Concentration of Measure Inequalities in Information Theory, Communications, and Coding: *Second Edition*

Concentration of Measure Inequalities in Information Theory, Communications, and Coding: *Second Edition*

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:*
Second Edition. Foundations and Trends® in Communications and Information Theory, vol. 10, no. 1-2, pp. 1–259, 2014.

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

ISBN: 978-1-60198-906-2

Foundations and Trends® in Communications and Information Theory
Volume 10, Issue 1-2, 2014

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, 2014, 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 (2014) 1–259

DOI: 10.1561/0100000064

now
the essence of knowledge

Concentration of Measure Inequalities in Information Theory, Communications, and Coding: *Second Edition*

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

Concentration-of-measure inequalities are studied in order to gain an understanding on the fluctuations of complicated random objects. These inequalities have been considerably developed during the last four decades, playing a significant role in various fields which include probability theory, functional analysis, geometry, high-dimensional statistics, information theory, learning theory, statistical physics, and theoretical computer science.

This monograph is focused on several key modern mathematical tools which are used for the derivation of concentration inequalities, on their links to information theory, and a sample of their applications to information theory, communications and coding. In addition to serving as a survey, it also includes new recent results derived by the authors, and new information-theoretic proofs of published results.

The first part of the monograph introduces classical concentration inequalities for martingales, including some of their recent refinements and extensions. The power and versatility of the martingale approach is mainly exemplified in the context of coding theory, random graphs, and codes defined on graphs and iterative decoding algorithms.

Its second part introduces the entropy method, an information-theoretic approach for the derivation of concentration inequalities. The basic ingredients of the entropy method are discussed 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 which is based on transportation-cost inequalities and probability in metric spaces. Representative results on concentration for dependent random variables are briefly summarized, with emphasis on their connections to the entropy method. We discuss applications of the entropy method to information theory, including strong converses, empirical distributions of good channel codes, and an information-theoretic converse for concentration of measure.

Foundations and Trends® in Communications and Information Theory
Vol. 10, No. 1-2, pp. 1–259 (2014)

DOI: 10.1561/0100000064

1

Introduction

1.1 An overview and a brief history

Concentration-of-measure inequalities provide upper bounds on the probability that a random variable X deviates from its mean, median or any other typical value $\overline{x}$ by a given amount. These inequalities have been studied for several decades, with several fundamental and substantial contributions. Very roughly speaking, the concentration of measure phenomenon was stated by Talagrand 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 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 $t > 0$, 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, 7].

Concentration-of-measure inequalities are studied in order to gain an understanding on the fluctuations of complicated random objects. These inequalities have been considerably developed during the last four decades, playing a significant role in various fields which include probability theory, functional analysis, geometry, high-dimensional statistics, information theory, learning theory, statistical physics, and theoretical computer science. Several techniques have been developed so far to prove concentration of measure inequalities. 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 is covered in Chapter 2, which is focused on concentration inequalities for discrete-time martingales with bounded differences, as well as on some of their applications in information theory, coding and communications. A recent interesting avenue that follows from the martingale-based concentration inequalities, presented in Chapter 2, is their generalization to random matrices (see, e.g., [15, 16]).

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

- Talagrand's inequalities for product measures (see, e.g., [1], [6, Chapter 4], [7] and [18, Chapter 6]) and their information-theoretic links [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) has been recently used to prove concentration inequalities (see, e.g., [20, 21, 22, 23, 24, 25, 26, 27, 28]).
- Concentration inequalities that follow from rigorous methods in statistical physics (see, e.g., [29, 30, 31, 32, 33, 34, 35, 36]).
- The so-called reverse Lyapunov inequalities were recently used to derive concentration inequalities for multi-dimensional log-concave distributions [37] (see also a related work in [38]). The concentration inequalities in [37] imply an extension of the Shannon–McMillan–Breiman strong ergodic theorem to the class of discrete-time processes with log-concave marginals.

The last three items are not addressed in this monograph. We now give a synopsis of some of the main ideas underlying the martingale approach (Chapter 2) and the entropy method (Chapter 3).

The Azuma–Hoeffding inequality, which is introduced in Chapter 2, is a useful tool for establishing concentration results for *discrete-time bounded-difference martingales*. This inequality is due to Hoeffding [9], who proved it for sums of independent and bounded random variables, and to Azuma [8], who extended it to bounded-difference martingales. This inequality was introduced into the computer science literature by Shamir and Spencer [39], proving the concentration of the chromatic number[1] for ensembles of random graphs. More specifically, they proved in [39] a concentration result of the chromatic number for the *Erdös–Rényi* ensemble of random graphs, characterized by the property that any pair of vertices in this graph is, independently to all other pairs of vertices, connected by an edge in probability $p \in (0, 1)$. This approach has been imported into coding theory in [40], [41] and [42], especially for exploring concentration of measure phenomena pertaining to codes defined on graphs (e.g., turbo and low-density parity-check codes) and their iterative message-passing decoding algorithms. The last decade has seen an ever-expanding use of the Azuma–Hoeffding inequality for

[1]The chromatic number of a graph is defined as the minimal number of colors required to color all the vertices of this graph such that no two adjacent vertices have the same color.

proving concentration inequalities in coding theory (see, e.g., [13] and references therein). All these concentration inequalities serve to justify theoretically the ensemble approach to codes defined on graphs; much stronger concentration of measure phenomena are, however, observed in practice.

Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a function that has *bounded differences*, i.e., the value of f changes by a bounded amount whenever any of its n input variables is changed while the others are held fixed. A common method for proving concentration of such a function of n independent random variables, around its expected value $\mathbb{E}[f]$, revolves around the so-called McDiarmid's inequality or "independent bounded-differences inequality" [6]. This inequality, introduced in Chapter 2, was originally proved via the martingale approach [6]; although its proof has some similarity to the proof of the Azuma–Hoeffding inequality, the bounded-difference assumption on f that is used for the derivation of the former inequality yields an improvement in the exponent by a factor of 4. Nice applications of martingale-based concentration inequalities in discrete mathematics and random graphs, based on the Azuma–Hoeffding and McDiarmid inequalities, are exemplified in [6], [10], [13] and [18].

Although the martingale approach can be used to assert a large variety of concentration of measure phenomena, as it was pointed out in [1, p. 10] "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." Chapter 3 of this monograph focuses on a different set of techniques, fundamentally rooted in information theory, which provide very strong concentration inequalities. These techniques, commonly referred to as the *entropy method*, have originated in the work of Michel Ledoux [43] by finding a different 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 technique, which bounds 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.

One of the best known log-Sobolev inequalities, first referred to as such by Leonard Gross [44], pertains to the standard n-dimensional Gaussian measure on the Euclidean space $\mathbb{R}^n$. This inequality gives an upper bound on the relative entropy $D(P\|G_n)$ between an arbitrary probability distribution P on $\mathbb{R}^n$ and the standard Gaussian measure G_n, expressed in terms of an "energy-like" quantity which is related to the squared norm of the gradient of the density of P with respect to G_n. Based on a clever analytic argument which Gross attributed to an unpublished note by Ira Herbst, he used his Gaussian log-Sobolev inequality to show that the logarithmic MGF $\Lambda(\lambda) \triangleq \ln \mathbb{E}[\exp(\lambda U)]$ of $U = f(X^n)$, where $X^n \sim G_n$ and $f\colon \mathbb{R}^n \to \mathbb{R}$ is an arbitrary 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 $X^n \sim G_n$ and $t > 0$. It should be pointed out that the Gaussian log-Sobolev inequality has a curious history, and it 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 the concavity of entropy power [47, 48, 49, 50].

In a nutshell, the entropy method takes this successful idea and applies it beyond the Gaussian case. In abstract terms, the log-Sobolev inequalities are functional inequalities that relate the relative entropy between an arbitrary distribution Q with respect to the distribution P of interest to some "energy functional" of the density $\frac{\mathrm{d}Q}{\mathrm{d}P}$. If one is interested in studying the concentration properties of some function $U = f(Z)$ with $Z \sim P$, then the core of the entropy method consists in applying an appropriate log-Sobolev inequality to the *tilted probability distributions* $P^{(\lambda f)}$ with $\frac{\mathrm{d}P^{(\lambda f)}}{\mathrm{d}P} \propto \exp(\lambda f)$ for $\lambda \in \mathbb{R}$. 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 to derive 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, 51, 52, 53], as well as in recent papers [54, 55, 56, 57, 58] and references therein.

Around the same time that Michel Ledoux introduced the entropy method [43], Katalin Marton showed in a breakthrough paper [59] that one can bypass functional inequalities and work directly on the level of probability measures (see also [60], written by Marton in the occasion of her 2013 Shannon Award lecture). More specifically, she showed that Gaussian concentration bounds can be deduced from the so-called *transportation-cost inequalities*. These inequalities, studied in Section 3.4, relate information-theoretic measures, such as the relative entropy, to a 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 a subject of intense research activity which play a prominent role in probability theory, statistics, functional analysis, dynamical systems, partial differential equations, statistical physics, differential geometry, and they have been also used in information theory (see, e.g., [61, 62, 63, 64, 65]). A great deal of information on the field of *optimal transportation* can be found in two books by Cédric Villani — [66] offers a concise and fairly elementary introduction, while the more recent book [67] is a lot more detailed and encyclopedic. Various connections between optimal transportation, concentration of measure, and information theory are also explored in [17, 19, 68, 69, 70, 71, 72].

The first explicit invocation of concentration inequalities in an information-theoretic context appears in the work of Ahlswede *et al.* [73, 74]. These authors demonstrated that a delicate probabilistic inequality, which was referred to as the "blowing up lemma", and which we now (thanks to the contributions by Marton [59, 75]) recognize as a Gaussian concentration bound in the Hamming space, can be used

to derive strong converses for a wide variety of information-theoretic problems, including multi-terminal scenarios. The importance of sharp concentration inequalities for characterizing fundamental limitations 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., [76, 77, 78, 79, 80, 81, 82, 83]). 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 graduate students and researchers in information theory and related fields.

1.2 A reader's guide

This monograph is focused on the interplay between concentration of measure inequalities and information theory, and on applications of concentration phenomena to problems related to information theory, communications and coding. For this reason, it is primarily aimed at researchers and graduate students working in information theory and related fields. The necessary mathematical background is real analysis, a first graduate course in probability theory and stochastic processes, and elementary functional analysis. As a refresher textbook for this mathematical background, the reader is referred, e.g., to [84].

Chapter 2 is focused on the derivation of concentration inequalities for martingales, and the use of these inequalities in communication and information-theoretic applications. This chapter has the following structure: Section 2.1 lists key definitions and basic facts pertaining to discrete-time sub/super-martingales, Section 2.2 provides the two basic ingredients for the derivation of concentration inequalities by using martingales, Section 2.3 introduces the Efron-Stein-Steele inequalities, and Section 2.4 presents basic inequalities that form the basis of the considered approach to concentration of measure. The concentration inequalities in Section 2.4 include the celebrated Azuma–Hoeffding and McDiarmid inequalities, and Section 2.5 is focused on the derivation of refined versions of the former inequality. Section 2.6 discusses the

connections of the concentration inequalities introduced in Section 2.5 to classical limit theorems in probability theory. Section 2.7 forms the second part of Chapter 2, applying earlier concentration inequalities to problems in information theory, communications and coding. A brief summary of Chapter 2 is given in Section 2.8.

Chapter 3 on the entropy method, log-Sobolev and transportation-cost inequalities is structured as follows: Section 3.1 introduces the main ingredients of the entropy method, and it 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. 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 [59, 75] 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 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. Applications of concentration of measures to problems in information theory are surveyed in Section 3.6, and finally 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 differences, and it provides several of their potential applications in information theory, digital communications and coding. It introduces the basic concentration inequalities of Efron-Stein-Steele, Azuma–Hoeffding, McDiarmid, and various refinements. It then moves to applications, which include concentration for random binary linear block codes, concentration for random regular bipartite graphs, concentration for low-density parity-check (LDPC) codes, and concentration for orthogonal-frequency-division-multiplexing (OFDM) signals.

2.1 Discrete-time martingales

This section provides a brief review of martingales to set definitions and notation.

Definition 2.1 (Discrete-time martingales)**.** Consider a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. A sequence $\{X_i, \mathscr{F}_i\}_{i=0}^{n}$, $n \in \mathbb{N}$, where the X_i's are random variables and the $\mathscr{F}_i$'s are σ-algebras, is a discrete-time martingale if the following conditions are satisfied:

1. The $\mathscr{F}_i$'s form a *filtration*, i.e., $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \ldots \subseteq \mathscr{F}_n \subseteq \mathscr{F}$; usually, $\mathscr{F}_0 = \{\emptyset, \Omega\}$ and $\mathscr{F}_n$ is the full σ-algebra $\mathscr{F}$.
2. $X_i \in \mathbb{L}^1(\Omega, \mathscr{F}_i, \mathbb{P})$ for every $i \in \{0, \ldots, n\}$; this means that each X_i is defined on the same sample space Ω, it is $\mathscr{F}_i$-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 following equality holds almost surely:

$$X_{i-1} = \mathbb{E}[X_i | \mathscr{F}_{i-1}]. \tag{2.1.1}$$

In general, relations between random variables such as $X = Y$, $X \leq Y$ or $X \geq Y$ are assumed to hold almost surely (a.s.).

Here are some useful facts about martingales.

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

$$X_j = \mathbb{E}[X_i | \mathscr{F}_j], \quad \forall\, i > j. \tag{2.1.2}$$

Also $\mathbb{E}[X_i] = \mathbb{E}\big[\mathbb{E}[X_i | \mathscr{F}_{i-1}]\big] = \mathbb{E}[X_{i-1}]$, so, it follows from (2.1.2) that the expectations of all X_i's of a martingale sequence are equal to $\mathbb{E}[X_0]$. Note that, since X_i is $\mathscr{F}_i$-measurable, (2.1.2) also holds for $i = j$.

Fact 2. It is possible to generate martingale sequences by the following procedure: Given a random variable $X \in \mathbb{L}^1(\Omega, \mathscr{F}, \mathbb{P})$ and an arbitrary filtration $\{\mathscr{F}_i\}_{i=0}^n$, let

$$X_i = \mathbb{E}[X | \mathscr{F}_i], \quad \forall\, i \in \{0, 1, \ldots\, n\}. \tag{2.1.3}$$

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

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

$$\begin{aligned} \mathbb{E}[X_i | \mathscr{F}_{i-1}] &= \mathbb{E}\big[\mathbb{E}[X | \mathscr{F}_i] | \mathscr{F}_{i-1}\big] && (2.1.4) \\ &= \mathbb{E}[X | \mathscr{F}_{i-1}] \quad (\text{since } \mathscr{F}_{i-1} \subseteq \mathscr{F}_i) && (2.1.5) \\ &= X_{i-1}. && (2.1.6) \end{aligned}$$

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

$$X_0 = \mathbb{E}[X|\mathscr{F}_0] = \mathbb{E}[X], \quad X_n = \mathbb{E}[X|\mathscr{F}_n] = X. \tag{2.1.7}$$

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 $U_1, \ldots, U_n$ be independent random variables which are defined on a common probability space $(\Omega, \mathscr{F}, \mathbb{P})$, and assume that $\mathbb{E}[U_k] = 0$ and $\mathbb{E}[|U_k|] < \infty$ for every k. Let us define

$$X_k = \sum_{j=1}^{k} U_j, \quad \forall\, k \in \{1, \ldots, n\} \tag{2.1.8}$$

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

$$\mathscr{F}_k = \sigma(X_1, \ldots, X_k) \tag{2.1.9}$$

$$= \sigma(U_1, \ldots, U_k), \quad \forall\, k \in \{1, \ldots, n\}. \tag{2.1.10}$$

Note that $\mathscr{F}_k = \sigma(X_1, \ldots, X_k)$ denotes the minimal σ-algebra that includes all the sets of the form

$$\mathcal{D}(\alpha_1, \ldots, \alpha_k) = \{\omega \in \Omega \colon X_1(\omega) \leq \alpha_1, \ldots, X_k(\omega) \leq \alpha_k\} \tag{2.1.11}$$

where $\alpha_j \in \mathbb{R} \cup \{-\infty, +\infty\}$ for $j \in \{1, \ldots, k\}$. It is easy to verify that $\{X_k, \mathscr{F}_k\}_{k=0}^n$ 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 the equality in (2.1.1) is relaxed, we obtain sub- and super-martingales. More precisely, to define sub- and super-martingales, we keep the first two conditions in Definition 2.1, and (2.1.1) is replaced by one of the following:

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

From the tower property for conditional expectations, it follows that

$$\mathbb{E}[X_i|\mathscr{F}_j] \geq X_j, \quad \forall\, i > j \tag{2.1.12}$$

for sub-martingales, and

$$\mathbb{E}[X_i|\mathscr{F}_j] \leq X_j, \quad \forall\, i > j \tag{2.1.13}$$

for super-martingales. By taking expectations on both sides of (2.1.12) or (2.1.13), it follows that the expectations of the terms of a sub/super-martingale form, respectively, a monotonically increasing/decreasing sequence. Clearly, every random process that is both a sub- and super-martingale is a martingale, and vice versa. Furthermore, $\{X_i, \mathscr{F}_i\}$ is a sub-martingale if and only if $\{-X_i, \mathscr{F}_i\}$ is a super-martingale. The following properties are direct consequences of Jensen's inequality for conditional expectations.

Theorem 2.1.1. The following holds for mappings of martingales or sub/ super martingales:

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

Example 2.2. The following are special cases of Theorem 2.1.1:

- If $\{X_i, \mathscr{F}_i\}$ is a martingale, then $\{|X_i|, \mathscr{F}_i\}$ is a sub-martingale.
- If $\{X_i, \mathscr{F}_i\}$ is a martingale and $X_i \in \mathbb{L}^2(\Omega, \mathscr{F}_i, \mathbb{P})$, then $\{X_i^2, \mathscr{F}_i\}$ is a sub-martingale.
- If $\{X_i, \mathscr{F}_i\}$ is a non-negative sub-martingale which satisfies that $X_i \in \mathbb{L}^2(\Omega, \mathscr{F}_i, \mathbb{P})$ (i.e., for every i, the random variable X_i is defined on the same sample space Ω, it is $\mathscr{F}_i$-measurable, and $\mathbb{E}[X_i^2] < \infty$), then $\{X_i^2, \mathscr{F}_i\}$ is a sub-martingale.

2.2 The main ingredients of the martingale method

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 is a real-valued 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 construct a suitable filtration $\{\mathscr{F}_i\}_{i=0}^n$ on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ that carries U, where $\mathscr{F}_0 = \{\emptyset, \Omega\}$ is the trivial σ-algebra, and $\mathscr{F}_n = \mathscr{F}$. Then, we decompose the difference $U - \mathbb{E}U$ as

$$U - \mathbb{E}U = \mathbb{E}[U|\mathscr{F}_n] - \mathbb{E}[U|\mathscr{F}_0] \tag{2.2.1}$$

$$= \sum_{i=1}^{n} \left(\mathbb{E}[U|\mathscr{F}_i] - \mathbb{E}[U|\mathscr{F}_{i-1}]\right). \tag{2.2.2}$$

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

2. **The Chernoff bound** — 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

$$\Lambda(t) = t\mathbb{E}[U] + \ln \mathbb{E}\left[\prod_{i=1}^{n} \exp(t\xi_i)\right], \tag{2.2.3}$$

   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 [85], 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 the occasion to see later on, but the above two ingredients are a good starting point. In the remainder of this chapter, we will elaborate on these ideas and examine their basic consequences.

2.3 Bounding the variance: Efron-Stein-Steele Inequalities

The basic idea behind the martingale method is to start with the *Doob martingale decomposition*

$$Z - \mathbb{E}[Z] = \sum_{k=1}^{n} \xi_k, \tag{2.3.1}$$

where

$$\xi_k \triangleq \mathbb{E}[Z|X^k] - \mathbb{E}[Z|X^{k-1}] \tag{2.3.2}$$

with

$$X^k \triangleq (X_1, \ldots, X_k) \tag{2.3.3}$$

and to exploit any information about the sensitivity of f to local changes in its arguments in order to control the sizes of the increments ξ_k. The following inequality was first obtained in a restricted setting by Efron and Stein [86], and it was then generalized by Steele [87].

Theorem 2.3.1 (Efron–Stein–Steele inequality)**.** If $Z = f(X^n)$ where f is a real-valued function and $X_1, \ldots, X_n$ are independent random variables, then

$$\mathrm{Var}[Z] \leq \sum_{k=1}^{n} \mathbb{E}\left[\mathrm{Var}[Z|\overline{X}^k]\right]. \tag{2.3.4}$$

where for $k \in \{1, \ldots, n\}$

$$\overline{X}^k \triangleq (X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n). \tag{2.3.5}$$

is the vector whose entries are $X_1, \ldots, X_n$ with X_k being excluded.

Proof. From (2.3.1), we have

$$\mathrm{Var}(Z) = \mathbb{E}\left[\left(\sum_{k=1}^{n} \xi_k\right)^2\right] = \sum_{k=1}^{n} \mathbb{E}[\xi_k^2] + 2\sum_{l>k} \mathbb{E}[\xi_l\, \xi_k]. \tag{2.3.6}$$

We exploit the fact that $\{\xi_k\}_{k=1}^n$ in (2.3.2) is a *martingale difference sequence* with respect to X^n, i.e., for all $k \in \{1, \ldots, n\}$

$$\mathbb{E}[\xi_k | X^{k-1}] = 0. \tag{2.3.7}$$

Hence, for all l and k such that $1 \leq k < l \leq n$, (2.3.2) and (2.3.7) yield

$$\begin{aligned} \mathbb{E}[\xi_l \xi_k] &= \mathbb{E}\left[\mathbb{E}[\xi_l \xi_k \,|\, X^{l-1}]\right] && (2.3.8)\\ &= \mathbb{E}\left[\xi_k \, \mathbb{E}[\xi_l \,|\, X^{l-1}]\right] && (2.3.9)\\ &= \mathbb{E}[\xi_k \cdot 0] = 0 && (2.3.10) \end{aligned}$$

which, from (2.3.6) and (2.3.8)–(2.3.10), imply that

$$\mathrm{Var}[Z] = \sum_{k=1}^{n} \mathbb{E}[\xi_k^2]. \tag{2.3.11}$$

The independence of $X_1, \ldots, X_n$ implies that (2.3.2) can be rewritten in the form

$$\xi_k = \mathbb{E}\left[Z - \mathbb{E}[Z|\overline{X}^k] \,|\, X^k\right] \tag{2.3.12}$$

and, from Jensen's inequality,

$$\xi_k^2 \leq \mathbb{E}\left[\left(Z - \mathbb{E}[Z|\overline{X}^k]\right)^2 \,|\, X^k\right]. \tag{2.3.13}$$

Using again the independence of $X_1, \ldots, X_n$ yields

$$\begin{aligned} \mathbb{E}[\xi_k^2] &\leq \mathbb{E}\left[\left(Z - \mathbb{E}[Z|\overline{X}^k]\right)^2\right] \\ &= \mathbb{E}\left[\mathrm{Var}[Z|\overline{X}^k]\right], && (2.3.14) \end{aligned}$$

and substituting (2.3.14) into (2.3.11) yields (2.3.4). □

The Efron–Stein–Steele inequality (2.3.4) is our first example of *tensorization*: it upper-bounds the variance of $Z = f(X_1, \ldots, X_n)$ by the sum of the expected values of the conditional variances of Z given all but one of the variables. In other words, we say that $\mathrm{Var}[f(X_1, \ldots, X_n)]$ tensorizes. This fact has immediate useful consequences. For example, we can use any convenient technique for upper-bounding variances to control each term on the right-hand side of (2.3.11), and thus obtain many useful variants of the Efron–Stein–Steele inequality:

1. For every random variable U with a finite second moment,

$$\mathrm{Var}[U] = \tfrac{1}{2}\,\mathbb{E}[(U-U')^2] \tag{2.3.15}$$

where U' is an i.i.d. copy of U (i.e., U and U' are i.i.d.). Thus, if we let

$$Z'_k = f(X_1, \ldots, X_{k-1}, X'_k, X_{k+1}, \ldots, X_n), \tag{2.3.16}$$

where X'_k is an i.i.d. copy of X_k, then Z and Z'_k are i.i.d. given $\overline{X}^k$. This implies that

$$\mathrm{Var}[Z|\overline{X}^k] = \tfrac{1}{2}\,\mathbb{E}\Big[(Z-Z'_k)^2 \,\Big|\, \overline{X}^k\Big] \tag{2.3.17}$$

for $k \in \{1, \ldots, n\}$, yielding the following variant of (2.3.4):

$$\mathrm{Var}[Z] \le \tfrac{1}{2}\sum_{i=1}^{n} \mathbb{E}[(Z-Z'_k)^2]. \tag{2.3.18}$$

Inequality (2.3.18) is sharp: if $Z = \sum_{k=1}^n X_k$, then

$$\mathbb{E}[(Z-Z'_k)^2] = 2\mathrm{Var}[X_k], \tag{2.3.19}$$

and (2.3.18) holds with equality. This therefore shows that for independent random variables $X_1, \ldots, X_n$, their sum is the least concentrated among all functions of X^n.

2. For every random variable U with a finite second moment and for all $c \in \mathbb{R}$,

$$\mathrm{Var}[U] \le \mathbb{E}[(U-c)^2]. \tag{2.3.20}$$

Thus, by conditioning on $\overline{X}^k$, we let $Z_k = f_k(\overline{X}^k)$ for arbitrary real-valued functions f_k (with $k \in \{1, \ldots, n\}$) of $n-1$ variables to obtain

$$\mathrm{Var}[Z|\overline{X}^k] \le \mathbb{E}[(Z-Z_k)^2|\overline{X}^k]. \tag{2.3.21}$$

From (2.3.14), this yields another variant of (2.3.4):

$$\mathrm{Var}[Z] \le \sum_{i=1}^{n} \mathbb{E}[(Z-Z_k)^2]. \tag{2.3.22}$$

3. Suppose we know that, by varying just one of the arguments of f while holding all others fixed, we cannot change the value of f by more than some bounded amount. More precisely, suppose that there exist finite constants $c_1, \ldots, c_n \geq 0$, such that

$$\sup_x f(x_1, \ldots, x_{k-1}, x, x_{k+1}, \ldots, x_n) - \inf_x f(x_1, \ldots, x_{k-1}, x, x_{k+1}, \ldots, x_n) \leq c_k \tag{2.3.23}$$

for all $k \in \{1, \ldots, n\}$ and all $x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n$. Then, for all k,

$$\mathrm{Var}[Z|\overline{X}^k] \leq \tfrac{1}{4}\, c_k^2 \tag{2.3.24}$$

and therefore from (2.3.4) and (2.3.14)

$$\mathrm{Var}[Z] \leq \tfrac{1}{4} \sum_{k=1}^{n} c_k^2. \tag{2.3.25}$$

Example 2.3 (Kernel Density Estimation)**.** As an example of the use of the Efron–Stein–Steele inequalities, the *kernel density estimation* (KDE) is considered. It is a nonparametric procedure for estimating an unknown probability density function (pdf) ϕ of a real-valued random variable X, based on observing n i.i.d. samples $X_1, \ldots, X_n$ drawn from ϕ [88, Chap. 9]. A *kernel* is a function $K \colon \mathbb{R} \to \mathbb{R}^+$ satisfying the following conditions:

1. It is integrable and normalized:

$$\int_{-\infty}^{\infty} K(u)\mathrm{d}u = 1. \tag{2.3.26}$$

2. It is an even function: $K(u) = K(-u)$ for all $u \in \mathbb{R}$.

3. $$\lim_{h \downarrow 0} \frac{1}{h}\, K\left(\frac{x-u}{h}\right) = \delta(x-u), \tag{2.3.27}$$

where δ denotes the Dirac delta function.

The KDE is given by

$$\phi_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right), \tag{2.3.28}$$

where $h_n > 0$ is called the *smoothing parameter*. From the properties of the kernel K, for each $x \in \mathbb{R}$ we have

$$\mathbb{E}[\phi_n(x)] = \frac{1}{h_n} \int_{-\infty}^{\infty} K\left(\frac{x-u}{h_n}\right) \phi(u)\, \mathrm{d}u \tag{2.3.29}$$

$$\xrightarrow{h_n \downarrow 0} \phi(x) \tag{2.3.30}$$

where (2.3.30) follows from (2.3.27). Thus, we expect the KDE ϕ_n to concentrate around the true pdf ϕ whenever $h_n \downarrow 0$. To quantify this, let us examine the L_1 error

$$Z_n = f(X_1, \ldots, X_n) = \int_{-\infty}^{\infty} |\phi_n(x) - \phi(x)|\, \mathrm{d}x. \tag{2.3.31}$$

A simple calculation shows that

$$\begin{aligned}
&\left|f(x_1, \ldots, x_{k-1}, x_k, x_{k+1}, \ldots, x_n) - f(x_1, \ldots, x_{k-1}, x_k', x_{k+1}, \ldots, x_n)\right| \\
&\leq \frac{1}{nh_n} \int_{-\infty}^{\infty} \left| K\left(\frac{x - x_k}{h_n}\right) - K\left(\frac{x - x_k'}{h_n}\right)\right| \mathrm{d}x && (2.3.32) \\
&\leq \frac{1}{nh_n} \left[\int_{-\infty}^{\infty} K\left(\frac{x - x_k}{h_n}\right) \mathrm{d}x + \int_{-\infty}^{\infty} K\left(\frac{x - x_k'}{h_n}\right) \mathrm{d}x \right] && (2.3.33) \\
&= \frac{2}{n} && (2.3.34)
\end{aligned}$$

for all $x_1, \ldots, x_{k-1}, x_k, x_k', x_{k+1}, \ldots, x_n \in \mathbb{R}$ where (2.3.32) follows from the triangle inequality and (2.3.28); (2.3.33) holds due to another use of the triangle inequality, and (2.3.34) is due to (2.3.26). This implies that the function f satisfies (2.3.23) with

$$c_1 = \ldots = c_n = \frac{2}{n}, \tag{2.3.35}$$

and, consequently, (2.3.25) and (2.3.35) yield

$$\mathrm{Var}[Z_n] \leq \frac{1}{n}. \tag{2.3.36}$$

2.4 Basic concentration inequalities

2.4.1 The Chernoff bounding technique and the Hoeffding lemma

An ingredient of the martingale method is the well-known Chernoff bounding technique[1]: Using Markov's inequality, for every $t > 0$ and $r \in \mathbb{R}$,

$$\mathbb{P}(U \geq r) = \mathbb{P}\big(\exp(tU) \geq \exp(tr)\big) \tag{2.4.1}$$

$$\leq \exp(-tr)\, \mathbb{E}[\exp(tU)]. \tag{2.4.2}$$

Equivalently, if we define the *logarithmic moment generating function*

$$\Lambda(t) \triangleq \ln \mathbb{E}[\exp(tU)], \quad \forall\, t \in \mathbb{R}, \tag{2.4.3}$$

we can write

$$\mathbb{P}(U \geq r) \leq \exp\big(\Lambda(t) - tr\big), \qquad \forall\, t > 0. \tag{2.4.4}$$

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 2.4.1 (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 every $t \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\big(t(U - \mathbb{E}U)\big)\right] \leq \exp\left(\tfrac{1}{8}\, t^2 (b-a)^2\right). \tag{2.4.5}$$

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

$$H_p(\lambda) \triangleq \ln\left(p e^{\lambda(1-p)} + (1-p) e^{-\lambda p}\right). \tag{2.4.6}$$

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

[1] The name of H. Chernoff is associated with this technique because of his 1952 paper [89]; however, its roots go back to S.N. Bernstein's 1927 textbook on the theory of probability [90].

the exponential function, we can write

$$\begin{aligned}
&\exp(t\xi) \\
&= \exp\left(\frac{U-a}{b-a}\cdot t(b-\mathbb{E}U) + \frac{b-U}{b-a}\cdot t(a-\mathbb{E}U)\right) \qquad (2.4.7)\\
&\leq \left(\frac{U-a}{b-a}\right)\exp\big(t(b-\mathbb{E}U)\big) + \left(\frac{b-U}{b-a}\right)\exp\big(t(a-\mathbb{E}U)\big). \qquad (2.4.8)
\end{aligned}$$

Taking expectations on both sides of (2.4.7) and (2.4.8) yields

$$\begin{aligned}
&\mathbb{E}[\exp(t\xi)] \\
&\leq \left(\frac{\mathbb{E}U-a}{b-a}\right)\exp\big(t(b-\mathbb{E}U)\big) + \left(\frac{b-\mathbb{E}U}{b-a}\right)\exp\big(t(a-\mathbb{E}U)\big) \qquad (2.4.9)\\
&= \exp\big(H_p(\lambda)\big) \qquad (2.4.10)
\end{aligned}$$

where (2.4.10) holds in view of (2.4.6) with

$$p = \frac{\mathbb{E}U-a}{b-a} \qquad \text{and} \qquad \lambda = t(b-a). \qquad (2.4.11)$$

In the following, we show that for every $\lambda \in \mathbb{R}$

$$H_p(\lambda) \leq \tfrac{1}{8}\lambda^2, \quad \forall\, p \in [0,1]. \qquad (2.4.12)$$

From (2.4.6), we have

$$H_p(\lambda) = -\lambda p + \ln\big(pe^{\lambda} + (1-p)\big), \qquad (2.4.13)$$

$$H_p'(\lambda) = -p + \frac{pe^{\lambda}}{pe^{\lambda}+1-p}, \qquad (2.4.14)$$

$$H_p''(\lambda) = \frac{p(1-p)e^{\lambda}}{\big(pe^{\lambda}+(1-p)\big)^2}. \qquad (2.4.15)$$

From (2.4.13)–(2.4.15), we have $H_p(0) = H_p'(0) = 0$, and

$$H_p''(\lambda) \leq \tfrac{1}{4}, \quad \forall\, \lambda \in \mathbb{R},\ p \in [0,1] \qquad (2.4.16)$$

where the last inequality holds since $ab \leq \frac{1}{4}(a+b)^2$ for all $a, b \in \mathbb{R}$. Using a Taylor's series expansion, there exists an intermediate value $\theta \in [0,\lambda]$ (or $\theta \in [\lambda, 0]$ if $t < 0$) such that

$$H_p(\lambda) = H_p(0) + H_p'(0)\lambda + \tfrac{1}{2}\, H_p''(\theta)\, \lambda^2 \qquad (2.4.17)$$

so, consequently, (2.4.12) holds. Substituting (2.4.12) into (2.4.10) and using the definitions of p and λ in (2.4.11), we get (2.4.5). □

2.4.2 The Azuma–Hoeffding inequality

The Azuma–Hoeffding inequality, stated in Theorem 2.4.2 below, is a useful concentration inequality for bounded-difference martingales. It was first proved by Hoeffding [9] for sums of independent and bounded random variables, followed by a short note on the suitability of this result in the generalized setting of a sum of bounded random variables which forms a martingale. This inequality was independently obtained four years later by Azuma [8] in this generalized 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 Section 2.2. We will have several occasions to revisit this proof in order to obtain various refinements of the Azuma–Hoeffding inequality.

Theorem 2.4.2 (The Azuma–Hoeffding inequality)**.** Let $\{X_k, \mathscr{F}_k\}_{k=0}^n$ be a real-valued martingale sequence. Suppose that there exist non-negative 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 $r > 0$,

$$\mathbb{P}(|X_n - X_0| \geq r) \leq 2 \exp\left(-\frac{r^2}{2\sum_{k=1}^n d_k^2}\right). \tag{2.4.18}$$

Proof. For an arbitrary $r > 0$,

$$\mathbb{P}(|X_n - X_0| \geq r) = \mathbb{P}(X_n - X_0 \geq r) + \mathbb{P}(X_n - X_0 \leq -r). \tag{2.4.19}$$

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

We now apply the Chernoff bounding technique:

$$\begin{aligned}
&\mathbb{P}(X_n - X_0 \geq r) \\
&= \mathbb{P}\left(\sum_{k=1}^n \xi_k \geq r\right) \\
&\leq \exp(-tr)\, \mathbb{E}\left[\exp\left(t \sum_{k=1}^n \xi_k\right)\right], \quad \forall\, t \geq 0.
\end{aligned} \tag{2.4.20}$$

By the law of iterated expectations, the expectation on the right side of (2.4.20) is equal to

$$\begin{aligned}
&\mathbb{E}\Big[\exp\Big(t\sum_{k=1}^{n}\xi_k\Big)\Big] \\
&= \mathbb{E}\Bigg[\mathbb{E}\Big[\exp\Big(t\sum_{k=1}^{n}\xi_k\Big)\,\Big|\,\mathscr{F}_{n-1}\Big]\Bigg] \\
&= \mathbb{E}\Bigg[\exp\Big(t\sum_{k=1}^{n-1}\xi_k\Big)\,\mathbb{E}\big[\exp(t\xi_n)\,|\,\mathscr{F}_{n-1}\big]\Bigg] \qquad (2.4.21)
\end{aligned}$$

where the last equality holds since $Y_n \triangleq \exp(t\sum_{k=1}^{n-1}\xi_k)$ is $\mathscr{F}_{n-1}$-measurable. We now apply the Hoeffding lemma with the conditioning on $\mathscr{F}_{n-1}$. Indeed, we know that $\mathbb{E}[\xi_n|\mathscr{F}_{n-1}] = 0$ and that $\xi_n \in [-d_n, d_n]$ a.s., so Lemma 2.4.1 gives that

$$\mathbb{E}\big[\exp(t\xi_n)\,|\,\mathscr{F}_{n-1}\big] \leq \exp\left(\frac{t^2\,d_n^2}{2}\right). \qquad (2.4.22)$$

Continuing recursively in a similar manner, we can bound the quantity in (2.4.21) by

$$\mathbb{E}\Big[\exp\Big(t\sum_{k=1}^{n}\xi_k\Big)\Big] \leq \prod_{k=1}^{n}\exp\left(\frac{t^2\,d_k^2}{2}\right) = \exp\left(\frac{t^2}{2}\sum_{k=1}^{n}d_k^2\right). \qquad (2.4.23)$$

Substituting this bound into (2.4.20), we obtain

$$\mathbb{P}(X_n - X_0 \geq r) \leq \exp\left(-tr + \frac{t^2}{2}\sum_{k=1}^{n}d_k^2\right), \quad \forall\, t \geq 0. \qquad (2.4.24)$$

Finally, choosing $t = r\,(\sum_{k=1}^{n}d_k^2)^{-1}$ to minimize the right side of (2.4.24) yields

$$\mathbb{P}(X_n - X_0 \geq r) \leq \exp\left(-\frac{r^2}{2\sum_{k=1}^{n}d_k^2}\right). \qquad (2.4.25)$$

Since $\{X_k, \mathscr{F}_k\}$ is a martingale with bounded differences, so is $\{-X_k, \mathscr{F}_k\}$ (with the same bounds on its differences). This implies that the same bound is also valid for the probability $\mathbb{P}(X_n - X_0 \leq -r)$. Using these bounds in (2.4.19) completes the proof of Theorem 2.4.2. □

Remark 2.1. In [6, Theorem 3.13], the Azuma–Hoeffding inequality is stated as follows: Let $\{Y_k, \mathscr{F}_k\}_{k=0}^n$ be a martingale-difference sequence with $Y_0 = 0$ (i.e., Y_k is $\mathscr{F}_k$-measurable, $\mathbb{E}[|Y_k|] < \infty$ and $\mathbb{E}[Y_k|\mathscr{F}_{k-1}] = 0$ 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 $r \geq 0$,

$$\mathbb{P}\left(\left|\sum_{k=1}^n Y_k\right| \geq r\right) \leq 2\exp\left(-\frac{2r^2}{\sum_{k=1}^n (b_k - a_k)^2}\right). \tag{2.4.26}$$

Consider a real-valued martingale sequence $\{X_k, \mathscr{F}_k\}_{k=0}^n$ for which there exist $a_k, b_k \in \mathbb{R}$ such that $a_k \leq X_k - X_{k-1} \leq b_k$ a.s. for all k. Let $Y_k \triangleq X_k - X_{k-1}$ for every $k \in \{1, \ldots, n\}$. It is easy to verify that $\{Y_k, \mathscr{F}_k\}_{k=0}^n$ is a martingale-difference sequence. Since $\sum_{k=1}^n Y_k = X_n - X_0$, it follows from (2.4.26) 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 2.4. 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 $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \mathscr{F}_2 \ldots$ where

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

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

$$\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) \leq 2\exp\left(-\frac{\alpha^2}{2d^2}\right) \tag{2.4.27}$$

for every $\alpha \geq 0$ and $n \in \mathbb{N}$. Since the random variables $\{Y_i\}_{i=0}^\infty$ 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 $\alpha \geq 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) = 2\,Q\Big(\frac{\alpha}{d}\Big) \tag{2.4.28}$$

where

$$Q(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t, \quad \forall\, x \in \mathbb{R} \tag{2.4.29}$$

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., [91, Section 3.3]):

$$\frac{1}{\sqrt{2\pi}}\,\frac{x}{1+x^2}\cdot\exp\left(-\frac{x^2}{2}\right) < Q(x) < \frac{1}{\sqrt{2\pi}\,x}\cdot\exp\left(-\frac{x^2}{2}\right),\quad \forall\, x>0. \tag{2.4.30}$$

From (2.4.28) and (2.4.30), it follows that the exponent on the right side of (2.4.27) is exact.

Example 2.5. Fix some $\gamma \in (0,1]$. Let us generalize Example 2.4 above by considering the case where the i.i.d. binary random variables $\{Y_i\}_{i=0}^{\infty}$ 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 $\{X_k, \mathscr{F}_k\}_{k=0}^{\infty}$ as in Example 2.4. 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 $\alpha \geq 0$

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) = 2\,Q\bigg(\frac{\alpha}{\sqrt{\gamma}\,d}\bigg). \tag{2.4.31}$$

From the bounds on the Q-function given in (2.4.30), it follows that the right side of (2.4.31) scales exponentially like $e^{-\frac{\alpha^2}{2\gamma d^2}}$. Hence, the exponent in this example is improved by a factor of $\frac{1}{\gamma}$ in comparison to the Azuma–Hoeffding inequality (which gives the same bound as in Example 2.4 since $|X_k - X_{k-1}| \leq d$ for every $k \in \mathbb{N}$). This indicates that a refinement of the Azuma–Hoeffding 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 2.5.2.

2.4.3 McDiarmid's inequality

A prominent application of the martingale approach is the derivation of a powerful inequality due to McDiarmid (see [92, Theorem 3.1] or

[93]), also known as the *bounded-difference inequality*. Let $\mathcal{X}$ be a set, and let $f\colon \mathcal{X}^n \to \mathbb{R}$ be a function that satisfies the *bounded difference assumption*

$$\sup_{x_1,\ldots,x_n,x_i' \in \mathcal{X}} \Big| f(x_1,\ldots,x_{i-1},x_i,x_{i+1}\ldots,x_n) - f(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n) \Big| \le d_i \tag{2.4.32}$$

for every $1 \le i \le n$, where $d_1,\ldots,d_n$ are arbitrary nonnegative real constants. This is equivalent to saying that, for every given i, the variation of the function f with respect to its i-th 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 easy.)

Theorem 2.4.3 (McDiarmid's inequality)**.** Let $\{X_k\}_{k=1}^n$ be independent (though not necessarily i.i.d.) random variables taking values in a set $\mathcal{X}$. Consider a random variable $U = f(X^n)$ where $f\colon \mathcal{X}^n \to \mathbb{R}$ is a measurable function that satisfies the bounded difference assumption (2.4.32), and $X^n \triangleq (X_1,\ldots,X_n)$. Then, for every $r \ge 0$,

$$\mathbb{P}\left(\big|U - \mathbb{E}U\big| \ge r\right) \le 2\exp\left(-\frac{2r^2}{\sum_{k=1}^n d_k^2}\right). \tag{2.4.33}$$

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

$$\xi_k \triangleq \mathbb{E}\big[f(X^n)\,\big|\,\mathscr{F}_k\big] - \mathbb{E}\big[f(X^n)\,\big|\,\mathscr{F}_{k-1}\big]. \tag{2.4.34}$$

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

$$\begin{aligned} &\mathbb{E}\big[f(X^n)\,\big|\,\mathscr{F}_0\big] = \mathbb{E}\big[f(X^n)\big], \\ &\mathbb{E}\big[f(X^n)\,\big|\,\mathscr{F}_n\big] = f(X^n). \end{aligned} \tag{2.4.35}$$

From the last three equalities, it follows that

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

In the following, we need a lemma:

Lemma 2.4.4. For every $k \in \{1, \ldots, n\}$, the following properties hold a.s.:

1. $\mathbb{E}[\xi_k \,|\, \mathscr{F}_{k-1}] = 0$ and ξ_k is $\mathscr{F}_k$-measurable, so $\{\xi_k, \mathscr{F}_k\}$ is a martingale-difference.

2. $|\xi_k| \leq d_k$.

3. $\xi_k \in [A_k, A_k + d_k]$ where A_k is a non-positive and $\mathscr{F}_{k-1}$-measurable random variable.

Proof. The random variable ξ_k, defined in (2.4.34), is $\mathscr{F}_k$-measurable since $\mathscr{F}_{k-1} \subseteq \mathscr{F}_k$, and ξ_k is a difference of two functions where one is $\mathscr{F}_k$-measurable and the other is $\mathscr{F}_{k-1}$-measurable. Furthermore, since $\{\mathscr{F}_i\}$ is a filtration, it follows from (2.4.34) and the tower principle for conditional expectations that $\mathbb{E}[\xi_k \,|\, \mathscr{F}_{k-1}] = 0$. This proves the first item. The second item follows from the first and third items since the latter two items imply that

$$\begin{aligned} A_k &= \mathbb{E}[A_k|\mathscr{F}_{k-1}] \\ &\leq \mathbb{E}[\xi_k|\mathscr{F}_{k-1}] = 0 \\ &\leq \mathbb{E}[A_k + d_k|\mathscr{F}_{k-1}] \\ &= A_k + d_k \end{aligned} \tag{2.4.36}$$

where the first and last equalities hold since A_k is $\mathscr{F}_{k-1}$-measurable. Hence, $0 \in [A_k, A_k + d_k]$ which implies that $[A_k, A_k + d_k] \subseteq [-d_k, d_k]$; consequently, it follows from the third item that $|\xi_k| \leq d_k$.

To prove the third item, note that $\xi_k = f_k(X_1, \ldots, X_k)$ holds a.s. for the $\mathscr{F}_k$-measurable function $f_k \colon \mathcal{X}^k \to \mathbb{R}$ which is given by

$$\begin{aligned} &f_k(x_1, \ldots, x_k) \\ &= \mathbb{E}\big[f(x_1, \ldots, x_k, X_{k+1}, \ldots, X_n)\big] - \mathbb{E}\big[f(x_1, \ldots, x_{k-1}, X_k, \ldots, X_n)\big]. \end{aligned} \tag{2.4.37}$$

Equality (2.4.37) holds due to the definition of $\{\xi_k\}$ in (2.4.34) with $\mathscr{F}_k = \sigma(X_1, \ldots, X_k)$ for $k \in \{1, \ldots, n\}$, and the independence of the random variables $\{X_k\}_{k=1}^n$. Let us define, for every $k \in \{1, \ldots, n\}$,

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

which are $\mathscr{F}_{k-1}$-measurable[2], 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 obtain from (2.4.37) that

$$\begin{aligned}
&\sup_{x \in \mathcal{X}} f_k(x_1, \ldots, x_{k-1}, x) - \inf_{x' \in \mathcal{X}} f_k(x_1, \ldots, x_{k-1}, x') \\
&= \sup_{x, x' \in \mathcal{X}} \big\{ f_k(x_1, \ldots, x_{k-1}, x) - f_k(x_1, \ldots, x_{k-1}, x') \big\} \\
&= \sup_{x, x' \in \mathcal{X}} \Big\{ \mathbb{E}\big[f(x_1, \ldots, x_{k-1}, x, X_{k+1}, \ldots, X_n) \big] \\
&\qquad\qquad - \mathbb{E}\big[f(x_1, \ldots, x_{k-1}, x', X_{k+1}, \ldots, X_n) \big] \Big\} \qquad (2.4.38) \\
&= \sup_{x, x' \in \mathcal{X}} \Big\{ \mathbb{E}\big[f(x_1, \ldots, x_{k-1}, x, X_{k+1}, \ldots, X_n) \\
&\qquad\qquad - f(x_1, \ldots, x_{k-1}, x', X_{k+1}, \ldots, X_n) \big] \Big\} \\
&\leq d_k \qquad (2.4.39)
\end{aligned}$$

where (2.4.38) follows from (2.4.37), and (2.4.39) follows from the bounded-difference condition in (2.4.32). Hence, $B_k - A_k \leq d_k$ a.s., which implies that $\xi_k \in [A_k, A_k + d_k]$. 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 (i.e., item 2 asserts that ξ_k is contained in the interval $[-d_k, d_k]$ which is twice longer than the sub-interval $[A_k, A_k + d_k]$ in the third item, see (2.4.36)). □

[2]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 [94] contains a thorough treatment of these issues.

We now proceed in the same manner as in the proof of the Azuma–Hoeffding inequality. Specifically, for $k \in \{1, \ldots, n\}$, $\xi_k \in [A_k, A_k + d_k]$ a.s., where A_k is $\mathscr{F}_{k-1}$-measurable, and $\mathbb{E}[\xi_k | \mathscr{F}_{k-1}] = 0$. Thus, we may apply the Hoeffding lemma (see Lemma 2.4.1) with a conditioning on $\mathscr{F}_{k-1}$ to get

$$\mathbb{E}\Big[e^{t\xi_k}\Big|\mathscr{F}_{k-1}\Big] \leq \exp\left(\frac{t^2 d_k^2}{8}\right). \tag{2.4.40}$$

Similarly to the proof of the Azuma–Hoeffding inequality, by repeatedly using the recursion in (2.4.21), the last inequality implies that

$$\mathbb{E}\left[\exp\Big(t\sum_{k=1}^{n}\xi_k\Big)\right] \leq \exp\left(\frac{t^2}{8}\sum_{k=1}^{n} d_k^2\right) \tag{2.4.41}$$

and, from (2.4.20),

$$\begin{aligned}
&\mathbb{P}(f(X^n) - \mathbb{E}[f(X^n)] \geq r) \\
&= \mathbb{P}\left(\sum_{k=1}^{n}\xi_k \geq r\right) \\
&\leq \exp\left(-tr + \frac{t^2}{8}\sum_{k=1}^{n} d_k^2\right), \quad \forall\, t \geq 0.
\end{aligned} \tag{2.4.42}$$

The choice $t = 4r\left(\sum_{k=1}^{n} d_k^2\right)^{-1}$ minimizes the expression in (2.4.42), so

$$\mathbb{P}\Big(f(X^n) - \mathbb{E}[f(X^n)] \geq r\Big) \leq \exp\left(-\frac{2r^2}{\sum_{k=1}^{n} d_k^2}\right). \tag{2.4.43}$$

By replacing f with $-f$, it follows that this bound is also valid for the probability $\mathbb{P}(f(X^n) - \mathbb{E}[f(X^n)] \leq -r)$, so

$$\begin{aligned}
&\Pr\Big(\Big|f(X^n) - \mathbb{E}[f(X^n)]\Big| \geq r\Big) \\
&= \Pr\Big(f(X^n) - \mathbb{E}[f(X^n)] \geq r\Big) + \mathbb{P}\Big(f(X^n) - \mathbb{E}[f(X^n)] \leq -r\Big) \\
&\leq 2\exp\left(-\frac{2r^2}{\sum_{k=1}^{n} d_k^2}\right)
\end{aligned}$$

which gives the bound in (2.4.33). □

Example 2.6. Let $g\colon \{1,\ldots,n\} \to \{1,\ldots,n\}$ be a random where all n^n such possible functions are equally likely. Let $L(g)$ denote the number of elements $y \in \{1,\ldots,n\}$ such that $g(x) \neq y$ for all $x \in \{1,\ldots,n\}$. By the linearity of the expectation, we have

$$\mathbb{E}[L(g)] = n\left(1-\frac{1}{n}\right)^n, \tag{2.4.44}$$

and consequently, it is easy to show that for every $n \in \mathbb{N}$,

$$\frac{n-1}{e} < \mathbb{E}[L(g)] < \frac{n}{e}. \tag{2.4.45}$$

In [10, Theorem 7.5.1], the following concentration result for $L(g)$ around its expected value is introduced. Construct a martingale sequence $\{X_k, \mathscr{F}_k\}_{k=0}^n$ (see Fact 2) by

$$X_k = \mathbb{E}[L(g) \,|\, \mathscr{F}_k], \quad \forall\, k \in \{0,\ldots,n\}$$

with the natural filtration $\mathscr{F}_k = \sigma(g(1),\ldots,g(k))$ which denotes the σ-algebra that is generated by revealing the first k values of the random function g, for $k \in \{1,\ldots,n\}$, and $\mathscr{F}_0$ is the minimal σ-algebra that includes the empty set and the sample space. By this construction, $X_0 = \mathbb{E}[L(g)]$ and $X_n = L(g)$. Since a modification of one value of g cannot change $L(g)$ by more than 1, it follows that $|X_k - X_{k-1}| \leq 1$ for every $k \in \{1,\ldots,n\}$. From the Azuma-Hoeffding inequality and (2.4.45), it follows that

$$\mathbb{P}\left(\left|L(g)-\frac{n}{e}\right| > \alpha\sqrt{n}+1\right) \leq 2\exp\left(-\frac{\alpha^2}{2}\right), \quad \forall\, \alpha > 0. \tag{2.4.46}$$

The concentration result in (2.4.46), as given in [10, Theorem 7.5.1], can be improved as follows: let $f\colon \{1,\ldots,n\}^n \to \{1,\ldots,n\}$ be defined by $L(g) \triangleq f(g(1),\ldots,g(n))$ so, the function f maps the n-length vector $(g(1),\ldots,g(n))$ to the number of elements $y \in \{1,\ldots,n\}$ where $g(x) \neq y$ for every $x \in \{1,\ldots,n\}$. Since by assumption $g(1),\ldots,g(n)$ are independent random variables, the variation of f with respect to each of its arguments (while all the other $n-1$ arguments of f are kept fixed) is no more than 1. Consequently, from McDiarmid's inequality,

$$\mathbb{P}\left(\left|L(g)-\frac{n}{e}\right| > \alpha\sqrt{n}+1\right) \leq 2\exp(-2\alpha^2), \quad \forall\, \alpha > 0, \tag{2.4.47}$$

which implies that the exponent of the concentration inequality (2.4.46) is improved by a factor of 4.

Example 2.7. Let B be a normed space, and $\{\underline{v}_k\}_{k=1}^n$ be n vectors in B. Let $\{\Theta_k\}_{k=1}^n$ be independent Bernoulli$(\frac{1}{2})$ random variables with $\mathbb{P}(\Theta_k = 1) = \mathbb{P}(\Theta_k = -1) = \frac{1}{2}$, and let $X = \left\|\sum_{k=1}^n \Theta_k \, \underline{v}_k\right\|$. By setting

$$f(\theta_1, \ldots, \theta_n) = \left\|\sum_{k=1}^n \theta_k \, \underline{v}_k\right\|, \quad \forall \, \theta_k \in \{-1, +1\}, \; k \in \{1, \ldots, n\}$$

the variation of f with respect to its k-th argument is upper bounded by $2\|\underline{v}_k\|$. Consequently, since $\{\Theta_k\}$ are independent, it follows from McDiarmid's inequality that

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \alpha) \leq 2 \exp\left(-\frac{\alpha^2}{2\sum_{k=1}^n \|\underline{v}_k\|^2}\right), \quad \forall \, \alpha > 0.$$

Remark 2.3. Due to the large applicability of McDiarmid's inequality, there is an interest to improve this inequality for sub-classes of Lipschitz functions of independent random variables. An improvement of this inequality for separately Lipschitz functions of independent random variables has been recently derived in [95] (see also a recent follow-up paper in [96]).

2.4.4 Hoeffding's inequality and its improved versions

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

Theorem 2.4.5 (Hoeffding's inequality)**.** Let $\{U_k\}_{k=1}^n$ be a sequence of independent and bounded random variables where, for $k \in \{1, \ldots, n\}$, $U_k \in [a_k, b_k]$ holds a.s. for some finite constants $a_k, b_k \in \mathbb{R}$ $(a_k < b_k)$. Let $\mu_n \triangleq \sum_{k=1}^n \mathbb{E}[U_k]$. Then,

$$\mathbb{P}\left(\left|\sum_{k=1}^n U_k - \mu_n\right| \geq r\right) \leq 2 \exp\left(-\frac{2r^2}{\sum_{k=1}^n (b_k - a_k)^2}\right), \quad \forall \, r \geq 0. \tag{2.4.48}$$

Proof. Apply Theorem 2.4.3 to the function

$$f(u^n) \triangleq \sum_{k=1}^{n} u_k, \qquad \forall\, u^n \in \prod_{k=1}^{n} [a_k, b_k].$$

An alternative elementary proof combines the Chernoff bound with Lemma 2.4.1 to get

$$\begin{aligned}
&\mathbb{P}\left(\sum_{k=1}^{n} U_k - \mu_n \geq r\right) \\
&= \mathbb{P}\left(\sum_{k=1}^{n} (U_k - \mathbb{E}[U_k]) \geq r\right) \\
&\leq \exp(-tr)\, \mathbb{E}\left[\exp\left(t \sum_{k=1}^{n} (U_k - \mathbb{E}[U_k])\right)\right] \qquad \forall\, t \geq 0 \\
&= \exp(-tr) \prod_{k=1}^{n} \mathbb{E}\left[\exp\Big(t(U_k - \mathbb{E}[U_k])\Big)\right] \\
&\leq \exp(-tr) \prod_{k=1}^{n} \exp\left(\frac{t^2 (b_k - a_k)^2}{8}\right) \\
&= \exp\Big(-tr + \frac{t^2}{8} \sum_{k=1}^{n} (b_k - a_k)^2\Big). \qquad (2.4.49)
\end{aligned}$$

Optimization of the right side of (2.4.49) with respect to t gives

$$t = \frac{4r}{\sum_{k=1}^{n} (b_k - a_k)^2}$$

and its substitution into (2.4.49) yields that, for every $r \geq 0$,

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

The same bound holds for $\mathbb{P}\left(\sum_{k=1}^{n} U_k - \mu_n \leq -r\right)$, which leads to the inequality in (2.4.48). □

Recall that a key step in the proof of McDiarmid's inequality is to invoke Hoeffding's lemma (Lemma 2.4.1). However, a careful look at the proof of Lemma 2.4.1 reveals a potential source of slack in the bound

$$\ln \mathbb{E}\Big[\exp\Big(t(U - \mathbb{E}[U])\Big)\Big] \leq \frac{t^2 (b-a)^2}{8}$$

— 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. An improved version of Hoeffding's inequality was derived by Kearns and Saul [97], and it has been recently further improved by Berend and Kontorovich [98]. The following improvement of Hoeffding's inequality (Lemma 2.4.1) is obtained in [98]:

Lemma 2.4.6 (Berend and Kontorovich)**.** Let U be a real-valued random variable, such that $U \in [a, b]$ a.s. for finite $a < b$. Then, for every $t \geq 0$,

$$\mathbb{E}\left[\exp\left(t(U - \mathbb{E}U)\right)\right] \leq \exp\left(c_{\mathrm{BK}}(p)\, t^2 (b-a)^2\right) \tag{2.4.50}$$

where

$$c_{\mathrm{BK}}(p) = \begin{cases} 0, & \text{if } p = 0 \\ \dfrac{1-2p}{4 \ln\left(\frac{1-p}{p}\right)}, & \text{if } 0 < p < \dfrac{1}{2} \\ \dfrac{p(1-p)}{2}, & \text{if } \dfrac{1}{2} \leq p \leq 1 \end{cases} \tag{2.4.51}$$

with

$$p = \frac{\mathbb{E}[U] - a}{b - a}. \tag{2.4.52}$$

Proof. Recall the definition of $H_p(\lambda)$ in (2.4.6). We deviate from the proof of Lemma 2.4.1 at the point where the bound $H_p(\lambda) \leq \frac{\lambda^2}{8}$ in (2.4.12) is replaced by the improved bound

$$H_p(\lambda) \leq c_{\mathrm{BK}}(p)\, \lambda^2, \qquad \forall\, \lambda \geq 0,\ p \in [0, 1]. \tag{2.4.53}$$

where $c_{\mathrm{BK}}(p)$ is introduced in (2.4.51); for a proof of (2.4.53), the reader is referred to the proofs of [98, Theorem 3.2] and [98, Lemma 3.3]. □

Remark 2.4. The bound on the right side of (2.4.50) depends on the location of $\mathbb{E}[U]$ in the interval $[a, b]$, and it therefore refines Hoeffding's inequality in Lemma 2.4.1. The worst case where $p = \frac{1}{2}$ (i.e., if $\mathbb{E}[U] = \frac{a+b}{2}$ is in the middle of the interval $[a, b]$) coincides however

with Hoeffding's inequality (since, from (2.4.51), $c_{\mathrm{BK}}(p) = \frac{1}{8}$ if $p = \frac{1}{2}$). The bound on $H_p(\lambda)$ in (2.4.53) can be weakened to

$$H_p(\lambda) \leq c_{\mathrm{KS}}(p)\,\lambda^2, \qquad \forall\, \lambda \in \mathbb{R},\ p \in [0,1] \tag{2.4.54}$$

where the abbreviation 'KS' on the right side of (2.4.54) stands for the Kearns-Saul inequality in [97], and it is given by

$$c_{\mathrm{KS}}(p) = \begin{cases} 0, & \text{if } p = 0, 1 \\ \frac{1}{8}, & \text{if } p = \frac{1}{2} \\ \frac{1-2p}{4\ln\left(\frac{1-p}{p}\right)}, & \text{if } p \in (0,1) \setminus \{\frac{1}{2}\}. \end{cases} \tag{2.4.55}$$

From (2.4.51) and (2.4.55), we have

$$\begin{aligned} c_{\mathrm{BK}}(p) = c_{\mathrm{KS}}(p), &\qquad \forall\, p \in \left[0, \tfrac{1}{2}\right] \\ 0 \leq c_{\mathrm{BK}}(p) \leq c_{\mathrm{KS}}(p) \leq \tfrac{1}{8}, &\qquad \forall\, p \in [0,1] \end{aligned}$$

where the equality $c_{\mathrm{BK}}(p) = c_{\mathrm{KS}}(p) = \frac{1}{8}$ holds if and only if $p = \frac{1}{2}$ (see Figure 2.1). Note that

$$\lim_{p \to \frac{1}{2}} c_{\mathrm{BK}}(p) = \lim_{p \to \frac{1}{2}} c_{\mathrm{KS}}(p) = \tfrac{1}{8}$$

which implies the continuity of $c_{\mathrm{BK}}(\cdot)$ and $c_{\mathrm{KS}}(\cdot)$ over the interval $[0,1]$.

