) is achieved with equality in this case. □

Lemma 3.3.5. Let f be a real-valued function such that $f - \mathbb{E}_\mu[f] \leq C$ for some $C \in \mathbb{R}$. Then, for every $t > 0$,

$$\mathsf{var}_\mu^{(tf)}[f] \leq \exp(tC)\, \mathsf{var}_\mu[f].$$

Proof.

$$\begin{aligned}
\mathsf{var}_\mu^{(tf)}[f] &= \mathsf{var}_\mu^{(tf)}\Big\{f - \mathbb{E}_\mu\left[f\right]\Big\} && (3.3.23)\\
&\leq \mathbb{E}_\mu^{(tf)}\left[\left(f - \mathbb{E}_\mu[f]\right)^2\right] && (3.3.24)\\
&= \mathbb{E}_\mu\left[\frac{\exp(tf)\left(f - \mathbb{E}_\mu[f]\right)^2}{\mathbb{E}_\mu[\exp(tf)]}\right] && (3.3.25)\\
&\leq \mathbb{E}_\mu\Big\{\left(f - \mathbb{E}_\mu[f]\right)^2 \exp\left[t\left(f - \mathbb{E}_\mu[f]\right)\right]\Big\} && (3.3.26)\\
&\leq \exp(tC)\, \mathbb{E}_\mu\left[\left(f - \mathbb{E}_\mu\left[f\right]\right)^2\right], && (3.3.27)
\end{aligned}$$

where:

- (3.3.23) holds since $\mathsf{var}[f] = \mathsf{var}[f + c]$ for any constant $c \in \mathbb{R}$;
- (3.3.24) uses the bound $\mathsf{var}[U] \leq \mathbb{E}[U^2]$;
- (3.3.25) is by definition of the tilted distribution $\mu^{(tf)}$;
- (3.3.26) follows from applying Jensen's inequality to the denominator; and
- (3.3.27) uses the assumption that $f - \mathbb{E}_\mu[f] \leq C$ and the monotonicity of $\exp(\cdot)$ (note that $t > 0$).

This completes the proof of Lemma 3.3.5. □

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 $f\colon \mathcal{X}^n \to \mathbb{R}$ define

$$\Gamma f(x^n) \triangleq \sqrt{\sum_{i=1}^n \big(f(x^n \oplus e_i) - f(x^n)\big)^2}, \qquad \forall x^n \in \mathcal{X}^n, \qquad (3.3.28)$$

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 3.3.6 (Discrete log-Sobolev inequality for the symmetric Bernoulli measure)**.** Let $\mathcal{A}$ be the set of all functions $f\colon \mathcal{X}^n \to \mathbb{R}$. Then, the pair $(\mathcal{A}, \Gamma)$ with Γ defined in (3.3.28) 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) with respect to $(\mathcal{A}, \Gamma)$. In other words, for any $f : \mathcal{X}^n \to \mathbb{R}$,

$$D(P^{(f)} \| P) \le \frac{1}{8}\, \mathbb{E}_P^{(f)} \left[(\Gamma f)^2 \right]. \tag{3.3.29}$$

Proof. Let $\mathcal{A}_0$ be the set of all functions $g\colon \{0,1\} \to \mathbb{R}$, and let Γ_0 be the operator that maps every $g \in \mathcal{A}_0$ to

$$\Gamma g \triangleq |g(0) - g(1)| = |g(x) - g(x \oplus 1)|, \quad \forall\, x \in \{0,1\}. \tag{3.3.30}$$

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 3.3.4, we have

$$\mathsf{var}_{P_i}^{(sg)}[g(X_i)|\bar{X}^i = \bar{x}^i] \le \frac{\big(g(0) - g(1)\big)^2}{4} = \frac{(\Gamma_i g)^2}{4} \tag{3.3.31}$$

for all $g \in \mathcal{A}_i$, $\bar{x}^i \in \mathcal{X}^{n-1}$. In other words, the condition (3.3.21) of Theorem 3.3.3 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 (3.3.11) is precisely the one in (3.3.28). Therefore, by Theorem 3.3.3, the pair $(\mathcal{A}, \Gamma)$ satisfies LSI(1/4) with respect to P, which proves (3.3.29). This completes the proof of Theorem 3.3.6. □

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 [51, Corollary 5.9]:

Theorem 3.3.7. Consider any function $f\colon \{0,1\}^n \to \mathbb{R}$ with the property that there exists some $c > 0$ such that

$$\max_{i\in\{1,\ldots,n\}} |f(x^n \oplus e_i) - f(x^n)| \leq c \tag{3.3.32}$$

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

$$D(P^{(f)}\|P) \leq pq\left(\frac{(c-1)\exp(c)+1}{c^2}\right)\mathbb{E}_P^{(f)}\left[(\Gamma f)^2\right], \tag{3.3.33}$$

where $q \triangleq 1-p$.

Proof. Following the usual route, we will establish the $n = 1$ case first, and then scale up to an arbitrary n by tensorization. In order to capture the correct dependence on the Bernoulli parameter p, we will use a more refined, distribution-dependent variance bound of Lemma 3.3.5, as opposed to the cruder bound of Lemma 3.3.4 that does not depend on the underlying distribution. Maurer's paper [134] has other examples.

Let $a = |\Gamma(f)| = |f(0) - f(1)|$, where Γ is defined as in (3.3.30). Without loss of generality, we may assume that $f(0) = 0$ and $f(1) = a$. Then

$$\mathbb{E}[f] = pa \qquad \text{and} \qquad \mathsf{var}[f] = pqa^2. \tag{3.3.34}$$

Using (3.3.34) and Lemma 3.3.5, since $f - \mathbb{E}[f] \leq a - pa = qa$, it follows that for every $t > 0$

$$\mathsf{var}_P^{(tf)}[f] \leq pqa^2\exp(tqa).$$

Therefore, by Lemma 3.3.2 we have

$$\begin{aligned} D(P^{(f)}\|P) &\leq pqa^2\int_0^1\int_t^1 \exp(sqa)\,\mathrm{d}s\,\mathrm{d}t \\ &= pqa^2\left(\frac{(qa-1)\exp(qa)+1}{(qa)^2}\right) \\ &\leq pqa^2\left(\frac{(c-1)\exp(c)+1}{c^2}\right), \end{aligned}$$

where the last step follows from the fact that the function $u \mapsto u^{-2}[(u-1)\exp(u)+1]$ is nondecreasing in $u \geq 0$, and $0 \leq qa \leq a \leq c$. Since

$a^2 = (\Gamma f)^2$, we can write

$$D(P^{(f)}\|P) \le pq\left(\frac{(c-1)\exp(c)+1}{c^2}\right)\mathbb{E}_P^{(f)}\left[(\Gamma f)^2\right],$$

so we have established (3.3.33) for $n = 1$.

Now consider an arbitrary $n \in \mathbb{N}$. Since the condition in (3.3.32) can be expressed as

$$\left|f_i(0|\bar{x}^i) - f_i(1|\bar{x}^i)\right| \le c, \qquad \forall\, i \in \{1,\ldots,n\},\ \bar{x}^i \in \{0,1\}^{n-1},$$

we can use (3.3.33) to write

$$D\Big(P_{X_i}^{(f_i(\cdot|\bar{x}^i))}\Big\|P_{X_i}\Big) \le pq\left(\frac{(c-1)\exp c+1}{c^2}\right)\mathbb{E}_P^{(f_i(\cdot|\bar{x}^i))}\Big[\left(\Gamma_i f_i(X_i|\bar{X}^i)\right)^2\Big|\bar{X}^i = \bar{x}^i\Big]$$

for every $i = 1,\ldots,n$ and all $\bar{x}^i \in \{0,1\}^{n-1}$. With this, the same sequence of steps that led to (3.3.14) in the proof of Theorem 3.3.1 can be used to complete the proof of (3.3.33) for arbitrary n. □

In Appendix 3.C, we comment on the relations between the log-Sobolev inequalities for the Bernoulli and the Gaussian measures.

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 (recall that the original proof in [6, 39] used the martingale method; the reader is referred to the derivation of McDiarmid's inequality via the martingale approach in Theorem 2.2.3 of the preceding chapter). Following the exposition in [134, Section 4.1], we have the following re-statement of McDiarmid's inequality in Theorem 2.2.3:

Theorem 3.3.8. Let $X_1,\ldots,X_n \in \mathcal{X}$ be independent random variables. Consider a function $f\colon \mathcal{X}^n \to \mathbb{R}$ with $\mathbb{E}[f(X^n)] = 0$, and also suppose that there exist some constants $0 \le c_1,\ldots,c_n < +\infty$ such that, for each $i \in \{1,\ldots,n\}$,

$$\left|f_i(x|\bar{x}^i) - f_i(y|\bar{x}^i)\right| \le c_i, \qquad \forall x,y \in \mathcal{X},\ \bar{x}^i \in \mathcal{X}^{n-1}. \tag{3.3.35}$$

Then, for any $r \geq 0$,

$$\mathbb{P}\Big(f(X^n) \geq r\Big) \leq \exp\left(-\frac{2r^2}{\sum_{i=1}^n c_i^2}\right). \tag{3.3.36}$$

Proof. Let $\mathcal{A}_0$ be the set of all bounded measurable functions $g\colon \mathcal{X} \to \mathbb{R}$, 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 $a \geq 0$ and $b \in \mathbb{R}$. Now, 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 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 3.3.4 gives

$$\mathsf{var}_i^{(sg)}[g(X_i)|\bar{X}^i = \bar{x}^i] \leq \frac{(\Gamma_i g)^2}{4}, \qquad \forall\, g \in \mathcal{A}_i,\ \bar{x}^i \in \mathcal{X}^{n-1}.$$

Hence, the condition (3.3.21) of Theorem 3.3.3 holds with $c = 1/4$. Now let $\mathcal{A}$ be the set of all bounded measurable functions $f\colon \mathcal{X}^n \to \mathbb{R}$. Then, for any $f \in \mathcal{A}$, $i \in \{1,\ldots,n\}$ and $x^n \in \mathcal{X}^n$, we have

$$\begin{aligned}
&\sup_{x_i\in\mathcal{X}_i} f(x_1,\ldots,x_i,\ldots,x_n) - \inf_{x_i\in\mathcal{X}_i} f(x_1,\ldots,x_i,\ldots,x_n)\\
&= \sup_{x_i\in\mathcal{X}_i} f_i(x_i|\bar{x}^i) - \inf_{x_i\in\mathcal{X}_i} f_i(x_i|\bar{x}^i)\\
&= \Gamma_i f_i(\cdot|\bar{x}^i).
\end{aligned}$$

Thus, if we construct an operator Γ on $\mathcal{A}$ from $\Gamma_1,\ldots,\Gamma_n$ according to (3.3.11), the pair $(\mathcal{A},\Gamma)$ will satisfy the conditions of Theorem 3.3.1. Therefore, by Theorem 3.3.3, it follows that the pair $(\mathcal{A},\Gamma)$ satisfies LSI(1/4) for *any* product probability measure on $\mathcal{X}^n$. Hence, inequality (3.3.9) implies that

$$\mathbb{P}\Big(f(X^n) \geq r\Big) \leq \exp\left(-\frac{2r^2}{\|\Gamma f\|_\infty^2}\right) \tag{3.3.37}$$

holds for any $r \geq 0$ and bounded f with $\mathbb{E}[f] = 0$. Now, if f satisfies

(3.3.35), then

$$\begin{aligned}\|\Gamma f\|_\infty^2 &= \sup_{x^n\in\mathcal{X}^n}\sum_{i=1}^n \big(\Gamma_i f_i(x_i|\bar{x}^i)\big)^2\\ &\le \sum_{i=1}^n \sup_{x^n\in\mathcal{X}^n}\big(\Gamma_i f_i(x_i|\bar{x}^i)\big)^2\\ &= \sum_{i=1}^n \sup_{x^n\in\mathcal{X}^n,\,y\in\mathcal{X}} |f_i(x_i|\bar{x}^i) - f_i(y|\bar{x}^i)|^2\\ &\le \sum_{i=1}^n c_i^2.\end{aligned}$$

Substituting this bound into the right-hand side of (3.3.37) gives (3.3.36). □

Note that Maurer's method gives the correct constant in the exponent, in contrast to an earlier approach by Boucheron, Lugosi and Massart [136] who also used the entropy method. For the reader's benefit, we give a more detailed discussion of this in Appendix 3.D.

3.3.5 Log-Sobolev inequalities for Poisson and compound Poisson measures

Let P_λ denote, for some $\lambda > 0$, the Poisson(λ) measure, i.e., $\mathsf{P}_\lambda(n) \triangleq \frac{e^{-\lambda}\lambda^n}{n!}$ for every $n \in \mathbb{N}_0$, where $\mathbb{N}_0 \triangleq \mathbb{N}\cup\{0\}$ is the set of the nonnegative integers. Bobkov and Ledoux [54] have established the following log-Sobolev inequality: for any function $f\colon \mathbb{N}_0 \to \mathbb{R}$,

$$D\Big(\mathsf{P}_\lambda^{(f)}\Big\|\mathsf{P}_\lambda\Big) \le \lambda\,\mathbb{E}_{\mathsf{P}_\lambda}^{(f)}\Big[(\Gamma f)\,e^{\Gamma f} - e^{\Gamma f} + 1\Big], \tag{3.3.38}$$

where Γ is the modulus of the discrete gradient:

$$\Gamma f(x) \triangleq |f(x) - f(x+1)|, \qquad \forall x \in \mathbb{N}_0. \tag{3.3.39}$$

(The inequality (3.3.38) can be obtained by combining the log-Sobolev inequality in [54, Corollary 7] with equality (3.3.4).) Using tensorization of (3.3.38), Kontoyiannis and Madiman [137] gave a simple proof of a log-Sobolev inequality for the *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

$$Z = \sum_{i=1}^{N} X_i, \tag{3.3.40}$$

where $N \sim \mathsf{P}_\lambda$ and $X_1, X_2, \ldots$ are i.i.d. random variables with distribution μ, independent of N (if N takes the value zero, then Z is defined to be zero).

Theorem 3.3.9 (Log-Sobolev inequality for compound Poisson measures [137])**.** For an arbitrary probability measure μ on $\mathbb{N}$ and an arbitrary bounded function $f\colon \mathbb{N}_0 \to \mathbb{R}$, and for every $\lambda > 0$,

$$D\Big(\mathsf{CP}_{\lambda,\mu}^{(f)}\Big\|\mathsf{CP}_{\lambda,\mu}\Big) \leq \lambda \sum_{k=1}^{\infty} \mu(k)\, \mathbb{E}_{\mathsf{CP}_{\lambda,\mu}}^{(f)}\left[(\Gamma_k f)\, e^{\Gamma_k f} - e^{\Gamma_k f} + 1\right], \tag{3.3.41}$$

where $\Gamma_k f(x) \triangleq |f(x) - f(x+k)|$ for each $k, x \in \mathbb{N}_0$.

Proof. The proof relies on the following alternative representation of the $\mathsf{CP}_{\lambda,\mu}$ probability measure:

Lemma 3.3.10. If $Z \sim \mathsf{CP}_{\lambda,\mu}$, then

$$Z \stackrel{\mathrm{d}}{=} \sum_{k=1}^{\infty} k Y_k, \qquad Y_k \sim \mathsf{P}_{\lambda\mu(k)},\ \forall\, k \in \mathbb{N} \tag{3.3.42}$$

where $\{Y_k\}_{k=1}^{\infty}$ are independent random variables, and $\stackrel{\mathrm{d}}{=}$ means equality in distribution.

Proof. The characteristic function of Z in (3.3.42) is equal to

$$\varphi_Z(\nu) \triangleq \mathbb{E}[\exp(j\nu Z)] = \exp\left\{\lambda\left(\sum_{k=1}^{\infty} \mu(k)\exp(j\nu k) - 1\right)\right\}, \quad \forall\, \nu \in \mathbb{R}$$

which coincides with the characteristic function of $Z \sim \mathsf{CP}_{\lambda,\mu}$ in (3.3.40). The statement of the lemma follows from the fact that two random variables are equal in distribution if and only if their characteristic functions coincide. □

For each $n \in \mathbb{N}$, let P_n denote the product distribution of $Y_1, \ldots, Y_n$. Consider an arbitrary function $f : \mathbb{N}_0 \to \mathbb{R}$, and define the function $g\colon (\mathbb{N}_0)^n \to \mathbb{R}$ by

$$g(y_1, \ldots, y_n) \triangleq f\left(\sum_{k=1}^{n} k y_k\right), \qquad \forall y_1, \ldots, y_n \in \mathbb{N}_0.$$

If we now denote by $\bar{P}_n$ the distribution of the sum $S_n \triangleq \sum_{k=1}^{n} k Y_k$, then

$$\begin{aligned} D\Big(\bar{P}_n^{(f)} \| \bar{P}_n\Big) &= \mathbb{E}_{\bar{P}_n}\left[\left(\frac{\exp\big(f(S_n)\big)}{\mathbb{E}_{\bar{P}_n}[\exp\big(f(S_n)\big)]}\right) \ln\left(\frac{\exp\big(f(S_n)\big)}{\mathbb{E}_{\bar{P}_n}[\exp\big(f(S_n)\big)]}\right)\right] \\ &= \mathbb{E}_{P_n}\left[\left(\frac{\exp\big(g(Y^n)\big)}{\mathbb{E}_{P_n}[\exp\big(g(Y^n)\big)]}\right) \ln\left(\frac{\exp\big(g(Y^n)\big)}{\mathbb{E}_{P_n}[\exp\big(g(Y^n)\big)]}\right)\right] \\ &= D\big(P_n^{(g)} \big\| P_n\big) \\ &\le \sum_{k=1}^{n} D\Big(P^{(g)}_{Y_k|\bar{Y}^k} \Big\| P_{Y_k} \Big| P^{(g)}_{\bar{Y}^k}\Big), \end{aligned} \tag{3.3.43}$$

where the last line uses Proposition 3.2 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 (3.3.38) to P_{Y_k} and all functions of the form $g_k(\cdot|\bar{y}^k)$, we can write

$$\begin{aligned} &D\Big(P^{(g)}_{Y_k|\bar{Y}^k} \| P_{Y_k} \mid P^{(g)}_{\bar{Y}^k}\Big) \\ &\quad \le \lambda\mu(k)\, \mathbb{E}^{(g)}_{P_n}\Big[\big(\Gamma g_k(Y_k|\bar{Y}^k)\big)\, e^{\Gamma g_k(Y_k|\bar{Y}^k)} - e^{\Gamma g_k(Y_k|\bar{Y}^k)} + 1\Big] \end{aligned} \tag{3.3.44}$$

where Γ is the absolute value of the "one-dimensional" discrete gradient

in (3.3.39). For any $y^n \in (\mathbb{N}_0)^n$, we have

$$\begin{aligned}
\Gamma g_k(y_k|\bar{y}^k) &= \left|g_k(y_k|\bar{y}^k) - g_k(y_k+1|\bar{y}^k)\right| \\
&= \left| f\left(ky_k + \sum_{j\in\{1,\ldots,n\}\setminus\{k\}} jy_j\right)\right. \\
&\qquad\qquad \left. - f\left(k(y_k+1) + \sum_{j\in\{1,\ldots,n\}\setminus\{k\}} jy_j\right)\right| \\
&= \left| f\left(\sum_{j=1}^n jy_j\right) - f\left(\sum_{j=1}^n jy_j + k\right)\right| \\
&= \Gamma_k f\left(\sum_{j=1}^n jy_j\right).
\end{aligned}$$

Using this in (3.3.44) and performing the reverse change of measure from P_n to $\bar{P}_n$, we can write

$$\begin{aligned}
&D\Big(P^{(g)}_{Y_k|\bar{Y}^k}\Big\|P_{Y_k}\,\Big|\,P^{(g)}_{\bar{Y}^k}\Big) \\
&\qquad \le \lambda\mu(k)\,\mathbb{E}^{(f)}_{\bar{P}_n}\Big[(\Gamma_k f(S_n))e^{\Gamma_k f(S_n)} - e^{\Gamma_k f(S_n)} + 1\Big]. \qquad (3.3.45)
\end{aligned}$$

Therefore, the combination of (3.3.43) and (3.3.45) gives

$$\begin{aligned}
D(\bar{P}^{(f)}_n\|\bar{P}_n) &\le \lambda\sum_{k=1}^n \mu(k)\,\mathbb{E}^{(f)}_{\bar{P}_n}\Big[(\Gamma_k f)\,e^{\Gamma_k f} - e^{\Gamma_k f} + 1\Big] \\
&\le \lambda\sum_{k=1}^\infty \mu(k)\,\mathbb{E}^{(f)}_{\bar{P}_n}\Big[(\Gamma_k f)\,e^{\Gamma_k f} - e^{\Gamma_k f} + 1\Big] \qquad (3.3.46)
\end{aligned}$$

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 (3.3.46). For the left-hand side, we use the fact that, by (3.3.42), $\bar{P}_n$ converges in distribution to $\mathsf{CP}_{\lambda,\mu}$ as $n \to \infty$. Since f is bounded, $\bar{P}^{(f)}_n \to \mathsf{CP}^{(f)}_{\lambda,\mu}$ in distribution. Therefore, by the bounded convergence theorem, we have

$$\lim_{n\to\infty} D(\bar{P}^{(f)}_n\|\bar{P}_n) = D\Big(\mathsf{CP}^{(f)}_{\lambda,\mu}\,\Big\|\,\mathsf{CP}_{\lambda,\mu}\Big). \qquad (3.3.47)$$

For the right-hand side, we have

$$\begin{aligned}
&\sum_{k=1}^{\infty} \mu(k)\, \mathbb{E}^{(f)}_{\bar{P}_n} \left[(\Gamma_k f)\, e^{\Gamma_k f} - e^{\Gamma_k f} + 1 \right] \\
&\quad = \mathbb{E}^{(f)}_{\bar{P}_n} \left\{ \sum_{k=1}^{\infty} \mu(k) \left[(\Gamma_k f)\, e^{\Gamma_k f} - e^{\Gamma_k f} + 1 \right] \right\} \\
&\quad \xrightarrow{n\to\infty} \mathbb{E}^{(f)}_{\mathsf{CP}_{\lambda,\mu}} \left[\sum_{k=1}^{\infty} \mu(k) \left((\Gamma_k f)\, e^{\Gamma_k f} - e^{\Gamma_k f} + 1 \right) \right] \\
&\quad = \sum_{k=1}^{\infty} \mu(k)\, \mathbb{E}^{(f)}_{\mathsf{CP}_{\lambda,\mu}} \left[(\Gamma_k f)\, e^{\Gamma_k f} - e^{\Gamma_k f} + 1 \right] \qquad (3.3.48)
\end{aligned}$$

where the first and the last steps follow from Fubini's theorem, and the second step follows from the bounded convergence theorem. Putting (3.3.46)–(3.3.48) together, we get the inequality in (3.3.41). This completes the proof of Theorem 3.3.9. □

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 $\mathbb{P}\big(f(X^n) \geq \mathbb{E}f(X^n) + r\big)$ for $r \geq 0$. It turns out that the reverse is also true: assuming that f has Gaussian-like concentration behavior,

$$\mathbb{P}\big(f(X^n) \geq \mathbb{E}f(X^n) + r\big) \leq K \exp\big(-\kappa r^2\big), \qquad \forall r \geq 0$$

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 [138], with subsequent refinements by Steele [139]:

Theorem 3.3.11 (Efron–Stein–Steele inequality)**.** Let $X_1, \ldots, X_n$ be independent $\mathcal{X}$-valued random variables. Consider any function $f\colon \mathcal{X}^n \to \mathbb{R}$, such that its scaled versions tf are exponentially integrable for all sufficiently small $t > 0$. Then

$$\mathsf{var}[f(X^n)] \leq \sum_{i=1}^{n} \mathbb{E}\left\{ \mathsf{var}\big[f(X^n) \big| \bar{X}^i\big] \right\}. \qquad (3.3.49)$$

Proof. By Proposition 3.2, for any $t > 0$, we have

$$D(P^{(tf)}\|P) \le \sum_{i=1}^{n} D(P^{(tf)}_{X_i|\bar{X}^i}\|P_{X_i}|P^{(tf)}_{\bar{X}^i}).$$

Using Lemma 3.3.2, 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 (3.3.49). □

Next, we discuss the connection between log-Sobolev inequalities and another class of functional inequalities, the so-called *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 $c \ge 0$ if

$$\mathsf{var}_\mu[f] \le c\,\mathbb{E}_\mu\left[(\Gamma f)^2\right], \qquad \forall f \in \mathcal{A}. \tag{3.3.50}$$

Theorem 3.3.12. Suppose that μ satisfies LSI(c) with respect to $(\mathcal{A}, \Gamma)$. 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 3.3.2 to express the corresponding LSI(c) for the function tf as

$$\int_0^t \int_s^t \mathsf{var}_\mu^{(\tau f)}[f]\,\mathrm{d}\tau\,\mathrm{d}s \le \frac{ct^2}{2} \cdot \mathbb{E}_\mu^{(tf)}\left[(\Gamma f)^2\right]. \tag{3.3.51}$$

Proceeding exactly as in the proof of Theorem 3.3.11 above (i.e., by dividing both sides of the above inequality by t^2 and passing to the limit as $t \to 0$), we obtain

$$\frac{1}{2}\mathsf{var}_\mu[f] \le \frac{c}{2} \cdot \mathbb{E}_\mu\left[(\Gamma f)^2\right].$$

Multiplying both sides by 2, we see that μ indeed satisfies (3.3.50). □

Moreover, Poincaré inequalities tensorize, as the following analogue of Theorem 3.3.1 shows:

Theorem 3.3.13. 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 $f \colon \mathcal{X}^n \to \mathbb{R}$, such that, for every i,

$$f_i(\cdot|\bar{x}^i) \in \mathcal{A}_i, \qquad \forall \bar{x}^i \in \mathcal{X}^{n-1} \tag{3.3.52}$$

Define the operator Γ that maps each $f \in \mathcal{A}$ to

$$\Gamma f = \sqrt{\sum_{i=1}^{n} (\Gamma_i f_i)^2}, \tag{3.3.53}$$

which is shorthand for

$$\Gamma f(x^n) = \sqrt{\sum_{i=1}^{n} \big(\Gamma_i f_i(x_i|\bar{x}^i)\big)^2}, \qquad \forall x^n \in \mathcal{X}^n. \tag{3.3.54}$$

Suppose that, for every $i \in \{1, \ldots, n\}$, P_{X_i} satisfies a Poincaré inequality with constant c with respect to $(\mathcal{A}_i, \Gamma_i)$. Then P satisfies a Poincaré inequality with constant c with respect to $(\mathcal{A}, \Gamma)$.