The improved bound in Lemma 2.4.6 (cf. Lemma 2.4.1) leads to the following improvement of Hoeffding's inequality (Theorem 2.4.5):

Theorem 2.4.7 (Berend and Kontorovich inequality)**.** Let $\{U_k\}_{k=1}^n$ be a sequence of independent and bounded random variables such that, for every $k \in \{1, \ldots, n\}$, $U_k \in [a_k, b_k]$ holds a.s. for some constants $a_k, b_k \in \mathbb{R}$. Let $\mu_n \triangleq \sum_{k=1}^n \mathbb{E}[U_k]$. Then,

$$\mathbb{P}\left(\left|\sum_{k=1}^n U_k - \mu_n\right| \geq r\right) \leq 2\exp\left(-\frac{r^2}{4\sum_{k=1}^n c_k\,(b_k - a_k)^2}\right), \quad \forall\, r \geq 0 \tag{2.4.56}$$

where $c_k \triangleq c_{\mathrm{BK}}(p_k)$ (see (2.4.51)) with

$$p_k = \frac{\mathbb{E}[U_k] - a_k}{b_k - a_k}. \tag{2.4.57}$$

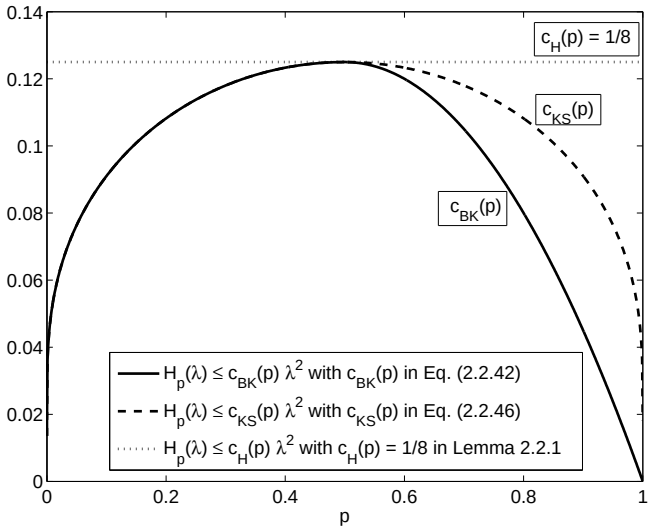

Figure 2.1: A comparison between upper bounds on the Hoeffding function $H_p(\lambda)$ in (2.4.6); these bounds are of the type $H_p(\lambda) \leq c(p)\,\lambda^2$ for every $p \in [0,1]$ and $\lambda \geq 0$ (see Eqs. (2.4.12), (2.4.53) and (2.4.54) with $c(p) = \frac{1}{8}$ or $c(p)$ in (2.4.51) and (2.4.55), respectively; these values of $c(p)$ correspond to the dotted, solid and dashed lines, respectively, as a function of $p \in [0,1]$.)

Proof. Inequality (2.4.56) follows from a combination of the Chernoff bound and Lemma 2.4.6 (similarly to the proof of Theorem 2.4.5 that relies on the Chernoff bound and Lemma 2.4.1). □

A loosening of the bound in Theorem 2.4.56, by a replacement of $c_k \triangleq c_{\mathrm{BK}}(p_k)$ with $\widetilde{c}_k \triangleq c_{\mathrm{KS}}(p_k)$ (see (2.4.51), (2.4.52) and (2.4.55)), gives the Kearns-Saul inequality in [97]:

Corollary 2.4.8 (Kearns–Saul inequality)**.** Let $\{U_k\}_{k=1}^n$ be a sequence of independent and bounded random variables such that, for every $k \in \{1,\ldots,n\}$, $U_k \in [a_k, b_k]$ holds a.s. for some constants $a_k, b_k \in \mathbb{R}$. Let $\mu_n \triangleq \sum_{k=1}^n \mathbb{E}[U_k]$. Then, for every $r \geq 0$,

$$\mathbb{P}\left(\left|\sum_{k=1}^n U_k - \mu_n\right| \geq r\right) \leq 2\exp\left(-\frac{r^2}{4\sum_{k=1}^n c_k\,(b_k-a_k)^2}\right) \qquad (2.4.58)$$

where $c_k = c_{\mathrm{KS}}(p_k)$ with $c_{\mathrm{KS}}(\cdot)$ and p_k in (2.4.55) and (2.4.57), respectively. The bound in (2.4.58) improves Hoeffding's inequality in (2.4.48) unless $p_k = \frac{1}{2}$ (i.e., if $\mathbb{E}[U_k] = \frac{a_k+b_k}{2}$) for every $k \in \{1, \ldots, n\}$; in the latter case, both bounds coincide.

An information-theoretic proof of the basic inequality that leads to the Kearns-Saul inequality is given in Section 3.4.3. Another recent refinement of Hoeffding's inequality is provided in [99].

2.5 Refined versions of the Azuma–Hoeffding inequality

The following section considers generalized and refined versions of the Azuma-Hoeffding inequality (see Sections 2.5.1 and 2.5.2). A derivation of one-sided inequalities for sub and super martingales is considered as well (see Section 2.5.3).

2.5.1 A generalization of the Azuma–Hoeffding inequality

The following theorem generalizes the Azuma-Hoeffding inequality for real-valued martingale sequences $\{X_k, \mathscr{F}_k\}_{k=0}^n$ with bounded differences in the case where the differences $\xi_k \triangleq X_k - X_{k-1}$ are bounded between the endpoints of *asymmetric* intervals around zero. Furthermore, it states that the same bound holds not only for the probability of the event where $|X_n - X_0| \geq r$, for some $r \geq 0$, but also for the probability of the more likely event where there exists an index $k \in \{1, \ldots, n\}$ such that $|X_k - X_0| \geq r$; the idea that strengthens the bound to hold for the latter event applies to all the concentration inequalities derived in this chapter.

Theorem 2.5.1 (A generalization of the Azuma-Hoeffding Inequality)**.** Let $\{X_k, \mathscr{F}_k\}_{k=0}^n$ be a real-valued martingale sequence. Suppose that $a_1, b_1, \ldots, a_n, b_n$ are constants such that $a_k \leq X_k - X_{k-1} \leq b_k$ holds a.s. for every $k \in \{1, \ldots, n\}$. Then, for every $r \geq 0$,

$$\mathbb{P}\Big(\max_{k \in \{1,\ldots,n\}} |X_k - X_0| \geq r\Big) \leq 2 \exp\left(-\frac{r^2}{4 \sum_{k=1}^n c_k \, (b_k - a_k)^2}\right) \quad (2.5.1)$$

where $c_k = c(p_k)$ with

$$p_k = -\frac{a_k}{b_k - a_k} \in [0,1], \quad \forall\, k \in \{1,\ldots,n\},$$

and $c(\cdot) = c_{\mathrm{BK}}(\cdot)$ is introduced in (2.4.51) over the interval $[0,1]$.

Remark 2.5. In the following, it is shown that the Azuma-Hoeffding inequality (Theorem 2.4.2) is a special case of Theorem 2.5.1. Consider the setting in the Azuma-Hoeffding inequality where the intervals $[a_k, b_k]$ in Theorem 2.5.1 are symmetric around zero, i.e., $b_k = -a_k = d_k$ for every $k \in \{1,\ldots,n\}$, and for some non-negative reals $d_1,\ldots,d_n$. In this special case, it follows from Theorem 2.5.1 that $p_k = \frac{1}{2}$, and $c(p_k) = \frac{1}{8}$ for every k. Hence, from (2.5.1), we have

$$\begin{aligned}\mathbb{P}(|X_n - X_0| \geq r) &\leq \mathbb{P}\Big(\max_{k\in\{1,\ldots,n\}} |X_k - X_0| \geq r\Big) \\ &\leq 2\exp\left(-\frac{r^2}{2\sum_{k=1}^n d_k^2}\right), \quad \forall\, r \geq 0,\end{aligned}$$

which gives the Azuma-Hoeffding inequality in (2.4.18).

Proof. In the following, the proof of the Azuma-Hoeffding inequality is modified for a derivation of the generalized inequality in (2.5.1). As a first step, the equality in (2.4.19) is replaced by the equality

$$\begin{aligned}&\mathbb{P}\Big(\max_{k\in\{1,\ldots,n\}} |X_k - X_0| \geq r\Big) \\ &= \mathbb{P}\Big(\max_{k\in\{1,\ldots,n\}} (X_k - X_0) \geq r\Big) + \mathbb{P}\Big(\max_{k\in\{1,\ldots,n\}} (X_0 - X_k) \geq r\Big). \quad (2.5.2)\end{aligned}$$

Let $\xi_k = X_k - X_{k-1}$ be the differences of the martingale sequence, then $\mathbb{E}[\xi_k|\mathscr{F}_{k-1}] = 0$ and $a_k \leq \xi_k \leq b_k$ hold a.s. for every $k \in \{1,\ldots,n\}$.

Recall that a composition of a convex function with a martingale gives a sub-martingale with respect to the same filtration (see Theorem 2.1.1). Since $\{X_k - X_0, \mathscr{F}_k\}_{k=1}^n$ is a martingale and $f_t(x) = \exp(tx)$ is a convex function over $\mathbb{R}$ for every $t \in \mathbb{R}$, it follows that $\{\exp(t(X_k - X_0)), \mathscr{F}_k\}_{k=1}^n$ is a sub-martingale for every $t \in \mathbb{R}$. From the maximal inequality for sub-martingales (a.k.a. the Doob-

Kolmogorov inequality), which states that if $\{Y_k, \mathscr{F}_k\}_{k=1}^n$ is a sub-martingale then

$$\mathbb{P}\Big(\max_{1\leq k\leq n} Y_k \geq \lambda\Big) \leq \frac{\mathbb{E}[|Y_n|]}{\lambda}, \quad \forall\, \lambda > 0$$

(see, e.g., [84, Theorem 14.3.1]), it follows that for every $t \geq 0$

$$\begin{aligned}
&\mathbb{P}\Big(\max_{k\in\{1,\ldots,n\}} (X_k - X_0) \geq r\Big) \\
&= \mathbb{P}\Big(\max_{k\in\{1,\ldots,n\}} \exp\big(t(X_k - X_0)\big) \geq \exp(tr)\Big) \\
&\leq \exp(-tr)\, \mathbb{E}\left[\exp\big(t(X_k - X_0)\big)\right] \\
&= \exp(-tr)\, \mathbb{E}\left[\exp\left(t \sum_{k=1}^n \xi_k\right)\right]. \qquad (2.5.3)
\end{aligned}$$

Hence, by applying the maximal inequality for sub-martingales instead of the Chernoff bound, inequality (2.4.20) is replaced with the stronger result in (2.5.3). Similarly to the proof of the Azuma-Hoeffding inequality, by the law of iterated expectations, we have from (2.4.21)

$$\mathbb{E}\Big[\exp\Big(t \sum_{k=1}^n \xi_k\Big)\Big] = \mathbb{E}\Big[\exp\Big(t \sum_{k=1}^{n-1} \xi_k\Big)\, \mathbb{E}\big[\exp(t\xi_n) \,|\, \mathscr{F}_{n-1}\big]\Big].$$

In the following, Lemma 2.4.6 is applied with the conditioning on $\mathscr{F}_{n-1}$. Based on the information that $\mathbb{E}[\xi_n|\mathscr{F}_{n-1}] = 0$ and $\xi_n \in [a_n, b_n]$ a.s., it follows that

$$\mathbb{E}\big[\exp(t\xi_n) \,|\, \mathscr{F}_{n-1}\big] \leq \exp\Big(c_n (b_n - a_n)^2 t^2\Big) \qquad (2.5.4)$$

where $c_n \triangleq c_{\mathrm{BK}}(p_n)$ is given in (2.4.51) with (see (2.4.52))

$$p_n = \frac{\mathbb{E}[\xi_n|\mathscr{F}_{n-1}] - a_n}{b_n - a_n} = -\frac{a_n}{b_n - a_n}.$$

(If $b_n = -a_n \triangleq d_n$ for a non-negative real number d_n then $p_n = \frac{1}{2}$ and $c_n = c_{\mathrm{BK}}(p_n) = \frac{1}{8}$, and inequality (2.5.4) is particularized to (2.4.22); the latter inequality can be obtained by applying Hoeffding's lemma, as in the proof of the Azuma-Hoeffding lemma.) Continuing recursively

in a similar manner, in parallel to (2.4.23), the quantity in (2.4.21) is upper-bounded by

$$\mathbb{E}\Big[\exp\Big(t\sum_{k=1}^{n}\xi_k\Big)\Big] \leq \exp\left(t^2\sum_{k=1}^{n} c_k(b_k-a_k)^2\right).$$

The combination of this bound with (2.5.3) gives that, for every $r \geq 0$,

$$\begin{aligned}&\mathbb{P}\Big(\max_{k\in\{1,\ldots,n\}}(X_k - X_0) \geq r\Big)\\ &\leq \exp\left(-tr + t^2\sum_{k=1}^{n} c_k(b_k-a_k)^2\right), \quad \forall\, t \geq 0. \end{aligned} \tag{2.5.5}$$

An optimization with respect to the non-negative parameter t gives

$$t = \frac{r}{2\sum_{k=1}^{n} c_k(b_k-a_k)^2}$$

and the substitution of this optimized value into (2.5.5) yields that, for every $r \geq 0$,

$$\mathbb{P}\Big(\max_{k\in\{1,\ldots,n\}}(X_k - X_0) \geq r\Big) \leq \exp\left(-\frac{r^2}{4\sum_{k=1}^{n} c_k(b_k-a_k)^2}\right). \tag{2.5.6}$$

The same bound as in (2.5.6) holds for $\mathbb{P}\Big(\max_{k\in\{1,\ldots,n\}}(X_0 - X_k) \geq r\Big)$. Using these two bounds on the right side of (2.5.2) completes the proof of Theorem 2.5.1. □

2.5.2 On martingales with uniformly bounded differences

Example 2.5 serves to motivate a derivation of an improvement of the Azuma-Hoeffding inequality with a constraint on the conditional variance of the martingale sequence. In the following, we 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 differences of the martingale). A new condition is added for the derivation of the next concentration inequality, where it is assumed that a.s.

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

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

One of the weaknesses of the Azuma–Hoeffding and McDiarmid's inequalities (see Theorems 2.4.2 and 2.4.3) is their insensitivity to the variance, leading to suboptimal exponents compared to the central limit theorem (CLT) and moderate deviation principle (MDP). The following result in [93] (see also [85, Corollary 2.4.7]) relies on a constraint on the conditional variance:

Theorem 2.5.2. Let $\{X_k, \mathscr{F}_k\}_{k=0}^n$ 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 $k \in \{1, \ldots, n\}$:

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

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

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) \leq 2 \exp\left(-n\, H\left(\frac{\delta+\gamma}{1+\gamma}\middle\| \frac{\gamma}{1+\gamma}\right)\right) \tag{2.5.7}$$

where

$$\gamma \triangleq \frac{\sigma^2}{d^2}, \quad \delta \triangleq \frac{\alpha}{d} \tag{2.5.8}$$

and

$$H(p\|q) \triangleq p \ln\Big(\frac{p}{q}\Big) + (1-p)\ln\Big(\frac{1-p}{1-q}\Big), \quad \forall\, p, q \in [0,1] \tag{2.5.9}$$

denotes the binary relative entropy. If $\delta > 1$, the probability on the left side of (2.5.7) 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 (2.4.21). The new ingredient in this proof is the use of the so-called Bennett's inequality (see, e.g., [85, Lemma 2.4.1]), which improves upon Lemma 2.4.1 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

$$\mathbb{E}\big[e^{\lambda X}\big] \leq \frac{e^{\lambda\overline{x}}\left[(b-\overline{x})^2 e^{-\frac{\lambda\sigma^2}{b-\overline{x}}} + \sigma^2 e^{\lambda(b-\overline{x})}\right]}{(b-\overline{x})^2 + \sigma^2}. \tag{2.5.10}$$

The proof of (2.5.10) is provided in Appendix 2.A for completeness.

We now apply Bennett's inequality (2.5.10) to the conditional law of ξ_k given the σ-algebra $\mathscr{F}_{k-1}$. Since $\mathbb{E}[\xi_k|\mathscr{F}_{k-1}] = 0$, $\mathsf{var}[\xi_k|\mathscr{F}_{k-1}] \leq \sigma^2$ and $\xi_k \leq d$ a.s. for $k \in \mathbb{N}$, we have

$$\mathbb{E}\left[\exp(t\xi_k) \,|\, \mathscr{F}_{k-1}\right] \leq \frac{\sigma^2 \exp(td) + d^2 \exp\left(-\frac{t\sigma^2}{d}\right)}{d^2 + \sigma^2}, \qquad \text{a.s..} \quad (2.5.11)$$

From (2.4.21) and (2.5.11) 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 recursively, 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 (2.5.8), we can rewrite this inequality as

$$\mathbb{E}\Big[\exp\Big(t\sum_{k=1}^{n} \xi_k\Big)\Big] \leq \Big(\frac{\gamma \exp(td) + \exp(-\gamma td)}{1+\gamma}\Big)^n, \quad \forall\, t \geq 0. \quad (2.5.12)$$

Let $x \triangleq td$ (so $x \geq 0$). We can now use (2.5.12) with the Chernoff bounding technique to get that for every $\alpha \geq 0$ (from the definition of δ in (2.5.8), $\alpha t = \delta x$)

$$\begin{aligned}
&\mathbb{P}(X_n - X_0 \geq \alpha n) \\
&\leq \exp(-\alpha n t)\, \mathbb{E}\Big[\exp\Big(t\sum_{k=1}^{n} \xi_k\Big)\Big] \\
&\leq \left(\frac{\gamma \exp\big((1-\delta)x\big) + \exp\big(-(\gamma+\delta)x\big)}{1+\gamma}\right)^n, \quad \forall\, x \geq 0. \qquad (2.5.13)
\end{aligned}$$

Consider first the case where $\delta = 1$ (i.e., $\alpha = d$). Then (2.5.13) becomes

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

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

$$\mathbb{P}(X_n - X_0 \geq dn) \leq \left(\frac{\gamma}{1+\gamma}\right)^n. \tag{2.5.14}$$

Otherwise, if $\delta \in [0, 1)$, we minimize the base of the exponent on the right side of (2.5.13) with respect to the free parameter $x \geq 0$. Setting the derivative of this exponent to zero yields that the optimal value of x is given by

$$x = \left(\frac{1}{1+\gamma}\right) \ln\left(\frac{\gamma+\delta}{\gamma(1-\delta)}\right). \tag{2.5.15}$$

Substituting (2.5.15) into the right side of (2.5.13) gives that, for every $\alpha \geq 0$,

$$\begin{aligned} \mathbb{P}(X_n - X_0 \geq \alpha n) &\leq \left[\left(\frac{\gamma+\delta}{\gamma}\right)^{-\frac{\gamma+\delta}{1+\gamma}} (1-\delta)^{-\frac{1-\delta}{1+\gamma}}\right]^n \\ &= \exp\left(-n\, H\left(\frac{\delta+\gamma}{1+\gamma}\middle\|\frac{\gamma}{1+\gamma}\right)\right) \end{aligned} \tag{2.5.16}$$

where $H(\cdot\|\cdot)$ is introduced in (2.5.9). Finally, if $\delta > 1$ (i.e., $\alpha > d$), the exponent is equal to $+\infty$. The application of inequality (2.5.16) to the martingale $\{-X_k, \mathscr{F}_k\}_{k=0}^{\infty}$ gives the same upper bound for the other tail probability $\mathbb{P}(X_n - X_0 \leq -\alpha n)$. Overall, we get the bound (2.5.7), which completes the proof of Theorem 2.5.2. □

Remark 2.6. The divergence (a.k.a. Kullback-Leibler distance or relative entropy) between two probability measures P and Q is denoted, throughout this monograph, by $D(P\|Q)$. The notation $H(p\|q)$ is used in (2.5.9) for the divergence in the special case where P and Q are Bernoulli(p) and Bernoulli(q), respectively. In this case, where $P = \text{Bernoulli}(p)$ and $Q = \text{Bernoulli}(q)$, we have $D(P\|Q) \triangleq H(p\|q)$.

Here is an illustration of how one can use Theorem 2.5.2 for getting better bounds in comparison to the Azuma–Hoeffding inequality:

Example 2.8. Let $d > 0$ and $\varepsilon \in (0, \frac{1}{2}]$ be some constants. Consider a discrete-time real-valued martingale $\{X_k, \mathscr{F}_k\}_{k=0}^{\infty}$ where a.s. $X_0 = 0$,

and for every $m \in \mathbb{N}$

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

This implies that $\mathbb{E}[X_m - X_{m-1} \,|\, \mathscr{F}_{m-1}] = 0$ a.s. for every $m \in \mathbb{N}$, and, since X_{m-1} is $\mathscr{F}_{m-1}$-measurable, we have $\mathbb{E}[X_m \,|\, \mathscr{F}_{m-1}] = X_{m-1}$ 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 the Azuma–Hoeffding inequality gives

$$\mathbb{P}(X_k \geq kx) \leq \exp\left(-\frac{kx^2}{2d^2}\right), \qquad \forall\, x \geq 0 \tag{2.5.17}$$

independently of the value of ε (note that $X_0 = 0$ a.s.). However, we can use Theorem 2.5.2 to get a better bound; since for every $m \in \mathbb{N}$

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

it follows from (2.5.16) that

$$\mathbb{P}(X_k \geq kx) \leq \exp\left(-k\, H\left(\frac{x(1-\varepsilon)}{d} + \varepsilon \,\Big\|\, \varepsilon\right)\right), \qquad \forall\, x \geq 0. \tag{2.5.18}$$

Consider the case where $\varepsilon \to 0$. Then, for arbitrary $x > 0$ and $k \in \mathbb{N}$, the Azuma–Hoeffding inequality in (2.5.17) provides an upper bound that is strictly positive independently of ε, whereas the one-sided concentration inequality of Theorem 2.5.2 implies a bound in (2.5.18) that tends to zero.

Corollary 2.5.3. Let $\{X_k, \mathscr{F}_k\}_{k=0}^n$ 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 $\alpha \geq 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) \leq 2\exp\left(-n f(\delta)\right) \tag{2.5.19}$$

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

$$f(\delta) = \begin{cases} \ln(2)\Big[1 - h_2\left(\frac{1-\delta}{2}\right)\Big], & 0 \leq \delta \leq 1 \\ +\infty, & \delta > 1 \end{cases} \tag{2.5.20}$$

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 2.5.2 (since there is no constraint on the conditional variance, one can take $\sigma^2 = d^2$), the corresponding exponent in (2.5.7) is equal to

$$H\Big(\frac{1+\delta}{2}\Big\|\frac{1}{2}\Big) = f(\delta), \tag{2.5.21}$$

since, from (2.5.9), it is easy to verify that $H(p\|\frac{1}{2}) = \ln 2\,[1 - h_2(p)]$ for every $p \in [0,1]$. □

An alternative proof of Corollary 2.5.3, which provides some further insight, is suggested in the following.

Proof. As a first step, a refined version of Hoeffding's lemma is provided (cf. Lemma 2.4.1).

Lemma 2.5.4. Let $U \in \mathbb{R}$ be a random variable, such that $U \in [a,b]$ a.s. for some finite $a < b$, and $\mathbb{E}U = \frac{a+b}{2}$. Then, for every $t \geq 0$,

$$\mathbb{E}\left[\exp\left(t(U - \mathbb{E}U)\right)\right] \leq \cosh\left(\frac{t(b-a)}{2}\right). \tag{2.5.22}$$

Proof. This refinement of (2.4.5), if $\mathbb{E}U = \frac{a+b}{2}$, follows from (2.4.10). □

The proof of Corollary 2.5.3 continues by following the proof of the Azuma–Hoeffding inequality. Recall that $\xi_k = X_k - X_{k-1}$, for all $k \in \mathbb{N}$, form the differences of the martingale sequence with $|\xi_k| \leq d$ (in the case where $d_k = d$, independently of k) and $\mathbb{E}[\xi_k|\mathscr{F}_{k-1}] = 0$. Using a conditional version of Lemma 2.5.4, the bound in (2.4.22) is improved to

$$\mathbb{E}\big[\exp(t\xi_n)\,|\,\mathscr{F}_{n-1}\big] \leq \cosh(td), \quad \forall\, t \geq 0 \tag{2.5.23}$$

and continuing recursively, the quantity in (2.4.21) is upper bounded by

$$\mathbb{E}\Big[\exp\Big(t\sum_{k=1}^{n}\xi_k\Big)\Big] \leq \cosh^n(td), \quad \forall\, t \geq 0.$$

Based on Chernoff's inequality, the following refinement of (2.4.24) holds

$$\begin{aligned}\mathbb{P}(X_n - X_0 \geq \alpha n) &\leq \exp(-\alpha n t)\, \cosh^n(td) \\ &= \exp\Big(-n\big[\alpha t - \ln\cosh(td)\big]\Big), \quad \forall\, t \geq 0. \quad (2.5.24)\end{aligned}$$

Due to the bounded differences assumption, we have (a.s.)

$$|X_n - X_0| \leq \sum_{k=1}^{n} |X_k - X_{k-1}| \leq nd$$

so, if $\alpha > d$, we have $\mathbb{P}(X_n - X_0 \geq \alpha n) = 0$. If $0 \leq \alpha < d$, an optimization of the free parameter t on the right side of (2.5.24) gives $t = \frac{1}{d}\tanh^{-1}\left(\frac{\alpha}{d}\right)$. Substituting this optimized value of t into (2.5.24), combined with the use of the following two identities for hyperbolic functions:

$$\begin{aligned}\tanh^{-1}(x) &= \frac{1}{2}\ln\left(\frac{1+x}{1-x}\right), \quad \forall\, |x| < 1, \\ \cosh(x) &= \frac{1}{\sqrt{1-\tanh^2(x)}}, \quad \forall\, x \in \mathbb{R},\end{aligned}$$

yield that the exponent on the right side of (2.5.24) is equal to

$$\begin{aligned}&\alpha t - \ln\cosh(td) \\ &= \frac{\alpha}{2d}\ln\left(\frac{1+\frac{\alpha}{d}}{1-\frac{\alpha}{d}}\right) + \frac{1}{2}\ln\left(1-\frac{\alpha^2}{d^2}\right) \\ &= \frac{1}{2}\left(1+\frac{\alpha}{d}\right)\ln\left(1+\frac{\alpha}{d}\right) + \frac{1}{2}\left(1-\frac{\alpha}{d}\right)\ln\left(1-\frac{\alpha}{d}\right) \\ &= \ln 2\,\left[1 - h_2\left(\frac{1}{2}\,\left(1-\frac{\alpha}{d}\right)\right)\right] \\ &= f(\delta)\end{aligned}$$

where the last equality follows from (2.5.8) and (2.5.20). This gives the exponential bound in Corollary 2.5.3 for $\alpha \in [0, d)$. Finally, the result

of this corollary for $\alpha = d$ is obtained by letting t tend to infinity in the exponential bound on the right side of (2.5.24). This gives

$$\lim_{t\to\infty} \big(td - \ln\cosh(td)\big) = \ln 2, \quad \forall\, d > 0$$

and, consequently,

$$\mathbb{P}(X_n - X_0 \geq dn) \leq 2^{-n}$$

which proves Corollary 2.5.3 for $\alpha = d$. Note that the factor 2 in the bound of (2.5.19) was justified in the proof of Theorem 2.4.2. □

Remark 2.7. Corollary 2.5.3, which is a special case of Theorem 2.5.2 with $\gamma = 1$, forms a tightening of the Azuma–Hoeffding inequality for the case where $d_k = d$ (independently of k). This follows from Pinsker's inequality, which implies that $f(\delta) > \frac{\delta^2}{2}$ for $\delta > 0$. Figure 2.2 plots the two exponents of the Azuma–Hoeffding inequality and its improvement in Corollary 2.5.3, and they nearly coincide for $\delta \leq 0.4$. The exponential bound of Theorem 2.5.2 is improved as the value of $\gamma \in (0, 1)$ is reduced (see Figure 2.2); this holds since the additional constraint on the conditional variance in Theorem 2.5.2 has a growing effect by reducing the value of γ.

Theorem 2.5.2 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 $\alpha \geq 0$ (in contrast to *large-deviation* events of the form $\{|X_n - X_0| \geq \alpha n\}$):

Proposition 2.1. Let $\{X_k, \mathscr{F}_k\}$ be a discrete-time real-valued martingale that satisfies the conditions of Theorem 2.5.2. Then, for every $\alpha \geq 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) \leq 2\exp\Big(-\frac{\delta^2}{2\gamma}\Big)\Big(1 + O\big(n^{-\frac{1}{2}}\big)\Big). \tag{2.5.25}$$

Remark 2.8. From Proposition 2.1, for an arbitrary $\alpha \geq 0$, the upper bound on $\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n})$ improves the exponent of the Azuma–Hoeffding inequality by a factor of $\frac{1}{\gamma}$.

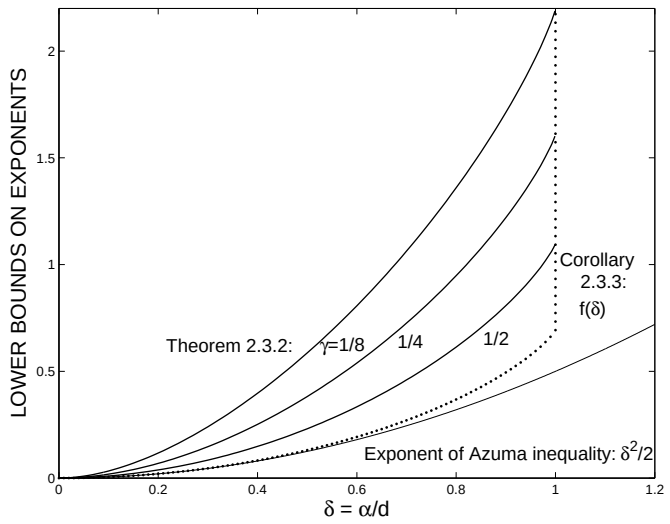

Figure 2.2: Plot of the lower bounds on the exponents in the Azuma–Hoeffding inequality and the improved bounds in Theorem 2.5.2 and Corollary 2.5.3. The pointed line refers to the exponent in Corollary 2.5.3, and the three solid lines for $\gamma = \frac{1}{8}, \frac{1}{4}$ and $\frac{1}{2}$ refer to the exponents in Theorem 2.5.2.

Proof. Let $\{X_k, \mathscr{F}_k\}_{k=0}^{\infty}$ be a discrete-time martingale that satisfies the conditions in Theorem 2.5.2. From (2.5.7), for every $\alpha \geq 0$ and $n \in \mathbb{N}$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) \leq 2\exp\left(-n\, H\left(\frac{\delta_n + \gamma}{1+\gamma}\middle\|\frac{\gamma}{1+\gamma}\right)\right) \tag{2.5.26}$$

where, following (2.5.8),

$$\gamma \triangleq \frac{\sigma^2}{d^2}, \quad \delta_n \triangleq \frac{\frac{\alpha}{\sqrt{n}}}{d} = \frac{\delta}{\sqrt{n}}\,. \tag{2.5.27}$$

With these definitions, we have

$$H\left(\frac{\delta_n + \gamma}{1+\gamma}\middle\|\frac{\gamma}{1+\gamma}\right) = \frac{\gamma}{1+\gamma}\left[\left(1 + \frac{\delta}{\gamma\sqrt{n}}\right)\ln\left(1 + \frac{\delta}{\gamma\sqrt{n}}\right) + \frac{1}{\gamma}\left(1 - \frac{\delta}{\sqrt{n}}\right)\ln\left(1 - \frac{\delta}{\sqrt{n}}\right)\right]. \tag{2.5.28}$$

Using the power series expansion

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

in (2.5.28), it follows that for every $n > \frac{\delta^2}{\gamma^2}$

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

Substituting this into the exponent on the right side of (2.5.26) gives (2.5.25). □

2.5.3 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 $\{X_k, \mathscr{F}_k\}_{k=0}^n$ with $\{-X_k, \mathscr{F}_k\}_{k=0}^n$, 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 2.5.2 to sub- and super-martingales gives the following inequality:

Corollary 2.5.5. Let $\{X_k, \mathscr{F}_k\}_{k=0}^{\infty}$ be a discrete-time real-valued super-martingale. Assume that, for some constants $d, \sigma > 0$, the following two requirements are satisfied a.s.:

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

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

$$\mathbb{P}(X_n - X_0 \geq \alpha n) \leq \exp\left(-n\, H\left(\frac{\delta+\gamma}{1+\gamma}\middle\|\frac{\gamma}{1+\gamma}\right)\right) \tag{2.5.29}$$

where γ and δ are defined in (2.5.8), and the binary divergence $H(p\|q)$ is introduced in (2.5.9). Alternatively, if $\{X_k, \mathscr{F}_k\}_{k=0}^{\infty}$ is a sub-martingale, the same upper bound in (2.5.29) holds for the probability $\mathbb{P}(X_n - X_0 \leq -\alpha n)$. If $\delta > 1$, these two probabilities are zero.

Proof. It is similar to the proof of Theorem 2.5.2; 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 $\xi_k \triangleq X_k - \mathbb{E}[X_k \,|\, \mathscr{F}_{k-1}]$ is $\mathscr{F}_k$-measurable. Therefore, we have $\mathbb{P}(X_n - X_0 \geq \alpha n) \leq \mathbb{P}(\sum_{k=1}^{n} \xi_k \geq \alpha n)$ where, a.s., $\xi_k \leq d$, $\mathbb{E}[\xi_k \,|\, \mathscr{F}_{k-1}] = 0$, and $\mathsf{var}(\xi_k \,|\, \mathscr{F}_{k-1}) \leq \sigma^2$. The rest of the proof coincides with the proof of Theorem 2.5.2 (starting from (2.4.20)). The other inequality for sub-martingales holds due to the fact that if $\{X_k, \mathscr{F}_k\}$ is a sub-martingale then $\{-X_k, \mathscr{F}_k\}$ is a super-martingale. □

The reader is referred to [100] for an extension of Hoeffding's inequality to super-martingales with differences bounded from above (or sub-martingales with differences bounded from below), and to [101] for large deviation exponential inequalities for super-martingales.

2.6 Relations to classical results in probability theory

2.6.1 The martingale central limit theorem

A relation between Proposition 2.1 and the martingale central limit theorem (CLT) is considered in the following.

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

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

Let

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

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

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

exists in probability and is positive. The martingale CLT asserts that, under the above conditions, $\left\{\frac{S_n}{\sqrt{n}}\right\}$ 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., [102, pp. 475–478]).

Let $\{X_k, \mathscr{F}_k\}_{k=0}^{\infty}$ be a real-valued martingale with bounded differences where there exists a constant d such that a.s.

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

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

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

and $Y_0 \triangleq 0$. Then $\{Y_k, \mathscr{F}_k\}_{k=0}^{\infty}$ 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,

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

Consequently, from the martingale CLT, it follows that

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

so, 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 (2.4.29). In terms of the notation in (2.5.8), we have $\frac{\alpha}{\sigma} = \frac{\delta}{\sqrt{\gamma}}$, so that

$$\lim_{n\to\infty} \mathbb{P}\Big(|X_n - X_0| \geq \alpha\sqrt{n}\Big) = 2\,Q\bigg(\frac{\delta}{\sqrt{\gamma}}\bigg). \tag{2.6.1}$$

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 in the pre-exponent (see the right side of (2.5.25)). 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 (2.4.30) and from (2.6.1), it follows that the exponent in the concentration inequality (2.5.25) cannot be improved without imposing additional conditions on the martingale sequence.

2.6.2 The moderate deviations principle

The moderate deviations principle (MDP) on the real line (see, e.g., [85, Theorem 3.7.1]) states the following: Let $\{X_i\}_{i=1}^n$ be a sequence of real-valued i.i.d. random variables such that $\Lambda_X(\lambda) \triangleq \ln \mathbb{E}[e^{\lambda X_i}] < \infty$ 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

$$Z_n \triangleq \sqrt{\frac{a_n}{n}} \sum_{i=1}^n X_i, \quad \forall\, n \in \mathbb{N}. \tag{2.6.2}$$

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

$$\begin{aligned} -\frac{1}{2\sigma^2} \inf_{x \in \Gamma^0} x^2 &\le \liminf_{n\to\infty} a_n \ln \mathbb{P}(Z_n \in \Gamma) \\ &\le \limsup_{n\to\infty} a_n \ln \mathbb{P}(Z_n \in \Gamma) \\ &\le -\frac{1}{2\sigma^2} \inf_{x \in \overline{\Gamma}} x^2 \end{aligned} \tag{2.6.3}$$

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 (2.6.2),

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| \ge \alpha n^\eta\right) = \mathbb{P}(Z_n \in \Gamma) \tag{2.6.4}$$

so, by the MDP,

$$\lim_{n\to\infty} n^{1-2\eta} \ln \mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq \alpha n^{\eta}\right) = -\frac{\alpha^2}{2\sigma^2}, \qquad \forall\, \alpha \geq 0. \tag{2.6.5}$$

We show in Appendix 2.B that, in contrast to the Azuma–Hoeffding inequality, Theorem 2.5.2 provides an upper bound on the left side of (2.6.4) which coincides with the asymptotic limit in (2.6.5). The analysis in Appendix 2.B provides another interesting link between Theorem 2.5.2 and a classical result in probability theory, and thus emphasizes the significance of the refinements of the Azuma–Hoeffding inequality.

2.6.3 Functions of discrete-time Markov chains

An interesting relation between discrete-time Markov chains and martingales is the following (see, e.g., [103, 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 $\psi\colon \mathcal{S} \to \mathbb{R}$ be a *harmonic function* of the Markov chain, i.e.,

$$\sum_{s\in\mathcal{S}} p_{s',s}\psi(s) = \psi(s'), \quad \forall\, s' \in \mathcal{S} \tag{2.6.6}$$

and assume also that ψ is a measurable and bounded function. Let $Y_n \triangleq \psi(X_n)$ for every $n \geq 0$, and let $\{\mathscr{F}_n\}$ be the natural filtration where $\mathscr{F}_n = \sigma(X_0, \ldots, X_n)$. It is a remarkable fact that $\{Y_n, \mathscr{F}_n\}$ is a martingale; this property holds since Y_n is $\mathscr{F}_n$-measurable, $\mathbb{E}[|Y_n|] < \infty$ (due to the requirement that ψ is bounded), and from (2.6.6)

$$\mathbb{E}[Y_n \,|\, \mathscr{F}_{n-1}] = \sum_{s\in\mathcal{S}} p_{X_{n-1},s}\, \psi(s) = \psi(X_{n-1}) = Y_{n-1}, \quad \forall\, n \in \mathbb{N}. \tag{2.6.7}$$

This relation between Markov chains and martingales enables to apply the concentration inequalities of this chapter to the composition of a bounded harmonic function and a Markov chain; note that the boundedness of ψ implies that the differences of the martingale sequence are uniformly bounded (this holds since, for every n, we have $|Y_n - Y_{n-1}| \leq 2\|\psi\|_{\infty} < \infty$).

More generally, let $\underline{\psi}$ be a right eigenvector of the transition matrix P such that $\|\underline{\psi}\|_\infty < \infty$, and let λ be its corresponding eigenvalue such that $|\lambda| \geq 1$. Let $\mathcal{S} = \{s_1, s_2, \ldots\}$ be the countable state space of the Markov chain, and let $\psi \colon \mathcal{S} \to \mathbb{R}$ be a real-valued function such that $\psi(s_i)$ is equal to the i-th entry of the vector $\underline{\psi}$. Then, the following equality holds:

$$\sum_{s \in \mathcal{S}} p_{s',s} \psi(s) = \lambda \, \psi(s'), \quad \forall \, s' \in \mathcal{S}$$

which generalizes (2.6.6) (i.e., if $\lambda = 1$, the function ψ is harmonic). Similarly to (2.6.7), for every $n \geq 1$,

$$\mathbb{E}[\psi(X_n) \,|\, \mathscr{F}_{n-1}] = \lambda \, \psi(X_{n-1}).$$

Defining $Y_n = \lambda^{-n} \, \psi(X_n)$, for $n \geq 0$, implies that $\mathbb{E}[Y_n | \mathscr{F}_{n-1}] = Y_{n-1}$. Since $|\lambda| \geq 1$ and $\|\underline{\psi}\|_\infty < \infty$ then $\mathbb{E}[|Y_n|] < \infty$. Consequently, $\{Y_n, \mathscr{F}_n\}$ is a martingale sequence, and its differences are uniformly bounded. The latter property holds since, for every $n \geq 1$,

$$\begin{aligned}
&|Y_n - Y_{n-1}| \\
&\leq |\lambda|^{-n} \, |\psi(X_n)| + |\lambda|^{-(n-1)} \, |\psi(X_{n-1})| \\
&\leq |\psi(X_n)| + |\psi(X_{n-1})| \\
&\leq 2 \|\underline{\psi}\|_\infty < \infty.
\end{aligned}$$

Since $\{Y_n, \mathscr{F}_n\}$ is demonstrated to be a discrete-time martingale with uniformly bounded differences, the concentration inequalities of this chapter are applicable here as well.

Exponential deviation bounds for an important class of Markov chains, so-called Doeblin chains, were derived by Kontoyiannis [104]. These bounds are essentially identical to the Hoeffding inequality in the special case of i.i.d. random variables (see [104, Remark 1]).

2.7 Applications in information theory and coding

This section is focused on applications of the concentration inequalities, derived in this chapter via the martingale approach, in information theory, communications and coding.

2.7.1 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. The asymptotic average value of the normalized minimum distance is equal to (see [105, Section 2.C])

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

where $h_2^{-1}\colon [0,1] \to [0,\frac{1}{2}]$ denotes the inverse of the binary entropy function to the base 2.

Let $\mathtt{H}$ 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 $\mathtt{H}$ 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 of integer-valued random variables $\{X_i\}_{i=0}^n$ where X_i is defined to be the minimal number of linearly dependent columns of a random parity-check matrix $\mathbf{H}$, chosen uniformly from the ensemble, given that the first i columns of $\mathbf{H}$ are already revealed; this refers to a random process where sequentially, at every time instant, a new column of the random parity-check matrix $\mathbf{H}$ is revealed.

Recalling Fact 2 from Section 2.1, we see that this is a martingale sequence with the natural filtration $\{\mathscr{F}_i\}_{i=0}^n$ where $\mathscr{F}_i$ is the σ-algebra that is generated by all subsets of $n(1-R)\times n$ binary parity-check matrices whose first i columns are fixed. This martingale sequence has bounded differences, and it satisfies $|X_i - X_{i-1}| \leq 1$ for $i \in \{1,\ldots,n\}$; this can be verified by noticing that the observation of a new column of the random parity-check matrix $\mathbf{H}$ can change the minimal number of linearly dependent columns by at most 1. Note that the random variable 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 all the n columns of $\mathtt{H}$ are revealed, the code is known exactly). Hence, by the Azuma–Hoeffding 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 concentration theorem of the minimum distance around the expected value:

Theorem 2.7.1. 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 $[n\, h_2^{-1}(1-R) - \alpha\sqrt{n},\ n\, h_2^{-1}(1-R) + \alpha\sqrt{n}]$.

Remark 2.9. Note that some well-known capacity-approaching families of binary linear block codes have a 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 [106].

2.7.2 Expansion properties of random regular bipartite graphs

The Azuma–Hoeffding inequality is useful for analyzing the expansion properties of random bipartite graphs. The following theorem was proved by Sipser and Spielman [42, Theorem 25] in the context of bit-flipping decoding algorithms for expander codes. It is stated, in the following, 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 2.7.2. Let $\mathcal{G}$ be a bipartite graph that is chosen uniformly at random from the ensemble of bipartite graphs with n vertices on the left, a left degree l, and a right degree r. Let $\alpha \in (0,1)$ and $\delta > 0$ be fixed numbers. Then, with probability at least $1 - \exp(-\delta n)$, all sets of αn vertices on the left side of $\mathcal{G}$ are connected to at least

$$n\left[\frac{l\big(1-(1-\alpha)^r\big)}{r} - \sqrt{2l\alpha\,\big(h(\alpha)+\delta\big)}\right] \tag{2.7.1}$$

vertices (neighbors) on the right side of $\mathcal{G}$, where h is the binary entropy function to base e (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.

Let $\mathcal{V}$ denote a given set of $n\alpha$ vertices on the left side of the selected bipartite graph $\mathcal{G}$. The set $\mathcal{V}$ has $n\alpha l$ outgoing edges in $\mathcal{G}$. Let $X(\mathcal{G})$ be a random variable which denotes the number of neighbors of $\mathcal{V}$ on the right side of $\mathcal{G}$, and let $\mathbb{E}[X(\mathcal{G})]$ be the expected value of neighbors of $\mathcal{V}$ where all the bipartite graphs are chosen uniformly at random from the ensemble. This expected number is equal to

$$\mathbb{E}[X(\mathcal{G})] = \frac{nl\big(1-(1-\alpha)^r\big)}{r} \tag{2.7.2}$$

since, for each of the $\frac{nl}{r}$ vertices on the right side of $\mathcal{G}$, the probability that it has at least one edge in the subset of $n\alpha$ chosen vertices on the left side of $\mathcal{G}$ is $1-(1-\alpha)^r$.

Let us form a martingale sequence to estimate, via the Azuma–Hoeffding inequality, the probability that the actual number of neighbors deviates by a certain amount from the expected value in (2.7.2).

The set of $n\alpha$ vertices in $\mathcal{V}$ 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 random variable denoting the vertex on the right side of $\mathcal{G}$ which the i-th edge is connected to, where $i \in \{1,\dots,n\alpha l\}$. Let us define, for $i \in \{0,\dots,n\alpha l\}$,

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

Note that this forms a martingale sequence where $X_0 = \mathbb{E}[X(\mathcal{G})]$ and $X_{n\alpha l} = X(\mathcal{G})$. For every $i \in \{1,\dots,n\alpha l\}$, we have $|X_i - X_{i-1}| \le 1$ since every time only one connected vertex on the right side of $\mathcal{G}$ is revealed, so the number of neighbors of the chosen set $\mathcal{V}$ cannot change by more than 1 at every single time. Hence, from the one-sided Azuma–Hoeffding inequality in Section 2.4.2,

$$\mathbb{P}\Big(\mathbb{E}[X(\mathcal{G})] - X(\mathcal{G}) \ge \lambda\sqrt{l\alpha n}\Big) \le \exp\left(-\frac{\lambda^2}{2}\right), \quad \forall\,\lambda > 0. \tag{2.7.3}$$

Since there are $\binom{n}{n\alpha}$ choices for the set $\mathcal{V}$, the event that there exists a set of size $n\alpha$ with less than $\mathbb{E}[X(\mathcal{G})] - \lambda\sqrt{l\alpha n}$ neighbors occurs

with probability at most $\binom{n}{n\alpha}\exp(-\frac{\lambda^2}{2})$, by the union bound. Based on the inequality $\binom{n}{n\alpha} \leq e^{nh(\alpha)}$, we get the exponential upper bound $\exp(nh(\alpha) - \frac{\lambda^2}{2})$. Finally, choosing $\lambda = \sqrt{2n(h(\alpha)+\delta)}$ in (2.7.3) gives the bound in (2.7.1). □

2.7.3 Concentration of the crest factor for OFDM signals

Orthogonal-frequency-division-multiplexing (OFDM) is a widely used modulation scheme that converts a high-rate data stream into a large number of closely spaced orthogonal sub-carrier signals. These sub-carriers are used to transmit data steams over parallel narrow-band channels. OFDM signals are used in various international standards for digital television and audio broadcasting, DSL internet access, wireless networks, and the fourth generation (4G) mobile communications. For a textbook treatment of OFDM, the reader is referred to, e.g., [107, Chapter 19].

The primary advantage of OFDM signals over single-carrier modulation schemes is in their immunity to severe channel conditions (e.g., attenuation of high frequencies in a long copper wire, narrowband interference and frequency-selective fading due to multipath propagation) without using complex equalization filters. This important advantage arises from the fact that channel equalization is significantly simplified due to the fact that the OFDM modulation scheme can be viewed as using many slowly-varying modulated narrowband signals rather than one rapidly-varying modulated wideband signal. Nevertheless, one of the significant problems of OFDM signals is that the peak amplitude of such a signal is typically much larger than its average amplitude. The high peak-to-average power ratio (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, linear transmitter circuitry is required for OFDM signals, which suffers from a poor power efficiency. For a recent comprehensive tutorial that considers this long-lasting problem of the high PAPR, and some related issues, the reader is referred to [108].

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

symbol is described by

$$s(t) = \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} X_i \exp\Big(\frac{j\,2\pi i t}{T}\Big), \quad 0 \le t \le T. \tag{2.7.4}$$

Let us assume that $X_0, \ldots, X_{n-1}$ are complex random variables, and $|X_i| = 1$ a.s. (for the moment, these random variables may be dependent; however, later in this section, some concentration inequalities are derived for the case where these random variables are independent). Since the sub-carriers are orthonormal over $[0, T]$, the signal power over the interval $[0, T]$ is 1 a.s.:

$$\frac{1}{T} \int_0^T |s(t)|^2 dt = 1. \tag{2.7.5}$$

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

$$\mathrm{CF}_n(s) \triangleq \max_{0 \le t \le T} |s(t)|. \tag{2.7.6}$$

Commonly, the impact of nonlinearities is described by the distribution of the CF of the transmitted signal [109], but its calculation involves time-consuming simulations even for a small number of sub-carriers. From [110, Section 4] and [111], it follows that the CF scales with high probability like $\sqrt{\ln n}$ for large n. In [109, Theorem 3 and Corollary 5], a concentration inequality was derived for the CF of OFDM signals. It states that, for an arbitrary $c \ge 2.5$,

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

Remark 2.10. The analysis used to derive this rather strong concentration inequality (see [109, Appendix C]) requires some assumptions on the distribution of the X_i's (see the two conditions in [109, Theorem 3] followed by [109, 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 [109, Theorem 3].

In the following, the concentration of the crest factor of OFDM signals is studied via the Azuma–Hoeffding inequality, its refinement in Proposition 2.1, and McDiarmid's inequality. It is assumed in the following that the symbols $\{X_j\}_{j=0}^{n-1}$ are independent complex-valued random variables with magnitude 1, attaining the M points of an M-ary PSK constellation with equal probability. The material in this section presents in part the work in [112].

Concentration via the Azuma–Hoeffding inequality: Let us define the random variables

$$Y_i = \mathbb{E}[\,\mathrm{CF}_n(s)\,|\,X_0,\ldots,X_{i-1}], \quad i = 0,\ldots,n. \tag{2.7.7}$$

Based on a standard construction of martingales, $\{Y_i, \mathscr{F}_i\}_{i=0}^n$ is a martingale, where $\mathscr{F}_i$ is the σ-algebra generated by the first i symbols $(X_0,\ldots,X_{i-1})$ in (2.7.4). Hence, $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \ldots \subseteq \mathscr{F}_n$ is a filtration. This martingale also has bounded differences:

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

since revealing the additional i-th coordinate X_i affects the CF, as defined in (2.7.6), by at most $\frac{2}{\sqrt{n}}$ (see the first part of Appendix 2.C). It therefore follows from the Azuma–Hoeffding inequality that, for every $\alpha > 0$,

$$\mathbb{P}(|\mathrm{CF}_n(s) - \mathbb{E}[\mathrm{CF}_n(s)]| \geq \alpha) \leq 2\exp\left(-\frac{\alpha^2}{8}\right), \tag{2.7.8}$$

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 (2.7.7), Appendix 2.C gives that a.s.

$$|Y_i - Y_{i-1}| \leq \frac{2}{\sqrt{n}}\,, \quad \mathbb{E}\big[(Y_i - Y_{i-1})^2 | \mathscr{F}_{i-1}\big] \leq \frac{2}{n} \tag{2.7.9}$$

for every $i \in \{1,\ldots,n\}$. Note that the conditioning on the σ-algebra $\mathscr{F}_{i-1}$ 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 $\alpha > 0$,

$$\begin{aligned}
&\mathbb{P}(|\mathrm{CF}_n(s) - \mathbb{E}[\mathrm{CF}_n(s)]| \geq \alpha) \\
&= \mathbb{P}(|Y_n - Y_0| \geq \alpha) \\
&= \mathbb{P}(|Z_n - Z_0| \geq \alpha\sqrt{n}) \\
&\leq 2 \exp\left(-\frac{\alpha^2}{4}\left(1 + O\Big(\frac{1}{\sqrt{n}}\Big)\right)\right) \qquad (2.7.10)
\end{aligned}$$

(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 (2.7.8) via the Azuma–Hoeffding inequality, and the term that scales like $O\Big(\frac{1}{\sqrt{n}}\Big)$ on the right side of (2.7.10) is expressed explicitly for finite n (see the proof of Proposition 2.1).

Establishing concentration via McDiarmid's inequality: We use in the following McDiarmid's inequality (see Theorem 2.4.3) in order to prove a concentration inequality for the crest factor of OFDM signals. To this end, let us define

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

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

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

where the last inequality holds since $|X_{i-1}| = |X'_{i-1}| = 1$. Hence, Mc-

Diarmid's inequality in Theorem 2.4.3 implies that, for every $\alpha \geq 0$,

$$\mathbb{P}(|\mathrm{CF}_n(s) - \mathbb{E}[\mathrm{CF}_n(s)]| \geq \alpha) \leq 2\exp\Big(-\frac{\alpha^2}{2}\Big) \tag{2.7.11}$$

which demonstrates concentration of the CF around its expected value. The improvement of McDiarmid's inequality is by a factor of 2 in comparison to the refined version of the Azuma–Hoeffding inequality in Proposition 2.1. As will be seen in Chapter 3, there are some deep connections between McDiarmid's inequality and information-theoretic aspects; McDiarmid's inequality will be proved in Chapter 3 by the use of the entropy method and information-theoretic tools, and it will be proved useful in information-theoretic problems.

To conclude, three concentration inequalities for the crest factor (CF) of OFDM signals have been derived in this section under the assumption that the symbols are independent. The first two concentration inequalities rely on the Azuma–Hoeffding inequality and its refinement in Proposition 2.1, whereas the third bound is based on McDiarmid's inequality. Although these concentration results are weaker than some existing results in the literature (see [109] and [111]), they establish concentration in a rather simple way and provide some additional insight to the problem. McDiarmid's inequality improves the exponent of the Azuma–Hoeffding inequality by a factor of 4, and the exponent of the refined version of the Azuma–Hoeffding 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 (this happens to be the case if $\gamma < \frac{1}{4}$ in the setting of Proposition 2.1).

2.7.4 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 [113]. 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 distributions 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 [114]. 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 preliminary material, the reader is referred to Sections II-A and II-E of [115]. In [115] and [116], the following question is addressed:

> Consider a sequence of LDPC code ensembles with block lengths tending to infinity, and let the communication take place over a memoryless binary-input output-symmetric channel. Assume that this sequence achieves a rate which is equal to a fraction $1 - \varepsilon$ of the channel capacity with a block error probability that tends to zero. Then, how small can the number of loops in their bipartite graphs be as a function of ε (i.e., the fractional gap in rate to capacity)?

An information-theoretic lower bound on the average cardinality of the fundamental system of cycles was derived in [115, 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 [114] 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 provide 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 every 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 $|E(\mathcal{G})| = m a_{\mathrm{R}}$. Therefore, for a code from the (n, λ, ρ) LDPC code ensemble, the cardinality of the fundamental system of cycles satisfies the equality

$$\beta(\mathcal{G}) = n\big[(1 - R_{\mathrm{d}}) a_{\mathrm{R}} - (2 - R_{\mathrm{d}})\big] + c(\mathcal{G}) \tag{2.7.12}$$

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

$$E \triangleq |E(\mathcal{G})| = n(1 - R_{\mathrm{d}})a_{\mathrm{R}} \tag{2.7.13}$$

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 random variable 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 $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \ldots \subseteq \mathscr{F}_E$ in this case is defined so that $\mathscr{F}_i$ 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 2.5.3, it follows that for every $\alpha \geq 0$

$$\begin{aligned} &\mathbb{P}\left(|c(\mathcal{G}) - \mathbb{E}_{\mathrm{LDPC}(n,\lambda,\rho)}[c(\mathcal{G})]| \geq \alpha E\right) \leq 2e^{-f(\alpha)E} \\ \Rightarrow\ &\mathbb{P}\left(|\beta(\mathcal{G}) - \mathbb{E}_{\mathrm{LDPC}(n,\lambda,\rho)}[\beta(\mathcal{G})]| \geq \alpha E\right) \leq 2e^{-f(\alpha)E} \end{aligned} \tag{2.7.14}$$

where the implication is a consequence of (2.7.12), and the function f was defined in (2.5.20). Hence, for $\alpha > 1$, this probability is zero (since $f(\alpha) = +\infty$ for $\alpha > 1$). Note that, from (2.7.12), $\mathbb{E}_{\mathrm{LDPC}(n,\lambda,\rho)}[\beta(\mathcal{G})]$ scales linearly with n. The combination of Eqs. (2.5.20), (2.7.13), (2.7.14) gives the following statement:

Theorem 2.7.3. 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:

$$\mathbb{P}\left(|\beta(\mathcal{G}) - \mathbb{E}_{\mathrm{LDPC}(n,\lambda,\rho)}[\beta(\mathcal{G})]| \geq \alpha n\right) \leq 2 \cdot 2^{-\left[1 - h_2\left(\frac{1-\eta}{2}\right)\right]\frac{\alpha n}{\eta}} \tag{2.7.15}$$

where h_2 is 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 2.11. We can obtain the following weakened version of (2.7.15) from the Azuma–Hoeffding inequality: for every $\alpha \geq 0$,

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

where η is defined in Theorem 2.7.3 (note that $\frac{\alpha}{\eta} = \frac{E}{n}$ is equal to the average degree of the variable nodes). The exponential decay of the last two bounds is similar for values of α close to zero (see the exponents of the Azuma–Hoeffding inequality and Corollary 2.5.3 in Figure 2.2).

Remark 2.12. 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., $R_{\mathrm{d}} = (1-\varepsilon)C$) [40]. Therefore, for small values of α, the exponential decay rate in the inequality of Theorem 2.7.3 scales like $\left(\log \frac{1}{\varepsilon}\right)^{-2}$. This large-deviations result complements the result in [115, 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 2.13. Consider small deviations from the expected value that scale like $\sqrt{n}$. Note that Corollary 2.5.3 is a special case of Theorem 2.5.2 when $\gamma = 1$ (i.e., when only an upper bound on the differences 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 2.5.3 does not provide any improvement in the exponent of the concentration inequality (in comparison to the Azuma–Hoeffding inequality) when small deviations are considered.

2.7.5 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

[41] and [117] 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 tends to infinity, and that this average performance converges asymptotically to the performance in the cycle-free case (when the bipartite graph that represents a linear code contains no cycles, the messages that are delivered by the message-passing decoder through the edges of the graph are statistically independent [13]). These concentration results were later generalized in [118] for intersymbol-interference (ISI) channels. The proofs of [118, 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 in comparison to [41] and [117]. The presentation in the following is based on [119].

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 [118]. 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 $\{N_j\}$ is a sequence of i.i.d. Gaussian random variables with zero mean and variance σ^2. 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 [118] and [120];

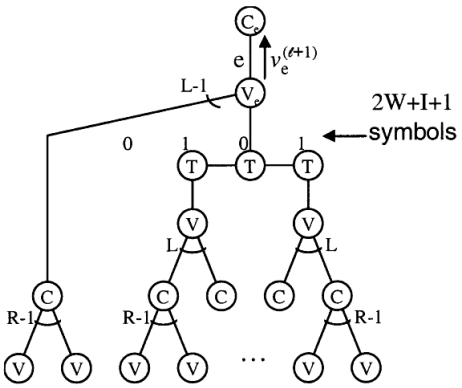

Figure 2.3: Message flow neighborhood of depth 1. This figure corresponds to the parameters $(I, W, d_{\mathrm{v}} = L, d_{\mathrm{c}} = R) = (1, 1, 2, 3)$.

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 message that is received from the channel at a variable node is not only a function of the channel output that corresponds to the considered symbol, but it is also a function of the $2W$ neighboring channel outputs and $2W$ neighboring variables nodes (as is illustrated in Fig. 2.3).

Concentration

We 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 neighborhood 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 $N_{\mathrm{e}}^{(\ell)}$ 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 (explicit expressions are given in Appendix 2.D).

Theorem 2.7.4. Let $P_{\overline{\mathrm{t}}}^{(\ell)} \equiv \Pr\left\{\mathcal{N}_{\vec{e}}^{(\ell)} \text{ not a tree}\right\}$ denote the probability that the sub-graph $\mathcal{N}_{\vec{e}}^{(\ell)}$ is not a tree (i.e., it contains cycles). Then, there exists a positive constant $\gamma \triangleq \gamma(d_{\mathrm{v}}, d_{\mathrm{c}}, \ell)$ that does not depend on the block-length n, such that $P_{\overline{\mathrm{t}}}^{(\ell)} \leq \frac{\gamma}{n}$. 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 [41] (for binary-input output-symmetric memoryless channels) to binary-input ISI channels. A detailed proof is available in [119]. □

The following concentration inequalities follow from Theorem 2.7.4 and the Azuma–Hoeffding inequality:

Theorem 2.7.5. 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 $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{Z^{(\ell)}(\underline{s})}{n d_{\mathrm{v}}} - \frac{\mathbb{E}[Z^{(\ell)}(\underline{s})]}{n d_{\mathrm{v}}}\right| > \varepsilon/2\right) \leq 2e^{-\beta\varepsilon^2 n}. \tag{2.7.16}$$

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

$$\left| \frac{\mathbb{E}[Z^{(\ell)}(\underline{s})]}{nd_{\mathrm{v}}} - p^{(\ell)}(\underline{s}) \right| \leq \varepsilon/2. \tag{2.7.17}$$

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

$$\mathbb{P}\left(\left| \frac{Z^{(\ell)}(\underline{s})}{nd_{\mathrm{v}}} - p^{(\ell)}(\underline{s}) \right| > \varepsilon \right) \leq 2e^{-\beta\varepsilon^2 n}. \tag{2.7.18}$$

More explicitly, the above statements hold for

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

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.D. □

The concentration inequalities in Theorem 2.7.5 extend the results in [41] from the special setting of memoryless binary-input output-symmetric (MBIOS) channels to ISI channels. One can particularize the above expression for β to MBIOS channels by setting $W = 0$ and $I = 0$. Since the proof of Theorem 2.7.5 uses exact expressions for $N_{\mathrm{e}}^{(\ell)}$ and $N_Y^{(\ell)}$, one would expect a tighter bound in comparison to the value of β in [41], which is given by $\frac{1}{\beta} = 544 d_{\mathrm{v}}^{2\ell-1} d_{\mathrm{c}}^{2\ell}$. As an example, for $(d_{\mathrm{v}}, d_{\mathrm{c}}, \ell) = (3, 4, 10)$, one gets an improvement by a factor of about 1 million. However, even with this improvement, the required size of n according to the analysis in this section 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 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 Section 2.7.4):

Theorem 2.7.6. 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} (this denotes the conditional *pdf* of the log-likelihood ratio $L \triangleq l(Y) = \ln\left(\frac{p_{Y|X}(Y|1)}{p_{Y|X}(Y|-1)}\right)$, given that $X = 1$ is the transmitted symbol). Assume that the decoder performs l iterations of message-passing decoding, and let $P_{\mathrm{b}}(\mathcal{C}, a_{\mathrm{MBIOS}}, l)$ denote the resulting bit error probability. Then, for every $\delta > 0$, there exists a positive α where $\alpha = \alpha(\lambda, \rho, \delta, l)$ is *independent of the block length* n, such that the following concentration inequality holds:

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

This theorem asserts that 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. Hence, assuming a sufficiently large block length, the ensemble average is a good indicator for the performance of individual codes; it is therefore reasonable to focus on the design and analysis of capacity-approaching ensembles (via the density evolution technique [41]). This forms a fundamental result in the theory of codes on graphs and iterative decoding.

2.7.6 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 [121, Theorem 4] and [35, Theorem 1]. The following theorem is proved in [121, 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 2.7.7. 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 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 every $\xi > 0$,

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

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 2.7.7, originally given in [121, 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 random variable

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

i.e., the conditional entropy when the transmission is 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 $\{\mathscr{F}_t\}_{t\in\{0,1,\ldots,m\}}$ form a filtration of σ-algebras $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \ldots \subseteq \mathscr{F}_m$ where $\mathscr{F}_t$ (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 $\mathscr{F}_0 = \{\emptyset, \Omega\}$ 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:

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

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 random variable that is equal a.s. to the conditional entropy of the particular code from the ensemble. Similarly to [121, Appendix I], we obtain upper bounds on the differences $|Z_{t+1} - Z_t|$ and then rely on the Azuma–Hoeffding inequality in Theorem 2.4.2.

Without loss of generality, we can order the parity-check nodes by increasing degree, as done in [121, 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 [121, Appendix I],

$$|Z_{t+1} - Z_t| \leq (r+1)\, H_{\mathcal{G}}(\tilde{X}|\mathbf{Y}) \tag{2.7.19}$$

where $H_{\mathcal{G}}(\tilde{X}|\mathbf{Y})$ is a random variable that is 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 [121, 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

$$|Z_{t+1} - Z_t| \leq d_{\mathrm{c}}^{\max} + 1 \quad t = 0, 1, \ldots, m-1 \tag{2.7.20}$$

which, together with the Azuma–Hoeffding inequality, completes the proof of Theorem 2.7.7. Note that the d_i's in Theorem 2.4.2 are equal to $d_{\mathrm{c}}^{\max} + 1$, and n in Theorem 2.4.2 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).

Based on [116], a refined analysis is provided; it departs from the analysis in [121, 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 (2.7.19), 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 [121, 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 side of (2.7.19) with $d_{\mathrm{c}}^{\max}+1$, we propose to leave it as is, since the Azuma–Hoeffding inequality applies to the case when the bounded differences of the martingale sequence are not fixed (see Theorem 2.4.2), 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 2.10.

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

$$H_{\mathcal{G}}(\tilde{X}|\mathbf{Y}) \leq h_2\left(\frac{1-C^{\frac{r}{2}}}{2}\right). \tag{2.7.21}$$

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

$$H_{\mathcal{G}}(\tilde{X}|\mathbf{Y}) \leq h_2\left(\frac{1-\left[1-2h_2^{-1}(1-C)\right]^r}{2}\right) \tag{2.7.22}$$

and

$$H_{\mathcal{G}}(\tilde{X}|\mathbf{Y}) \leq 1-C^r \tag{2.7.23}$$

respectively, where h_2^{-1} in (2.7.22) 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 (2.7.21) holds with equality (where both sides of (2.7.21) are zero), whereas the trivial upper bound is 1.

Proof. Since conditioning reduces the entropy, we have

$$H(\tilde{X}|\mathbf{Y}) \leq H(\tilde{X}|Y_{i_1},\ldots,Y_{i_r}).$$

Note that $Y_{i_1}, \ldots, Y_{i_r}$ are the channel outputs that correspond 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 [115, Eq. (17) and Appendix I], we can show that

$$H(\tilde{X}|\mathbf{Y}) \leq 1 - \frac{1}{2\ln 2} \sum_{k=1}^{\infty} \frac{(g_k)^r}{k(2k-1)} \tag{2.7.24}$$

where (see [115, Eq. (19)])

$$g_k \triangleq \int_0^{\infty} a(l)(1+e^{-l}) \tanh^{2k}\left(\frac{l}{2}\right) \mathrm{d}l, \quad \forall\, k \in \mathbb{N} \tag{2.7.25}$$

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 [115, Lemmas 4 and 5], it follows that $g_k \geq C^k$ for every $k \in \mathbb{N}$. Substituting this inequality in (2.7.24) gives

$$\begin{aligned} H(\tilde{X}|\mathbf{Y}) &\leq 1 - \frac{1}{2\ln 2} \sum_{k=1}^{\infty} \frac{C^{kr}}{k(2k-1)} \\ &= h_2\left(\frac{1-C^{\frac{r}{2}}}{2}\right) \end{aligned} \tag{2.7.26}$$

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

$$h_2(x) = 1 - \frac{1}{2\ln 2} \sum_{k=1}^{\infty} \frac{(1-2x)^{2k}}{k(2k-1)}, \quad 0 \leq x \leq 1. \tag{2.7.27}$$

This proves the result in (2.7.21).

The tightened bound on the conditional entropy for the BSC is obtained from (2.7.24) 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 [115, Eq. (97)]). This replaces C on the right side of (2.7.26) with $\left(1 - 2h_2^{-1}(1-C)\right)^2$, thus leading to the tightened bound in (2.7.22).

The tightened result for the BEC follows from (2.7.24) where, from (2.7.25),

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

(see [115, Appendix II]). Substituting g_k into the right side of (2.7.24) gives (2.7.22) (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 (2.7.19), we get

$$|Z_{t+1} - Z_t| \leq (r+1)\, h_2\left(\frac{1 - C^{\frac{r}{2}}}{2}\right), \tag{2.7.28}$$

where the two improvements for the BSC and BEC are obtained by replacing the second term, $h_2(\cdot)$, on the right side of (2.7.28) by (2.7.22) and (2.7.23), respectively. This improves upon the earlier bound of $(d_{\mathrm{c}}^{\max} + 1)$ in [121, Appendix I]. From (2.7.28) and Theorem 2.4.2, we obtain the following tightened version of the concentration inequality in Theorem 2.7.7:

Theorem 2.7.8. 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 $H(\mathbf{X}|\mathbf{Y})$ designate the conditional entropy of the transmitted codeword $\mathbf{X}$ given the received sequence $\mathbf{Y}$ at the channel output. Then, for every $\xi > 0$,

$$\mathbb{P}\big(\big|H(\mathbf{X}|\mathbf{Y}) - \mathbb{E}_{\mathrm{LDPC}(n,\lambda,\rho)}[H(\mathbf{X}|\mathbf{Y})]\big| \geq \xi\sqrt{n}\,\big) \leq 2\exp(-B\xi^2), \tag{2.7.29}$$

where

$$B \triangleq \frac{1}{2(1-R_{\mathrm{d}}) \displaystyle\sum_{i=1}^{d_{\mathrm{c}}^{\max}} \left\{(i+1)^2\, \Gamma_i \left[h_2\left(\frac{1 - C^{\frac{i}{2}}}{2}\right)\right]^2\right\}}, \tag{2.7.30}$$

$d_{\mathrm{c}}^{\max}$ 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 side of (2.7.29) 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

$$B_{\mathrm{BEC}} \triangleq \frac{1}{2(1-R_{\mathrm{d}}) \sum\limits_{i=1}^{d_{\mathrm{c}}^{\max}} \left\{(i+1)^2 \, \Gamma_i \, (1-C^i)^2\right\}}. \tag{2.7.31}$$

Remark 2.14. From (2.7.30), Theorem 2.7.8 indeed yields a stronger concentration inequality than the one in Theorem 2.7.7.

Remark 2.15. In the limit where $C \to 1$ bit per channel use, it follows from (2.7.30) that, if $d_{\mathrm{c}}^{\max} < \infty$, then $B \to \infty$. This is in contrast to the value of B in Theorem 2.7.7, 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 2.7.8 is tight in this case. This is in contrast to the value of B in Theorem 2.7.7, which vanishes when $d_{\mathrm{c}}^{\max} = \infty$, making it useless in this case (see Example 2.10).

Example 2.9 (Comparison of Theorems 2.7.7 and 2.7.8 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 (note that the above calculation of p_{BP} for the BEC follows from the fixed-point characterization of the threshold in [13, Theorem 3.59] with the pair of degree distributions $\lambda(x) = x$ and $\rho(x) = x^{19}$). For the BIAWGNC, the threshold under BP decoding is equal to $\sigma_{\mathrm{BP}} = 0.4156590$ (this numerical result is based on a computation that follows from [41, Example 11]). 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 code ensemble is right-regular with $d_{\mathrm{c}} = 20$, the value of B in Theorem 2.7.8 is improved by a factor of $\left[h_2\left(\frac{1-C^{\frac{d_{\mathrm{c}}}{2}}}{2}\right)\right]^{-2} = 5.134$. For the BEC, the result is improved by a factor of $\left(1-C^{d_{\mathrm{c}}}\right)^{-2} = 9.051$; this follows from the tightened value of B in (2.7.31), which improves the concentration inequality in Theorem 2.7.7.

Example 2.10 (Comparison of Theorems 2.7.7 and 2.7.8 for a heavy-tail Poisson distribution (Tornado codes))**.** The capacity-achieving sequence of the so-called Tornado codes for the BEC was introduced in [40, Section IV], [122] (see also [13, Problem 3.20]).

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 2.7.7, which renders this theorem useless for the considered code ensemble. On the other hand, from Theorem 2.7.8,

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

with the following justification:

- 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 a BEC with an erasure probability p, which states that satisfying the inequality $p\lambda'(0)\rho'(1) < 1$ is a necessary condition for reliable communication under BP decoding (see [13, Theorem 3.65]);
- equality (f) follows from the analysis in [115, Appendix VI] (an upper bound on λ_2 is derived in [115, Eq. (120)], and the average parity-check node degree scales like $\log \frac{1}{\varepsilon}$).

It therefore follows from the above chain of inequalities and (2.7.30) that, for a small gap to capacity, the parameter B in Theorem 2.7.8 scales (at least) like

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

Theorem 2.7.8 is therefore useful for the analysis of this LDPC code ensemble. As is shown above, the parameter B in (2.7.30) tends to zero rather slowly as we let the fractional gap ε tend to zero (which therefore demonstrates a rather fast concentration in Theorem 2.7.8).

Example 2.11. Here, we continue with the setting of Example 2.9 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 2.7.7 gives

$$\begin{aligned}&\mathbb{P}\big(\big|H(\mathbf{X}|\mathbf{Y}) - \mathbb{E}_{\mathrm{LDPC}(n,\lambda,\rho)}[H(\mathbf{X}|\mathbf{Y})]\big| \geq \xi\sqrt{n}\,\big)\\ &\leq 2\exp(-0.0113\,\xi^2), \quad \forall\,\xi > 0. \end{aligned} \tag{2.7.32}$$

As was mentioned already in Example 2.9, the exponential inequalities in Theorem 2.7.8 achieve an improvement in the exponent of Theorem 2.7.7 by factors of 5.134 and 9.051 for the BIAWGNC and BEC, respectively. One therefore obtains from the concentration inequalities in Theorem 2.7.8 that, for every $\xi > 0$,

$$\begin{aligned}&\mathbb{P}\big(\big|H(\mathbf{X}|\mathbf{Y}) - \mathbb{E}_{\mathrm{LDPC}(n,\lambda,\rho)}[H(\mathbf{X}|\mathbf{Y})]\big| \geq \xi\sqrt{n}\,\big)\\ &\leq \begin{cases} 2\exp(-0.0580\,\xi^2), & \text{(BIAWGNC)}\\ 2\exp(-0.1023\,\xi^2), & \text{(BEC)} \end{cases}. \end{aligned} \tag{2.7.33}$$

2.8 Summary

This chapter introduces several classical concentration inequalities for discrete-time martingales with bounded differences, and some of their applications in information theory, communications and coding.

The exposition starts with the martingale decomposition of Doob, the Chernoff bound, and the Hoeffding Lemma (see Section 2.4); these form basic ingredients for the derivation of concentration inequalities via the martingale approach. This chapter derives the Azuma–Hoeffding inequality for discrete-time martingales with bounded differences ([8], [9]), and some of its refined versions (see Sections 2.4.2 and 2.5). The martingale approach also serves as a useful tool for establishing concentration of a function $f\colon \mathbb{R}^n \to \mathbb{R}$ whose value changes by a bounded amount whenever any of its n input variables is changed arbitrarily while the other variables are held fixed. A common method for proving concentration of such a function of n independent random variables around its expected value $\mathbb{E}[f]$ revolves around McDiarmid's inequality or the "independent bounded-differences inequality" [6]. McDiarmid's inequality was originally proved via the martingale approach,

as it is derived in Section 2.4.3. 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.

The presentation of the concentration inequalities in this chapter is followed by a short discussion on their relations to some selected classical results in probability theory (see Section 2.6); these include the central limit theorem for discrete-time martingales, the moderate deviations principle, and the suitability of the concentration inequalities derived in this chapter for harmonic and bounded functions of discrete-time Markov chains.

Section 2.7 is focused on the applications of the concentration inequalities in information theory, communication, and coding theory. These include the establishment of concentration results for the minimum distance of random binary linear codes, expansion properties of random bipartite graphs, the crest factor (or peak to average power ratio) of OFDM signals, and concentration results for LDPC code ensembles. Additional concentration results have been established by Richardson and Urbanke for LDPC code ensembles under MAP and iterative message-passing decoding [41]. These martingale inequalities also prove to be useful for the derivation of achievable rates and random coding error exponents, under ML decoding, when transmission takes place over linear or nonlinear additive white Gaussian noise channels with or without memory ([123]–[124]). Nice and interesting applications of these concentration inequalities to discrete mathematics and random graphs were provided, e.g., in [6, Section 3], [10, Chapter 7] and [18, Chapters 1 and 2].

A recent interesting avenue that follows from the inequalities that are introduced in this chapter is their generalization to random matrices (see, e.g., [15] and [16]). The interested reader is also referred to [125] for a derivation of concentration inequalities that refer to martingales whose differences are not necessarily bounded, followed by some applications to graph theory.

2.A Proof of Bennett's inequality

The inequality in (2.5.10) is trivial for $\lambda = 0$, so we 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 such 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); this minimal value is zero. Furthermore, f is concave on $[y_0, \infty)$ (since its second derivative is non-positive on this interval) and it attains its maximal value on this interval at $y = y_1$. By construction, $f(b_Y) = 0$; this implies that $f \geq 0$ on the interval $(-\infty, b_Y]$, so $\mathbb{E}[f(Y)] \geq 0$ for an arbitrary 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 $\phi''(y) - e^y = f''(y) \geq 0$ for $y < y_0$, so $\phi''(0) = \phi''(y) > 0$ (recall that ϕ'' is constant since ϕ is a parabola). Hence, for every random variable Y of zero mean, $\mathbb{E}[\phi(Y)]$, 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 in (2.5.10).

2.B On the moderate deviations principle

Here we show that, in contrast to the Azuma–Hoeffding inequality, Theorem 2.5.2 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 (2.6.5) 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 $\{S_k, \mathscr{F}_k\}_{k=0}^n$ where

$$S_k \triangleq \sum_{i=1}^{k} X_i, \quad \mathscr{F}_k \triangleq \sigma(X_1, \ldots, X_k)$$

for every $k \in \{1, \ldots, n\}$ with $S_0 = 0$ and $\mathscr{F}_0 = \{\emptyset, \mathscr{F}\}$. This martingale sequence has uniformly bounded differences: $|S_k - S_{k-1}| = |X_k| \leq d$ a.s. for every $k \in \{1, \ldots, n\}$. Hence, it follows from the Azuma–Hoeffding 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

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

This differs from the limit in (2.6.5) where σ^2 is replaced by d^2, so the Azuma–Hoeffding inequality does not provide the asymptotic limit in (2.6.5) (unless $\sigma^2 = d^2$, i.e., $|X_k| = d$ a.s. for every k).

An analysis that follows from Theorem 2.5.2: The following analysis 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 2.5.2 that, for every $\alpha \geq 0$,

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

where γ is introduced in (2.5.8), $H(p\|q)$ is the divergence in (2.5.9) between the Bernoulli(p) and Bernoulli(q) probability measures, and δ_n in (2.5.27) is replaced with

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

due to the definition of δ in (2.5.8). 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 (2.5.8), $\frac{\delta^2}{\gamma} = \frac{\alpha^2}{\sigma^2}$)

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

Hence, this upper bound coincides with the exact asymptotic result in (2.6.5).

2.C Proof of (2.7.9) for OFDM signals

Consider an OFDM signal from Section 2.7.3. The sequence in (2.7.7) is a martingale. From (2.7.6), for every $i \in \{0, \ldots, n\}$,

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

The conditional expectation for the random variable 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 $i \in \{1, \ldots, n\}$

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

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 (2.7.4)

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

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 $\mathsf{var}(Y_i \,|\, \mathscr{F}_{i-1}) = \mathbb{E}\big[(Y_i - Y_{i-1})^2 \,|\, \mathscr{F}_{i-1}\big]$. Since $\big(\mathbb{E}(Z)\big)^2 \le \mathbb{E}(Z^2)$ for a real-valued random variable Z, from (2.C.1), (2.C.2) and the tower property for conditional expectations, it follows that

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

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

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

The last equality holds since

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

2.D Proof of Theorem 2.7.5

From the triangle inequality, we have

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

If inequality (2.7.17) 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.D.1), we deduce that (2.7.18) follows from (2.7.16) and (2.7.17) for any $\varepsilon > 0$ and $n > \frac{2\gamma}{\varepsilon}$. We start by proving (2.7.16).

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 ℓ-th 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 random variable $\widetilde{Z}_i \triangleq \mathbb{E}[Z^{(\ell)}(\underline{s})|a_1, \ldots, a_i]$ be defined as the conditional expectation of $Z^{(\ell)}(\underline{s})$ given the first i elements of the sequence $\underline{a}$. Note that it forms a martingale sequence (see 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 $|\widetilde{Z}_{i+1} - \widetilde{Z}_i|$ enables to apply the Azuma–Hoeffding inequality for proving concentration around the expected value $\widetilde{Z}_0$. To this end, let us 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 of

$$\alpha \triangleq (d_{\mathrm{v}} - 1 + 2Wd_{\mathrm{v}})(d_{\mathrm{c}} - 1),$$

there are a total of

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

edges related to the code structure (variable-to-check node edges or vice versa) in the neighborhood $\mathcal{N}_{\vec{e}}^{(\ell)}$. By symmetry, the two edges can affect at most $2N_{\mathrm{e}}^{(\ell)}$ neighbors (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 $|\widetilde{Z}_{i+1} - \widetilde{Z}_i| \leq 2N_{\mathrm{e}}^{(\ell)}$ 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

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

and a comparison of this concentration inequality with (2.7.16) gives that

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

Next, proving inequality (2.7.17) relies on concepts from [41] and [118]. Let $\mathbb{E}[Z_i^{(\ell)}(\underline{s})]$, for $i \in \{1, \ldots, nd_{\mathrm{v}}\}$, 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

$$\mathbb{E}[Z^{(\ell)}(\underline{s})] = \sum_{i=1}^{nd_{\mathrm{v}}} \mathbb{E}[Z_i^{(\ell)}(\underline{s})] = nd_{\mathrm{v}}\mathbb{E}[Z_1^{(\ell)}(\underline{s})], \tag{2.D.3}$$

and

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

where $P_{\mathrm{t}}^{(\ell)}$ and $P_{\bar{\mathrm{t}}}^{(\ell)} \triangleq 1 - P_{\mathrm{t}}^{(\ell)}$ denote the probabilities that the subgraph $\mathcal{N}_{\vec{e}}^{(\ell)}$ is or, respectively, is not a tree. From Theorem 2.7.4, we have $P_{\bar{\mathrm{t}}}^{(\ell)} \leq \frac{\gamma}{n}$, where γ is a positive constant which is independent of n. Furthermore, $\mathbb{E}[Z_1^{(\ell)}(\underline{s}) \,|\, \text{neighborhood is a tree}] = p^{(\ell)}(\underline{s})$, so

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

Using (2.D.3), (2.D.4) and the inequality $P_{\bar{\mathrm{t}}}^{(\ell)} \leq \frac{\gamma}{n}$ gives that

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

Hence, if $n > \frac{2\gamma}{\varepsilon}$, then (2.7.17) holds.

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 it exhibits its fundamental connections to information theory. The chapter is structured as follows. Sections 3.1–3.3 introduce the basic ingredients of the entropy method, and the closely related logarithmic Sobolev inequalities. This underlies the so-called functional approach for the derivation of concentration inequalities. Section 3.4 is devoted to a related viewpoint which is 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 briefly summarizes results on concentration inequalities for functions of 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 independent random variables, each taking values in a set $\mathcal{X}$. Given a function $f\colon \mathcal{X}^n \to \mathbb{R}$, we wish to find tight upper bounds on the *deviation probabilities* for the random variable $U = f(X^n)$, i.e., we are interested to bound from above the probability $\mathbb{P}(|U - \mathbb{E}U| \geq r)$ for $r > 0$. If U has finite variance, then Chebyshev's inequality 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 and Hoeffding (Theorem 2.4.2) and McDiarmid (Theorem 2.4.3). 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 it is specified otherwise) that the following conditions are satisfied by the function $f\colon \mathcal{X}^n \to \mathbb{R}$ and the probability distribution P of X^n:

- $U = f(X^n)$ 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 bound** — 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* $\Lambda\colon \mathbb{R} \to \mathbb{R}$ defined by

$$\Lambda(\lambda) = \ln \mathbb{E}[\exp(\lambda U)], \quad \lambda \in \mathbb{R}. \tag{3.1.4}$$

(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)

$$D(P^{(\lambda f)}\|P) = \mathbb{E}_{P^{(\lambda f)}}\left[\ln \frac{\mathrm{d}P^{(\lambda f)}}{\mathrm{d}P}\right] \tag{3.1.5}$$

$$= \mathbb{E}_P\left[\frac{\mathrm{d}P^{(\lambda f)}}{\mathrm{d}P}\,\ln \frac{\mathrm{d}P^{(\lambda f)}}{\mathrm{d}P}\right], \tag{3.1.6}$$

where $P = P_{X^n}$ is the probability distribution of X^n, and $P^{(\lambda f)}$ is the *tilted probability distribution* defined by

$$\frac{\mathrm{d}P^{(\lambda f)}}{\mathrm{d}P} = \frac{\exp(\lambda f)}{\mathbb{E}[\exp(\lambda f)]} = \exp\big(\lambda f - \Lambda(\lambda)\big). \tag{3.1.7}$$

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

$$D(P^{(\lambda f)}\|P) \leq \frac{c\lambda^2}{2} \tag{3.1.8}$$

for some $c > 0$, then the Gaussian-type bound (3.1.2) holds with $K = 2$ and $\kappa = \frac{1}{2c}$. The standard way to establish (3.1.8) is through the so-called *logarithmic Sobolev inequalities*.

3. **Tensorization of the entropy** — it is in general difficult to derive a bound like (3.1.8) directly. Instead, one typically takes a divide-and-conquer approach: relying on 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

$$D\Big(P^{(\lambda f)}_{X_i|\overline{X}^i}\|P_{X_i}|P^{(\lambda f)}_{\overline{X}^i}\Big), \qquad i=1,\ldots,n \tag{3.1.9}$$

where, for each i, $\overline{X}^i \in \mathcal{X}^{n-1}$ denotes the $(n-1)$-tuple obtained from X^n by removing the i-th coordinate, i.e.,

$$\overline{X}^i = (X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n). \tag{3.1.10}$$

Despite their formidable appearance, the conditional divergences in (3.1.9) are easier to handle because, for each given realization $\overline{X}^i = \bar{x}^i$, the i-th such term involves a single-variable function $f_i(\cdot|\bar{x}^i)\colon \mathcal{X}\to\mathbb{R}$ defined by

$$f_i(x|\bar{x}^i) \triangleq f(x_1,\ldots,x_{i-1},x,x_{i+1},\ldots,x_n), \quad x\in\mathcal{X}, \tag{3.1.11}$$

and the corresponding tilted distribution $P^{(\lambda f)}_{X_i|\overline{X}^i=\bar{x}^i}$, where

$$\frac{\mathrm{d}P^{(\lambda f)}_{X_i|\overline{X}^i=\bar{x}^i}}{\mathrm{d}P_{X_i}} = \frac{\exp\big(\lambda f_i(\cdot|\bar{x}^i)\big)}{\mathbb{E}\left[\exp\big(\lambda f_i(X_i|\bar{x}^i)\big)\right]}, \qquad \forall\,\bar{x}^i\in\mathcal{X}^{n-1}. \tag{3.1.12}$$

From (3.1.7) and (3.1.12), it is observed that the conditional distribution $P^{(\lambda f)}_{X_i|\overline{X}^i=\bar{x}^i}$ 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 $c_1,\ldots,c_n>0$ for which

$$D\Big(P^{(\lambda f_i(\cdot|\bar{x}^i))}_{X_i}\|P_{X_i}\Big) \le \frac{c_i\lambda^2}{2} \tag{3.1.13}$$

holds for every $i\in\{1,\ldots,n\}$ and $\bar{x}^i\in\mathcal{X}^{n-1}$, then (3.1.8) holds with $c=\sum_{i=1}^n c_i$ (to be shown explicitly later), which in turn gives the following bound for all $r>0$:

$$\mathbb{P}\Big(\,|f(X^n)-\mathbb{E}f(X^n)|\ge r\Big) \le 2\exp\left(-\frac{r^2}{2\sum_{i=1}^n c_i}\right). \tag{3.1.14}$$

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

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

Again, one would typically use logarithmic Sobolev inequalities to verify (3.1.13). Conceptually, the tensorization step is similar to "single-letter" techniques which are common in information theory.

In the following, we elaborate on these three ingredients. Logarithimic Sobolev inequalities and their various applications to concentration of measure inequalities are described in detail in Sections 3.2 and 3.3.

3.1.1 The Chernoff bounding technique revisited

The Chernoff bound in Section 2.4.1 reduces the problem of bounding the deviation probability $\mathbb{P}(U \geq r)$ to the analysis of the logarithmic moment-generating function:

$$\mathbb{P}(U \geq r) \leq \exp\Big(\Lambda(\lambda) - \lambda r\Big), \qquad \forall\, \lambda > 0. \tag{3.1.15}$$

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

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

$$\Lambda'(\lambda) = \frac{\mathbb{E}[U \exp(\lambda U)]}{\mathbb{E}[\exp(\lambda U)]}, \tag{3.1.16}$$

$$\Lambda''(\lambda) = \frac{\mathbb{E}[U^2 \exp(\lambda U)]}{\mathbb{E}[\exp(\lambda U)]} - \left(\frac{\mathbb{E}[U \exp(\lambda U)]}{\mathbb{E}[\exp(\lambda U)]}\right)^2. \tag{3.1.17}$$

  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

$$\lim_{\lambda \to 0} \frac{\Lambda(\lambda)}{\lambda} = 0. \tag{3.1.18}$$

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 [126]).

Given a function $g\colon \mathcal{X}^n \to \mathbb{R}$ such that $\mathbb{E}[\exp(g(X^n))] < \infty$ with $X^n \sim P$, 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 $t \in \mathbb{R}$, then

$$\begin{aligned} D(P^{(tf)}\|P) &= \frac{t\,\mathbb{E}[f\exp(tf)]}{\mathbb{E}[\exp(tf)]} - \ln \mathbb{E}[\exp(tf)] \\ &= t\Lambda'(t) - \Lambda(t) \\ &= t^2 \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\Lambda(t)}{t}\right), \end{aligned} \tag{3.1.19}$$

where in the second line we have used (3.1.16). Integrating from $t = 0$ to $t = \lambda$ and using (3.1.18), we get

$$\Lambda(\lambda) = \lambda \int_0^\lambda \frac{D(P^{(tf)}\|P)}{t^2}\, \mathrm{d}t. \tag{3.1.20}$$

Combining (3.1.15) and (3.1.20), 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 every $r \geq 0$,

$$\mathbb{P}(U \geq r) \leq \exp\left(\lambda \int_0^\lambda \frac{D(P^{(tf)}\|P)}{t^2}\,\mathrm{d}t - \lambda r\right), \qquad \forall\,\lambda > 0. \quad (3.1.21)$$

Consequently, the problem of bounding the deviation probabilities $\mathbb{P}(U \geq r)$ is reduced to the problem of bounding the relative entropies $D(P^{(tf)}\|P)$. In particular, we have

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

$$D(P^{(tf)}\|P) \leq \frac{ct^2}{2}, \qquad \forall\,t > 0 \quad (3.1.22)$$

for some constant $c > 0$, then

$$\mathbb{P}(U \geq r) \leq \exp\left(-\frac{r^2}{2c}\right), \quad \forall\,r \geq 0. \quad (3.1.23)$$

Proof. Using (3.1.22) to upper-bound the integrand on the right side of (3.1.21), we get

$$\mathbb{P}(U \geq r) \leq \exp\left(\frac{c\lambda^2}{2} - \lambda r\right), \qquad \forall\,\lambda > 0. \quad (3.1.24)$$

Optimizing over $\lambda > 0$ to get the tightest bound gives $\lambda = \frac{r}{c}$, and its substitution in (3.1.24) gives the bound in (3.1.23). □

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 divergence directly is difficult in general. This is where the third ingredient of the entropy method, the so-called *tensorization step*, enters to the picture. 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 probability measures on the product space $\mathcal{X}^n$ where P is a product measure. For every $i \in \{1, \ldots, n\}$, let $\overline{X}^i$ denote the $(n-1)$-tuple given in (3.1.10). Then

$$D(Q\|P) \leq \sum_{i=1}^{n} D(Q_{X_i|\overline{X}^i}\|P_{X_i}|Q_{\overline{X}^i}). \tag{3.1.25}$$

Proof.

$$D(Q\|P) = \sum_{i=1}^{n} D(Q_{X_i\,|\,X^{i-1}} \,\|\, P_{X_i|X^{i-1}} \,|\, Q_{X^{i-1}}) \tag{3.1.26}$$

$$= \sum_{i=1}^{n} D(Q_{X_i\,|\,X^{i-1}} \,\|\, P_{X_i} \,|\, Q_{X^{i-1}}) \tag{3.1.27}$$

where (3.1.26) is due to the chain rule for the relative entropy; (3.1.27) holds since $X_1, X_2, \ldots, X_n$ are independent random variables under P, implying that $P_{X_i|X^{i-1}} = P_{X_i|\overline{X}^i} = P_{X_i}$). Moreover, for $i \in \{1, \ldots, n\}$,

$$\begin{aligned} &D(Q_{X_i|\overline{X}^i}\|P_{X_i}|Q_{\overline{X}^i}) - D(Q_{X_i|X^{i-1}}\|P_{X_i}|Q_{X^{i-1}}) \\ &\quad = \mathbb{E}_Q\left[\ln \frac{\mathrm{d}Q_{X_i|\overline{X}^i}}{\mathrm{d}P_{X_i}}\right] - \mathbb{E}_Q\left[\ln \frac{\mathrm{d}Q_{X_i|X^{i-1}}}{\mathrm{d}P_{X_i}}\right] \\ &\quad = \mathbb{E}_Q\left[\ln \frac{\mathrm{d}Q_{X_i|\overline{X}^i}}{\mathrm{d}Q_{X_i|X^{i-1}}}\right] \\ &\quad = D(Q_{X_i|\overline{X}^i}\|Q_{X_i|X^{i-1}}|Q_{\overline{X}^i}) \geq 0. \end{aligned} \tag{3.1.28}$$

Hence, by combining (3.1.26)–(3.1.28), we get (3.1.25). □

Remark 3.1. The quantity on the right side of (3.1.25) is actually the so-called *erasure divergence* $D^-(Q\|P)$ between Q and P (see [127, Definition 4]), which in the case of arbitrary Q and P is defined by

$$D^-(Q\|P) \triangleq \sum_{i=1}^{n} D(Q_{X_i|\overline{X}^i}\|P_{X_i|\overline{X}^i}|Q_{\overline{X}^i}). \tag{3.1.29}$$

Because P is assumed to be a product measure in (3.1.25), we can replace $P_{X_i|\overline{X}^i}$ 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 (3.1.29),

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

On the other hand, if $X_1 = X_2$ under 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 (3.1.21), we obtain from Proposition 3.1 the following:

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

$$\mathbb{P}(U \geq r) \leq \exp\left(\lambda \sum_{i=1}^{n} \int_0^\lambda \frac{D(P^{(tf)}_{X_i|\overline{X}^i}\|P_{X_i}|P^{(tf)}_{\overline{X}^i})}{t^2}\,\mathrm{d}t - \lambda r\right). \quad (3.1.30)$$

The conditional divergences in the integrand in (3.1.30) may look formidable, but the remarkable thing is that, for each i and a given $\overline{X}^i = \bar{x}^i$, 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 the function $f_i(\cdot|\bar{x}^i)\colon \mathcal{X} \to \mathbb{R}$ be given in (3.1.11). Then,

$$\frac{\mathrm{d}P^{(f)}_{X_i|\overline{X}^i=\bar{x}^i}}{\mathrm{d}P_{X_i}} = \frac{\exp\left(f_i(\cdot|\bar{x}^i)\right)}{\mathbb{E}\left[\exp\left(f_i(X_i|\bar{x}^i)\right)\right]}. \quad (3.1.31)$$

In other words, $P^{(f)}_{X_i|\overline{X}^i=\bar{x}^i}$ is the $f_i(\cdot|\bar{x}^i)$-tilting of P_{X_i}. This is the essence of tensorization: the n-dimensional problem of bounding $D(P^{(tf)}\|P)$ is effectively decomposed into n one-dimensional problems, where the i-th problem involves the tilting of the marginal distribution P_{X_i} by functions of the form $f_i(\cdot|\bar{x}^i)\colon \mathcal{X} \to \mathbb{R}$. In particular, we get the following:

Corollary 3.1.2. Let the function f and the probability distribution P of X^n satisfy the condition that there exist constants $c_1, \ldots, c_n > 0$ such that, for every $t > 0$,

$$D\Big(P_{X_i}^{(tf_i(\cdot|\bar{x}^i))} \| P_{X_i}\Big) \leq \frac{c_i t^2}{2}, \qquad \forall\, i \in \{1, \ldots, n\},\ \bar{x}^i \in \mathcal{X}^{n-1}. \quad (3.1.32)$$

Then

$$\mathbb{P}\Big(f(X^n) - \mathbb{E}f(X^n) \geq r\Big) \leq \exp\left(-\frac{r^2}{2\sum_{i=1}^n c_i}\right), \quad \forall\, r > 0. \quad (3.1.33)$$

Remark 3.2. Note the obvious similarity between the bound (3.1.33) and McDiarmid's inequality (2.4.33). 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 every $t > 0$

$$\begin{aligned}
&D(P^{(tf)} \| P) \\
&\leq \sum_{i=1}^n D\Big(P_{X_i|\overline{X}^i}^{(tf)} \| P_{X_i} \,|\, P_{\overline{X}^i}^{(tf)}\Big) && (3.1.34)\\
&= \sum_{i=1}^n \int_{\mathcal{X}^{n-1}} D\Big(P_{X_i|\overline{X}^i=\bar{x}^i}^{(tf)} \| P_{X_i}\Big)\, \mathrm{d}P_{\overline{X}^i}^{(tf)}(\bar{x}^i) && (3.1.35)\\
&= \sum_{i=1}^n \int_{\mathcal{X}^{n-1}} D\Big(P_{X_i}^{(tf_i(\cdot|\bar{x}^i))} \| P_{X_i}\Big)\, \mathrm{d}P_{\overline{X}^i}^{(tf)}(\bar{x}^i) && (3.1.36)\\
&\leq \tfrac{1}{2} \sum_{i=1}^n c_i\, t^2 && (3.1.37)
\end{aligned}$$

where (3.1.34) follows from the tensorization of the relative entropy, (3.1.35) holds since P is a product measure (so $P_{X_i} = P_{X_i|\overline{X}^i}$) and by the definition of the conditional relative entropy, (3.1.36) follows from (3.1.11) and (3.1.31) which implies that $P_{X_i|\overline{X}^i=\bar{x}^i}^{(tf)} = P_{X_i}^{(tf_i(\cdot|\bar{x}^i))}$, and inequality (3.1.37) holds by the assumption in (3.1.32). Finally, the inequality in (3.1.33) follows from (3.1.37) 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 (3.1.32) 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 every $a \geq 0$ 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 every $a \geq 0$ and $g \in \mathcal{A}$, we have

$$E(ag) = a^2 E(g). \tag{3.1.38}$$

We are interested in the case when there exists a constant $c > 0$, such that the inequality

$$D(\mu^{(g)} \| \mu) \leq \tfrac{1}{2}\, c\, E(g) \tag{3.1.39}$$

holds for every $g \in \mathcal{A}$. Now suppose that, for each $i \in \{1, \ldots, n\}$, inequality (3.1.39) holds with $\mu = P_{X_i}$ and some constant $c_i > 0$. Let $f\colon \mathcal{X}^n \to \mathbb{R}$ be a function such that, for every $\bar{x}^i \in \mathcal{X}^{n-1}$ 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 $f_i\colon \mathcal{X} \to \mathbb{R}$ is defined in (3.1.11). Then, the bounds in (3.1.32) hold, since from (3.1.39) and the above properties of the functional E it follows that for every $t > 0$ and $\bar{x}^i \in \mathcal{X}^{n-1}$

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

Consequently, the Gaussian concentration inequality in (3.1.33) follows from Corollary 3.1.2.

3.2 The Gaussian logarithmic Sobolev inequality

Before turning to general 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 log-Sobolev inequality (LSI) in this instance comes from a seminal paper of Gross [44], and it connects two key information-theoretic quantities, namely the relative entropy and relative Fisher information. There exist deep links between Gross's LSI and other fundamental information-theoretic inequalities, such as Stam's inequality and the entropy power inequality. Some of these links are considered in this section.

For every $n \in \mathbb{N}$ and for every symmetric and 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; it will be written G_s for $n = 1$. 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, γ^n, and γ the corresponding probability 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 every 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 [128], 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 by assuming in (3.2.1) that $\mathbb{E}[\phi^2(Z)] = 1$. Then (3.2.1) can be rewritten as

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

[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.

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

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

We provide an information-theoretic proof of the Gaussian LSI (Theorem 3.2.1) later in this section; the reader is referred to [129] as one of the typical proofs using techniques from functional analysis.

From an information-theoretic point of view, the Gaussian LSI (Theorem 3.2.1) relates two fundamental measures of (dis)similarity between probability measures: the relative entropy (a.k.a. Kullback-Leibler divergence, KL distance/divergence, information divergence), and *relative Fisher information* (a.k.a. Fisher information distance). The second information measure is defined as follows. Let P_1 and P_2 be Borel probability measures on $\mathbb{R}^n$ with differentiable densities p_1 and p_2, and suppose that the Radon–Nikodym derivative $\frac{\mathrm{d}P_1}{\mathrm{d}P_2} \equiv \frac{p_1}{p_2}$ is differentiable P_2-a.s. Then, the relative Fisher information between P_1 and P_2 is defined as (see [130, Eq. (6.4.12)])

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

whenever the integral converges. Under suitable regularity conditions, $I(P_1 \| P_2)$ admits the equivalent form (see [131, Eq. (1.108)])

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

Remark 3.5. A 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 $\sqrt{\frac{\mathrm{d}P_1}{\mathrm{d}P_2}} = \sqrt{\frac{p_1}{p_2}}$, such that the equality

$$\int_{-\infty}^{\infty} \sqrt{\frac{p_1(z)}{p_2(z)}} \partial_i \psi(z) \mathrm{d}z = -\int_{-\infty}^{\infty} \xi_i(z) \psi(z) \mathrm{d}z \tag{3.2.6}$$

holds for all $i \in \{1, \dots, n\}$ and all test functions $\psi \in C_c^\infty(\mathbb{R}^n)$ [132, Sec. 6.6]. (Here, $\partial_i \psi$ denotes the i-th coordinate of $\nabla \psi$). Then (3.2.5) holds, provided $\xi \in L^2(P_2)$.

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

$$\frac{\mathrm{d}P_Y}{\mathrm{d}P_Z} = \frac{\mathrm{d}P_Y}{\mathrm{d}G^n} = \phi^2. \tag{3.2.7}$$

Then, on one hand, we have

$$\begin{aligned} \mathbb{E}\left[\phi^2(Z) \ln \phi^2(Z)\right] &= \mathbb{E}\left[\left(\frac{\mathrm{d}P_Y}{\mathrm{d}P_Z}(Z)\right) \ln \left(\frac{\mathrm{d}P_Y}{\mathrm{d}P_Z}(Z)\right)\right] \\ &= D(P_Y \| P_Z), \end{aligned} \tag{3.2.8}$$

and on the other hand, from (3.2.5),

$$\begin{aligned} \mathbb{E}\left[\|\nabla \phi(Z)\|^2\right] &= \mathbb{E}\left[\left\|\nabla \sqrt{\frac{\mathrm{d}P_Y}{\mathrm{d}P_Z}(Z)}\right\|^2\right] \\ &= \tfrac{1}{4} \, I(P_Y \| P_Z). \end{aligned} \tag{3.2.9}$$

Substituting (3.2.8) and (3.2.9) into (3.2.2), we obtain the inequality

$$D(P_Y \| P_Z) \leq \tfrac{1}{2} \, I(P_Y \| P_Z), \qquad P_Z = G^n, \tag{3.2.10}$$

holding for every P_Y such that $P_Y \ll G^n$ and $\nabla \sqrt{\frac{\mathrm{d}P_Y}{\mathrm{d}G^n}} \in L^2(G^n)$. Conversely, for every $P_Y \ll G^n$ satisfying (3.2.10), we can derive (3.2.2) by letting $\phi = \sqrt{\frac{\mathrm{d}P_Y}{\mathrm{d}G^n}}$, provided $\nabla \phi$ exists (e.g., in the distributional sense). Similarly, for every $s > 0$, (3.2.3) can be written as

$$D(P_Y \| P_Z) \leq \tfrac{1}{2} \, s \, I(P_Y \| P_Z), \qquad P_Z = G_s^n. \tag{3.2.11}$$

We next apply the Gaussian LSI (see (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

$$\mathbb{E}\left[\exp\big(g(Z)\big) \ln \frac{\exp\big(g(Z)\big)}{\mathbb{E}[\exp\big(g(Z)\big)]}\right] \leq \tfrac{1}{2} \, \mathbb{E}\left[\|\nabla g(Z)\|^2 \exp\big(g(Z)\big)\right], \tag{3.2.12}$$

where $Z \sim G^n$. If we let $P = G^n$ and denote by $P^{(g)}$ the g-tilting of P, the left side of (3.2.12) is recognized as $\mathbb{E}[\exp(g(Z))] \cdot D(P^{(g)}\|P)$. Similarly, the right side is equal to $\mathbb{E}[\exp(g(Z))] \cdot \mathbb{E}_P^{(g)}[\|\nabla g\|^2]$ with $\mathbb{E}_P^{(g)}[\cdot]$ denoting expectation with respect to $P^{(g)}$. We therefore obtain the so-called *modified LSI* for the standard Gaussian measure:

$$D(P^{(g)}\|P) \le \tfrac{1}{2}\,\mathbb{E}_P^{(g)}\left[\|\nabla g\|^2\right], \qquad P = G^n \tag{3.2.13}$$

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 (3.2.13) implies (3.1.39) with $\mu = G^n$, $c = 1$, and $E(g) = \|\nabla g\|_\infty^2$.

In the remainder of this section, we provide an information-theoretic proof of Theorem 3.2.1, and discuss several applications of the modified LSI (3.2.13) for the derivation of Gaussian concentration inequalities via the Herbst argument.

3.2.1 An information-theoretic proof of Gross's LSI

In accordance with our general theme, we will prove Theorem 3.2.1 via tensorization: we first show that the satisfiability of the theorem for $n = 1$ implies that it holds for every $n \ge 2$ by scaling up to general n using suitable (sub)additivity properties, and then establish the $n = 1$ case. Indeed, suppose that (3.2.1) holds in dimension 1. For $n \ge 2$, let $X^n = (X_1, \ldots, X_n)$ 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 $\mathbb{E}_P[\phi^2(X^n)] = 1$, where $P = P_{X^n} = G^n$ is the product of n copies of the standard Gaussian distribution G. If we define a probability measure $Q = Q_{X^n}$ with $\mathrm{d}Q_{X^n}/\mathrm{d}P_{X^n} = \phi^2$, then using Proposition 3.2 we can write

$$\begin{aligned}
\mathbb{E}_P\left[\phi^2(X^n)\ln\phi^2(X^n)\right] &= \mathbb{E}_P\left[\frac{\mathrm{d}Q}{\mathrm{d}P}\ln\frac{\mathrm{d}Q}{\mathrm{d}P}\right] \\
&= D(Q\|P) \\
&\le \sum_{i=1}^n D(Q_{X_i|\overline{X}^i}\|P_{X_i}|Q_{\overline{X}^i}).
\end{aligned} \tag{3.2.14}$$

Following the same steps as the ones that led to (3.1.12), 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

$$\phi_i(x|\bar{x}^i) \triangleq \phi(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n), \qquad \forall\, \bar{x}^i \in \mathbb{R}^{n-1},\, x \in \mathbb{R}.$$

Then

$$\frac{\mathrm{d}Q_{X_i|\overline{X}^i=\bar{x}^i}}{\mathrm{d}P_{X_i}} = \frac{\phi_i^2(\cdot|\bar{x}^i)}{\mathbb{E}_P[\phi_i^2(X_i|\bar{x}^i)]}$$

for all $i \in \{1, \ldots, n\}$ and $\bar{x}^i \in \mathbb{R}^{n-1}$. With this, we can write

$$\begin{aligned}
&D(Q_{X_i|\overline{X}^i} \| P_{X_i} | Q_{\overline{X}^i}) \\
&\quad = \mathbb{E}_Q\left[\ln \frac{\mathrm{d}Q_{X_i|\overline{X}^i}}{\mathrm{d}P_{X_i}}\right] \\
&\quad = \mathbb{E}_P\left[\frac{\mathrm{d}Q}{\mathrm{d}P} \ln \frac{\mathrm{d}Q_{X_i|\overline{X}^i}}{\mathrm{d}P_{X_i}}\right] \\
&\quad = \mathbb{E}_P\left[\phi^2(X) \ln \frac{\phi_i^2(X_i|\overline{X}^i)}{\mathbb{E}_P[\phi_i^2(X_i|\overline{X}^i)|\overline{X}^i]}\right] \\
&\quad = \mathbb{E}_P\left[\phi_i^2(X_i|\overline{X}^i) \ln \frac{\phi_i^2(X_i|\overline{X}^i)}{\mathbb{E}_P[\phi_i^2(X_i|\overline{X}^i)|\overline{X}^i]}\right] \\
&\quad = \int_{\mathbb{R}^{n-1}} \mathbb{E}_P\left[\phi_i^2(X_i|\bar{x}^i) \ln \frac{\phi_i^2(X_i|\bar{x}^i)}{\mathbb{E}_P[\phi_i^2(X_i|\bar{x}^i)]}\right] \mathrm{d}P_{\overline{X}^i}(\bar{x}^i). \qquad (3.2.15)
\end{aligned}$$

Since each $X_i \sim G$, we can apply the Gaussian LSI (3.2.1) to the univariate functions $\phi_i(\cdot|\bar{x}^i)$ (note that we currently assume that the Gaussian LSI holds for $n = 1$) to get

$$\mathbb{E}_P\left[\phi_i^2(X_i|\bar{x}^i) \ln \frac{\phi_i^2(X_i|\bar{x}^i)}{\mathbb{E}_P[\phi_i^2(X_i|\bar{x}^i)]}\right] \le 2\,\mathbb{E}_P\left[\left(\phi_i'(X_i|\bar{x}^i)\right)^2\right] \qquad (3.2.16)$$

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

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

Since $X_1, \ldots, X_n$ are i.i.d. under P, we can express (3.2.16) as

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

where $\partial_i \phi$ denotes the i-th coordinate of the gradient $\nabla \phi$. Substituting this bound into (3.2.15), we have

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

Using this to bound each term in the sum on the right side of (3.2.14) together with the equality $\sum_{i=1}^n \left(\partial_i \phi(x^n) \right)^2 = \|\nabla \phi(x^n)\|^2$, we get

$$\mathbb{E}_P \left[\phi^2(X^n) \ln \phi^2(X^n) \right] \leq 2\, \mathbb{E}_P \left[\|\nabla \phi(X^n)\|^2 \right], \tag{3.2.17}$$

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

In view of the above argument, the Gaussian LSI is next proved for $n = 1$. To that end, it is useful to express it in an equivalent form that relates the Fisher information and the entropy power of a real-valued random variable with a sufficiently regular density. The Gaussian LSI in this form was derived by Stam [45], and the equivalence between Stam's inequality and (3.2.1) was only noted much later by Carlen [128]. We first establish this equivalence following Carlen's argument, and then give a new information-theoretic proof of Stam's inequality that, unlike existing proofs [48, 133], does not directly rely on de Bruijn's identity or on the entropy-power inequality.

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

$$h(Y) = h(p_Y) \triangleq -\int_{-\infty}^{\infty} p_Y(y) \ln p_Y(y) \mathrm{d}y, \tag{3.2.18}$$

provided the integral exists, and the *entropy power* of Y is given by

$$N(Y) \triangleq \frac{\exp(2h(Y))}{2\pi e}. \tag{3.2.19}$$

Moreover, if the density p_Y is differentiable, the *Fisher information* is given by

$$J(Y) = J(p_Y) = \int_{-\infty}^{\infty} \left(\frac{\mathrm{d}}{\mathrm{d}y} \ln p_Y(y) \right)^2 p_Y(y) \mathrm{d}y \tag{3.2.20}$$

$$= \mathbb{E}[\rho_Y^2(Y)], \tag{3.2.21}$$

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. An alternative definition of the Fisher information of a real-valued random variable Y is (see [134, Definition 4.1])

$$J(Y) \triangleq \sup\Big\{\big(\mathbb{E}\psi'(Y)\big)^2 : \psi \in C^1,\, \mathbb{E}[\psi^2(Y)] = 1\Big\} \tag{3.2.22}$$

where the supremum in the right side of (3.2.22) 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 it assumes that such a density even exists). It can be shown that the quantity defined in (3.2.22) exists and is finite if and only if Y has an absolutely continuous density p_Y, in which case $J(Y)$ is equal to (3.2.20) (see [134, Theorem 4.2]).

We will need the following facts:

1. If $D(P_Y\|G_s) < \infty$, then

$$D(P_Y\|G_s) = \frac{1}{2}\ln\frac{1}{N(Y)} + \frac{1}{2}\ln s - \frac{1}{2} + \frac{1}{2s}\,\mathbb{E}Y^2. \tag{3.2.23}$$

   This is proved by direct calculation: since $D(P_Y\|G_s) < \infty$, we have $P_Y \ll G_s$ and $\mathrm{d}P_Y/\mathrm{d}G_s = p_Y/\gamma_s$. Consequently,

$$\begin{aligned} D(P_Y\|G_s) &= \int_{-\infty}^{\infty} p_Y(y)\ln\frac{p_Y(y)}{\gamma_s(y)}\mathrm{d}y \\ &= -h(Y) + \frac{1}{2}\ln(2\pi s) + \frac{1}{2s}\,\mathbb{E}Y^2 \\ &= -\frac{1}{2}\big(2h(Y) - \ln(2\pi e)\big) + \frac{1}{2}\ln s - \frac{1}{2} + \frac{1}{2s}\,\mathbb{E}Y^2 \\ &= \frac{1}{2}\ln\tfrac{1}{N(Y)} + \frac{1}{2}\ln s - \frac{1}{2} + \frac{1}{2s}\,\mathbb{E}Y^2, \end{aligned}$$

   which is (3.2.23).

2. If $J(Y) < \infty$ and $\mathbb{E}Y^2 < \infty$, then for every $s > 0$

$$I(P_Y\|G_s) = J(Y) + \frac{1}{s^2}\,\mathbb{E}Y^2 - \frac{2}{s} < \infty, \tag{3.2.24}$$

   where $I(\cdot\|\cdot)$ is the relative Fisher information (see (3.2.4)). This

equality is verified as follows:

$$\begin{aligned} I(P_Y\|G_s) &= \int_{-\infty}^{\infty} p_Y(y)\left(\frac{\mathrm{d}}{\mathrm{d}y}\ln p_Y(y) - \frac{\mathrm{d}}{\mathrm{d}y}\ln \gamma_s(y)\right)^2 \mathrm{d}y \\ &= \int_{-\infty}^{\infty} p_Y(y)\left(\rho_Y(y) + \frac{y}{s}\right)^2 \mathrm{d}y \\ &= \mathbb{E}[\rho_Y^2(Y)] + \frac{2}{s}\,\mathbb{E}[Y\rho_Y(Y)] + \frac{1}{s^2}\,\mathbb{E}Y^2 \\ &= J(Y) + \frac{2}{s}\,\mathbb{E}[Y\rho_Y(Y)] + \frac{1}{s^2}\,\mathbb{E}Y^2. \end{aligned} \tag{3.2.25}$$

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

$$\begin{aligned} \mathbb{E}[Y\rho_Y(Y)] &= \int_{-\infty}^{\infty} y\,\rho_Y(y)\,p_Y(y)\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} y\,p_Y'(y)\,\mathrm{d}y \\ &= \left(\lim_{y\to\infty} y\,p_Y(y) - \lim_{y\to-\infty} y\,p_Y(y)\right) - \int_{-\infty}^{\infty} p_Y(y)\,\mathrm{d}y \\ &= -1 \end{aligned} \tag{3.2.26}$$

(see [135, Lemma A1] for another proof). Substituting (3.2.26) into (3.2.25), we get (3.2.24).

We are now in a position to prove the following result of Carlen [128]:

Proposition 3.4. The following statements are equivalent to hold for the class of real-valued random variables Y with a smooth density p_Y, such that $J(Y) < \infty$ and $\mathbb{E}Y^2 < \infty$:

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

Remark 3.7. Carlen's original derivation in [128] 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 for every real-valued random variable Y as in Proposition 3.4, it follows that

$$D(P_Y\|G_s) \leq \frac{s}{2}\, I(P_Y\|G_s), \qquad \forall\, s > 0. \tag{3.2.27}$$

Inequality (3.2.27) can be verified from (3.2.7)–(3.2.9), together with the equivalence of (3.2.2) and (3.2.3). Since $J(Y)$ and $\mathbb{E}Y^2$ are finite by assumption, the right side of (3.2.27) is finite and equal to (3.2.24). Therefore, $D(P_Y\|G_s)$ is also finite, and it is equal to (3.2.23). Hence, we can rewrite (3.2.27) as

$$\frac{1}{2}\ln\frac{1}{N(Y)} + \frac{1}{2}\ln s - \frac{1}{2} + \frac{1}{2s}\,\mathbb{E}Y^2 \leq \frac{s}{2}J(Y) + \frac{1}{2s}\,\mathbb{E}Y^2 - 1.$$

Since $\mathbb{E}Y^2 < \infty$, 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. Note that it is first required to show that $D(P_Y\|G_s) < \infty$. Since by assumption $J(Y)$ is finite and 2) holds, also $\frac{1}{N(Y)}$ is finite; since both $\mathbb{E}[Y^2]$ and $\frac{1}{N(Y)}$ are finite, it follows from (3.2.23) that $D(P_Y\|G_s)$ is finite. $\square$

We now turn to prove 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 [136], that expresses the divergence between two probability measures 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, the minimum MSE (MMSE) at SNR s is defined as

$$\mathsf{mmse}(Y,s) \triangleq \inf_{\varphi} \mathbb{E}[(Y-\varphi(\sqrt{s}Y+Z))^2], \tag{3.2.28}$$

where the infimum is taken over all measurable functions (estimators) $\varphi\colon \mathbb{R}\to\mathbb{R}$. It is well-known that the infimum in (3.2.28) 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 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 computed assuming that $Y\sim Q$. Then, the resulting *mismatched* MSE is given by

$$\mathsf{mse}_Q(Y,s) = \mathbb{E}\left[\left(Y-\mathbb{E}_Q[Y|\sqrt{s}Y+Z]\right)^2\right], \tag{3.2.29}$$

where the outer expectation on the right side is computed using the correct distribution P of Y. Then, the following relation holds for the divergence between P and Q (see [136, Theorem 1]):

$$D(P\|Q) = \frac{1}{2}\int_0^\infty [\mathsf{mse}_Q(Y,s) - \mathsf{mmse}(Y,s)]\,\mathrm{d}s. \tag{3.2.30}$$

We will apply the formula (3.2.30) 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 every $s>0$,

$$\mathsf{mse}_Q(Y,s) = \mathsf{mse}_G(Y,s) = \mathsf{lmmse}(Y,s), \tag{3.2.31}$$

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

$$\mathsf{lmmse}(Y,s) = \inf_{a,b\in\mathbb{R}} \mathbb{E}\left[\left(Y-a(\sqrt{s}Y+Z)-b\right)^2\right]. \tag{3.2.32}$$

The infimum in (3.2.32) is achieved by $a^*=\sqrt{\frac{s}{1+s}}$ and $b=0$, giving

$$\mathsf{lmmse}(Y,s) = \frac{1}{1+s}. \tag{3.2.33}$$

Moreover, $\mathsf{mmse}(Y,s)$ can be bounded from below using the so-called *van Trees inequality* [137] (see also Appendix 3.A):

$$\mathsf{mmse}(Y,s) \geq \frac{1}{J(Y)+s}. \tag{3.2.34}$$

Then

$$\begin{aligned} D(P_Y\|G) &= \tfrac{1}{2}\int_0^\infty (\mathsf{Immse}(Y,s) - \mathsf{mmse}(Y,s))\,\mathrm{d}s && (3.2.35)\\ &\leq \tfrac{1}{2}\int_0^\infty \left(\frac{1}{1+s} - \frac{1}{J(Y)+s}\right)\mathrm{d}s && (3.2.36)\\ &= \tfrac{1}{2}\lim_{\lambda\to\infty}\int_0^\lambda \left(\frac{1}{1+s} - \tfrac{1}{J(Y)+s}\right)\mathrm{d}s \\ &= \tfrac{1}{2}\lim_{\lambda\to\infty}\ln\left(\frac{J(Y)\,(1+\lambda)}{J(Y)+\lambda}\right) \\ &= \tfrac{1}{2}\ln J(Y), && (3.2.37) \end{aligned}$$

where (3.2.35) holds by combining (3.2.30) and (3.2.31) with $P = P_Y$ and $Q = G$; (3.2.36) holds by using (3.2.33) and (3.2.34). On the other hand, using (3.2.23) with $s = \mathbb{E}Y^2 = 1$ yields

$$D(P_Y\|G) = \frac{1}{2}\ln\frac{1}{N(Y)}. \tag{3.2.38}$$

Combining (3.2.37) and (3.2.38), we recover Stam's inequality $N(Y)J(Y) \geq 1$. Moreover, the van Trees bound (3.2.34) is achieved with equality if and only if Y is a standard Gaussian random variable.

3.2.2 From Gaussian LSI 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 an arbitrary 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

$\|\nabla f(X^n)\| \le 1$ almost surely, we have

$$\mathbb{P}\Big(f(X^n) \ge \mathbb{E}f(X^n) + r\Big) \le \exp\left(-\frac{r^2}{2}\right), \quad \forall\, r \ge 0 \tag{3.2.39}$$

Proof. Let $P = G^n$ denote the distribution of X^n, and let Q be a probability measure such that $P \ll\gg Q$ (i.e., P and Q are mutually absolutely continuous). Then, every event that has P-probability 1 also has Q-probability 1 and vice versa. Since f is differentiable, it is finite everywhere, so $P^{(f)}$ and P are mutually absolutely continuous. Hence, every event that occurs P-a.s. also occurs $P^{(tf)}$-a.s. for all $t \in \mathbb{R}$. In particular, $\|\nabla f(X^n)\| \le 1$ $P^{(tf)}$-a.s. for all $t > 0$. Therefore, applying the modified LSI (3.2.13) to $g = tf$ for some $t > 0$ gives

$$D(P^{(tf)}\|P) \le \left(\frac{t^2}{2}\right) \mathbb{E}_P^{(tf)}\Big[\|\nabla f(X^n)\|^2\Big] \le \frac{t^2}{2}. \tag{3.2.40}$$

Finally, in view of the Herbst argument, using Corollary 3.1.1 with $U = f(X^n) - \mathbb{E}f(X^n)$ yields (3.2.39). □

Remark 3.8. Corollary 3.1.1 and inequality (3.2.13) with $g = tf$ imply that, for every smooth function f with $\|\nabla f(X^n)\|^2 \le L$ a.s.,

$$\mathbb{P}\Big(f(X^n) \ge \mathbb{E}f(X^n) + r\Big) \le \exp\left(-\frac{r^2}{2L}\right), \quad \forall\, r \ge 0. \tag{3.2.41}$$

Thus, the constant κ in the Gaussian concentration bound (3.1.2) is controlled by the sensitivity of f to modifications of its coordinates.