Proof. The proof is conceptually similar to the proof of Theorem 3.3.1 (which refers to the tensorization of the logarithmic Sobolev inequality), except that now we use the Efron–Stein–Steele inequality of Theorem 3.3.11 to tensorize the variance of f. □

3.4 Transportation-cost inequalities

So far, we have been looking at 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 [44]). 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 $f\colon \mathcal{X}^n \to \mathbb{R}$:

$$|\nabla f|(x^n) \triangleq \limsup_{y^n : d(x^n,y^n)\downarrow 0} \frac{|f(x^n)-f(y^n)|}{d(x^n,y^n)}. \tag{3.4.1}$$

If we also define the Lipschitz constant of f by

$$\|f\|_{\mathrm{Lip}} \triangleq \sup_{x^n \neq y^n} \frac{|f(x^n)-f(y^n)|}{d(x^n,y^n)} \tag{3.4.2}$$

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, suppose that a given probability distribution P for a random n-tuple $X^n \in \mathcal{X}^n$ satisfies LSI(c) with respect to the pair $(\mathcal{A},\Gamma)$. The use of (3.3.9) and the inequality $\|\Gamma f\|_\infty \leq \|f\|_{\mathrm{Lip}}$, which follows directly from (3.4.1) and (3.4.2), gives the concentration inequality

$$\mathbb{P}\Big(f(X^n) \geq \mathbb{E}f(X^n) + r\Big) \leq \exp\left(-\frac{r^2}{2c\|f\|^2_{\mathrm{Lip}}}\right), \quad \forall\, r > 0. \tag{3.4.3}$$

All the examples of concentration we have discussed so far in this chapter 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 $\|f\|_{\text{Lip}}$ of any function $f\colon \mathcal{X}^n \to \mathbb{R}$ is given by

$$\begin{aligned}\|f\|_{\text{Lip}} &\triangleq \sup_{x^n \neq y^n} \frac{|f(x^n) - f(y^n)|}{d(x^n, y^n)} \\ &= \sup_{x^n \neq y^n} \frac{|f(x^n) - f(y^n)|}{\|x^n - y^n\|}, \end{aligned} \tag{3.4.4}$$

and for any probability measure P on $\mathbb{R}^n$ that satisfies LSI(c) we have the bound (3.4.3). We have already seen in (3.2.11) 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 $\|f\|_{\text{Lip}}$, which we also denote by $\|f\|_{\text{Lip},\, c^n}$ to emphasize the role of the weights $\{c_i\}_{i=1}^n$, is given by

$$\|f\|_{\text{Lip}, c^n} \triangleq \sup_{x^n \neq y^n} \frac{|f(x^n) - f(y^n)|}{d_{c^n}(x^n, y^n)}.$$

Then it is easy to see that the condition $\|f\|_{\text{Lip},\, c^n} \leq 1$ is equivalent to (3.3.35). 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) with respect to

$$\mathcal{A} = \Big\{ f : \|f\|_{\text{Lip},\, c^n} < \infty \Big\}$$

and $\Gamma f(\cdot) = |\nabla f|(\cdot)$ with $|\nabla f|$ given by (3.4.1) with $d = d_{c^n}$. In this case, the concentration inequality (3.4.3) (with $c = 1/4$) is precisely McDiarmid's inequality (3.3.36).

The above two examples suggest that the metric structure plays the primary role, while the functional concentration inequalities like (3.4.3) 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 can then be related to the divergence via so-called *transportation-cost inequalities*. The pioneering work by K. Marton in [58] and [71] 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 [140]. 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

$$A_r \triangleq \{x \in \mathcal{X} : d(x, A) < r\}, \tag{3.4.5}$$

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

$$\mu(A) \geq 1/2 \qquad \Longrightarrow \qquad \mu(A_r) \geq 1 - Ke^{-\kappa r^2}, \ \forall\, r > 0. \tag{3.4.6}$$

Remark 3.14. Of the two constants K and κ in (3.4.6), 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 (3.4.6) holds with that value of κ and some $K > 0$.

Remark 3.15. The concentration condition (3.4.6) is often weakened to the following: there exists some $r_0 > 0$, such that

$$\mu(A) \geq 1/2 \qquad \Longrightarrow \qquad \mu(A_r) \geq 1 - Ke^{-\kappa(r - r_0)^2}, \ \forall\, r \geq r_0 \tag{3.4.7}$$

(see, for example, [60, Remark 22.23] or [64, Proposition 3.3]). It is not hard to pass from (3.4.7) to the stronger statement (3.4.6), possibly with degraded constants (i.e., larger K and/or smaller κ). However, since we mainly care about sufficiently large values of r, (3.4.7) with *sharper constants* is preferable. In the sequel, therefore, whenever we talk about Gaussian concentration with constant $\kappa > 0$, we will normally refer to (3.4.7), unless stated otherwise.

Here are a few standard examples (see [3, Section 1.1]):

1. **Standard Gaussian distribution** — if $\mathcal{X} = \mathbb{R}^n$, $d(x, y) = \|x - y\|$ is the standard Euclidean metric, and $\mu = G^n$ is the standard Gaussian distribution, then for any Borel set $A \subseteq \mathbb{R}^n$ with $G^n(A) \geq 1/2$ we have

$$\begin{aligned} G^n(A_r) &\geq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{r} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t \\ &\geq 1 - \frac{1}{2} \exp\left(-\frac{r^2}{2}\right), \qquad \forall\, r > 0 \end{aligned} \tag{3.4.8}$$

   i.e., (3.4.6) holds with $K = \frac{1}{2}$ and $\kappa = \frac{1}{2}$.

2. **Uniform distribution on the unit sphere** — if $\mathcal{X} = \mathbb{S}^n \equiv \{x \in \mathbb{R}^{n+1} : \|x\| = 1\}$, 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 $\sigma^n(A) \geq 1/2$ we have

$$\sigma^n(A_r) \geq 1 - \exp\left(-\frac{(n-1)r^2}{2}\right), \qquad \forall\, r > 0. \tag{3.4.9}$$

   In this instance, (3.4.6) 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

$$B^n(A_r) \geq 1 - \exp\left(-2nr^2\right), \qquad \forall\, r > 0 \tag{3.4.10}$$

so (3.4.6) holds with $K = 1$ and $\kappa = 2n$.

Remark 3.16. Gaussian concentration of the form (3.4.6) is often discussed in the context of 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 [3, Section 2.1])

$$\mu^+(A) \triangleq \liminf_{r \to 0} \frac{\mu(A_r \setminus A)}{r} = \liminf_{r \to 0} \frac{\mu(A_r) - \mu(A)}{r}. \tag{3.4.11}$$

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

$$(G^n)^+(A) \geq (G^n)^+(H), \tag{3.4.12}$$

with equality if and only if A is a half-space. In other words, the Gaussian isoperimetric inequality (3.4.12) 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 (3.4.12) says the following (see, e.g., [141]): Consider a Borel set A in $\mathbb{R}^n$ and a half-space $H = \{x : \langle x, u\rangle < c\}$ with $\|u\| = 1$, $c \geq 0$ and $G^n(A) = G^n(H)$. Then, for any $r > 0$, we have

$$G^n(A_r) \geq G^n(H_r),$$

with equality if and only if A is itself a half-space. Moreover, an easy calculation shows that

$$\begin{aligned} G^n(H_r) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{c+r} \exp\left(-\frac{\xi^2}{2}\right) \mathrm{d}\xi \\ &\geq 1 - \frac{1}{2} \exp\left(-\frac{(r+c)^2}{2}\right), \quad \forall\, r > 0. \end{aligned}$$

So, if $G(A) \geq 1/2$, we can always choose $c = 0$ and get (3.4.8).

Intuitively, what (3.4.6) 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 $f\colon \mathcal{X} \to \mathbb{R}$: we say that a real number m_f is a median of f with respect to μ (or a μ-*median* of f) if

$$\mathbb{P}_\mu(f(X) \geq m_f) \geq \frac{1}{2} \qquad \text{and} \qquad \mathbb{P}_\mu(f(X) \leq m_f) \geq \frac{1}{2} \tag{3.4.13}$$

(note that a median of f may not be unique). The precise result is as follows:

Theorem 3.4.1. Let $(\mathcal{X}, d, \mu)$ be a metric probability space. Then μ has the normal concentration property (3.4.6) (with arbitrary constants $K, \kappa > 0$) if and only if for every Lipschitz function $f\colon \mathcal{X} \to \mathbb{R}$ (where the Lipschitz property is defined with respect to the metric d) we have

$$\mathbb{P}_\mu\Big(f(X) \geq m_f + r\Big) \leq K \exp\bigg(-\frac{\kappa r^2}{\|f\|^2_{\text{Lip}}}\bigg), \qquad \forall\, r > 0 \tag{3.4.14}$$

where m_f is any μ-median of f.

Proof. Suppose that μ satisfies (3.4.6). Fix any Lipschitz function f, where, without loss of generality, we may assume that $\|f\|_{\text{Lip}} = 1$. Let m_f be any median of f, and define the set $A^f \triangleq \Big\{x \in \mathcal{X} : f(x) \leq m_f\Big\}$. By definition of the median in (3.4.13), $\mu(A^f) \geq 1/2$. Consequently, by (3.4.6), we have

$$\begin{aligned} \mu(A^f_r) &\equiv \mathbb{P}_\mu\Big(d(X, A^f) < r\Big) \\ &\geq 1 - K\exp(-\kappa r^2), \qquad \forall\, r > 0. \end{aligned} \tag{3.4.15}$$

By the Lipschitz property of f, for any $y \in A^f$ we have $f(X) - m_f \leq f(X) - f(y) \leq d(X, y)$, so $f(X) - m_f \leq d(X, A^f)$. This, together with

(3.4.15), implies that

$$\begin{aligned}\mathbb{P}_\mu\Big(f(X) - m_f < r\Big) &\geq \mathbb{P}_\mu\Big(d(X, A^f) < r\Big) \\ &\geq 1 - K\exp(-\kappa r^2), \qquad \forall\, r > 0\end{aligned}$$

which is (3.4.14).

Conversely, suppose (3.4.14) holds for every Lipschitz f. Choose any Borel set A with $\mu(A) \geq 1/2$ and define the function $f_A(x) \triangleq d(x, A)$ for every $x \in \mathcal{X}$. Then f_A is 1-Lipschitz, since

$$\begin{aligned}|f_A(x) - f_A(y)| &= \Big|\inf_{u\in A} d(x,u) - \inf_{u\in A} d(y,u)\Big| \\ &\leq \sup_{u\in A} |d(x,u) - d(y,u)| \\ &\leq d(x,y),\end{aligned}$$

where the last step is by the triangle inequality. Moreover, zero is a median of f_A, since

$$\mathbb{P}_\mu(f_A(X) \leq 0) = \mathbb{P}_\mu(X \in A) \geq \frac{1}{2} \qquad \text{and} \qquad \mathbb{P}_\mu(f_A(X) \geq 0) \geq \frac{1}{2},$$

where the second bound is vacuously true since $f_A \geq 0$ everywhere. Consequently, with $m_f = 0$, we get

$$\begin{aligned}1 - \mu(A_r) &= \mathbb{P}_\mu(d(X, A) \geq r) \\ &= \mathbb{P}_\mu(f_A(X) \geq m_f + r) \\ &\leq K\exp\left(-\kappa r^2\right), \qquad \forall\, r > 0\end{aligned}$$

which gives (3.4.6). □

In fact, for Lipschitz functions, normal concentration around the mean also implies normal concentration around any median, but possibly with worse constants [3, Proposition 1.7]:

Theorem 3.4.2. Let $(\mathcal{X}, d, \mu)$ be a metric probability space, such that for any 1-Lipschitz function $f\colon \mathcal{X} \to \mathbb{R}$ we have

$$\mathbb{P}_\mu\Big(f(X) \geq \mathbb{E}_\mu[f(X)] + r\Big) \leq K_0\exp\big(-\kappa_0 r^2\big), \qquad \forall\, r > 0 \quad (3.4.16)$$

with some constants $K_0, \kappa_0 > 0$. Then, μ has the normal concentration property (3.4.6) with $K = K_0$ and $\kappa = \kappa_0/4$. Consequently, the concentration inequality in (3.4.14) 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 $f_{A,r}(x) \triangleq \min\{d(x, A), r\}$. From the triangle inequality, $\|f_{A,r}\|_{\mathrm{Lip}} \leq 1$ and

$$\begin{aligned}
\mathbb{E}_\mu[f_{A,r}(X)] &= \int_{\mathcal{X}} \min\{d(x, A), r\}\, \mu(\mathrm{d}x) \\
&= \underbrace{\int_A \min\{d(x, A), r\}\, \mu(\mathrm{d}x)}_{=0} + \int_{A^c} \min\{d(x, A), r\}\, \mu(\mathrm{d}x) \\
&\leq r\, \mu(A^c) \\
&= (1 - \mu(A))\, r. \qquad (3.4.17)
\end{aligned}$$

Then

$$\begin{aligned}
1 - \mu(A_r) &= \mathbb{P}_\mu\Big(d(X, A) \geq r\Big) \\
&= \mathbb{P}_\mu\Big(f_{A,r}(X) \geq r\Big) \\
&\leq \mathbb{P}_\mu\Big(f_{A,r}(X) \geq \mathbb{E}_\mu[f_{A,r}(X)] + r\mu(A)\Big) \\
&\leq K_0 \exp\Big(-\kappa_0 \left(\mu(A) r\right)^2\Big),
\end{aligned}$$

where the first two steps use the definition of $f_{A,r}$, the third step uses (3.4.17), and the last step uses (3.4.16). Consequently, if $\mu(A) \geq 1/2$, we get (3.4.6) with $K = K_0$ and $\kappa = \kappa_0/4$. Theorem 3.4.1 therefore implies that the concentration inequality in (3.4.14) holds for any median m_f with the same constants of κ and K. □

Remark 3.17. Let $(\mathcal{X}, d, \mu)$ be a metric probability space, and suppose that μ has the normal concentration property (3.4.6) (with arbitrary constants $K, \kappa > 0$). Let $f\colon \mathcal{X} \to \mathbb{R}$ be an arbitrary Lipschitz function (with respect to the metric d). Then we can upper-bound the distance between the mean and any μ-median of f in terms of the parameters K, κ and K and the Lipschitz constant of f. From Theorem 3.4.1, we

have

$$\begin{aligned}|\mathbb{E}_\mu[f(X)] - m_f| &\le \mathbb{E}_\mu[|f(X) - m_f|] \\ &= \int_0^\infty \mathbb{P}_\mu(|f(X) - m_f| \ge r)\,\mathrm{d}r \\ &\le \int_0^\infty 2K \exp\left(-\frac{\kappa r^2}{\|f\|_{\mathrm{Lip}}^2}\right)\mathrm{d}r \\ &= \sqrt{\frac{\pi}{\kappa}}\,K\|f\|_{\mathrm{Lip}}\end{aligned}$$

where the last inequality follows from the (one-sided) concentration inequality in (3.4.14) applied to f and $-f$ (both are Lipschitz functions with the same constant). We have already seen a special case of this relation in Chapter 2 (cf. Corollary 2.5.11 and Remark 2.12).

3.4.2 Marton's argument: from transportation to concentration

As we have just seen, the phenomenon of concentration is fundamentally geometric in nature, as captured by the isoperimetric inequality (3.4.6). Once we have established (3.4.6) on a given metric probability space $(\mathcal{X}, d, \mu)$, we immediately obtain Gaussian concentration for all Lipschitz functions $f\colon \mathcal{X} \to \mathbb{R}$ by Theorem 3.4.1.

There is a powerful information-theoretic technique for deriving concentration inequalities like (3.4.6). This technique, first introduced by Marton (see [71] and [58]), 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 $p \ge 1$, let $\mathcal{P}_p(\mathcal{X})$ denote the space of all Borel probability measures μ on $\mathcal{X}$, such that the moment bound

$$\mathbb{E}_\mu[d^p(X, x_0)] < \infty \tag{3.4.18}$$

holds for some (and hence all) $x_0 \in \mathcal{X}$.

Definition 3.1. Given $p \ge 1$, the L^p *Wasserstein distance* between any pair $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$ is defined as

$$W_p(\mu, \nu) \triangleq \inf_{\pi \in \Pi(\mu,\nu)} \left(\int_{\mathcal{X}\times\mathcal{X}} d^p(x, y)\pi(\mathrm{d}x, \mathrm{d}y)\right)^{1/p}, \tag{3.4.19}$$

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 3.18. Another equivalent way of writing down the definition of $W_p(\mu,\nu)$ is

$$W_p(\mu,\nu) = \inf_{X\sim\mu, Y\sim\nu} \left\{\mathbb{E}[d^p(X,Y)]\right\}^{1/p}, \tag{3.4.20}$$

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 $P_Y=\nu$.

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 measurable function $c\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$. 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 its destination in the final sand pile, then the minimum transportation cost is given by

$$C^*(\mu,\nu) \triangleq \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} c(x,y)\pi(\mathrm{d}x,\mathrm{d}y). \tag{3.4.21}$$

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 (3.4.21) has a rich history, dating back to a 1781 essay by Gaspard Monge, who has considered a particular special case of the problem

$$C_0^*(\mu,\nu) \triangleq \inf_{\varphi:\mathcal{X}\to\mathcal{X}} \left\{\int_{\mathcal{X}} c(x,\varphi(x))\mu(\mathrm{d}x) : \mu\circ\varphi^{-1}=\nu\right\}. \tag{3.4.22}$$

Here, the infimum is over all *deterministic* transportation policies, i.e., measurable mappings $\varphi\colon\mathcal{X}\to\mathcal{X}$, 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 (3.4.22) (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 (3.4.21), was considered in 1942 by Leonid Kantorovich (see [142] for a recent reprint). We recommend the books by Villani [59, 60] for a detailed historical overview and rigorous treatment of optimal transportation.

Lemma 3.4.3. 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 and bounded function $\varphi \colon \mathcal{X} \to \mathbb{R}$.
   (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 (3.4.19) [and therefore in (3.4.20)] 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., $\mu(\{x\}) = 0$ 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., $\mathsf{F}_\nu^{-1}(\alpha) \triangleq \inf\{x \in \mathbb{R} : \mathsf{F}_\nu(x) \geq \alpha\}$.

Definition 3.2. We say that a probability measure μ on $(\mathcal{X}, d)$ satisfies an L^p *transportation cost inequality with constant* $c > 0$, or a $\mathrm{T}_p(c)$ inequality for short, if for any probability measure $\nu \ll \mu$ we have

$$W_p(\mu, \nu) \leq \sqrt{2c\, D(\nu \| \mu)}. \tag{3.4.23}$$

Example 3.1 (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

$$\|\mu - \nu\|_{\mathrm{TV}} \triangleq \sup_{A \subseteq \mathcal{X}} |\mu(A) - \nu(A)| = \frac{1}{2} \sum_{x \in \mathcal{X}} |\mu(x) - \nu(x)| \tag{3.4.24}$$

(we are slightly abusing 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

$$\mu(x) = \sum_{y\in\mathcal{X}} \pi(x,y) \geq \pi(x,x),$$

and the same goes for ν. Thus, $\pi(x,x) \leq \min\{\mu(x),\nu(x)\}$, and so

$$\begin{aligned}
\mathbb{E}_\pi[d(X,Y)] &= \mathbb{E}_\pi[1_{\{X\neq Y\}}] && (3.4.25)\\
&= \mathbb{P}(X\neq Y) && (3.4.26)\\
&= 1 - \sum_{x\in\mathcal{X}} \pi(x,x) && (3.4.27)\\
&\geq 1 - \sum_{x\in\mathcal{X}} \min\{\mu(x),\nu(x)\}. && (3.4.28)
\end{aligned}$$

On the other hand, if we define the set $A \triangleq \{x\in\mathcal{X} : \mu(x)\geq\nu(x)\}$, then

$$\begin{aligned}
\|\mu-\nu\|_{\mathrm{TV}} &= \frac{1}{2}\sum_{x\in A}|\mu(x)-\nu(x)| + \frac{1}{2}\sum_{x\in A^c}|\mu(x)-\nu(x)|\\
&= \frac{1}{2}\sum_{x\in A}[\mu(x)-\nu(x)] + \frac{1}{2}\sum_{x\in A^c}[\nu(x)-\mu(x)]\\
&= \frac{1}{2}\Big(\mu(A)-\nu(A)+\nu(A^c)-\mu(A^c)\Big)\\
&= \mu(A)-\nu(A)
\end{aligned}$$

and

$$\begin{aligned}
\sum_{x\in\mathcal{X}}\min\{\mu(x),\nu(x)\} &= \sum_{x\in A}\nu(x) + \sum_{x\in A^c}\mu(x)\\
&= \nu(A)+\mu(A^c)\\
&= 1-\Big(\mu(A)-\nu(A)\Big)\\
&= 1-\|\mu-\nu\|_{\mathrm{TV}}. \qquad (3.4.29)
\end{aligned}$$

Consequently, for any $\pi\in\Pi(\mu,\nu)$ we see from (3.4.25)–(3.4.29) that

$$\mathbb{P}(X\neq Y) = \mathbb{E}_\pi[d(X,Y)] \geq \|\mu-\nu\|_{\mathrm{TV}}. \qquad (3.4.30)$$

Moreover, the lower bound in (3.4.30) is actually achieved by π^* taking

$$\pi^*(x,y) = \min\{\mu(x),\nu(x)\}\,1_{\{x=y\}} + \frac{(\mu(x)-\nu(x))1_{\{x\in A\}}(\nu(y)-\mu(y))1_{\{y\in A^c\}}}{\mu(A)-\nu(A)}1_{\{x\neq y\}}. \tag{3.4.31}$$

Now that we have expressed the total variation distance $\|\mu-\nu\|_{\mathrm{TV}}$ as the L^1 Wasserstein distance induced by the Hamming metric on $\mathcal{X}$, the well-known Pinsker's inequality

$$\|\mu-\nu\|_{\mathrm{TV}} \le \sqrt{\frac{1}{2}D(\nu\|\mu)} \tag{3.4.32}$$

can be identified as a $\mathrm{T}_1(1/4)$ inequality that holds for any probability measure μ on $\mathcal{X}$.

Remark 3.19. It should be pointed out that the constant $c = 1/4$ in Pinsker's inequality (3.4.32) is not necessarily the best possible *for a given distribution* μ. Ordentlich and Weinberger [143] have obtained the following *distribution-dependent* refinement of Pinsker's inequality. Let the function $\varphi\colon [0,1/2] \to \mathbb{R}^+$ be defined by

$$\varphi(p) \triangleq \begin{cases} \left(\dfrac{1}{1-2p}\right)\ln\left(\dfrac{1-p}{p}\right), & \text{if } p \in [0,\frac{1}{2}) \\ 2, & \text{if } p = \frac{1}{2} \end{cases} \tag{3.4.33}$$

(in fact, $\varphi(p) \to 2$ as $p \uparrow 1/2$, $\varphi(p) \to \infty$ as $p \downarrow 0$, and φ is a monotonically decreasing and convex function). Let $\mathcal{X}$ be a discrete set. For any $P \in \mathcal{P}(\mathcal{X})$, define the *balance coefficient*

$$\pi_P \triangleq \max_{A\subseteq\mathcal{X}} \min\{P(A),1-P(A)\} \quad\Longrightarrow\quad \pi_P \in \Big[0,\frac{1}{2}\Big].$$

Then, for any $Q \in \mathcal{P}(\mathcal{X})$,

$$\|P-Q\|_{\mathrm{TV}} \le \sqrt{\frac{1}{\varphi(\pi_P)}D(Q\|P)} \tag{3.4.34}$$

(see [143, Theorem 2.1]; related results have also appeared in a technical report by Weissman et al. [144]). 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

$$\varphi(\pi_P) = \inf_{Q \in \mathcal{P}(\mathcal{X})} \frac{D(Q\|P)}{\|P-Q\|_{\mathrm{TV}}^2}. \tag{3.4.35}$$

For instance, if $\mathcal{X} = \{0,1\}$ and P is the distribution of a Bernoulli(p) random variable, then $\pi_P = \min\{p, 1-p\} \in \left[0, \frac{1}{2}\right]$,

$$\varphi(\pi_P) = \begin{cases} \left(\frac{1}{1-2p}\right) \ln\left(\frac{1-p}{p}\right), & \text{if } p \neq \frac{1}{2} \\ 2, & \text{if } p = \frac{1}{2} \end{cases}$$

and for any other $Q \in \mathcal{P}(\{0,1\})$ we have, from (3.4.34),

$$\|P-Q\|_{\mathrm{TV}} \leq \begin{cases} \sqrt{\frac{1-2p}{\ln[(1-p)/p]} D(Q\|P)}, & \text{if } p \neq \frac{1}{2} \\ \sqrt{\frac{1}{2}\, D(Q\|P)}, & \text{if } p = \frac{1}{2}. \end{cases} \tag{3.4.36}$$

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 it is easy to come up with examples where the total variation distance is arbitrarily close to zero, whereas the divergence is equal to infinity. However, consider an i.i.d. sample of size n drawn 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 $\varepsilon \in (0,1]$ scales asymptotically like $\exp\bigl(-n\, D^*(P,\varepsilon)\bigr)$, where

$$D^*(P,\varepsilon) \triangleq \inf_{Q:\, \|P-Q\|_{\mathrm{TV}} \geq \varepsilon} D(Q\|P).$$

Although a reverse form of Pinsker's inequality (or its probability-dependent refinement in [143]) cannot be derived, it was recently proved in [145] 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 (3.4.34) 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 (3.4.34) also enables us 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 [146] in the context of the Poisson approximation, where (3.4.34) 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. Note that, for a sum of i.i.d. Bernoulli random variables, the lower bound on this relative entropy (see [146]) scales similarly to the upper bound on this relative entropy derived by Kontoyiannis et al. (see [147, Theorem 1]) using the Bobkov–Ledoux logarithmic Sobolev inequality for the Poisson distribution [54] (see also Section 3.3.5 here).

Marton's procedure for deriving Gaussian concentration from a transportation cost inequality [71, 58] can be distilled in the following:

Proposition 3.6. Suppose μ satisfies a $\mathrm{T}_1(c)$ inequality. Then, the Gaussian concentration inequality in (3.4.7) holds with $\kappa = 1/(2c)$, $K = 1$, and $r_0 = \sqrt{2c\ln 2}$.