Having established a concentration result for a smooth f with a bounded gradient, we proceed to the more general case where f is Lipschitz.

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

$$\mathbb{P}\Big(f(X^n) \ge \mathbb{E}f(X^n) + r\Big) \le \exp\left(-\frac{r^2}{2}\right), \quad \forall\, r \ge 0. \tag{3.2.42}$$

Proof. By Rademacher's theorem (see, e.g., [138, Section 3.1.2]), the assumption that $f\colon \mathbb{R}^n \to \mathbb{R}$ is 1-Lipschitz yields its differentiability almost everywhere in $\mathbb{R}^n$ with $\|\nabla f\| \leq 1$. Hence, $\|\nabla f(X^n)\| \leq 1$ almost surely (since $X_1, \ldots, X_n$ are i.i.d. standard Gaussian random variables). Consequently, (3.2.42) follows from Proposition 3.5. □

3.2.3 Hypercontractivity, Gaussian log-Sobolev inequality, and Rényi divergence

We close our treatment of the Gaussian LSI with a striking result, proved by Gross [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 [139].

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

$$Y = e^{-t}X + \sqrt{1 - e^{-2t}}Z \tag{3.2.43}$$

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 (3.2.43) 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 every $t \geq 0$.
3. Using the terminology of [13, Chapter 4], the channel family $\{\mathrm{OU}(t)\}_{t=0}^{\infty}$ is *ordered by degradation*: for every $t_1, t_2 \geq 0$ we have

$$\mathrm{OU}(t_1 + t_2) = \mathrm{OU}(t_2) \circ \mathrm{OU}(t_1) = \mathrm{OU}(t_1) \circ \mathrm{OU}(t_2), \tag{3.2.44}$$

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

$$\begin{aligned} & e^{-(t_1+t_2)}X + \sqrt{1-e^{-2(t_1+t_2)}}Z \\ & \quad \stackrel{\mathrm{d}}{=} e^{-t_2}\left[e^{-t_1}X + \sqrt{1-e^{-2t_1}}Z_1\right] + \sqrt{1-e^{-2t_2}}Z_2 \\ & \quad \stackrel{\mathrm{d}}{=} e^{-t_1}\left[e^{-t_2}X + \sqrt{1-e^{-2t_2}}Z_1\right] + \sqrt{1-e^{-2t_1}}Z_2 \end{aligned} \tag{3.2.45}$$

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 $(\Omega, \mathscr{F}, \mathbb{P})$, 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 an arbitrary real-valued random variable X, it is possible to 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)

$$\mathrm{d}Y_t = -Y_t\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \qquad t \geq 0 \tag{3.2.46}$$

with the initial condition $Y_0 \stackrel{\mathrm{d}}{=} X$, where $\{B_t\}_{t\geq 0}$ denotes the standard one-dimensional Wiener process (Brownian motion). The solution of the SDE (3.2.46), known as the *Langevin equation* [140, p. 75], is given by the so-called *Ornstein–Uhlenbeck process*

$$Y_t = Xe^{-t} + \sqrt{2}\int_0^t e^{-(t-s)}\,\mathrm{d}B_s, \qquad t \geq 0$$

where, by the Itô isometry property, the variance of the zero-mean

additive Gaussian noise is indeed

$$\mathbb{E}\left[\left(\sqrt{2}\int_0^t e^{-(t-s)}\,\mathrm{d}B_s\right)^2\right] = 2\int_0^t e^{-2(t-s)}\mathrm{d}s$$
$$= 1 - e^{-2t}, \quad \forall\, t \geq 0$$

(see, e.g., [141, p. 358] or [142, p. 127]). This explains our choice of the name "Ornstein–Uhlenbeck channel" for the random transformation (3.2.43).

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

$$D_\alpha(P\|Q) \triangleq \frac{1}{\alpha - 1}\,\ln\left(\int \mathrm{d}\mu \left(\frac{\mathrm{d}P}{\mathrm{d}\mu}\right)^\alpha \left(\frac{\mathrm{d}Q}{\mathrm{d}\mu}\right)^{1-\alpha}\right), \tag{3.2.47}$$

where μ is an arbitrary σ-finite measure that dominates both P and Q. If $P \ll Q$, we have the equivalent form

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1}\,\ln\left(\mathbb{E}_Q\left[\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)^\alpha\right]\right). \tag{3.2.48}$$

We recall several key properties of the Rényi divergence (see [143]):

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,

$$D(P\|Q) = \sup_{0<\alpha<1} D_\alpha(P\|Q) \leq \inf_{\alpha>1} D_\alpha(P\|Q)$$

and, if $D(P\|Q) = \infty$ or there exists some $\beta > 1$ such that $D_\beta(P\|Q) < \infty$, then also

$$D(P\|Q) = \lim_{\alpha\downarrow 1} D_\alpha(P\|Q). \tag{3.2.49}$$

2. By defining $D_1(P\|Q)$ as $D(P\|Q)$, 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 every channel (stochastic transformation) T that takes U as input we have

$$D_\alpha(\tilde{P}\|\tilde{Q}) \le D_\alpha(P\|Q), \qquad \forall\, \alpha > 0 \tag{3.2.50}$$

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 every 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 every $t \ge 0$, 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 (3.2.50), we have

$$D_\alpha(P_t\|G) \le D_\alpha(P\|G), \qquad \forall\, t \ge 0,\ \alpha > 0. \tag{3.2.51}$$

This is, of course, nothing but the second law of thermodynamics for Markov chains (see, e.g., [144, Section 4.4] or [139]) applied to the continuous-time Markov process governed by the Langevin equation (3.2.46). We will now show, however, that the Gaussian LSI of Gross (see Theorem 3.2.1) implies a stronger statement: for every $\alpha > 1$ and $\varepsilon \in (0, 1)$, there exists a positive constant $\tau = \tau(\alpha, \varepsilon)$, such that

$$D_\alpha(P_t\|G) \le \varepsilon D_\alpha(P\|G), \qquad \forall\, t \ge \tau. \tag{3.2.52}$$

By increasing the parameter t, the output distribution P_t resembles the invariant distribution G more and more, where the measure of similarity is given by a Rényi divergence. Here is the precise result:

Theorem 3.2.3. The Gaussian LSI of Theorem 3.2.1 is equivalent to the following statement: for every α, β such that $1 < \beta < \alpha < \infty$

$$D_\alpha(P_t\|G) \le \left(\frac{\alpha(\beta-1)}{\beta(\alpha-1)}\right) D_\beta(P\|G),\ \forall\, t \ge \tfrac{1}{2}\ln\left(\frac{\alpha-1}{\beta-1}\right). \tag{3.2.53}$$

The proof of Theorem 3.2.3 is provided in Appendix 3.B (with a certain equality, involved in this proof, that is proved separately in Appendix 3.C).

Remark 3.9. The original hypercontractivity result of Gross is stated as an inequality relating suitable norms of $g_t = \frac{\mathrm{d}P_t}{\mathrm{d}G}$ and $g = \frac{\mathrm{d}P}{\mathrm{d}G}$; we refer the reader to the original paper [44] or to the lecture notes of Guionnet and Zegarlinski [51] for the traditional treatment of hypercontractivity.

Remark 3.10. To see that Theorem 3.2.3 implies (3.2.52), fix $\alpha > 1$ and $\varepsilon \in (0, 1)$. Let

$$\beta = \beta(\varepsilon, \alpha) \triangleq \frac{\alpha}{\alpha - \varepsilon(\alpha - 1)}.$$

It is easy to see that $1 < \beta < \alpha$ and $\frac{\alpha(\beta-1)}{\beta(\alpha-1)} = \varepsilon$. Hence, Theorem 3.2.3 implies that

$$D_\alpha(P_t \| P) \leq \varepsilon D_\beta(P\|G), \quad \forall\, t \geq \tfrac{1}{2} \ln\left(1 + \frac{\alpha(1-\varepsilon)}{\varepsilon}\right) \triangleq \tau(\alpha, \varepsilon).$$

Since the Rényi divergence $D_\alpha(\cdot\|\cdot)$ is non-decreasing in the parameter α, and $1 < \beta < \alpha$, it follows that $D_\beta(P\|G) \leq D_\alpha(P\|G)$. Therefore, the last inequality implies that

$$D_\alpha(P_t \| P) \leq \varepsilon D_\alpha(P\|G), \quad \forall\, t \geq \tau(\alpha, \varepsilon).$$

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 every $t \geq 0$

$$D(P_t\|G) \leq e^{-2t} D(P\|G). \tag{3.2.54}$$

Proof. Let $\alpha = 1 + \varepsilon e^{2t}$ and $\beta = 1 + \varepsilon$ for some $\varepsilon > 0$. Then using Theorem 3.2.3, we have

$$D_{1+\varepsilon e^{2t}}(P_t\|G) \leq \left(\frac{e^{-2t} + \varepsilon}{1 + \varepsilon}\right) D_{1+\varepsilon}(P\|G), \qquad \forall\, t \geq 0. \tag{3.2.55}$$

Taking the limit of both sides of (3.2.55) as $\varepsilon \downarrow 0$ and using (3.2.49) (note that $D_\alpha(P\|G) < \infty$ for $\alpha > 1$), we get (3.2.54). □

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 [54] and consider a probability space $(\Omega, \mathscr{F}, \mu)$ 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 $af + b \in \mathcal{A}$ for every $a \geq 0$ 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 every $f \in \mathcal{A}$, $a \geq 0$, and $b \in \mathbb{R}$,

$$\Gamma(af + b) = a\,\Gamma f. \tag{3.3.1}$$

We say that μ satisfies a *logarithmic Sobolev inequality* with constant $c \geq 0$, or LSI(c) for short, if

$$D(\mu^{(f)} \| \mu) \leq \frac{c}{2}\,\mathbb{E}_\mu^{(f)}\left[(\Gamma f)^2\right], \qquad \forall\, f \in \mathcal{A}. \tag{3.3.2}$$

Here, as before, $\mu^{(f)}$ denotes the f-tilting of μ, i.e.,

$$\frac{\mathrm{d}\mu^{(f)}}{\mathrm{d}\mu} = \frac{\exp(f)}{\mathbb{E}_\mu[\exp(f)]}, \tag{3.3.3}$$

and $\mathbb{E}_\mu^{(f)}[\cdot]$ in the right side of (3.3.2) denotes expectation with respect to $\mu^{(f)}$.

Remark 3.11. We have expressed the LSI 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*

$$\begin{aligned}\mathrm{Ent}_\mu(g) &\triangleq \int g \ln g \,\mathrm{d}\mu - \int g \,\mathrm{d}\mu \cdot \ln\left(\int g \,\mathrm{d}\mu\right) \\ &\equiv \mathbb{E}_\mu[g \ln g] - \mathbb{E}_\mu[g]\,\ln \mathbb{E}_\mu[g]\end{aligned} \tag{3.3.4}$$

with the convention that $0 \ln 0 \triangleq 0$. Due to the convexity of the function $f(t) = t \ln t$ $(t \geq 0)$, Jensen's inequality gives $\mathrm{Ent}_\mu(g) \geq 0$. The LSI($c$) condition in (3.3.2) can be equivalently written as (see [54, (1.1)])

$$\mathrm{Ent}_\mu\big(\exp(f)\big) \leq \frac{c}{2} \int (\Gamma f)^2 \exp(f)\, \mathrm{d}\mu. \tag{3.3.5}$$

To see the equivalence of (3.3.2) and (3.3.5), note that

$$\begin{aligned} \mathrm{Ent}_\mu\big(\exp(f)\big) &= \int \exp(f) \ln \left(\frac{\exp(f)}{\int \exp(f) \mathrm{d}\mu} \right) \mathrm{d}\mu \\ &= \mathbb{E}_\mu[\exp(f)] \int \frac{\mathrm{d}\mu^{(f)}}{\mathrm{d}\mu} \ln \left(\frac{\mathrm{d}\mu^{(f)}}{\mathrm{d}\mu} \right) \mathrm{d}\mu \\ &= \mathbb{E}_\mu[\exp(f)]\, D(\mu^{(f)} \| \mu) \end{aligned} \tag{3.3.6}$$

and

$$\begin{aligned} \int (\Gamma f)^2 \exp(f)\, \mathrm{d}\mu &= \mathbb{E}_\mu[\exp(f)] \int (\Gamma f)^2\, \mathrm{d}\mu^{(f)} \\ &= \mathbb{E}_\mu[\exp(f)]\, \mathbb{E}_\mu^{(f)} \left[(\Gamma f)^2 \right]. \end{aligned} \tag{3.3.7}$$

Substituting (3.3.6) and (3.3.7) into (3.3.5), we obtain (3.3.2). We note that the entropy functional Ent is homogeneous of degree 1: for every g such that $\mathrm{Ent}_\mu(g) < \infty$ and $a > 0$, we have

$$\mathrm{Ent}_\mu(ag) = a\, \mathbb{E}_\mu \left[g \ln \frac{g}{\mathbb{E}_\mu[g]} \right] = a\, \mathrm{Ent}_\mu(g). \tag{3.3.8}$$

Remark 3.12. Strictly speaking, (3.3.2) should be called a modified (or exponential) logarithmic Sobolev inequality. The ordinary LSI takes the form (see [54, (1.2)])

$$\mathrm{Ent}_\mu(g^2) \leq 2c \int (\Gamma g)^2\, \mathrm{d}\mu \tag{3.3.9}$$

for all strictly positive $g \in \mathcal{A}$. If the pair $(\mathcal{A}, \Gamma)$ is such that $\psi \circ g \in \mathcal{A}$ for every $g \in \mathcal{A}$ and for every C^∞ function $\psi \colon \mathbb{R} \to \mathbb{R}$, and Γ obeys the chain rule

$$\Gamma(\psi \circ g) = |\psi' \circ g|\, \Gamma g \qquad \forall\, g \in \mathcal{A},\, \psi \in C^\infty, \tag{3.3.10}$$

then (3.3.2) and (3.3.9) are equivalent. In order to show this, recall the equivalence of (3.3.2) and (3.3.5) (see Remark 3.11); the equivalence of (3.3.5) and (3.3.9) is proved in the following when the mapping Γ satisfies the chain rule in (3.3.10). Indeed, if (3.3.9) holds then using it with $g = \exp(f/2)$ gives

$$\begin{aligned}\mathrm{Ent}_\mu\big(\exp(f)\big) &\le 2c \int \Big(\Gamma\big(\exp(f/2)\big)\Big)^2 \,\mathrm{d}\mu \\ &= \frac{c}{2} \int (\Gamma f)^2 \exp(f)\,\mathrm{d}\mu\end{aligned}$$

which is (3.3.5). The last equality in the above display follows from (3.3.10) which implies that

$$\Gamma\big(\exp(f/2)\big) = \tfrac{1}{2}\,\exp(f/2)\cdot \Gamma f.$$

Conversely, the combination of (3.3.5) with $f = 2\ln g$ and (3.3.10) with $\psi(t) = t\ln t$ for $t > 0$ (by continuous extension $\psi(0) = 0$) 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 (3.3.9). Again, the last equality is a consequence of (3.3.10), which gives $\Gamma(2\ln g) = \frac{2\,\Gamma g}{g}$ for all strictly positive $g \in \mathcal{A}$. In fact, the Gaussian LSI we have looked at in Section 3.2 is an instance of the LSI in (3.3.2), for which $\Gamma f = \|\nabla f\|$ satisfies the chain rule in (3.3.10).

Recalling the discussion of Section 3.1.4, we now show how to 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) with 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 every $c > 0$). Then $tf \in \mathcal{A}$ for every $t \ge 0$, so applying (3.3.2) to $g = tf$ we get

$$\begin{aligned}D\big(P^{(tf)}\|P\) &\le \frac{c}{2}\,\mathbb{E}_P^{(tf)}\Big[\big(\Gamma(tf)\big)^2\Big] \\ &= \frac{ct^2}{2}\,\mathbb{E}_P^{(tf)}\Big[(\Gamma f)^2\Big] \\ &\le \frac{c\|\Gamma f\|_\infty^2 t^2}{2}, \qquad (3.3.11)\end{aligned}$$

where the second step uses the fact that $\Gamma(tf) = t\,\Gamma f$ for every $f \in \mathcal{A}$ and $t \geq 0$. In other words, P satisfies the bound (3.1.39) for every $g \in \mathcal{A}$ with $E(g) = \|\Gamma g\|_\infty^2$. Therefore, using the bound (3.3.11) together with Corollary 3.1.1 gives

$$\mathbb{P}\big(f(X^n) \geq \mathbb{E}f(X^n) + r\big) \leq \exp\left(-\frac{r^2}{2c\|\Gamma f\|_\infty^2}\right), \qquad \forall\, r \geq 0. \quad (3.3.12)$$

3.3.1 Tensorization of the logarithmic Sobolev inequality

In the above demonstration, we have capitalized on an appropriate LSI in order to derive a concentration inequality. Showing that a LSI 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 $\mathcal{X}$-valued random variables $X_1, \ldots, X_n$ are independent under P, we can use once again the "divide-and-conquer" tensorization strategy: we break the original n-dimensional problem into n one-dimensional subproblems, demonstrate that each marginal distribution P_{X_i} $(i = 1, \ldots, n)$ satisfies a LSI 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 an arbitrary function $f\colon \mathcal{X}^n \to \mathbb{R}$, for $i \in \{1, \ldots, n\}$, and for an arbitrary $(n-1)$-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 $X_1, \ldots, X_n$ be n independent $\mathcal{X}$-valued 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 $i \in \{1, \ldots, n\}$,

$$f_i(\cdot|\bar{x}^i) \in \mathcal{A}_i, \qquad \forall\, \bar{x}^i \in \mathcal{X}^{n-1}. \quad (3.3.13)$$

Define the operator Γ that maps each $f \in \mathcal{A}$ to

$$\Gamma f = \sqrt{\sum_{i=1}^n (\Gamma_i f_i)^2}, \quad (3.3.14)$$

which is shorthand for

$$\Gamma f(x^n) = \sqrt{\sum_{i=1}^{n} \Big(\Gamma_i f_i(x_i|\bar{x}^i) \Big)^2}, \qquad \forall\, x^n \in \mathcal{X}^n. \tag{3.3.15}$$

Then, the following statements hold:

1. If there exists a constant $c \geq 0$ such that, for every $i \in \{1, \ldots, n\}$, 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 every $f \in \mathcal{A}$ with $\mathbb{E}[f(X^n)] = 0$, and every $r \geq 0$,
$$\mathbb{P}(f(X^n) \geq r) \leq \exp\left(-\frac{r^2}{2c\|\Gamma f\|_\infty^2} \right). \tag{3.3.16}$$

Proof. We first verify 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, $g_i(\cdot|\bar{x}^i) = af_i(\cdot|\bar{x}^i) + b \in \mathcal{A}_i$ for every $i \in \{1, \ldots, n\}$ and an arbitrary $\bar{x}^i$, which relies on (3.3.13) and the property (LSI-1) of the pair $(\mathcal{A}_i, \Gamma_i)$. Hence, $f \in \mathcal{A}$ implies that $g = af + b \in \mathcal{A}$ for every $a \geq 0$ and $b \in \mathbb{R}$, so (LSI-1) holds. From the definition of Γ in (3.3.14) and (3.3.15), it is readily seen that (LSI-2) and (LSI-3) hold as well.

Next, for every $f \in \mathcal{A}$ and $t \geq 0$, we have

$$\begin{aligned}
D(P^{(tf)} \| P) &\leq \sum_{i=1}^{n} D\Big(P^{(tf)}_{X_i|\overline{X}^i} \Big\| P_{X_i} \Big| P^{(tf)}_{\overline{X}^i} \Big) \\
&= \sum_{i=1}^{n} \int D\Big(P^{(tf)}_{X_i|\overline{X}^i = \bar{x}^i} \Big\| P_{X_i} \Big)\, \mathrm{d}P^{(tf)}_{\overline{X}^i}(\bar{x}^i) \\
&= \sum_{i=1}^{n} \int D\Big(P^{(tf_i(\cdot|\bar{x}^i))}_{X_i} \Big\| P_{X_i} \Big)\, \mathrm{d}P^{(tf)}_{\overline{X}^i}(\bar{x}^i) \\
&\leq \frac{ct^2}{2} \sum_{i=1}^{n} \int \mathbb{E}^{(tf_i(\cdot|\bar{x}^i))}_{P_{X_i}} \left[\Big(\Gamma_i f_i(X_i|\bar{x}^i) \Big)^2 \right]\, \mathrm{d}P^{(tf)}_{\overline{X}^i}(\bar{x}^i) \\
&= \frac{ct^2}{2} \sum_{i=1}^{n} \mathbb{E}^{(tf)}_{P_{\overline{X}^i}} \left\{ \mathbb{E}^{(tf)}_{P_{X_i|\overline{X}^i}} \Big[\Big(\Gamma_i f_i(X_i|\overline{X}^i) \Big)^2 \Big] \right\} \\
&= \frac{ct^2}{2} \cdot \mathbb{E}^{(tf)}_{P} \Big[(\Gamma f)^2 \Big],
\end{aligned} \tag{3.3.17}$$

where the first step uses Proposition 3.2 with $Q = P^{(tf)}$, the second is by the definition of conditional divergence where $P_{X_i} = P_{X_i|\overline{X}^i}$, the third is due to (3.1.31), 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, and (3.3.14). 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 (3.3.12). □

3.3.2 Maurer's thermodynamic method

Theorem 3.3.1 enables to derive concentration inequalities in product spaces whenever an appropriate LSI 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 [145], it is possible to derive log-Sobolev inequalities in a wide variety of settings by means of a single unified method. This method has two basic ingredients:

1. A certain "thermodynamic" representation of the divergence $D(\mu^{(f)}\|\mu)$, $f \in \mathcal{A}$, as an integral of the *variances* of f 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 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 $(\Omega, \mathscr{F}, \mu)$ 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 an arbitrary $h\colon \Omega \to \mathbb{R}$ 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 [145, Theorem 3]):

Lemma 3.3.2. Let $f\colon \Omega \to \mathbb{R}$ be a function such that $\mathbb{E}_{\mu}[\exp(\lambda f)] < \infty$ for all $\lambda > 0$. Then, the following equality holds:

$$D(\mu^{(\lambda f)}\|\mu) = \int_0^{\lambda}\int_t^{\lambda} \mathsf{var}_{\mu}^{(sf)}[f]\,\mathrm{d}s\,\mathrm{d}t, \quad \forall\,\lambda > 0. \tag{3.3.18}$$

Proof. We start by noting that (see (3.1.16) and (3.1.17))

$$\Lambda'(t) = \mathbb{E}_{\mu}^{(tf)}[f] \qquad \text{and} \qquad \Lambda''(t) = \mathsf{var}_{\mu}^{(tf)}[f], \tag{3.3.19}$$

and, in particular, $\Lambda'(0) = \mathbb{E}_{\mu}[f]$. Moreover, from (3.1.19), we get

$$D(\mu^{(\lambda f)}\|\mu) = \lambda^2\,\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{\Lambda(\lambda)}{\lambda}\right) = \lambda\Lambda'(\lambda) - \Lambda(\lambda). \tag{3.3.20}$$

Now, using (3.3.19), we get

$$\begin{aligned}\lambda\Lambda'(\lambda) &= \int_0^{\lambda}\Lambda'(\lambda)\,\mathrm{d}t\\ &= \int_0^{\lambda}\left(\int_0^{\lambda}\Lambda''(s)\,\mathrm{d}s + \Lambda'(0)\right)\mathrm{d}t\\ &= \int_0^{\lambda}\left(\int_0^{\lambda}\mathsf{var}_{\mu}^{(sf)}[f]\,\mathrm{d}s + \mathbb{E}_{\mu}[f]\right)\mathrm{d}t\end{aligned} \tag{3.3.21}$$

and

$$\begin{aligned}\Lambda(\lambda) &= \int_0^{\lambda}\Lambda'(t)\,\mathrm{d}t\\ &= \int_0^{\lambda}\left(\int_0^{t}\Lambda''(s)\,\mathrm{d}s + \Lambda'(0)\right)\mathrm{d}t\\ &= \int_0^{\lambda}\left(\int_0^{t}\mathsf{var}_{\mu}^{(sf)}[f]\,\mathrm{d}s + \mathbb{E}_{\mu}[f]\right)\mathrm{d}t.\end{aligned} \tag{3.3.22}$$

Substituting (3.3.21) and (3.3.22) into (3.3.20), we get (3.3.18). □

Remark 3.13. The thermodynamic interpretation of (3.3.18) stems from the fact that the tilted measures $\mu^{(tf)}$ can be viewed as the *Gibbs measures*, 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 a physical system Σ, the elements $x \in \Omega$ are the states (or configurations) of Σ, μ is a base (or reference) measure, and f is the energy function. We can view μ as an 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 globally minimizes the *free energy functional*

$$\Psi_\theta(\nu) \triangleq \mathbb{E}_\nu[f] + \theta D(\nu\|\mu). \tag{3.3.23}$$

It is claimed that $\Psi_\theta(\nu)$ is uniquely minimized by $\nu^* = \mu^{(-tf)}$, where $t = 1/\theta$ is the *inverse temperature*. To see this, consider an arbitrary ν, where we may assume, without loss of generality, that $\nu \ll \mu$. Let $\psi \triangleq \mathrm{d}\nu/\mathrm{d}\mu$. Then

$$\frac{\mathrm{d}\nu}{\mathrm{d}\mu^{(-tf)}} = \frac{\frac{\mathrm{d}\nu}{\mathrm{d}\mu}}{\frac{\mathrm{d}\mu^{(-tf)}}{\mathrm{d}\mu}} = \frac{\psi}{\frac{\exp(-tf)}{\mathbb{E}_\mu[\exp(-tf)]}} = \psi\, \exp(tf)\, \mathbb{E}_\mu[\exp(-tf)]$$

and

$$\begin{aligned}
\Psi_\theta(\nu) &= \frac{1}{t}\, \mathbb{E}_\nu[tf + \ln\psi] \\
&= \frac{1}{t}\, \mathbb{E}_\nu\left[\ln\big(\psi \exp(tf)\big)\right] \\
&= \frac{1}{t}\, \mathbb{E}_\nu\left[\ln \frac{\mathrm{d}\nu}{\mathrm{d}\mu^{(-tf)}} - \Lambda(-t)\right] \\
&= \frac{1}{t}\left[D(\nu\|\mu^{(-tf)}) - \Lambda(-t)\right],
\end{aligned}$$

where, as before, $\Lambda(-t) \triangleq \ln\big(\mathbb{E}_\mu[\exp(-tf)]\big)$. Therefore, we have $\Psi_\theta(\nu) = \Psi_{1/t}(\nu) \geq -\Lambda(-t)/t$ with equality if and only if $\nu = \mu^{(-tf)}$.

The reader is referred to a recent monograph by Merhav [146], which highlights interesting relations between information theory and statistical physics.

Now the whole affair hinges on the second step, which involves bounding the variances $\mathsf{var}_\mu^{(tf)}[f]$ from above, for $t > 0$, 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 $(\mathcal{A}, \Gamma)$ and $\{(\mathcal{A}_i, \Gamma_i)\}_{i=1}^n$ be constructed as in the statement of Theorem 3.3.1. 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

$$\mathsf{var}_i^{(sg)}[g(X_i)|\overline{X}^i = \bar{x}^i] \le c\,(\Gamma_i g)^2\,, \qquad \forall\, \bar{x}^i \in \mathcal{X}^{n-1} \tag{3.3.24}$$

for all $i \in \{1, \ldots, n\}$, $s > 0$, and $g \in \mathcal{A}_i$, where $\mathsf{var}_i^{(g)}[\cdot|\overline{X}^i = \bar{x}^i]$ denotes the (conditional) variance with respect to $P^{(g)}_{X_i|\overline{X}^i=\bar{x}^i}$. 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

$$\begin{aligned}
&D\Big(P^{(f)}_{X_i|\overline{X}^i=\bar{x}^i} \,\|\, P_{X_i}\Big) \\
&= D\Big(P_{X_i}^{(f_i(\cdot|\bar{x}^i))} \,\|\, P_{X_i}\Big) && (3.3.25)\\
&= \int_0^1 \int_t^1 \mathsf{var}_i^{(sf_i(\cdot|\bar{x}^i))}[f_i(X_i|\overline{X}^i)|\overline{X}^i = \bar{x}^i]\,\mathrm{d}s\,\mathrm{d}t && (3.3.26)\\
&\le c\,(\Gamma_i f_i)^2 \int_0^1 \int_t^1 \mathrm{d}s\,\mathrm{d}t && (3.3.27)\\
&= \frac{c(\Gamma_i f_i)^2}{2} && (3.3.28)
\end{aligned}$$

where (3.3.25) uses the fact that $P^{(f)}_{X_i|\overline{X}^i=\bar{x}^i}$ is equal to the $f_i(\cdot|\bar{x}^i)$-tilting of P_{X_i}; (3.3.26) uses Lemma 3.3.2; (3.3.27) uses (3.3.24) with $g = f_i(\cdot|\bar{x}^i) \in \mathcal{A}_i$; (3.3.28) holds since $\int_0^1 \int_t^1 \mathrm{d}s\,\mathrm{d}t = \int_0^1 (1-t)\mathrm{d}t = \frac{1}{2}$. We have therefore established that, for each $i \in \{1, \ldots, n\}$, 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 will be useful for establishing bounds like (3.3.24):

Lemma 3.3.4. Let U be a random variable such that $U \in [a, b]$ a.s. for some $-\infty < a \leq b < +\infty$. Then

$$\mathsf{var}[U] \leq \tfrac{1}{4}(b-a)^2. \tag{3.3.29}$$

Proof. $\mathsf{var}[U]$ is maximized when $\mathbb{P}(U = a) = \mathbb{P}(U = b) = \frac{1}{2}$, which attains the bound in the right side of (3.3.29). □

Lemma 3.3.5. [145] Let f be a real-valued function, for which there exists $d \in \mathbb{R}$ such that $f - \mathbb{E}_\mu[f] \leq d$ almost surely. Then, for all $t > 0$,

$$\mathsf{var}_\mu^{(tf)}[f] \leq \exp(td)\, \mathsf{var}_\mu[f]. \tag{3.3.30}$$

Proof.

$$\begin{aligned}
\mathsf{var}_\mu^{(tf)}[f] &= \mathsf{var}_\mu^{(tf)}\Big\{f - \mathbb{E}_\mu\left[f\right]\Big\} && (3.3.31)\\
&\leq \mathbb{E}_\mu^{(tf)}\left[\left(f - \mathbb{E}_\mu[f]\right)^2\right] && (3.3.32)\\
&= \mathbb{E}_\mu\left[\frac{\exp(tf)\left(f - \mathbb{E}_\mu[f]\right)^2}{\mathbb{E}_\mu[\exp(tf)]}\right] && (3.3.33)\\
&\leq \mathbb{E}_\mu\Big\{\left(f - \mathbb{E}_\mu[f]\right)^2 \exp\left[t\left(f - \mathbb{E}_\mu[f]\right)\right]\Big\} && (3.3.34)\\
&\leq \exp(td)\, \mathbb{E}_\mu\left[\left(f - \mathbb{E}_\mu\left[f\right]\right)^2\right], && (3.3.35)
\end{aligned}$$

where

- (3.3.31) holds since $\mathsf{var}[f] = \mathsf{var}[f + c]$ for $c \in \mathbb{R}$;
- (3.3.32) uses the bound $\mathsf{var}[U] \leq \mathbb{E}[U^2]$;
- (3.3.33) is by definition of the tilted distribution $\mu^{(tf)}$;
- (3.3.34) is verified by applying Jensen's inequality to the denominator;
- (3.3.35) relies on the assumption that $f - \mathbb{E}_\mu[f] \leq d$, and the monotonicity of the exponential function (note that $t > 0$).

□

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 i-th position and zeros elsewhere. For every $f\colon \mathcal{X}^n \to \mathbb{R}$, define

$$\Gamma f(x^n) \triangleq \sqrt{\sum_{i=1}^{n} \big(f(x^n \oplus e_i) - f(x^n)\big)^2}, \qquad \forall\, x^n \in \mathcal{X}^n, \tag{3.3.36}$$

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$(\frac{1}{2})$, case first:

Theorem 3.3.6 (Discrete log-Sobolev inequality for the symmetric Bernoulli measure)**.** Let $\mathcal{A}$ be the set of all functions $f\colon \mathcal{X}^n \to \mathbb{R}$ with $\mathcal{X} = \{0,1\}$, and let $\Gamma\colon \mathcal{A} \to [0,\infty)$ be as defined in (3.3.36). Moreover, let $X_1, \ldots, X_n$ be i.i.d. Bernoulli$(\frac{1}{2})$ random variables, and let P denote their joint distribution. Then, P satisfies LSI$(\frac{1}{4})$ with respect to $(\mathcal{A}, \Gamma)$. In other words, for every $f\colon \mathcal{X}^n \to \mathbb{R}$,

$$D(P^{(f)} \| P) \le \tfrac{1}{8}\, \mathbb{E}_P^{(f)} \left[(\Gamma f)^2 \right]. \tag{3.3.37}$$

Proof. It is easy to verify that $(\mathcal{A}, \Gamma)$ satisfies the conditions (LSI-1)–(LSI-3).

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

$$\Gamma_0\, g \triangleq |g(0) - g(1)| = |g(x) - g(x \oplus 1)|, \quad \forall\, x \in \{0,1\}. \tag{3.3.38}$$

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 every $g \in \mathcal{A}_i$, the random variable $U_i = g(X_i)$ is bounded between $g(0)$ and $g(1)$. Hence, by Lemma 3.3.4, we have

$$\mathsf{var}_i^{(sg)}[g(X_i) | \overline{X}^i = \bar{x}^i] \le \tfrac{1}{4}\big(g(0) - g(1)\big)^2 = \tfrac{1}{4}(\Gamma_i g)^2 \tag{3.3.39}$$

for all $g \in \mathcal{A}_i$, $\bar{x}^i \in \mathcal{X}^{n-1}$. In other words, the condition (3.3.24) of Theorem 3.3.3 holds with $c = \frac{1}{4}$. In addition, it is easy to see that

the operator Γ constructed from $\Gamma_1, \ldots, \Gamma_n$ according to (3.3.14) is precisely the one in (3.3.36). Therefore, by Theorem 3.3.3, the pair $(\mathcal{A}, \Gamma)$ satisfies LSI($\frac{1}{4}$) with respect to P, which proves (3.3.37). $\square$

Now let us consider the general case where $X_1, \ldots, X_n$ are i.i.d. Bernoulli(p) random variables with $p \in (0,1)$. We will use Maurer's method to give an alternative, simpler proof of the following result by Ledoux [52, Corollary 5.9] (it actually suggests a sharpened version of the latter result, as it is explained in Remark 3.14):

Theorem 3.3.7. Consider an arbitrary function $f\colon \{0,1\}^n \to \mathbb{R}$ with the property that there exists some $c > 0$ such that

$$\max_{i \in \{1,\ldots,n\}} |f(x^n \oplus e_i) - f(x^n)| \leq c \tag{3.3.40}$$

for all $x^n \in \{0,1\}^n$. Let $\{X_i\}_{i=1}^n$ be i.i.d. Bernoulli(p) random variables with $p \in (0,1)$, and let P be their joint distribution. Then

$$D(P^{(f)} \| P) \leq pq \left(\frac{(qc-1)\exp(qc)+1}{(qc)^2} \right) \mathbb{E}_P^{(f)} \left[(\Gamma f)^2 \right], \tag{3.3.41}$$

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 [145] has other examples. For $n = 1$, let

$$a = |\Gamma f| = |f(0) - f(1)|, \tag{3.3.42}$$

where Γ is defined as in (3.3.38). Without loss of generality, let $f(0) = 0$ and $f(1) = a$. Then

$$\mathbb{E}[f] = pa, \quad \mathsf{var}[f] = pqa^2. \tag{3.3.43}$$

Using (3.3.43) 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

$$\begin{aligned} D(P^{(f)}\|P) &\le pqa^2 \int_0^1 \int_t^1 \exp(sqa)\,\mathrm{d}s\,\mathrm{d}t \\ &= pqa^2 \left(\frac{(qa-1)\exp(qa)+1}{(qa)^2}\right) \\ &\le pqa^2 \left(\frac{(qc-1)\exp(qc)+1}{(qc)^2}\right), \end{aligned} \tag{3.3.44}$$

where the last step follows from the fact that the function

$$u \mapsto u^{-2}[(u-1)\exp(u)+1]$$

(defined, for continuity, to be $\frac{1}{2}$ at $u=0$) is monotonically increasing in $[0,\infty)$, and $0 \le qa \le qc$ due to (3.3.40) and (3.3.42). Since $a^2 = (\Gamma f)^2$, (3.3.44) gives

$$D(P^{(f)}\|P) \le pq \left(\frac{(qc-1)\exp(qc)+1}{(qc)^2}\right) \mathbb{E}_P^{(f)}\left[(\Gamma f)^2\right],$$

so we have established (3.3.41) for $n=1$.

Now consider an arbitrary $n \in \mathbb{N}$. Since the condition in (3.3.40) can be expressed as $\left|f_i(0|\bar{x}^i) - f_i(1|\bar{x}^i)\right| \le c$ for all $i \in \{1,\dots,n\}$ and $\bar{x}^i \in \{0,1\}^{n-1}$, we can use (3.3.44) to write

$$\begin{aligned} &D\Big(P_{X_i}^{(f_i(\cdot|\bar{x}^i))}\|P_{X_i}\Big) \\ &\le pq \left(\frac{(qc-1)\exp(qc)+1}{(qc)^2}\right) \mathbb{E}_{P_{X_i}}^{(f_i(\cdot|\bar{x}^i))}\Big[\left(\Gamma_i f_i(X_i|\bar{x}^i)\right)^2\Big] \end{aligned}$$

for every $i = 1,\dots,n$ and all $\bar{x}^i \in \{0,1\}^{n-1}$. With this, the same sequence of steps that led to (3.3.17) in the proof of Theorem 3.3.1 can be used to complete the proof of (3.3.41) for an arbitrary n. □

In Appendix 3.D, we comment on the relations between the log-Sobolev inequalities for Bernoulli and Gaussian measures.

Remark 3.14. Note that (3.3.41) improves the bound by Ledoux in [52, Corollary 5.9], which is equivalent to

$$D(P^{(f)}\|P) \le pq \left(\frac{(c-1)\exp(c)+1}{c^2}\right) \mathbb{E}_P^{(f)}\left[(\Gamma f)^2\right]. \tag{3.3.45}$$

The improvement in (3.3.41) follows from a replacement of c on the right side of (3.3.45) with qc (recall that $q \in (0,1)$); this can be verified due the fact that the function

$$u \mapsto u^{-2}[(u-1)\exp(u)+1], \quad u \in (0,\infty)$$

is monotonically increasing.

3.3.4 The method of bounded differences revisited

As our second illustration of Maurer's method, we give an information-theoretic proof of McDiarmid's inequality in Theorem 2.4.3. Following the exposition in [145, Section 4.1], we have the following re-statement of McDiarmid's inequality:

Theorem 3.3.8. Let $X_1, \ldots, X_n$ be independent $\mathcal{X}$-valued 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 $0 \leq c_1, \ldots, c_n < +\infty$ such that, for each $i \in \{1, \ldots, n\}$,

$$\left| f_i(x|\bar{x}^i) - f_i(y|\bar{x}^i) \right| \leq c_i, \qquad \forall\, x, y \in \mathcal{X},\ \bar{x}^i \in \mathcal{X}^{n-1}. \tag{3.3.46}$$

Then, for every $r \geq 0$,

$$\mathbb{P}\Big(f(X^n) \geq r\Big) \leq \exp\left(-\frac{2r^2}{\sum_{i=1}^n c_i^2}\right). \tag{3.3.47}$$

Proof. Let $\mathcal{A}_0$ be the set of 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

$$\Gamma_0\, g \triangleq \sup_{x \in \mathcal{X}} g(x) - \inf_{x \in \mathcal{X}} g(x). \tag{3.3.48}$$

It is easy to verify that properties (LSI-1)–(LSI-3) hold for the pair $(\mathcal{A}_0, \Gamma_0)$ since in particular

$$\Gamma_0(ag + b) = a\,\Gamma_0\, g, \quad \forall\, a \geq 0,\ b \in \mathbb{R}.$$

For all $i \in \{1, \ldots, n\}$, 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 non-negative constant (3.3.48). Moreover, for every $g \in \mathcal{A}_i$, 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 yields

$$\mathsf{var}_i^{(sg)}[g(X_i)|\overline{X}^i = \bar{x}^i] \le \tfrac{1}{4}(\Gamma_i g)^2, \qquad \forall\, g \in \mathcal{A}_i,\ \bar{x}^i \in \mathcal{X}^{n-1}, \quad (3.3.49)$$

which implies that the condition (3.3.24) of Theorem 3.3.3 holds with $c = \frac{1}{4}$.

Let $\mathcal{A}$ be the set of all bounded measurable functions $f\colon \mathcal{X}^n \to \mathbb{R}$. Then, for every $f \in \mathcal{A}$, $i \in \{1,\dots,n\}$ and $x^n \in \mathcal{X}^n$,

$$\begin{aligned}
&\sup_{x_i\in\mathcal{X}_i} f(x_1,\dots,x_i,\dots,x_n) - \inf_{x_i\in\mathcal{X}_i} f(x_1,\dots,x_i,\dots,x_n)\\
&= \sup_{x_i\in\mathcal{X}_i} f_i(x_i|\bar{x}^i) - \inf_{x_i\in\mathcal{X}_i} f_i(x_i|\bar{x}^i)\\
&= \Gamma_i f_i(\cdot|\bar{x}^i). \qquad (3.3.50)
\end{aligned}$$

By constructing $\Gamma\colon \mathcal{A} \to [0,\infty)$ from $\Gamma_1,\dots,\Gamma_n$, according to (3.3.14) and (3.3.50), Theorem 3.3.1 can be applied to $(\mathcal{A},\Gamma)$. By Theorem 3.3.3 and (3.3.49), it follows that $(\mathcal{A},\Gamma)$ satisfies LSI($\frac{1}{4}$) for every product probability measure on $\mathcal{X}^n$. Hence, (3.3.12) implies that, for every $r \ge 0$ and bounded f with $\mathbb{E}[f(X^n)] = 0$,

$$\mathbb{P}\Big(f(X^n) \ge r\Big) \le \exp\left(-\frac{2r^2}{\|\Gamma f\|_\infty^2}\right). \qquad (3.3.51)$$

If f satisfies (3.3.46), then

$$\begin{aligned}
\|\Gamma f\|_\infty^2 &= \sup_{x^n\in\mathcal{X}^n} \sum_{i=1}^n \big(\Gamma_i f_i(x_i|\bar{x}^i)\big)^2\\
&\le \sum_{i=1}^n \sup_{x^n\in\mathcal{X}^n} \big(\Gamma_i f_i(x_i|\bar{x}^i)\big)^2\\
&= \sum_{i=1}^n \sup_{x^n\in\mathcal{X}^n,\, y\in\mathcal{X}} |f_i(x_i|\bar{x}^i) - f_i(y|\bar{x}^i)|^2\\
&\le \sum_{i=1}^n c_i^2. \qquad (3.3.52)
\end{aligned}$$

Substituting the bound in the right side of (3.3.52) into the right side of (3.3.51) gives (3.3.47). □

Note that Maurer's method yields the constant in the exponent of McDiarmid's inequality; it is instructive to compare it to an earlier approach in [147] which, by also using the entropy method, gave an exponent that is smaller by a factor of 8.

3.3.5 Log-Sobolev inequalities for Poisson and compound Poisson measures

Let P_λ denote the probability measure of a Poisson random variable with parameter $\lambda > 0$, i.e., $\mathsf{P}_\lambda(n) \triangleq \frac{e^{-\lambda}\lambda^n}{n!}$ for $n \in \mathbb{N}_0 \triangleq \{0, 1, 2, \ldots, \}$. Bobkov and Ledoux [55] derived the following LSI: for every function $f \colon \mathbb{N}_0 \to \mathbb{R}$,

$$D(\mathsf{P}_\lambda^{(f)} \| \mathsf{P}_\lambda) \leq \lambda \, \mathbb{E}_{\mathsf{P}_\lambda}^{(f)} \left[(\Gamma f) \, e^{\Gamma f} - e^{\Gamma f} + 1 \right], \tag{3.3.53}$$

where Γ is the modulus of the discrete gradient:

$$\Gamma f(x) \triangleq |f(x) - f(x+1)|, \qquad \forall \, x \in \mathbb{N}_0. \tag{3.3.54}$$

(The inequality (3.3.53) can be obtained by combining the LSI in [55, Corollary 7] with equality (3.3.6).)

Using tensorization of (3.3.53), Kontoyiannis and Madiman [148] derived an LSI for the *compound Poisson distribution*. We recall that a compound Poisson distribution is defined as follows: given $\lambda > 0$ and a probability measure μ, let $N \sim \mathsf{P}_\lambda$ and let $X_1, X_2, \ldots$ be i.i.d. random variables with distribution μ, which are independent of N; the compound Poisson distribution, denoted by $\mathsf{CP}_{\lambda,\mu}$, is the probability distribution of the random sum

$$Z = \sum_{i=1}^{N} X_i, \tag{3.3.55}$$

with the convention that $Z = 0$ whenever $N = 0$.

For the purpose of proving the log-Sobolev inequality for compound Poisson measures in [148, Theorem 1], we need the following result:

Lemma 3.3.9. If $Z \sim \mathsf{CP}_{\lambda,\mu}$, then

$$Z \stackrel{\mathrm{d}}{=} \sum_{k=1}^{\infty} k Y_k, \qquad Y_k \sim \mathsf{P}_{\lambda\mu(k)}, \;\; \forall \, k \in \mathbb{N} \tag{3.3.56}$$

where $\{Y_k\}_{k=1}^{\infty}$ are independent, and $\stackrel{\mathrm{d}}{=}$ means equality in distribution.

Proof. The characteristic function of Z in (3.3.56) 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 is equal to the characteristic function of $Z \sim \mathsf{CP}_{\lambda,\mu}$ as defined in (3.3.55). The lemma follows from the fact that equality in distribution holds if and only if the characteristic functions coincide. □

Theorem 3.3.10 (Log-Sobolev inequality for compound Poisson measures [148])**.** For an arbitrary probability measure μ on $\mathbb{N} = \{1, 2, \ldots\}$ and an arbitrary bounded function $f\colon \mathbb{N}_0 \to \mathbb{R}$, and for every $\lambda > 0$,

$$D(\mathsf{CP}^{(f)}_{\lambda,\mu}\|\mathsf{CP}_{\lambda,\mu}) \le \lambda\sum_{k=1}^{\infty}\mu(k)\,\mathbb{E}^{(f)}_{\mathsf{CP}_{\lambda,\mu}}\left[(\Gamma_k f)\,e^{\Gamma_k f} - e^{\Gamma_k f} + 1\right] \quad (3.3.57)$$

where, for every $k \in \mathbb{N}$ and $x \in \mathbb{N}_0$,

$$\Gamma_k f(x) \triangleq |f(x) - f(x+k)|. \quad (3.3.58)$$

Proof. For $n \in \mathbb{N}$, let P_n be the joint distribution of the independent random variables $Y_1, \ldots, Y_n$ in (3.3.56). Consider an arbitrary bounded function $f\colon \mathbb{N}_0 \to \mathbb{R}$, and define the function g by

$$g(y_1, \ldots, y_n) \triangleq f\left(\sum_{k=1}^{n} k y_k\right), \qquad \forall\, y_1, \ldots, y_n \in \mathbb{N}_0.$$

Denote by $\bar{P}_n$ the distribution of the sum $S_n \triangleq \sum_{k=1}^{n} kY_k$, then

$$\begin{aligned}
D(\bar{P}^{(f)}_n\|\bar{P}_n) &= \mathbb{E}_{\bar{P}_n}\left[\left(\frac{\exp\big(f(S_n)\big)}{\mathbb{E}_{\bar{P}_n}[\exp\big(f(S_n)\big)]}\right)\,\ln\left(\frac{\exp\big(f(S_n)\big)}{\mathbb{E}_{\bar{P}_n}[\exp\big(f(S_n)\big)]}\right)\right] \\
&= \mathbb{E}_{P_n}\left[\left(\frac{\exp\big(g(Y^n)\big)}{\mathbb{E}_{P_n}[\exp\big(g(Y^n)\big)]}\right)\,\ln\left(\frac{\exp\big(g(Y^n)\big)}{\mathbb{E}_{P_n}[\exp\big(g(Y^n)\big)]}\right)\right] \\
&= D(P^{(g)}_n\|P_n) \\
&\le \sum_{k=1}^{n} D(P^{(g)}_{Y_k|\bar{Y}^k}\|P_{Y_k}\,|\,P^{(g)}_{\bar{Y}^k}), \qquad (3.3.59)
\end{aligned}$$

where the last line uses Proposition 3.2 and the fact that P_n is a product measure. Due to the equality

$$\frac{\mathrm{d}P^{(g)}_{Y_k|\bar{Y}^k=\bar{y}^k}}{\mathrm{d}P_{Y_k}} = \frac{\exp\left(g_k(\cdot|\bar{y}^k)\right)}{\mathbb{E}_{\mathsf{P}_{\lambda\mu(k)}}[\exp\left(g_k(Y_k|\bar{y}^k)\right)]} \tag{3.3.60}$$

with $P_{Y_k} = \mathsf{P}_{\lambda\mu(k)}$, applying the Bobkov–Ledoux inequality (3.3.53) to P_{Y_k} and all functions of the form $g_k(\cdot|\bar{y}^k)$ gives

$$\begin{aligned} &D\Big(P^{(g)}_{Y_k|\bar{Y}^k}\|P_{Y_k}\,|\,P^{(g)}_{\bar{Y}^k}\Big) \\ &\leq \lambda\mu(k)\,\mathbb{E}^{(g)}_{P_n}\left[\left(\Gamma g_k(Y_k|\bar{Y}^k)\right)e^{\Gamma g_k(Y_k|\bar{Y}^k)} - e^{\Gamma g_k(Y_k|\bar{Y}^k)} + 1\right] \end{aligned} \tag{3.3.61}$$

where Γ in (3.3.54) is the absolute value of the one-dimensional discrete gradient. For every $y^n \in \{0,1,2,\ldots\}^n$ and $k \in \{1,\ldots,n\}$, we have

$$\begin{aligned} \Gamma g_k(y_k|\bar{y}^k) &= \left|g_k(y_k|\bar{y}^k) - g_k(y_k+1|\bar{y}^k)\right| \\ &= \left|f\left(ky_k + \sum_{j\in\{1,\ldots,n\}\setminus\{k\}} jy_j\right)\right. \\ &\qquad \left. - f\left(k(y_k+1) + \sum_{j\in\{1,\ldots,n\}\setminus\{k\}} jy_j\right)\right| \\ &= \left|f\left(\sum_{j=1}^n jy_j\right) - f\left(\sum_{j=1}^n jy_j + k\right)\right| \\ &= \Gamma_k f\left(\sum_{j=1}^n jy_j\right) = \Gamma_k f(S_n). \end{aligned}$$

Using this in (3.3.61) and performing the reverse change of measure from P_n to $\bar{P}_n$, we can write

$$\begin{aligned} &D(P^{(g)}_{Y_k|\bar{Y}^k}\|P_{Y_k}\,|\,P^{(g)}_{\bar{Y}^k}\Big) \\ &\leq \lambda\mu(k)\,\mathbb{E}^{(f)}_{\bar{P}_n}\left[\left(\Gamma_k f(S_n)\right)e^{\Gamma_k f(S_n)} - e^{\Gamma_k f(S_n)} + 1\right]. \end{aligned} \tag{3.3.62}$$

Therefore, the combination of (3.3.59) and (3.3.62) gives

$$\begin{aligned} D(\bar{P}_n^{(f)} \| \bar{P}_n) &\leq \lambda \sum_{k=1}^{n} \mu(k)\, \mathbb{E}_{\bar{P}_n}^{(f)} \left[(\Gamma_k f)\, e^{\Gamma_k f} - e^{\Gamma_k f} + 1 \right] \\ &\leq \lambda \sum_{k=1}^{\infty} \mu(k)\, \mathbb{E}_{\bar{P}_n}^{(f)} \left[(\Gamma_k f)\, e^{\Gamma_k f} - e^{\Gamma_k f} + 1 \right] \end{aligned} \tag{3.3.63}$$

where the second line follows from the inequality $xe^x - e^x + 1 \geq 0$ that holds for all $x \geq 0$.

Now we will take the limit as $n \to \infty$ of both sides of (3.3.63). For the left side, we use the fact that, by (3.3.56), $\bar{P}_n$ converges in distribution to $\mathsf{CP}_{\lambda,\mu}$ as $n \to \infty$. Since f is bounded, $\bar{P}_n^{(f)} \to \mathsf{CP}_{\lambda,\mu}^{(f)}$ in distribution. Therefore, by the bounded convergence theorem, we have

$$\lim_{n\to\infty} D(\bar{P}_n^{(f)} \| \bar{P}_n) = D(\mathsf{CP}_{\lambda,\mu}^{(f)} \,\|\, \mathsf{CP}_{\lambda,\mu}). \tag{3.3.64}$$

For the right side of (3.3.63), we have

$$\begin{aligned} &\sum_{k=1}^{\infty} \mu(k)\, \mathbb{E}_{\bar{P}_n}^{(f)} \left[(\Gamma_k f)\, e^{\Gamma_k f} - e^{\Gamma_k f} + 1 \right] \\ &= \mathbb{E}_{\bar{P}_n}^{(f)} \left\{ \sum_{k=1}^{\infty} \mu(k) \left[(\Gamma_k f)\, e^{\Gamma_k f} - e^{\Gamma_k f} + 1 \right] \right\} \\ &\xrightarrow{n\to\infty} \mathbb{E}_{\mathsf{CP}_{\lambda,\mu}}^{(f)} \left[\sum_{k=1}^{\infty} \mu(k) \left((\Gamma_k f)\, e^{\Gamma_k f} - e^{\Gamma_k f} + 1 \right) \right] \\ &= \sum_{k=1}^{\infty} \mu(k)\, \mathbb{E}_{\mathsf{CP}_{\lambda,\mu}}^{(f)} \left[(\Gamma_k f)\, e^{\Gamma_k f} - e^{\Gamma_k f} + 1 \right] \end{aligned} \tag{3.3.65}$$

where the first and last steps follow from Fubini's theorem, and the second step follows from the bounded convergence theorem. Putting (3.3.63)–(3.3.65) together, we get the inequality in (3.3.57). □

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 in obtaining tight bounds on the deviation probabilities $\mathbb{P}(f(X^n) \geq \mathbb{E}f(X^n) + r)$ 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 [86], with subsequent refinements by Steele [87]:

Theorem 3.3.11 (Efron–Stein–Steele inequality [86], [87])**.** Let $X_1, \ldots, X_n$ be independent $\mathcal{X}$-valued random variables. Consider a function $f\colon \mathcal{X}^n \to \mathbb{R}$ whose scaled versions tf, for all sufficiently small $t > 0$, are exponentially integrable. Then

$$\mathsf{var}[f(X^n)] \leq \sum_{i=1}^{n} \mathbb{E}\left\{\mathsf{var}\big[f(X^n)\big|\overline{X}^i\big]\right\}. \tag{3.3.66}$$

Proof. Let $P = P_{X_1} \otimes \ldots \otimes P_{X_n}$ be the joint probability distribution of $X_1, \ldots, X_n$. By Proposition 3.2, for every $t > 0$, we have

$$D\big(P^{(tf)}\|P\big) \leq \sum_{i=1}^{n} D\big(P^{(tf)}_{X_i|\overline{X}^i}\|P_{X_i}|P^{(tf)}_{\overline{X}^i}\big).$$

Using Lemma 3.3.2, we can rewrite this inequality as

$$\begin{aligned}&\int_0^t \int_s^t \mathsf{var}_P^{(\tau f)}[f]\,\mathrm{d}\tau\,\mathrm{d}s \\ &\leq \sum_{i=1}^{n} \mathbb{E}_{P^{(tf)}_{\overline{X}^i}}\left[\int_0^t \int_s^t \mathsf{var}_{P_{X_i|\overline{X}^i}}^{(\tau f_i(\cdot|\overline{X}^i))}[f]\,\mathrm{d}\tau\,\mathrm{d}s\right].\end{aligned} \tag{3.3.67}$$

Dividing both sides by t^2, and passing to the limit as $t \to 0$, we get from L'Hôpital's rule

$$\lim_{t\to 0} \frac{1}{t^2} \int_0^t \int_s^t \mathsf{var}_P^{(\tau f)}[f]\,\mathrm{d}\tau\,\mathrm{d}s = \tfrac{1}{2}\,\mathsf{var}_P[f] = \tfrac{1}{2}\,\mathsf{var}[f(X^n)], \tag{3.3.68}$$

and

$$\begin{aligned}
&\lim_{t\to 0}\frac{1}{t^2}\sum_{i=1}^{n}\mathbb{E}_{P_{\overline{X}^i}^{(tf)}}\left[\int_0^t\int_s^t \mathsf{var}_{P_{X_i|\overline{X}^i}}^{(\tau f_i(\cdot|\overline{X}^i))}[f]\,\mathrm{d}\tau\,\mathrm{d}s\right]\\
&=\sum_{i=1}^{n}\mathbb{E}_{P_{\overline{X}^i}}\left\{\lim_{t\to 0}\frac{1}{t^2}\int_0^t\int_s^t \mathsf{var}_{P_{X_i|\overline{X}^i}}^{(\tau f_i(\cdot|\overline{X}^i))}[f]\,\mathrm{d}\tau\,\mathrm{d}s\right\}\\
&=\sum_{i=1}^{n}\mathbb{E}_{P_{\overline{X}^i}}\left\{\tfrac{1}{2}\,\mathsf{var}_{P_{X_i|\overline{X}^i}}[f]\right\}\\
&=\tfrac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left\{\mathsf{var}\big[f(X^n)|\overline{X}^i\big]\right\} \qquad (3.3.69)
\end{aligned}$$

where the first equality in (3.3.69) is justified by invoking the dominated convergence theorem (recall the pointwise convergence of $P_{\overline{X}^i}^{(tf)}$ to $P_{\overline{X}^i}$, as $t\to 0$, which holds under the assumption that the scaled functions tf are exponentially integrable for all sufficiently small $t>0$), and the second equality holds due to L'Hôpital's rule. Inequality (3.3.66) finally follows from (3.3.67)–(3.3.69). □

The following result considers the connection between log-Sobolev inequalities and another class of functional inequalities, the so-called *Poincaré inequalities*. Consider, as before, a probability space $(\Omega,\mathscr{F},\mu)$ and a pair $(\mathcal{A},\Gamma)$ satisfying the conditions (LSI-1)–(LSI-3). Then, we say that μ satisfies a *Poincaré inequality* with constant $c\geq 0$ if

$$\mathsf{var}_\mu[f]\leq c\,\mathbb{E}_\mu\left[(\Gamma f)^2\right],\qquad \forall\, f\in\mathcal{A}. \qquad (3.3.70)$$

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 every $f\in\mathcal{A}$ and $t>0$, we can use Lemma 3.3.2 to express the corresponding LSI(c) for the function tf as

$$\int_0^t\int_s^t \mathsf{var}_\mu^{(\tau f)}[f]\,\mathrm{d}\tau\,\mathrm{d}s\leq\frac{ct^2}{2}\,\mathbb{E}_\mu^{(tf)}\left[(\Gamma f)^2\right]. \qquad (3.3.71)$$

By dividing both sides of (3.3.71) by t^2 and passing to the limit as $t\to 0$, we obtain (see (3.3.68))

$$\tfrac{1}{2}\,\mathsf{var}_\mu[f]\leq\tfrac{1}{2}\,c\,\mathbb{E}_\mu\left[(\Gamma f)^2\right].$$

Multiplying both sides by 2, we see that μ indeed satisfies (3.3.70). □

The following analogue of Theorem 3.3.1 shows that the Poincaré inequalities tensorize.

Theorem 3.3.13. Let $X_1, \ldots, X_n$ be independent $\mathcal{X}$-valued 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

$$f_i(\cdot|\bar{x}^i) \in \mathcal{A}_i, \qquad \forall\, \bar{x}^i \in \mathcal{X}^{n-1} \tag{3.3.72}$$

Define the operator Γ that maps each $f \in \mathcal{A}$ to Γf in (3.3.14). Suppose that, for every $i \in \{1, \ldots, n\}$, P_{X_i} satisfies a Poincaré inequality with constant $c \geq 0$ with respect to $(\mathcal{A}_i, \Gamma_i)$ (see (3.3.70)). Consequently, P satisfies a Poincaré inequality with constant c with respect to $(\mathcal{A}, \Gamma)$.

Proof. It 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

We have been looking so far at concentration of measure through the lens of *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 with $\frac{\mathrm{d}P^{(f)}}{\mathrm{d}P} \propto \exp(f)$, 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 LSI of Gross [44]). The common theme underlying these functional inequalities is that every 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 a metric $d(\cdot, \cdot)$, and consider the following generalized definition of the modulus of the gradient of an arbitrary function $f\colon \mathcal{X}^n \to \mathbb{R}$:

$$|\nabla f|(x^n) \triangleq \limsup_{y^n\colon d(x^n, y^n) \to 0} \frac{|f(x^n) - f(y^n)|}{d(x^n, y^n)}, \quad \forall\, x^n \in \mathcal{X}^n. \tag{3.4.1}$$

Let 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 $\|f\|_{\mathrm{Lip}} < \infty$. It is easy to verify that the pair $(\mathcal{A}, \Gamma)$ with

$$\Gamma f(x^n) \triangleq |\nabla f|(x^n), \quad \forall\, x^n \in \mathcal{X}^n \tag{3.4.3}$$

satisfies the conditions (LSI-1)–(LSI-3) in Section 3.3. 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 (3.3.12) and the inequality $\|\Gamma f\|_\infty \leq \|f\|_{\mathrm{Lip}}$, which follows from (3.4.1)–(3.4.3), yields the concentration inequality

$$\mathbb{P}\Big(f(X^n) \geq \mathbb{E}f(X^n) + r\Big) \leq \exp\left(-\frac{r^2}{2c\|f\|_{\mathrm{Lip}}^2}\right), \quad \forall\, r > 0. \tag{3.4.4}$$

Some examples of concentration we have discussed so far in this chapter can be seen to fit this theme. Consider, for instance, the following case:

Example 3.1 (Euclidean metric)**.** For $\mathcal{X} = \mathbb{R}$, equip the product space $\mathcal{X}^n = \mathbb{R}^n$ with the ordinary Euclidean metric:

$$\begin{aligned} d(x^n, y^n) &= \|x^n - y^n\| \\ &= \sqrt{\sum_{i=1}^n (x_i - y_i)^2}, \quad \forall\, x^n, y^n \in \mathbb{R}^n. \end{aligned} \tag{3.4.5}$$

From (3.4.2) and (3.4.5), the Lipschitz constant $\|f\|_{\mathrm{Lip}}$ of a function $f\colon \mathcal{X}^n \to \mathbb{R}$ is given by

$$\|f\|_{\mathrm{Lip}} = \sup_{x^n \neq y^n} \frac{|f(x^n) - f(y^n)|}{\|x^n - y^n\|}, \tag{3.4.6}$$

and, for every probability measure P on $\mathbb{R}^n$ that satisfies LSI(c), the concentration inequality (3.4.4) holds. We have already seen in (3.2.13) an instance of this with $P = G^n$, which satisfies LSI(1).