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

$$W_1(\mu_A, \mu_B) \le W_1(\mu, \mu_A) + W_1(\mu, \mu_B) \tag{3.4.37}$$

$$\le \sqrt{2cD(\mu_A\|\mu)} + \sqrt{2cD(\mu_B\|\mu)}, \tag{3.4.38}$$

where (3.4.37) is by the triangle inequality, while (3.4.38) is because μ satisfies $\mathrm{T}_1(c)$. 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 $\mathrm{d}\mu_A/\mathrm{d}\mu = 1_A/\mu(A)$, and the same holds for μ_B. Therefore,

$$D(\mu_A\|\mu) = \mathbb{E}_\mu\left[\frac{\mathrm{d}\mu_A}{\mathrm{d}\mu}\ln\frac{\mathrm{d}\mu_A}{\mathrm{d}\mu}\right] = \ln\frac{1}{\mu(A)}, \tag{3.4.39}$$

and an analogous formula holds for μ_B in place of μ_A. Substituting this into (3.4.38) gives

$$W_1(\mu_A,\mu_B) \le \sqrt{2c\ln\frac{1}{\mu(A)}} + \sqrt{2c\ln\frac{1}{\mu(B)}}. \tag{3.4.40}$$

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 $\pi\in\Pi(\mu_A,\mu_B)$ is supported on $A\times B$. Consequently, for any such π we have

$$\begin{aligned}
\int_{\mathcal{X}\times\mathcal{X}} d(x,y)\,\pi(\mathrm{d}x,\mathrm{d}y) &= \int_{A\times B} d(x,y)\,\pi(\mathrm{d}x,\mathrm{d}y)\\
&\ge \int_{A\times B}\inf_{y\in B} d(x,y)\,\pi(\mathrm{d}x,\mathrm{d}y)\\
&= \int_A d(x,B)\,\mu_A(\mathrm{d}x)\\
&\ge \inf_{x\in A} d(x,B)\,\mu_A(A)\\
&= d(A,B),
\end{aligned} \tag{3.4.41}$$

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 (3.4.41) holds for every $\pi\in\Pi(\mu_A,\mu_B)$, we can take the infimum over all such π and get $W_1(\mu_A,\mu_B)\ge d(A,B)$. Combining this with (3.4.40) gives the inequality

$$d(A,B) \le \sqrt{2c\ln\frac{1}{\mu(A)}} + \sqrt{2c\ln\frac{1}{\mu(B)}}, \tag{3.4.42}$$

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 $d(A,B) \geq r$. Consequently, (3.4.42) gives

$$r \leq \sqrt{2c \ln \frac{1}{\mu(A)}} + \sqrt{2c \ln \frac{1}{1-\mu(A_r)}}. \tag{3.4.43}$$

If $\mu(A) \geq 1/2$ and $r \geq \sqrt{2c \ln 2}$, then (3.4.43) gives

$$\mu(A_r) \geq 1 - \exp\left(-\frac{1}{2c}\left(r - \sqrt{2c \ln 2}\right)^2\right). \tag{3.4.44}$$

Hence, the Gaussian concentration inequality in (3.4.7) indeed holds with $\kappa = 1/(2c)$ and $K = 1$ for all $r \geq r_0 = \sqrt{2c \ln 2}$. □

Remark 3.20. The exponential inequality (3.4.44) has appeared earlier in the work of McDiarmid [85] and Talagrand [7]. The major innovation that came from Marton's work was her use of optimal transportation ideas to derive a more general "symmetric" form (3.4.42).

Remark 3.21. The formula (3.4.39), apparently first used explicitly by Csiszár [148, 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 3.6 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, [60, Proposition 22.5]):

Proposition 3.7 (Tensorization of transportation cost inequalities)**.** For any $p \in [1,2]$, the following statement is true: If μ satisfies $\mathrm{T}_p(c)$ on $(\mathcal{X}, d)$, then, for any $n \in \mathbb{N}$, the product measure $\mu^{\otimes n}$ satisfies $\mathrm{T}_p(cn^{2/p-1})$ on $(\mathcal{X}^n, d_{p,n})$ with the metric

$$d_{p,n}(x^n, y^n) \triangleq \left(\sum_{i=1}^n d^p(x_i, y_i)\right)^{1/p}, \qquad \forall x^n, y^n \in \mathcal{X}^n. \tag{3.4.45}$$

Proof. Suppose μ satisfies $\mathrm{T}_p(c)$. 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

$$P_{X^n} = P_{X_1} \otimes P_{X_2|X_1} \otimes \ldots \otimes P_{X_n|X^{n-1}} = \nu \tag{3.4.46}$$

$$P_{Y^n} = P_{Y_1} \otimes P_{Y_2} \otimes \ldots \otimes P_{Y_n} = \mu^{\otimes n}. \tag{3.4.47}$$

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

$$\begin{aligned} W_p^p(\nu, \mu^{\otimes n}) &= W_p^p(P_{X^n}, P_{Y^n}) \\ &\leq \sum_{i=1}^n W_p^p(P_{X_i|X^{i-1}}, P_{Y_i}|P_{X^{i-1}}), \end{aligned} \tag{3.4.48}$$

where the L^p Wasserstein distance on the left-hand side is computed with respect to the $d_{p,n}$ metric. By Lemma 3.4.3, 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 3.4.3, 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

$$W_p^p(P_{X_i|X^{i-1}=x^{i-1}}, P_{Y_i}) = \mathbb{E}[d^p(X_i^*(x^{i-1}), Y_i^*(x^{i-1}))]. \tag{3.4.49}$$

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 $C \subseteq \mathcal{X} \times \mathcal{X}$ [60]. 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

$$P_{X^{*n}Y^{*n}} = P_{X_1^*Y_1^*} \otimes P_{X_2^*Y_2^*|X_1^*} \otimes \ldots \otimes P_{X_n^*Y_n^*|X^{*(n-1)}}$$

is a coupling of $P_{X^n} = \nu$ and $P_{Y^n} = \mu^{\otimes n}$, and

$$W_p^p(P_{X_i|X^{i-1}}, P_{Y_i}) = \mathbb{E}[d^p(X_i^*, Y_i^*)|X^{*(i-1)}],\ i = 1, \ldots, n. \tag{3.4.50}$$

By definition of W_p, we then have

$$\begin{aligned} W_p^p(\nu, \mu^{\otimes n}) &\leq \mathbb{E}[d_{p,n}^p(X^{*n}, Y^{*n})] && (3.4.51)\\ &= \sum_{i=1}^n \mathbb{E}[d^p(X_i^*, Y_i^*)] && (3.4.52)\\ &= \sum_{i=1}^n \mathbb{E}\Big[\mathbb{E}[d^p(X_i^*, Y_i^*)|X^{*(i-1)}]\Big] && (3.4.53)\\ &= \sum_{i=1}^n W_p^p(P_{X_i|X^{i-1}}, P_{Y_i}|P_{X^{i-1}}), && (3.4.54) \end{aligned}$$

where:

- (3.4.51) is due to the fact that (X^{*n}, Y^{*n}) is a (not necessarily optimal) coupling of $P_{X^n} = \nu$ and $P_{Y^n} = \mu^{\otimes n}$;
- (3.4.52) is by the definition (3.4.45) of $d_{p,n}$;
- (3.4.53) is by the law of iterated expectation; and
- (3.4.54) is by (3.4.49).

We have thus proved (3.4.48). By hypothesis, μ satisfies $\mathrm{T}_p(c)$ on $(\mathcal{X}, d)$. Therefore, since $P_{Y_i} = \mu$ for every i, we can write

$$\begin{aligned} &W_p^p(P_{X_i|X^{i-1}}, P_{Y_i}|P_{X^{i-1}})\\ &\quad= \int_{\mathcal{X}^{i-1}} W_p^p(P_{X_i|X^{i-1}=x^{i-1}}, P_{Y_i}) P_{X^{i-1}}(\mathrm{d}x^{i-1})\\ &\quad\leq \int_{\mathcal{X}^{i-1}} \Big(2cD(P_{X_i|X^{i-1}=x^{i-1}}\|P_{Y_i})\Big)^{p/2} P_{X^{i-1}}(\mathrm{d}x^{i-1})\\ &\quad\leq (2c)^{p/2} \left(\int_{\mathcal{X}^{i-1}} D(P_{X_i|X^{i-1}=x^{i-1}}\|P_{Y_i}) P_{X^{i-1}}(\mathrm{d}x^{i-1})\right)^{p/2}\\ &\quad= (2c)^{p/2} \Big(D(P_{X_i|X^{i-1}}\|P_{Y_i}|P_{X^{i-1}})\Big)^{p/2}, \end{aligned} \tag{3.4.55}$$

where the second inequality follows from Jensen's inequality and the concavity of the function $t \mapsto t^{p/2}$ for $p \in [1,2]$. Consequently, it follows that

$$\begin{aligned} W_p^p(\nu, \mu^{\otimes n}) &\leq (2c)^{p/2} \sum_{i=1}^n \left(D(P_{X_i|X^{i-1}} \| P_{Y_i} | P_{X^{i-1}}) \right)^{p/2} \\ &\leq (2c)^{p/2} n^{1-p/2} \left(\sum_{i=1}^n D(P_{X_i|X^{i-1}} \| P_{Y_i} | P_{X^{i-1}}) \right)^{p/2} \\ &= (2c)^{p/2} n^{1-p/2} \left(D(P_{X^n} \| P_{Y^n}) \right)^{p/2} \\ &= (2c)^{p/2} n^{1-p/2} D(\nu \| \mu^{\otimes n})^{p/2}, \end{aligned}$$

where the first line holds by summing from $i = 1$ to $i = n$ and using (3.4.48) and (3.4.55), the second line is by Hölder's inequality, the third line is by the chain rule for the divergence and since P_{Y^n} is a product probability measure, and the fourth line is by (3.4.46) and (3.4.47). This finally gives

$$W_p(\nu, \mu^{\otimes n}) \leq \sqrt{2cn^{2/p-1} D(\nu \| \mu^{\otimes n})},$$

i.e., $\mu^{\otimes n}$ indeed satisfies the $\mathrm{T}_p(cn^{2/p-1})$ inequality. □

Since W_2 dominates W_1 (cf. Item 2 of Lemma 3.4.3), a $\mathrm{T}_2(c)$ inequality is stronger than a $\mathrm{T}_1(c)$ inequality (for an arbitrary $c > 0$). Moreover, as Proposition 3.7 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 $\mathrm{T}_1(c)$, 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 $\mathrm{T}_p(c)$ 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 3.7 and conse-

quently also apply Proposition 3.6. If we go through with this approach, we will see that:

- If μ satisfies $\mathrm{T}_1(c)$ on $(\mathcal{X}, d)$, then $\mu^{\otimes n}$ satisfies $\mathrm{T}_1(cn)$ on $(\mathcal{X}^n, d_{1,n})$, which is equivalent to Gaussian concentration with constant $1/(2cn)$. Hence, in this case, the concentration phenomenon is weakened by increasing the dimension n.

- If, on the other hand, μ satisfies $\mathrm{T}_2(c)$ on $(\mathcal{X}, d)$, then $\mu^{\otimes n}$ satisfies $\mathrm{T}_2(c)$ 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 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 $f_n(x^n) \triangleq (1/n) \sum_{i=1}^n f(x_i)$,

$$\|f_n\|_{\mathrm{Lip},1} \triangleq \sup_{x^n \neq y^n} \frac{|f_n(x^n) - f_n(y^n)|}{d_{1,n}(x^n, y^n)} \le \frac{\|f\|_{\mathrm{Lip}}}{n}$$

and

$$\|f_n\|_{\mathrm{Lip},2} \triangleq \sup_{x^n \neq y^n} \frac{|f_n(x^n) - f_n(y^n)|}{d_{2,n}(x^n, y^n)} \le \frac{\|f\|_{\mathrm{Lip}}}{\sqrt{n}}\,.$$

Therefore, both $\mathrm{T}_1(c)$ and $\mathrm{T}_2(c)$ give

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n f(X_i) \ge r \right) \le \exp\left(-\frac{nr^2}{2c\|f\|_{\mathrm{Lip}}^2} \right), \quad \forall\, r > 0$$

where $X_1, \ldots, X_n \in \mathcal{X}$ are i.i.d. random variables whose common marginal μ satisfies either $\mathrm{T}_2(c)$ or $\mathrm{T}_1(c)$, 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 "non-linear" functions of $X_1, \ldots, X_n$. However, in practice it is often easier to work with T_1 inequalities than with T_2 inequalities.

The same strategy as above can be used to prove the following generalization of Proposition 3.7:

Proposition 3.8. 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 $\mathrm{T}_p(c_i)$ for some $c_i > 0$, for each $i = 1, \ldots, n$. Let $c \triangleq \max_{1 \leq i \leq n} c_i$. Then $\mu = \mu_1 \otimes \ldots \otimes \mu_n$ satisfies $\mathrm{T}_p(cn^{2/p-1})$ 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 was obtained by Bobkov and Götze [53], and it relies on the following variational representations of the L^1 Wasserstein distance and the divergence:

1. **Kantorovich–Rubinstein theorem** [59, Theorem 1.14]: For any $\mu, \nu \in \mathcal{P}_1(\mathcal{X})$ on a Polish probability space $(\mathcal{X}, d)$,
$$W_1(\mu, \nu) = \sup_{f: \|f\|_{\mathrm{Lip}} \leq 1} |\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f]|. \tag{3.4.56}$$

2. **Donsker–Varadhan lemma** [81, Lemma 6.2.13]: For any two Borel probability measures μ, ν on a Polish probability space $(\mathcal{X}, d)$ such that $\nu \ll \mu$, the following variational representation of the divergence holds:
$$D(\nu\|\mu) = \sup_{g \in C_{\mathrm{b}}(\mathcal{X})} \{\mathbb{E}_\nu[g] - \ln \mathbb{E}_\mu[\exp(g)]\} \tag{3.4.57}$$
where the supremization in (3.4.57) is over the set $C_{\mathrm{b}}(\mathcal{X})$ of continuous and bounded real-valued functions on $\mathcal{X}$. Furthermore,

for any measurable function g such that $\mathbb{E}_\mu[\exp(g)] < \infty$,

$$\mathbb{E}_\nu[g] \leq D(\nu\|\mu) + \ln \mathbb{E}_\mu[\exp(g)]. \tag{3.4.58}$$

(In fact, the supremum in (3.4.57) can be extended to bounded Borel-measurable functions g [149, Lemma 1.4.3].)

Theorem 3.4.4 (Bobkov–Götze [53])**.** A Borel probability measure $\mu \in \mathcal{P}_1(\mathcal{X})$ satisfies $\mathrm{T}_1(c)$ if and only if the inequality

$$\mathbb{E}_\mu\left\{\exp[tf(X)]\right\} \leq \exp\left(\frac{ct^2}{2}\right) \tag{3.4.59}$$

holds for all 1-Lipschitz functions $f\colon \mathcal{X} \to \mathbb{R}$ with $\mathbb{E}_\mu[f(X)] = 0$, and all $t \in \mathbb{R}$.

Remark 3.22. The moment condition $\mathbb{E}_\mu[d(X, x_0)] < \infty$ is needed to ensure that every Lipschitz function $f\colon \mathcal{X} \to \mathbb{R}$ is μ-integrable:

$$\begin{aligned}\mathbb{E}_\mu\big[|f(X)|\big] &\leq |f(x_0)| + \mathbb{E}_\mu\big[|f(X) - f(x_0)|\big] \\ &\leq |f(x_0)| + \|f\|_{\mathrm{Lip}}\, \mathbb{E}_\mu\big[d(X, x_0)\big] < \infty.\end{aligned}$$

Proof. Without loss of generality, we may consider (3.4.59) only for $t \geq 0$.

Suppose first that μ satisfies $\mathrm{T}_1(c)$. Consider some $\nu \ll \mu$. Using the $\mathrm{T}_1(c)$ property of μ together with the Kantorovich–Rubinstein formula (3.4.56), we can write

$$\int_{\mathcal{X}} f \mathrm{d}\nu \leq W_1(\nu, \mu) \leq \sqrt{2cD(\nu\|\mu)}$$

for any 1-Lipschitz $f\colon \mathcal{X} \to \mathbb{R}$ with $\mathbb{E}_\mu[f] = 0$. Next, from the fact that

$$\inf_{t>0}\left(\frac{a}{t} + \frac{bt}{2}\right) = \sqrt{2ab} \tag{3.4.60}$$

for any $a, b \geq 0$, we see that any such f must satisfy

$$\int_{\mathcal{X}} f\, \mathrm{d}\nu \leq \frac{D(\nu\|\mu)}{t} + \frac{ct}{2}, \qquad \forall\, t > 0.$$

Rearranging, we obtain

$$\int_{\mathcal{X}} tf\, \mathrm{d}\nu - \frac{ct^2}{2} \leq D(\nu\|\mu), \qquad \forall\, t > 0.$$

Applying this inequality to $\nu = \mu^{(g)}$ (the g-tilting of μ) where $g \triangleq tf$, and using the fact that

$$\begin{aligned} D(\mu^{(g)}\|\mu) &= \int_{\mathcal{X}} g \,\mathrm{d}\mu^{(g)} - \ln \int_{\mathcal{X}} \exp(g) \,\mathrm{d}\mu \\ &= \int_{\mathcal{X}} tf \,\mathrm{d}\nu - \ln \int_{\mathcal{X}} \exp(tf) \,\mathrm{d}\mu \end{aligned}$$

we deduce that

$$\ln \left(\int_{\mathcal{X}} \exp(tf) \,\mathrm{d}\mu \right) \leq \frac{ct^2}{2}$$

for all $t \geq 0$, and all f with $\|f\|_{\mathrm{Lip}} \leq 1$ and $\mathbb{E}_\mu[f] = 0$, which is precisely (3.4.59).

Conversely, assume that μ satisfies (3.4.59) for all 1-Lipschitz functions $f \colon \mathcal{X} \to \mathbb{R}$ with $\mathbb{E}_\mu[f(X)] = 0$ and all $t \in \mathbb{R}$, and let ν be an arbitrary Borel probability measure such that $\nu \ll \mu$. Consider any function of the form $g \triangleq tf$ where $t > 0$. By the assumption in (3.4.59), $\mathbb{E}_\mu[\exp(g)] < \infty$; furthermore, g is a Lipschitz function, so it is also measurable. Hence, (3.4.58) gives

$$\begin{aligned} D(\nu\|\mu) &\geq \int_{\mathcal{X}} tf \,\mathrm{d}\nu - \ln \int_{\mathcal{X}} \exp(tf) \,\mathrm{d}\mu \\ &\geq \int_{\mathcal{X}} tf \,\mathrm{d}\nu - \int_{\mathcal{X}} tf \,\mathrm{d}\mu - \frac{ct^2}{2} \end{aligned}$$

where in the second step we have used the fact that $\int_{\mathcal{X}} f \,\mathrm{d}\mu = 0$ by hypothesis, as well as (3.4.59). Rearranging gives

$$\left| \int_{\mathcal{X}} f \,\mathrm{d}\nu - \int_{\mathcal{X}} f \,\mathrm{d}\mu \right| \leq \frac{D(\nu\|\mu)}{t} + \frac{ct}{2}, \qquad \forall\, t > 0 \tag{3.4.61}$$

(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 (3.4.60), we see that the bound

$$\left| \int_{\mathcal{X}} f \,\mathrm{d}\nu - \int_{\mathcal{X}} f \,\mathrm{d}\mu \right| \leq \sqrt{2cD(\nu\|\mu)} \tag{3.4.62}$$

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 (3.4.62) and using the Kantorovich–Rubinstein formula (3.4.56), we conclude that $W_1(\mu,\nu) \le \sqrt{2cD(\nu\|\mu)}$ for every $\nu \ll \mu$, i.e., μ satisfies $\mathrm{T}_1(c)$. This completes the proof of Theorem 3.4.4. □

Theorem 3.4.4 gives us an alternative way of deriving Gaussian concentration for Lipschitz functions (compare with earlier derivations using the entropy method):

Corollary 3.4.5. Let $\mathcal{A}$ be the space of all Lipschitz functions on $\mathcal{X}$, and let $\mu \in \mathcal{P}_1(\mathcal{X})$ be a Borel probability measure that satisfies $\mathrm{T}_1(c)$. Then, the following inequality holds for every $f \in \mathcal{A}$:

$$\mathbb{P}\Big(f(X) \ge \mathbb{E}_\mu[f(X)] + r\Big) \le \exp\left(-\frac{r^2}{2c\|f\|^2_{\mathrm{Lip}}}\right), \qquad \forall\, r > 0. \quad (3.4.63)$$

Proof. The result follows from the Chernoff bound and (3.4.59). □

As another illustration, we prove the following bound, which includes the Kearns–Saul inequality (cf. Theorem 2.2.6) as a special case:

Theorem 3.4.6. Let $\mathcal{X}$ be the Hamming space $\{0,1\}^n$, equipped with the metric

$$d(x^n, y^n) = \sum_{i=1}^n 1_{\{x_i \ne y_i\}}. \quad (3.4.64)$$

Let $X_1, \ldots, X_n \in \{0,1\}$ be i.i.d. Bernoulli(p) random variables. Then, for every Lipschitz function $f : \{0,1\}^n \to \mathbb{R}$,

$$\mathbb{P}\left(f(X^n) - \mathbb{E}[f(X^n)] \ge r\right) \le \exp\left(-\frac{\ln[(1-p)/p]\, r^2}{n\|f\|^2_{\mathrm{Lip}}(1-2p)}\right), \qquad \forall\, r > 0. \quad (3.4.65)$$

Remark 3.23. In the limit as $p \to 1/2$, the right-hand side of (3.4.65) becomes $\exp\left(-\frac{2r^2}{n\|f\|^2_{\mathrm{Lip}}}\right)$.

Proof. Taking into account Remark 3.23, we may assume without loss of generality that $p \ne 1/2$. From the distribution-dependent refinement

of Pinsker's inequality (3.4.36), it follows that the Bernoulli(p) measure satisfies $\mathrm{T}_1(1/(2\varphi(p)))$ with respect to the Hamming metric, where $\varphi(p)$ is defined in (3.4.33). By Proposition 3.7, the product of n Bernoulli(p) measures satisfies $\mathrm{T}_1(n/(2\varphi(p)))$ with respect to the metric (3.4.64). The bound (3.4.65) then follows from Corollary 3.4.5. □

Remark 3.24. If $\|f\|_{\mathrm{Lip}} \leq C/n$ for some $C > 0$, then (3.4.65) implies that

$$\mathbb{P}\left(f(X^n) - \mathbb{E}[f(X^n)] \geq r\right) \leq \exp\left(-\frac{\ln[(1-p)/p]}{C^2(1-2p)} \cdot nr^2\right), \qquad \forall\, r > 0.$$

This will be the case, for instance, if $f(x^n) = (1/n)\sum_{i=1}^n f_i(x_i)$ for some functions $f_1, \ldots, f_n : \{0,1\} \to \mathbb{R}$ satisfying $|f_i(0) - f_i(1)| \leq C$ for all $i = 1, \ldots, n$. More generally, any f satisfying (3.3.35) with $c_i = c_i'/n$, $i = 1, \ldots, n$, for some constants $c_1', \ldots, c_n' \geq 0$, satisfies

$$\mathbb{P}\left(f(X^n) - \mathbb{E}[f(X^n)] \geq r\right) \leq \exp\left(-\frac{C^2 \ln[(1-p)/p]}{(1-2p)} \cdot nr^2\right)$$

for all $r > 0$, with $C = \max_{1 \leq i \leq n} c_i'$.

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 3.7 shows that the transportation cost inequalities tensorize, so in principle this property can be used to derive concentration inequalities for such functions. However, as suggested by Proposition 3.7 and the discussion following it, T_1 inequalities are not very useful in this regard, since the resulting concentration bounds will deteriorate as n increases. Indeed, if μ satisfies $\mathrm{T}_1(c)$ on $(\mathcal{X}, d)$, then the product measure $\mu^{\otimes n}$ satisfies $\mathrm{T}_1(cn)$ on the product space $(\mathcal{X}^n, d_{1,n})$, which is equivalent to the Gaussian concentration property

$$\mathbb{P}\Big(f(X^n) \geq \mathbb{E}f(X^n) + r\Big) \leq K \exp\left(-\frac{r^2}{2cn}\right)$$

for any $f\colon \mathcal{X}^n \to \mathbb{R}$ with Lipschitz constant 1 with respect to $d_{1,n}$. Since the exponent is inversely proportional to the dimension n, we need to have r grow at least as $\sqrt{n}$ in order to guarantee a given value for the deviation probability. In particular, the higher the dimension n, the more we will need to "inflate" a given set $A \subset \mathcal{X}^n$ to capture most of the probability mass. For these reasons, we seek a direct characterization of a much stronger concentration property, the so-called *dimension-free Gaussian concentration*.

Once again, 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 $k \in \mathbb{N}$ and any $r > 0$

$$A \subseteq \mathcal{X}^k \text{ and } \mu^{\otimes k}(A) \geq 1/2 \quad \Longrightarrow \quad \mu^{\otimes k}(A_r) \geq 1 - Ke^{-\kappa r^2}. \tag{3.4.66}$$

where the isoperimetric enlargement A_r of a Borel set $A \subseteq \mathcal{X}^k$ is defined with respect to the metric $d_k \equiv d_{2,k}$ defined according to (3.4.45):

$$A_r \triangleq \left\{ y^k \in \mathcal{X}^k : \exists x^k \in A \text{ s.t. } \sum_{i=1}^{k} d^2(x_i, y_i) < r^2 \right\}.$$

Remark 3.25. 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 (3.4.66) holds with that κ and some $K > 0$.

Remark 3.26. In the same spirit as Remark 3.15, it may be desirable to relax (3.4.66) to the following: there exists some $r_0 > 0$, such that for any $k \in \mathbb{N}$ and any $r > 0$,

$$A \subseteq \mathcal{X}^k \text{ and } \mu^{\otimes k}(A) \geq 1/2 \qquad \Longrightarrow \qquad \mu^{\otimes k}(A_r) \geq 1 - Ke^{-\kappa(r-r_0)^2}. \tag{3.4.67}$$

(see, for example, [60, Remark 22.23] or [64, Proposition 3.3]). The same considerations about (possibly) sharper constants that were stated in Remark 3.15 also apply here.

In this section, we will show that dimension-free Gaussian concentration and T_2 are equivalent. Before we get to that, here is an example of a T_2 inequality:

Theorem 3.4.7 (Talagrand [150])**.** Let $\mathcal{X} = \mathbb{R}^n$ and $d(x, y) = \|x - y\|$. Then $\mu = G^n$ satisfies a $\mathrm{T}_2(1)$ inequality.