The above example suggests that the metric structure plays the primary role, while the functional concentration inequalities like (3.4.4)

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. The key tool underlying this approach is the notion of *transportation cost*, which can be used to define a metric on probability measures over the space of interest in terms of a given base metric on this space. This metric on distributions can be related to the divergence via the so-called *transportation-cost inequalities*. The pioneering work by K. Marton in [59] and [75] has shown that one can use these inequalities to deduce concentration.

3.4.1 Concentration and isoperimetry

We next give 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 [149]. 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 an arbitrary subset $\mathcal{C} \subseteq \mathcal{X}$ and every $r > 0$, define the *r-blowup of* $\mathcal{C}$ by

$$\mathcal{C}_r \triangleq \{x \in \mathcal{X} \colon d(x, \mathcal{C}) < r\}, \tag{3.4.7}$$

where

$$d(x, \mathcal{C}) \triangleq \inf_{y \in \mathcal{C}} d(x, y) \tag{3.4.8}$$

is the distance from the point $x \in \mathcal{X}$ to the subset $\mathcal{C}$. We say that the probability measure μ has *normal* (or *Gaussian*) *concentration* on $(\mathcal{X}, d)$ if there exist positive constants K and κ such that

$$\mu(\mathcal{C}) \geq \tfrac{1}{2} \implies \mu(\mathcal{C}_r) \geq 1 - Ke^{-\kappa r^2}, \ \forall\, r > 0. \tag{3.4.9}$$

Remark 3.15. Of the two constants K and κ in (3.4.9), 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 (3.4.9) holds with that value of κ and some $K > 0$.

Remark 3.16. The concentration condition (3.4.9) is often weakened to the requirement that there exists $r_0 > 0$ such that

$$\mu(\mathcal{C}) \geq \tfrac{1}{2} \implies \mu(\mathcal{C}_r) \geq 1 - Ke^{-\kappa(r-r_0)^2}, \, \forall \, r \geq r_0 \tag{3.4.10}$$

(see, for example, [67, Remark 22.23] or [71, Proposition 3.3]). It is not hard to pass from (3.4.10) to the stronger statement (3.4.9), possibly with loosened constants (i.e., larger K and/or smaller κ). However, since we mainly care about sufficiently large values of r, (3.4.10) with *sharper constants* is preferable. In the sequel, whenever we talk about Gaussian concentration with constant $\kappa > 0$, we refer to (3.4.10) unless stated otherwise.

Here are three standard examples (see [3, Section 1.1]):

1. *Standard Gaussian distribution*: if $\mathcal{X} = \mathbb{R}^n$, $d(x, y) = \|x - y\|$ is the Euclidean metric in (3.4.5), and $\mu = G^n$ is the standard n-dimensional Gaussian distribution, then for an arbitrary Borel set $\mathcal{C} \subseteq \mathbb{R}^n$ with $G^n(\mathcal{C}) \geq \frac{1}{2}$ and for all $r > 0$

$$\begin{aligned} G^n(\mathcal{C}_r) &\geq \tfrac{1}{\sqrt{2\pi}} \int_{-\infty}^{r} \exp\left(-\tfrac{t^2}{2}\right) \mathrm{d}t \\ &\geq 1 - \tfrac{1}{2} \exp\left(-\tfrac{r^2}{2}\right), \end{aligned} \tag{3.4.11}$$

   i.e., (3.4.9) holds with $K = \frac{1}{2}$ and $\kappa = \frac{1}{2}$.

2. *Uniform distribution on the unit sphere*: if

$$\mathcal{X} = \mathbb{S}^n \equiv \left\{ x \in \mathbb{R}^{n+1} : \|x\| = 1 \right\}, \tag{3.4.12}$$

   $d(\cdot, \cdot)$ is the geodesic distance on $\mathbb{S}^n$, and $\mu = \sigma^n$ is a uniform distribution on $\mathbb{S}^n$, then for every Borel set $\mathcal{C} \subseteq \mathbb{S}^n$ with $\sigma^n(\mathcal{C}) \geq \frac{1}{2}$ and for all $r > 0$

$$\sigma^n(\mathcal{C}_r) \geq 1 - \exp\left(-\tfrac{1}{2}\,(n-1)r^2\right). \tag{3.4.13}$$

   In this instance, (3.4.9) holds with $K = 1$ and $\kappa = \frac{1}{2}\,(n-1)$.

3. *Equiprobable distribution on the Hamming cube*: if $\mathcal{X} = \{0,1\}^n$, $d(\cdot, \cdot)$ is the normalized Hamming distance

$$d(x, y) = \tfrac{1}{n} \sum_{i=1}^{n} 1_{\{x_i \neq y_i\}}, \quad \forall \, x, y \in \{0,1\}^n, \tag{3.4.14}$$

and $\mu = B^n$ is the equiprobable distribution on $\{0,1\}^n$ (i.e., $B = \text{Bernoulli}(\frac{1}{2})$), then for every $\mathcal{C} \subseteq \{0,1\}^n$ with $B^n(\mathcal{C}) \geq \frac{1}{2}$ and for all $r > 0$

$$B^n(\mathcal{C}_r) \geq 1 - \exp(-2nr^2), \tag{3.4.15}$$

yielding (3.4.9) with $K = 1$ and $\kappa = 2n$.

Remark 3.17. The Gaussian concentration in (3.4.9) is often discussed in the context of the so-called *isoperimetric inequalities*, which relate the 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 an arbitrary Borel subset $\mathcal{C} \subseteq \mathcal{X}$ define its *surface measure* as (see [3, Section 2.1])

$$\mu^+(\mathcal{C}) \triangleq \liminf_{r \to 0} \frac{\mu(\mathcal{C}_r \setminus \mathcal{C})}{r} = \liminf_{r \to 0} \frac{\mu(\mathcal{C}_r) - \mu(\mathcal{C})}{r}. \tag{3.4.16}$$

Then, the classical Gaussian isoperimetric inequality can be stated as follows: If $\mathcal{H}$ is a half-space in $\mathbb{R}^n$, i.e., $\mathcal{H} = \{x \in \mathbb{R}^n : \langle x, u \rangle < c\}$ for some $u \in \mathbb{R}^n$ with $\|u\| = 1$ and some $c \in \mathbb{R}$, and if $\mathcal{C} \subseteq \mathbb{R}^n$ is a Borel set with $G^n(\mathcal{C}) = G^n(\mathcal{H})$, then

$$(G^n)^+(\mathcal{C}) \geq (G^n)^+(\mathcal{H}), \tag{3.4.17}$$

with equality in (3.4.17) if and only if $\mathcal{C}$ is a half-space. In other words, the Gaussian isoperimetric inequality (3.4.17) 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 (3.4.17) says the following (see, e.g., [150]): consider a Borel set $\mathcal{C}$ in $\mathbb{R}^n$ and a half-space $\mathcal{H} = \{x \in \mathbb{R}^n : \langle x, u \rangle < c\}$ with $\|u\| = 1$, $c \geq 0$ and $G^n(\mathcal{C}) = G^n(\mathcal{H})$. Then, for every $r > 0$,

$$G^n(\mathcal{C}_r) \geq G^n(\mathcal{H}_r), \tag{3.4.18}$$

with equality in (3.4.18) if and only if $\mathcal{C}$ is an n-dimensional half-space. It can be also shown that for all $r > 0$

$$\begin{aligned} G^n(\mathcal{H}_r) &= \tfrac{1}{\sqrt{2\pi}} \int_{-\infty}^{c+r} \exp\left(-\tfrac{\xi^2}{2}\right) \mathrm{d}\xi \\ &\geq 1 - \tfrac{1}{2} \exp\left(-\tfrac{1}{2}(r+c)^2\right). \end{aligned}$$

If $G(\mathcal{C}) \geq \frac{1}{2}$, we can choose $c = 0$ and get (3.4.11).

Intuitively, (3.4.9) says that if μ has Gaussian concentration on $(\mathcal{X}, d)$, then most of the probability mass in $\mathcal{X}$ is concentrated around any set with probability at least $\frac{1}{2}$. At first glance, this seems to have nothing to do with what we have been looking at so far, namely the concentration of Lipschitz functions around their mean. However, as we next show, the geometric and functional pictures of the concentration of measure phenomenon are, in fact, equivalent. To that end, we 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

$$\mathbb{P}_\mu(f(X) \geq m_f) \geq \tfrac{1}{2}, \;\; \mathbb{P}_\mu(f(X) \leq m_f) \geq \tfrac{1}{2} \tag{3.4.19}$$

(note that a median of f may not be unique). The precise result is as follows:

Theorem 3.4.1. If $(\mathcal{X}, d, \mu)$ is a metric probability space, then μ has the normal concentration property (3.4.9) if and only if for every Lipschitz function $f\colon \mathcal{X} \to \mathbb{R}$

$$\mathbb{P}_\mu(f(X) \geq m_f + r) \leq K \exp\left(-\frac{\kappa r^2}{\|f\|_{\mathrm{Lip}}^2}\right), \qquad \forall\, r > 0 \tag{3.4.20}$$

where m_f is a μ-median of f.

Proof. Let μ satisfy (3.4.9), and fix an arbitrary Lipschitz function f which can be assumed without loss of generality to satisfy $\|f\|_{\mathrm{Lip}} = 1$. Let m_f be a μ-median of f, and define the set

$$\mathcal{C}^f \triangleq \Big\{x \in \mathcal{X}\colon f(x) \leq m_f\Big\}. \tag{3.4.21}$$

By definition (see (3.4.19)), $\mu(\mathcal{C}^f) \geq \frac{1}{2}$. Consequently, from (3.4.9), for all $r > 0$

$$\begin{aligned} \mu(\mathcal{C}_r^f) &\equiv \mathbb{P}_\mu\left(d(X, \mathcal{C}^f) < r\right) \\ &\geq 1 - K\exp(-\kappa r^2). \end{aligned} \tag{3.4.22}$$

For every $y \in \mathcal{C}^f$

$$f(X) - m_f \leq f(X) - f(y) \tag{3.4.23}$$

$$\leq d(X, y), \tag{3.4.24}$$

where (3.4.23) is due to (3.4.21), and (3.4.24) holds by the assumption that $\|f\|_{\mathrm{Lip}} = 1$ (see (3.4.6)). Taking an infimum on the right side of (3.4.24) over all $y \in \mathcal{C}^f$ gives

$$f(X) - m_f \leq d(X, \mathcal{C}^f). \tag{3.4.25}$$

Combining (3.4.22) and (3.4.25) implies that for all $r > 0$

$$\begin{aligned}\mathbb{P}_\mu\Big(f(X) - m_f < r\Big) &\geq \mathbb{P}_\mu\Big(d(X, \mathcal{C}^f) < r\Big) \\ &\geq 1 - K \exp(-\kappa r^2),\end{aligned}$$

which is (3.4.20).

Conversely, suppose that (3.4.20) holds for every Lipschitz f. Choose an arbitrary Borel set $\mathcal{C}$ with $\mu(\mathcal{C}) \geq \frac{1}{2}$, and define the function $f_{\mathcal{C}}(x) \triangleq d(x, \mathcal{C})$ for every $x \in \mathcal{X}$. Then $f_{\mathcal{C}}$ is 1-Lipschitz, since

$$\begin{aligned}|f_{\mathcal{C}}(x) - f_{\mathcal{C}}(y)| &= \Big|\inf_{u \in \mathcal{C}} d(x,u) - \inf_{u \in \mathcal{C}} d(y,u)\Big| \\ &\leq \sup_{u \in \mathcal{C}} |d(x,u) - d(y,u)| \\ &\leq d(x,y),\end{aligned}$$

where the last step is by the triangle inequality. Moreover, zero is a median of $f_{\mathcal{C}}$ since

$$\mathbb{P}_\mu(f_{\mathcal{C}}(X) \leq 0) = \mathbb{P}_\mu(X \in \mathcal{C}) \geq \tfrac{1}{2}, \;\; \mathbb{P}_\mu(f_{\mathcal{C}}(X) \geq 0) \geq \tfrac{1}{2},$$

where the second bound is vacuously true since $f_{\mathcal{C}} \geq 0$ everywhere. Consequently, with $m_f = 0$, we get that for all $r > 0$

$$\begin{aligned}1 - \mu(\mathcal{C}_r) &= \mathbb{P}_\mu(d(X, \mathcal{C}) \geq r) \\ &= \mathbb{P}_\mu(f_{\mathcal{C}}(X) \geq m_f + r) \\ &\leq K \exp(-\kappa r^2),\end{aligned}$$

which gives (3.4.9). □

In fact, for Lipschitz functions, Gaussian concentration around the mean also implies Gaussian concentration around every median, though 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 every 1-Lipschitz function $f\colon \mathcal{X} \to \mathbb{R}$ we have

$$\mathbb{P}_\mu\Big(f(X) \geq \mathbb{E}_\mu[f(X)] + r\Big) \leq K_0 \exp\big(-\kappa_0 r^2\big), \qquad \forall\, r > 0 \quad (3.4.26)$$

with some constants $K_0, \kappa_0 > 0$. Then, μ has the normal concentration property (3.4.9) with $K = K_0$ and $\kappa = \frac{\kappa_0}{4}$. Hence, the concentration inequality in (3.4.20) around every median m_f is satisfied with the same constants of κ and K.

Proof. Let $\mathcal{C} \subseteq \mathcal{X}$ be an arbitrary Borel set with $\mu(\mathcal{C}) \geq \frac{1}{2}$, and fix some $r > 0$. Let the function $f_{\mathcal{C},r}\colon \mathcal{X} \to [0, r]$ be defined as

$$f_{\mathcal{C},r}(x) = \min\left\{d(x, \mathcal{C}), r\right\}, \quad x \in \mathcal{X}. \quad (3.4.27)$$

It is easy to verify from (3.4.27) and the triangle inequality that

$$\|f_{\mathcal{C},r}\|_{\mathrm{Lip}} \leq 1, \quad (3.4.28)$$

and

$$\begin{aligned}
&\mathbb{E}_\mu[f_{\mathcal{C},r}(X)] \\
&= \int_{\mathcal{X}} \min\left\{d(x, \mathcal{C}), r\right\}\, \mu(\mathrm{d}x) && (3.4.29)\\
&= \underbrace{\int_{\mathcal{C}} \min\left\{d(x, \mathcal{C}), r\right\}\, \mu(\mathrm{d}x)}_{=0} + \int_{\mathcal{C}^c} \min\left\{d(x, \mathcal{C}), r\right\}\, \mu(\mathrm{d}x) && (3.4.30)\\
&\leq r\, \mu(\mathcal{C}^c) = \big(1 - \mu(\mathcal{C})\big)\, r && (3.4.31)
\end{aligned}$$

where (3.4.29) is due to (3.4.27); (3.4.30) holds since, by (3.4.8), $d(x, \mathcal{C}) = 0$ for all $x \in \mathcal{C}$. This consequently implies that

$$\begin{aligned}
1 - \mu(\mathcal{C}_r) &= \mathbb{P}_\mu\Big(d(X, \mathcal{C}) \geq r\Big) && (3.4.32)\\
&= \mathbb{P}_\mu\Big(f_{\mathcal{C},r}(X) \geq r\Big) && (3.4.33)\\
&\leq \mathbb{P}_\mu\Big(f_{\mathcal{C},r}(X) \geq \mathbb{E}_\mu[f_{\mathcal{C},r}(X)] + r\mu(\mathcal{C})\Big) && (3.4.34)\\
&\leq K_0 \exp\Big(-\kappa_0 \big(r\mu(\mathcal{C})\big)^2\Big) && (3.4.35)\\
&\leq K_0 \exp\Big(-\tfrac{1}{4}\, \kappa_0\, r^2\Big) && (3.4.36)
\end{aligned}$$

where (3.4.32) is due to (3.4.7); (3.4.33) is due to (3.4.27); (3.4.34) holds due to (3.4.29)–(3.4.30); (3.4.35) relies on (3.4.26) and since $f_{\mathcal{C},r}$ is 1-Lipschitz (see (3.4.28)); finally, (3.4.36) holds since by assumption $\mu(\mathcal{C}) \geq \frac{1}{2}$. Consequently, we get (3.4.9) with $K = K_0$ and $\kappa = \frac{1}{4}\kappa_0$. Theorem 3.4.1 implies that the concentration inequality in (3.4.20) holds for every μ-median m_f with the same constants of κ and K. $\square$

Remark 3.18. Let $(\mathcal{X}, d, \mu)$ be a metric probability space, and let μ have the Gaussian concentration property (3.4.9). Let $f\colon \mathcal{X} \to \mathbb{R}$ be an arbitrary Lipschitz function. The distance between the mean and an arbitrary μ-median of f can be upper bounded as follows:

$$\begin{aligned} |\mathbb{E}_\mu[f(X)] - m_f| &\leq \mathbb{E}_\mu[|f(X) - m_f|] && (3.4.37)\\ &= \int_0^\infty \mathbb{P}_\mu(|f(X) - m_f| \geq r)\,\mathrm{d}r && (3.4.38)\\ &\leq \int_0^\infty 2K \exp\left(-\frac{\kappa r^2}{\|f\|_{\mathrm{Lip}}^2}\right)\mathrm{d}r && (3.4.39)\\ &= \sqrt{\frac{\pi}{\kappa}}\, K\|f\|_{\mathrm{Lip}} && (3.4.40) \end{aligned}$$

where (3.4.38) holds since $\mathbb{E}[U] = \int_0^\infty \mathbb{P}(U \geq r)\,\mathrm{d}r$ for every non-negative random variable U with $\mathbb{E}|U| < \infty$; (3.4.39) holds by the assumption (3.4.9), which is equivalent to (3.4.20) for any μ-median of f; consequently, applying the (one-sided) concentration inequality in (3.4.20) to f and $-f$ (having the same Lipschitz constant) yields (3.4.39).

3.4.2 Marton's argument: from transportation to concentration

The concentration of measure phenomenon is fundamentally geometric, as it is captured by the isoperimetric inequality (3.4.9). Once (3.4.9) is established on a given metric probability space $(\mathcal{X}, d, \mu)$, Gaussian concentration is obtained 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 (3.4.9). This technique, first introduced by Marton (see [59] and [75]), hinges on a certain type of inequality that

relates the divergence between two probability measures to a quantity called the *transportation cost*. Let $(\mathcal{X}, d)$ be a Polish space. Given $p \geq 1$, let $\mathcal{P}_p(\mathcal{X})$ denote the space of all Borel probability measures μ on $\mathcal{X}$, such that the moment bound

$$\mathbb{E}_\mu[d^p(X, x_0)] < \infty \tag{3.4.41}$$

holds for some (and hence all) $x_0 \in \mathcal{X}$.

Definition 3.1. Given $p \geq 1$, the L^p *Wasserstein distance* (a.k.a. the *Wasserstein distance of order* p) between $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$ is defined as

$$W_p(\mu, \nu) \triangleq \inf_{\pi \in \Pi(\mu,\nu)} \left(\int_{\mathcal{X}\times\mathcal{X}} d^p(x, y)\, \pi(\mathrm{d}x, \mathrm{d}y) \right)^{1/p}, \tag{3.4.42}$$

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 3.19. Another equivalent way of writing down the definition of $W_p(\mu, \nu)$ in (3.4.42) is

$$W_p(\mu, \nu) = \inf_{X\sim\mu,\, Y\sim\nu} \mathbb{E}^{1/p}[d^p(X, Y)] \tag{3.4.43}$$

where the infimum in the right side of (3.4.43) is over all pairs (X, Y) of jointly distributed random variables taking values in $\mathcal{X}$ such that $P_X = \mu$ and $P_Y = \nu$.

The name "transportation cost" is interpreted as follows: let μ and ν represent, respectively, the initial and desired distributions 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 a 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

$$C^*(\mu, \nu) \triangleq \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} c(x, y)\, \pi(\mathrm{d}x, \mathrm{d}y). \tag{3.4.44}$$

If the cost function is given by $c = d^p$ for $p \geq 1$ and a metric d defined on $\mathcal{X} \times \mathcal{X}$, then it follows from (3.4.43) and (3.4.44) that

$$C^*(\mu, \nu) = W_p^p(\mu, \nu). \tag{3.4.45}$$

The optimal transportation problem (3.4.44) dates back to a 1781 essay by Gaspard Monge who considered a special case of the problem

$$C_0^*(\mu, \nu) \triangleq \inf_{\varphi\colon \mathcal{X} \to \mathcal{X}} \left\{ \int_{\mathcal{X}} c(x, \varphi(x)) \, \mathrm{d}\mu(x) : \mu \circ \varphi^{-1} = \nu \right\}. \tag{3.4.46}$$

The infimum in the right side of (3.4.46) is over all the *deterministic* transportation policies, i.e., the measurable mappings $\varphi\colon \mathcal{X} \to \mathcal{X}$ such that if $X \sim \mu$, then $Y = \varphi(X) \sim \nu$. The problem (3.4.46), known as the *Monge optimal transportation problem*, does not necessarily admit a solution (incidentally, an optimal mapping does exist in the special case considered by Monge, namely $\mathcal{X} = \mathbb{R}^3$ and $c(x, y) = \|x - y\|$). A stochastic relaxation of (3.4.46), given by (3.4.44), was considered in 1942 by Leonid Kantorovich (see [151] for a recent reprint). The books by Villani [66, 67] are recommended for a detailed historical overview and rigorous treatment of optimal transportation.

Lemma 3.4.3. The following properties are satisfied by the Wasserstein distances:

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 every $\mu, \nu \in \mathcal{P}_q(\mathcal{X})$.
3. W_p metrizes weak convergence plus convergence of p-th 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 every continuous and bounded function $\varphi\colon \mathcal{X} \to \mathbb{R}$.
   (b) For some (and hence all) $x_0 \in \mathcal{X}$,

$$\int_{\mathcal{X}} d^p(x, x_0) \, \mu_n(\mathrm{d}x) \xrightarrow{n \to \infty} \int_{\mathcal{X}} d^p(x, x_0) \, \mu(\mathrm{d}x). \tag{3.4.47}$$

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 $\nu_n \to \nu$ converge weakly in $\mathcal{P}_p(\mathcal{X})$, then $W_p(\mu_n,\nu_n) \to W_p(\mu,\nu)$. However, it is *lower semicontinuous* in the usual weak topology (without the convergence of p-th order moments): if $\mu_n \to \mu$ and $\nu_n \to \nu$ converge weakly, then
$$\liminf_{n\to\infty} W_p(\mu_n,\nu_n) \geq W_p(\mu,\nu). \tag{3.4.48}$$
5. The infimum in (3.4.42) [and therefore in (3.4.43)] is actually a minimum; i.e., there exists an *optimal coupling* $\pi^* \in \Pi(\mu,\nu)$, such that
$$W_p^p(\mu,\nu) = \int_{\mathcal{X}\times\mathcal{X}} d^p(x,y)\,\pi^*(\mathrm{d}x,\mathrm{d}y). \tag{3.4.49}$$
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
$$W_p^p(\mu,\nu) = \mathbb{E}[d^p(X^*,Y^*)]. \tag{3.4.50}$$
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 every ν is given by the deterministic mapping
$$Y = \mathsf{F}_\nu^{-1} \circ \mathsf{F}_\mu(X) \tag{3.4.51}$$
for $X \sim \mu$, where F_μ denotes the cumulative distribution function (cdf) 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., $\mathsf{F}_\nu^{-1}(\alpha) \triangleq \inf\{x \in \mathbb{R}\colon \mathsf{F}_\nu(x) \geq \alpha\}$.

Proof. See [67, Chapter 6]. □

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 every probability measure $\nu \ll \mu$ we have
$$W_p(\mu,\nu) \leq \sqrt{2c\,D(\nu\|\mu)}. \tag{3.4.52}$$

Example 3.2 (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 discrete probability measures μ and ν on $\mathcal{X}$ admits the simple form

$$W_1(\mu,\nu) = \inf_{X\sim\mu, Y\sim\nu} \mathbb{P}[X \neq Y]. \tag{3.4.53}$$

As we next show, (3.4.53) turns out to be the *total variation distance*

$$\|\mu - \nu\|_{\mathrm{TV}} \triangleq \sup_{\mathcal{C}\subseteq\mathcal{X}} |\mu(\mathcal{C}) - \nu(\mathcal{C})|. \tag{3.4.54}$$

Proposition 3.6.

$$W_1(\mu,\nu) = \|\mu - \nu\|_{\mathrm{TV}} \tag{3.4.55}$$

$$= \tfrac{1}{2} \sum_{x\in\mathcal{X}} |\mu(x) - \nu(x)| \tag{3.4.56}$$

(we are slightly abusing notation, writing $\mu(x)$ for the μ-probability of the singleton $\{x\}$).

Proof. Consider a probability measure $\pi \in \Pi(\mu,\nu)$. For every $x \in \mathcal{X}$, $\mu(x) = \sum_{y\in\mathcal{X}} \pi(x,y) \geq \pi(x,x)$, and the same holds by replacing μ with ν. Consequently,

$$\pi(x,x) \leq \min\{\mu(x), \nu(x)\}, \tag{3.4.57}$$

and

$$\mathbb{E}_\pi[d(X,Y)] = \mathbb{E}_\pi[1_{\{X\neq Y\}}] \tag{3.4.58}$$

$$= \mathbb{P}[X \neq Y] \tag{3.4.59}$$

$$= 1 - \sum_{x\in\mathcal{X}} \pi(x,x) \tag{3.4.60}$$

$$\geq 1 - \sum_{x\in\mathcal{X}} \min\{\mu(x), \nu(x)\} \tag{3.4.61}$$

$$= \tfrac{1}{2} \sum_{x\in\mathcal{X}} (\mu(x) + \nu(x)) - \sum_{x\in\mathcal{X}} \min\{\mu(x), \nu(x)\} \tag{3.4.62}$$

$$= \tfrac{1}{2} \sum_{x\in\mathcal{X}} |\mu(x) - \nu(x)| \tag{3.4.63}$$

where (3.4.63) holds due to the equality $\min\{a,b\} = \frac{1}{2}(a+b-|a-b|)$ for all $a, b \in \mathbb{R}$. From (3.4.43) and (3.4.58)–(3.4.63), we get

$$W_1(\mu,\nu) \geq \tfrac{1}{2} \sum_{x \in \mathcal{X}} |\mu(x) - \nu(x)|. \tag{3.4.64}$$

Furthermore, (3.4.61) holds with equality for the probability measure $\pi^*\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which is defined as follows:

$$\begin{aligned} \pi^*(x,y) = \min\{\mu(x),\nu(x)\}\, 1_{\{x=y\}} \\ + \frac{(\mu(x)-\nu(x))1_{\{x\in\mathcal{C}\}}(\nu(y)-\mu(y))1_{\{y\in\mathcal{C}^{\mathrm{c}}\}}}{\mu(\mathcal{C})-\nu(\mathcal{C})} \end{aligned} \tag{3.4.65}$$

with $\mathcal{C} \subseteq \mathcal{X}$ defined as

$$\mathcal{C} \triangleq \{x \in \mathcal{X}\colon \mu(x) \geq \nu(x)\}. \tag{3.4.66}$$

This can be verified by noticing that

$$\pi^*(x,x) = \min\{\mu(x),\, \nu(x)\}, \quad \forall\, x \in \mathcal{X}$$

which, due to (3.4.57), is the necessary and sufficient condition for (3.4.61) to hold with equality; it is also easy to verify that π^* is indeed a probability measure with marginals μ and ν. Since $\pi = \pi^* \in \Pi(\mu,\nu)$ achieves (3.4.58)–(3.4.63) with equality then (3.4.64) is satisfied with equality.

We next prove (3.4.56). For an arbitrary $\mathcal{C} \subseteq \mathcal{X}$,

$$\begin{aligned} \mu(\mathcal{C}) - \nu(\mathcal{C}) &= (1-\mu(\mathcal{C}^{\mathrm{c}})) - (1-\nu(\mathcal{C}^{\mathrm{c}})) \\ &= \nu(\mathcal{C}^{\mathrm{c}}) - \mu(\mathcal{C}^{\mathrm{c}}) \end{aligned} \tag{3.4.67}$$

and, from the triangle inequality,

$$\begin{aligned} &|\mu(\mathcal{C}) - \nu(\mathcal{C})| + |\mu(\mathcal{C}^{\mathrm{c}}) - \nu(\mathcal{C}^{\mathrm{c}})| \\ &\leq \sum_{x\in\mathcal{C}} |\mu(x)-\nu(x)| + \sum_{x\in\mathcal{C}^{\mathrm{c}}} |\mu(x)-\nu(x)| \\ &= \sum_{x\in\mathcal{X}} |\mu(x)-\nu(x)|. \end{aligned} \tag{3.4.68}$$

Combining (3.4.67) and (3.4.68) gives that, for every $A \subseteq \mathcal{X}$,

$$|\mu(\mathcal{C}) - \nu(\mathcal{C})| \leq \tfrac{1}{2} \sum_{x\in\mathcal{X}} |\mu(x)-\nu(x)|. \tag{3.4.69}$$

By taking a supremum on the left side of (3.4.69) over all $\mathcal{C} \subseteq \mathcal{X}$, we get from (3.4.54)

$$\|\mu - \nu\|_{\mathrm{TV}} \leq \tfrac{1}{2} \sum_{x \in \mathcal{X}} \big|\mu(x) - \nu(x)\big|. \tag{3.4.70}$$

Conversely, for the subset $\mathcal{C}$ given in (3.4.66), we get from (3.4.67)

$$\begin{aligned} \mu(\mathcal{C}) - \nu(\mathcal{C}) &= \tfrac{1}{2} \Big[\big(\mu(\mathcal{C}) - \nu(\mathcal{C})\big) + \big(\nu(\mathcal{C}^{\mathrm{c}}) - \mu(\mathcal{C}^{\mathrm{c}})\big) \Big] \\ &= \tfrac{1}{2} \left[\sum_{x \in \mathcal{C}} \big(\mu(x) - \nu(x)\big) + \sum_{x \in \mathcal{C}^{\mathrm{c}}} \big(\nu(x) - \mu(x)\big) \right] \\ &= \tfrac{1}{2} \sum_{x \in \mathcal{X}} \big|\mu(x) - \nu(x)\big|. \end{aligned} \tag{3.4.71}$$

Hence, from (3.4.54) and (3.4.71),

$$\|\mu - \nu\|_{\mathrm{TV}} \geq \tfrac{1}{2} \sum_{x \in \mathcal{X}} \big|\mu(x) - \nu(x)\big|, \tag{3.4.72}$$

yielding (3.4.56) by combining (3.4.70) and (3.4.72). □

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

$$\|\mu - \nu\|_{\mathrm{TV}} \leq \sqrt{\tfrac{1}{2} D(\nu \| \mu)} \tag{3.4.73}$$

can be identified as a $\mathrm{T}_1(\frac{1}{4})$ inequality that holds for every probability measure μ on $\mathcal{X}$.

Remark 3.20. It should be pointed out that the constant $c = \frac{1}{4}$ in Pinsker's inequality (3.4.73) is not necessarily the best possible *for a given distribution* μ. Ordentlich and Weinberger [152] have obtained the following *distribution-dependent* refinement of Pinsker's inequality. Let the function $\varphi \colon [0, \frac{1}{2}] \to \mathbb{R}^+$ be defined by

$$\varphi(p) \triangleq \begin{cases} \left(\dfrac{1}{1-2p}\right) \ln\left(\dfrac{1-p}{p}\right), & \text{if } p \in [0, \frac{1}{2}) \\ 2, & \text{if } p = \frac{1}{2} \end{cases} \tag{3.4.74}$$

(note that $\varphi(p) \to 2$ as $p \uparrow \frac{1}{2}$; $\varphi(p) \to \infty$ as $p \downarrow 0$, and the function φ is monotonically decreasing and convex). Let $\mathcal{X}$ be a discrete set, and let $\mathcal{P}(\mathcal{X})$ be the set of all probability distributions defined on the set $\mathcal{X}$. For every $P \in \mathcal{P}(\mathcal{X})$, let the *balance coefficient* be defined as

$$\pi_P \triangleq \max_{A \subseteq \mathcal{X}} \min\{P(A), 1 - P(A)\}, \tag{3.4.75}$$

which yields $\pi_P \in [0, \frac{1}{2}]$. Then, for every $Q \in \mathcal{P}(\mathcal{X})$,

$$\|P - Q\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{\varphi(\pi_P)} \cdot D(Q\|P)} \tag{3.4.76}$$

(see [152, Theorem 2.1]; related results have been considered in [153]). The smaller is the value of the balance coefficient in (3.4.75), the more significant is the refinement of Pinsker's inequality in (3.4.76). The bound in (3.4.76) is also optimal for a given P in the sense that

$$\varphi(\pi_P) = \inf_{Q \in \mathcal{P}(\mathcal{X})} \frac{D(Q\|P)}{\|P - Q\|^2_{\mathrm{TV}}}. \tag{3.4.77}$$

For instance, if $\mathcal{X} = \{0, 1\}$ and P is the distribution of a Bernoulli(p) random variable, then $\pi_P = \min\{p, 1-p\} \in [0, \frac{1}{2}]$,

$$\varphi(\pi_P) = \begin{cases} \left(\frac{1}{1-2p}\right) \ln\left(\frac{1-p}{p}\right), & \text{if } p \neq \frac{1}{2} \\ 2, & \text{if } p = \frac{1}{2} \end{cases}$$

and, from (3.4.76), for every $Q \in \mathcal{P}(\{0, 1\})$

$$\|P - Q\|_{\mathrm{TV}} \leq \begin{cases} \sqrt{\frac{1-2p}{\ln\left(\frac{1-p}{p}\right)} \cdot D(Q\|P)}, & \text{if } p \neq \frac{1}{2} \\ \sqrt{\frac{1}{2} D(Q\|P)}, & \text{if } p = \frac{1}{2}. \end{cases} \tag{3.4.78}$$

Inequality (3.4.78) provides an upper bound on the total variation distance in terms of the divergence. A bound in the reverse direction cannot hold in general 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.

Marton's procedure for deriving Gaussian concentration from a transportation-cost inequality [59, 75] can be distilled as follows:

Proposition 3.7. Suppose μ satisfies a $\mathrm{T}_1(c)$ inequality. Then, the Gaussian concentration inequality in (3.4.10) holds with $\kappa = \frac{1}{2c}$, $K = 1$, and $r_0 = \sqrt{2c \ln 2}$.

Proof. Fix two Borel sets $\mathcal{C}, \mathcal{D} \subseteq \mathcal{X}$ with $\mu(\mathcal{C}), \mu(\mathcal{D}) > 0$. Define the conditional probability measures

$$\mu_{\mathcal{C}}(\mathcal{E}) \triangleq \frac{\mu(\mathcal{C} \cap \mathcal{E})}{\mu(\mathcal{C})} \qquad \text{and} \qquad \mu_{\mathcal{D}}(\mathcal{E}) \triangleq \frac{\mu(\mathcal{D} \cap \mathcal{E})}{\mu(\mathcal{D})},$$

where $\mathcal{E}$ is an arbitrary Borel subset of $\mathcal{X}$. Then $\mu_{\mathcal{C}}, \mu_{\mathcal{D}} \ll \mu$, and

$$W_1(\mu_{\mathcal{C}}, \mu_{\mathcal{D}}) \leq W_1(\mu, \mu_{\mathcal{C}}) + W_1(\mu, \mu_{\mathcal{D}}) \tag{3.4.79}$$

$$\leq \sqrt{2cD(\mu_{\mathcal{C}} \| \mu)} + \sqrt{2cD(\mu_{\mathcal{D}} \| \mu)}, \tag{3.4.80}$$

where (3.4.79) is by the triangle inequality, while (3.4.80) is because μ satisfies $\mathrm{T}_1(c)$. Now, for an arbitrary Borel subset $\mathcal{E} \subseteq \mathcal{X}$, we have

$$\mu_{\mathcal{C}}(\mathcal{E}) = \int_{\mathcal{E}} \frac{1_{\mathcal{C}}(x)}{\mu(\mathcal{C})} \, \mathrm{d}\mu(x),$$

so it follows that $\frac{\mathrm{d}\mu_{\mathcal{C}}}{\mathrm{d}\mu} = \frac{1_{\mathcal{C}}}{\mu(\mathcal{C})}$, and the same holds for $\mu_{\mathcal{D}}$. Therefore,

$$D(\mu_{\mathcal{C}} \| \mu) = \mathbb{E}_{\mu}\left[\frac{\mathrm{d}\mu_{\mathcal{C}}}{\mathrm{d}\mu} \ln \frac{\mathrm{d}\mu_{\mathcal{C}}}{\mathrm{d}\mu}\right] = \ln \frac{1}{\mu(\mathcal{C})}, \tag{3.4.81}$$

and an analogous formula holds for $\mu_{\mathcal{D}}$ in place of $\mu_{\mathcal{C}}$. Substituting this into (3.4.80) gives

$$W_1(\mu_{\mathcal{C}}, \mu_{\mathcal{D}}) \leq \sqrt{2c \ln \frac{1}{\mu(\mathcal{C})}} + \sqrt{2c \ln \frac{1}{\mu(\mathcal{D})}}. \tag{3.4.82}$$

We now obtain a lower bound on $W_1(\mu_{\mathcal{C}}, \mu_{\mathcal{D}})$. Since the probability measures $\mu_{\mathcal{C}}$ and $\mu_{\mathcal{D}}$ are, respectively, supported on $\mathcal{C}$ and $\mathcal{D}$ then every $\pi \in \Pi(\mu_{\mathcal{C}}, \mu_{\mathcal{D}})$ is supported on $\mathcal{C} \times \mathcal{D}$. Consequently, for every

such π,

$$\begin{aligned}\int_{\mathcal{X}\times\mathcal{X}} d(x,y)\,\mathrm{d}\pi(x,y) &= \int_{\mathcal{C}\times\mathcal{D}} d(x,y)\,\mathrm{d}\pi(x,y)\\ &\geq \int_{\mathcal{C}\times\mathcal{D}} \inf_{y\in\mathcal{D}} d(x,y)\,\mathrm{d}\pi(x,y)\\ &= \int_{\mathcal{C}} d(x,\mathcal{D})\,\mathrm{d}\mu_{\mathcal{C}}(x)\\ &\geq \inf_{x\in\mathcal{C}} d(x,\mathcal{D})\;\mu_{\mathcal{C}}(\mathcal{C})\\ &= d(\mathcal{C},\mathcal{D}), \end{aligned} \tag{3.4.83}$$

where (3.4.83) holds since $\mu_{\mathcal{C}}(\mathcal{C}) = 1$, and $d(\mathcal{C},\mathcal{D}) \triangleq \inf_{x\in\mathcal{C},\,y\in\mathcal{D}} d(x,y)$. Since (3.4.83) holds for all $\pi \in \Pi(\mu_{\mathcal{C}},\mu_{\mathcal{D}})$, we can take the infimum over all such π to get

$$W_1(\mu_{\mathcal{C}},\mu_{\mathcal{D}}) \geq d(\mathcal{C},\mathcal{D}). \tag{3.4.84}$$

Combining (3.4.82) with (3.4.84) gives the inequality

$$d(\mathcal{C},\mathcal{D}) \leq \sqrt{2c\ln\frac{1}{\mu(\mathcal{C})}} + \sqrt{2c\ln\frac{1}{\mu(\mathcal{D})}}, \tag{3.4.85}$$

which holds for all Borel sets $\mathcal{C}$ and $\mathcal{D}$ having positive μ-probabilities.

Let $\mathcal{D} = (\mathcal{C}_r)^{\mathrm{c}}$, then

$$\mu(\mathcal{D}) = 1 - \mu(\mathcal{C}_r), \tag{3.4.86}$$

$$d(\mathcal{C},\mathcal{D}) \geq r. \tag{3.4.87}$$

Consequently, combining (3.4.85)–(3.4.87) gives

$$r \leq \sqrt{2c\ln\frac{1}{\mu(\mathcal{C})}} + \sqrt{2c\ln\frac{1}{1-\mu(\mathcal{C}_r)}}. \tag{3.4.88}$$

If $\mu(A) \geq 1/2$ and $r \geq \sqrt{2c\ln 2}$, then (3.4.88) gives

$$\mu(\mathcal{C}_r) \geq 1 - \exp\left(-\frac{1}{2c}\left(r - \sqrt{2c\ln 2}\right)^2\right). \tag{3.4.89}$$

Hence, the Gaussian concentration inequality in (3.4.10) indeed holds with $\kappa = \frac{1}{2c}$ and $K = 1$ for all $r \geq r_0 = \sqrt{2c\ln 2}$. □

Remark 3.21. The exponential inequality (3.4.89) has appeared earlier in the work of McDiarmid [93] 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 (3.4.85).

Remark 3.22. The formula (3.4.81), apparently first used explicitly by Csiszár [154, Eq. (4.13)], is actually quite remarkable: it states that the probability of an arbitrary event can be expressed as an exponential of a divergence.

While the method described in the proof of Proposition 3.7 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 result that, for $p \in [1,2]$, the T_p inequalities tensorize (see, for example, [67, Proposition 22.5]):

Proposition 3.8 (Tensorization of transportation-cost inequalities)**.** If μ satisfies $\mathrm{T}_p(c)$ on $(\mathcal{X}, d)$ for an arbitrary $p \in [1,2]$, then, for every $n \in \mathbb{N}$, the product measure $\mu^{\otimes n}$ satisfies $\mathrm{T}_p(cn^{2/p-1})$ on $(\mathcal{X}^n, d_{p,n})$ with the metric

$$d_{p,n}(x^n, y^n) \triangleq \left(\sum_{i=1}^n d^p(x_i, y_i)\right)^{1/p}, \qquad \forall\, x^n, y^n \in \mathcal{X}^n. \tag{3.4.90}$$

Proof. Suppose μ satisfies $\mathrm{T}_p(c)$. For $n \in \mathbb{N}$, fix 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

$$P_{X^n} = P_{X_1} \otimes P_{X_2|X_1} \otimes \ldots \otimes P_{X_n|X^{n-1}} = \nu \tag{3.4.91}$$

$$P_{Y^n} = P_{Y_1} \otimes P_{Y_2} \otimes \ldots \otimes P_{Y_n} = \mu^{\otimes n}. \tag{3.4.92}$$

For each $i \in \{1, \ldots, n\}$, let the conditional W_p distance be defined as follows:

$$\begin{aligned} &W_p(P_{X_i|X^{i-1}}, P_{Y_i}|P_{X^{i-1}}) \\ &\triangleq \left(\int_{\mathcal{X}^{i-1}} W_p^p(P_{X_i|X^{i-1}=x^{i-1}}, P_{Y_i})\, \mathrm{d}P_{X^{i-1}}(x^{i-1})\right)^{1/p}. \end{aligned} \tag{3.4.93}$$

We next prove that

$$W_p^p(\nu, \mu^{\otimes n}) = W_p^p(P_{X^n}, P_{Y^n}) \\ \leq \sum_{i=1}^{n} W_p^p(P_{X_i|X^{i-1}}, P_{Y_i}|P_{X^{i-1}}), \tag{3.4.94}$$

where the L^p Wasserstein distance on the left 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

$$W_p^p(P_{X_1}, P_{Y_1}) = \mathbb{E}[d^p(X_1^*, Y_1^*)]. \tag{3.4.95}$$

For $i \in \{2, \ldots, n\}$ and for every 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

$$P_{X_i^*(x^{i-1})} = P_{X_i|X^{i-1}=x^{i-1}}, \quad P_{Y_i^*(x^{i-1})} = P_{Y_i}, \tag{3.4.96}$$

and

$$W_p^p(P_{X_i|X^{i-1}=x^{i-1}}, P_{Y_i}) = \mathbb{E}[d^p(X_i^*(x^{i-1}), Y_i^*(x^{i-1}))]. \tag{3.4.97}$$

Moreover, since by assumption $(\mathcal{X}, d)$ is a Polish space, all couplings can be constructed in such a way that the mapping

$$x^{i-1} \mapsto \mathbb{P}\big((X_i^*(x^{i-1}), Y_i^*(x^{i-1})) \in \mathcal{C}\big)$$

is measurable for each Borel set $\mathcal{C} \subseteq \mathcal{X} \times \mathcal{X}$ [67]. 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 for all $x^{i-1} \in \mathcal{X}^{i-1}$ and $i \in \{1, \ldots, n\}$

$$W_p^p(P_{X_i|X^{i-1}=x^{i-1}}, P_{Y_i}) = \mathbb{E}[d^p(X_i^*, Y_i^*)|X^{*(i-1)} = x^{i-1}]. \tag{3.4.98}$$

We have

$$W_p^p(\nu, \mu^{\otimes n}) \le \mathbb{E}[d_{p,n}^p(X^{*n}, Y^{*n})] \tag{3.4.99}$$

$$= \sum_{i=1}^n \mathbb{E}[d^p(X_i^*, Y_i^*)] \tag{3.4.100}$$

$$= \sum_{i=1}^n \mathbb{E}\Big[\mathbb{E}\big[d^p(X_i^*, Y_i^*)|X^{*(i-1)}\big]\Big] \tag{3.4.101}$$

$$= \sum_{i=1}^n W_p^p(P_{X_i|X^{i-1}}, P_{Y_i}|P_{X^{i-1}}), \tag{3.4.102}$$

where

- (3.4.99) is due to the facts that $W_p(\nu, \mu^{\otimes n})$ is the L^p Wasserstein distance with respect to the $d_{p,n}$ metric, and (X^{*n}, Y^{*n}) is a (not necessarily optimal) coupling of $P_{X^n} = \nu$ and $P_{Y^n} = \mu^{\otimes n}$;
- (3.4.100) is by the definition (3.4.90) of $d_{p,n}$;
- (3.4.101) is by the law of iterated expectations; and
- (3.4.102) is by (3.4.93) and (3.4.98).

We have thus proved (3.4.94). By hypothesis, μ satisfies $\mathrm{T}_p(c)$ on $(\mathcal{X}, d)$. Therefore, since $P_{Y_i} = \mu$ for every i, we can write

$$\begin{aligned}
&W_p^p(P_{X_i|X^{i-1}}, P_{Y_i}|P_{X^{i-1}}) \\
&= \int_{\mathcal{X}^{i-1}} W_p^p(P_{X_i|X^{i-1}=x^{i-1}}, P_{Y_i})\, \mathrm{d}P_{X^{i-1}}(x^{i-1}) \\
&\le \int_{\mathcal{X}^{i-1}} \Big(2cD(P_{X_i|X^{i-1}=x^{i-1}}\|P_{Y_i})\Big)^{p/2}\, \mathrm{d}P_{X^{i-1}}(x^{i-1}) \\
&\le (2c)^{p/2} \left(\int_{\mathcal{X}^{i-1}} D(P_{X_i|X^{i-1}=x^{i-1}}\|P_{Y_i})\, \mathrm{d}P_{X^{i-1}}(x^{i-1})\right)^{p/2} \qquad (3.4.103) \\
&= (2c)^{p/2} \Big(D(P_{X_i|X^{i-1}}\|P_{Y_i}|P_{X^{i-1}})\Big)^{p/2}, \qquad (3.4.104)
\end{aligned}$$

where (3.4.103) 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

$$W_p^p(\nu, \mu^{\otimes n}) \leq (2c)^{p/2} \sum_{i=1}^n \left(D(P_{X_i|X^{i-1}} \| P_{Y_i} | P_{X^{i-1}}) \right)^{p/2} \tag{3.4.105}$$

$$\leq (2c)^{p/2} n^{1-p/2} \left(\sum_{i=1}^n D(P_{X_i|X^{i-1}} \| P_{Y_i} | P_{X^{i-1}}) \right)^{p/2} \tag{3.4.106}$$

$$= (2c)^{p/2} n^{1-p/2} \left(D(P_{X^n} \| P_{Y^n}) \right)^{p/2} \tag{3.4.107}$$

$$= (2c)^{p/2} n^{1-p/2} \left(D(\nu \| \mu^{\otimes n}) \right)^{p/2}, \tag{3.4.108}$$

where (3.4.105) is due to (3.4.99)–(3.4.104); (3.4.106) follows from Hölder's inequality; (3.4.107) is by the chain rule for the divergence and since P_{Y^n} is a product probability measure; (3.4.108) is by (3.4.91) and (3.4.92). This finally gives

$$W_p(\nu, \mu^{\otimes n}) \leq \sqrt{2cn^{2/p-1} D(\nu \| \mu^{\otimes n})},$$

i.e., $\mu^{\otimes n}$ satisfies the $\mathrm{T}_p(cn^{2/p-1})$ inequality. □

Since the metric W_2 dominates W_1 (see 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 3.8 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 next two sections, we 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 $\kappa = \frac{1}{2c}$. 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 3.8 and consequently also apply Proposition 3.7. If we go through with this approach, we show 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 $\kappa_n = \frac{1}{2cn}$. Consequently, 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 $\kappa = \frac{1}{2c}$ *independently of the dimension* n.

These two results give the same constants in concentration inequalities for sums of independent random variables: to this end, note that by (3.4.90) and the Cauchy-Schwarz inequality

$$\begin{aligned} d_{1,n}(x^n, y^n) &= \sum_{i=1}^n d(x_i, y_i) \\ &\le \sqrt{n} \left(\sum_{i=1}^n d^2(x_i, y_i) \right)^{\frac{1}{2}} \\ &= \sqrt{n}\, d_{2,n}(x^n, y^n). \end{aligned} \tag{3.4.109}$$

Let $f \colon \mathcal{X} \to \mathbb{R}$ be a Lipschitz function on $(\mathcal{X}, d)$, and let $f_n \colon \mathcal{X}^n \to \mathbb{R}$ be defined as

$$f_n(x^n) \triangleq \frac{1}{n} \sum_{i=1}^n f(x_i), \quad x^n \in \mathcal{X}^n. \tag{3.4.110}$$

Then, we can conclude that

$$\begin{aligned} \|f_n\|_{\mathrm{Lip},1} &\triangleq \sup_{x^n \neq y^n} \frac{|f_n(x^n) - f_n(y^n)|}{d_{1,n}(x^n, y^n)} \\ &\le \frac{1}{n} \sup_{x^n \neq y^n} \frac{\sum_{i=1}^n |f(x_i) - f(y_i)|}{\sum_{i=1}^n d(x_i, y_i)} \\ &\le \frac{1}{n} \sup_{u \neq v} \frac{|f(u) - f(v)|}{d(u, v)} \\ &= \frac{\|f\|_{\mathrm{Lip}}}{n}, \end{aligned} \tag{3.4.111}$$

and, from (3.4.109) and (3.4.111),

$$\begin{aligned}\|f_n\|_{\mathrm{Lip},2} &\triangleq \sup_{x^n \neq y^n} \frac{|f_n(x^n) - f_n(y^n)|}{d_{2,n}(x^n, y^n)} \\ &\leq \sqrt{n} \sup_{x^n \neq y^n} \frac{|f_n(x^n) - f_n(y^n)|}{d_{1,n}(x^n, y^n)} \\ &= \sqrt{n}\, \|f_n\|_{\mathrm{Lip},1} \leq \frac{\|f\|_{\mathrm{Lip}}}{\sqrt{n}}. \end{aligned} \tag{3.4.112}$$

Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{X}$-valued random variables whose common marginal is a probability measure μ, satisfying either $\mathrm{T}_1(c)$ or $\mathrm{T}_2(c)$. In view of (3.4.111), (3.4.112), and Corollary 3.4.5 in the next section, both $\mathrm{T}_1(c)$ and $\mathrm{T}_2(c)$ yield

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) \geq \mathbb{E}[f(X_1)] + r\right) \leq \exp\left(-\frac{nr^2}{2c\|f\|^2_{\mathrm{Lip}}}\right), \quad \forall\, r > 0. \tag{3.4.113}$$

However, in general, the difference between concentration inequalities that are derived from T_1 and T_2 inequalities becomes quite pronounced. Note that, in practice, it is often easier to work with T_1 inequalities.

The same strategy as above can be used to prove the following generalization of Proposition 3.8:

Proposition 3.9. 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 $i \in \{1, \ldots, n\}$. Let $c \triangleq \max_{1 \leq i \leq n} c_i$. Then, for an arbitrary $p \in [1, 2]$, the probability measure $\mu = \mu_1 \otimes \ldots \otimes \mu_n$ satisfies $\mathrm{T}_p(cn^{2/p-1})$ on $(\mathcal{X}^n, d_{p,n})$ (with the metric $d_{p,n}$ in (3.4.90)).

3.4.3 Gaussian concentration and T_1 inequalities

As we saw in Proposition 3.7, Marton's argument can be used to deduce Gaussian concentration from a transportation-cost inequality. As we demonstrate here and in the following section, in certain cases these properties are *equivalent*. We first consider the case where μ satisfies a T_1 inequality. The first proof of the equivalence between T_1 and Gaussian concentration was obtained by Bobkov and Götze [54], and it

relies on the following variational representations of the L^1 Wasserstein distance and the divergence:

1. **Kantorovich–Rubinstein theorem** [67, Theorem 5.10]: For every $\mu, \nu \in \mathcal{P}_1(\mathcal{X})$ on a Polish probability space $(\mathcal{X}, d)$,
$$W_1(\mu, \nu) = \sup_{f\colon \|f\|_{\mathrm{Lip}} \leq 1} \Big| \mathbb{E}_\mu[f] - \mathbb{E}_\nu[f] \Big|. \tag{3.4.114}$$

2. **Donsker–Varadhan lemma** [85, Lemma 6.2.13]: For every 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:
$$D(\nu\|\mu) = \sup_{g \in \mathcal{C}_{\mathrm{b}}(\mathcal{X})} \big\{ \mathbb{E}_\nu[g] - \ln \mathbb{E}_\mu[\exp(g)] \big\} \tag{3.4.115}$$
where the supremization in (3.4.115) is over the set $\mathcal{C}_{\mathrm{b}}(\mathcal{X})$ of all continuous and bounded real-valued functions on $\mathcal{X}$. Moreover, for every measurable function g such that $\mathbb{E}_\mu[\exp(g)] < \infty$,
$$\mathbb{E}_\nu[g] \leq D(\nu\|\mu) + \ln \mathbb{E}_\mu[\exp(g)]. \tag{3.4.116}$$
(In fact, the supremum in (3.4.115) can be extended to bounded Borel-measurable functions g [155, Lemma 1.4.3].)

The following theorem was introduced by Bobkov and Götze [54, Theorem 3.1]:

Theorem 3.4.4. Let $\mu \in \mathcal{P}_1(\mathcal{X})$ be a Borel probability measure, and suppose that there exists $x_0 \in \mathcal{X}$ such that $\mathbb{E}_\mu[d(X, x_0)] < \infty$. Then, μ satisfies $\mathrm{T}_1(c)$ if and only if the inequality
$$\mathbb{E}_\mu \left\{ \exp[t f(X)] \right\} \leq \exp\left(\frac{c t^2}{2} \right) \tag{3.4.117}$$
holds for all 1-Lipschitz functions $f\colon \mathcal{X} \to \mathbb{R}$ with $\mathbb{E}_\mu[f(X)] = 0$, and all $t \in \mathbb{R}$.

Proof. The condition $\mathbb{E}_\mu[d(X, x_0)] < \infty$, for some $x_0 \in \mathcal{X}$, ensures that every Lipschitz function $f\colon \mathcal{X} \to \mathbb{R}$ is μ-integrable:
$$\begin{aligned} \mathbb{E}_\mu\big[|f(X)|\big] &\leq |f(x_0)| + \mathbb{E}_\mu\big[|f(X) - f(x_0)|\big] \\ &\leq |f(x_0)| + \|f\|_{\mathrm{Lip}}\, \mathbb{E}_\mu\big[d(X, x_0)\big] < \infty. \end{aligned} \tag{3.4.118}$$

Without loss of generality, we may consider (3.4.117) only for $t \geq 0$ (otherwise, replace f with $-f$ for $t < 0$).

Suppose first that μ satisfies $\mathrm{T}_1(c)$, and let $\nu \ll \mu$. Using the $\mathrm{T}_1(c)$ property of μ together with the Kantorovich–Rubinstein formula (3.4.114), we can write

$$\int_{\mathcal{X}} f \,\mathrm{d}\nu \leq W_1(\mu, \nu) \leq \sqrt{2cD(\nu\|\mu)}$$

for every 1-Lipschitz $f\colon \mathcal{X} \to \mathbb{R}$ with $\mathbb{E}_\mu[f] = 0$. Next, since

$$\inf_{t>0} \left(\frac{a}{t} + \frac{bt}{2} \right) = \sqrt{2ab} \tag{3.4.119}$$

for every $a, b \geq 0$, we see that every such f satisfies

$$\int_{\mathcal{X}} f \,\mathrm{d}\nu \leq \frac{D(\nu\|\mu)}{t} + \frac{ct}{2}, \qquad \forall\, t > 0.$$

Rearranging, we obtain

$$\int_{\mathcal{X}} tf \,\mathrm{d}\nu - \frac{ct^2}{2} \leq D(\nu\|\mu), \qquad \forall\, t \geq 0. \tag{3.4.120}$$

Applying this inequality to $\nu = \mu^{(g)}$ (the g-tilting of μ) where $g \triangleq tf$, and using the fact that

$$\begin{aligned} D(\mu^{(g)}\|\mu) &= \int_{\mathcal{X}} g \,\mathrm{d}\mu^{(g)} - \ln \int_{\mathcal{X}} \exp(g) \,\mathrm{d}\mu \\ &= \int_{\mathcal{X}} tf \,\mathrm{d}\nu - \ln \int_{\mathcal{X}} \exp(tf) \,\mathrm{d}\mu \end{aligned} \tag{3.4.121}$$

we deduce from (3.4.120) and (3.4.121) that

$$\ln \left(\int_{\mathcal{X}} \exp(tf) \,\mathrm{d}\mu \right) \leq \frac{ct^2}{2}$$

for all $t \geq 0$, and f with $\|f\|_{\mathrm{Lip}} \leq 1$ and $\mathbb{E}_\mu[f] = 0$, which is (3.4.117).

Conversely, suppose that μ satisfies (3.4.117) 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 an arbitrary function of the form $g \triangleq tf$ with $t > 0$. By the assumption

in (3.4.117), $\mathbb{E}_\mu[\exp(g)] < \infty$; furthermore, g is a Lipschitz function, so it is also measurable. Hence, (3.4.116) gives

$$\begin{aligned} D(\nu\|\mu) &\geq \int_{\mathcal{X}} tf \,\mathrm{d}\nu - \ln \int_{\mathcal{X}} \exp(tf) \,\mathrm{d}\mu \\ &\geq \int_{\mathcal{X}} tf \,\mathrm{d}\nu - \int_{\mathcal{X}} tf \,\mathrm{d}\mu - \frac{ct^2}{2} \end{aligned} \tag{3.4.122}$$

where (3.4.122) relies on (3.4.117) and the assumption that $\int_{\mathcal{X}} f \,\mathrm{d}\mu = 0$. Rearranging gives

$$\left| \int_{\mathcal{X}} f \,\mathrm{d}\nu - \int_{\mathcal{X}} f \,\mathrm{d}\mu \right| \leq \frac{D(\nu\|\mu)}{t} + \frac{ct}{2}, \qquad \forall\, t > 0 \tag{3.4.123}$$

(the absolute value in the left side of (3.4.123) is a consequence of the fact that the same argument goes through with $-f$ replaced by f). Minimizing the right side of (3.4.123) over $t > 0$ and applying (3.4.119), we get that the inequality

$$\left| \int_{\mathcal{X}} f \,\mathrm{d}\nu - \int_{\mathcal{X}} f \,\mathrm{d}\mu \right| \leq \sqrt{2cD(\nu\|\mu)} \tag{3.4.124}$$

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 functions f on the left side of (3.4.124) and using the Kantorovich–Rubinstein formula (3.4.114), we conclude that $W_1(\mu,\nu) \leq \sqrt{2cD(\nu\|\mu)}$ 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 inequalities for Lipschitz functions:

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 $f \in \mathcal{A}$:

$$\mathbb{P}\Big(f(X) \geq \mathbb{E}_\mu[f(X)] + r\Big) \leq \exp\left(-\frac{r^2}{2c\|f\|_{\mathrm{Lip}}^2}\right), \qquad \forall\, r > 0. \tag{3.4.125}$$

Proof. The result follows from the Chernoff bound and (3.4.117). □

As another illustration of the use of T_1 inequalities, the following concentration inequality is proved (cf. Kearns–Saul inequality [97], see Corollary 2.4.8):

Theorem 3.4.6. Let $\mathcal{X}$ be the Hamming space $\{0,1\}^n$, equipped with the metric

$$d_n(x^n, y^n) = \sum_{i=1}^{n} 1_{\{x_i \neq y_i\}}. \tag{3.4.126}$$

Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli(p) random variables. Then, for every Lipschitz function $f\colon \{0,1\}^n \to \mathbb{R}$ and $r > 0$,

$$\mathbb{P}\left(f(X^n) - \mathbb{E}[f(X^n)] \geq r\right) \leq \exp\left(-\frac{\ln\left(\frac{1-p}{p}\right) r^2}{n\|f\|_{\mathrm{Lip}}^2 (1-2p)}\right). \tag{3.4.127}$$

Remark 3.23. By letting $p \to \frac{1}{2}$, the right side of (3.4.127) tends to $\exp\left(-\frac{2r^2}{n\|f\|_{\mathrm{Lip}}^2}\right)$ for all $r > 0$.