Proof. The proof starts for $n = 1$: let $\mu = G$, let $\nu \in \mathcal{P}(\mathbb{R})$ have density f with respect to 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 3.4.3) the optimal coupling of $\mu = G$ and ν, i.e., the one that achieves the infimum in

$$W_2(\nu, \mu) = W_2(\nu, G) = \inf_{X \sim G,\, Y \sim \nu} \Big(\mathbb{E}[(X - Y)^2]\Big)^{1/2}$$

is given by $Y = h(X)$ with $h = \mathsf{F}_\nu^{-1} \circ \Phi$. Consequently,

$$\begin{aligned} W_2^2(\nu, G) &= \mathbb{E}[(X - h(X))^2] \\ &= \int_{-\infty}^{\infty} \big(x - h(x)\big)^2 \gamma(x)\, \mathrm{d}x. \end{aligned} \tag{3.4.68}$$

Since $\mathrm{d}\nu = f\, \mathrm{d}\mu$ with $\mu = G$, and $\mathsf{F}_\nu(h(x)) = \Phi(x)$ for every $x \in \mathbb{R}$, we have

$$\begin{aligned} \int_{-\infty}^{x} \gamma(y)\, \mathrm{d}y &= \Phi(x) \\ &= \mathsf{F}_\nu(h(x)) \\ &= \int_{-\infty}^{h(x)} \mathrm{d}\nu \\ &= \int_{-\infty}^{h(x)} f\, \mathrm{d}\mu \\ &= \int_{-\infty}^{h(x)} f(y)\gamma(y)\, \mathrm{d}y. \end{aligned} \tag{3.4.69}$$

Differentiating both sides of (3.4.69) with respect to x gives

$$h'(x) f(h(x)) \gamma(h(x)) = \gamma(x), \quad \forall\, x \in \mathbb{R}. \tag{3.4.70}$$

Since $h = \mathsf{F}_\nu^{-1} \circ \Phi$, h is a monotonically increasing function, and

$$\lim_{x \to -\infty} h(x) = -\infty, \quad \lim_{x \to \infty} h(x) = \infty.$$

Moreover,

$$\begin{aligned} D(\nu\|G) &= D(\nu\|\mu) \\ &= \int_{\mathbb{R}} \mathrm{d}\nu \, \ln \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \\ &= \int_{-\infty}^{\infty} \ln\big(f(x)\big) \, \mathrm{d}\nu(x) \\ &= \int_{-\infty}^{\infty} f(x) \, \ln\big(f(x)\big) \, \mathrm{d}\mu(x) \\ &= \int_{-\infty}^{\infty} f(x) \, \ln\big(f(x)\big) \, \gamma(x) \, \mathrm{d}x \\ &= \int_{-\infty}^{\infty} f\big(h(x)\big) \, \ln\big(f\big(h(x)\big)\big) \, \gamma\big(h(x)\big) \, h'(x) \, \mathrm{d}x \\ &= \int_{-\infty}^{\infty} \ln\big(f(h(x))\big) \, \gamma(x) \, \mathrm{d}x, \end{aligned} \tag{3.4.71}$$

where we have used (3.4.70) to get the last equality. From (3.4.70), we have

$$\ln f(h(x)) = \ln\left(\frac{\gamma(x)}{h'(x)\,\gamma\big(h(x)\big)}\right) = \frac{h^2(x) - x^2}{2} - \ln h'(x).$$

Upon substituting this into (3.4.71), we get

$$\begin{aligned} D(\nu\|\mu) &= \frac{1}{2} \int_{-\infty}^{\infty} \Big[h^2(x) - x^2\Big] \, \gamma(x) \, \mathrm{d}x - \int_{-\infty}^{\infty} \ln h'(x) \, \gamma(x) \, \mathrm{d}x \\ &= \frac{1}{2} \int_{-\infty}^{\infty} \big(x - h(x)\big)^2 \, \gamma(x) \, \mathrm{d}x + \int_{-\infty}^{\infty} x\big(h(x) - x\big) \, \gamma(x) \, \mathrm{d}x \\ &\qquad\qquad - \int_{-\infty}^{\infty} \ln h'(x) \, \gamma(x) \, \mathrm{d}x \\ &= \frac{1}{2} \int_{-\infty}^{\infty} \big(x - h(x)\big)^2 \, \gamma(x) \, dx + \int_{-\infty}^{\infty} \big(h'(x) - 1\big) \, \gamma(x) \, \mathrm{d}x \\ &\qquad\qquad - \int_{-\infty}^{\infty} \ln h'(x) \, \gamma(x) \, \mathrm{d}x \\ &\geq \frac{1}{2} \int_{-\infty}^{\infty} \big(x - h(x)\big)^2 \, \gamma(x) \, \mathrm{d}x \\ &= \frac{1}{2} \, W_2^2(\nu, \mu) \end{aligned}$$

where the third line relies on integration by parts, the fourth line follows from the inequality $\ln t \leq t - 1$ for $t > 0$, and the last line holds due

to (3.4.68). This shows that $\mu = G$ satisfies $\mathrm{T}_2(1)$, so the proof of Theorem 3.4.7 for $n = 1$ is complete. Finally, this theorem is generalized for an arbitrary n by tensorization via Proposition 3.7. □

We now get to the main result of this section, namely that dimension-free Gaussian concentration and T_2 are equivalent:

Theorem 3.4.8. Let $(\mathcal{X}, d, \mu)$ be a metric probability space. Then, the following statements are equivalent:

1. μ satisfies $\mathrm{T}_2(c)$.
2. μ has dimension-free Gaussian concentration with constant $\kappa = 1/(2c)$.

Remark 3.27. As we will see, the implication 1) $\Rightarrow$ 2) follows easily from the tensorization property of transportation cost inequalities (Proposition 3.7). The reverse implication 2) $\Rightarrow$ 1) is a nontrivial result, which was proved by Gozlan [64] using an elegant probabilistic approach relying on the theory of large deviations [81].

Proof. We first prove that 1) $\Rightarrow$ 2). Assume that μ satisfies $\mathrm{T}_2(c)$ 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 (3.4.45) with $p = 2$. By the tensorization property of transportation cost inequalities (Proposition 3.7), the product measure $\mu^{\otimes k}$ satisfies $\mathrm{T}_2(c)$ on $(\mathcal{X}^k, d_{2,k})$. Because the L^2 Wasserstein distance dominates the L^1 Wasserstein distance (by item 2 of Lemma 3.4.3), $\mu^{\otimes k}$ also satisfies $\mathrm{T}_1(c)$ on $(\mathcal{X}^k, d_{2,k})$. Therefore, by Proposition 3.6, $\mu^{\otimes k}$ has Gaussian concentration (3.4.7) with respect to $d_{2,k}$ with constants $\kappa = 1/(2c), K = 1, r_0 = \sqrt{2c \ln 2}$. 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) $\Rightarrow$ 1). Suppose that μ has dimension-free Gaussian concentration with constant $\kappa > 0$, where for simplicity we assume that $r_0 = 0$ (the argument for the general case of $r_0 > 0$ is slightly more involved, and does not contribute much in the way of insight). 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.,

$$\mathsf{P}_{x^k} = \frac{1}{k}\sum_{i=1}^{k} \delta_{x_i}, \tag{3.4.72}$$

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 $f_\nu \colon \mathcal{X}^k \to \mathbb{R}$ 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 with respect to $d_{2,k}$ with Lipschitz constant $\frac{1}{\sqrt{k}}$. To verify this, note that

$$\begin{aligned}
|f_\nu(x^k) - f_\nu(y^k)| &= \left|W_2(\mathsf{P}_{x^k}, \nu) - W_2(\mathsf{P}_{y^k}, \nu)\right| \\
&\le W_2(\mathsf{P}_{x^k}, \mathsf{P}_{y^k}) && (3.4.73)\\
&= \inf_{\pi \in \Pi(\mathsf{P}_{x^k}, \mathsf{P}_{y^k})} \left(\int_{\mathcal{X}} d^2(x,y)\, \pi(\mathrm{d}x, \mathrm{d}y)\right)^{1/2} && (3.4.74)\\
&\le \left(\frac{1}{k}\sum_{i=1}^{k} d^2(x_i, y_i)\right)^{1/2} && (3.4.75)\\
&= \frac{1}{\sqrt{k}}\, d_{2,k}(x^k, y^k), && (3.4.76)
\end{aligned}$$

where

- (3.4.73) is by the triangle inequality;
- (3.4.74) is by definition of W_2;
- (3.4.75) uses 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 (3.4.72), the marginals of the above measure are indeed P_{x^k} and P_{y^k}); and
- (3.4.76) uses the definition (3.4.45) of $d_{2,k}$.

Now let us consider the function $f_k(x^k) \triangleq W_2(\mathsf{P}_{x^k}, \mu)$, for which, as we have just seen, we have $\|f_k\|_{\mathrm{Lip},2} \le 1/\sqrt{k}$. Let $X_1, \ldots, X_k$ be i.i.d. draws from μ. Let m_k denote some $\mu^{\otimes k}$-median of f_k. Then, by the

assumed dimension-free Gaussian concentration property of μ, we have from Theorem 3.4.1

$$\mathbb{P}\Big(f_k(X^k) \geq m_k + r\Big) \leq \exp\left(-\frac{\kappa r^2}{\|f_k\|^2_{\mathrm{Lip},2}}\right) \leq \exp\left(-\kappa k r^2\right), \qquad \forall\, r \geq 0; k \in \mathbb{N} \quad (3.4.77)$$

where the second inequality follows from the fact that $\|f_k\|^2_{\mathrm{Lip},2} \leq \frac{1}{k}$.

We now claim that any sequence $\{m_k\}_{k=1}^{\infty}$ of medians of the f_k's converges to zero. 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 [151, Theorem 11.4.1]). Therefore, since W_2 metrizes the topology of weak convergence together with the convergence of second moments (cf. Lemma 3.4.3), $\lim_{k\to\infty} W_2(\mathsf{P}_{X^k}, \mu) = 0$ almost surely. Hence, using the fact that convergence almost surely implies convergence in probability, we have

$$\lim_{k\to\infty} \mathbb{P}(W_2(\mathsf{P}_{X^k}, \mu) \geq t) = 0, \qquad \forall t > 0.$$

Consequently, any sequence $\{m_k\}$ of medians of the f_k's converges to zero, as claimed. Combined with (3.4.77), this implies that

$$\limsup_{k\to\infty} \frac{1}{k} \ln \mathbb{P}\Big(W_2(\mathsf{P}_{X^k}, \mu) \geq r\Big) \leq -\kappa r^2. \quad (3.4.78)$$

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 3.4.3). Consequently, the set $\{\mu : W_2(\mathsf{P}_{X^k}, \mu) > r\}$ is open in the weak topology, so by Sanov's theorem [81, Theorem 6.2.10]

$$\liminf_{k\to\infty} \frac{1}{k} \ln \mathbb{P}\Big(W_2(\mathsf{P}_{X^k}, \mu) \geq r\Big) \geq -\inf\left\{D(\nu\|\mu) : W_2(\mu, \nu) > r\right\}. \quad (3.4.79)$$

Combining (3.4.78) and (3.4.79), we get that

$$\inf\{D(\nu\|\mu) : W_2(\mu, \nu) > r\} \geq \kappa r^2$$

which then implies that $D(\nu\|\mu) \geq \kappa\, W_2^2(\mu, \nu)$. Upon rearranging, we obtain $W_2(\mu, \nu) \leq \sqrt{(\frac{1}{\kappa})\, D(\nu\|\mu)}$, which is a $\mathrm{T}_2(c)$ inequality with $c = \frac{1}{2\kappa}$. This completes the proof of Theorem 3.4.8. □

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 [152], 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 [153]:

Theorem 3.4.9. Let G be the standard Gaussian probability distribution on $\mathbb{R}$. Then, the inequality

$$D(P\|G) \leq W_2(P,G)\sqrt{I(P\|G)}, \tag{3.4.80}$$

where W_2 is the L^2 Wasserstein distance with respect to 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 (3.4.80) is finite.

Proof. We first show the following:

Lemma 3.4.10. 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 $t > 0$,

$$D(P_{X+\sqrt{t}N}\|P_{Y+\sqrt{t}N}) \leq \frac{1}{2t}\, W_2^2(P_X, P_Y). \tag{3.4.81}$$

Proof. Using the chain rule for divergence, we can expand the divergence $D(P_{X,Y,X+\sqrt{t}N} \| P_{X,Y,Y+\sqrt{t}N})$ in two ways:

$$\begin{aligned} &D(P_{X,Y,X+\sqrt{t}N} \| P_{X,Y,Y+\sqrt{t}N}) \\ &= D(P_{X+\sqrt{t}N} \| P_{Y+\sqrt{t}N}) + D(P_{X,Y|X+\sqrt{t}N} \| P_{X,Y|Y+\sqrt{t}N} | P_{X+\sqrt{t}N}) \\ &\geq D(P_{X+\sqrt{t}N} \| P_{Y+\sqrt{t}N}) \end{aligned} \tag{3.4.82}$$

and

$$\begin{aligned} &D(P_{X,Y,X+\sqrt{t}N} \| P_{X,Y,Y+\sqrt{t}N}) \\ &\qquad = D(P_{X+\sqrt{t}N}|X,Y \,\|\, P_{Y+\sqrt{t}N}|X,Y|P_{X,Y}) \\ &\qquad \stackrel{\text{(a)}}{=} \mathbb{E}[D(\mathcal{N}(X,t) \,\|\, \mathcal{N}(Y,t)) \,|\, X,Y] \\ &\qquad \stackrel{\text{(b)}}{=} \frac{1}{2t}\,\mathbb{E}[(X-Y)^2]. \end{aligned} \tag{3.4.83}$$

Note that equality (a) holds since N is independent of (X,Y), and equality (b) is a special case of the equality

$$D\big(\mathcal{N}(m_1,\sigma_1^2) \,\|\, \mathcal{N}(m_2,\sigma_2^2)\big) = \frac{1}{2}\,\ln\left(\frac{\sigma_2^2}{\sigma_1^2}\right) + \frac{1}{2}\left(\frac{(m_1-m_2)^2}{\sigma_2^2} + \frac{\sigma_1^2}{\sigma_2^2} - 1\right).$$

It therefore follows from (3.4.82) and (3.4.83) that

$$D(P_{X+\sqrt{t}N} \| P_{Y+\sqrt{t}N}) \leq \frac{1}{2t}\,\mathbb{E}[(X-Y)^2] \tag{3.4.84}$$

where the left-hand side of this inequality only depends on the marginal distributions of X and Y (due to the independence of (X,Y) and $N \sim G$). Hence, taking the infimum of the right-hand side of (3.4.84) w.r.t. all couplings of P_X and P_Y (i.e., all $\mu \in \Pi(P_X, P_Y)$), we get (3.4.81) (see (3.4.20)). □

We now proceed with the proof of Theorem 3.4.9. Let X have distribution P, and Y have distribution G. For simplicity, we focus on the case when X has zero mean and unit variance; the general case can be handled similarly. We define the function

$$F(t) \triangleq D(P_{X+\sqrt{t}Z} \| P_{Y+\sqrt{t}Z}), \quad \forall\, t > 0$$

where $Z \sim G$ is independent of (X, Y). Then $F(0) = D(P\|G)$, and from (3.4.81) we have

$$F(t) \le \frac{1}{2t} W_2^2(P_X, P_Y) = \frac{1}{2t} W_2^2(P, G). \tag{3.4.85}$$

Moreover, the function $F(t)$ is differentiable, and it follows from [125, Eq. (32)] that

$$F'(t) = \frac{1}{2t^2}[\mathsf{mmse}(X, t^{-1}) - \mathsf{lmmse}(X, t^{-1})] \tag{3.4.86}$$

where $\mathsf{mmse}(X, \cdot)$ and $\mathsf{lmmse}(X, \cdot)$ have been defined in (3.2.24) and (3.2.26), respectively. For any $t > 0$,

$$\begin{aligned}
D(P||G) &= F(0) \\
&= -\big(F(t) - F(0)\big) + F(t) \\
&= -\int_0^t F'(s)\mathrm{d}s + F(t) \\
&= \frac{1}{2}\int_0^t \frac{1}{s^2}\left(\mathsf{lmmse}(X, s^{-1}) - \mathsf{mmse}(X, s^{-1})\right)\mathrm{d}s + F(t) && (3.4.87) \\
&\le \frac{1}{2}\int_0^t \left(\frac{1}{s(s+1)} - \frac{1}{s(sJ(X)+1)}\right)\mathrm{d}s + \frac{1}{2t} W_2^2(P, G) && (3.4.88) \\
&= \frac{1}{2}\left(\ln \frac{tJ(X)+1}{t+1} + \frac{W_2^2(P, G)}{t}\right) && (3.4.89) \\
&\le \frac{1}{2}\left(\frac{t(J(X)-1)}{t+1} + \frac{W_2^2(P, G)}{t}\right) && (3.4.90) \\
&\le \frac{1}{2}\left(I(P\|G)\, t + \frac{W_2^2(P, G)}{t}\right) && (3.4.91)
\end{aligned}$$

where

- (3.4.87) uses (3.4.86);
- (3.4.88) uses (3.2.27), the Van Trees inequality (3.2.28), and (3.4.85);
- (3.4.89) is an exercise in calculus;

- (3.4.90) uses the inequality $\ln x \leq x - 1$ for $x > 0$; and
- (3.4.91) uses the formula (3.2.21) (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 (3.2.21) to get $G_s = G$), and the fact that $t \geq 0$.

Optimizing the choice of t in (3.4.91), we get (3.4.80). □

Remark 3.28. Note that the HWI inequality (3.4.80) together with the T_2 inequality for the Gaussian distribution imply a weaker version of the log-Sobolev inequality (3.2.8) (i.e., with a larger constant). Indeed, using the T_2 inequality of Theorem 3.4.7 on the right-hand side of (3.4.80), we get

$$\begin{aligned} D(P\|G) &\leq W_2(P,G)\sqrt{I(P\|G)} \\ &\leq \sqrt{2D(P\|G)}\sqrt{I(P\|G)}, \end{aligned}$$

which gives $D(P\|G) \leq 2I(P\|G)$. It is not surprising that we end up with a suboptimal constant here as compared to (3.2.8): the series of bounds leading up to (3.4.91) contributes a lot more slack than the single use of the van Trees inequality (3.2.28) in our proof of Stam's inequality (which, due to Proposition 3.4, is equivalent to the Gaussian log-Sobolev inequality of Gross).

We are now ready to state the HWI inequality in its strong form:

Theorem 3.4.11 (Otto–Villani [152])**.** Let P be a Borel probability measure on $\mathbb{R}^n$ that is absolutely continuous with respect to the Lebesgue measure, and let the corresponding pdf p be such that

$$\nabla^2 \ln\left(\frac{1}{p}\right) \succeq K I_n \tag{3.4.92}$$

for some $K \in \mathbb{R}$ (where ∇^2 denotes the Hessian, and the matrix inequality $A \succeq B$ means that $A - B$ is non-negative semidefinite). Then, any probability measure $Q \ll P$ satisfies

$$D(Q\|P) \leq W_2(Q,P)\sqrt{I(Q\|P)} - \frac{K}{2}\, W_2^2(Q,P). \tag{3.4.93}$$

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 with respect to the Euclidean norm in $\mathbb{R}^n$. (An alternative simpler proof was given later by Cordero–Erausquin [154].) We can, however, highlight a couple of key consequences (see [152]):

1. Suppose that P, in addition to satisfying the conditions of Theorem 3.4.11, also satisfies a $\mathrm{T}_2(c)$ inequality. Using this fact in (3.4.93), we get

$$D(Q\|P) \le \sqrt{2cD(Q\|P)}\sqrt{I(Q\|P)} - \frac{K}{2}\, W_2^2(Q,P). \quad (3.4.94)$$

   If the pdf p of P is log-concave, so that (3.4.92) holds with $K = 0$, then (3.4.94) implies the inequality

$$D(Q\|P) \le 2c\, I(Q\|P) \quad (3.4.95)$$

   for any $Q \ll P$. This is an Euclidean log-Sobolev inequality that is similar to the one satisfied by $P = G^n$ (see Remark 3.28). However, note that the constant in front of the Fisher information distance $I(\cdot\|\cdot)$ on the right-hand side of (3.4.95) is suboptimal, as can be verified by letting $P = G^n$, which satisfies $\mathrm{T}_2(1)$; going through the above steps, as we know from Section 3.2 (in particular, see (3.2.8)), the optimal constant should be $\frac{1}{2}$, so the one in (3.4.95) 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 (3.4.92) holds with some $K > 0$, then P satisfies the Euclidean log-Sobolev inequality with constant $\frac{1}{K}$. Indeed, using Young's inequality $ab \le \frac{a^2+b^2}{2}$, we have from (3.4.93)

$$\begin{aligned} D(Q\|P) &\le \sqrt{K} W_2(Q,P)\sqrt{\frac{I(Q\|P)}{K}} - \frac{K}{2}\, W_2^2(Q,P) \\ &\le \frac{1}{2K}\, I(Q\|P), \end{aligned}$$

which shows that P satisfies the Euclidean LSI($\frac{1}{K}$) inequality. In particular, the standard Gaussian distribution $P = G^n$ satisfies (3.4.92) with $K = 1$, so we even get the right constant. In fact, the statement that (3.4.92) with $K > 0$ implies Euclidean LSI($\frac{1}{K}$) was first proved in 1985 by Bakry and Emery [155] 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 transportation cost inequalities for dependent random variables

Samson [156] 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 $(X_1, \ldots, X_n)$, 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 $i < j$

$$\Delta_{i,j} = \sup_{x_i, x_i'} \sup_{x^{i-1}} \sqrt{\left\| P_{X_j^n | X_i = x_i, X^{i-1} = x^{i-1}} - P_{X_j^n | X_i = x_i', X^{i-1} = x^{i-1}} \right\|_{\mathrm{TV}}}. \tag{3.5.1}$$

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, i.e.,

$$\|\Delta\| \triangleq \sup_{v\in\mathbb{R}^n:\, v\neq 0} \frac{\|\Delta v\|}{\|v\|} = \sup_{v\in\mathbb{R}^n:\, \|v\|=1} \|\Delta v\|.$$

Following Marton [157], 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 $\alpha = (\alpha_1,\ldots,\alpha_n) : \mathcal{X}^n \to \mathbb{R}^n$, such that

$$\mathbb{E}_Q\left[\|\alpha(Y^n)\|^2\right] \leq 1.$$

The main result of [156] goes as follows:

Theorem 3.5.1. The probability distribution P of X^n satisfies the following transportation cost inequality:

$$d_2(Q,P) \leq \|\Delta\|\sqrt{2D(Q\|P)} \tag{3.5.2}$$

for all $Q \ll P$.

Let us examine some implications:

1. Let $\mathcal{X} = [0,1]$. Then Theorem 3.5.1 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 $f\colon [0,1]^n \to \mathbb{R}$,

$$D\big(P^{(f)}\|P\big) \leq 2\|\Delta\|^2\, \mathbb{E}^{(f)}\left[\|\nabla f(X^n)\|^2\right] \tag{3.5.3}$$

(see [156, Corollary 1]). The same method as the one we used to prove Proposition 3.5 and Theorem 3.2.2 can be applied to obtain from (3.5.3) the following concentration inequality for any convex function $f\colon [0,1]^n \to \mathbb{R}$ with $\|f\|_{\mathrm{Lip}} \leq 1$:

$$\mathbb{P}\Big(f(X^n) \geq \mathbb{E}f(X^n) + r\Big) \leq \exp\left(-\frac{r^2}{2\|\Delta\|^2}\right), \qquad \forall r \geq 0. \tag{3.5.4}$$

2. While (3.5.2) and its corollaries, (3.5.3) and (3.5.4), 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 $X_1, \ldots, X_n$ are independent, then $\Delta = I_{n\times n}$. In this case, (3.5.2) 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 (3.5.1), 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 [156, 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 [157], 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

$$\|\Delta\| \leq \frac{1}{1-\delta^{1/2}}, \qquad \text{where } \delta \triangleq \max_{i=1,\ldots,n} \delta_i.$$

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 [158, 159], 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 [158, 159] are set in the Euclidean space $\mathbb{R}^n$, and center around a transportation cost inequality for the L^2 Wasserstein distance

$$W_2(P,Q) \triangleq \inf_{X^n \sim P, Y^n \sim Q} \sqrt{\mathbb{E}\|X^n - Y^n\|^2}, \tag{3.5.5}$$

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 with respect to 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 [158], we say that P is δ-*contractive*, with $0 < \delta < 1$, if for any $y^n, z^n \in \mathbb{R}^n$

$$\sum_{i=1}^n W_2^2(P_{X_i|\bar{X}^i=\bar{y}^i}, P_{X_i|\bar{X}^i=\bar{z}^i}) \leq (1-\delta)\|y^n - z^n\|_2. \tag{3.5.6}$$

Remark 3.29. Marton's contractivity condition (3.5.6) is closely related to the so-called *Dobrushin–Shlosman mixing condition* from mathematical statistical physics.

Theorem 3.5.2 (Marton [158, 159])**.** Suppose that P is absolutely continuous with respect to 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 $\mathrm{T}_2(c)$ with respect to the L^2 Wasserstein distance (3.5.5) (cf. Definition 3.2)

Then for any probability measure Q on $\mathbb{R}^n$ we have

$$W_2(Q, P) \le \left(\frac{K}{\sqrt{\delta}} + 1\right) \sqrt{2cD(Q\|P)}, \tag{3.5.7}$$

where $K > 0$ is an absolute constant. In other words, any P satisfying the conditions of the theorem admits a $\mathrm{T}_2(c')$ inequality with $c' = (K/\sqrt{\delta} + 1)^2 c$.

The contractivity criterion (3.5.6) is not easy to verify in general. Let us mention one sufficient condition [158]. Let p denote the probability density of P, and suppose that it takes the form

$$p(x^n) = \frac{1}{Z} \exp\left(-\Psi(x^n)\right) \tag{3.5.8}$$

for some C^2 function $\Psi \colon \mathbb{R}^n \to \mathbb{R}$, 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

$$B_{ij}(x^n, y^n) \triangleq \begin{cases} \nabla^2_{ij} \Psi(x_i \odot \bar{y}^i), & i \ne j \\ 0, & i = j \end{cases} \tag{3.5.9}$$

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 [158]) the conditions of Theorem 3.5.2 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

$$D(Q\|P_{X_i|\bar{X}^i=\bar{x}^i}) \leq \frac{c}{2} I(Q\|P_{X_i|\bar{X}^i=\bar{x}^i}),$$

   where $I(\cdot\|\cdot)$ is the Fisher information distance, cf. (3.2.4) for the definition.