Proof. Taking into account Remark 3.23, we may assume without loss of generality that $p \neq \frac{1}{2}$. From the distribution-dependent refinement of Pinsker's inequality (3.4.78), it follows that the Bernoulli(p) measure satisfies $\mathrm{T}_1\left(\frac{1}{2\varphi(p)}\right)$ with respect to the Hamming metric, where $\varphi(\cdot)$ is defined in (3.4.74). By Proposition 3.8, the probability measure of a sum of n independent Bernoulli(p) random variables satisfies $\mathrm{T}_1\left(\frac{n}{2\varphi(p)}\right)$ with respect to the metric (3.4.126). The bound (3.4.127) then follows from Corollary 3.4.5. □

Remark 3.24. If $\|f\|_{\mathrm{Lip}} \leq \frac{c}{n}$ for an arbitrary $c > 0$, then (3.4.127) implies that for every $r > 0$

$$\mathbb{P}\left(f(X^n) - \mathbb{E}[f(X^n)] \geq r\right) \leq \exp\left(-\frac{\ln\left(\frac{1-p}{p}\right)}{c^2(1-2p)} \cdot nr^2\right). \tag{3.4.128}$$

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 $i \in \{1, \ldots, n\}$. More generally, every f satisfying (3.3.46) with $c_i = c_i'/n$, $i \in \{1, \ldots, n\}$, for some constants $c_1', \ldots, c_n' \geq 0$, satisfies (3.4.128) for all $r > 0$ with $c = \max_{1 \leq i \leq n} c_i'$.

In the following, we provide Marton's coupling inequality, which forms a slightly stronger form of the original result of Marton [75] (see [2, Theorem 8.2] for the following stronger statement):

Theorem 3.4.7 (Marton's coupling inequality)**.** Let $\mu = \mu_1 \otimes \ldots \mu_n$ be a product probability measure of $X^n \in \mathcal{X}^n$, and let ν (where $\nu \ll \mu$) be a probability measure of $Y^n \in \mathcal{X}^n$. Then,

$$\min_{\pi \in \Pi(\mu,\nu)} \sum_{i=1}^{n} \mathbb{P}^2(X_i \neq Y_i) \leq \tfrac{1}{2}\, D(\nu \| \mu) \tag{3.4.129}$$

where the relative entropy is expressed in nats.

Proof. See [2, p. 241]. □

We provide in the following an alternative proof of McDiarmid's inequality (3.3.47), based on the earlier material in this chapter about transportation-cost inequalities (recall the two previous proofs of this inequality in Sections 2.4.3 and 3.3.4).

An alternative proof of McDiarmid's inequality: For every $n \in \mathbb{N}$, constants $c_1, \ldots, c_n > 0$, and a measurable space $\mathcal{X}$, let us equip the product space $\mathcal{X}^n$ with the weighted Hamming metric

$$d_n(x^n, y^n) \triangleq \sum_{i=1}^{n} c_i 1_{\{x_i \neq y_i\}}. \tag{3.4.130}$$

Let $f \colon \mathcal{X}^n \to \mathbb{R}$ be a Lipschitz function (with respect to the metric d), and suppose that it satisfies the condition of the bounded differences in (3.3.46). The corresponding Lipschitz constant $\|f\|_{\mathrm{Lip}}$ is given by

$$\|f\|_{\mathrm{Lip}} = \sup_{x^n \neq y^n} \frac{|f(x^n) - f(y^n)|}{d_n(x^n, y^n)}. \tag{3.4.131}$$

It is easy to verify that the condition $\|f\|_{\mathrm{Lip}} \leq 1$ is equivalent to the condition in (3.3.46).

Let $\mu_1, \ldots, \mu_n$ be arbitrary n probability measures on $\mathcal{X}$, and let $\mu = \mu_1 \otimes \ldots \mu_n$ be a product probability measure of $X^n \in \mathcal{X}^n$. Let ν

be an arbitrary (not necessarily a product) probability measure on $\mathcal{X}^n$, where $\nu \ll \mu$, and let Y^n be a random vector that is drawn from ν.

$$\begin{aligned}
&\Big|\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f]\Big| \\
&= \Big|\mathbb{E}\big[f(X^n) - f(Y^n)\big]\Big| && (3.4.132)\\
&\le \mathbb{E}\Big|f(X^n) - f(Y^n)\Big| && (3.4.133)\\
&\le \sum_{i=1}^n \mathbb{E}\big[c_i\, 1_{\{X_i \neq Y_i\}}\big] && (3.4.134)\\
&\le \left(\sum_{i=1}^n c_i^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^n \mathbb{E}^2\big[1_{\{X_i \neq Y_i\}}\big]\right)^{\frac{1}{2}} && (3.4.135)\\
&= \left(\sum_{i=1}^n c_i^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^n \mathbb{P}^2(X_i \neq Y_i)\right)^{\frac{1}{2}}, && (3.4.136)
\end{aligned}$$

where (3.4.132) holds since, by construction, $X^n \sim \mu$ and $Y^n \sim \nu$; (3.4.133) holds since $\big|\mathbb{E}[Z]\big| \le \mathbb{E}|Z|$; (3.4.134) follows from the bounded differences condition in (3.3.46), combined the triangle inequality; (3.4.135) holds by using the Cauchy-Schwarz inequality; (3.4.136) holds because the expectation of the indicator function of an event is equal to the probability of the event. By minimizing the right side of (3.4.136) over all the couplings $\pi \in \Pi(\mu, \nu)$, it follows from (3.4.129) that

$$\Big|\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f]\Big| \le \sqrt{\frac{1}{2}\sum_{i=1}^n c_i^2\; D(\nu\|\mu)}. \tag{3.4.137}$$

By supremizing the left side of (3.4.137) over all the Lipschitz functions $f\colon \mathcal{X}^n \to \mathbb{R}$ such that $\|f\|_{\mathrm{Lip}} \le 1$, it follows from the Kantorovich–Rubinstein theorem (see (3.4.114)) that

$$W_1(\mu, \nu) \le \sqrt{\frac{1}{2}\sum_{i=1}^n c_i^2\; D(\nu\|\mu)}\,. \tag{3.4.138}$$

Hence, from (3.4.138), μ satisfies $\mathrm{T}_1(c)$ (with the weighted Hamming metric d_n in (3.4.130)) where $c = \frac{1}{4}\sum_{i=1}^n c_i^2$. From Corollary 3.4.5, this provides an alternative proof of McDiarmid's inequality (3.3.47) which is different from the martingale approach (see Section 2.4.3).

3.4.4 Dimension-free Gaussian concentration and T_2 inequalities

Our discussion so far has been mostly confined to 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 3.8 demonstrates 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 3.8 and the discussion following it, T_1 inequalities are not very useful in this regard, since the resulting concentration inequalities 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

$$\mathbb{P}\Big(f(X^n) \geq \mathbb{E}f(X^n) + r\Big) \leq K \exp\left(-\frac{r^2}{2cn}\right), \quad r > 0 \tag{3.4.139}$$

for every $f\colon \mathcal{X}^n \to \mathbb{R}$ 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 is, the more we will need to "inflate" a given set $\mathcal{C} \subset \mathcal{X}^n$ 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 every $k \in \mathbb{N}$ and $r > 0$,

$$\mathcal{C} \subseteq \mathcal{X}^k, \quad \mu^{\otimes k}(\mathcal{C}) \geq \tfrac{1}{2} \quad \Longrightarrow \quad \mu^{\otimes k}(\mathcal{C}_r) \geq 1 - Ke^{-\kappa r^2} \tag{3.4.140}$$

where the isoperimetric enlargement $\mathcal{C}_r$ of a Borel set $\mathcal{C} \subseteq \mathcal{X}^k$ is defined in (3.4.7) with respect to the metric $d_{2,k}$ defined according to (3.4.90), i.e.,

$$\mathcal{C}_r \triangleq \left\{y^k \in \mathcal{X}^k\colon \exists\, x^k \in \mathcal{C} \text{ such that } \sum_{i=1}^k d^2(x_i, y_i) < r^2\right\}. \tag{3.4.141}$$

Remark 3.25. As before, we are mainly interested in the constant κ in the exponent. Thus, it is said that μ has dimension-free Gaussian

concentration with constant $\kappa > 0$ if (3.4.140) holds with that κ and some $K > 0$.

Remark 3.26. In the same spirit as Remark 3.16, it may be desirable to relax (3.4.140) to the following: there exists some $r_0 > 0$ such that, for every $k \in \mathbb{N}$ and $r \geq r_0$,

$$\mathcal{C} \subseteq \mathcal{X}^k, \quad \mu^{\otimes k}(\mathcal{C}) \geq \tfrac{1}{2} \quad \Longrightarrow \quad \mu^{\otimes k}(A_r) \geq 1 - Ke^{-\kappa(r-r_0)^2} \tag{3.4.142}$$

(see, for example, [67, Remark 22.23] or [71, Proposition 3.3]). The same considerations about (possibly) sharper constants that were stated in Remark 3.16 also apply here.

In this section, we show that dimension-free Gaussian concentration and T_2 inequalities are equivalent. Before we get to that, here is an example of a T_2 inequality:

Theorem 3.4.8 (Talagrand [156])**.** 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 with the one-dimensional case where $n = 1$: 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.,

$$\begin{aligned} \Phi(x) &= \int_{-\infty}^{x} \gamma(y)\mathrm{d}y \\ &= \tfrac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\tfrac{y^2}{2}\right) \mathrm{d}y, \quad \forall\, x \in \mathbb{R}. \end{aligned} \tag{3.4.143}$$

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

$$W_2(\nu, \mu) = W_2(\nu, G) = \inf_{X \sim G,\, Y \sim \nu} \Big(\mathbb{E}[(X - Y)^2]\Big)^{1/2} \tag{3.4.144}$$

is given by $Y = h(X)$ with $h = \mathsf{F}_\nu^{-1} \circ \Phi$. Consequently,

$$\begin{aligned} W_2^2(\nu, G) &= \mathbb{E}[(X - h(X))^2] \\ &= \int_{-\infty}^{\infty} \big(x - h(x)\big)^2 \gamma(x)\, \mathrm{d}x. \end{aligned} \tag{3.4.145}$$

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

$$\begin{aligned}\int_{-\infty}^{x} \gamma(y)\,\mathrm{d}y &= \Phi(x)\\ &= \mathsf{F}_\nu(h(x))\\ &= \int_{-\infty}^{h(x)} f\,\mathrm{d}\mu\\ &= \int_{-\infty}^{h(x)} f(y)\gamma(y)\,\mathrm{d}y. \end{aligned} \tag{3.4.146}$$

Differentiating both sides of (3.4.146) with respect to x gives

$$\gamma(x) = h'(x)\, f(h(x))\, \gamma(h(x)), \quad \forall\, x \in \mathbb{R}. \tag{3.4.147}$$

Since $h = \mathsf{F}_\nu^{-1} \circ \Phi$, h is a monotonically increasing function, and

$$\lim_{x\to-\infty} h(x) = -\infty, \quad \lim_{x\to\infty} h(x) = \infty.$$

Consequently, we get

$$\begin{aligned} D(\nu\|G) &= D(\nu\|\mu)\\ &= \int_{\mathbb{R}} \mathrm{d}\nu\, \ln\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\\ &= \int_{-\infty}^{\infty} \ln\big(f(x)\big)\,\mathrm{d}\nu(x)\\ &= \int_{-\infty}^{\infty} f(x)\, \ln\big(f(x)\big)\,\mathrm{d}\mu(x)\\ &= \int_{-\infty}^{\infty} f(x)\, \ln\big(f(x)\big)\,\gamma(x)\,\mathrm{d}x\\ &= \int_{-\infty}^{\infty} f\big(h(x)\big)\, \ln\big(f\big(h(x)\big)\big)\,\gamma\big(h(x)\big)\,h'(x)\,\mathrm{d}x\\ &= \int_{-\infty}^{\infty} \ln\big(f(h(x))\big)\,\gamma(x)\,\mathrm{d}x, \end{aligned} \tag{3.4.148}$$

where we have used (3.4.147) to get the last equality. From (3.4.147)

$$\begin{aligned}\ln\big(f\big(h(x)\big)\big) &= \ln\left(\frac{\gamma(x)}{h'(x)\,\gamma\big(h(x)\big)}\right)\\ &= \frac{h^2(x) - x^2}{2} - \ln h'(x), \end{aligned} \tag{3.4.149}$$

which yields

$$
\begin{aligned}
&D(\nu\|\mu) \\
&= \tfrac{1}{2} \int_{-\infty}^{\infty} \left[h^2(x) - x^2\right] \gamma(x)\,\mathrm{d}x - \int_{-\infty}^{\infty} \ln h'(x)\,\gamma(x)\,\mathrm{d}x && (3.4.150) \\
&= \tfrac{1}{2} \int_{-\infty}^{\infty} \big(x - h(x)\big)^2 \gamma(x)\,\mathrm{d}x + \int_{-\infty}^{\infty} x\big(h(x) - x\big)\,\gamma(x)\,\mathrm{d}x && (3.4.151) \\
&\qquad\qquad - \int_{-\infty}^{\infty} \ln h'(x)\,\gamma(x)\,\mathrm{d}x \\
&= \tfrac{1}{2} \int_{-\infty}^{\infty} \big(x - h(x)\big)^2 \gamma(x)\,\mathrm{d}x + \int_{-\infty}^{\infty} \big(h'(x) - 1\big)\,\gamma(x)\,\mathrm{d}x \\
&\qquad\qquad - \int_{-\infty}^{\infty} \ln h'(x)\,\gamma(x)\,\mathrm{d}x && (3.4.152) \\
&\geq \tfrac{1}{2} \int_{-\infty}^{\infty} \big(x - h(x)\big)^2 \gamma(x)\,\mathrm{d}x && (3.4.153) \\
&= \tfrac{1}{2}\, W_2^2(\nu, \mu) && (3.4.154)
\end{aligned}
$$

where (3.4.150) follows from the substitution of (3.4.149) into (3.4.148); (3.4.151) is due to the identity $\frac{1}{2}(a^2 - b^2) = \frac{1}{2}(a-b)^2 + b(b-a)$ for all $a, b \in \mathbb{R}$; (3.4.152) relies on integration by parts; (3.4.153) holds due to the inequality $\ln t \leq t - 1$ for $t > 0$, and because $h\colon \mathbb{R} \to \mathbb{R}$ is monotonically increasing and differentiable; (3.4.154) holds due to (3.4.145). This shows that $\mu = G$ satisfies $\mathrm{T}_2(1)$, which completes the proof for the special case where $n = 1$.

This theorem is generalized for an arbitrary $n \in \mathbb{N}$ by tensorization via Proposition 3.8 (with $p = 2$). □

We get in the following to the main result of this section, namely that dimension-free Gaussian concentration and T_2 inequalities are equivalent:

Theorem 3.4.9. 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 $\kappa = \frac{1}{2c}$.

Remark 3.27. As we next show, the implication 1) ⇒ 2) follows easily from Propositions 3.7 and 3.8). The reverse implication 2) ⇒ 1) is a nontrivial result, which was proved by Gozlan [71], using an elegant probabilistic approach relying on the theory of large deviations.

Proof. We first prove that 1) ⇒ 2). Assume that μ satisfies $\mathrm{T}_2(c)$ on $(\mathcal{X}, d)$. Let $k \in \mathbb{N}$ be fixed, and consider the metric probability space $(\mathcal{X}^k, d_{2,k}, \mu^{\otimes k})$ where the metric $d_{2,k}$ is defined in (3.4.90) with $p = 2$. By the tensorization property of transportation-cost inequalities (Proposition 3.8), the product measure $\mu^{\otimes k}$ satisfies $\mathrm{T}_2(c)$ on $(\mathcal{X}^k, d_{2,k})$. Since the Wasserstein distance of order 2 dominates its order–1 distance (by item 2 of Lemma 3.4.3), $\mu^{\otimes k}$ also satisfies $\mathrm{T}_1(c)$ on $(\mathcal{X}^k, d_{2,k})$. Hence, by Proposition 3.7, $\mu^{\otimes k}$ satisfies the Gaussian concentration property (3.4.10) with the constants $\kappa = \frac{1}{2c}, K = 1, r_0 = \sqrt{2c \ln 2}$. Since this holds for every $k \in \mathbb{N}$, by definition μ has dimension-free Gaussian concentration with constant $\kappa = \frac{1}{2c}$.

We next prove the converse implication 2) ⇒ 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 $k \in \mathbb{N}$ be fixed, 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 *empirical measure*

$$\mathsf{P}_{x^k} = \frac{1}{k} \sum_{i=1}^{k} \delta_{x_i}, \tag{3.4.155}$$

where δ_x denotes a unit mass concentrated at $x \in \mathcal{X}$. Now consider a probability measure ν on $\mathcal{X}$ with $\nu \ll \mathsf{P}_{x^k}$, and define a function $f_{\nu,k} \colon \mathcal{X}^k \to \mathbb{R}$ by

$$f_{\nu,k}(x^k) \triangleq W_2(\mathsf{P}_{x^k}, \nu), \quad \forall\, x^k \in \mathcal{X}^k. \tag{3.4.156}$$

We claim that f_ν is Lipschitz with respect to the metric $d_{2,k}$. To verify

this property, note that

$$
\begin{aligned}
&\left|f_{\nu,k}(x^k) - f_{\nu,k}(y^k)\right| \\
&= \left|W_2(\mathsf{P}_{x^k}, \nu) - W_2(\mathsf{P}_{y^k}, \nu)\right| && (3.4.157)\\
&\le W_2(\mathsf{P}_{x^k}, \mathsf{P}_{y^k}) && (3.4.158)\\
&= \inf_{\pi \in \Pi(\mathsf{P}_{x^k}, \mathsf{P}_{y^k})} \left(\int_{\mathcal{X}} d^2(x,y)\, \mathrm{d}\pi(x,y)\right)^{1/2} && (3.4.159)\\
&\le \left(\frac{1}{k}\sum_{i=1}^{k} d^2(x_i, y_i)\right)^{1/2} && (3.4.160)\\
&= \frac{1}{\sqrt{k}}\, d_{2,k}(x^k, y^k), && (3.4.161)
\end{aligned}
$$

where

- (3.4.157) is by the definition in (3.4.156);
- (3.4.158) is by the triangle inequality for the Wasserstein distance;
- (3.4.159) is by definition of W_2 (note that the empirical measure P_{x^k} is defined on $\mathcal{X}$);
- (3.4.160) uses the fact that the measure that places mass $\frac{1}{k}$ on each (x_i, y_i) for $i \in \{1, \dots, k\}$, is an element of $\Pi(\mathsf{P}_{x^k}, \mathsf{P}_{y^k})$ (since due to the definition of an empirical distribution in (3.4.155), the marginals of the above measure are indeed P_{x^k} and P_{y^k}); and
- (3.4.161) uses the definition (3.4.90) of $d_{2,k}$.

Hence, from (3.4.157)–(3.4.161), $f_{\nu,k}$ is Lipschitz with respect to the metric $d_{2,k}$, and its Lipschitz constant satisfies

$$\|f_{\nu,k}\|_{\mathrm{Lip},2} \le \frac{1}{\sqrt{k}}. \tag{3.4.162}$$

Let $X_1, \dots, X_k$ be i.i.d. random variables with $X_1 \sim \mu$, and let m_k be an arbitrary $\mu^{\otimes k}$-median of $f_{\nu,k}$. Then, by the assumed dimension-free Gaussian concentration property of μ, Theorem 3.4.1 implies that for

every $r > 0$ and $k \in \mathbb{N}$

$$\mathbb{P}(f_{\nu,k}(X^k) \geq m_k + r) \leq \exp\left(-\frac{\kappa r^2}{\|f_{\nu,k}\|^2_{\mathrm{Lip},2}}\right) \tag{3.4.163}$$

$$\leq \exp(-\kappa k r^2), \tag{3.4.164}$$

where (3.4.164) follows from (3.4.162).

We prove that every sequence $\{m_k\}_{k=1}^{\infty}$ of $\mu^{\otimes k}$-medians of the $f_{\nu,k}$'s converges to zero. Since by construction $X_1, X_2, \ldots$ are i.i.d. draws from μ, the sequence of empirical distributions $\{\mathsf{P}_{X^k}\}_{k=1}^{\infty}$ converges almost surely to μ (it is Varadarajan's theorem [157, Theorem 11.4.1]). Hence, since W_2 metrizes the topology of weak convergence together with the convergence of second order moments (by Item 3 of Lemma 3.4.3), $\lim_{k\to\infty} W_2(\mathsf{P}_{X^k}, \mu) = 0$ almost surely. Convergence almost surely yields convergence in probability, which implies that

$$\lim_{k\to\infty} \mathbb{P}(W_2(\mathsf{P}_{X^k}, \mu) \geq t) = 0, \qquad \forall\, t > 0. \tag{3.4.165}$$

In view of (3.4.165), it follows that every sequence $\{m_k\}$ of medians of the $f_{\nu,k}$'s converges to zero, as claimed. By combining (3.4.156), (3.4.163) and (3.4.164), it follows that for all $r > 0$

$$\limsup_{k\to\infty} \frac{1}{k} \ln \mathbb{P}\Big(W_2(\mathsf{P}_{X^k}, \mu) \geq r\Big) \leq -\kappa r^2. \tag{3.4.166}$$

On the other hand, the mapping $\nu \mapsto W_2(\nu, \mu)$ is lower semicontinuous in the topology of weak convergence of probability measures (see Item 4 of Lemma 3.4.3). Consequently, the set $\{\mu\colon W_2(\mathsf{P}_{X^k}, \mu) > r\}$ is open in the weak topology, which implies by Sanov's theorem (see,e.g., [85, Theorem 6.2.10]) that for all $r > 0$

$$\liminf_{k\to\infty} \frac{1}{k} \ln \mathbb{P}\Big(W_2(\mathsf{P}_{X^k}, \mu) \geq r\Big) \geq -\inf\Big\{D(\nu\|\mu)\colon W_2(\mu, \nu) > r\Big\}. \tag{3.4.167}$$

Combining (3.4.166) and (3.4.167), we get that

$$\inf\Big\{D(\nu\|\mu)\colon\ W_2(\mu, \nu) > r\Big\} \geq \kappa r^2, \tag{3.4.168}$$

which implies that $D(\nu\|\mu) \geq \kappa\, W_2^2(\mu, \nu)$. Upon rearranging terms, we get $W_2(\mu, \nu) \leq \sqrt{\frac{1}{\kappa}\, D(\nu\|\mu)}$, which is a $\mathrm{T}_2(c)$ inequality with $c = \frac{1}{2\kappa}$. □

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 log-Sobolev inequalities) and probabilistic (through transportation-cost inequalities). We next 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 [158], relating three measures of similarity between probability measures: the divergence, quadratic Wasserstein distance, and Fisher information distance. In the literature on optimal transport, the divergence (relative entropy) between probability measures P and Q is often denoted by $H(P\|Q)$ or $H(P,Q)$, due to its close links to the Boltzmann H-functional of statistical physics. For this reason, the inequality we alluded to above is dubbed the *HWI inequality*, with H standing for the divergence, W for the Wasserstein distance, and I for the Fisher information distance (see (3.2.4) and (3.2.5)).

We first state the strong version of the HWI inequality which is specialized to the Gaussian distribution, and give a self-contained information-theoretic proof following [159]:

Theorem 3.4.10 (Gaussian HWI inequality)**.** Let G denote the standard Gaussian probability distribution on $\mathbb{R}$. Then, the inequality

$$D(P\|G) \leq W_2(P,G)\sqrt{I(P\|G)} - \tfrac{1}{2}\, W_2^2(P,G), \tag{3.4.169}$$

where W_2 is the quadratic Wasserstein distance with respect to the absolute-value metric $d(x,y) = |x-y|$, holds for every Borel probability distribution P on $\mathbb{R}$, for which the right side of (3.4.169) is finite.

Proof. We first show the following:

Lemma 3.4.11. Let X and Y be a pair of real-valued random variables, and let $N \sim G$ be independent of (X,Y). Then, for every $t > 0$,

$$D(P_{X+\sqrt{t}N}\|P_{Y+\sqrt{t}N}) \leq \frac{1}{2t}\, W_2^2(P_X, P_Y). \tag{3.4.170}$$

Proof. From the chain rule for divergence (see [144, Theorem 2.5.3]), we have

$$D(P_{X,Y,X+\sqrt{t}N}\|P_{X,Y,Y+\sqrt{t}N}) \geq D(P_{X+\sqrt{t}N}\|P_{Y+\sqrt{t}N}) \tag{3.4.171}$$

and

$$\begin{aligned} & D(P_{X,Y,X+\sqrt{t}N} \| P_{X,Y,Y+\sqrt{t}N}) \\ & = D(P_{X+\sqrt{t}N|X,Y} \,\|\, P_{Y+\sqrt{t}N|X,Y} | P_{X,Y}) \\ & = \mathbb{E}[D(\mathcal{N}(X,t) \,\|\, \mathcal{N}(Y,t)) \,|\, X,Y] \qquad (3.4.172) \\ & = \frac{1}{2t}\,\mathbb{E}[(X-Y)^2]. \qquad (3.4.173) \end{aligned}$$

Note that (3.4.172) holds since $N \sim G$ is independent of (X,Y), and (3.4.173) is a special case of the identity

$$\begin{aligned} & D(\mathcal{N}(m_1,\sigma_1^2) \,\|\, \mathcal{N}(m_2,\sigma_2^2)) \\ & = \frac{1}{2}\left[\ln\left(\frac{\sigma_2^2}{\sigma_1^2}\right) + \frac{(m_1-m_2)^2}{\sigma_2^2} + \frac{\sigma_1^2}{\sigma_2^2} - 1\right]. \qquad (3.4.174) \end{aligned}$$

It therefore follows from (3.4.171) and (3.4.173) that

$$D(P_{X+\sqrt{t}N} \| P_{Y+\sqrt{t}N}) \le \frac{1}{2t}\,\mathbb{E}[(X-Y)^2]. \qquad (3.4.175)$$

The left side of (3.4.175) only depends on the marginal distributions of X and Y (due to the independence of (X,Y) and $N \sim G$). Hence, by taking the infimum of the right side of (3.4.175) with respect to all $\mu \in \Pi(P_X, P_Y)$, we get (3.4.170) (see (3.4.43)). □

We now proceed with the proof of Theorem 3.4.10. Let X and Y have distributions P and $Q = G$, respectively. For simplicity, we focus on the case where X has zero mean and unit variance; the general case can be handled similarly. Let

$$F(t) \triangleq D(P_{X+\sqrt{t}N} \| P_{Y+\sqrt{t}N}), \quad \forall\, t > 0, \qquad (3.4.176)$$

where $N \sim G$ is independent of the pair (X,Y). Then, we have

$$F(0) = D(P\|G), \qquad (3.4.177)$$

and from (3.4.170)

$$F(t) \le \frac{1}{2t}\, W_2^2(P_X, P_Y) = \frac{1}{2t}\, W_2^2(P,G), \quad \forall\, t > 0. \qquad (3.4.178)$$

Moreover, the function $F(t)$ is differentiable, and it follows from a result by Verdú [136, Eq. (32)] that

$$\begin{aligned} F'(t) &= \frac{1}{2t^2}\left[\mathsf{mmse}(X,t^{-1}) - \mathsf{mse}_Q(X,t^{-1})\right] \\ &= \frac{1}{2t^2}\left[\mathsf{mmse}(X,t^{-1}) - \mathsf{lmmse}(X,t^{-1})\right], \quad \forall\, t>0 \end{aligned} \tag{3.4.179}$$

where $\mathsf{mmse}(X,\cdot)$, $\mathsf{mse}_Q(X,\cdot)$ and $\mathsf{lmmse}(X,\cdot)$ have been defined in (3.2.28), (3.2.29) and (3.2.32), respectively. The second equality in (3.4.179) holds due to (3.2.31) with $Q=G$ (recall that in the Gaussian setting, the optimal estimator for minimizing the mean square error is linear). For every $t>0$,

$$\begin{aligned} &D(P\|G) \\ &= F(0) && (3.4.180)\\ &= -\int_0^t F'(s)\mathrm{d}s + F(t) && (3.4.181)\\ &= \tfrac{1}{2}\int_0^t \frac{1}{s^2}\left(\mathsf{lmmse}(X,s^{-1}) - \mathsf{mmse}(X,s^{-1})\right)\mathrm{d}s + F(t) && (3.4.182)\\ &\le \tfrac{1}{2}\int_0^t \left(\frac{1}{s(s+1)} - \frac{1}{s(sJ(X)+1)}\right)\mathrm{d}s + \tfrac{1}{2t}\,W_2^2(P,G) && (3.4.183)\\ &= \tfrac{1}{2}\left(\ln\frac{tJ(X)+1}{t+1} + \frac{W_2^2(P,G)}{t}\right) && (3.4.184)\\ &= \tfrac{1}{2}\left(\ln\frac{t\big(I(P\|G)+1\big)+1}{t+1} + \frac{W_2^2(P,G)}{t}\right) && (3.4.185)\\ &\le \tfrac{1}{2}\left(\frac{t\,I(P\|G)}{t+1} + \frac{W_2^2(P,G)}{t}\right) && (3.4.186) \end{aligned}$$

where

- (3.4.180) is (3.4.177);
- (3.4.181) uses the identity $\int_0^t F'(s)\mathrm{d}s = F(t) - F(0)$;
- (3.4.182) uses (3.4.179);
- (3.4.183) uses (3.2.33), the Van Trees inequality (3.2.34), and (3.4.178);

- (3.4.184) is an exercise in calculus;
- (3.4.185) uses the formula (3.2.24) (so $I(P\|G) = J(X) - 1$ since $X \sim P$ has zero mean and unit variance; one needs to substitute $s = 1$ in (3.2.24) to get $G_s = G$), and the fact that $t \geq 0$;
- (3.4.186) uses the inequality $\ln x \leq x - 1$ for $x > 0$; and

Optimizing t to minimize the bound in (3.4.186) yields

$$t_{\text{opt}} = \frac{W_2(P, G)}{\sqrt{I(P\|G)} - W_2(P, G)}. \tag{3.4.187}$$

Note that, due to the celebrated Talagrand quadratic transportation-cost inequality [156]

$$I(P\|G) \geq W_2^2(P, G), \tag{3.4.188}$$

which is obtained by combining the Gaussian-LSI in Proposition 3.4 with (3.4.150)–(3.4.154), it follows that the optimized value of t in (3.4.187) is indeed positive whenever $I(P\|G) > 0$ (i.e., if P is not the standard Gaussian measure). Consequently, the substitution of $t = t_{\text{opt}}$ in (3.4.187) into (3.4.186) gives (3.4.169). □

A stronger version of the Gaussian HWI in Theorem 3.4.10 can be obtained by optimizing the parameter $t > 0$ in the right side of (3.4.185), yielding

$$D(P\|G) \leq \tfrac{1}{2} \inf_{t>0} \left(\ln \frac{t\big(I(P\|G) + 1\big) + 1}{t + 1} + \frac{W_2^2(P, G)}{t} \right). \tag{3.4.189}$$

Note that, in the proof of Theorem 3.4.10, the Gaussian HWI (3.4.169) is derived by optimizing $t > 0$ in the right side of (3.4.186) which is looser than (3.4.185). The minimization in the right side of (3.4.189) yields the following closed-form result:

- If

$$\frac{I(P\|G)}{1 + I(P\|G)} \leq W_2^2(P, G) \leq I(P\|G) \tag{3.4.190}$$

(note that the right side of (3.4.190) is due to (3.4.188)), then $t \to \infty$ is optimal, yielding (cf. Proposition 3.4)

$$D(P\|G) \leq \tfrac{1}{2} \log(1 + I(P\|G)). \tag{3.4.191}$$

- Otherwise, if

$$W_2^2(P,G) < \frac{I(P\|G)}{1+I(P\|G)}, \tag{3.4.192}$$

then the optimized $t \in (0,\infty)$ in the right side of (3.4.185) is given by

$$t_{\text{opt}} = \frac{\sqrt{b^2-4ac}-b}{2a} \tag{3.4.193}$$

with

$$a = I(P\|G) - W_2^2(P,G)\left(I(P\|G)+1\right) > 0, \tag{3.4.194}$$
$$b = -\big(I(P\|G)+2\big)\,W_2^2(P,G), \tag{3.4.195}$$
$$c = -W_2^2(P,G). \tag{3.4.196}$$

Consequently, if the condition (3.4.192) holds, then a tightened version of the Gaussian HWI (3.4.169) is obtained by substituting (3.4.193) into the right side of (3.4.185).

Remark 3.28. Note that the HWI inequality (3.4.169) together with the T_2 inequality for the Gaussian distribution imply a weaker version of the LSI (3.2.10) (i.e., with a larger constant). Indeed, using the T_2 inequality of Theorem 3.4.8 on the right side of (3.4.169), we get

$$D(P\|G) \le W_2(P,G)\sqrt{I(P\|G)} \tag{3.4.197}$$
$$\le \sqrt{2D(P\|G)}\sqrt{I(P\|G)}, \tag{3.4.198}$$

which gives $D(P\|G) \le 2I(P\|G)$. It is not surprising that we end up with a suboptimal constant here as compared to (3.2.10): the series of bounds leading up to (3.4.186) contributes a lot more slack than the single use of the van Trees inequality (3.2.34) in our proof of Stam's inequality (which, due to Proposition 3.4, is equivalent to the Gaussian LSI of Gross).

We are now ready to state the HWI inequality in its general form:

Theorem 3.4.12 (Otto–Villani, Theorem 3 in [158])**.** 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

$$\nabla^2 \ln\left(\frac{1}{p}\right) \succeq K I_n \tag{3.4.199}$$

for some $K \in \mathbb{R}$ (where ∇^2 denotes the Hessian matrix, and the matrix inequality $A \succeq B$ means that $A - B$ is non-negative semidefinite). Then, every probability measure $Q \ll P$ satisfies

$$D(Q\|P) \leq W_2(Q,P)\sqrt{I(Q\|P)} - \tfrac{K}{2}\, W_2^2(Q,P). \tag{3.4.200}$$

We omit the proof of Theorem 3.4.12, which relies on some deep structural properties of optimal transportation mappings achieving the infimum in the definition of the quadratic Wasserstein distance with respect to the Euclidean norm in $\mathbb{R}^n$. (An alternative simpler proof was given later by Cordero–Erausquin [160].) We can, however, highlight a couple of key consequences (see [158]):

1. Let P, in addition to satisfying the conditions of Theorem 3.4.12, satisfy a $\mathrm{T}_2(c)$ inequality. Using this T_2 inequality and (3.4.200) yields

$$D(Q\|P) \leq \sqrt{2cD(Q\|P)}\sqrt{I(Q\|P)} - \tfrac{K}{2}\, W_2^2(Q,P). \tag{3.4.201}$$

   If the pdf p of P is log-concave, so that (3.4.199) holds with $K = 0$, then (3.4.201) implies the inequality

$$D(Q\|P) \leq 2c\, I(Q\|P) \tag{3.4.202}$$

   where $Q \ll P$. This is an Euclidean LSI that is similar to the one satisfied by $P = G^n$ (see Remark 3.28). Note, however, that the coefficient in the right side of (3.4.202) (i.e., the constant in front of the Fisher information distance) is suboptimal; this can be verified by letting $P = G^n$, which satisfies $\mathrm{T}_2(1)$. Going through the above steps, as we know from (3.2.10), the optimal constant should be $\frac{1}{2}$, so the one in (3.4.202) 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 (3.4.199) holds with some $K > 0$, then P satisfies the Euclidean LSI with constant $\frac{1}{K}$. Indeed, we have from (3.4.200)

$$D(Q\|P) \leq \sqrt{K} W_2(Q,P) \sqrt{\tfrac{1}{K} I(Q\|P)} - \tfrac{K}{2} W_2^2(Q,P) \quad (3.4.203)$$

$$\leq \tfrac{1}{2K} I(Q\|P), \quad (3.4.204)$$

where (3.4.203) is (3.4.201), and (3.4.204) relies on the simple inequality $ab \leq \frac{a^2+b^2}{2}$ for all $a, b \in \mathbb{R}$. This shows that P satisfies the Euclidean LSI $\left(\frac{1}{K}\right)$ inequality. In particular, the standard n-dimensional Gaussian distribution $P = G^n$ satisfies (3.4.199) with $K = 1$, so we even get the right constant in (3.4.204). In fact, the statement that (3.4.199) with $K > 0$ yields the Euclidean LSI $\left(\frac{1}{K}\right)$ was first proved in 1985 by Bakry and Emery [161] using very different means.

Introduced independently by Ali-Silvey [162] and Csiszár [163, 164], a useful generalization of the relative entropy, which retains some of its major properties (and, in particular, the data processing inequality [165]), is the class of f-divergences. In [166], Sason and Verdú study several approaches to derive f-divergence inequalities. By combining f-divergence inequalities and the WHI inequality in Theorem 3.4.12, this enables to derive upper bounds on various f-divergences as a function of the relative Fisher information and the quadratic Wasserstein distance.

3.5 Extension to non-product distributions

Our focus in this chapter is mostly on functions of independent random variables. However, there is extensive literature on the concentration of measure inequalities for weakly dependent random variables. In this section, we describe (without proof) some 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 as before: $X_1, \ldots, X_n$ are random variables 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 a function $f(X^n)$.

3.5.1 Samson's approach for dependent random variables

In [167], Samson developed an 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.,

$$\|\Delta\| \triangleq \sup_{v \in \mathbb{R}^n \setminus \{0\}} \frac{\|\Delta v\|}{\|v\|} \tag{3.5.2}$$

$$= \sup_{v \in \mathbb{R}^n \colon \|v\| = 1} \|\Delta v\|. \tag{3.5.3}$$

Following Marton [168], Samson [167] considered a Wasserstein-type distance on the space of probability measures on $\mathcal{X}^n$. For every pair of probability measures Q and R on $\mathcal{X}^n$, let $\Pi(Q, R)$ denote the set of all probability measures on $\mathcal{X}^n \times \mathcal{X}^n$ with marginals Q and R; the following non-negative quantity is defined in [167]:

$$d_2(Q, R) \triangleq \inf_{\pi \in \Pi(Q,R)} \sup_{\alpha} \int \sum_{i=1}^n \alpha_i(y^n) 1_{\{x_i \neq y_i\}} \, \mathrm{d}\pi(x^n, y^n), \tag{3.5.4}$$

where $\sup_\alpha$ refers to the supremum over all vector-valued functions $\alpha \colon \mathcal{X}^n \to \mathbb{R}^n$ where $\alpha = (\alpha_1, \ldots, \alpha_n)$ is a vector of positive functions, and

$$\mathbb{E}_R \left[\|\alpha(Y^n)\|^2 \right] = \int_{\mathcal{X}^n} \sum_{i=1}^n \alpha_i^2(y^n) \, \mathrm{d}R(y^n) \leq 1. \tag{3.5.5}$$

Remark 3.29. Note that $d_2(Q,Q) = 0$; however, in general, we have $d_2(Q,R) \neq d_2(R,Q)$ due to the difference in the two conditions $\mathbb{E}_R\left[\|\alpha(Y^n)\|^2\right] \leq 1$ and $\mathbb{E}_Q\left[\|\alpha(Y^n)\|^2\right] \leq 1$ involved in the definition of $d_2(Q,R)$ and $d_2(R,Q)$, respectively. Therefore, d_2 is *not* a distance.

The main result of [167] is the following Pinsker-type inequalities (see [167, Theorem 1]).

Theorem 3.5.1. Let P and Q be probability measures on $\mathbb{R}^n$ such that $Q \ll P$. Then, the following inequalities hold:

$$d_2(Q,P) \leq \|\Delta\|\sqrt{2D(Q\|P)}, \tag{3.5.6}$$

and

$$d_2(P,Q) \leq \|\Delta\|\sqrt{2D(Q\|P)}. \tag{3.5.7}$$

In the following, we examine some implications of Theorem 3.5.1.

1. Let $\mathcal{X} = [0,1]$, and let P be a probability measure defined on the unit cube $\mathcal{X}^n = [0,1]^n$. Theorem 3.5.1 implies that P satisfies the following Euclidean LSI (see [167, Corollary 1]): for an arbitrary smooth convex function $f\colon [0,1]^n \to \mathbb{R}$,

$$D(P^{(f)}\|P) \leq 2\|\Delta\|^2\,\mathbb{E}_P^{(f)}\left[\|\nabla f(X^n)\|^2\right] \tag{3.5.8}$$

(note that the equivalence of (3.5.8) and [167, Eq. (2.13)] follows from (3.3.6) and (3.3.7)). The same method as the one we used to prove Proposition 3.5 and Theorem 3.2.2 can be applied to obtain, from (3.5.8), the following concentration inequality for every convex function $f\colon [0,1]^n \to \mathbb{R}$ with $\|f\|_{\mathrm{Lip}} \leq 1$:

$$\mathbb{P}\Big(f(X^n) \geq \mathbb{E}f(X^n) + r\Big) \leq \exp\left(-\frac{r^2}{8\|\Delta\|^2}\right), \tag{3.5.9}$$

which holds for all $r \geq 0$. An adaptation of the approach by Bobkov and Götze [54] (recall that this approach is used to prove Theorem 3.4.4 and Corollary 3.4.5) gives, however, the following improved concentration inequality which holds for every smooth

convex function $f\colon [0,1]^n \to \mathbb{R}$ with $\|\nabla f\| \leq 1$ P-a.s. and for all $r \geq 0$ (see [167, Corollary 3]):

$$\mathbb{P}\Big(f(X^n) \geq \mathbb{E}f(X^n) + r\Big) \leq \exp\left(-\frac{r^2}{2\|\Delta\|^2}\right). \tag{3.5.10}$$

The same inequality in (3.5.10) also holds for an arbitrary smooth concave function $f\colon [0,1]^n \to \mathbb{R}$ such that $\mathbb{E}_P\big[\|\nabla f\|^2\big] \leq 1$.

2. The operator norm $\|\Delta\|$ in (3.5.6)–(3.5.10) is weakly dependent on n whenever the dependence between the X_i's is sufficiently weak. For instance, if $X_1, \ldots, X_n$ are independent then $\Delta = I_{n\times n}$, and $\|\Delta\| = 1$ for all n. In this case, (3.5.6) becomes

$$d_2(Q,P) \leq \sqrt{2D(Q\|P)}, \tag{3.5.11}$$

enabling to recover the concentration inequalities for Lipschitz functions. For examples with dependent random variables, let $X_1, \ldots, X_n$ be a Markov chain; by definition, 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 gets the simplified form

$$\Delta_{i,j} = \sup_{x_i, x_i'} \sqrt{\Big\|P_{X_j|X_i=x_i} - P_{X_j|X_i=x_i'}\Big\|_{\mathrm{TV}}}, \qquad \forall\, i < j. \tag{3.5.12}$$

The norm $\|\Delta\|$ is bounded in n under suitable assumptions on the Markov chain $\{X_k\}_{k=1}^n$. For instance, suppose that the Markov chain is homogeneous (i.e., $P_{X_i|X_{i-1}}$ is independent of i), and

$$\sup_{x_i, x_i'} \|P_{X_{i+1}|X_i=x_i} - P_{X_{i+1}|X_i=x_i'}\|_{\mathrm{TV}} \leq 2\rho \tag{3.5.13}$$

with $\rho < 1$. Then, it can be shown that (see [167, Eq. (2.5)])

$$\|\Delta\| \leq \sqrt{2}\left(1 + \sum_{k=1}^{n-1} \rho^{k/2}\right) \leq \frac{\sqrt{2}}{1-\sqrt{\rho}}. \tag{3.5.14}$$

More generally, a non-necessarily homogeneous Markov chain $\{X_k\}$ is said to be *contracting* if for every i (see [167, (2.7)])

$$\delta_i \triangleq \sup_{x_i, x_i'} \|P_{X_{i+1}|X_i=x_i} - P_{X_{i+1}|X_i=x_i'}\|_{\mathrm{TV}} < 1. \tag{3.5.15}$$

In this case, $\|\Delta\| \leq \frac{1}{1-\sqrt{\delta}}$ with $\delta \triangleq \max \delta_i$ [167, pp. 422–424].

3.5.2 Marton's transportation-cost inequalities for L^2 Wasserstein distance

Another approach to obtain concentration of measure inequalities for dependent random variables, due to Marton ([169], [170]), relies on another measure of dependence that pertains to the sensitivity of the conditional distributions of X_i given $\overline{X}^i$ to the particular realization $\bar{x}^i$ of $\overline{X}^i$. These results are set in the Euclidean space $\mathbb{R}^n$, and center around a transportation-cost inequality for the L^2 Wasserstein distance

$$W_2(P,Q) \triangleq \inf_{X^n \sim P,\, Y^n \sim Q} \sqrt{\mathbb{E}\|X^n - Y^n\|^2}, \tag{3.5.16}$$

where $\|\cdot\|$ denotes the Euclidean norm.

We next state a special case of Marton's results (a more general development considers conditional distributions of $(X_i \colon i \in S)$ given $(X_j \colon j \in S^c)$ 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 all $x^n \in \mathbb{R}^n$ and $i \in \{1,\ldots,n\}$, denote by $\bar{x}^i$ the vector in $\mathbb{R}^{n-1}$ obtained by deleting the i-th coordinate of x^n:

$$\bar{x}^i = (x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n). \tag{3.5.17}$$

Following Marton [169], the probability measure P is $(1-\delta)$-*contractive*, with $\delta \in (0,1)$, if for every $y^n, z^n \in \mathbb{R}^n$

$$\sum_{i=1}^{n} W_2^2(P_{X_i|\overline{X}^i=\bar{y}^i}, P_{X_i|\overline{X}^i=\bar{z}^i}) \le (1-\delta)\|y^n - z^n\|^2. \tag{3.5.18}$$

Remark 3.30. The contractivity condition (3.5.18) is closely related to the so-called *Dobrushin–Shlosman's strong mixing condition* [171] from statistical physics.

Theorem 3.5.2 (Marton [169, 170])**.** Let P be a probability measure which is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^n$, and let it be $(1-\delta)$-contractive with $\delta \in (0,1)$. Suppose also that, for all $i \in \{1,\ldots,n\}$, the following properties hold:

1. The function $x^n \mapsto p_{X_i|\overline{X}^i}(x_i|\bar{x}^i)$ is continuous, where $p_{X_i|\overline{X}^i}(\cdot|\bar{x}^i)$ denotes the univariate probability density function of $P_{X_i|\overline{X}^i=\bar{x}^i}$.

2. For every $\bar{x}^i \in \mathbb{R}^{n-1}$, $P_{X_i|\overline{X}^i=\bar{x}^{i-1}}$ satisfies $\mathrm{T}_2(c)$ with respect to the L^2 Wasserstein distance (3.5.16).

Then, for every probability measure Q on $\mathbb{R}^n$, we have

$$W_2(Q,P) \le \left(\frac{K}{\sqrt{\delta}}+1\right)\sqrt{2cD(Q\|P)}, \tag{3.5.19}$$

where $K>0$ is an absolute constant. In other words, every P satisfying the conditions of the theorem admits a $\mathrm{T}_2(c')$ inequality with

$$c' = \left(\frac{K}{\sqrt{\delta}}+1\right)^2 c. \tag{3.5.20}$$

The contractivity criterion (3.5.18) is not easy to verify in general. Let us mention a sufficient condition [169]. Let p denote the probability density of P, and suppose that it takes the form

$$p(x^n) = \frac{1}{Z}\exp(-\Psi(x^n)) \tag{3.5.21}$$

for some C^2 function $\Psi\colon \mathbb{R}^n \to \mathbb{R}$, where Z is the normalization factor. For every $x^n, y^n \in \mathbb{R}^n$, let the matrix $B(x^n,y^n) \in \mathbb{R}^{n\times n}$ be defined as

$$B_{i,j}(x^n,y^n) \triangleq \begin{cases} \nabla^2_{i,j}\Psi(x_i \odot \bar{y}^i), & i \ne j \\ 0, & i=j \end{cases} \tag{3.5.22}$$

where $\nabla^2_{i,j}F$ denotes the (i,j) entry of the Hessian matrix of a function $F \in C^2(\mathbb{R}^n)$, and $x_i \odot \bar{y}^i$ denotes the n-tuple obtained by replacing the deleted i-th coordinate in $\bar{y}^i$ with x_i:

$$x_i \odot \bar{y}^i = (y_1,\ldots,y_{i-1},x_i,y_{i+1},\ldots,y_n). \tag{3.5.23}$$

For example, if Ψ is a sum of one-variable and two-variable terms

$$\Psi(x^n) = \sum_{i=1}^n V_i(x_i) + \sum_{i<j} b_{i,j}x_ix_j \tag{3.5.24}$$

for smooth functions $V_i\colon \mathbb{R}\to\mathbb{R}$ and constants $b_{i,j}\in\mathbb{R}$, which is often the case in statistical physics, then the matrix B is independent of x^n and y^n, and it has off-diagonal entries $b_{i,j}$ if $i<j$ or $b_{j,i}$ if $i>j$. In view of [169, Theorem 2], the conditions of Theorem 3.5.2 are satisfied provided the following holds:

1. For each $i \in \{1, \ldots, n\}$ and $\bar{x}^i \in \mathbb{R}^{n-1}$, the conditional probability distributions $P_{X_i|\overline{X}^i=\bar{x}^i}$ satisfy the Euclidean LSI

$$D(Q\|P_{X_i|\overline{X}^i=\bar{x}^i}) \leq \frac{c}{2}\, I(Q\|P_{X_i|\overline{X}^i=\bar{x}^i}), \tag{3.5.25}$$

   where $I(\cdot\|\cdot)$ is the Fisher information distance (3.2.4).

2. The operator norms of $B(x^n, y^n)$ are uniformly bounded as

$$\sup_{x^n, y^n} \|B(x^n, y^n)\|^2 \leq \frac{1-\delta}{c^2}. \tag{3.5.26}$$

We refer the reader to a follow-up work by Marton [172], which further elaborates on the theme of studying the concentration properties of dependent random variables by focusing on the conditional probability distributions $P_{X_i|\overline{X}^i}$ for $i \in \{1, \ldots, n\}$. This paper describes sufficient conditions on the joint distribution P of $X_1, \ldots, X_n$ such that, for every other distribution Q,

$$D(Q\|P) \leq K(P)\, D^-(Q\|P), \tag{3.5.27}$$

where $D^-(\cdot\|\cdot)$ is the erasure divergence (defined in (3.1.29)), 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 (3.5.27) should be clear: each term in the erasure divergence

$$D^-(Q\|P) = \sum_{i=1}^{n} D(Q_{X_i|\overline{X}^i}\|P_{X_i|\overline{X}^i}|Q_{\overline{X}^i}) \tag{3.5.28}$$

can be handled by appealing to appropriate log-Sobolev or transportation-cost inequalities for probability measures on $\mathcal{X}$ (indeed, one can treat $P_{X_i|\overline{X}^i=\bar{x}^i}$ 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 to derive a concentration result for the original n-dimensional distribution.

3.6 Applications in information theory and related topics

3.6.1 The blowing-up lemma

An explicit invocation of the concentration of measure phenomenon in an information-theoretic context appeared for the first time in the work by Ahlswede et al. [73, 74]. These papers show that the following result, known to-date as the *blowing-up lemma* (see, e.g., [173, Lemma 5.4]), provides a versatile tool for proving strong converses in a variety of scenarios, including some multiterminal problems. Informally, it says that if we enlarge any set of not too small probability with a thin layer then the enlarged set shall have probability almost one.

In the sequel, let $\mathcal{Y}$ be a finite set, $n \in \mathbb{N}$, and $r > 0$. Given a set $\mathcal{B} \subset \mathcal{Y}^n$ and $r > 0$, the set $\mathcal{B}_r$ denotes the r-blowup of $\mathcal{B}$, i.e.,

$$\mathcal{B}_r \triangleq \{y^n \in \mathcal{Y}^n : d_n(y^n, \mathcal{B}) < r\} \tag{3.6.1}$$

where

$$d_n(y^n, \mathcal{B}) \triangleq \min_{\hat{y}^n \in \mathcal{B}} d_n(y^n, \hat{y}^n) \tag{3.6.2}$$

with the Hamming metric

$$d_n(y^n, \hat{y}^n) \triangleq \sum_{i=1}^{n} 1_{\{y_i \neq \hat{y}_i\}}, \qquad \forall\, y^n, \hat{y}^n \in \mathcal{Y}^n. \tag{3.6.3}$$

Lemma 3.6.1 (the blowing-up lemma)**.** For every positive sequence $\xi_n \to 0$, there exist positive sequences $\delta_n \to 0$ and $\eta_n \to 0$ such that the following property holds: for every discrete memoryless channel (DMC) with finite input alphabet $\mathcal{X}$, finite output alphabet $\mathcal{Y}$, and transition probabilities $T(y|x)$ with $(x, y) \in \mathcal{X} \times \mathcal{Y}$, and for every $x^n \in \mathcal{X}^n$, and $\mathcal{B} \subset \mathcal{Y}^n$,

$$T^n(\mathcal{B}|x^n) \geq \exp(-n\xi_n) \quad \Longrightarrow \quad T^n(\mathcal{B}_{n\delta_n}|x^n) \geq 1 - \eta_n. \tag{3.6.4}$$

The proof in [73] of the blowing-up lemma is rather technical and it makes use of a delicate isoperimetric inequality for discrete probability measures on a Hamming space, due to Margulis [174]. Later, the same result was obtained by Marton [75] using purely information-theoretic

methods. We use here a sharper non-asymptotic version of the blowing-up lemma (see also Marton's follow-up paper in [59], extending the result for some non-product measures):

Lemma 3.6.2. Let $X_1, \ldots, X_n$ be independent random variables taking values in a finite set $\mathcal{X}$. Then, for every $\mathcal{B} \subseteq \mathcal{X}^n$ with $P_{X^n}(\mathcal{B}) > 0$,

$$P_{X^n}(\mathcal{B}_r) \geq 1 - \exp\left[-\frac{2}{n}\left(r - \sqrt{\frac{n}{2}\ln\frac{1}{P_{X^n}(\mathcal{B})}}\right)^2\right], \tag{3.6.5}$$

for all $r > \sqrt{\frac{n}{2}\ln\frac{1}{P_{X^n}(\mathcal{B})}}$.

Proof. Let P_n denote the product measure $P_{X^n} = P_{X_1} \otimes \ldots \otimes P_{X_n}$. By Pinsker's inequality, every $\mu \in \mathcal{P}(\mathcal{X})$ satisfies $\mathrm{T}_1(\frac{1}{4})$ on $(\mathcal{X}, d)$ where $d = d_1$ is the Hamming metric (see Example 3.2). By Proposition 3.9, P_n satisfies $\mathrm{T}_1(\frac{n}{4})$ on $(\mathcal{X}^n, d_n)$ with the Hamming metric d_n in (3.6.3), i.e., for all $\mu_n \in \mathcal{P}(\mathcal{X}^n)$,

$$W_1(\mu_n, P_n) \leq \sqrt{\frac{n}{2}\,D(\mu_n \| P_n)}. \tag{3.6.6}$$

The statement of the lemma follows from the proof of Proposition 3.7. More precisely, applying (3.4.88) to the probability measure P_{X^n} with $c = \frac{n}{4}$ gives

$$r \leq \sqrt{\frac{n}{2}\,\ln\frac{1}{P_{X^n}(\mathcal{B})}} + \sqrt{\frac{n}{2}\,\ln\frac{1}{1 - P_{X^n}(\mathcal{B}_r)}}\,, \quad \forall\, r > 0, \tag{3.6.7}$$

and (3.6.5) follows by rearranging terms. □

We next prove Lemma 3.6.1.

Proof. Given a positive sequence $\{\xi_n\}_{n=1}^{\infty}$ which converges to zero, let $\{\delta_n\}_{n=1}^{\infty}$ converge to zero such that

$$\delta_n > \sqrt{\frac{\xi_n}{2}}, \tag{3.6.8}$$

$$\eta_n \triangleq \exp\left(-2n\left(\delta_n - \sqrt{\frac{\xi_n}{2}}\right)^2\right) \xrightarrow{n\to\infty} 0. \tag{3.6.9}$$

These requirements can be satisfied, e.g., by setting

$$\delta_n \triangleq \sqrt{\frac{\xi_n}{2}} + \sqrt{\frac{\alpha \ln n}{n}}, \quad \forall\, n \in \mathbb{N}, \tag{3.6.10}$$

where $\alpha > 0$ is a fixed constant which can be made arbitrarily small, and from (3.6.9) and (3.6.10)

$$\eta_n = \frac{1}{n^{2\alpha}}, \quad \forall\, n \in \mathbb{N}. \tag{3.6.11}$$

Let $T^n(\cdot|x^n) = P_{X^n}$, then (3.6.5) and (3.6.10) yield (3.6.4). □

Remark 3.31. The sequences $\{\delta_n\}$ and $\{\eta_n\}$ in the blowing-up property (3.6.4), as specified in (3.6.10) and (3.6.11), only depend on $\{\xi_n\}$ in the left side of (3.6.4).

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 [173], we use the notation $T\colon \mathcal{U} \to \mathcal{V}$ for a DMC with finite input alphabet $\mathcal{U}$, finite output alphabet $\mathcal{V}$, and transition probabilities $T(v|u)$ for $(u, v) \in \mathcal{U} \times \mathcal{V}$.

Consider the problem of characterizing the capacity region of a 2-user discrete memoryless degraded broadcast channel (DM-DBC) with independent messages, defined as follows:

Definition 3.3 (DM-DBC)**.** Let $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ be finite sets. A DM-DBC is specified by a pair of DMCs $T_1\colon \mathcal{X} \to \mathcal{Y}$ and $T_2\colon \mathcal{X} \to \mathcal{Z}$ where there exists a DMC $T_3\colon \mathcal{Y} \to \mathcal{Z}$ such that

$$T_2(z|x) = \sum_{y \in \mathcal{Y}} T_1(y|x)\, T_3(z|y), \qquad \forall\, (x, z) \in \mathcal{X} \times \mathcal{Z}. \tag{3.6.12}$$

(More precisely, this is a *stochastically degraded* broadcast channel – see, e.g., [144, Section 15.6] and [175, Section 5.4]; a *physically degraded* broadcast channel has the probability law

$$\mathbb{P}(y, z|x) = T_1(y|x)\, T_3(z|y), \quad \forall\, (x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \tag{3.6.13}$$

so, to every DM-DBC, there is a corresponding physically degraded broadcast channel with the same conditional marginal distributions.

Definition 3.4 (Codes). Given $n, M_1, M_2 \in \mathbb{N}$, an (n, M_1, M_2)-*code* $\mathcal{C}$ for the broadcast channel consists of the following objects:

1. An *encoding map* $f_n \colon \{1, \dots, M_1\} \times \{1, \dots, M_2\} \to \mathcal{X}^n$;
2. A collection $\mathcal{D}_1$ of M_1 disjoint *decoding sets* for receiver 1
$$D_{1,i} \subset \mathcal{Y}^n, \quad i \in \{1, \dots, M_1\}$$
and a collection $\mathcal{D}_2$ of M_2 disjoint decoding sets for receiver 2
$$D_{2,j} \subset \mathcal{Z}^n, \quad j \in \{1, \dots, M_2\}.$$

Given $\varepsilon_1, \varepsilon_2 \in (0, 1)$, we say that the code $\mathcal{C} = (f_n, \mathcal{D}_1, \mathcal{D}_2)$ is an $(n, M_1, M_2, \varepsilon_1, \varepsilon_2)$-*code* if

$$\max_{1 \le i \le M_1} \max_{1 \le j \le M_2} T_1^n\Big(D_{1,i}^c \Big| f_n(i,j)\Big) \le \varepsilon_1, \tag{3.6.14a}$$

$$\max_{1 \le i \le M_1} \max_{1 \le j \le M_2} T_2^n\Big(D_{2,j}^c \Big| f_n(i,j)\Big) \le \varepsilon_2. \tag{3.6.14b}$$

In other words, the maximal probability of error criterion is used in Definition 3.4. Note that, for general multiuser channels, the capacity region with respect to the maximal probability of error may be strictly smaller than the capacity region with respect to the average probability of error [176]; nevertheless, these two capacity regions are identical for discrete memoryless broadcast channels [177].

Definition 3.5 (Achievable rates). A pair of rates (R_1, R_2) (in nats per channel use) is said to be $(\varepsilon_1, \varepsilon_2)$-*achievable* if for every $\delta > 0$, there exists an $(n, M_1, M_2, \varepsilon_1, \varepsilon_2)$-code (for a sufficiently large block length n) such that

$$\tfrac{1}{n} \ln M_k \ge R_k - \delta, \qquad k \in \{1, 2\}. \tag{3.6.15}$$

Likewise, (R_1, R_2) is said to be *achievable* if it is $(\varepsilon_1, \varepsilon_2)$-achievable for all $0 < \varepsilon_1, \varepsilon_2 \le 1$ (according to the criterion of the maximal probability of error in Definition 3.4, this is equivalent to the requirement that (R_1, R_2) is $(\varepsilon_1, \varepsilon_2)$-achievable for arbitrarily small values of $\varepsilon_1, \varepsilon_2 > 0$). 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,

$$\mathcal{R} = \bigcap_{(\varepsilon_1, \varepsilon_2) \in (0,1)^2} \mathcal{R}(\varepsilon_1, \varepsilon_2) \tag{3.6.16}$$

is the capacity region.

The capacity region of a discrete memoryless broadcast channel only depends on its conditional marginal distributions (see, e.g., [175, Lemma 5.1]). This observation implies that the capacity region of a DM-DBC is identical to the capacity region of a discrete memoryless physically degraded broadcast channel when both channels have the same conditional marginal distributions. Consequently, for a DM-DBC, it can be assumed w.l.o.g. that $X \to Y_1 \to Y_2$ is a Markov chain (see, e.g., [175, Section 5.4]).

The capacity region of the DM-DBC is fully characterized, and its achievability was demonstrated by Cover [178] and Bergmans [179] via the use of superposition coding. Weak converses were proved by Wyner [180], Gallager [181], and Ahlswede and Körner [182], and a strong converse was proved by Ahlswede, Gács and Körner [73].

In the absence of a common message, the capacity region of the DM-DBC is provided as follows (see, e.g., [175, Theorem 5.2]).

Theorem 3.6.3. A rate pair (R_1, R_2) is achievable for the DM-DBC (T_1, T_2), characterized by (3.6.12) with $P_{Y|X} = T_1$ and $P_{Z|X} = T_2$, if and only if

$$R_1 \leq I(X;Y|U), \quad R_2 \leq I(U;Z) \tag{3.6.17}$$

for an auxiliary random variable U taking its values in $\mathcal{U}$ such that $U \to X \to Y \to Z$ is a Markov chain, and $|\mathcal{U}| \leq \min\{|\mathcal{X}|, |\mathcal{Y}|, |\mathcal{Z}|\} + 1$.

The *strong converse* for the DM-DBC, due to Ahlswede, Gács and Körner [73], states that allowing for nonvanishing probabilities of error does not enlarge the achievable region:

Theorem 3.6.4 (Strong converse for the DM-DBC)**.**

$$\mathcal{R}(\varepsilon_1, \varepsilon_2) = \mathcal{R}, \qquad \forall\, (\varepsilon_1, \varepsilon_2) \in (0,1)^2. \tag{3.6.18}$$

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 an arbitrary code, one can "blow up" the decoding sets in such a way that the probability of decoding error can be as small as desired (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 Definition 3.4. Note, however, that these blown-up sets transform the original code into a *list code* with a subexponential list size, and one can use a generalization of Fano's inequality for list decoding (see Appendix 3.E) to get nontrivial converse bounds.

Proof (Theorem 3.6.4). Given $\widetilde{\varepsilon}_1, \widetilde{\varepsilon}_2 \in (0,1)$, 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 DM-DBC (T_1, T_2) with

$$\widetilde{\mathcal{D}}_1 = \left\{\widetilde{D}_{1,i}\right\}_{i=1}^{M_1}, \qquad \widetilde{\mathcal{D}}_2 = \left\{\widetilde{D}_{2,j}\right\}_{j=1}^{M_2}.$$

By hypothesis, the decoding sets in $\widetilde{\mathcal{D}}_1$ and $\widetilde{\mathcal{D}}_2$ satisfy

$$\min_{1\le i\le M_1} \min_{1\le j\le M_2} T_1^n\left(\widetilde{D}_{1,i} \,|\, f_n(i,j)\right) \ge 1-\widetilde{\varepsilon}_1, \tag{3.6.19a}$$

$$\min_{1\le i\le M_1} \min_{1\le j\le M_2} T_2^n\left(\widetilde{D}_{2,j} \,|\, f_n(i,j)\right) \ge 1-\widetilde{\varepsilon}_2. \tag{3.6.19b}$$

For an arbitrary $\alpha > 0$, let $\{\delta_n\}$ be the following positive sequence:

$$\delta_n = \sqrt{\frac{1}{2n}\ln\left(\frac{1}{1-\max\{\widetilde{\varepsilon}_1,\widetilde{\varepsilon}_2\}}\right)} + \sqrt{\frac{\alpha\ln n}{n}}, \quad \forall\, n\in\mathbb{N}. \tag{3.6.20}$$

Note that, as $n\to\infty$,

$$n^\beta\,\delta_n \to 0, \quad \forall\,\beta < \tfrac{1}{2}, \tag{3.6.21}$$

$$\sqrt{n}\,\delta_n \to \infty. \tag{3.6.22}$$

For each $i\in\{1,\ldots,M_1\}$ and $j\in\{1,\ldots,M_2\}$, define the "blown-up" decoding sets

$$D_{1,i} \triangleq \left[\widetilde{D}_{1,i}\right]_{n\delta_n}, \qquad D_{2,j} \triangleq \left[\widetilde{D}_{2,j}\right]_{n\delta_n}. \tag{3.6.23}$$

We rely in the following on Lemma 3.6.1 with (3.6.10) and (3.6.11). Note that the correspondence between (3.6.10) and (3.6.20) is

$$\xi_n = \frac{1}{n}\ln\left(\frac{1}{1-\max\{\widetilde{\varepsilon}_1,\widetilde{\varepsilon}_2\}}\right), \quad \forall\, n\in\mathbb{N} \tag{3.6.24}$$

which follows by comparing the condition in the left side of (3.6.4) and (3.6.19), yielding the equation $\exp(-n\xi_n) = 1-\max\{\widetilde{\varepsilon}_1,\widetilde{\varepsilon}_2\}$. From

(3.6.19), the blown-up decoding sets in (3.6.23) with the sequence $\{\delta_n\}$ defined in (3.6.20) imply that, for every $n \in \mathbb{N}$,

$$\min_{1\le i\le M_1} \min_{1\le j\le M_2} T_1^n\left(D_{1,i}\middle|f_n(i,j)\right) \ge 1 - n^{-2\alpha}, \tag{3.6.25a}$$

$$\min_{1\le i\le M_1} \min_{1\le j\le M_2} T_2^n\left(D_{2,j}\middle|f_n(i,j)\right) \ge 1 - n^{-2\alpha}. \tag{3.6.25b}$$

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 (3.6.25). Note, however, that this new object is not a code because the blown-up sets $\mathcal{D}_1$ are not disjoint, and the same holds for the blown-up sets $\mathcal{D}_2$. On the other hand, each given n-tuple $y^n \in \mathcal{Y}^n$ belongs to a subexponential number of the $D_{1,i}$'s, and the same applies to $D_{2,j}$'s. More precisely, let us define the sets

$$\mathcal{N}_1(y^n) \triangleq \{i\colon y^n \in D_{1,i}\}, \quad \forall\, y^n \in \mathcal{Y}^n, \tag{3.6.26a}$$

$$\mathcal{N}_2(z^n) \triangleq \{j\colon z^n \in D_{2,j}\}, \quad \forall\, z^n \in \mathcal{Z}^n. \tag{3.6.26b}$$

A simple combinatorial argument (to be explained) shows that there exists a positive sequence $\{\eta_n\}_{n=1}^{\infty}$ such that $\eta_n \to 0$ as $n \to \infty$, and

$$|\mathcal{N}_1(y^n)| \le \exp(n\eta_n), \qquad \forall\, y^n \in \mathcal{Y}^n, \tag{3.6.27a}$$

$$|\mathcal{N}_2(z^n)| \le \exp(n\eta_n), \qquad \forall\, z^n \in \mathcal{Z}^n. \tag{3.6.27b}$$

In order to get an explicit expression for $\{\eta_n\}$, for every $y^n \in \mathcal{Y}^n$ and $r \ge 0$, let $\mathcal{B}_r(y^n) \subseteq \mathcal{Y}^n$ denote the ball of d_n-radius r centered at y^n:

$$\mathcal{B}_r(y^n) \triangleq \{\hat{y}^n \in \mathcal{Y}^n\colon d_n(\hat{y}^n, y^n) \le r\} \equiv \{y^n\}_r \tag{3.6.28}$$

where d_n is the Hamming metric (3.6.3), and $\{y^n\}_r$ denotes the r-blowup of the singleton set $\{y^n\}$. Since $\delta_n \to 0$ as $n \to \infty$, there exists $n_0 \in \mathbb{N}$ such that $\delta_n + \frac{1}{n} \le \frac{1}{2}$ for every $n \ge n_0$. Consequently, it follows that for every $n \ge n_0$ and $y^n \in \mathcal{Y}^n$,

$$|\mathcal{N}_1(y^n)| \le |\mathcal{B}_{n\delta_n}(y^n)| \tag{3.6.29}$$

$$= \sum_{i=0}^{\lceil n\delta_n\rceil} \binom{n}{i} |\mathcal{Y}|^i \tag{3.6.30}$$

$$\le (\lceil n\delta_n\rceil + 1)\binom{n}{\lceil n\delta_n\rceil} |\mathcal{Y}|^{\lceil n\delta_n\rceil} \tag{3.6.31}$$

$$\le (n\delta_n + 2)\, \exp\left(n\, h\Big(\delta_n + \frac{1}{n}\Big)\right) |\mathcal{Y}|^{n\delta_n+1}, \tag{3.6.32}$$

where (3.6.29) and (3.6.30) hold, respectively, due to (3.6.26) and (3.6.28); (3.6.31) holds since, for $n \geq n_0$, we have $\lceil n\delta_n \rceil \leq \lfloor \frac{n}{2} \rfloor$, and the binomial coefficients $\{\binom{n}{k}\}$ are monotonically increasing in k if $k \leq \lfloor \frac{n}{2} \rfloor$; (3.6.32) holds since $\binom{n}{k} \leq \exp\left(n\, h\left(\frac{k}{n}\right)\right)$ if $k \leq \lfloor \frac{n}{2} \rfloor$ where h denotes the binary entropy function; similarly,

$$|\mathcal{N}_2(z^n)| \leq (n\delta_n + 2)\, \exp\left(n\, h\left(\delta_n + \frac{1}{n}\right)\right)\, |\mathcal{Z}|^{n\delta_n + 1} \tag{3.6.33}$$

for all $n \geq n_0$ and $z^n \in \mathcal{Z}^n$. From (3.6.27), (3.6.32) and (3.6.33), the sequence $\{\eta_n\}$ can be defined such that for all $n \geq n_0$

$$\eta_n = \frac{\ln(n\delta_n + 2)}{n} + h\left(\delta_n + \frac{1}{n}\right) + \left(\delta_n + \frac{1}{n}\right) a, \tag{3.6.34}$$

with $a \triangleq \log\left(\max\{|\mathcal{Y}|, |\mathcal{Z}|\}\right)$; hence, by letting $n \to \infty$, $\eta_n \to 0$.

We are now ready to apply a generalization of Fano's inequality for list decoding [182]. To this end, for every $j \in \{1, \ldots, M_2\}$, let

$$\mathcal{T}(j) \triangleq \{f_n(i, j) \colon 1 \leq i \leq M_1\}, \tag{3.6.35}$$

let U be an equiprobable random variable on $\{1, \ldots, M_2\}$, and let $X^n \in \mathcal{X}^n$ be an equiprobable random variable on $\mathcal{T}(U)$. Finally, let $Y^n \in \mathcal{Y}^n$ and $Z^n \in \mathcal{Z}^n$ be generated from X^n via the DMCs T_1^n and T_2^n, respectively. Now, consider the error event of the second receiver (which corresponds to the degraded channel T_2^n); the error event of a list decoder for the second receiver refers to the case where $U \notin \mathcal{N}_2(Z^n)$. Let ζ_n be the error probability of the list decoder for the blown-up sets $\mathcal{D}_2$. Consequently, we get

$$\begin{aligned}
\tfrac{1}{n} \ln M_2 &= \tfrac{1}{n}\, H(U) && (3.6.36)\\
&= \tfrac{1}{n}\, [I(U; Z^n) + H(U|Z^n)] && (3.6.37)\\
&\leq \tfrac{1}{n} \left[I(U; Z^n) + h(\zeta_n) + \zeta_n \ln M_2\right] + (1 - \zeta_n)\eta_n && (3.6.38)\\
&= \tfrac{1}{n}\, I(U; Z^n) + o(1), && (3.6.39)
\end{aligned}$$

where (3.6.38) follows from a modification of Fano's inequality for list decoding (see Appendix 3.E) together with (3.6.27); (3.6.39) uses the fact that $\eta_n \to 0$ and, by (3.6.25), $\zeta_n \leq n^{-2\alpha}$ for some $\alpha > 0$, so also

$\zeta_n \to 0$ as $n \to \infty$. Using a similar argument, we can also prove that

$$\tfrac{1}{n} \ln M_1 \leq \tfrac{1}{n} I(X^n; Y^n|U) + o(1). \tag{3.6.40}$$

By the weak converse for the DM-DBC [182], the rate pair (R_1, R_2) with $R_1 = \frac{1}{n} I(X^n; Y^n|U)$ and $R_2 = \frac{1}{n} I(U; Z^n)$ is included in the achievable region $\mathcal{R}$. Since every element of $\mathcal{R}(\varepsilon_1, \varepsilon_2)$ can be expressed as a limit of rates $\left(\frac{1}{n} \ln M_1, \frac{1}{n} \ln M_2\right)$ in the region $\mathcal{R}$, 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 the reverse inclusion is trivial by (3.6.16). $\square$

3.6.3 Lossless source coding with side information

Our second example of the use of the blowing-up lemma to prove a strong converse is a bit more sophisticated, and concerns the problem of lossless source coding with side information. Let $\mathcal{X}$ and $\mathcal{Y}$ be finite sets, and $\{(X_i, Y_i)\}_{i=1}^{\infty}$ be a sequence of i.i.d. samples drawn from a given joint distribution $P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. The $\mathcal{X}$-valued and the $\mathcal{Y}$-valued parts of this sequence are observed by two independent encoders. An (n, M_1, M_2)-code is a triple $\mathcal{C} = (f_n^{(1)}, f_n^{(2)}, g_n)$, where

$$f_n^{(1)} \colon \mathcal{X}^n \to \{1, \ldots, M_1\} \tag{3.6.41}$$

and

$$f_n^{(2)} \colon \mathcal{Y}^n \to \{1, \ldots, M_2\} \tag{3.6.42}$$

are the encoding maps and

$$g_n \colon \{1, \ldots, M_1\} \times \{1, \ldots, M_2\} \to \mathcal{Y}^n \tag{3.6.43}$$

is the decoding map. The decoder observes

$$J_n^{(1)} = f_n^{(1)}(X^n), \quad J_n^{(2)} = f_n^{(2)}(Y^n), \tag{3.6.44}$$

and it wishes to reconstruct Y^n with a small probability of error. The reconstruction is given by

$$\begin{aligned} \widehat{Y}^n &= g_n(J_n^{(1)}, J_n^{(2)}) \\ &= g_n(f_n^{(1)}(X^n), f_n^{(2)}(Y^n)). \end{aligned} \tag{3.6.45}$$

We say that $\mathcal{C} = \left(f_n^{(1)}, f_n^{(2)}, g_n\right)$ is an $(n, M_1, M_2, \varepsilon)$-*code* if

$$\mathbb{P}\left(\widehat{Y}^n \neq Y^n\right) = \mathbb{P}\left(g_n(f_n^{(1)}(X^n), f_n^{(2)}(Y^n)) \neq Y^n\right) \leq \varepsilon. \tag{3.6.46}$$

We say that a rate pair (R_1, R_2) is ε-*achievable* if, for any $\delta > 0$ and sufficiently large $n \in \mathbb{N}$, there exists an $(n, M_1, M_2, \varepsilon)$-code $\mathcal{C}$ with

$$\frac{1}{n} \ln M_k \leq R_k + \delta, \qquad k = 1, 2. \tag{3.6.47}$$

A rate pair (R_1, R_2) is *achievable* if it is ε-achievable for all $\varepsilon \in (0, 1]$. Again, let $\mathcal{R}(\varepsilon)$ (resp., $\mathcal{R}$) denote the set of all ε-achievable (resp., achievable) rate pairs. Clearly,

$$\mathcal{R} = \bigcap_{\varepsilon \in (0,1]} \mathcal{R}(\varepsilon). \tag{3.6.48}$$

The following characterization of the achievable region was obtained by Ahlswede and Körner [182]:

Theorem 3.6.5. A rate pair (R_1, R_2) is achievable if and only if there exist random variables $U \in \mathcal{U}$, $X \in \mathcal{X}$, $Y \in \mathcal{Y}$ such that $U \to X \to Y$ is a Markov chain, (X, Y) has the given joint distribution P_{XY}, and

$$R_1 \geq I(X; U), \tag{3.6.49}$$

$$R_2 \geq H(Y|U). \tag{3.6.50}$$

Moreover, the domain $\mathcal{U}$ of U can be chosen so that $|\mathcal{U}| \leq |\mathcal{X}| + 2$.

Our goal is to prove the corresponding *strong converse*, originally established by Ahlswede, Gács and J. Körner [73], which states that allowing for a nonvanishing error probability, as in (3.6.46), does not asymptotically enlarge the achievable region:

Theorem 3.6.6 (Strong converse theorem for lossless source coding with side information)**.**

$$\mathcal{R}(\varepsilon) = \mathcal{R}, \qquad \forall\, \varepsilon \in (0, 1). \tag{3.6.51}$$

In preparation for the proof of Theorem 3.6.6, we need to introduce some additional terminology and definitions.

Definition 3.6. [173, Chapter 6] Given two finite sets $\mathcal{U}$ and $\mathcal{V}$, a DMC $S\colon \mathcal{U} \to \mathcal{V}$, and a parameter $\eta \in (0,1]$, we say that a set $\mathcal{B} \subseteq \mathcal{V}$ is a *η-image of $u \in \mathcal{U}$ under S* if $S(\mathcal{B}|u) \geq \eta$.

For any $\mathcal{B} \subseteq \mathcal{V}$, let $\mathcal{D}_\eta(\mathcal{B}; S) \subseteq \mathcal{U}$ denote the set of all $u \in \mathcal{U}$ such that $\mathcal{B}$ is a η-image of u under S:

$$\mathcal{D}_\eta(\mathcal{B}; S) \triangleq \Big\{ u \in \mathcal{U} : S(\mathcal{B}|u) \geq \eta \Big\}. \tag{3.6.52}$$

Definition 3.7. Let P be a probability measure on $\mathcal{X}$, and $n \in \mathbb{N}$. A sequence $x^n \in \mathcal{X}^n$ is a *(P,δ)-typical sequence* if, for every $a \in \mathcal{X}$,

$$\left| \frac{1}{n} N(a|x^n) - P(a) \right| < \delta \tag{3.6.53}$$

where $N(a|x^n)$ denotes the number of appearances of a in x^n. The *(P,δ)-typical set* is the set of all these typical sequences.

Let $P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, and let $T\colon \mathcal{X} \to \mathcal{Y}$ be a DMC with the conditional probability distribution $P_{Y|X}$. For $Q_Y \in \mathcal{P}(\mathcal{Y})$ which is strictly positive, and parameters $c \geq 0$ and $\varepsilon \in (0,1]$, let

$$\begin{aligned} &\widehat{\Gamma}_n(c, \varepsilon; Q_Y) \\ &\triangleq \min_{\mathcal{B} \subseteq \mathcal{Y}^n} \left\{ \frac{1}{n} \ln Q_Y^n(\mathcal{B}) : \frac{1}{n} \ln P_X^n \left(\mathcal{D}_{1-\varepsilon}(\mathcal{B}; T^n) \cap \mathcal{T}^n_{[X,\delta_n]} \right) \geq -c \right\} \end{aligned} \tag{3.6.54}$$

where P_X is the marginal distribution of X, and $\mathcal{T}^n_{[X,\delta_n]} \subset \mathcal{X}^n$ denotes the (P_X, δ_n)-typical set with an arbitrary sequence $\{\delta_n\}$ that satisfies

$$\delta_n \to 0, \quad \sqrt{n}\delta_n \to \infty. \tag{3.6.55}$$

Theorem 3.6.7. For any $c \geq 0$ and any $\varepsilon \in (0,1]$,

$$\lim_{n\to\infty} \widehat{\Gamma}_n(c, \varepsilon; Q_Y) = \Gamma(c; Q_Y), \tag{3.6.56}$$

where

$$\Gamma(c; Q_Y) \triangleq - \max_{\substack{U \in \mathcal{U}, \\ |\mathcal{U}| \leq |\mathcal{X}|+2}} \Big\{ D(P_{Y|U} \| Q_Y | P_U) : U \to X \to Y,\, I(X;U) \leq c \Big\}. \tag{3.6.57}$$

Moreover, the function $c \mapsto \Gamma(c; Q_Y)$ is continuous.

Proof. The proof consists of two major steps. The first is to show that (3.6.56) holds for a suitable sequence $\varepsilon_n \to 0$, and that the limit $\Gamma(c; Q_Y)$ is equal to (3.6.57). We omit the details of this step and refer the reader to the original paper by Ahlswede, Gács and Körner [73]. The second step, which relies on the blowing-up lemma, is to show that for all $\varepsilon \in (0, 1]$

$$\lim_{n\to\infty} \left[\widehat{\Gamma}_n(c, \varepsilon; Q_Y) - \widehat{\Gamma}_n(c, \varepsilon_n; Q_Y)\right] = 0. \tag{3.6.58}$$

To that end, let ε be fixed and let $\{\delta_n\}$ be chosen to satisfy (3.6.55). For a fixed n, let $\mathcal{B} \subseteq \mathcal{Y}^n$ be a set such that $T^n(\mathcal{B}|x^n) \geq 1 - \varepsilon$ for some $x^n \in \mathcal{X}^n$. Then, we get

$$T^n(\mathcal{B}_{n\delta_n}|x^n) \geq 1 - \exp\left[-2\left(\sqrt{n}\,\delta_n - \sqrt{\tfrac{1}{2}\ln\tfrac{1}{1-\varepsilon}}\right)^2\right] \tag{3.6.59}$$

$$\triangleq 1 - \varepsilon_n \tag{3.6.60}$$

where (3.6.59) holds by Lemma 3.6.2 with $r = n\delta_n$; the sequence $\{\varepsilon_n\}$ in (3.6.60) converges to zero in view of (3.6.55) where $\sqrt{n}\delta_n \to \infty$. Consequently, it follows from (3.6.52) that for all sufficiently large n

$$\mathcal{D}_{1-\varepsilon_n}(\mathcal{B}_{n\delta_n}; T^n) \cap \mathcal{T}^n_{[X,\delta_n]} \supseteq \mathcal{D}_{1-\varepsilon}(\mathcal{B}; T^n) \cap \mathcal{T}^n_{[X,\delta_n]}. \tag{3.6.61}$$

On the other hand, since Q_Y is strictly positive,

$$Q_Y^n(\mathcal{B}_{n\delta_n}) = \sum_{y^n \in \mathcal{B}_{n\delta_n}} Q_Y^n(y^n) \tag{3.6.62}$$

$$\leq \sum_{y^n \in \mathcal{B}} Q_Y^n\left(\mathcal{B}_{n\delta_n}(y^n)\right) \tag{3.6.63}$$

$$\leq \sup_{y^n \in \mathcal{Y}^n} \frac{Q_Y^n\left(\mathcal{B}_{n\delta_n}(y^n)\right)}{Q_Y^n(y^n)} \sum_{y^n \in \mathcal{B}} Q_Y^n(y^n) \tag{3.6.64}$$

$$= \sup_{y^n \in \mathcal{Y}^n} \frac{Q_Y^n\left(\mathcal{B}_{n\delta_n}(y^n)\right)}{Q_Y^n(y^n)} \cdot Q_Y^n(\mathcal{B}). \tag{3.6.65}$$

Using (3.6.65) together with the fact that (see [73, Lemma 5])

$$\lim_{n\to\infty} \frac{1}{n} \ln \sup_{y^n \in \mathcal{Y}^n} \frac{Q_Y^n\left(\mathcal{B}_{n\delta_n}(y^n)\right)}{Q_Y^n(y^n)} = 0 \tag{3.6.66}$$

yields

$$\lim_{n\to\infty} \sup_{\mathcal{B}\subseteq\mathcal{Y}^n} \frac{1}{n} \ln \frac{Q_Y^n\left(\mathcal{B}_{n\delta_n}\right)}{Q_Y^n(\mathcal{B})} = 0. \tag{3.6.67}$$

From (3.6.61) and (3.6.67), we get (3.6.58). The bound on $|\mathcal{U}|$ in (3.6.57) follows from Carathéodory's theorem [73]. □

We are now ready to prove Theorem 3.6.6.

Proof. Let $\mathcal{C} = (f_n^{(1)}, f_n^{(2)}, g_n)$ be an arbitrary $(n, M_1, M_2, \varepsilon)$-code. For a given index $j \in \{1, \ldots, M_1\}$, we define the set

$$\mathcal{Y}^n(j) \triangleq \left\{ y^n \in \mathcal{Y}^n : y^n = g_n\left(j, f_n^{(2)}(y^n)\right) \right\}, \tag{3.6.68}$$

which consists of all the sequences $y^n \in \mathcal{Y}^n$ that are correctly decoded for any $x^n \in \mathcal{X}^n$ such that $f_n^{(1)}(x^n) = j$. Using this notation, we can write

$$\mathbb{E}\left[T^n(\mathcal{Y}^n(f_n^{(1)}(X^n))|X^n)\right] \geq 1 - \varepsilon. \tag{3.6.69}$$

If we define the set

$$\mathcal{A}_n \triangleq \left\{ x^n \in \mathcal{X}^n : T^n(\mathcal{Y}^n(f_n^{(1)}(x^n))|x^n) \geq 1 - \sqrt{\varepsilon} \right\}, \tag{3.6.70}$$

then, using the so-called "reverse Markov inequality"[3] and (3.6.69), we see that

$$P_X^n(\mathcal{A}_n) = 1 - P_X^n(\mathcal{A}_n^{\mathrm{c}}) \tag{3.6.71}$$

$$= 1 - P_X^n\left(\underbrace{T^n\Big(\mathcal{Y}^n(f_n^{(1)}(X^n)) \,|\, X^n\Big)}_{\leq 1} < 1 - \sqrt{\varepsilon} \right) \tag{3.6.72}$$

$$\geq 1 - \frac{1 - \mathbb{E}\left[T^n(\mathcal{Y}^n(f_n^{(1)}(X^n))|X^n)\right]}{1 - (1 - \sqrt{\varepsilon})} \tag{3.6.73}$$

$$\geq 1 - \frac{1 - (1 - \varepsilon)}{\sqrt{\varepsilon}} = 1 - \sqrt{\varepsilon}. \tag{3.6.74}$$

[3] The reverse Markov inequality states that if Y is a random variable such that $Y \leq b$ a.s. for some constant b, then for all $a < b$

$$\mathbb{P}(Y \leq a) \leq \frac{b - \mathbb{E}[Y]}{b - a}.$$

Consequently, for all sufficiently large n, we have

$$P_X^n\left(\mathcal{A}_n \cap \mathcal{T}_{[X,\delta_n]}^n\right) \geq 1 - 2\sqrt{\varepsilon}. \tag{3.6.75}$$

This implies, in turn, that there exists some $j^* \in f_n^{(1)}(\mathcal{X}^n)$, such that

$$P_X^n\left(\mathcal{D}_{1-\sqrt{\varepsilon}}(\mathcal{Y}^n(j^*)) \cap \mathcal{T}_{[X,\delta_n]}^n\right) \geq \frac{1-2\sqrt{\varepsilon}}{M_1}. \tag{3.6.76}$$

On the other hand,

$$M_2 = \left|f_n^{(2)}(\mathcal{Y}^n)\right| \geq |\mathcal{Y}^n(j^*)|. \tag{3.6.77}$$

We are now in a position to apply Theorem 3.6.7. If we choose Q_Y to be the equiprobable distribution on $\mathcal{Y}$, then it follows from (3.6.76) and (3.6.77) that

$$\begin{aligned}
\tfrac{1}{n}\ln M_2 &\geq \tfrac{1}{n}\ln|\mathcal{Y}^n(j^*)| && (3.6.78)\\
&= \tfrac{1}{n}\ln Q_Y^n(\mathcal{Y}^n(j^*)) + \ln|\mathcal{Y}| && (3.6.79)\\
&\geq \widehat{\Gamma}_n\left(-\tfrac{1}{n}\ln(1-2\sqrt{\varepsilon}) + \tfrac{1}{n}\ln M_1, \sqrt{\varepsilon};\, Q_Y\right) + \ln|\mathcal{Y}|. && (3.6.80)
\end{aligned}$$

Using Theorem 3.6.7, we conclude that the bound

$$\tfrac{1}{n}\ln M_2 \geq \Gamma\left(-\tfrac{1}{n}\ln(1-2\sqrt{\varepsilon}) + \tfrac{1}{n}\ln M_1;\, Q_Y\right) + \ln|\mathcal{Y}| + o(1) \tag{3.6.81}$$

holds for any $(n, M_1, M_2, \varepsilon)$-code. If $(R_1, R_2) \in \mathcal{R}(\varepsilon)$, then there exists a sequence $\{\mathcal{C}_n\}_{n=1}^{\infty}$, where each $\mathcal{C}_n = (f_n^{(1)}, f_n^{(2)}, g_n)$ is an $(n, M_{1,n}, M_{2,n}, \varepsilon)$-code, and

$$\lim_{n\to\infty} \tfrac{1}{n}\ln M_{k,n} = R_k, \qquad k = 1, 2. \tag{3.6.82}$$

Using this in (3.6.81), together with the continuity of the mapping $c \mapsto \Gamma(c; Q_Y)$, we get

$$R_2 \geq \Gamma(R_1; Q_Y) + \ln|\mathcal{Y}|, \qquad \forall (R_1, R_2) \in \mathcal{R}(\varepsilon). \tag{3.6.83}$$

By the definition of Γ in (3.6.57), there exists a triple $U \to X \to Y$ such that $I(X;U) \leq R_1$ and

$$\Gamma(R_1; Q_Y) = -D(P_{Y|U}\|Q_Y|P_U) = -\ln|\mathcal{Y}| + H(Y|U), \tag{3.6.84}$$

where the last equality is due to the fact that $U \to X \to Y$ is a Markov chain, and Q_Y is an equiprobable distribution on $\mathcal{Y}$. Therefore, (3.6.83) and (3.6.84) imply that

$$R_2 \geq H(Y|U). \tag{3.6.85}$$

Consequently, the triple $(U, X, Y) \in \mathcal{R}$ by Theorem 3.6.5, and hence $\mathcal{R}(\varepsilon) \subseteq \mathcal{R}$ for all $\varepsilon > 0$. Since $\mathcal{R} \subseteq \mathcal{R}(\varepsilon)$ by definition, the proof of Theorem 3.6.6 is completed. □

3.6.4 The empirical distribution of good channel codes with non-vanishing error probability

A more recent application of concentration of measure inequalities to information theory has to do with the characterization of the stochastic behavior of output sequences of good channel codes. Conceptually, the random coding argument, originally used by Shannon and many times since, to show the existence of good channel codes suggests that the input (respectively, output) sequence of such a code should resemble a typical realization of a sequence of i.i.d. random variables sampled from a capacity-achieving input (respectively, output) distribution. For capacity-achieving sequences of codes with asymptotically vanishing probability of error, this intuition has been rigorously analyzed by Shamai and Verdú who proved the following remarkable statement [183, Theorem 2]: given a DMC $T\colon \mathcal{X} \to \mathcal{Y}$, every capacity-achieving sequence of channel codes with asymptotically vanishing (maximal or average) probability of error has the property that

$$\lim_{n\to\infty} \frac{1}{n} D(P_{Y^n} \| P^*_{Y^n}) = 0, \tag{3.6.86}$$

where, for each n, P_{Y^n} denotes the output distribution on $\mathcal{Y}^n$ induced by the code (assuming that the messages are equiprobable), while $P^*_{Y^n}$ denotes the product of n copies of the single-letter capacity-achieving output distribution. In fact, the convergence in (3.6.86) holds not just for DMCs, but for arbitrary channels satisfying the condition

$$C = \lim_{n\to\infty} \frac{1}{n} \sup_{P_{X^n} \in \mathcal{P}(\mathcal{X}^n)} I(X^n; Y^n). \tag{3.6.87}$$

(These ideas go back to the work of Han and Verdú on approximation theory of output statistics, see [184, Theorem 15]). In a recent paper [63], Polyanskiy and Verdú extended the results of [183] for codes with *nonvanishing* probability of error, provided one uses the maximal probability of error criterion and deterministic encoders.

In this section, we present some of the results from [63, 185] in the context of the material covered earlier in this chapter. To keep things simple, we 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 calculated by solving the optimization problem

$$C = \max_{P_X \in \mathcal{P}(\mathcal{X})} I(X;Y), \tag{3.6.88}$$

where X and Y are related via T. Let $P_X^* \in \mathcal{P}(\mathcal{X})$ be a capacity-achieving input distribution (there may be several). It can be shown [186, 187] that the corresponding output distribution $P_Y^* \in \mathcal{P}(\mathcal{Y})$ is unique, and for every $n \in \mathbb{N}$, the product distribution $P_{Y^n}^* \equiv (P_Y^*)^{\otimes n}$ has the key property

$$D(P_{Y^n|X^n=x^n} \| P_{Y^n}^*) \le nC, \qquad \forall\, x^n \in \mathcal{X}^n \tag{3.6.89}$$

where $P_{Y^n|X^n=x^n}$ is shorthand for the product distribution $T^n(\cdot|x^n)$. From (3.6.89), we see that the capacity-achieving output distribution $P_{Y^n}^*$ dominates every output distribution P_{Y^n} induced by an arbitrary input distribution $P_{X^n} \in \mathcal{P}(\mathcal{X}^n)$:

$$P_{Y^n|X^n=x^n} \ll P_{Y^n}^*,\ \forall\, x^n \in \mathcal{X}^n \implies P_{Y^n} \ll P_{Y^n}^*,\ \forall\, P_{X^n} \in \mathcal{P}(\mathcal{X}^n).$$

This has two important consequences:

1. The information density is well-defined for every $x^n \in \mathcal{X}^n$ and $y^n \in \mathcal{Y}^n$:

$$i_{X^n;Y^n}^*(x^n;y^n) \triangleq \ln \frac{\mathrm{d}P_{Y^n|X^n=x^n}}{\mathrm{d}P_{Y^n}^*}(y^n). \tag{3.6.90}$$

2. For an input distribution P_{X^n}, the respective output distribution P_{Y^n} satisfies

$$D(P_{Y^n} \| P_{Y^n}^*) \le nC - I(X^n;Y^n). \tag{3.6.91}$$

Indeed, by the chain rule for divergence, it follows that for every input distribution $P_{X^n} \in \mathcal{P}(\mathcal{X}^n)$

$$\begin{aligned} I(X^n;Y^n) &= D(P_{Y^n|X^n}\|P_{Y^n}|P_{X^n}) \\ &= D(P_{Y^n|X^n}\|P^*_{Y^n}|P_{X^n}) - D(P_{Y^n}\|P^*_{Y^n}) \\ &\le nC - D(P_{Y^n}\|P^*_{Y^n}). \end{aligned} \tag{3.6.92}$$

Inequality (3.6.91) follows upon rearranging terms in (3.6.92).

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

$$\max_{1\le i\le M} \mathbb{P}(g_n(Y^n) \ne i \,\big|\, X^n = f_n(i)) \le \varepsilon. \tag{3.6.93}$$

Remark 3.32. Polyanskiy and Verdú use a more precise nomenclature in [63] and say that every such $\mathcal{C} = (f_n, g_n)$ satisfying (3.6.93) 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 only consider codes of this type, so we adhere to our simplified terminology.

Consider an arbitrary (n, M)-code $\mathcal{C} = (f_n, g_n)$ for T, and let J be a random variable having an equiprobable distribution on $\{1,\ldots,M\}$. Hence, we can think of every $i \in \{1\ldots,M\}$ 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 [63] is that the output distribution $P^{(\mathcal{C})}_{Y^n}$ of every (n, M, ε)-code satisfies

$$D(P^{(\mathcal{C})}_{Y^n}\|P^*_{Y^n}) \le nC - \ln M + o(n); \tag{3.6.94}$$

moreover, the $o(n)$ term was refined in [63, Theorem 5] to $O(\sqrt{n})$ for every 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 determine an explicit form of the term that scales like $O(\sqrt{n})$.

Just as in [63], the proof of (3.6.94) with the $O(\sqrt{n})$ term uses the following strong converse for channel codes due to Augustin [188] (see also [63, Theorem 1] and [189, Section 2]):

Theorem 3.6.8 (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 every $M \in \mathbb{N}$ and $0 < \varepsilon \leq 1$, let $f\colon \{1, \dots, M\} \to \mathcal{U}$ and $g\colon \mathcal{V} \to \{1, \dots, M\}$ be mappings, such that

$$\max_{1 \leq i \leq M} \mathbb{P}(g(V) \neq i \,|\, U = f(i)) \leq \varepsilon. \tag{3.6.95}$$

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 result holds:

$$M \leq \frac{\exp\bigl(\mathbb{E}[\gamma(U)]\bigr)}{\inf\limits_{u \in \mathcal{U}} P_{V|U=u}\left(\ln \dfrac{\mathrm{d}P_{V|U=u}}{\mathrm{d}Q_V}\,(V) < \gamma(u)\right) - \varepsilon}, \tag{3.6.96}$$

provided that the denominator in the right side of (3.6.96) is strictly positive. The expectation in the numerator is taken with respect to the distribution of $U = f(J)$ where J is equiprobable on $\{1, \dots, M\}$.

We first establish the bound (3.6.94) for the case when the DMC T is such that

$$C_1 \triangleq \max_{x, x' \in \mathcal{X}} D(P_{Y|X=x} \| P_{Y|X=x'}) < \infty. \tag{3.6.97}$$

Note that $C_1 < \infty$ if and only if the transition matrix of T does not have any zeroes. Consequently,

$$c(T) \triangleq 2 \max_{x \in \mathcal{X}} \max_{y, y' \in \mathcal{Y}} \left| \ln \frac{P_{Y|X}(y|x)}{P_{Y|X}(y'|x)} \right| < \infty. \tag{3.6.98}$$

We can now establish the following sharpened version of the bound in [63, Theorem 5]:

Theorem 3.6.9. Let $T\colon \mathcal{X} \to \mathcal{Y}$ be a DMC with $C > 0$ satisfying (3.6.97). Then, every (n, M, ε)-code $\mathcal{C}$ for T with $0 < \varepsilon < 1/2$ satisfies

$$D\bigl(P_{Y^n}^{(\mathcal{C})} \| P_{Y^n}^{*}\bigr) \leq nC - \ln M + \ln \frac{1}{\varepsilon} + c(T) \sqrt{\frac{n}{2} \ln \frac{1}{1 - 2\varepsilon}}. \tag{3.6.99}$$

Remark 3.33. As it is shown in [63], the restriction to codes with deterministic encoders and to the maximal probability of error criterion is necessary both for Theorems 3.6.9 and 3.6.10.

Proof. Fix an input sequence $x^n \in \mathcal{X}^n$, and consider the function $h_{x^n}\colon \mathcal{Y}^n \to \mathbb{R}$ defined for the DMC as follows:

$$h_{x^n}(y^n) \triangleq \ln \frac{\mathrm{d}P_{Y^n|X^n=x^n}}{\mathrm{d}P_{Y^n}^{(\mathcal{C})}}\,(y^n) = \ln \frac{P_{Y^n|X^n=x^n}(y^n)}{P_{Y^n}^{(\mathcal{C})}(y^n)} \tag{3.6.100}$$

for every $y^n \in \mathcal{Y}^n$, which yields

$$\mathbb{E}[h_{x^n}(Y^n)|X^n = x^n] = D(P_{Y^n|X^n=x^n}\|P_{Y^n}^{(\mathcal{C})}). \tag{3.6.101}$$

Moreover, for every $i \in \{1,\ldots,n\}$, $y, y' \in \mathcal{Y}$, and $\overline{y}^i \in \mathcal{Y}^{n-1}$, we have (see the notation used in (3.1.11))

$$\begin{aligned}
&\left|h_{i,x^n}(y|\overline{y}^i) - h_{i,x^n}(y'|\overline{y}^i)\right| \\
&\le \left|\ln P_{Y^n|X^n=x^n}(y^{i-1}, y, y_{i+1}^n) - \ln P_{Y^n|X^n=x^n}(y^{i-1}, y', y_{i+1}^n)\right| \\
&\quad + \left|\ln P_{Y^n}^{(\mathcal{C})}(y^{i-1}, y, y_{i+1}^n) - \ln P_{Y^n}^{(\mathcal{C})}(y^{i-1}, y', y_{i+1}^n)\right| \\
&\le \left|\ln \frac{P_{Y_i|X_i=x_i}(y)}{P_{Y_i|X_i=x_i}(y')}\right| + \left|\ln \frac{P^{(\mathcal{C})}_{Y_i|\overline{Y}^i}(y|\overline{y}^i)}{P^{(\mathcal{C})}_{Y_i|\overline{Y}^i}(y'|\overline{y}^i)}\right| \\
&\le 2 \max_{x\in\mathcal{X}} \max_{y,y'\in\mathcal{Y}} \left|\ln \frac{P_{Y|X}(y|x)}{P_{Y|X}(y'|x')}\right| && (3.6.102) \\
&= c(T) < \infty && (3.6.103)
\end{aligned}$$

where (3.6.102) is justified in Appendix 3.F, and (3.6.103) is due to (3.6.98). Hence, for every $x^n \in \mathcal{X}^n$, the function $h_{x^n}\colon \mathcal{Y}^n \to \mathbb{R}$ satisfies the bounded differences condition (3.3.46) with $c_1 = \ldots = c_n = c(T)$. In view of (3.6.101) and (3.6.103), Theorem 3.3.8 (a restatement of McDiarmid's inequality) implies that for all $r \ge 0$ and $x^n \in \mathcal{X}^n$

$$\begin{aligned}
P_{Y^n|X^n=x^n}\left(\ln \frac{\mathrm{d}P_{Y^n|X^n=x^n}}{\mathrm{d}P_{Y^n}^{(\mathcal{C})}}(Y^n) \ge D(P_{Y^n|X^n=x^n}\|P_{Y^n}^{(\mathcal{C})}) + r\right) \\
\le \exp\left(-\frac{2r^2}{nc^2(T)}\right). && (3.6.104)
\end{aligned}$$

(In fact, the above derivation goes through for every possible output distribution P_{Y^n}, not necessarily one induced by a code). This is where we depart from the original proof by Polyanskiy and Verdú [63]: we use McDiarmid's inequality to control the deviation probability for the "conditional" information density $h_{x^n}(Y^n)$ directly, whereas they bounded the *variance* of $h_{x^n}(Y^n)$ using a Poincaré inequality, and then bounded the deviation probability using Chebyshev's inequality. As it is shown in the sequel, the concentration inequality (3.6.104) allows us to explicitly identify the dependence of the constant multiplying $\sqrt{n}$ in (3.6.99) 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, let $\mathcal{U} = \mathcal{X}^n$, $\mathcal{V} = \mathcal{Y}^n$, and consider the DMC $T^n \colon \mathcal{U} \to \mathcal{V}$ together with an (n, M, ε)-code $\mathcal{C} = (f_n, g_n)$. Furthermore, let

$$\zeta_n = \zeta_n(\varepsilon) \triangleq c(T)\sqrt{\frac{n}{2}\ln\frac{1}{1-2\varepsilon}}, \tag{3.6.105}$$

$$\gamma(x^n) \triangleq D(P_{Y^n|X^n=x^n} \| P_{Y^n}^{(\mathcal{C})}) + \zeta_n. \tag{3.6.106}$$

Using (3.6.96) with the auxiliary distribution $Q_V = P_{Y^n}^{(\mathcal{C})}$, we get

$$M \leq \frac{\exp(\mathbb{E}[\gamma(X^n)])}{\inf\limits_{x^n \in \mathcal{X}^n} P_{Y^n|X^n=x^n}\left(\ln \dfrac{\mathrm{d}P_{Y^n|X^n=x^n}}{\mathrm{d}P_{Y^n}^{(\mathcal{C})}}(Y^n) < \gamma(x^n)\right) - \varepsilon} \tag{3.6.107}$$

where, from (3.6.106),

$$\mathbb{E}[\gamma(X^n)] = D(P_{Y^n|X^n} \| P_{Y^n}^{(\mathcal{C})} \,|\, P_{X^n}^{(\mathcal{C})}) + \zeta_n. \tag{3.6.108}$$

The concentration inequality in (3.6.104) with (3.6.105) and (3.6.106) give that, for every $x^n \in \mathcal{X}^n$,

$$\begin{aligned} P_{Y^n|X^n=x^n}\left(\ln \frac{\mathrm{d}P_{Y^n|X^n=x^n}}{\mathrm{d}P_{Y^n}^{(\mathcal{C})}}(Y^n) \geq \gamma(x^n)\right) &\leq \exp\left(-\frac{2\zeta_n^2}{nc^2(T)}\right) \\ &= 1 - 2\varepsilon, \end{aligned} \tag{3.6.109}$$

which implies that

$$\inf_{x^n \in \mathcal{X}^n} P_{Y^n|X^n=x^n}\left(\ln \frac{\mathrm{d}P_{Y^n|X^n=x^n}}{\mathrm{d}P_{Y^n}^{(\mathcal{C})}}(Y^n) < \gamma(x^n)\right) \geq 2\varepsilon. \tag{3.6.110}$$

Hence, from (3.6.107), (3.6.108) and (3.6.110), it follows that

$$M \leq \frac{1}{\varepsilon} \exp\Big(D(P_{Y^n|X^n} \| P_{Y^n}^{(\mathcal{C})} \,|\, P_{X^n}^{(\mathcal{C})}) + \zeta_n\Big) \tag{3.6.111}$$

so, taking logarithms on both sides of (3.6.111) and rearranging terms gives

$$\begin{aligned} D(P_{Y^n|X^n} \| P_{Y^n}^{(\mathcal{C})} \,|\, P_{X^n}^{(\mathcal{C})}) &\geq \ln M + \ln \varepsilon - \zeta_n \\ &= \ln M + \ln \varepsilon - c(T) \sqrt{\frac{n}{2} \ln \frac{1}{1-2\varepsilon}} \end{aligned} \tag{3.6.112}$$

where (3.6.112) is due to (3.6.105). We are now ready to derive (3.6.99):

$$\begin{aligned} &D(P_{Y^n}^{(\mathcal{C})} \| P_{Y^n}^*) \\ &\quad = D(P_{Y^n|X^n} \| P_{Y^n}^* | P_{X^n}^{(\mathcal{C})}) - D(P_{Y^n|X^n} \| P_{Y^n}^{(\mathcal{C})} | P_{X^n}^{(\mathcal{C})}) && (3.6.113) \\ &\quad \leq nC - \ln M + \ln \frac{1}{\varepsilon} + c(T) \sqrt{\frac{n}{2} \ln \frac{1}{1-2\varepsilon}} && (3.6.114) \end{aligned}$$

where (3.6.113) uses the chain rule for divergence, while (3.6.114) relies on (3.6.89) and (3.6.112). □

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 [63, Theorem 6]:

Theorem 3.6.10. Let $T\colon \mathcal{X} \to \mathcal{Y}$ be a DMC with capacity $C > 0$, and let $\varepsilon \in (0,1)$. Every (n, M, ε)-code $\mathcal{C}$ for T satisfies

$$D(P_{Y^n}^{(\mathcal{C})} \| P_{Y^n}^*) \leq nC - \ln M + O\left(\sqrt{n}\,(\ln n)^{3/2}\right) \tag{3.6.115}$$

and, more precisely, for every such code

$$\begin{aligned} D\big(P_{Y^n}^{(\mathcal{C})} \| P_{Y^n}^*\big) &\leq nC - \ln M \\ &\quad + \sqrt{2n}\,(\ln n)^{3/2} \left(1 + \sqrt{\frac{1}{\ln n} \ln \frac{1}{1-\varepsilon}}\right) \left(1 + \frac{\ln |\mathcal{Y}|}{\ln n}\right) \\ &\quad + 3 \ln n + \ln(2|\mathcal{X}||\mathcal{Y}|^2). \end{aligned} \tag{3.6.116}$$

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 $\widetilde{\mathcal{D}}_1, \ldots, \widetilde{\mathcal{D}}_M \subseteq \mathcal{Y}^n$ be the respective decoding sets:

$$\widetilde{\mathcal{D}}_i = g_n^{-1}(i) \equiv \{y^n \in \mathcal{Y}^n \colon g_n(y^n) = i\}, \quad i \in \{1, \ldots, M\}. \tag{3.6.117}$$

Define

$$\delta_n = \delta_n(\varepsilon) = \frac{1}{n}\left[\sqrt{\frac{n \ln n}{2}} + \sqrt{\frac{n}{2} \ln \frac{1}{1-\varepsilon}}\right] \tag{3.6.118}$$

(note that $n\delta_n$ is an integer) then, by Lemma 3.6.2, the "blown-up" decoding sets $\mathcal{D}_i \triangleq \left[\widetilde{\mathcal{D}}_i\right]_{n\delta_n}$ satisfy

$$\begin{aligned} P_{Y^n|X^n=c_i}(\mathcal{D}_i^c) &\le \exp\left[-2n\left(\delta_n - \sqrt{\frac{1}{2n}\ln\frac{1}{1-\varepsilon}}\right)^2\right] \\ &\le \frac{1}{n}, \qquad \forall\, i \in \{1, \ldots, M\} \end{aligned} \tag{3.6.119}$$

where the last inequality holds since, from (3.6.118),

$$\delta_n \ge \sqrt{\frac{\ln n}{2n}} + \sqrt{\frac{1}{2n}\ln\frac{1}{1-\varepsilon}}. \tag{3.6.120}$$

We now complete the proof by a random coding argument. For

$$N \triangleq \left\lceil \frac{M}{n\binom{n}{n\delta_n}|\mathcal{Y}|^{n\delta_n}} \right\rceil, \tag{3.6.121}$$

let $U_1, \ldots, U_N$ be i.i.d. random variables which are equiprobable on $\{1, \ldots, M\}$. 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 equiprobable on the set $\{U_1, \ldots, U_N\}$, and let $P_{Y^n(V)}$ denote the output distribution

$$P_{Y^n(V)} = \frac{1}{N}\sum_{i=1}^{N} P_{Y^n|X^n=c_{U_i}}. \tag{3.6.122}$$

It can be easily verified that

$$\mathbb{E}\left[P_{Y^n(V)}\right] = P_{Y^n}^{(\mathcal{C})}, \tag{3.6.123}$$

which is the output distribution of the code $\mathcal{C}$, where the expectation in the left side of (3.6.123) 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 is included in the blown-up decoding sets $\mathcal{D}_{U_1}, \ldots, \mathcal{D}_{U_N}$:

$$\mathcal{N}_V(y^n) \triangleq \left\{ j \in \{1, \ldots, N\} : y^n \in \mathcal{D}_{U_j} \right\}. \tag{3.6.124}$$

Consider the list decoder $Y^n \mapsto \mathcal{N}_V(Y^n)$, and let

$$\varepsilon(V) \triangleq P(J \notin \mathcal{N}_V(Y^n)|V) \tag{3.6.125}$$

denote the conditional decoding error probability. Eq. (3.6.121) yields

$$\begin{aligned} \ln N &\geq \ln M - \ln n - \ln \binom{n}{n\delta_n} - n\delta_n \ln |\mathcal{Y}| \\ &\geq \ln M - \ln n - n\delta_n \left(\ln n + \ln |\mathcal{Y}| \right) \end{aligned} \tag{3.6.126}$$

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 the gain in using instead the inequality $\binom{n}{n\delta_n} \leq \exp\big(n\, h(\delta_n)\big)$ is asymptotically marginal for large n). Moreover, since each $y^n \in \mathcal{Y}^n$ belongs to at most $\binom{n}{n\delta_n} |\mathcal{Y}|^{n\delta_n}$ blown-up decoding sets then

$$\begin{aligned} \ln |\mathcal{N}_V(y^n)| &\leq \ln \binom{n}{n\delta_n} + n\delta_n \ln |\mathcal{Y}| \\ &\leq n\delta_n \left(\ln n + \ln |\mathcal{Y}| \right), \quad \forall\, y^n \in \mathcal{Y}^n. \end{aligned} \tag{3.6.127}$$

Now, for each realization of V, we have

$$\begin{aligned} &D\big(P_{Y^n(V)} \big\| P^*_{Y^n}\big) \\ &= D\big(P_{Y^n(V)|X^n(V)} \big\| P^*_{Y^n} \big| P_{X^n(V)}\big) - I(X^n(V); Y^n(V)) && (3.6.128) \\ &\leq nC - I(X^n(V); Y^n(V)) && (3.6.129) \\ &\leq nC - I(J; Y^n(V)) && (3.6.130) \\ &= nC - H(J) + H(J|Y^n(V)) && (3.6.131) \\ &\leq nC - \ln N + (1 - \varepsilon(V)) \max_{y^n \in \mathcal{Y}^n} \ln |\mathcal{N}_V(y^n)| \\ &\quad + n\, \varepsilon(V) \ln |\mathcal{X}| + \ln 2 && (3.6.132) \end{aligned}$$

with the following reasoning:

- (3.6.128) is by the chain rule for divergence;
- (3.6.129) is by (3.6.89);
- (3.6.130) is by the data processing inequality, and the fact that $J \to X^n(V) \to Y^n(V)$ is a Markov chain;
- (3.6.131) holds since $I(J; Y^n(V)) = H(J) - H(J|Y^n(V))$;
- (3.6.132) holds due to the generalization of Fano's inequality for list decoding (see Appendix 3.E), and since (i) $N \leq |\mathcal{X}|^n$, (ii) J is equiprobable on $\{U_1, \ldots, U_N\}$, so $H(J|U_1, \ldots, U_N) = \ln N$ and $H(J) \geq \ln N$.

Note that the quantities which are indexed by V in (3.6.128)–(3.6.132) are random variables since they depend on the realization $V = U^N$. Assembling (3.6.126), (3.6.127) and (3.6.132) yields

$$D(P_{Y^n(V)} \| P^*_{Y^n}) \leq nC - \ln M + \ln n + 2n\delta_n \left(\ln n + \ln |\mathcal{Y}|\right) + n\varepsilon(V) \ln |\mathcal{X}| + \ln 2. \tag{3.6.133}$$

In view of (3.6.122) and the equiprobable distribution of the messages, we get

$$\mathbb{E}\left[P_{Y^n(V)}\right] = P^{(\mathcal{C})}_{Y^n}, \tag{3.6.134}$$

and, from Jensen's inequality and the convexity of the relative entropy,

$$\mathbb{E}\left[D(P_{Y^n(V)} \| P^*_{Y^n})\right] \geq D(P^{(\mathcal{C})}_{Y^n} \| P^*_{Y^n}). \tag{3.6.135}$$

By taking expectations on both sides of (3.6.133), we get from (3.6.135)

$$D(P^{(\mathcal{C})}_{Y^n} \| P^*_{Y^n}) \leq nC - \ln M + \ln n + 2n\delta_n \left(\ln n + \ln |\mathcal{Y}|\right) + n\,\mathbb{E}\left[\varepsilon(V)\right] \ln |\mathcal{X}| + \ln 2, \tag{3.6.136}$$

and, in view of (3.6.119) and (3.6.125), we get

$$\mathbb{E}\left[\varepsilon(V)\right] \leq \max_{1 \leq i \leq M} P_{Y^n|X^n=c_i}\left(\mathcal{D}_i^c\right) \leq \frac{1}{n}. \tag{3.6.137}$$

Eq. (3.6.116) is finally derived by assembling (3.6.136), (3.6.137), and the following simple bound on δ_n (obtained from (3.6.118)):

$$\delta_n < \sqrt{\frac{\ln n}{2n}} + \sqrt{\frac{1}{2n} \ln \frac{1}{1-\varepsilon}} + \frac{1}{n}. \tag{3.6.138}$$

□

We are now ready to examine some consequences of Theorems 3.6.9 and 3.6.10. To start with, consider a sequence $\{\mathcal{C}_n\}_{n=1}^{\infty}$ where each code $\mathcal{C}_n = (f_n, g_n)$ is an (n, M_n, ε)-code for a DMC $T\colon \mathcal{X} \to \mathcal{Y}$ with $C > 0$ and $\varepsilon \in (0,1)$. We say that $\{\mathcal{C}_n\}_{n=1}^{\infty}$ is *capacity-achieving* if

$$\lim_{n\to\infty} \tfrac{1}{n} \ln M_n = C. \tag{3.6.139}$$

From Theorem 3.6.10, it follows that every such sequence satisfies

$$\lim_{n\to\infty} \tfrac{1}{n} D\big(P_{Y^n}^{(\mathcal{C}_n)} \| P_{Y^n}^*\big) = 0. \tag{3.6.140}$$

Moreover, as it is shown in [63], if the restriction to either deterministic encoding maps or to the maximal probability of error criterion is lifted, then the convergence in (3.6.140) may no longer hold. This is in sharp contrast to [183, Theorem 2], which states that (3.6.140) holds for *every* capacity-achieving sequence of codes with vanishing probability of error (maximal or average).

Another remarkable fact that follows from Theorems 3.6.9 and 3.6.10 is that a broad class of functions evaluated at the output of a good channel code concentrate sharply around their expectations with respect to the capacity-achieving output distribution. Specifically, we have the following version of [63, Proposition 11] (again, we have streamlined the statement and the proof a bit to relate them to earlier material in this chapter):

Theorem 3.6.11. Let $T\colon \mathcal{X} \to \mathcal{Y}$ be a DMC with $C > 0$ and $C_1 < \infty$ (see (3.6.97)). 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 $(\mathcal{Y}^n, d)$. For $\varepsilon \in \big(0, \frac{1}{2}\big)$, let

$$a \triangleq c(T)\sqrt{\frac{1}{2}\ln\frac{1}{1-2\varepsilon}} \tag{3.6.141}$$

with $c(T)$ defined in (3.6.98). Then, for every (n, M, ε)-code $\mathcal{C}$ for T and every Lipschitz function $f\colon \mathcal{Y}^n \to \mathbb{R}$ with respect to the metric d

$$\begin{aligned} &P_{Y^n}^{(\mathcal{C})}\Big(\,|f(Y^n) - \mathbb{E}[f(Y^{*n})]| \ge r\Big) \\ &\qquad \le \frac{4}{\varepsilon}\cdot \exp\left(nC - \ln M + a\sqrt{n} - \frac{r^2}{8c\|f\|_{\mathrm{Lip}}^2}\right), \quad \forall\, r > 0 \end{aligned} \tag{3.6.142}$$

where $\mathbb{E}[f(Y^{*n})]$ designates the expected value of $f(Y^n)$ with respect to the capacity-achieving output distribution $P^*_{Y^n}$, and $\|f\|_{\mathrm{Lip}}$ is the Lipschitz constant of f as defined in (3.4.2).

Remark 3.34. Our sharpening of the corresponding result from [63, Proposition 11] consists mainly in identifying an explicit form for the constant in front of $\sqrt{n}$ in the bound (3.6.142); this provides a closed-form expression for the concentration of measure inequality.

Proof. For an arbitrary Lipschitz function $f\colon \mathcal{Y}^n \to \mathbb{R}$, define

$$\mu_f^* \triangleq \mathbb{E}[f(Y^{*n})], \tag{3.6.143}$$

$$\phi(x^n) \triangleq \mathbb{E}[f(Y^n)|X^n = x^n] \tag{3.6.144}$$

for $x^n \in \mathcal{X}^n$. Since (by assumption) each $P_{Y^n|X^n=x^n}$ satisfies $\mathrm{T}_1(c)$, by Corollary 3.4.5,

$$\mathbb{P}\Big(\big|f(Y^n) - \phi(x^n)\big| \geq r \,\big|\, X^n = x^n\Big) \leq 2\exp\left(-\frac{r^2}{2c\,\|f\|^2_{\mathrm{Lip}}}\right) \tag{3.6.145}$$

for every $x^n \in \mathcal{X}^n$ and $r \geq 0$. Now, given $\mathcal{C}$, consider a subcode $\mathcal{C}'$ with the codewords $x^n \in \mathcal{X}^n$ satisfying $\phi(x^n) \geq \mu_f^* + r$ ($r \geq 0$). The number of codewords M' of $\mathcal{C}'$ satisfies

$$M' = MP_{X^n}^{(\mathcal{C})}\left(\phi(X^n) \geq \mu_f^* + r\right). \tag{3.6.146}$$

Let $Q = P_{Y^n}^{(\mathcal{C}')}$ be the output distribution induced by $\mathcal{C}'$. Then

$$\mu_f^* + r \leq \frac{1}{M'}\sum_{x^n \in \mathcal{C}'} \phi(x^n) \tag{3.6.147}$$

$$= \mathbb{E}_Q[f(Y^n)] \tag{3.6.148}$$

$$\leq \mathbb{E}[f(Y^{*n})] + \|f\|_{\mathrm{Lip}}\sqrt{2cD(Q_{Y^n}\|P^*_{Y^n})} \tag{3.6.149}$$

$$\leq \mu_f^* + \|f\|_{\mathrm{Lip}}\sqrt{2c\left(nC - \ln M' + a\sqrt{n} + \ln\frac{1}{\varepsilon}\right)}, \tag{3.6.150}$$

where

- (3.6.147) is by definition of $\mathcal{C}'$;
- (3.6.148) is by definition of ϕ in (3.6.144);

- (3.6.149) follows from the assumption that $P^*_{Y^n}$ satisfies $\mathrm{T}_1(c)$ and from the Kantorovich–Rubinstein formula (3.4.114); and
- (3.6.150) holds for the constant $a = a(T, \varepsilon) > 0$ in (3.6.141) due to Theorem 3.6.9 (see (3.6.99)) and since $\mathcal{C}'$ is an (n, M', ε)-code for T. The constant μ^*_f in (3.6.150) is defined in (3.6.143).