2. The operator norms of $B(x^n, y^n)$ are uniformly bounded as

$$\sup_{x^n, y^n} \|B(x^n, y^n)\|^2 \leq \frac{1-\delta}{c^2}.$$

We also refer the reader to more recent follow-up work by Marton [160, 161], which further elaborates on the theme of studying the concentration properties of dependent random variables by focusing on the conditional probability distributions $P_{X_i|\bar{X}^i}$, $i = 1, \ldots, n$. These papers describe sufficient conditions on the joint distribution P of $X_1, \ldots, X_n$, such that, for any other distribution Q,

$$D(Q\|P) \leq K(P) \cdot D^-(Q\|P), \tag{3.5.10}$$

where $D^-(\cdot\|\cdot)$ is the erasure divergence (cf. (3.1.21) 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 (3.5.10) 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

The first explicit invocation of the concentration of measure phenomenon in an information-theoretic context appears in the work of Ahlswede et al. [69, 70]. These authors have shown that the following result, now known as the "blowing up lemma" (see, e.g., [162, Lemma 1.5.4]), provides a versatile tool for proving strong converses in a variety of scenarios, including some multiterminal problems:

Lemma 3.6.1. 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 $B \subseteq \mathcal{Y}^n$,

$$T^n(B|x^n) \geq \exp(-n\varepsilon_n) \qquad \Longrightarrow \qquad T^n(B_{n\delta_n}|x^n) \geq 1 - \eta_n. \quad (3.6.1)$$

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 (3.4.5)) w.r.t. the Hamming metric

$$d_n(y^n, u^n) \triangleq \sum_{i=1}^{n} 1_{\{y_i \neq u_i\}}, \qquad \forall y^n, u^n \in \mathcal{Y}^n.$$

The proof of the blowing-up lemma, given in [69], was rather technical and made use of a delicate isoperimetric inequality for discrete probability measures on a Hamming space, due to Margulis [163]. Later, the same result was obtained by Marton [71] 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 [58]):

Lemma 3.6.2. 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 $P_{X^n}(A) > 0$,

$$P_{X^n}(A_r) \geq 1 - \exp\left[-\frac{2}{n}\left(r - \sqrt{\frac{n}{2}\ln\left(\frac{1}{P_{X^n}(A)}\right)}\right)^2\right], \quad \forall\, r > \sqrt{\frac{n}{2}\ln\left(\frac{1}{P_{X^n}(A)}\right)}. \tag{3.6.2}$$

Proof. 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 $\mathrm{T}_1(1/4)$ on $(\mathcal{X}, d)$ where $d = d_1$ is the Hamming metric. By Proposition 3.8, the product measure P_n satisfies $\mathrm{T}_1(n/4)$ on the product space $(\mathcal{X}^n, d_n)$, i.e., for any $\mu_n \in \mathcal{P}(\mathcal{X}^n)$,

$$W_1(\mu_n, P_n) \leq \sqrt{\frac{n}{2}\, D(\mu_n \| P_n)}. \tag{3.6.3}$$

The statement of the lemma then follows from Proposition 3.6. □

We can now easily prove Lemma 3.6.1. 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 (3.6.2), we get (3.6.1) 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 3.6.1 and 3.6.2 is given by $P_{X^n} = T^n(\cdot|x^n)$ where $x^n \in \mathcal{X}^n$ is arbitrary).

3.6.2 Strong converse for the degraded broadcast channel

We are now ready to demonstrate how the blowing-up lemma can be used to obtain strong converses. Following [162], from this point on, we will use the notation $T\colon \mathcal{U} \to \mathcal{V}$ for a DMC with input alphabet $\mathcal{U}$, output alphabet $\mathcal{V}$, and transition probabilities $T(v|u), u \in \mathcal{U}, v \in \mathcal{V}$.

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 $T_1\colon \mathcal{X} \to \mathcal{Y}$ and $T_2\colon \mathcal{X} \to \mathcal{Z}$ where there exists a DMC $T_3\colon \mathcal{Y} \to \mathcal{Z}$ such that

$$\mathrm{T}_2(z|x) = \sum_{y \in \mathcal{Y}} T_3(z|y)\mathrm{T}_1(y|x), \qquad \forall x \in \mathcal{X}, z \in \mathcal{Z}. \tag{3.6.4}$$

(More precisely, this is an instance of a *stochastically degraded* broadcast channel – see, e.g., [133, Section 5.6] and [164, 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* $f_n\colon \{1, \ldots, M_1\} \times \{1, \ldots, M_2\} \to \mathcal{X}^n$;
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 with respect to the maximal probability of error is strictly smaller than the capacity region with respect to the average probability of error [165], these two capacity regions are identical for broadcast channels [166]. 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

$$\frac{1}{n} \ln M_k \geq R_k - \delta, \qquad k = 1, 2.$$

Likewise, we say that (R_1, R_2) is *achievable* if it is $(\varepsilon_1, \varepsilon_2)$-achievable for all $0 < \varepsilon_1, \varepsilon_2 \leq 1$. 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 [167]:

Theorem 3.6.3. 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 (3.6.4)), and

$$R_1 \leq I(X; Y|U), \quad R_2 \leq I(U; Z).$$

Moreover, the domain $\mathcal{U}$ of U can be chosen so that $|\mathcal{U}| \leq \min\{|\mathcal{X}|, |\mathcal{Y}|, |\mathcal{Z}|\}$.

The *strong converse* for the DBC, due to Ahlswede, Gács and Körner [69], states that allowing for nonvanishing probabilities of error does not enlarge the achievable region:

Theorem 3.6.4 (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 3.6.4). 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 $i \in \{1, \ldots, M_1\}$ and $j \in \{1, \ldots, M_2\}$, 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 3.6.2, we can find a sequence $\varepsilon_n \to 0$, such that

$$\min_{1\le i\le M_1} \min_{1\le j\le M_2} T_1^n\Big(D_{1,i}\Big|f_n(i,j)\Big) \ge 1-\varepsilon_n \tag{3.6.5a}$$
$$\min_{1\le i\le M_1} \min_{1\le j\le M_2} T_2^n\Big(D_{2,j}\Big|f_n(i,j)\Big) \ge 1-\varepsilon_n. \tag{3.6.5b}$$

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 (3.6.5). 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 [69, 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

$$|\mathcal{N}_1(y^n)| \leq |\mathcal{B}_{n\delta_n}(y^n)| \leq \exp(n\eta_n), \qquad \forall y^n \in \mathcal{Y}^n \tag{3.6.6a}$$

$$|\mathcal{N}_2(z^n)| \leq |\mathcal{B}_{n\delta_n}(z^n)| \leq \exp(n\eta_n), \qquad \forall z^n \in \mathcal{Z}^n \tag{3.6.6b}$$

where, for any $y^n \in \mathcal{Y}^n$ and any $r \geq 0$, $\mathcal{B}_r(y^n) \subseteq \mathcal{Y}^n$ denotes the ball of d_n-radius r centered at y^n:

$$\mathcal{B}_r(y^n) \triangleq \{v^n \in \mathcal{Y}^n : d_n(v^n, y^n) \leq r\} \equiv \{y^n\}_r$$

(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 [167]. Specifically, let U have a uniform distribution over $\{1, \ldots, M_2\}$, and let $X^n \in \mathcal{X}^n$ have a uniform distribution over the set $\mathcal{T}(U)$, where for each $j \in \{1, \ldots, M_2\}$ we let

$$\mathcal{T}(j) \triangleq \{f_n(i, j) : 1 \leq i \leq M_1\}.$$

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.E) together with (3.6.6), we get

$$H(U|Z^n) \leq h(\zeta_n) + (1 - \zeta_n)n\eta_n + \zeta_n \ln M_2. \tag{3.6.7}$$

On the other hand, $\ln M_2 = H(U) = I(U; Z^n) + H(U|Z^n)$, so

$$\begin{aligned}\frac{1}{n}\ln M_2 &\leq \frac{1}{n}\Big[I(U; Z^n) + h(\zeta_n) + \zeta_n \ln M_2\Big] + (1 - \zeta_n)\eta_n \\ &= \frac{1}{n}I(U; Z^n) + o(1),\end{aligned}$$

where the second step uses the fact that, by (3.6.5), $\zeta_n \leq \varepsilon_n$, which converges to zero. Using a similar argument, we can also prove that

$$\frac{1}{n}\ln M_1 \leq \frac{1}{n}I(X^n; Y^n|U) + o(1).$$

By the weak converse for the DBC [167], 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 $\mathcal{R}(\varepsilon_1, \varepsilon_2) \subseteq \mathcal{R}$ for all $\varepsilon_1, \varepsilon_2 \in (0, 1)$, and Theorem 3.6.4 is proved. □

3.6.3 The empirical distribution of good channel codes with non-vanishing 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 [168], who have proved the following remarkable statement [168, Theorem 2]: given a DMC $T\colon \mathcal{X} \to \mathcal{Y}$, any capacity-achieving sequence of channel codes with asymptotically vanishing probability of error (maximal or average) has the property that

$$\lim_{n\to\infty} \frac{1}{n} D(P_{Y^n} \| P^*_{Y^n}) = 0, \tag{3.6.8}$$

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 (3.6.8) holds not just for DMC's, but for arbitrary channels satisfying the condition

$$C = \lim_{n\to\infty} \frac{1}{n} \sup_{P_{X^n} \in \mathcal{P}(\mathcal{X}^n)} I(X^n; Y^n).$$

(These ideas go back to the work of Han and Verdú on approximation theory of output statistics, see Theorem 15 in [169]). In a recent

preprint [170], Polyanskiy and Verdú extended the results of [168] for codes with *nonvanishing* probability of error, provided one uses the maximal probability of error criterion and deterministic encoders.

In this section, we will present some of the results from [170] 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 $T\colon \mathcal{X} \to \mathcal{Y}$. 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 [171, 172] 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

$$D(P_{Y^n|X^n=x^n} \| P_{Y^n}^*) \leq nC, \qquad \forall x^n \in \mathcal{X}^n \tag{3.6.9}$$

where $P_{Y^n|X^n=x^n}$ is shorthand for the product distribution $T^n(\cdot|x^n)$. From the bound (3.6.9), 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 $y^n \in \mathcal{Y}^n$:

$$i_{X^n;Y^n}^*(x^n;y^n) \triangleq \ln \frac{\mathrm{d}P_{Y^n|X^n=x^n}}{\mathrm{d}P_{Y^n}^*}(y^n).$$

2. For any input distribution P_{X^n}, the corresponding output distribution P_{Y^n} satisfies

$$D(P_{Y^n} \| P_{Y^n}^*) \leq nC - I(X^n;Y^n).$$

Indeed, by the chain rule for divergence, it follows that for any input distribution $P_{X^n} \in \mathcal{P}(\mathcal{X}^n)$

$$\begin{aligned} I(X^n;Y^n) &= D(P_{Y^n|X^n}\|P_{Y^n}|P_{X^n}) \\ &= D(P_{Y^n|X^n}\|P^*_{Y^n}|P_{X^n}) - D(P_{Y^n}\|P^*_{Y^n}) \\ &\le nC - D(P_{Y^n}\|P^*_{Y^n}). \end{aligned}$$

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* $f_n \colon \{1, \ldots, M\} \to \mathcal{X}^n$ and a *decoding map* $g_n \colon \mathcal{Y}^n \to \{1, \ldots, M\}$. Given $0 < \varepsilon \le 1$, we say that $\mathcal{C}$ is an (n, M, ε)-*code* if

$$\max_{1 \le i \le M} \mathbb{P}\big(g_n(Y^n) \neq i \big| X^n = f_n(i)\big) \le \varepsilon. \tag{3.6.10}$$

Remark 3.30. Polyanskiy and Verdú [170] use a more precise nomenclature and say that any such $\mathcal{C} = (f_n, g_n)$ satisfying (3.6.10) is an $(n, M, \varepsilon)_{\max,\det}$-*code* to indicate explicitly that the encoding map f_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 $\{1, \ldots, M\}$. 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 [170] is that the output distribution $P^{(\mathcal{C})}_{Y^n}$ of any (n, M, ε)-code satisfies

$$D(P^{(\mathcal{C})}_{Y^n}\|P^*_{Y^n}) \le nC - \ln M + o(n); \tag{3.6.11}$$

moreover, the $o(n)$ term was refined in [170, Theorem 5] to $O(\sqrt{n})$ for any DMC, except those that have zeroes in their transition matrix. In the following, we present a sharpened bound with a modified proof, in which we specify an explicit form for the term that scales like $O(\sqrt{n})$.

Just as in [170], the proof of (3.6.11) with the $O(\sqrt{n})$ term uses the following strong converse for channel codes due to Augustin [173] (see also [170, Theorem 1] and [174, Section 2]):

Theorem 3.6.5 (Augustin)**.** Let $S\colon \mathcal{U} \to \mathcal{V}$ 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 $0 < \varepsilon \leq 1$, let $f\colon \{1,\dots,M\} \to \mathcal{U}$ and $g\colon \mathcal{V} \to \{1,\dots,M\}$ be two mappings, such that

$$\max_{1\leq i\leq M} \mathbb{P}\big(g(V) \neq i \big| U = f(i)\big) \leq \varepsilon.$$

Let $Q_V \in \mathcal{P}(\mathcal{V})$ be an auxiliary output distribution, and fix an arbitrary mapping $\gamma\colon \mathcal{U} \to \mathbb{R}$. Then, the following inequality holds:

$$M \leq \frac{\exp\{\mathbb{E}[\gamma(U)]\}}{\inf\limits_{u\in\mathcal{U}} P_{V|U=u}\left(\ln \dfrac{\mathrm{d}P_{V|U=u}}{\mathrm{d}Q_V} < \gamma(u)\right) - \varepsilon}, \tag{3.6.12}$$

provided the denominator is strictly positive. The expectation in the numerator is taken with respect to the distribution of $U = f(J)$ with $J \sim \mathrm{Uniform}\{1,\dots,M\}$.

We first establish the bound (3.6.11) for the case when the DMC T is such that

$$C_1 \triangleq \max_{x,x'\in\mathcal{X}} D(P_{Y|X=x} \| P_{Y|X=x'}) < \infty. \tag{3.6.13}$$

Note that $C_1 < \infty$ if and only if the transition matrix of T does not have any zeroes. Consequently,

$$c(T) \triangleq 2 \max_{x\in\mathcal{X}} \max_{y,y'\in\mathcal{Y}} \left|\ln \frac{P_{Y|X}(y|x)}{P_{Y|X}(y'|x)}\right| < \infty. \tag{3.6.14}$$

We can now establish the following sharpened version of the bound in [170, Theorem 5]:

Theorem 3.6.6. Let $T\colon \mathcal{X} \to \mathcal{Y}$ be a DMC with $C > 0$ satisfying (3.6.13). Then, any (n, M, ε)-code $\mathcal{C}$ for T with $0 < \varepsilon < 1/2$ satisfies

$$D\big(P_{Y^n}^{(\mathcal{C})} \big\| P_{Y^n}^*\big) \leq nC - \ln M + \ln\frac{1}{\varepsilon} + c(T)\sqrt{\frac{n}{2}\ln\frac{1}{1-2\varepsilon}}. \tag{3.6.15}$$

Remark 3.31. As shown in [170], the restriction to codes with deterministic encoders 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

$$h_{x^n}(y^n) \triangleq \ln \frac{\mathrm{d}P_{Y^n|X^n=x^n}}{\mathrm{d}P_{Y^n}^{(\mathcal{C})}}(y^n).$$

Then $\mathbb{E}[h_{x^n}(Y^n)|X^n = x^n] = D(P_{Y^n|X^n=x^n} \| P_{Y^n}^{(\mathcal{C})})$. Moreover, for any $i \in \{1, \ldots, n\}$, $y, y' \in \mathcal{Y}$, and $\overline{y}^i \in \mathcal{Y}^{n-1}$, we have (see the notation used in (3.1.23))

$$\begin{aligned}
&\left| h_{i,x^n}(y|\overline{y}^i) - h_{i,x^n}(y'|\overline{y}^i) \right| \\
&\leq \left| \ln P_{Y^n|X^n=x^n}(y^{i-1}, y, y_{i+1}^n) - \ln P_{Y^n|X^n=x^n}(y^{i-1}, y', y_{i+1}^n) \right| \\
&\qquad + \left| \ln P_{Y^n}^{(\mathcal{C})}(y^{i-1}, y, y_{i+1}^n) - \ln P_{Y^n}^{(\mathcal{C})}(y^{i-1}, y', y_{i+1}^n) \right| \\
&\leq \left| \ln \frac{P_{Y_i|X_i=x_i}(y)}{P_{Y_i|X_i=x_i}(y')} \right| + \left| \ln \frac{P^{(\mathcal{C})}_{Y_i|\overline{Y}^i}(y|\overline{y}^i)}{P^{(\mathcal{C})}_{Y_i|\overline{Y}^i}(y'|\overline{y}^i)} \right| \\
&\leq 2 \max_{x \in \mathcal{X}} \max_{y,y' \in \mathcal{Y}} \left| \ln \frac{P_{Y|X}(y|x)}{P_{Y|X}(y'|x')} \right| \qquad (3.6.16) \\
&= c(T) < \infty \qquad (3.6.17)
\end{aligned}$$

(see Appendix 3.F for a detailed derivation of the inequality in (3.6.16)). 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 (3.3.35) with $c_1 = \ldots = c_n = c(T)$. Theorem 3.3.8 therefore implies that, for any $r \geq 0$, we have

$$\begin{aligned}
P_{Y^n|X^n=x^n} &\left(\ln \frac{\mathrm{d}P_{Y^n|X^n=x^n}}{\mathrm{d}P_{Y^n}^{(\mathcal{C})}}(Y^n) \geq D(P_{Y^n|X^n=x^n} \| P_{Y^n}^{(\mathcal{C})}) + r \right) \\
&\leq \exp\left(-\frac{2r^2}{nc^2(T)} \right). \qquad (3.6.18)
\end{aligned}$$

(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 [170]: 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 (3.6.18) allows us to explicitly identify the dependence of the constant multiplying $\sqrt{n}$ in (3.6.15) 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

$$\zeta_n = \zeta_n(\varepsilon) \triangleq c(T)\sqrt{\frac{n}{2}\ln\frac{1}{1-2\varepsilon}} \tag{3.6.19}$$

and take $\gamma(x^n) = D(P_{Y^n|X^n=x^n}\|P_{Y^n}^{(\mathcal{C})}) + \zeta_n$. Using (3.6.12) with the auxiliary distribution $Q_V = P_{Y^n}^{(\mathcal{C})}$, we get

$$M \leq \frac{\exp\{\mathbb{E}[\gamma(X^n)]\}}{\inf\limits_{x^n\in\mathcal{X}^n} P_{Y^n|X^n=x^n}\left(\ln\dfrac{\mathrm{d}P_{Y^n|X^n=x^n}}{\mathrm{d}P_{Y^n}^{(\mathcal{C})}} < \gamma(x^n)\right) - \varepsilon} \tag{3.6.20}$$

where

$$\mathbb{E}[\gamma(X^n)] = D(P_{Y^n|X^n}\|P_{Y^n}^{(\mathcal{C})} \,|\, P_{X^n}^{(\mathcal{C})}) + \zeta_n. \tag{3.6.21}$$

The concentration inequality in (3.6.18) with ζ_n in (3.6.19) 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 (3.6.20), (3.6.21) and the last inequality, it follows that

$$M \leq \frac{1}{\varepsilon}\,\exp\Big(D(P_{Y^n|X^n}\|P_{Y^n}^{(\mathcal{C})} \,|\, P_{X^n}^{(\mathcal{C})}) + \zeta_n\Big)$$

so, by taking logarithms on both sides of the last inequality and rearranging terms, we get from (3.6.19) that

$$D(P_{Y^n|X^n}\|P_{Y^n}^{(\mathcal{C})} \,|\, P_{X^n}^{(\mathcal{C})}) \geq \ln M + \ln\varepsilon - \zeta_n$$
$$= \ln M + \ln\varepsilon - c(T)\sqrt{\frac{n}{2}\ln\frac{1}{1-2\varepsilon}}. \quad (3.6.22)$$

We are now ready to derive (3.6.15):

$$D(P_{Y^n}^{(\mathcal{C})}\|P_{Y^n}^*)$$
$$= D(P_{Y^n|X^n}\|P_{Y^n}^*|P_{X^n}^{(\mathcal{C})}) - D(P_{Y^n|X^n}\|P_{Y^n}^{(\mathcal{C})}|P_{X^n}^{(\mathcal{C})}) \quad (3.6.23)$$
$$\leq nC - \ln M + \ln\frac{1}{\varepsilon} + c(T)\sqrt{\frac{n}{2}\ln\frac{1}{1-2\varepsilon}} \quad (3.6.24)$$

where (3.6.23) uses the chain rule for divergence, while (3.6.24) uses (3.6.9) and (3.6.22). This completes the proof of Theorem 3.6.6. □

For an arbitrary DMC T with nonzero capacity and zeroes in its transition matrix, we have the following result which forms a sharpened version of the bound in [170, Theorem 6]:

Theorem 3.6.7. Let $T\colon \mathcal{X} \to \mathcal{Y}$ be a DMC with $C > 0$. Then, for any $0 < \varepsilon < 1$, any (n, M, ε)-code $\mathcal{C}$ for T satisfies

$$D(P_{Y^n}^{(\mathcal{C})}\|P_{Y^n}^*) \leq nC - \ln M + O\left(\sqrt{n}\,(\ln n)^{3/2}\right).$$

More precisely, for any such code we have

$$D(P_{Y^n}^{(\mathcal{C})}\|P_{Y^n}^*) \leq nC - \ln M$$
$$+ \sqrt{2n}\,(\ln n)^{3/2}\left(1 + \sqrt{\frac{1}{\ln n}\,\ln\left(\frac{1}{1-\varepsilon}\right)}\right)\left(1 + \frac{\ln|\mathcal{Y}|}{\ln n}\right)$$
$$+ 3\ln n + \ln(2|\mathcal{X}||\mathcal{Y}|^2). \quad (3.6.25)$$

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:

$$\widetilde{D}_i = g_n^{-1}(i) \equiv \{y^n \in \mathcal{Y}^n : g_n(y^n) = i\}, \qquad i = 1, \ldots, M.$$

If we choose

$$\delta_n = \delta_n(\varepsilon) = \frac{1}{n}\left\lceil n\left(\sqrt{\frac{\ln n}{2n}} + \sqrt{\frac{1}{2n}\ln\frac{1}{1-\varepsilon}}\right)\right\rceil \tag{3.6.26}$$

(note that $n\delta_n$ is an integer), then by Lemma 3.6.2 the "blown-up" decoding regions $D_i \triangleq \left[\widetilde{D}_i\right]_{n\delta_n}$ satisfy

$$\begin{aligned} P_{Y^n|X^n=c_i}(D_i^c) &\le \exp\left[-2n\left(\delta_n - \sqrt{\frac{1}{2n}\ln\frac{1}{1-\varepsilon}}\right)^2\right] \\ &\le \frac{1}{n}, \qquad\qquad \forall\, i \in \{1,\ldots,M\}. \end{aligned} \tag{3.6.27}$$

We now complete the proof by a random coding argument. For

$$N \triangleq \left\lceil \frac{M}{n\binom{n}{n\delta_n}|\mathcal{Y}|^{n\delta_n}} \right\rceil, \tag{3.6.28}$$

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 $Y^n(V)$:

$$P_{Y^n(V)} = \frac{1}{N}\sum_{i=1}^{N} P_{Y^n|X^n=c_{U_i}}. \tag{3.6.29}$$

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 with respect to 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 $y^n \in D_{U_j}$:

$$\mathcal{N}_V(y^n) \triangleq \left\{j : y^n \in D_{U_j}\right\}.$$

Consider the list decoder $Y^n \mapsto \mathcal{N}_V(Y^n)$, and let $\varepsilon(V)$ denote its conditional decoding error probability: $\varepsilon(V) \triangleq P(J \notin \mathcal{N}_V(Y^n)|V)$. From (3.6.28), it follows that

$$\begin{aligned} \ln N &\ge \ln M - \ln n - \ln\binom{n}{n\delta_n} - n\delta_n \ln|\mathcal{Y}| \\ &\ge \ln M - \ln n - n\delta_n(\ln n + \ln|\mathcal{Y}|) \end{aligned} \tag{3.6.30}$$

where the last inequality uses the simple inequality $\binom{n}{k} \leq n^k$ for $k \leq n$ with $k \triangleq n\delta_n$ (we note that any gain in using instead the inequality $\binom{n}{n\delta_n} \leq \exp(n\, h(\delta_n))$ 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

$$\begin{aligned}\ln|\mathcal{N}_V(Y^n = y^n)| &\leq \ln\binom{n}{n\delta_n} + n\delta_n \ln|\mathcal{Y}| \\ &\leq n\delta_n\left(\ln n + \ln|\mathcal{Y}|\right), \quad \forall\, y^n \in \mathcal{Y}^n. \end{aligned} \tag{3.6.31}$$

Now, for each realization of V, we have

$$\begin{aligned}
&D(P_{Y^n(V)}\|P^*_{Y^n}) \\
&= D(P_{Y^n(V)|X^n(V)}\|P^*_{Y^n}|P_{X^n(V)}) - I(X^n(V);Y^n(V)) && (3.6.32)\\
&\leq nC - I(X^n(V);Y^n(V)) && (3.6.33)\\
&\leq nC - I(J;Y^n(V)) && (3.6.34)\\
&= nC - H(J) + H(J|Y^n(V)) \\
&\leq nC - \ln N + (1-\varepsilon(V)) \max_{y^n\in\mathcal{Y}^n} \ln|\mathcal{N}_V(y^n)| + n\varepsilon(V)\ln|\mathcal{X}| + \ln 2 && (3.6.35)
\end{aligned}$$

where:

- (3.6.32) is by the chain rule for divergence;
- (3.6.33) is by (3.6.9);
- (3.6.34) is by the data processing inequality and the fact that $J \to X^n(V) \to Y^n(V)$ is a Markov chain; and
- (3.6.35) is by Fano's inequality for list decoding (see Appendix 3.E), 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 $V = U^N$.) Substituting (3.6.30) and (3.6.31) into (3.6.35), we

get

$$D\big(P_{Y^n(V)}\big\|P^*_{Y^n}\big) \leq nC - \ln M + \ln n + 2n\delta_n\left(\ln n + \ln|\mathcal{Y}|\right) + n\varepsilon(V)\ln|\mathcal{X}| + \ln 2. \tag{3.6.36}$$

Using the fact that $\mathbb{E}\left[P_{Y^n(V)}\right] = P^{(\mathcal{C})}_{Y^n}$, convexity of the relative entropy, and (3.6.36), we get

$$D\big(P^{(\mathcal{C})}_{Y^n}\|P^*_{Y^n}\big) \leq nC - \ln M + \ln n + 2n\delta_n\left(\ln n + \ln|\mathcal{Y}|\right) + n\,\mathbb{E}\left[\varepsilon(V)\right]\,\ln|\mathcal{X}| + \ln 2. \tag{3.6.37}$$

To finish the proof and get (3.6.25), 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 (3.6.27), as well as the substitution of (3.6.26) in (3.6.37) (note that, from (3.6.26), 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 3.6.7. □

We are now ready to examine some consequences of Theorems 3.6.6 and 3.6.7. 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 $T\colon \mathcal{X}\to\mathcal{Y}$ with $C > 0$. We say that $\{\mathcal{C}_n\}_{n=1}^{\infty}$ is *capacity-achieving* if

$$\lim_{n\to\infty}\frac{1}{n}\ln M_n = C. \tag{3.6.38}$$

Then, from Theorems 3.6.6 and 3.6.7, it follows that any such sequence satisfies

$$\lim_{n\to\infty}\frac{1}{n}D\big(P^{(\mathcal{C}_n)}_{Y^n}\|P^*_{Y^n}\big) = 0. \tag{3.6.39}$$

Moreover, as shown in [170], if the restriction to either deterministic encoding maps or to the maximal probability of error criterion is lifted, then the convergence in (3.6.39) may no longer hold. This is in sharp contrast to [168, Theorem 2], which states that (3.6.39) 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 [170, Proposition 10] (again, we have streamlined the statement and the proof a bit to relate them to earlier material in this chapter):

Theorem 3.6.8. Let $T\colon \mathcal{X} \to \mathcal{Y}$ be a DMC with $C > 0$ and $C_1 < \infty$. Let $d\colon \mathcal{Y}^n \times \mathcal{Y}^n \to \mathbb{R}_+$ 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 $\mathrm{T}_1(c)$ 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 (to be defined explicitly in the following), such that for any (n, M, ε)-code $\mathcal{C}$ for T and any function $f\colon \mathcal{Y}^n \to \mathbb{R}$ we have

$$\begin{aligned} &P^{(\mathcal{C})}_{Y^n}\Big(|f(Y^n) - \mathbb{E}[f(Y^{*n})]| \geq r\Big) \\ &\qquad \leq \frac{4}{\varepsilon} \cdot \exp\left(nC - \ln M + a\sqrt{n} - \frac{r^2}{8c\|f\|^2_{\mathrm{Lip}}}\right) \end{aligned} \tag{3.6.40}$$

for all $r > 0$, where $\mathbb{E}[f(Y^{*n})]$ designates the expected value of $f(Y^n)$ w.r.t. the capacity-achieving output distribution $P^*_{Y^n}$,

$$\|f\|_{\mathrm{Lip}} \triangleq \sup_{y^n \neq v^n} \frac{|f(y^n) - f(v^n)|}{d(y^n, v^n)}$$

is the Lipschitz constant of f w.r.t. the metric d, and

$$a \triangleq c(T)\sqrt{\frac{1}{2}\ln\frac{1}{1-2\varepsilon}} \tag{3.6.41}$$

with $c(T)$ in (3.6.14).

Remark 3.32. Our sharpening of the corresponding result from [170, Proposition 10] consists mainly in identifying an explicit form for the constant in front of $\sqrt{n}$ in the bound (3.6.40); this provides a closed-form expression for the concentration of measure inequality.

Proof. For any f, define

$$\mu_f^* \triangleq \mathbb{E}[f(Y^{*n})], \qquad \phi(x^n) \triangleq \mathbb{E}[f(Y^n)|X^n = x^n], \ \forall\, x^n \in \mathcal{X}^n. \tag{3.6.42}$$

Since each $P_{Y^n|X^n=x^n}$ satisfies $\mathrm{T}_1(c)$, by the Bobkov–Götze theorem (Theorem 3.4.4), we have

$$\mathbb{P}\Big(|f(Y^n) - \phi(x^n)| \geq r \Big| X^n = x^n \Big) \leq 2\exp\left(-\frac{r^2}{2c\|f\|_{\mathrm{Lip}}^2} \right), \ \forall\, r \geq 0. \tag{3.6.43}$$

Now, given $\mathcal{C}$, consider a subcode $\mathcal{C}'$ with codewords $x^n \in \mathcal{X}^n$ satisfying $\phi(x^n) \geq \mu_f^* + r$ for $r \geq 0$. The number of codewords M' of $\mathcal{C}'$ satisfies

$$M' = M P_{X^n}^{(\mathcal{C})} \left(\phi(X^n) \geq \mu_f^* + r \right). \tag{3.6.44}$$

Let $Q = P_{Y^n}^{(\mathcal{C}')}$ be the output distribution induced by $\mathcal{C}'$. Then

$$\begin{aligned}
\mu_f^* + r &\leq \frac{1}{M'} \sum_{x^n \in \text{ codewords}(\mathcal{C}')} \phi(x^n) && (3.6.45)\\
&= \mathbb{E}_Q[f(Y^n)] && (3.6.46)\\
&\leq \mathbb{E}[f(Y^{*n})] + \|f\|_{\mathrm{Lip}} \sqrt{2cD(Q_{Y^n} \| P_{Y^n}^*)} && (3.6.47)\\
&\leq \mu_f^* + \|f\|_{\mathrm{Lip}} \sqrt{2c\left(nC - \ln M' + a\sqrt{n} + \ln\frac{1}{\varepsilon} \right)}, && (3.6.48)
\end{aligned}$$

where:

- (3.6.45) is by definition of $\mathcal{C}'$;
- (3.6.46) is by definition of ϕ in (3.6.42);
- (3.6.47) follows from the fact that $P_{Y^n}^*$ satisfies $\mathrm{T}_1(c)$ and from the Kantorovich–Rubinstein formula (3.4.56); and
- (3.6.48) holds for the constant $a = a(T, \varepsilon) > 0$ in (3.6.41) due to Theorem 3.6.6 (see (3.6.15)) and because $\mathcal{C}'$ is an (n, M', ε)-code for T.

From this and (3.6.44), we get

$$r \le \|f\|_{\mathrm{Lip}} \sqrt{2c\left(nC - \ln M - \ln P_{X^n}^{(\mathcal{C})}\left(\phi(X^n) \ge \mu_f^* + r\right) + a\sqrt{n} + \ln\frac{1}{\varepsilon}\right)}$$

so, it follows that

$$P_{X^n}^{(\mathcal{C})}\Big(\phi(X^n) \ge \mu_f^* + r\Big) \le \exp\left(nC - \ln M + a\sqrt{n} + \ln\frac{1}{\varepsilon} - \frac{r^2}{2c\|f\|_{\mathrm{Lip}}^2}\right).$$

Following the same line of reasoning with $-f$ instead of f, we conclude that

$$\begin{aligned} &P_{X^n}^{(\mathcal{C})}\Big(\Big|\phi(X^n) - \mu_f^*\Big| \ge r\Big) \\ &\qquad \le 2\exp\left(nC - \ln M + a\sqrt{n} + \ln\frac{1}{\varepsilon} - \frac{r^2}{2c\|f\|_{\mathrm{Lip}}^2}\right). \end{aligned} \tag{3.6.49}$$

Finally, for every $r \ge 0$,

$$\begin{aligned} &P_{Y^n}^{(\mathcal{C})}\Big(\Big|f(Y^n) - \mu_f^*\Big| \ge r\Big) \\ &\le P_{X^n,Y^n}^{(\mathcal{C})}\Big(|f(Y^n) - \phi(X^n)| \ge r/2\Big) + P_{X^n}^{(\mathcal{C})}\Big(\Big|\phi(X^n) - \mu_f^*\Big| \ge r/2\Big) \\ &\le 2\exp\left(-\frac{r^2}{8c\|f\|_{\mathrm{Lip}}^2}\right) + 2\exp\left(nC - \ln M + a\sqrt{n} + \ln\frac{1}{\varepsilon} - \frac{r^2}{8c\|f\|_{\mathrm{Lip}}^2}\right) \qquad (3.6.50) \\ &\le 4\exp\left(nC - \ln M + a\sqrt{n} + \ln\frac{1}{\varepsilon} - \frac{r^2}{8c\|f\|_{\mathrm{Lip}}^2}\right), \qquad (3.6.51) \end{aligned}$$

where (3.6.50) is by (3.6.43) and (3.6.49), while (3.6.51) follows from the fact that

$$nC - \ln M + a\sqrt{n} + \ln\frac{1}{\varepsilon} \ge D(P_{Y^n}^{(\mathcal{C})} \| P_{Y^n}^*) \ge 0$$

by Theorem 3.6.6, and from (3.6.41). This proves (3.6.40). □

As an illustration, let us consider $\mathcal{Y}^n$ with the Hamming metric

$$d_n(y^n, v^n) = \sum_{i=1}^n 1_{\{y_i \ne v_i\}}. \tag{3.6.52}$$

Then, any function $f\colon \mathcal{Y}^n \to \mathbb{R}$ of the form

$$f(y^n) = \frac{1}{n}\sum_{i=1}^{n} f_i(y_i), \qquad \forall y^n \in \mathcal{Y}^n \tag{3.6.53}$$

where $f_1, \ldots, f_n\colon \mathcal{Y} \to \mathbb{R}$ are Lipschitz functions on $\mathcal{Y}$, will satisfy

$$\|f\|_{\mathrm{Lip}} \le \frac{L}{n}, \qquad L \triangleq \max_{1\le i\le n} \|f_i\|_{\mathrm{Lip}}.$$

Any probability distribution P on $\mathcal{Y}$ equipped with the Hamming metric satisfies $\mathrm{T}_1(1/4)$ (this is simply Pinsker's inequality); by Proposition 3.8, any product probability distribution on $\mathcal{Y}^n$ satisfies $\mathrm{T}_1(n/4)$ with respect to the product metric (3.6.52). Consequently, for any (n, M, ε)-code for T and any function $f\colon \mathcal{Y}^n \to \mathbb{R}$ of the form (3.6.53), Theorem 3.6.8 gives the concentration inequality

$$\begin{aligned} &P_{Y^n}^{(\mathcal{C})}\Big(\,|f(Y^n) - \mathbb{E}[f(Y^{*n})]| \ge r\Big) \\ &\qquad \le \frac{4}{\varepsilon}\,\exp\left(nC - \ln M + a\sqrt{n} - \frac{nr^2}{2L^2}\right) \end{aligned} \tag{3.6.54}$$

for all $r > 0$. Concentration inequalities like (3.6.40), or its more specialized version (3.6.54), can be very useful for assessing 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 3.6.8 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.

The bounds presented in Theorems 3.6.6 and 3.6.7 quantify the trade-offs between the minimal blocklength required for achieving a certain gap (in rate) to capacity with a fixed block error probability, and normalized divergence between the *output distribution* induced by the code and the (unique) capacity-achieving output distribution of the channel. Moreover, these bounds sharpen the asymptotic $O(\cdot)$ terms in the results of [170] for all finite blocklengths n.

These results are similar in spirit to 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 a code and the binomial distribution. Specifically, a combination of [175, Eqs. (A17) and (A19)] provides a lower bound on the rate loss with respect to fully random block codes in terms of the normalized divergence between the distance spectrum of the code and the binomial distribution where the latter result refers to the empirical *input distribution* of good codes.

3.6.4 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 "too small" a 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 [176] using information-theoretic techniques.

The following setting is considered in [176]: Let $\mathcal{X}$ be a finite set, together with a nonnegative distortion function $d\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ (which is not necessarily a metric) and a strictly positive mass function $M\colon \mathcal{X} \to (0, \infty)$ (which is not necessarily normalized to one). As before, let us extend the "single-letter" distortion d to $d_n\colon \mathcal{X}^n \to \mathbb{R}^+$, $n \in \mathbb{N}$, 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

$$M^n(x^n) \triangleq \prod_{i=1}^{n} M(x_i), \qquad \forall x^n \in \mathcal{X}^n.$$

We also recall the definition of the r-blowup of any set $A \subseteq \mathcal{X}^n$:

$$A_r \triangleq \{x^n \in \mathcal{X}^n : d_n(x^n, A) \leq r\},$$

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 $\delta \geq 0$, how small can their masses $M^n(A^{(n)})$ be?

In order to state and prove the main result of [176] 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 $\delta \geq 0$, let us define the set

$$\Pi_n(P_n, Q_n, \delta) \triangleq \left\{\pi_n \in \Pi_n(P_n, Q_n) : \frac{1}{n}\mathbb{E}_{\pi_n}\left[d_n(X^n, Y^n)\right] \leq \delta\right\} \tag{3.6.55}$$

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*:

$$\begin{aligned} R_n(\delta) &\equiv R_n(\delta; P_n, M^n) \\ &\triangleq \inf_{Q_n \in \mathcal{P}(\mathcal{X}^n)} \left\{I_n(P_n, Q_n, \delta) + \mathbb{E}_{Q_n}[\ln M^n(Y^n)]\right\} \\ &\equiv \inf_{P_{X^nY^n}} \Bigg\{I(X^n; Y^n) + \mathbb{E}[\ln M^n(Y^n)] \\ &\qquad\qquad : P_{X^n} = P_n, \frac{1}{n}\mathbb{E}[d_n(X^n, Y^n)] \leq \delta\Bigg\}. \end{aligned}$$

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

$$R(\delta) = \lim_{n\to\infty} \frac{1}{n} R_n(\delta) = \inf_{n \geq 1} \frac{1}{n} R_n(\delta) \tag{3.6.56}$$

(see [176, Lemma 2]). We are now ready to state the main result of [176]:

Theorem 3.6.9 (Kontoyiannis)**.** 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

$$\frac{1}{n}\ln M^n(A^{(n)}) \geq R(\delta; P, M). \tag{3.6.57}$$

Proof. Given $A^{(n)} \subseteq \mathcal{X}^n$, let $\varphi_n \colon \mathcal{X}^n \to A^{(n)}$ 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:

$$\pi_n(x^n, y^n) = P^{\otimes n}(x^n)1_{\{y^n = \varphi_n(x^n)\}}.$$

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,

$$\begin{aligned}
\ln M^n(A^{(n)}) &= \ln \sum_{y^n \in A^{(n)}} M^n(y^n) \\
&= \ln \sum_{y^n \in A^{(n)}} Q_n(y^n) \cdot \frac{M^n(y^n)}{Q_n(y^n)} \\
&\geq \sum_{y^n \in A^{(n)}} Q_n(y^n) \ln \frac{M^n(y^n)}{Q_n(y^n)} && (3.6.58) \\
&= \sum_{x^n \in \mathcal{X}^n, y^n \in A^{(n)}} \pi_n(x^n, y^n) \ln \frac{\pi_n(x^n, y^n)}{P^{\otimes n}(x^n) Q_n(y^n)} \\
&\qquad + \sum_{y^n \in A^{(n)}} Q_n(y^n) \ln M^n(y^n) && (3.6.59) \\
&= I(X^n; Y^n) + \mathbb{E}_{Q_n}[\ln M^n(Y^n)] && (3.6.60) \\
&\geq R_n(\delta), && (3.6.61)
\end{aligned}$$

where (3.6.58) is by Jensen's inequality, (3.6.59) and (3.6.60) use the fact that π_n is a coupling of $P^{\otimes n}$ and Q_n, and (3.6.61) is by definition of $R_n(\delta)$. Using (3.6.56), we get (3.6.57), and the theorem is proved. □

Remark 3.33. In the same paper [176], 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

$$\frac{1}{n} \ln M^n(A^{(n)}) \leq R(\delta) + \varepsilon, \qquad \forall n \in \mathbb{N} \tag{3.6.62}$$

and

$$\frac{1}{n} d_n(X^n, A^{(n)}) \leq \delta, \qquad \text{eventually a.s.} \tag{3.6.63}$$

We are now ready to use Theorem 3.6.9 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 3.6.10 (Converse concentration of measure)**.** If $A^{(n)} \subseteq \mathcal{X}^n$ is an arbitrary set, then

$$P^{\otimes n}\left(A^{(n)}\right) \geq \exp\left(n\, R_{\mathrm{c}}(\delta; P)\right), \tag{3.6.64}$$

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 $\delta \geq 0$, $P^{\otimes n}\left(A_{n\delta}^{(n)}\right) \to 1$ as $n \to \infty$, then

$$\liminf_{n\to\infty} \frac{1}{n} \ln P^{\otimes n}\left(A^{(n)}\right) \geq R_{\rm c}(\delta; P). \tag{3.6.65}$$

Remark 3.34. A moment of reflection shows that the concentration exponent $R_{\rm c}(\delta; P)$ is nonpositive. Indeed, from definitions,

$$\begin{aligned}
&R_{\rm c}(\delta; P)\\
&= R(\delta; P, P)\\
&= \inf_{P_{XY}} \Big\{ I(X;Y) + \mathbb{E}[\ln P(Y)] : P_X = P,\, \mathbb{E}[d(X,Y)] \leq \delta \Big\}\\
&= \inf_{P_{XY}} \Big\{ H(Y) - H(Y|X) + \mathbb{E}[\ln P(Y)] : P_X = P,\, \mathbb{E}[d(X,Y)] \leq \delta \Big\}\\
&= \inf_{P_{XY}} \Big\{ -D(P_Y\|P) - H(Y|X) : P_X = P,\, \mathbb{E}[d(X,Y)] \leq \delta \Big\}\\
&= -\sup_{P_{XY}} \Big\{ D(P_Y\|P) + H(Y|X) : P_X = P,\, \mathbb{E}[d(X,Y)] \leq \delta \Big\},
\end{aligned} \tag{3.6.66}$$

which proves the claim, since both the divergence and the (conditional) entropy are nonnegative.

Remark 3.35. Using the achievability result from [176] (cf. Remark 3.33), one can also prove that there exists a sequence of sets $\{A^{(n)}\}_{n=1}^{\infty}$, such that

$$\lim_{n\to\infty} P^{\otimes n}\left(A_{n\delta}^{(n)}\right) = 1 \qquad \text{and} \qquad \lim_{n\to\infty} \frac{1}{n} \ln P^{\otimes n}\left(A^{(n)}\right) \leq R_{\rm c}(\delta; P).$$

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 $\mathrm{T}_1\left(\frac{1}{2\varphi(p)}\right)$ transportation-cost inequality with respect to the L^1 Wasserstein distance induced by the Hamming metric, where $\varphi(p)$ is defined in (3.4.33). By Proposition 3.7, the product measure $P^{\otimes n}$ satisfies a $\mathrm{T}_1\left(\frac{n}{2\varphi(p)}\right)$ transportation-cost inequality on the product space $(\mathcal{X}^n, d_n)$. Consequently, it follows from (3.4.44) that

for any $\delta \geq 0$ and any $A^{(n)} \subseteq \mathcal{X}^n$,

$$P^{\otimes n}\left(A_{n\delta}^{(n)}\right) \geq 1 - \exp\left(-\frac{\varphi(p)}{n}\left(n\delta - \sqrt{\frac{n}{\varphi(p)}\ \ln\frac{1}{P^{\otimes n}\left(A^{(n)}\right)}}\right)^2\right)$$
$$= 1 - \exp\left(-n\,\varphi(p)\left(\delta - \sqrt{\frac{1}{n\,\varphi(p)}\ \ln\frac{1}{P^{\otimes n}\left(A^{(n)}\right)}}\right)^2\right). \tag{3.6.67}$$

Thus, if a sequence of sets $A^{(n)} \subseteq \mathcal{X}^n$, $n \in \mathbb{N}$, satisfies

$$\liminf_{n\to\infty} \frac{1}{n} \ln P^{\otimes n}\left(A^{(n)}\right) \geq -\varphi(p)\delta^2, \tag{3.6.68}$$

then

$$P^{\otimes n}\left(A_{n\delta}^{(n)}\right) \xrightarrow{n\to\infty} 1. \tag{3.6.69}$$

The converse result, Corollary 3.6.10, says that if a sequence of sets $A^{(n)} \subseteq \mathcal{X}^n$ satisfies (3.6.69), then (3.6.65) 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 (3.6.68):

Theorem 3.6.11. If P is the Bernoulli(p) measure with $p \in [0, 1/2]$, then the concentration exponent $R_{\mathrm{c}}(\delta; P)$ satisfies

$$R_{\mathrm{c}}(\delta; P) \leq -\varphi(p)\delta^2 - (1-p)h\left(\frac{\delta}{1-p}\right), \qquad \forall\, \delta \in [0, 1-p] \tag{3.6.70}$$

and

$$R_{\mathrm{c}}(\delta; P) = \ln p, \qquad \forall\, \delta \in [1-p, 1] \tag{3.6.71}$$

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 (3.6.66), we have

$$R_{\mathrm{c}}(\delta; P) = -\sup_{P_{XY}} \Big\{ D(P_Y \| P) + H(Y|X) : P_X = P,\, \mathbb{P}(X \neq Y) \leq \delta \Big\}. \tag{3.6.72}$$

For a given $\delta \in [0, 1-p]$, let us choose P_Y so that $\|P_Y - P\|_{\mathrm{TV}} = \delta$. Then from (3.4.35),

$$\begin{aligned}\frac{D(P_Y\|P)}{\delta^2} &= \frac{D(P_Y\|P)}{\|P_Y - P\|_{\mathrm{TV}}^2} \\ &\geq \inf_Q \frac{D(Q\|P)}{\|Q - P\|_{\mathrm{TV}}^2} \\ &= \varphi(p). \end{aligned} \tag{3.6.73}$$

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 $\mathbb{P}(\widetilde{X} \neq \widetilde{Y}) = \|P_Y - P\|_{\mathrm{TV}} = \delta$. Moreover, using (3.4.31), 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,

$$H(\widetilde{Y}|\widetilde{X}) = (1-p)H(\widetilde{Y}|\widetilde{X}=0) = (1-p)h\left(\frac{\delta}{1-p}\right). \tag{3.6.74}$$

From (3.6.72), (3.6.73) and (3.6.74), we obtain

$$\begin{aligned} R_{\mathrm{c}}(\delta; P) &\leq -D(P_{\widetilde{Y}}\|P) - H(\widetilde{Y}|\widetilde{X}) \\ &\leq -\varphi(p)\delta^2 - (1-p)h\left(\frac{\delta}{1-p}\right). \end{aligned}$$

To prove (3.6.71), 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 $p \leq 1/2$. □

3.7 Summary

In this chapter, we have covered the essentials of the entropy method, an information-theoretic technique for deriving concentration inequalities for functions of many independent random variables. As its very

name suggests, the entropy method revolves around the relative entropy (or information divergence), which in turn can be related to the logarithmic moment-generating function and its derivatives.

A key ingredient of the entropy method is *tensorization*, or the use of a certain subadditivity property of the divergence in order to break the original multidimensional problem up into simpler one-dimensional problems. Tensorization is used in conjunction with various inequalities relating the relative entropy to suitable energy-type functionals defined on the space of functions for which one wishes to establish concentration. These inequalities fall into two broad classes: functional inequalities (typified by the logarithmic Sobolev inequalities) and transportation-cost inequalities (such as Pinsker's inequality). We have examined the many deep and remarkable information-theoretic ideas that bridge these two classes of inequalities, and also showed some examples of their applications to problems in coding and information theory.

At this stage, the relationship between information theory and the study of measure concentration is heavily skewed towards the use of the former as a tool for the latter. By contrast, the range of applications of measure concentration ideas to problems in information theory has been somewhat limited. We hope that the present monograph may offer some inspiration for information and coding theorists to deepen the ties between their discipline and the fascinating realm of high-dimensional probability and concentration of measure.

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 $J(Y) < \infty$. Our goal is to prove the van Trees inequality (3.2.28) and to establish that equality in (3.2.28) 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

$$\mathbb{E}\left[(Y-\varphi(U))^2\right] \geq \frac{1}{s+J(Y)}, \tag{3.A.1}$$

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

$$\begin{aligned}
\Delta(U,Y) &\triangleq \varphi(U)-Y,\\
\Upsilon(U,Y) &\triangleq \frac{\mathrm{d}}{\mathrm{d}y}\ln\left[p_{U|Y}(U|y)p_Y(y)\right]\Big|_{y=Y}\\
&= \frac{\mathrm{d}}{\mathrm{d}y}\ln\left[\gamma(U-\sqrt{s}y)p_Y(y)\right]\Big|_{y=Y}\\
&= \sqrt{s}(U-\sqrt{s}Y)+\rho_Y(Y)\\
&= \sqrt{s}Z+\rho_Y(Y)
\end{aligned}$$

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

$$\begin{aligned}
1 &= |\mathbb{E}[\Delta(U,Y)\Upsilon(U,Y)]|^2\\
&\leq \mathbb{E}[\Delta^2(U,Y)]\cdot\mathbb{E}[\Upsilon^2(U,Y)]\\
&= \mathbb{E}[(\varphi(U)-Y)^2]\cdot\mathbb{E}[(\sqrt{s}Z+\rho_Y(Y))^2]\\
&= \mathbb{E}[(\varphi(U)-Y)^2]\cdot(s+J(Y)).
\end{aligned}$$

Upon rearranging, we obtain (3.A.1). Now, the fact that $J(Y)<\infty$ implies that the density p_Y is bounded (see [124, Lemma A.1]). Using this together with the rapid decay of the Gaussian density γ at infinity, we have

$$\int_{-\infty}^{\infty}\frac{\mathrm{d}}{\mathrm{d}y}\left[p_{U|Y}(u|y)p_Y(y)\right]\mathrm{d}y = \gamma(u-\sqrt{s}y)p_Y(y)\Big|_{-\infty}^{\infty} = 0. \tag{3.A.2}$$

Integration by parts gives

$$\begin{aligned}
&\int_{-\infty}^{\infty} y\frac{\mathrm{d}}{\mathrm{d}y}\left[p_{U|Y}(u|y)p_Y(y)\right]\mathrm{d}y \\
&\quad= y\gamma(u-\sqrt{s}y)p_Y(y)\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} p_{U|Y}(u|y)p_Y(y)\mathrm{d}y \\
&\quad= -\int_{-\infty}^{\infty} p_{U|Y}(u|y)p_Y(y)\mathrm{d}y \\
&\quad= -p_U(u). \qquad (3.A.3)
\end{aligned}$$

Using (3.A.2) and (3.A.3), we have

$$\begin{aligned}
&\mathbb{E}[\Delta(U,Y)\Upsilon(U,Y)] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (\varphi(u)-y)\,\frac{\mathrm{d}}{\mathrm{d}y}\ln\left[p_{U|Y}(u|y)p_Y(y)\right] p_{U|Y}(u|y)p_Y(y)\mathrm{d}u\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (\varphi(u)-y)\,\frac{\mathrm{d}}{\mathrm{d}y}\left[p_{U|Y}(u|y)p_Y(y)\right]\mathrm{d}u\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty} \varphi(u)\underbrace{\left(\int_{-\infty}^{\infty}\frac{\mathrm{d}}{\mathrm{d}y}\left[p_{U|Y}(u|y)p_Y(y)\right]\mathrm{d}y\right)}_{=0}\mathrm{d}u \\
&\qquad\qquad - \int_{-\infty}^{\infty}\underbrace{\left(\int_{-\infty}^{\infty} y\frac{\mathrm{d}}{\mathrm{d}y}\left[p_{U|Y}(u|y)p_Y(y)\right]\mathrm{d}y\right)}_{=-p_U(u)}\mathrm{d}u \\
&= \int_{-\infty}^{\infty} p_U(u)\mathrm{d}u \\
&= 1,
\end{aligned}$$

as was claimed. It remains to establish the necessary and sufficient condition for equality in (3.A.1). 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 (3.2.52) and (3.2.53) pertaining to the Ornstein–Uhlenbeck semigroup. We start with (3.2.52). 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

$$\begin{aligned}\dot{h}_t(x) &= \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[h\left(e^{-t}x + \sqrt{1-e^{-2t}}Z\right)\right] \\ &= -e^{-t}x\,\mathbb{E}\left[h'\left(e^{-t}x + \sqrt{1-e^{-2t}}Z\right)\right] \\ &\qquad + \frac{e^{-2t}}{\sqrt{1-e^{-2t}}} \cdot \mathbb{E}\left[Zh'\left(e^{-t}x + \sqrt{1-e^{-2t}}Z\right)\right].\end{aligned}$$

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,

$$\dot{h}_t(x) = -e^{-t}x \cdot K_t h'(x) + e^{-2t}K_t h''(x). \tag{3.B.1}$$

On the other hand,

$$\begin{aligned}\mathcal{L}h_t(x) &= h_t''(x) - xh_t'(x) \\ &= e^{-2t}\mathbb{E}\left[h''\left(e^{-t}x + \sqrt{1-e^{-2t}}Z\right)\right] \\ &\qquad\qquad - xe^{-t}\mathbb{E}\left[h'\left(e^{-t}x + \sqrt{1-e^{-2t}}Z\right)\right] \\ &= e^{-2t}K_t h''(x) - e^{-t}xK_t h'(x). \end{aligned} \tag{3.B.2}$$

Comparing (3.B.1) and (3.B.2), we get (3.2.52).

The proof of the integration-by-parts formula (3.2.53) 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 $t, t' \geq 0$, $(Y_t, Y_{t'}) \stackrel{\mathrm{d}}{=} (Y_{t'}, Y_t)$. 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 $\mathrm{OU}(t)$ 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 (3.2.48) and the reversibility property of the Ornstein–Uhlenbeck process. Taking the derivative of both sides with respect to t, we conclude that

$$\mathbb{E}[g(Z)\mathcal{L}h(Z)] = \mathbb{E}[\mathcal{L}g(Z)h(Z)]. \tag{3.B.3}$$

In particular, since $\mathcal{L}1 = 0$ (where on the left-hand side 1 denotes the constant function $x \mapsto 1$), we have

$$\mathbb{E}[\mathcal{L}g(Z)] = \mathbb{E}[1\mathcal{L}g(Z)] = \mathbb{E}[g(Z)\mathcal{L}1] = 0 \tag{3.B.4}$$

for all smooth g.

Remark 3.36. 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.B.3) expresses the fact that $\mathcal{L}$ is a self-adjoint linear operator on this space. Moreover, (3.B.4) 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 (3.2.53). To that end, let us first define the operator Γ on pairs of functions g, h by

$$\Gamma(g, h) \triangleq \frac{1}{2}\left[\mathcal{L}(gh) - g\mathcal{L}h - h\mathcal{L}g\right]. \tag{3.B.5}$$

Remark 3.37. 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

$$\begin{aligned}\Gamma(g,h)(x) &= \frac{1}{2}\Big[(gh)''(x) - x(gh)'(x) - g(x)\big(h''(x) - xh'(x)\big) \\ &\qquad - h(x)\big(g''(x) - xg'(x)\big)\Big] \\ &= \frac{1}{2}\Big[g''(x)h(x) + 2g'(x)h'(x) + g(x)h''(x) \\ &\qquad - xg'(x)h(x) - xg(x)h'(x) - g(x)h''(x) \\ &\qquad + xg(x)h'(x) - g''(x)h(x) + xg'(x)h(x)\Big] \\ &= g'(x)h'(x), \end{aligned} \tag{3.B.6}$$

or, more succinctly, $\Gamma(g,h) = g'h'$. Therefore,

$$\mathbb{E}[g(Z)\mathcal{L}h(Z)] = \frac{1}{2}\Big\{\mathbb{E}[g(Z)\mathcal{L}h(Z)] + \mathbb{E}[h(Z)\mathcal{L}g(Z)]\Big\} \tag{3.B.7}$$

$$= \frac{1}{2}\mathbb{E}[\mathcal{L}(gh)(Z)] - \mathbb{E}[\Gamma(g,h)(Z)] \tag{3.B.8}$$

$$= -\mathbb{E}[g'(Z)h'(Z)], \tag{3.B.9}$$

where (3.B.7) uses (3.B.3), (3.B.8) uses the definition (3.B.5) of Γ, and (3.B.9) uses (3.B.6) together with (3.B.4). This proves (3.2.53).

3.C Log-Sobolev inequalities for Bernoulli and Gaussian measures

The following log-Sobolev inequality was derived by Gross [44]:

$$\mathrm{Ent}_P[g^2] \le \frac{(g(0) - g(1))^2}{2}. \tag{3.C.1}$$

We will now show that (3.3.29) can be derived from (3.C.1). Let us define f by $e^f = g^2$, where we may assume without loss of generality that $0 < g(0) \le g(1)$. Note that

$$\begin{aligned}(g(0) - g(1))^2 &= \left(\exp\left(f(0)/2\right) - \exp\left(f(1)/2\right)\right)^2 \\ &\le \frac{1}{8}\left[\exp\left(f(0)\right) + \exp\left(f(1)\right)\right]\left(f(0) - f(1)\right)^2 \\ &= \frac{1}{4}\mathbb{E}_P\left[\exp(f)(\Gamma f)^2\right] \end{aligned} \tag{3.C.2}$$

with $\Gamma f = |f(0) - f(1)|$, where the inequality follows from the easily verified fact that $(1-x)^2 \leq \frac{(1+x^2)(\ln x)^2}{2}$ for all $x \geq 0$, which we apply to $x \triangleq g(1)/g(0)$. Therefore, the inequality in (3.C.1) implies the following:

$$D(P^{(f)}||P) = \frac{\mathrm{Ent}_P[\exp(f)]}{\mathbb{E}_P[\exp(f)]} \tag{3.C.3}$$

$$= \frac{\mathrm{Ent}_P[g^2]}{\mathbb{E}_P[\exp(f)]} \tag{3.C.4}$$

$$\leq \frac{\big(g(0) - g(1)\big)^2}{2\,\mathbb{E}_P[\exp(f)]} \tag{3.C.5}$$

$$\leq \frac{\mathbb{E}_P[\exp(f)\,(\Gamma f)^2]}{8\,\mathbb{E}_P[\exp(f)]} \tag{3.C.6}$$

$$= \frac{1}{8}\,\mathbb{E}_P^{(f)}\big[(\Gamma f)^2\big] \tag{3.C.7}$$

where equality (3.C.3) follows from (3.3.4), equality (3.C.4) holds due to the equality $e^f = g^2$, inequality (3.C.5) holds due to (3.C.1), inequality (3.C.6) follows from (3.C.2), and equality (3.C.7) follows by definition of the expectation with respect to the tilted probability measure $P^{(f)}$. Therefore, we conclude that indeed (3.C.1) implies (3.3.29).

Gross used (3.C.1) and the central limit theorem to establish his Gaussian log-Sobolev inequality (see Theorem 3.2.1). We can follow the same steps and arrive at (3.2.11) from (3.3.29). To that end, let $g\colon \mathbb{R} \to \mathbb{R}$ 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 $f\colon \{0,1\}^n \to \mathbb{R}$ 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 $n \to \infty$: $P_{Z_n} \Rightarrow G$. Therefore, by the assumed smoothness properties of g we have (see (3.3.2) and (3.3.4))

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(f(X^n)\right)\right] \cdot D(P_{X^n}^{(f)} \| P_{X^n}) \\
&= \mathbb{E}\left[f(X^n)\exp\left(f(X^n)\right)\right] - \mathbb{E}[\exp\left(f(X^n)\right)] \ln \mathbb{E}[\exp\left(f(X^n)\right)] \\
&= \mathbb{E}\left[g(Z_n)\exp\left(g(Z_n)\right)\right] - \mathbb{E}[\exp\left(g(Z_n)\right)] \ln \mathbb{E}[\exp\left(g(Z_n)\right)] \\
&\xrightarrow{n\to\infty} \mathbb{E}\left[g(Z)\exp\left(g(Z)\right)\right] - \mathbb{E}[\exp\left(g(Z)\right)] \ln \mathbb{E}[\exp\left(g(Z)\right)] \\
&= \mathbb{E}\left[\exp\left(g(Z)\right)\right] D(P_Z^{(g)} \| P_Z)
\end{aligned} \tag{3.C.8}$$

where $Z \sim G$ is a standard Gaussian random variable. Moreover, using the definition (3.3.28) 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,

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(f(X^n)\right)\right] \cdot \mathbb{E}^{(f)}\left[\left(\Gamma f(X^n)\right)^2\right] \\
&\quad = \mathbb{E}\left[\exp\left(f(X^n)\right)\left(\Gamma f(X^n)\right)^2\right] \\
&\quad = 4\,\mathbb{E}\left[\exp\left(g(Z_n)\right)\left(\left(g'(Z_n)\right)^2 + o(1)\right)\right] \\
&\xrightarrow{n\to\infty} 4\,\mathbb{E}\left[\exp\left(g(Z)\right)\left(g'(Z)\right)^2\right] \\
&\quad = 4\,\mathbb{E}\left[\exp\left(g(Z)\right)\right] \cdot \mathbb{E}_{P_Z}^{(g)}\left[\left(g'(Z)\right)^2\right].
\end{aligned} \tag{3.C.9}$$

Taking the limit of both sides of (3.3.29) as $n \to \infty$ and then using (3.C.8) and (3.C.9), we obtain

$$D(P_Z^{(g)} \| P_Z) \leq \frac{1}{2} \mathbb{E}_{P_Z}^{(g)} \left[\left(g'(Z) \right)^2 \right],$$

which is (3.2.11). The same technique applies in the case of an asymmetric Bernoulli measure: given a sufficiently smooth function $g \colon \mathbb{R} \to \mathbb{R}$, define $f \colon \{0,1\}^n \to \mathbb{R}$ by

$$f(x^n) \triangleq g\left(\frac{x_1 + \ldots + x_n - np}{\sqrt{npq}} \right).$$

and then apply (3.3.33) to it.

3.D On the sharp exponent in McDiarmid's inequality

It is instructive to compare the strategy used to prove Theorem 3.3.8 with an earlier approach by Boucheron, Lugosi and Massart [136] using the entropy method. Their starting point is the following lemma:

Lemma 3.D.1. Define the function $\psi \colon \mathbb{R} \to \mathbb{R}$ by $\psi(u) = \exp(u) - u - 1$. Consider a probability space $(\Omega, \mathcal{F}, \mu)$ and a measurable function $f \colon \Omega \to \mathbb{R}$ such that tf is exponentially integrable with respect to μ for all $t \in \mathbb{R}$. Then, the following inequality holds for any $c \in \mathbb{R}$:

$$D(\mu^{(tf)} \| \mu) \leq \mathbb{E}_{\mu}^{(tf)} \left[\psi(-t(f-c)) \right]. \tag{3.D.1}$$

Proof. Recall that

$$\begin{aligned} D(\mu^{(tf)} \| \mu) &= t \, \mathbb{E}_{\mu}^{(tf)} [f] + \ln \frac{1}{\mathbb{E}_{\mu}[\exp(tf)]} \\ &= t \, \mathbb{E}_{\mu}^{(tf)}[f] - tc + \ln \frac{\exp(tc)}{\mathbb{E}_{\mu}[\exp(tf)]} \end{aligned}$$

Using this together with the inequality $\ln u \leq u - 1$ for every $u > 0$, we can write

$$\begin{aligned} D(\mu^{(tf)} \| \mu) &\leq t \, \mathbb{E}_{\mu}^{(tf)}[f] - tc + \frac{\exp(tc)}{\mathbb{E}_{\mu}[\exp(tf)]} - 1 \\ &= t \, \mathbb{E}_{\mu}^{(tf)}[f] - tc + \mathbb{E}_{\mu} \left[\frac{\exp(t(f+c)) \exp(-tf)}{\mathbb{E}_{\mu}[\exp(tf)]} \right] - 1 \\ &= t \, \mathbb{E}_{\mu}^{(tf)}[f] + \exp(tc) \, \mathbb{E}_{\mu}^{(tf)} \left[\exp(-tf) \right] - tc - 1, \end{aligned}$$

and we get (3.D.1). This completes the proof of Lemma 3.D.1. □

Notice that, while (3.D.1) is an upper bound on $D(\mu^{(tf)}\|\mu)$, the thermal fluctuation representation (3.3.15) of Lemma 3.3.2 is an *exact* expression. Lemma 3.D.1 leads to the following inequality of log-Sobolev type:

Theorem 3.D.2. 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 $f\colon \mathcal{X}^n \to \mathbb{R}$. 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,

$$D(P^{(tf)}\|P) \leq \exp\big(-\Lambda(t)\big) \sum_{i=1}^{n} \mathbb{E}\left[\exp(tU)\,\psi\left(-t(U-U^{(i)})\right)\right], \tag{3.D.2}$$

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 with respect to X^n and $(X')^n$. Moreover, if we define the function $\tau\colon \mathbb{R} \to \mathbb{R}$ by $\tau(u) \triangleq u\big(\exp(u) - 1\big)$ for $u \in \mathbb{R}$, then

$$D(P^{(tf)}\|P) \leq \exp\big(-\Lambda(t)\big) \sum_{i=1}^{n} \mathbb{E}\left[\exp(tU)\,\tau\left(-t(U-U^{(i)})\right) 1_{\{U>U^{(i)}\}}\right] \tag{3.D.3}$$

and

$$D(P^{(tf)}\|P) \leq \exp\big(-\Lambda(t)\big) \sum_{i=1}^{n} \mathbb{E}\left[\exp(tU)\,\tau\left(-t(U-U^{(i)})\right) 1_{\{U<U^{(i)}\}}\right]. \tag{3.D.4}$$

Proof. Applying Proposition 3.2 to P and $Q = P^{(tf)}$, we have

$$\begin{aligned} D(P^{(tf)}\|P) &\le \sum_{i=1}^n D\Big(P^{(tf)}_{X_i|\bar{X}^i}\Big\|P_{X_i}\Big|P^{(tf)}_{\bar{X}^i}\Big) \\ &= \sum_{i=1}^n \int P^{(tf)}_{\bar{X}^i}(\mathrm{d}\bar{x}^i)\, D\Big(P^{(tf)}_{X_i|\bar{X}^i=\bar{x}^i}\Big\|P_{X_i}\Big) \\ &= \sum_{i=1}^n \int P^{(tf)}_{\bar{X}^i}(\mathrm{d}\bar{x}^i)\, D\Big(P^{(tf_i(\cdot|\bar{x}^i))}_{X_i}\Big\|P_{X_i}\Big). \end{aligned} \tag{3.D.5}$$

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 3.D.1 to the ith term of the summation in (3.D.5) 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.D.5), and then taking expectation with respect to both X^n and $(X')^n$, we get (3.D.2).

To prove (3.D.3) and (3.D.4), let us write (note that $\psi(0) = 0$)

$$\begin{aligned} &\exp(tU)\,\psi\Big(-t(U-U^{(i)})\Big) \\ &= \exp(tU)\,\psi\Big(-t(U-U^{(i)})\Big)\,1_{\{U>U^{(i)}\}} \\ &\qquad + \exp(tU)\,\psi\Big(t(U^{(i)}-U)\Big)\,1_{\{U<U^{(i)}\}}. \end{aligned} \tag{3.D.6}$$

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 distributions of U and $U^{(i)}$)

$$\begin{aligned} &\mathbb{E}\Big[\exp(tU)\,\psi\Big(t(U^{(i)}-U)\Big)\,1_{\{U<U^{(i)}\}}\Big|\bar{X}^i\Big] \\ &= \mathbb{E}\Big[\exp(tU^{(i)})\,\psi\Big(t(U-U^{(i)})\Big)\,1_{\{U>U^{(i)}\}}\Big|\bar{X}^i\Big] \\ &= \mathbb{E}\Big[\exp(tU)\exp\Big(t(U^{(i)}-U)\Big)\,\psi\Big(t(U-U^{(i)})\Big)\,1_{\{U>U^{(i)}\}}\Big|\bar{X}^i\Big]. \end{aligned}$$

Using this and (3.D.6), we can write

$$\begin{aligned}
&\mathbb{E}\left[\exp(tU)\psi\left(-t(U-U^{(i)})\right)\right] \\
&= \mathbb{E}\Bigg\{ \exp(tU)\Big[\psi\left(t(U^{(i)}-U)\right) \\
&\qquad\qquad + \exp\left(t(U^{(i)}-U)\right)\psi\left(t(U-U^{(i)})\right)\Big]\, 1_{\{U>U^{(i)}\}}\Bigg\}.
\end{aligned}$$

Using the equality $\psi(u)+\exp(u)\psi(-u)=\tau(u)$ for every $u\in\mathbb{R}$, we get (3.D.3). The proof of (3.D.4) is similar. □

Now suppose that f satisfies the bounded difference condition in (3.3.35). Using this together with the fact that $\tau(-u) = u\big(1-\exp(-u)\big) \le u^2$ for every $u\ge 0$, then for every $t>0$ we can write

$$\begin{aligned}
&D\big(P^{(tf)}\|P\) \\
&\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)\,(U-U^{(i)})^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] \\
&\le 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 3.1.1, we get

$$\mathbb{P}\Big(f(X^n)\ge \mathbb{E}f(X^n)+r\Big)\le \exp\left(-\frac{r^2}{4\sum_{i=1}^n c_i^2}\right),\quad \forall\, r>0$$

which is weaker than McDiarmid's inequality (3.3.36) since the constant in the exponent here is smaller by a factor of 8.

3.E Fano's inequality for list decoding

The following generalization of Fano's inequality has been used in the proof of Theorem 3.6.4: 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 $L\colon \mathcal{Y} \to 2^{\mathcal{X}}$ which maps any $y \in \mathcal{Y}$ to a set $L(y) \subseteq \mathcal{X}$, such that $|L(Y)| \le N$ a.s.. Let $P_{\mathrm{e}} = \mathbb{P}\,(X \notin L(Y))$. Then

$$H(X|Y) \le h(P_{\mathrm{e}}) + (1-P_{\mathrm{e}}) \ln N + P_{\mathrm{e}} \ln |\mathcal{X}| \tag{3.E.1}$$

(see, e.g., [167] or [177, Lemma 1]).

To prove (3.E.1), 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

$$\begin{aligned} H(E,X|Y) &= H(E|Y) + H(X|E,Y) && \text{(3.E.2a)} \\ &= H(X|Y) + H(E|X,Y). && \text{(3.E.2b)} \end{aligned}$$

Since X and Y uniquely determine E (for the given L), the quantity on the right-hand side of (3.E.2b) is equal to $H(X|Y)$. On the other hand, we can upper-bound the right-hand side of (3.E.2a) as

$$\begin{aligned} H(E|Y) + H(X|E,Y) &\le H(E) + H(X|E,Y) \\ &\le h(P_{\mathrm{e}}) + (1-P_{\mathrm{e}}) \ln N + P_{\mathrm{e}} \ln |\mathcal{X}|, \end{aligned}$$

where we have bounded the conditional entropy $H(X|E,Y)$ as follows:

$$\begin{aligned} H(X|E,Y) &= \sum_{y \in \mathcal{Y}} \mathbb{P}(E=0,Y=y) H(X|E=0,Y=y) \\ &\qquad + \sum_{y \in \mathcal{Y}} \mathbb{P}(E=1,Y=y) H(X|E=1,Y=y) \\ &\le \sum_{y \in \mathcal{Y}} \mathbb{P}(E=0,Y=y) H(X|E=0,Y=y) + P_{\mathrm{e}} \ln |\mathcal{X}| \\ &= (1-P_{\mathrm{e}}) \sum_{y \in \mathcal{Y}} H(X|E=0,Y=y) \mathbb{P}(Y=y|E=0) \\ &\qquad + P_{\mathrm{e}} \ln |\mathcal{X}| \\ &\le (1-P_{\mathrm{e}}) \mathbb{E}\big[\ln |L(Y)| \,\big|\, E=0 \big] + P_{\mathrm{e}} \ln |\mathcal{X}| \\ &\le (1-P_{\mathrm{e}}) \ln N + P_{\mathrm{e}} \ln |\mathcal{X}|, \end{aligned}$$

where in the first line we have used the standard log-cardinality bound on the entropy, while in the third line we have used the fact that, given $E = 0$ and $Y = y$, X is supported on the set $L(y)$. Since $H(X|Y) = H(E) + H(X|E,Y) \le H(E) + H(X|E,Y)$, we get (3.E.1).

Remark 3.38. If instead of assuming that $L(Y)$ is bounded a.s. we assume that it is bounded in expectation, i.e., if $\mathbb{E}[\ln|L(Y)|] < \infty$, then we can obtain a weaker inequality

$$H(X|Y) \le \mathbb{E}\,[\ln|L(Y)|] + h(P_{\mathrm{e}}) + P_{\mathrm{e}} \ln|\mathcal{X}|.$$

To get this, we follow the same steps as before, except the last step in the above series of bounds on $H(X|E,Y)$ is replaced by

$$\begin{aligned}(1-P_{\mathrm{e}})\mathbb{E}\big[\ln|L(Y)|\,\big|\,E=0\big] &\le (1-P_{\mathrm{e}})\,\mathbb{E}\big[\ln|L(Y)|\,\big|\,E=0\big] \\ &\quad + P_{\mathrm{e}}\,\mathbb{E}\big[\ln|L(Y)|\,\big|\,E=1\big] \\ &= \mathbb{E}\,[\ln|L(Y)|]\end{aligned}$$

(we assume, of course, that $L(y)$ is nonempty for all $y \in \mathcal{Y}$).

3.F Details for the derivation of (3.6.16)

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 $T\colon \mathcal{X} \to \mathcal{Y}$, 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

$$\begin{aligned}P_{Y^n|X^n}(y^n|x^n) &= \prod_{i=1}^{n} P_{Y|X}(y_i|x_i), \\ &\qquad \forall y^n \in \mathcal{Y}^n,\ \forall x^n \in \mathcal{X}^n \text{ s.t. } P_{X^n}(x^n) > 0.\end{aligned}$$

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 $i = 1, \dots, n$, so we can write

$$P_{Y_i|\overline{Y}^i}(y|\overline{y}^i) = \sum_{x\in\mathcal{X}} P_{Y_i|X_i}(y|x) P_{X_i|\overline{Y}^i}(x|\overline{y}^i) \tag{3.F.1}$$

$$= \sum_{x\in\mathcal{X}} P_{Y|X}(y|x) P_{X_i|\overline{Y}^i}(x|\overline{y}^i) \tag{3.F.2}$$

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

$$\begin{aligned}\ln \frac{P_{Y_i|\overline{Y}^i}(y|\overline{y}^i)}{P_{Y_i|\overline{Y}^i}(y'|\overline{y}^i)} &= \ln \frac{\sum_{x\in\mathcal{X}} P_{Y|X}(y|x) P_{X_i|\overline{Y}^i}(x|\overline{y}^i)}{\sum_{x\in\mathcal{X}} P_{Y|X}(y'|x) P_{X_i|\overline{Y}^i}(x|\overline{y}^i)} \\ &= \ln \frac{\sum_{x\in\mathcal{X}} P_{Y|X}(y'|x) P_{X_i|\overline{Y}^i}(x|\overline{y}^i) \frac{P_{Y|X}(y|x)}{P_{Y|X}(y'|x)}}{\sum_{x\in\mathcal{X}} P_{Y|X}(y'|x) P_{X_i|\overline{Y}^i}(x|\overline{y}^i)},\end{aligned}$$

where in the last line we have used the fact that $P_{Y|X}(\cdot|\cdot) > 0$. This shows that we can express the quantity $\ln \dfrac{P_{Y_i|\overline{Y}^i}(y|\overline{y}^i)}{P_{Y_i|\overline{Y}^i}(y'|\overline{y}^i)}$ as the logarithm of expectation of $\dfrac{P_{Y|X}(y|X)}{P_{Y|X}(y'|X)}$ with respect to the (conditional) probability measure

$$Q(x|\overline{y}^i, y') = \frac{P_{Y|X}(y'|x) P_{X_i|\overline{Y}^i}(x|\overline{y}^i)}{\sum_{x\in\mathcal{X}} P_{Y|X}(y'|x) P_{X_i|\overline{Y}^i}(x|\overline{y}^i)}, \qquad \forall x \in \mathcal{X}.$$

Therefore,

$$\ln \frac{P_{Y_i|\overline{Y}^i}(y|\overline{y}^i)}{P_{Y_i|\overline{Y}^i}(y'|\overline{y}^i)} \le \max_{x\in\mathcal{X}} \ln \frac{P_{Y|X}(y|x)}{P_{Y|X}(y'|x)}.$$

Interchanging the roles of y and y', we get

$$\ln \frac{P_{Y_i|\overline{Y}^i}(y'|\overline{y}^i)}{P_{Y_i|\overline{Y}^i}(y|\overline{y}^i)} \le \max_{x\in\mathcal{X}} \ln \frac{P_{Y|X}(y'|x)}{P_{Y|X}(y|x)}.$$

This implies, in turn, that

$$\left|\ln \frac{P_{Y_i|\overline{Y}^i}(y|\overline{y}^i)}{P_{Y_i|\overline{Y}^i}(y'|\overline{y}^i)}\right| \le \max_{x\in\mathcal{X}} \max_{y,y'\in\mathcal{Y}} \left|\ln \frac{P_{Y|X}(y|x)}{P_{Y|X}(y'|x)}\right| = \frac{1}{2} c(T)$$

for all $y, y' \in \mathcal{Y}$.

Acknowledgments

The authors would like to thank several individuals, who have carefully read parts of the manuscript in its multiple stages and provided multiple constructive comments, suggestions, and corrections. These include Ronen Eshel, Peter Harremoës, Eran Hof, Nicholas Kalouptsidis, Leor Kehaty, Aryeh Kontorovich, Ioannis Kontoyannis, Mokshay Madiman, Daniel Paulin, Yury Polyanskiy, Boaz Shuval, Emre Telatar, Tim van Erven, Sergio Verdú, Yihong Wu and Kostis Xenoulis. Among these people, Leor Kehaty is gratefully acknowledged for a detailed report on the initial manuscript. The authors wish also to gratefully acknowledge the three anonymous reviewers for their very constructive and detailed suggestions, which contributed a lot to the presentation of this manuscript in its advanced stages. The authors accept full responsibility for any remaining omissions or errors.

The work of M. Raginsky was supported in part by the U.S. National Science Foundation (NSF) under CAREER award no. CCF–1254041. The work of I. Sason was supported by the Israeli Science Foundation (ISF), grant number 12/12. The hospitality of the Bernoulli inter-faculty center at EPFL (in Lausanne, Switzerland) during the summer of 2011 is acknowledged by I. Sason.

References

[1] M. Talagrand. A new look at independence. *Annals of Probability*, 24(1):1–34, January 1996.

[2] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities - A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[3] M. Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2001.

[4] G. Lugosi. Concentration of measure inequalities - lecture notes, 2009. URL: `http://www.econ.upf.edu/~lugosi/anu.pdf`.

[5] P. Massart. *The Concentration of Measure Phenomenon*, volume 1896 of *Lecture Notes in Mathematics*. Springer, 2007.

[6] C. McDiarmid. Concentration. In *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 195–248. Springer, 1998.

[7] M. Talagrand. Concentration of measure and isoperimteric inequalities in product space. *Publications Mathématiques de l'I.H.E.S*, 81:73–205, 1995.

[8] K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 19:357–367, 1967.

[9] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963.

[10] N. Alon and J. H. Spencer. *The Probabilistic Method*. Wiley Series in Discrete Mathematics and Optimization, third edition, 2008.

[11] F. Chung and L. Lu. *Complex Graphs and Networks*, volume 107 of *Regional Conference Series in Mathematics*. Wiley, 2006.

[12] F. Chung and L. Lu. Concentration inequalities and martingale inequalities: a survey. *Internet Mathematics*, 3(1):79–127, March 2006. URL: `http://www.ucsd.edu/~fan/wp/concen.pdf`.

[13] T. J. Richardson and R. Urbanke. *Modern Coding Theory*. Cambridge University Press, 2008.

[14] Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer. PAC-Bayesian inequalities for martingales. *IEEE Trans. on Information Theory*, 58(12):7086–7093, December 2012.

[15] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, August 2012.

[16] J. A. Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, March 2011.

[17] N. Gozlan and C. Leonard. Transport inequalities: a survey. *Markov Processes and Related Fields*, 16(4):635–736, 2010.

[18] J. M. Steele. *Probability Theory and Combinatorial Optimization*, volume 69 of *CBMS–NSF Regional Conference Series in Applied Mathematics*. Siam, Philadelphia, PA, USA, 1997.

[19] A. Dembo. Information inequalities and concentration of measure. *Annals of Probability*, 25(2):927–939, 1997.

[20] S. Chatterjee. *Concentration Inequalities with Exchangeable Pairs*. PhD thesis, Stanford University, California, USA, February 2008. URL: `http://arxiv.org/abs/0507526`.

[21] S. Chatterjee. Stein's method for concentration inequalities. *Probability Theory and Related Fields*, 138:305–321, 2007.

[22] S. Chatterjee and P. S. Dey. Applications of Stein's method for concentration inequalities. *Annals of Probability*, 38(6):2443–2485, June 2010.

[23] N. Ross. Fundamentals of Stein's method. *Probability Surveys*, 8:210–293, 2011.

[24] E. Abbe and A. Montanari. On the concentration of the number of solutions of random satisfiability formulas, 2010. URL: `http://arxiv.org/abs/1006.3786`.

[25] 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*, pages 2801–2805, Nice, France, June 2007.

[26] S. B. Korada, S. Kudekar, and N. Macris. Concentration of magnetization for linear block codes. In *Proceedings of the 2008 IEEE International Symposium on Information Theory*, pages 1433–1437, Toronto, Canada, July 2008.

[27] S. Kudekar. *Statistical Physics Methods for Sparse Graph Codes*. PhD thesis, EPFL - Swiss Federal Institute of Technology, Lausanne, Switzeland, July 2009. URL: `http://infoscience.epfl.ch/record/138478/files/EPFL_TH4442.pdf`.

[28] S. Kudekar and N. Macris. Sharp bounds for optimal decoding of low-density parity-check codes. *IEEE Trans. on Information Theory*, 55(10):4635–4650, October 2009.

[29] S. B. Korada and N. Macris. Tight bounds on the capacity of binary input random CDMA systems. *IEEE Trans. on Information Theory*, 56(11):5590–5613, November 2010.

[30] A. Montanari. Tight bounds for LDPC and LDGM codes under MAP decoding. *IEEE Trans. on Information Theory*, 51(9):3247–3261, September 2005.

[31] M. Talagrand. *Mean Field Models for Spin Glasses*. Springer-Verlag, 2010.

[32] S. Bobkov and M. Madiman. Concentration of the information in data with log-concave distributions. *Annals of Probability*, 39(4):1528–1543, 2011.

[33] S. Bobkov and M. Madiman. The entropy per coordinate of a random vector is highly constrained under convexity conditions. *IEEE Trans. on Information Theory*, 57(8):4940–4954, August 2011.

[34] E. Shamir and J. Spencer. Sharp concentration of the chromatic number on random graphs. *Combinatorica*, 7(1):121–129, 1987.

[35] M. G. Luby, Mitzenmacher, M. A. Shokrollahi, and D. A. Spielmann. Efficient erasure-correcting codes. *IEEE Trans. on Information Theory*, 47(2):569–584, February 2001.

[36] T. J. Richardson and R. Urbanke. The capacity of low-density parity-check codes under message-passing decoding. *IEEE Trans. on Information Theory*, 47(2):599–618, February 2001.

[37] M. Sipser and D. A. Spielman. Expander codes. *IEEE Trans. on Information Theory*, 42(6):1710–1722, November 1996.

[38] A. B. Wagner, P. Viswanath, and S. R. Kulkarni. Probability estimation in the rare-events regime. *IEEE Trans. on Information Theory*, 57(6):3207–3229, June 2011.

[39] C. McDiarmid. Centering sequences with bounded differences. *Combinatorics, Probability and Computing*, 6(1):79–86, March 1997.

[40] 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*, pages 32–36, Volos, Greece, June 2009.

[41] K. Xenoulis and N. Kalouptsidis. Achievable rates for nonlinear Volterra channels. *IEEE Trans. on Information Theory*, 57(3):1237–1248, March 2011.

[42] 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*, pages 1430–1434, MIT, Boston, MA, USA, July 2012.

[43] M. Ledoux. On Talagrand's deviation inequalities for product measures. *ESAIM: Probability and Statistics*, 1:63–87, 1997.

[44] L. Gross. Logarithmic Sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.

[45] A. J. Stam. Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Information and Control*, 2:101–112, 1959.

[46] P. Federbush. A partially alternate derivation of a result of Nelson. *Journal of Mathematical Physics*, 10(1):50–52, 1969.

[47] A. Dembo, T. M. Cover, and J. A. Thomas. Information theoretic inequalities. *IEEE Trans. on Information Theory*, 37(6):1501–1518, November 1991.

[48] C. Villani. A short proof of the 'concavity of entropy power'. *IEEE Trans. on Information Theory*, 46(4):1695–1696, July 2000.

[49] G. Toscani. An information-theoretic proof of Nash's inequality. *Rendiconti Lincei: Matematica e Applicazioni*, 2012. in press.

[50] A. Guionnet and B. Zegarlinski. Lectures on logarithmic Sobolev inequalities. *Séminaire de probabilités (Strasbourg)*, 36:1–134, 2002.

[51] M. Ledoux. Concentration of measure and logarithmic Sobolev inequalities. In *Séminaire de Probabilités XXXIII*, volume 1709 of *Lecture Notes in Math.*, pages 120–216. Springer, 1999.

[52] G. Royer. *An Invitation to Logarithmic Sobolev Inequalities*, volume 14 of *SFM/AMS Texts and Monographs*. American Mathematical Society and Société Mathématiques de France, 2007.

[53] S. G. Bobkov and F. Götze. Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. *Journal of Functional Analysis*, 163:1–28, 1999.

[54] S. G. Bobkov and M. Ledoux. On modified logarithmic Sobolev inequalities for Bernoulli and Poisson measures. *Journal of Functional Analysis*, 156(2):347–365, 1998.

[55] S. G. Bobkov and P. Tetali. Modified logarithmic Sobolev inequalities in discrete settings. *Journal of Theoretical Probability*, 19(2):289–336, 2006.

[56] D. Chafaï. Entropies, convexity, and functional inequalities: Φ-entropies and Φ-Sobolev inequalities. *J. Math. Kyoto University*, 44(2):325–363, 2004.

[57] C. P. Kitsos and N. K. Tavoularis. Logarithmic Sobolev inequalities for information measures. *IEEE Trans. on Information Theory*, 55(6):2554–2561, June 2009.

[58] K. Marton. Bounding $\bar{d}$-distance by informational divergence: a method to prove measure concentration. *Annals of Probability*, 24(2):857–866, 1996.

[59] C. Villani. *Topics in Optimal Transportation*. American Mathematical Society, Providence, RI, 2003.

[60] C. Villani. *Optimal Transport: Old and New*. Springer, 2008.

[61] P. Cattiaux and A. Guillin. On quadratic transportation cost inequalities. *Journal de Matématiques Pures et Appliquées*, 86:342–361, 2006.

[62] A. Dembo and O. Zeitouni. Transportation approach to some concentration inequalities in product spaces. *Electronic Communications in Probability*, 1:83–90, 1996.

[63] H. Djellout, A. Guillin, and L. Wu. Transportation cost-information inequalities and applications to random dynamical systems and diffusions. *Annals of Probability*, 32(3B):2702–2732, 2004.

[64] N. Gozlan. A characterization of dimension free concentration in terms of transportation inequalities. *Annals of Probability*, 37(6):2480–2498, 2009.

[65] E. Milman. Properties of isoperimetric, functional and transport-entropy inequalities via concentration. *Probability Theory and Related Fields*, 152:475–507, 2012.

[66] R. M. Gray, D. L. Neuhoff, and P. C. Shields. A generalization of Ornstein's $\bar{d}$ distance with applications to information theory. *Annals of Probability*, 3(2):315–328, 1975.

[67] 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*, 21(5):524–532, September 1975.

[68] Y. Steinberg and S. Verdú. Simulation of random processes and rate-distortion theory. *IEEE Trans. on Information Theory*, 42(1):63–86, January 1996.

[69] R. Ahlswede, P. Gács, and J. Körner. Bounds on conditional probabilities with applications in multi-user communication. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 34:157–177, 1976. See correction in vol. 39, no. 4, pp. 353–354, 1977.

[70] R. Ahlswede and G. Dueck. Every bad code has a good subcode: a local converse to the coding theorem. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 34:179–182, 1976.

[71] K. Marton. A simple proof of the blowing-up lemma. *IEEE Trans. on Information Theory*, 32(3):445–446, May 1986.

[72] Y. Altuğ and A. B. Wagner. Refinement of the sphere-packing bound: asymmetric channels, 2012. URL: `http://arxiv.org/abs/1211.6997`.

[73] A. Amraoui, A. Montanari, T. Richardson, and R. Urbanke. Finite-length scaling for iteratively decoded LDPC ensembles. *IEEE Trans. on Information Theory*, 55(2):473–498, February 2009.

[74] T. Nozaki, K. Kasai, and K. Sakaniwa. Analytical solution of covariance evolution for irregular LDPC codes. *IEEE Trans. on Information Theory*, 58(7):4770–4780, July 2012.

[75] I. Kontoyiannis and S. Verdú. Lossless data compression at finite blocklengths, 2012. URL: `http://arxiv.org/abs/1212.2668`.

[76] V. Kostina and S. Verdú. Fixed-length lossy compression in the finite blocklength regime. *IEEE Trans. on Information Theory*, 58(6):3309–3338, June 2012.

[77] W. Matthews. A linear program for the finite block length converse of Polyanskiy-Poor-Verdú via nonsignaling codes. *IEEE Trans. on Information Theory*, (12):7036–7044, December 2012.

[78] Y. Polyanskiy, H. V. Poor, and S. Verdú. Channel coding rate in finite blocklength regime. *IEEE Trans. on Information Theory*, 56(5):2307–2359, May 2010.

[79] G. Wiechman and I. Sason. An improved sphere-packing bound for finite-length codes on symmetric channels. *IEEE Trans. on Information Theory*, 54(5):1962–1990, 2008.

[80] J. S. Rosenthal. *A First Look at Rigorous Probability Theory*. World Scientific, second edition, 2006.

[81] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Springer, second edition, 1997.

[82] H. Chernoff. A measure of asymptotic efficiency of tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23(4):493–507, 1952.

[83] S. N. Bernstein. *The Theory of Probability*. Gos. Izdat., Moscow/Leningrad, 1927. in Russian.

[84] S. Verdú. *Multiuser Detection*. Cambridge University Press, 1998.

[85] C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, volume 141, pages 148–188. Cambridge University Press, 1989.

[86] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

[87] 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*, pages 311–319, San-Francisco, CA, USA, March 16-18 1998.

[88] D. Berend and A. Kontorovich. On the concentration of the missing mass. *Electronic Communications in Probability*, 18(3):1–7, January 2013.

[89] S. G. From and A. W. Swift. A refinement of Hoeffding's inequality. *Journal of Statistical Computation and Simulation*, pages 1–7, December 2011.

[90] P. Billingsley. *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics, 3rd edition, 1995.

[91] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. Oxford University Press, third edition, 2001.

[92] 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*, pages 1563–1567, Adelaide, Australia, September 2005.

[93] A. Barg and G. D. Forney. Random codes: minimum distances and error exponents. *IEEE Trans. on Information Theory*, 48(9):2568–2573, September 2002.

[94] M. Breiling. A logarithmic upper bound on the minimum distance of turbo codes. *IEEE Trans. on Information Theory*, 50(8):1692–1710, August 2004.

[95] R. G. Gallager. *Low-Density Parity-Check Codes*. PhD thesis, MIT, Cambridge, MA, USA, 1963.

[96] T. Etzion, A. Trachtenberg, and A. Vardy. Which codes have cycle-free Tanner graphs? *IEEE Trans. on Information Theory*, 45(6):2173–2181, September 1999.

[97] I. Sason. On universal properties of capacity-approaching LDPC code ensembles. *IEEE Trans. on Information Theory*, 55(7):2956–2990, July 2009.

[98] 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*, 47(2):585–598, February 2001.

[99] A. Kavčić, X. Ma, and M. Mitzenmacher. Binary intersymbol interference channels: Gallager bounds, density evolution, and code performance bounds. *IEEE Trans. on Information Theory*, 49(7):1636–1652, July 2003.

[100] 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.

[101] J. Douillard, M. Jezequel, C. Berrou, A. Picart, P. Didier, and A. Glavieux. Iterative correction of intersymbol interference: turbo-equalization. *European Transactions on Telecommunications*, 6(1):507–511, September 1995.

[102] C. Méasson, A. Montanari, and R. Urbanke. Maxwell construction: the hidden bridge between iterative and maximum apposteriori decoding. *IEEE Trans. on Information Theory*, 54(12):5277–5307, December 2008.

[103] A. Shokrollahi. Capacity-achieving sequences. In *Volume in Mathematics and its Applications*, volume 123, pages 153–166, 2000.

[104] A. F. Molisch. *Wireless Communications*. John Wiley and Sons, 2005.

[105] 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.

[106] 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*, 52(3):992–1006, March 2006.

[107] R. Salem and A. Zygmund. Some properties of trigonometric series whose terms have random signs. *Acta Mathematica*, 91(1):245–301, 1954.

[108] G. Wunder and H. Boche. New results on the statistical distribution of the crest-factor of OFDM signals. *IEEE Trans. on Information Theory*, 49(2):488–494, February 2003.

[109] I. Sason. On the concentration of the crest factor for OFDM signals. In *Proceedings of the 8th International Symposium on Wireless Communication Systems (ISWCS '11)*, pages 784–788, Aachen, Germany, November 2011.

[110] S. Benedetto and E. Biglieri. *Principles of Digital Transmission with Wireless Applications*. Kluwer Academic/ Plenum Publishers, 1999.

[111] I. Sason. Tightened exponential bounds for discrete-time conditionally symmetric martingales with bounded jumps. *Statistics and Probability Letters*, 83(8):1928–1936, August 2013.

[112] X. Fan, I. Grama, and Q. Liu. Hoeffding's inequality for supermartingales, 2011. URL: `http://arxiv.org/abs/1109.4359.`

[113] X. Fan, I. Grama, and Q. Liu. The missing factor in Bennett's inequality, 2012. URL: `http://arxiv.org/abs/1206.2592.`

[114] I. Sason and S. Shamai. *Performance Analysis of Linear Codes under Maximum-Likelihood Decoding: A Tutorial*, volume 3 of Foundations and Trends in Communications and Information Theory. Now Publishers, Delft, the Netherlands, July 2006.

[115] E. B. Davies and B. Simon. Ultracontractivity and the heat kernel for Schrödinger operators and Dirichlet Laplacians. *Journal of Functional Analysis*, 59(335-395), 1984.

[116] S. Verdú and T. Weissman. The information lost in erasures. *IEEE Trans. on Information Theory*, 54(11):5030–5058, November 2008.

[117] E. A. Carlen. Superadditivity of Fisher's information and logarithmic Sobolev inequalities. *Journal of Functional Analysis*, 101:194–211, 1991.

[118] R. A. Adams and F. H. Clarke. Gross's logarithmic Sobolev inequality: a simple proof. *American Journal of Mathematics*, 101(6):1265–1269, December 1979.

[119] G. Blower. *Random Matrices: High Dimensional Phenomena*. London Mathematical Society Lecture Notes. Cambridge University Press, Cambridge, U.K., 2009.

[120] O. Johnson. *Information Theory and the Central Limit Theorem*. Imperial College Press, London, 2004.

[121] E. H. Lieb and M. Loss. *Analysis*. American Mathematical Society, Providence, RI, 2nd edition, 2001.

[122] 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*, 30(6):837–839, November 1984.

[123] P. J. Huber and E. M. Ronchetti. *Robust Statistics*. Wiley Series in Probability and Statistics, second edition, 2009.

[124] O. Johnson and A. Barron. Fisher information inequalities and the central limit theorem. *Probability Theory and Related Fields*, 129:391–409, 2004.

[125] S. Verdú. Mismatched estimation and relative entropy. *IEEE Trans. on Information Theory*, 56(8):3712–3720, August 2010.

[126] H. L. van Trees. *Detection, Estimation and Modulation Theory, Part I*. Wiley, 1968.

[127] L. C. Evans and R. F. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, 1992.

[128] M. C. Mackey. *Time's Arrow: The Origins of Thermodynamic Behavior*. Springer, New York, 1992.

[129] B. Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, Berlin, 5 edition, 1998.

[130] I. Karatzas and S. Shreve. *Brownian Motion and Stochastic Calculus*. Springer, second edition, 1988.

[131] F. C. Klebaner. *Introduction to Stochastic Calculus with Applications*. Imperial College Press, second edition, 2005.

[132] T. van Erven and P. Harremoës. Rényi divergence and Kullback–Leibler divergence. *IEEE Trans. on Information Theory*, 2012. submitted, 2012. URL: `http://arxiv.org/abs/1206.2459`.

[133] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley and Sons, second edition, 2006.

[134] A. Maurer. Thermodynamics and concentration. *Bernoulli*, 18(2):434–454, 2012.

[135] N. Merhav. *Statistical Physics and Information Theory*, volume 6 of Foundations and Trends in Communications and Information Theory. Now Publishers, Delft, the Netherlands, 2009.

[136] S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities using the entropy method. *Annals of Probability*, 31(3):1583–1614, 2003.

[137] I. Kontoyiannis and M. Madiman. Measure concentration for compound Poisson distributions. *Electronic Communications in Probability*, 11:45–57, 2006.

[138] B. Efron and C. Stein. The jackknife estimate of variance. *Annals of Statistics*, 9:586–596, 1981.

[139] J. M. Steele. An Efron–Stein inequality for nonsymmetric statistics. *Annals of Statistics*, 14:753–758, 1986.

[140] M. Gromov. *Metric Structures for Riemannian and Non-Riemannian Spaces*. Birkhäuser, 2001.

[141] S. Bobkov. A functional form of the isoperimetric inequality for the Gaussian measure. *Journal of Functional Analysis*, 135:39–49, 1996.

[142] L. V. Kantorovich. On the translocation of masses. *Journal of Mathematical Sciences*, 133(4):1381–1382, 2006.

[143] E. Ordentlich and M. Weinberger. A distribution dependent refinement of Pinsker's inequality. *IEEE Trans. on Information Theory*, 51(5):1836–1840, May 2005.

[144] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. J. Weinberger. Inequalities for the L_1 deviation of the empirical distribution. Technical Report HPL-2003-97 (R.1), Information Theory Research Group, HP Laboratories, Palo Alto, CA, June 2003.

[145] D. Berend, P. Harremoës, and A. Kontorovich. A reverse Pinsker inequality, 2012. URL: `http://arxiv.org/abs/1206.6544`.

[146] I. Sason. Improved lower bounds on the total variation distance and relative entropy for the Poisson approximation. In *Proceedings of the 2013 IEEE Information Theory and Applications (ITA) Workshop*, pages 1–4, San-Diego, California, USA, February 2013.

[147] I. Kontoyiannis, P. Harremoës, and O. Johnson. Entropy and the law of small numbers. *IEEE Trans. on Information Theory*, 51(2):466–472, February 2005.

[148] I. Csiszár. Sanov property, generalized I-projection and a conditional limit theorem. *Annals of Probability*, 12(3):768–793, 1984.

[149] P. Dupuis and R. S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations.* Wiley Series in Probability and Statistics, New York, 1997.

[150] M. Talagrand. Transportation cost for Gaussian and other product measures. *Geometry and Functional Analysis*, 6(3):587–600, 1996.

[151] R. M. Dudley. *Real Analysis and Probability.* Cambridge University Press, 2004.

[152] F. Otto and C. Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173:361–400, 2000.

[153] Y. Wu. On the HWI inequality. a work in progress.

[154] D. Cordero-Erausquin. Some applications of mass transport to Gaussian-type inequalities. *Arch. Rational Mech. Anal.*, 161:257–269, 2002.

[155] D. Bakry and M. Emery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX*, volume 1123 of *Lecture Notes in Mathematics*, pages 177–206. Springer, 1985.

[156] P.-M. Samson. Concentration of measure inequalities for Markov chains and ϕ-mixing processes. *Annals of Probability*, 28(1):416–461, 2000.

[157] K. Marton. A measure concentration inequality for contracting Markov chains. *Geometric and Functional Analysis*, 6:556–571, 1996. See also erratum in *Geometric and Functional Analysis*, vol. 7, pp. 609–613, 1997.

[158] K. Marton. Measure concentration for Euclidean distance in the case of dependent random variables. *Annals of Probability*, 32(3B):2526–2544, 2004.

[159] K. Marton. Correction to 'Measure concentration for Euclidean distance in the case of dependent random variables'. *Annals of Probability*, 38(1):439–442, 2010.

[160] K. Marton. Bounding relative entropy by the relative entropy of local specifications in product spaces, 2009. URL: `http://arxiv.org/abs/0907.4491`.

[161] K. Marton. An inequality for relative entropy and logarithmic Sobolev inequalities in Euclidean spaces. *Journal of Functional Analysis*, 264:34–61, 2013.

[162] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2nd edition, 2011.

[163] G. Margulis. Probabilistic characteristics of graphs with large connectivity. *Problems of Information Transmission*, 10(2):174–179, 1974.

[164] A. El Gamal and Y.-H. Kim. *Network Information Theory*. Cambridge University Press, 2011.

[165] G. Dueck. Maximal error capacity regions are smaller than average error capacity regions for multi-user channels. *Problems of Control and Information Theory*, 7(1):11–19, 1978.

[166] 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*, 19(4):339–347, 1990.

[167] R. Ahlswede and J. Körner. Source coding with side information and a converse for degraded broadcast channels. *IEEE Trans. on Information Theory*, 21(6):629–637, November 1975.

[168] S. Shamai and S. Verdú. The empirical distribution of good codes. *IEEE Trans. on Information Theory*, 43(3):836–846, May 1997.

[169] T. S. Han and S. Verdú. Approximation theory of output statistics. *IEEE Trans. on Information Theory*, 39(3):752–772, May 1993.

[170] Y. Polyanskiy and S. Verdú. Empirical distribution of good channel codes with non-vanishing error probability. To appear in the *IEEE Trans. on Information Theory*. URL: `http://arxiv.org/abs/1309.0141`.

[171] F. Topsøe. An information theoretical identity and a problem involving capacity. *Studia Scientiarum Mathematicarum Hungarica*, 2:291–292, 1967.

[172] J. H. B. Kemperman. On the Shannon capacity of an arbitrary channel. *Indagationes Mathematicae*, 36:101–115, 1974.

[173] U. Augustin. Gedächtnisfreie Kanäle für diskrete Zeit. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 6:10–61, 1966.

[174] R. Ahlswede. An elementary proof of the strong converse theorem for the multiple-access channel. *Journal of Combinatorics, Information and System Sciences*, 7(3):216–230, 1982.

[175] S. Shamai and I. Sason. Variations on the Gallager bounds, connections and applications. *IEEE Trans. on Information Theory*, 48(12):3029–3051, December 2001.

[176] Y. Kontoyiannis. Sphere-covering, measure concentration, and source coding. *IEEE Trans. on Information Theory*, 47(4):1544–1552, May 2001.

[177] Y. H. Kim, A. Sutivong, and T. M. Cover. State amplification. *IEEE Trans. on Information Theory*, 54(5):1850–1859, May 2008.