From (3.6.146)–(3.6.150), we get

$$r \le \|f\|_{\mathrm{Lip}} \sqrt{2c \left(nC - \ln M - \ln P^{(\mathcal{C})}_{X^n} \left(\phi(X^n) \ge \mu^*_f + r \right) + a\sqrt{n} + \ln \frac{1}{\varepsilon} \right)}$$

so, it follows that

$$P^{(\mathcal{C})}_{X^n} \Big(\phi(X^n) \ge \mu^*_f + r \Big) \le \exp \left(nC - \ln M + a\sqrt{n} + \ln \frac{1}{\varepsilon} - \frac{r^2}{2c \, \|f\|^2_{\mathrm{Lip}}} \right).$$

Following the same line of reasoning with $-f$ instead of f, we conclude that

$$\begin{aligned} & P^{(\mathcal{C})}_{X^n} \Big(|\phi(X^n) - \mu^*_f| \ge r \Big) \\ & \le 2 \exp \left(nC - \ln M + a\sqrt{n} + \ln \frac{1}{\varepsilon} - \frac{r^2}{2c \|f\|^2_{\mathrm{Lip}}} \right). \end{aligned} \tag{3.6.151}$$

Finally, for every $r \ge 0$,

$$\begin{aligned} & P^{(\mathcal{C})}_{Y^n} \Big(|f(Y^n) - \mu^*_f| \ge r \Big) \\ & \le P^{(\mathcal{C})}_{X^n, Y^n} \Big(|f(Y^n) - \phi(X^n)| \ge r/2 \Big) \\ & \quad + P^{(\mathcal{C})}_{X^n} \Big(\Big| \phi(X^n) - \mu^*_f \Big| \ge r/2 \Big) \qquad (3.6.152) \\ & \le 2 \exp \left(- \frac{r^2}{8c \|f\|^2_{\mathrm{Lip}}} \right) \\ & \quad + 2 \exp \left(nC - \ln M + a\sqrt{n} + \ln \frac{1}{\varepsilon} - \frac{r^2}{8c \|f\|^2_{\mathrm{Lip}}} \right) \qquad (3.6.153) \\ & \le 4 \exp \left(nC - \ln M + a\sqrt{n} + \ln \frac{1}{\varepsilon} - \frac{r^2}{8c \|f\|^2_{\mathrm{Lip}}} \right), \qquad (3.6.154) \end{aligned}$$

where (3.6.152) is due to the triangle inequality; (3.6.153) is by (3.6.145) and (3.6.151); (3.6.154) follows from the inequality

$$nC - \ln M + a\sqrt{n} + \ln\frac{1}{\varepsilon} \geq D(P_{Y^n}^{(\mathcal{C})} \| P_{Y^n}^*) \geq 0 \tag{3.6.155}$$

which holds by Theorem 3.6.9 with the constant a in (3.6.141). This proves the concentration inequality (3.6.142). □

As an illustration, let us consider $\mathcal{Y}^n$ with the Hamming metric

$$d_n(y^n, v^n) = \sum_{i=1}^n 1_{\{y_i \neq v_i\}}. \tag{3.6.156}$$

Then, every function $f\colon \mathcal{Y}^n \to \mathbb{R}$ of the form

$$f(y^n) = \frac{1}{n}\sum_{i=1}^n f_i(y_i), \qquad \forall\, y^n \in \mathcal{Y}^n \tag{3.6.157}$$

where $f_1, \ldots, f_n\colon \mathcal{Y} \to \mathbb{R}$ are Lipschitz functions on $\mathcal{Y}$, satisfies

$$\|f\|_{\mathrm{Lip}} \leq \frac{L}{n}, \qquad L \triangleq \max_{1\leq i\leq n} \|f_i\|_{\mathrm{Lip}}.$$

Every probability distribution P defined on $\mathcal{Y}$ and equipped with the Hamming metric satisfies $\mathrm{T}_1\big(\frac{1}{4}\big)$ (this is simply Pinsker's inequality); by Proposition 3.9, every product probability distribution on $\mathcal{Y}^n$ satisfies $\mathrm{T}_1\big(\frac{n}{4}\big)$ with respect to the metric (3.6.156). Consequently, for every (n, M, ε)-code for T and every Lipschitz function $f\colon \mathcal{Y}^n \to \mathbb{R}$ of the form (3.6.157), Theorem 3.6.11 yields the concentration inequality

$$\begin{aligned} &P_{Y^n}^{(\mathcal{C})}\Big(\big|f(Y^n) - \mathbb{E}[f(Y^{*n})]\big| \geq r\Big) \\ &\qquad \leq \frac{4}{\varepsilon}\,\exp\left(nC - \ln M + a\sqrt{n} - \frac{nr^2}{2L^2}\right) \end{aligned} \tag{3.6.158}$$

for every $r > 0$. Concentration inequalities like (3.6.142), or its more specialized version (3.6.158), 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 a Lipschitz function f of interest. Then, Theorem 3.6.11 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 3.6.9 and 3.6.10 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 [63] for all finite blocklengths n.

3.6.5 Information-theoretic converse for concentration of measure

If we were to summarize the main concept behind the concentration of measure phenomenon, it would be as follows: if a subset of a metric probability space does not have a too small probability mass, then its isoperimetric enlargements (or blowups) will eventually take up most of the probability mass. On the other hand, it makes sense to ask whether a converse of this statement is true, i.e.,

> 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 [190] using information-theoretic techniques.

The following setting is considered in [190]: 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). Let us extend the single-letter distortion d to $d_n\colon \mathcal{X}^n \times \mathcal{X}^n \to \mathbb{R}^+$ with $n \in \mathbb{N}$, and

$$d_n(x^n, y^n) \triangleq \sum_{i=1}^{n} d(x_i, y_i), \qquad \forall\, x^n, y^n \in \mathcal{X}^n. \tag{3.6.159}$$

For every set $\mathcal{C} \subseteq \mathcal{X}^n$, let

$$M^n(\mathcal{C}) \triangleq \sum_{x^n \in \mathcal{C}} M^n(x^n) \tag{3.6.160}$$

where

$$M^n(x^n) \triangleq \prod_{i=1}^{n} M(x_i), \qquad \forall\, x^n \in \mathcal{X}^n. \tag{3.6.161}$$

We also define the (closed) r-blowup of an arbitrary set $\mathcal{A} \subseteq \mathcal{X}^n$:

$$\mathcal{A}_r \triangleq \{x^n \in \mathcal{X}^n \colon d_n(x^n, \mathcal{A}) \leq r\}, \tag{3.6.162}$$

where

$$d_n(x^n, \mathcal{A}) = \min_{y^n \in \mathcal{A}} d_n(x^n, y^n). \tag{3.6.163}$$

Fix a probability distribution $P \in \mathcal{P}(\mathcal{X})$ where we assume without loss of generality that P is strictly positive on $\mathcal{X}$. We are interested in the following question:

> Given a sequence of sets $\{\mathcal{A}^{(n)}\}_{n\in\mathbb{N}}$ such that $\mathcal{A}^{(n)} \subseteq \mathcal{X}^n$ for every n, and
>
> $$P^{\otimes n}\left(\mathcal{A}^{(n)}_{n\delta}\right) \xrightarrow{n\to\infty} 1, \tag{3.6.164}$$
>
> for some $\delta \geq 0$, how small can their masses $M^n(\mathcal{A}^{(n)})$ be?

In order to state and prove the main result of [190] that answers this question, we need a few preliminary definitions. For every $n \in \mathbb{N}$, every pair (P_n, Q_n) of probability measures on $\mathcal{X}^n$, and every $\delta \geq 0$, we define the set

$$\begin{aligned}\Pi_n(P_n, Q_n, \delta) \triangleq \{\pi_n \colon\, & P_{X^nY^n} = \pi_n,\ P_{X^n} = P_n,\ P_{Y^n} = Q_n,\\ & \mathbb{E}_{\pi_n}\left[d_n(X^n, Y^n)\right] \leq n\delta\},\end{aligned} \tag{3.6.165}$$

which is the set 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

$$I_n(P_n, Q_n, \delta) \triangleq \inf_{\pi_n \in \Pi_n(P_n, Q_n, \delta)} D(\pi_n \| P_n \otimes Q_n), \tag{3.6.166}$$

and consider the sequence of functions $\{R_n(\cdot)\}_{n\in\mathbb{N}}$ defined as follows:

$$R_n(\delta) \equiv R_n(\delta; P_n, M^n) \tag{3.6.167}$$

$$\triangleq \inf_{Q_n\in\mathcal{P}(\mathcal{X}^n)} \Big\{ I_n(P_n, Q_n, \delta) + \mathbb{E}_{Q_n}[\ln M^n(Y^n)] \Big\} \tag{3.6.168}$$

$$= \inf_{P_{X^nY^n}} \Big\{ I(X^n;Y^n) + \mathbb{E}[\ln M^n(Y^n)]: \\ P_{X^n} = P_n,\ \tfrac{1}{n}\,\mathbb{E}[d_n(X^n,Y^n)] \le \delta \Big\} \tag{3.6.169}$$

for $\delta \ge 0$. In the special case where each P_n is the product measure $P^{\otimes n}$, the *rate function* $R(\cdot)$ is defined by (see [190, Eq. (10)])

$$R(\delta) \equiv R(\delta; P, M) \tag{3.6.170}$$

$$\triangleq \inf_{P_{XY}} \big\{ I(X;Y) + \mathbb{E}[\ln M(Y)]: \\ P_X = P,\ \mathbb{E}[d(X,Y)] \le \delta \big\}. \tag{3.6.171}$$

The rate function is monotonically non-increasing and convex in $\delta \ge 0$, and it satisfies (see [190, Lemmas 1 and 2])

$$R(\delta) = \lim_{n\to\infty} \tfrac{1}{n} R_n(\delta) \tag{3.6.172}$$

$$= \inf_{n\ge 1} \tfrac{1}{n}\, R_n(\delta). \tag{3.6.173}$$

We next state the main result of [190]:

Theorem 3.6.12 (Kontoyiannis)**.** Consider an arbitrary set $\mathcal{A}^{(n)} \subseteq \mathcal{X}^n$, and denote $\delta \triangleq \frac{1}{n}\,\mathbb{E}[d_n(X^n, \mathcal{A}^{(n)})]$. Then

$$\tfrac{1}{n} \ln M^n(\mathcal{A}^{(n)}) \ge R(\delta; P, M). \tag{3.6.174}$$

Furthermore, the following achievability result holds: for every $\delta \ge 0$ and $\varepsilon > 0$, there is a sequence of sets $\Big\{\mathcal{A}^{(n)}\Big\}_{n\in\mathbb{N}}$ such that $\mathcal{A}^{(n)} \subseteq \mathcal{X}^n$ for every n, and

$$\tfrac{1}{n} \ln M^n(\mathcal{A}^{(n)}) \le R(\delta) + \varepsilon, \qquad \tfrac{1}{n}\, d_n(X^n, \mathcal{A}^{(n)}) \le \delta \ \text{ a.s.} \tag{3.6.175}$$

Proof. We prove in the sequel the converse result in (3.6.174). Given $\mathcal{A}^{(n)} \subseteq \mathcal{X}^n$, let $\varphi_n \colon \mathcal{X}^n \to \mathcal{A}^{(n)}$ be the function that maps each $x^n \in \mathcal{X}^n$ to the closest element $y^n \in \mathcal{A}^{(n)}$, i.e.,

$$d_n(x^n, \varphi_n(x^n)) = d_n(x^n, \mathcal{A}^{(n)}), \quad \forall\, x^n \in \mathcal{X}^n \tag{3.6.176}$$

(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 $\pi_n \in \mathcal{P}(\mathcal{X}^n \times \mathcal{X}^n)$ be the following joint distribution of X^n and Y^n:

$$\pi_n(x^n, y^n) = P^{\otimes n}(x^n)\, 1_{\{y^n = \varphi_n(x^n)\}}, \quad \forall\, x^n, y^n \in \mathcal{X}^n. \tag{3.6.177}$$

Since $\Pi_n(P^{\otimes n}$ and Q_n are the marginal distributions of π_n, and

$$\begin{aligned}
\mathbb{E}_{\pi_n}[d_n(X^n, Y^n)] &= \mathbb{E}_{\pi_n}[d_n(X^n, \varphi_n(X^n))] && (3.6.178)\\
&= \mathbb{E}_{\pi_n}\left[d_n(X^n, \mathcal{A}^{(n)})\right] && (3.6.179)\\
&= n\delta && (3.6.180)
\end{aligned}$$

then $\pi_n \in \Pi_n(P^{\otimes n}, Q_n, \delta)$. Furthermore, we have

$$\begin{aligned}
\ln M^n(\mathcal{A}^{(n)}) &= \ln \sum_{y^n \in \mathcal{A}^{(n)}} M^n(y^n) && (3.6.181)\\
&= \ln \sum_{y^n \in \mathcal{A}^{(n)}} Q_n(y^n) \cdot \frac{M^n(y^n)}{Q_n(y^n)} && (3.6.182)\\
&\geq \sum_{y^n \in \mathcal{A}^{(n)}} Q_n(y^n) \ln \frac{M^n(y^n)}{Q_n(y^n)} && (3.6.183)\\
&= \sum_{x^n \in \mathcal{X}^n, y^n \in \mathcal{A}^{(n)}} \pi_n(x^n, y^n) \ln \frac{\pi_n(x^n, y^n)}{P^{\otimes n}(x^n) Q_n(y^n)}\\
&\qquad + \sum_{y^n \in \mathcal{A}^{(n)}} Q_n(y^n) \ln M^n(y^n) && (3.6.184)\\
&= I(X^n; Y^n) + \mathbb{E}_{Q_n}[\ln M^n(Y^n)] && (3.6.185)\\
&\geq R_n(\delta), && (3.6.186)
\end{aligned}$$

where (3.6.181) is by (3.6.160); (3.6.182) is trivial; (3.6.183) is by Jensen's inequality; (3.6.184) relies on (3.6.177); (3.6.185) uses the fact that the marginal distributions of π_n are $P^{\otimes n}$ and Q_n, and (3.6.186) relies on (3.6.167)–(3.6.169) and (3.6.180). Finally, using (3.6.167), (3.6.172) and (3.6.181)–(3.6.186), we get (3.6.174).

The reader is referred to [190, Theorem 2] for a proof of the direct part (achievability result) in (3.6.175). □

We are now ready to use Theorem 3.6.12 to answer the question posed at the beginning of this section. Specifically, we consider the case

when $M = P$. Defining the *concentration exponent*

$$R_{\mathrm{c}}(r; P) \triangleq R(r; P, P), \tag{3.6.187}$$

we get the following result:

Corollary 3.6.13 (Converse concentration of measure)**.** If $\mathcal{A}^{(n)} \subseteq \mathcal{X}^n$ is an arbitrary set, then

$$P^{\otimes n}\left(\mathcal{A}^{(n)}\right) \geq \exp\left(n\, R_{\mathrm{c}}(\delta; P)\right), \tag{3.6.188}$$

where

$$\delta = \frac{1}{n}\,\mathbb{E}\left[d_n\left(X^n, \mathcal{A}^{(n)}\right)\right]. \tag{3.6.189}$$

Moreover, if the sequence of sets $\{\mathcal{A}^{(n)}\}_{n=1}^{\infty}$ is such that, for some $\delta \geq 0$, $P^{\otimes n}\left(\mathcal{A}_{n\delta}^{(n)}\right) \to 1$ as $n \to \infty$, then

$$\liminf_{n\to\infty} \frac{1}{n}\,\ln P^{\otimes n}\left(\mathcal{A}^{(n)}\right) \geq R_{\mathrm{c}}(\delta; P). \tag{3.6.190}$$

Remark 3.35. A moment of reflection shows that the concentration exponent $R_{\mathrm{c}}(\delta; P)$ is nonpositive. Indeed, from definitions,

$$\begin{aligned}
&R_{\mathrm{c}}(\delta; P)\\
&= R(\delta; P, P)\\
&= \inf_{P_{XY}} \Big\{ I(X;Y) + \mathbb{E}[\ln P(Y)] \colon P_X = P,\, \mathbb{E}[d(X,Y)] \leq \delta \Big\}\\
&= \inf_{P_{XY}} \Big\{ H(Y) - H(Y|X) + \mathbb{E}[\ln P(Y)] \colon P_X = P,\, \mathbb{E}[d(X,Y)] \leq \delta \Big\}\\
&= \inf_{P_{XY}} \Big\{ -D(P_Y\|P) - H(Y|X) \colon P_X = P,\, \mathbb{E}[d(X,Y)] \leq \delta \Big\}\\
&= -\sup_{P_{XY}} \Big\{ D(P_Y\|P) + H(Y|X) \colon P_X = P,\, \mathbb{E}[d(X,Y)] \leq \delta \Big\},
\end{aligned} \tag{3.6.191}$$

which proves the claim, since both the divergence and the (conditional) entropy are nonnegative.

Remark 3.36. Using the achievability result (3.6.175), which appears in [190, Theorem 2], one can also prove that there exists a sequence of sets $\{\mathcal{A}^{(n)}\}_{n=1}^{\infty}$, such that

$$\lim_{n\to\infty} P^{\otimes n}\left(\mathcal{A}_{n\delta}^{(n)}\right) = 1, \quad \lim_{n\to\infty} \frac{1}{n}\,\ln P^{\otimes n}\left(\mathcal{A}^{(n)}\right) \leq R_{\mathrm{c}}(\delta; P). \tag{3.6.192}$$

As an illustration, consider the case where $\mathcal{X} = \{0,1\}$ and d is the Hamming distortion $d(x,y) = 1_{\{x \neq y\}}$. Let $P(0) = 1-p$ and $P(1) = p$ with $p \in [0, \frac{1}{2}]$, so P satisfies $\mathrm{T}_1\left(\frac{1}{2\varphi(p)}\right)$ with respect to the Hamming metric d, and with $\varphi(p)$ defined in (3.4.74). By Proposition 3.8, the product measure $P^{\otimes n}$ satisfies $\mathrm{T}_1\left(\frac{n}{2\varphi(p)}\right)$ on the product space $(\mathcal{X}^n, d_n)$. Consequently, it follows from (3.4.89) that for every $\mathcal{A}^{(n)} \subseteq \mathcal{X}^n$,

$$P^{\otimes n}\left(\mathcal{A}^{(n)}_{n\delta}\right) \geq 1 - \exp\left(-n\,\varphi(p)\,\left(\delta - \sqrt{\frac{1}{n\,\varphi(p)}\,\ln\frac{1}{P^{\otimes n}\left(\mathcal{A}^{(n)}\right)}}\right)^2\right) \tag{3.6.193}$$

provided that

$$\delta \geq \sqrt{\frac{1}{n\,\varphi(p)}\,\ln\frac{1}{P^{\otimes n}\left(\mathcal{A}^{(n)}\right)}}. \tag{3.6.194}$$

Thus, if a sequence of sets $\mathcal{A}^{(n)} \subseteq \mathcal{X}^n$, $n \in \mathbb{N}$, satisfies

$$\liminf_{n\to\infty} \frac{1}{n}\,\ln P^{\otimes n}\left(\mathcal{A}^{(n)}\right) > -\varphi(p)\delta^2, \tag{3.6.195}$$

then

$$P^{\otimes n}\left(\mathcal{A}^{(n)}_{n\delta}\right) \xrightarrow{n\to\infty} 1. \tag{3.6.196}$$

The converse result, Corollary 3.6.13, says that if a sequence of sets $\mathcal{A}^{(n)} \subseteq \mathcal{X}^n$ satisfies (3.6.196), then (3.6.190) 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 side of (3.6.195):

Theorem 3.6.14. If P is the Bernoulli(p) measure with $p \in [0, \frac{1}{2}]$, then the concentration exponent $R_{\mathrm{c}}(\delta; P)$ satisfies

$$R_{\mathrm{c}}(\delta; P) \leq -\varphi(p)\delta^2 - (1-p)\,h\left(\frac{\delta}{1-p}\right), \qquad \forall\,\delta \in [0, 1-p] \tag{3.6.197}$$

and

$$R_{\mathrm{c}}(\delta; P) = \ln p, \qquad \forall\,\delta \in [1-p, 1] \tag{3.6.198}$$

where

$$h(x) \triangleq -x \ln x - (1-x)\ln(1-x), \quad \forall\, x \in [0,1]$$

is the binary entropy function to base e (with the convention that $0 \log 0 = 0$).

Proof. From (3.6.191) with the Hamming metric $d(x,y) = 1_{\{x \neq y\}}$ for $x, y \in \{0,1\}$, we have

$$R_{\mathrm{c}}(\delta; P) = -\sup_{P_{XY}} \Big\{ D(P_Y \| P) + H(Y|X) \colon P_X = P,\, \mathbb{P}(X \neq Y) \leq \delta \Big\}. \tag{3.6.199}$$

For a given $\delta \in [0, 1-p]$, let us choose P_Y so that $\|P_Y - P\|_{\mathrm{TV}} = \delta$. Then, from (3.4.77),

$$\frac{D(P_Y\|P)}{\delta^2} = \frac{D(P_Y\|P)}{\|P_Y - P\|_{\mathrm{TV}}^2} \tag{3.6.200}$$

$$\geq \inf_Q \frac{D(Q\|P)}{\|Q-P\|_{\mathrm{TV}}^2} \tag{3.6.201}$$

$$= \varphi(p). \tag{3.6.202}$$

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 (3.4.65), it can be verified that

$$P_{\tilde{Y}|\tilde{X}=0} = \text{Bernoulli}\left(\frac{\delta}{1-p}\right), \tag{3.6.203}$$

$$P_{\tilde{Y}|\tilde{X}=1}(\tilde{y}) = \delta_1(\tilde{y}) \triangleq 1_{\{\tilde{y}=1\}}, \tag{3.6.204}$$

which yields

$$H(\widetilde{Y}|\widetilde{X}) = (1-p)H(\widetilde{Y}|\widetilde{X}=0) = (1-p)h\left(\frac{\delta}{1-p}\right). \tag{3.6.205}$$

From (3.6.199), (3.6.202) and (3.6.205), we obtain

$$R_{\mathrm{c}}(\delta; P) \leq -D(P_{\widetilde{Y}}\|P) - H(\widetilde{Y}|\widetilde{X}) \tag{3.6.206}$$

$$\leq -\varphi(p)\delta^2 - (1-p)h\left(\frac{\delta}{1-p}\right). \tag{3.6.207}$$

To prove (3.6.198), 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 $p \leq \frac{1}{2}$. □

3.7 Summary

This chapter covers the essentials of the entropy method, an information-theoretic technique to derive concentration inequalities for functions of independent random variables. As its 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 relative entropy in order to break the original multi-dimensional problem up into more simple 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 examined the several deep and remarkable information-theoretic ideas that bridge these two classes of inequalities, and also exemplified 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. Moreover, applications of concentration of measure inequalities to problems in information theory, coding and communications are exemplified in Chapters 2 and 3. We hope that the 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 (3.2.34) and to show that equality in (3.2.34) holds if and only if Y is Gaussian. To this end, we prove the following statement. Let $\varphi(U)$ be an arbitrary estimator of Y where $\varphi(\cdot)$ is a Borel-measurable real-valued function. 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 simple. Define two random variables

$$\Delta(U,Y) \triangleq \varphi(U) - Y, \tag{3.A.2}$$

$$\begin{aligned}\Upsilon(U,Y) &\triangleq \frac{\mathrm{d}}{\mathrm{d}y} \ln\left[p_{U|Y}(U|y)p_Y(y)\right]\Big|_{y=Y} \\ &= \frac{\mathrm{d}}{\mathrm{d}y} \ln\left[\gamma(U - \sqrt{s}y)p_Y(y)\right]\Big|_{y=Y} \\ &= \sqrt{s}(U - \sqrt{s}Y) + \rho_Y(Y) \\ &= \sqrt{s}Z + \rho_Y(Y)\end{aligned} \tag{3.A.3}$$

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 show below that

$$\mathbb{E}[\Delta(U,Y)\Upsilon(U,Y)] = 1. \tag{3.A.4}$$

Then, in view of (3.A.4), by applying the Cauchy–Schwarz inequality,

$$\begin{aligned}1 &= \mathbb{E}^2[\Delta(U,Y)\Upsilon(U,Y)] \\ &\leq \mathbb{E}[\Delta^2(U,Y)] \cdot \mathbb{E}[\Upsilon^2(U,Y)] \\ &= \mathbb{E}[(\varphi(U) - Y)^2] \cdot \mathbb{E}[(\sqrt{s}Z + \rho_Y(Y))^2] \\ &= \mathbb{E}[(\varphi(U) - Y)^2] \cdot (s + J(Y)).\end{aligned} \tag{3.A.5}$$

Upon rearranging, we obtain (3.A.1). We next prove (3.A.4). The fact that $J(Y) < \infty$ implies that the density p_Y is bounded (see [135, Lemma A.1]). Using this and the rapid decay of the Gaussian density γ at infinity, we have

$$\begin{aligned}
&\int_{-\infty}^{\infty} \frac{\mathrm{d}}{\mathrm{d}y} \left[p_{U|Y}(u|y)p_Y(y)\right] \mathrm{d}y \\
&\quad = \gamma(u - \sqrt{s}y)p_Y(y)\Big|_{-\infty}^{\infty} \\
&\quad = 0, \qquad (3.A.6)
\end{aligned}$$

and integration by parts gives

$$\begin{aligned}
&\int_{-\infty}^{\infty} y\frac{\mathrm{d}}{\mathrm{d}y} \left[p_{U|Y}(u|y)p_Y(y)\right] \mathrm{d}y \\
&\quad = y\gamma(u - \sqrt{s}y)p_Y(y)\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} p_{U|Y}(u|y)p_Y(y)\mathrm{d}y \\
&\quad = -\int_{-\infty}^{\infty} p_{U|Y}(u|y)p_Y(y)\mathrm{d}y \\
&\quad = -p_U(u). \qquad (3.A.7)
\end{aligned}$$

Using (3.A.6) and (3.A.7), we get

$$\begin{aligned}
&\mathbb{E}[\Delta(U,Y)\Upsilon(U,Y)] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (\varphi(u) - y) \frac{\mathrm{d}}{\mathrm{d}y} \ln \left[p_{U|Y}(u|y)p_Y(y)\right] p_{U|Y}(u|y)p_Y(y)\mathrm{d}u\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (\varphi(u) - y) \frac{\mathrm{d}}{\mathrm{d}y} \left[p_{U|Y}(u|y)p_Y(y)\right] \mathrm{d}u\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty} \varphi(u) \underbrace{\left(\int_{-\infty}^{\infty} \frac{\mathrm{d}}{\mathrm{d}y} \left[p_{U|Y}(u|y)p_Y(y)\right] \mathrm{d}y\right)}_{=0} \mathrm{d}u \\
&\qquad\qquad - \int_{-\infty}^{\infty} \underbrace{\left(\int_{-\infty}^{\infty} y\frac{\mathrm{d}}{\mathrm{d}y} \left[p_{U|Y}(u|y)p_Y(y)\right] \mathrm{d}y\right)}_{=-p_U(u)} \mathrm{d}u \\
&= \int_{-\infty}^{\infty} p_U(u)\mathrm{d}u = 1,
\end{aligned}$$

which establishes the equality in (3.A.4). 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 $\Delta(U,Y) = c\Upsilon(U,Y)$ holds almost surely for some constant $c \in \mathbb{R}$. In view of (3.A.2) and (3.A.3), the latter equality 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, as 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

$$\rho_Y(y) \triangleq \frac{p_Y'(y)}{p_Y(y)} = \frac{c'}{c} + \left(s - \tfrac{1}{c}\right) y.$$

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 The proof of Theorem 3.2.3

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 $p \geq 1$. It will be convenient to work with the following equivalent form of the Rényi divergence in (3.2.48): For every two random variables U and V such that $P_U \ll P_V$, we have

$$D_\alpha(P_U\|P_V) = \frac{\alpha}{\alpha-1}\ln\left\|\frac{\mathrm{d}P_U}{\mathrm{d}P_V}(V)\right\|_\alpha, \qquad \alpha > 1. \tag{3.B.1}$$

Let g denote 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, $g_0 = g$ (recall that $P_0 = P$ since, by the definition of the random transformation in (3.2.43), OU(0) is a perfect channel with $Y = X$ at $t = 0$). Also, let the function $\alpha\colon [0,\infty) \to [\beta,\infty)$ be defined as $\alpha(t) = 1 + (\beta-1)e^{2t}$ for some $\beta > 1$. Let $Z \sim G$. Using (3.B.1), it can be shown that the desired bound (3.2.53) is equivalent

to the statement that the function $F\colon [0,\infty) \to \mathbb{R}$, given by

$$F(t) \triangleq \ln \left\| \frac{\mathrm{d}P_t}{\mathrm{d}G}(Z) \right\|_{\alpha(t)} \equiv \ln \left\| g_t(Z) \right\|_{\alpha(t)}, \tag{3.B.2}$$

is monotonically decreasing. From now on, we adhere to the following notational convention: we use either 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

$$\begin{aligned} \dot{F}(t) &= \frac{\mathrm{d}}{\mathrm{d}t} \left\{ \frac{1}{\alpha(t)} \ln \mathbb{E}\left[\left(g_t(Z) \right)^{\alpha(t)} \right] \right\} \\ &= -\frac{\dot{\alpha}(t)}{\alpha^2(t)} \ln \mathbb{E}\left[\left(g_t(Z) \right)^{\alpha(t)} \right] + \frac{1}{\alpha(t)} \frac{\frac{\mathrm{d}}{\mathrm{d}t} \mathbb{E}\left[\left(g_t(Z) \right)^{\alpha(t)} \right]}{\mathbb{E}\left[\left(g_t(Z) \right)^{\alpha(t)} \right]}. \end{aligned} \tag{3.B.3}$$

To handle the derivative with respect to t in the second term in the right side of (3.B.3), 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 (3.2.43).

For every $t \geq 0$, define a linear operator K_t acting on an arbitrary sufficiently regular (e.g., $L^1(G)$) function h as

$$K_t h(x) \triangleq \mathbb{E}\left[h\left(e^{-t} x + \sqrt{1 - e^{-2t}} Z \right) \right] \tag{3.B.4}$$

with $Z \sim G$. The family of linear operators $\{K_t\}_{t=0}^{\infty}$ has the following properties:

1. K_0 is the identity operator, $K_0 h = h$ for every h.
2. Consider the OU(t) channel, for every $t \geq 0$, given by the random transformation (3.2.43). For every measurable function F such that $\mathbb{E}\left|F(Y)\right| < \infty$ with Y in (3.2.43), we can write

$$K_t F(x) = \mathbb{E}[F(Y)|X = x], \qquad \forall\, x \in \mathbb{R} \tag{3.B.5}$$

and

$$\mathbb{E}[F(Y)] = \mathbb{E}[K_t F(X)]. \tag{3.B.6}$$

Note that (3.B.5) holds since by assumption X and Z in (3.2.43) are independent random variables, and (3.B.6) holds by taking expectation on both sides of (3.B.5).

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 $\mathrm{OU}(t)$ channel. Then, as we have noted before, $P_t \ll G$; the corresponding density $g_t = \mathrm{d}P_t/\mathrm{d}G$ satisfies for all $t \geq 0$

$$g_t(x) = K_t g(x). \tag{3.B.7}$$

To prove (3.B.7), we can either use (3.B.5) and the fact that $g_t(x) = \mathbb{E}[g(Y)|X = x]$, or proceed directly from (3.2.43) to get

$$\begin{aligned} g_t(x) &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} g\left(e^{-t}x + \sqrt{1-e^{-2t}}z\right) \exp\left(-\tfrac{z^2}{2}\right) \mathrm{d}z \\ &= \mathbb{E}\left[g\left(e^{-t}x + \sqrt{1-e^{-2t}}Z\right)\right] \end{aligned} \tag{3.B.8}$$

with $Z \sim G$.

4. The family of operators $\{K_t\}_{t=0}^{\infty}$ forms a semigroup, i.e., for every $t_1, t_2 \geq 0$ we have

$$K_{t_1+t_2} = K_{t_1} \circ K_{t_2} = K_{t_2} \circ K_{t_1}, \tag{3.B.9}$$

which is shorthand for saying that, for every sufficiently regular function h,

$$K_{t_1+t_2}h = K_{t_2}(K_{t_1}h) = K_{t_1}(K_{t_2}h). \tag{3.B.10}$$

Eq. (3.B.9) follows from (3.B.5) and (3.B.6), and due to the fact that the channel family $\{\mathrm{OU}(t)\}_{t=0}^{\infty}$ is ordered by degradation as in (3.2.45). For this reason, the family of linear operators $\{K_t\}_{t=0}^{\infty}$ is referred to as the *Ornstein–Uhlenbeck semigroup*. Note that if $\{Y_t\}_{t=0}^{\infty}$ is the Ornstein–Uhlenbeck process, then every function $F \in L^1(G)$ satisfies

$$K_t F(x) = \mathbb{E}[F(Y_t)|Y_0 = x], \qquad \forall\, x \in \mathbb{R}.$$

Two deeper results concerning the Ornstein–Uhlenbeck semigroup, which will be needed, are as follows: let the second-order differential operator $\mathcal{L}$ be defined as

$$\mathcal{L}h(x) \triangleq h''(x) - xh'(x) \tag{3.B.11}$$

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)

$$\dot{h}_t = \mathcal{L}h_t. \tag{3.B.12}$$

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*

$$\mathbb{E}[g(Z)\mathcal{L}h(Z)] = \mathbb{E}[h(Z)\mathcal{L}g(Z)] = -\mathbb{E}[g'(Z)h'(Z)]. \tag{3.B.13}$$

We provide the proofs of (3.B.12) and (3.B.13) in Appendix 3.C.

We are now ready to tackle the second term in (3.B.3). Noting that the family of densities $\{g_t\}_{t=0}^{\infty}$ forms an Ornstein–Uhlenbeck flow with initial condition $g_0 = g$, we have

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\left(g_t(Z)\right)^{\alpha(t)}\right] \\
&= \mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}t}\left\{\left(g_t(Z)\right)^{\alpha(t)}\right\}\right] \\
&= \dot{\alpha}(t)\,\mathbb{E}\left[\left(g_t(Z)\right)^{\alpha(t)}\ln g_t(Z)\right] + \alpha(t)\,\mathbb{E}\left[\left(g_t(Z)\right)^{\alpha(t)-1}\frac{\mathrm{d}}{\mathrm{d}t}g_t(Z)\right] \\
&= \dot{\alpha}(t)\,\mathbb{E}\left[\left(g_t(Z)\right)^{\alpha(t)}\ln g_t(Z)\right] \\
&\qquad + \alpha(t)\,\mathbb{E}\left[\left(g_t(Z)\right)^{\alpha(t)-1}\mathcal{L}g_t(Z)\right] &\text{(3.B.14)} \\
&= \dot{\alpha}(t)\,\mathbb{E}\left[\left(g_t(Z)\right)^{\alpha(t)}\ln g_t(Z)\right] \\
&\qquad - \alpha(t)\,\mathbb{E}\left[\left(\left(g_t(Z)\right)^{\alpha(t)-1}\right)' g_t'(Z)\right] &\text{(3.B.15)} \\
&= \dot{\alpha}(t)\,\mathbb{E}\left[\left(g_t(Z)\right)^{\alpha(t)}\ln g_t(Z)\right] \\
&\qquad - \alpha(t)(\alpha(t)-1)\,\mathbb{E}\left[\left(g_t(Z)\right)^{\alpha(t)-2}\left(g_t'(Z)\right)^2\right] &\text{(3.B.16)}
\end{aligned}$$

where we use (3.B.12) to get (3.B.14), and (3.B.13) to get (3.B.15). (Referring back to (3.B.8), 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 $\phi_t(z) \triangleq \big(g_t(z)\big)^{\alpha(t)/2}$, then we can rewrite (3.B.16) as

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\big(g_t(Z)\big)^{\alpha(t)}\right] = \frac{\dot{\alpha}(t)}{\alpha(t)}\,\mathbb{E}\left[\phi_t^2(Z)\ln\phi_t^2(Z)\right] - \frac{4\big(\alpha(t)-1\big)}{\alpha(t)}\,\mathbb{E}\left[\big(\phi_t'(Z)\big)^2\right]. \tag{3.B.17}$$

Using the definition of ϕ_t and substituting (3.B.17) into the right side of (3.B.3), we get

$$\alpha^2(t)\,\mathbb{E}[\phi_t^2(Z)]\,\dot{F}(t) = \dot{\alpha}(t)\,\Big(\mathbb{E}[\phi_t^2(Z)\ln\phi_t^2(Z)] - \mathbb{E}[\phi_t^2(Z)]\ln\mathbb{E}[\phi_t^2(Z)]\Big) - 4(\alpha(t)-1)\mathbb{E}\left[\big(\phi_t'(Z)\big)^2\right]. \tag{3.B.18}$$

If we now apply the Gaussian LSI (3.2.1) to ϕ_t, then (3.B.18) yields

$$\alpha^2(t)\,\mathbb{E}[\phi_t^2(Z)]\,\dot{F}(t) \le 2\big(\dot{\alpha}(t) - 2(\alpha(t)-1)\big)\mathbb{E}\left[\big(\phi_t'(Z)\big)^2\right]. \tag{3.B.19}$$

Since $\alpha(t) = 1 + (\beta-1)e^{2t}$ then $\dot{\alpha}(t) - 2\big(\alpha(t)-1\big) = 0$ for all t, which implies that the right side of (3.B.19) 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.B.8) where g is a probability density function), we conclude that $\dot{F}(t) \le 0$. Hence, F is monotonically decreasing on $[0,\infty)$, and from (3.B.2)

$$F(t) = \frac{\alpha(t)-1}{\alpha(t)}\,D_{\alpha(t)}(P_t\|G), \quad t \ge 0, \tag{3.B.20}$$

$$F(0) = \frac{\beta-1}{\beta}\,D_\beta(P\|G). \tag{3.B.21}$$

Consequently, for every $\beta > 1$ and $t \ge 0$,

$$D_{\alpha(t)}(P_t\|G) \le \left(\frac{\alpha(t)(\beta-1)}{\beta(\alpha(t)-1)}\right) D_\beta(P\|G) \tag{3.B.22}$$

where $\alpha(t) = 1 + (\beta - 1)e^{2t}$ for all $t \geq 0$. Since the Rényi divergence is monotonically increasing in its order (see, e.g., [143, Theorem 3]), the left side of (3.B.22) is greater than or equal to $D_\alpha(P_t\|G)$ as soon as $\alpha \leq \alpha(t)$. By the same token, because the function $u \in (1,\infty) \mapsto \frac{u}{u-1}$ is strictly decreasing, the right side of (3.B.22) can be upper-bounded by $\left(\frac{\alpha(\beta-1)}{\beta(\alpha-1)}\right) D_\beta(P\|G)$ for all $\alpha \leq \alpha(t)$. Putting all these facts together, we conclude that the Gaussian LSI (3.2.1) implies (3.2.53).

We now show that (3.2.53) yields the LSI of Theorem 3.2.1. To that end, we recall that (3.2.53) is equivalent to the right side of (3.B.18) being less than or equal to zero for all $t \geq 0$ and all $\beta > 1$. Let us choose $t = 0$ and $\beta = 2$, for which

$$\alpha(0) = \dot{\alpha}(0) = 2, \qquad \phi_0 = g.$$

Using this in (3.B.18) for $t = 0$, we get

$$2\left(\mathbb{E}\left[g^2(Z)\ln g^2(Z)\right] - \mathbb{E}[g^2(Z)]\ln \mathbb{E}[g^2(Z)]\right) - 4\,\mathbb{E}\left[\left(g'(Z)\right)^2\right] \leq 0$$

which is precisely the LSI (3.2.1) with $\mathbb{E}[g(Z)] = \mathbb{E}_G\left[\frac{\mathrm{d}P}{\mathrm{d}G}\right] = 1$; recall, however, that the LSI (3.2.1) is invariant to a scaling of the real-valued function. This completes the proof of Theorem 3.2.3 (up to the proofs of (3.B.12) and (3.B.13), which are relegated to Appendix 3.C).

3.C Details on the Ornstein–Uhlenbeck semigroup

This appendix proves the formulas (3.B.12) and (3.B.13), pertaining to the Ornstein–Uhlenbeck semigroup. We start with (3.B.12). Recalling that

$$h_t(x) = K_t h(x) = \mathbb{E}\left[h\left(e^{-t}x + \sqrt{1-e^{-2t}}Z\right)\right], \tag{3.C.1}$$

we have

$$\begin{aligned}\dot{h}_t(x) &= \frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{E}\left[h\left(e^{-t}x + \sqrt{1-e^{-2t}}Z\right)\right] \\ &= -e^{-t}x\,\mathbb{E}\left[h'\left(e^{-t}x + \sqrt{1-e^{-2t}}Z\right)\right] \\ &\quad + \frac{e^{-2t}}{\sqrt{1-e^{-2t}}}\cdot\mathbb{E}\left[Zh'\left(e^{-t}x + \sqrt{1-e^{-2t}}Z\right)\right].\end{aligned} \tag{3.C.2}$$

For an arbitrary sufficiently smooth function h and every $m, \sigma \in \mathbb{R}$,

$$\mathbb{E}[Zh'(m+\sigma Z)] = \sigma \mathbb{E}[h''(m+\sigma Z)], \tag{3.C.3}$$

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 (3.C.3) yields

$$\begin{aligned} &\mathbb{E}\left[Zh'\left(e^{-t}x + \sqrt{1-e^{-2t}}Z\right)\right] \\ &= \sqrt{1-e^{-2t}}\,\mathbb{E}\left[h''\left(e^{-t}x + \sqrt{1-e^{-2t}}Z\right)\right]. \end{aligned} \tag{3.C.4}$$

Consequently, combining (3.C.2) and (3.C.4) yields

$$\dot{h}_t(x) = -e^{-t}x\; K_t h'(x) + e^{-2t} K_t h''(x). \tag{3.C.5}$$

On the other hand, from (3.B.11) and (3.C.1),

$$\begin{aligned} \mathcal{L}h_t(x) &= h_t''(x) - x h_t'(x) \\ &= e^{-2t}\mathbb{E}\left[h''\left(e^{-t}x + \sqrt{1-e^{-2t}}Z\right)\right] \\ &\quad - xe^{-t}\mathbb{E}\left[h'\left(e^{-t}x + \sqrt{1-e^{-2t}}Z\right)\right] \\ &= e^{-2t} K_t h''(x) - e^{-t}x\; K_t h'(x). \end{aligned} \tag{3.C.6}$$

Comparing (3.C.5) and (3.C.6), we get (3.B.12).

Proving the integration-by-parts formula (3.B.13) is more subtle, and it 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 every $t, t' \geq 0$,

$$(Y_t, Y_{t'}) \stackrel{\mathrm{d}}{=} (Y_{t'}, Y_t). \tag{3.C.7}$$

To see this, let

$$p^{(t)}(y|x) \triangleq \frac{1}{\sqrt{2\pi(1-e^{-2t})}} \exp\left(-\frac{(y-e^{-t}x)^2}{2(1-e^{-2t})}\right) \tag{3.C.8}$$

be the transition density of the OU(t) channel. It is easy to show that

$$p^{(t)}(y|x)\gamma(x) = p^{(t)}(x|y)\gamma(y), \qquad \forall\, x, y \in \mathbb{R} \tag{3.C.9}$$

(recall that γ denotes the standard Gaussian pdf). For $Z \sim G$ and every two smooth functions g, h, this implies that

$$\begin{aligned}
\mathbb{E}[g(Z)K_t h(Z)] &= \mathbb{E}[g(Y_0)K_t h(Y_0)] && \text{(3.C.10)}\\
&= \mathbb{E}[g(Y_0)\mathbb{E}[h(Y_t)|Y_0]] && \text{(3.C.11)}\\
&= \mathbb{E}[g(Y_0)h(Y_t)] && \text{(3.C.12)}\\
&= \mathbb{E}[g(Y_t)h(Y_0)] && \text{(3.C.13)}\\
&= \mathbb{E}[K_t g(Y_0)h(Y_0)] && \text{(3.C.14)}\\
&= \mathbb{E}[K_t g(Z)h(Z)], && \text{(3.C.15)}
\end{aligned}$$

where (3.C.10) and (3.C.15) are due to the assumption that $Y_0 \sim G$ and $Z \sim G$; (3.C.11) and (3.C.14) rely on (3.B.5); (3.C.12) relies on the tower principle for the conditional expectation; (3.C.13) holds due to the reversibility property of the Ornstein–Uhlenbeck process with $Y_0 \sim G$ (see (3.C.7)). Taking the derivative with respect to t of the left side in (3.C.10) and the right side in (3.C.15), we conclude from (3.B.12) that

$$\mathbb{E}[g(Z)\mathcal{L}h(Z)] = \mathbb{E}[\mathcal{L}g(Z)h(Z)]. \tag{3.C.16}$$

In particular, since $\mathcal{L}1 = 0$ (where on the left side 1 denotes the constant function $x \mapsto 1$), we have

$$\mathbb{E}[\mathcal{L}g(Z)] = \mathbb{E}[1\mathcal{L}g(Z)] = \mathbb{E}[g(Z)\mathcal{L}1] = 0 \tag{3.C.17}$$

for all smooth g.

We are now ready to prove (3.B.13). To that end, let us first define the operator Γ on pairs of functions g, h by

$$\Gamma(g,h) \triangleq \tfrac{1}{2}\left[\mathcal{L}(gh) - g\mathcal{L}h - h\mathcal{L}g\right]. \tag{3.C.18}$$

Now, for the specific definition of $\mathcal{L}$ in (3.B.11), we have

$$\begin{aligned}\Gamma(g,h)(x) &= \tfrac{1}{2}\Big[(gh)''(x) - x(gh)'(x) - g(x)\big(h''(x) - xh'(x)\big) \\ &\qquad - h(x)\big(g''(x) - xg'(x)\big)\Big] \\ &= \tfrac{1}{2}\Big[g''(x)h(x) + 2g'(x)h'(x) + g(x)h''(x) \\ &\qquad - xg'(x)h(x) - xg(x)h'(x) - g(x)h''(x) \\ &\qquad + xg(x)h'(x) - g''(x)h(x) + xg'(x)h(x)\Big] \\ &= g'(x)h'(x), \end{aligned} \tag{3.C.19}$$

or, more succinctly, $\Gamma(g,h) = g'h'$. Therefore,

$$\mathbb{E}[g(Z)\mathcal{L}h(Z)] = \tfrac{1}{2}\Big\{\mathbb{E}[g(Z)\mathcal{L}h(Z)] + \mathbb{E}[h(Z)\mathcal{L}g(Z)]\Big\} \tag{3.C.20}$$

$$= \tfrac{1}{2}\mathbb{E}[\mathcal{L}(gh)(Z)] - \mathbb{E}[\Gamma(g,h)(Z)] \tag{3.C.21}$$

$$= -\mathbb{E}[g'(Z)h'(Z)], \tag{3.C.22}$$

where (3.C.20) uses (3.C.16); (3.C.21) uses the definition (3.C.18) of Γ, and (3.C.22) uses (3.C.19) together with (3.C.17). This proves (3.B.13).

Remark 3.37. If we consider the Hilbert space $L^2(G)$ of all functions $g\colon \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}[g^2(Z)] < \infty$ with $Z \sim G$, then (3.C.16) expresses the fact that $\mathcal{L}$ is a self-adjoint linear operator on this space. Moreover, (3.C.17) 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 $\mathcal{L}g = 0$).

Remark 3.38. The operator in the left side of (3.C.18) was introduced into the study of Markov processes by Paul Meyer under the name "carré du champ" (French for "square of the field"). In general, $\mathcal{L}$ in the right side of (3.C.18) can be an arbitrary linear operator that serves as an infinitesimal generator of a Markov semigroup. Intuitively, Γ in the left side of (3.C.18) 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 every g, h in its domain,

$$\mathcal{L}(gh) = g\mathcal{L}h + h\mathcal{L}g. \tag{3.C.23}$$

An example of a derivation is the first-order linear differential operator $\mathcal{L}g = g'$, for which the Leibniz rule is simply the product rule of differential calculus.

3.D LSI for Bernoulli and Gaussian measures

The following LSI was derived by Gross [44]:

$$\mathrm{Ent}_P[g^2] \leq \frac{(g(0)-g(1))^2}{2}. \tag{3.D.1}$$

We will now show that (3.3.37) can be derived from (3.D.1). Let us define f by $e^f = g^2$, where we may assume without loss of generality that $0 < g(0) \leq g(1)$. Note that

$$\begin{aligned}(g(0)-g(1))^2 &= (\exp(f(0)/2) - \exp(f(1)/2))^2 \\ &\leq \tfrac{1}{8}\left[\exp(f(0)) + \exp(f(1))\right](f(0)-f(1))^2 \\ &= \tfrac{1}{4}\,\mathbb{E}_P\left[\exp(f)(\Gamma f)^2\right] \end{aligned} \tag{3.D.2}$$

with $\Gamma f = |f(0) - f(1)|$, where (3.D.2) follows from the inequality $(1-x)^2 \leq \frac{1}{2}(1+x^2)(\ln x)^2$ for all $x \geq 0$, which is applied to $x \triangleq \frac{g(1)}{g(0)}$. Consequently, we have

$$D(P^{(f)}\|P) = \frac{\mathrm{Ent}_P[\exp(f)]}{\mathbb{E}_P[\exp(f)]} \tag{3.D.3}$$

$$= \frac{\mathrm{Ent}_P[g^2]}{\mathbb{E}_P[\exp(f)]} \tag{3.D.4}$$

$$\leq \frac{\big(g(0)-g(1)\big)^2}{2\,\mathbb{E}_P[\exp(f)]} \tag{3.D.5}$$

$$\leq \frac{\mathbb{E}_P[\exp(f)\,(\Gamma f)^2]}{8\,\mathbb{E}_P[\exp(f)]} \tag{3.D.6}$$

$$= \tfrac{1}{8}\,\mathbb{E}_P^{(f)}\left[(\Gamma f)^2\right] \tag{3.D.7}$$

where (3.D.3) follows from (3.3.6); (3.D.4) holds due to the equality $e^f = g^2$; (3.D.5) holds due to (3.D.1); (3.D.6) follows from (3.D.2), and (3.D.7) holds by the definition of the expectation with respect to the

tilted probability measure $P^{(f)}$. Therefore, we conclude that (3.D.1) implies (3.3.37).

Gross used (3.D.1) and the central limit theorem (CLT) to establish his Gaussian LSI (see Theorem 3.2.1). We can follow the same steps and arrive at (3.2.13) from (3.3.37). 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

$$f(x_1, \ldots, x_n) \triangleq g\left(\frac{x_1 + x_2 + \ldots + x_n - n/2}{\sqrt{n/4}}\right). \tag{3.D.8}$$

If $X_1, \ldots, X_n$ are i.i.d. Bernoulli(1/2) random variables, then, by the CLT, the sequence of probability measures $\{P_{Z_n}\}_{n=1}^{\infty}$ with

$$Z_n \triangleq \frac{X_1 + \ldots + X_n - n/2}{\sqrt{n/4}} \tag{3.D.9}$$

converges weakly to the standard Gaussian distribution G as $n \to \infty$: $P_{Z_n} \Rightarrow G$. By the assumed smoothness properties of g we therefore have (see (3.3.4) and (3.3.6))

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(f(X^n)\right)\right] \cdot D\big(P_{X^n}^{(f)} \| P_{X^n}\big) \\
&= \mathbb{E}\left[f(X^n)\exp\left(f(X^n)\right)\right] - \mathbb{E}[\exp\left(f(X^n)\right)] \ln \mathbb{E}[\exp\left(f(X^n)\right)] \\
&= \mathbb{E}\left[g(Z_n)\exp\left(g(Z_n)\right)\right] - \mathbb{E}[\exp\left(g(Z_n)\right)] \ln \mathbb{E}[\exp\left(g(Z_n)\right)] \\
&\xrightarrow{n\to\infty} \mathbb{E}\left[g(Z)\exp\left(g(Z)\right)\right] - \mathbb{E}[\exp\left(g(Z)\right)] \ln \mathbb{E}[\exp\left(g(Z)\right)] \\
&= \mathbb{E}\left[\exp\left(g(Z)\right)\right] D\big(P_Z^{(g)} \| P_Z\big)
\end{aligned} \tag{3.D.10}$$

where $Z \sim G$ is a standard Gaussian random variable. Moreover, using the definition (3.3.36) of Γ and the smoothness of g, it follows that for every $i \in \{1, \ldots, n\}$ and $x^n \in \{0,1\}^n$

$$\begin{aligned}
&|f(x^n \oplus e_i) - f(x^n)|^2 \\
&= \left| g\left(\frac{x_1 + \ldots + x_n - n/2}{\sqrt{n/4}} + \frac{(-1)^{x_i}}{\sqrt{n/4}}\right) - g\left(\frac{x_1 + \ldots + x_n - n/2}{\sqrt{n/4}}\right)\right|^2 \\
&= \frac{4}{n}\left(g'\left(\frac{x_1 + \ldots + x_n - n/2}{\sqrt{n/4}}\right)\right)^2 + o\left(\frac{1}{n}\right),
\end{aligned} \tag{3.D.11}$$

which implies that

$$|\Gamma f(x^n)|^2 = \sum_{i=1}^{n} (f(x^n \oplus e_i) - f(x^n))^2$$
$$= 4\left(g'\left(\frac{x_1 + \ldots + x_n - n/2}{\sqrt{n/4}}\right)\right)^2 + o(1). \qquad (3.D.12)$$

Consequently,

$$\mathbb{E}\left[\exp\left(f(X^n)\right)\right] \cdot \mathbb{E}^{(f)}\left[\left(\Gamma f(X^n)\right)^2\right]$$
$$= \mathbb{E}\left[\exp\left(f(X^n)\right)\left(\Gamma f(X^n)\right)^2\right]$$
$$= 4\,\mathbb{E}\left[\exp\left(g(Z_n)\right)\left(\left(g'(Z_n)\right)^2 + o(1)\right)\right]$$
$$\xrightarrow{n\to\infty} 4\,\mathbb{E}\left[\exp\left(g(Z)\right)\left(g'(Z)\right)^2\right]$$
$$= 4\,\mathbb{E}\left[\exp\left(g(Z)\right)\right] \cdot \mathbb{E}_{P_Z}^{(g)}\left[\left(g'(Z)\right)^2\right]. \qquad (3.D.13)$$

Taking the limit of both sides of (3.3.37) as $n \to \infty$ and then using (3.D.10) and (3.D.13), we obtain

$$D(P_Z^{(g)} \| P_Z) \le \tfrac{1}{2}\,\mathbb{E}_{P_Z}^{(g)}\left[\left(g'(Z)\right)^2\right], \qquad (3.D.14)$$

which is (3.2.13). The same technique applies to 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

$$f(x^n) \triangleq g\left(\frac{x_1 + \ldots + x_n - np}{\sqrt{npq}}\right), \qquad (3.D.15)$$

and then apply (3.3.41) to it.

3.E Generalization of Fano's inequality for list decoding

The following generalization of Fano's inequality for list decoding 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 $\mathcal{L}\colon \mathcal{Y} \to 2^{\mathcal{X}}$ which maps every $y \in \mathcal{Y}$

to a set $\mathcal{L}(y) \subseteq \mathcal{X}$, such that $|\mathcal{L}(Y)| \leq N$ a.s.. Let $P_{\mathrm{e}} = \mathbb{P}\left(X \notin \mathcal{L}(Y)\right)$ designate the list decoding error. Then

$$H(X|Y) \leq h(P_{\mathrm{e}}) + (1 - P_{\mathrm{e}}) \ln N + P_{\mathrm{e}} \ln |\mathcal{X}| \tag{3.E.1}$$

(see, e.g., [182] or [191, Lemma 1]). For proving (3.E.1), define the indicator random variable $E \triangleq 1_{\{X \notin \mathcal{L}(Y)\}}$. Then we can expand the conditional entropy $H(E, X|Y)$ in two ways as

$$H(E, X|Y) = H(E|Y) + H(X|E, Y) \tag{3.E.2a}$$
$$= H(X|Y) + H(E|X, Y). \tag{3.E.2b}$$

Since X and Y uniquely determine E (for a given mapping $\mathcal{L}$), the quantity on the right side of (3.E.2b) is equal to $H(X|Y)$. On the other hand, we can upper-bound the right side of (3.E.2a) as

$$H(E|Y) + H(X|E, Y) \leq H(E) + H(X|E, Y) \tag{3.E.3}$$
$$\leq h(P_{\mathrm{e}}) + (1 - P_{\mathrm{e}}) \ln N + P_{\mathrm{e}} \ln |\mathcal{X}|, \tag{3.E.4}$$

where we have bounded the conditional entropy $H(X|E, Y)$ in (3.E.3) as follows:

$$\begin{aligned}
&H(X|E, Y) \\
&= \sum_{y \in \mathcal{Y}} \mathbb{P}(E = 0, Y = y)\, H(X|E = 0, Y = y) \\
&\qquad + \sum_{y \in \mathcal{Y}} \mathbb{P}(E = 1, Y = y)\, H(X|E = 1, Y = y) && \text{(3.E.5)} \\
&\leq \sum_{y \in \mathcal{Y}} \Big\{ \mathbb{P}(E = 0, Y = y)\, H(X|E = 0, Y = y) \Big\} + P_{\mathrm{e}} \ln |\mathcal{X}| && \text{(3.E.6)} \\
&= (1 - P_{\mathrm{e}}) \sum_{y \in \mathcal{Y}} \Big\{ \mathbb{P}(Y = y|E = 0)\, H(X|E = 0, Y = y) \Big\} \\
&\quad + P_{\mathrm{e}} \ln |\mathcal{X}| && \text{(3.E.7)} \\
&\leq (1 - P_{\mathrm{e}})\, \mathbb{E}\big[\ln |\mathcal{L}(Y)| \,\big|\, E = 0 \big] + P_{\mathrm{e}} \ln |\mathcal{X}| && \text{(3.E.8)} \\
&\leq (1 - P_{\mathrm{e}}) \ln N + P_{\mathrm{e}} \ln |\mathcal{X}|, && \text{(3.E.9)}
\end{aligned}$$

where (3.E.6) and (3.E.8) rely on the standard log-cardinality bound on the entropy; (3.E.8) uses the fact that, given $E = 0$ and $Y = y$, X

is supported on the set $\mathcal{L}(y)$. Since

$$H(X|Y) = H(E|Y) + H(X|E,Y) \le H(E) + H(X|E,Y), \quad (3.E.10)$$

we get (3.E.1).

Remark 3.39. If instead of assuming that $\mathcal{L}(Y)$ is bounded a.s. we assume that it is bounded in expectation, i.e., if $\mathbb{E}[\ln|\mathcal{L}(Y)|] < \infty$, then we can obtain a weaker inequality

$$H(X|Y) \le \mathbb{E}\,[\ln|\mathcal{L}(Y)|] + h(P_{\mathrm{e}}) + P_{\mathrm{e}} \ln|\mathcal{X}|. \quad (3.E.11)$$

To get this, we follow the same steps as before, except (3.E.9) in the above series of bounds on $H(X|E,Y)$ is replaced by

$$\begin{aligned}
&(1-P_{\mathrm{e}})\,\mathbb{E}\big[\ln|\mathcal{L}(Y)|\,\big|\,E=0\big] \\
&\le (1-P_{\mathrm{e}})\,\mathbb{E}\big[\ln|\mathcal{L}(Y)|\,\big|\,E=0\big] + P_{\mathrm{e}}\,\mathbb{E}\big[\ln|\mathcal{L}(Y)|\,\big|\,E=1\big] \quad (3.E.12)\\
&= \mathbb{E}\,[\ln|\mathcal{L}(Y)|] \quad (3.E.13)
\end{aligned}$$

(we assume, of course, that $\mathcal{L}(y)$ is a nonempty set for all $y \in \mathcal{Y}$).

3.F Details for the derivation of (3.6.102)

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, $(X^n, Y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ with $X^n \sim P_{X^n}$ and

$$\begin{aligned}
&P_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^{n} P_{Y|X}(y_i|x_i),\\
&\forall\, y^n \in \mathcal{Y}^n,\ \forall\, x^n \in \mathcal{X}^n \text{ s.t. } P_{X^n}(x^n) > 0.
\end{aligned}$$

Since the channel is memoryless and there is no feedback, the i-th output symbol $Y_i \in \mathcal{Y}$ depends only on the i-th input symbol $X_i \in \mathcal{X}$ and not on the rest of the input symbols $\overline{X}^i$. Hence, $\overline{Y}^i \to X_i \to Y_i$ is a Markov chain for every $i \in \{1,\dots,n\}$, so we can write

$$\begin{aligned}
P_{Y_i|\overline{Y}^i}(y|\overline{y}^i) &= \sum_{x\in\mathcal{X}} P_{Y_i|X_i}(y|x) P_{X_i|\overline{Y}^i}(x|\overline{y}^i) \quad (3.F.1)\\
&= \sum_{x\in\mathcal{X}} P_{Y|X}(y|x) P_{X_i|\overline{Y}^i}(x|\overline{y}^i) \quad (3.F.2)
\end{aligned}$$

for all $y \in \mathcal{Y}$ and $\overline{y}^i \in \mathcal{Y}^{n-1}$ such that $P_{\overline{Y}^i}(\overline{y}^i) > 0$. Therefore, for every $y, y' \in \mathcal{Y}$, we have

$$\ln \frac{P_{Y_i|\overline{Y}^i}(y|\overline{y}^i)}{P_{Y_i|\overline{Y}^i}(y'|\overline{y}^i)} = \ln \frac{\sum_{x\in\mathcal{X}} P_{Y|X}(y|x) P_{X_i|\overline{Y}^i}(x|\overline{y}^i)}{\sum_{x\in\mathcal{X}} P_{Y|X}(y'|x) P_{X_i|\overline{Y}^i}(x|\overline{y}^i)} \tag{3.F.3}$$

$$= \ln \frac{\sum_{x\in\mathcal{X}} P_{Y|X}(y'|x) P_{X_i|\overline{Y}^i}(x|\overline{y}^i) \frac{P_{Y|X}(y|x)}{P_{Y|X}(y'|x)}}{\sum_{x\in\mathcal{X}} P_{Y|X}(y'|x) P_{X_i|\overline{Y}^i}(x|\overline{y}^i)}, \tag{3.F.4}$$

where in the last line we have used the fact that $P_{Y|X}(\cdot|\cdot) > 0$. This shows that we can express the left side of (3.F.3) as the logarithm of expectation of $\frac{P_{Y|X}(y|X)}{P_{Y|X}(y'|X)}$ with respect to the (conditional) probability measure

$$Q(x|\overline{y}^i, y') = \frac{P_{Y|X}(y'|x) P_{X_i|\overline{Y}^i}(x|\overline{y}^i)}{\sum_{x\in\mathcal{X}} P_{Y|X}(y'|x) P_{X_i|\overline{Y}^i}(x|\overline{y}^i)}, \qquad \forall\, x \in \mathcal{X}. \tag{3.F.5}$$

Therefore,

$$\ln \frac{P_{Y_i|\overline{Y}^i}(y|\overline{y}^i)}{P_{Y_i|\overline{Y}^i}(y'|\overline{y}^i)} \le \max_{x\in\mathcal{X}} \ln \frac{P_{Y|X}(y|x)}{P_{Y|X}(y'|x)}. \tag{3.F.6}$$

Interchanging the roles of y and y', we get

$$\ln \frac{P_{Y_i|\overline{Y}^i}(y'|\overline{y}^i)}{P_{Y_i|\overline{Y}^i}(y|\overline{y}^i)} \le \max_{x\in\mathcal{X}} \ln \frac{P_{Y|X}(y'|x)}{P_{Y|X}(y|x)}. \tag{3.F.7}$$

Combining (3.F.6) and (3.F.7) with the definition in (3.6.98), it follows that for all $y, y' \in \mathcal{Y}$

$$\left| \ln \frac{P_{Y_i|\overline{Y}^i}(y|\overline{y}^i)}{P_{Y_i|\overline{Y}^i}(y'|\overline{y}^i)} \right| \le \max_{x\in\mathcal{X}} \max_{y,y'\in\mathcal{Y}} \left| \ln \frac{P_{Y|X}(y|x)}{P_{Y|X}(y'|x)} \right| = \tfrac{1}{2}\, c(T). \tag{3.F.8}$$

Similarly, we have

$$\left| \ln \frac{P^{(\mathcal{C})}_{Y_i|\overline{Y}^i}(y|\overline{y}^i)}{P^{(\mathcal{C})}_{Y_i|\overline{Y}^i}(y'|\overline{y}^i)} \right| \le \max_{x\in\mathcal{X}} \max_{y,y'\in\mathcal{Y}} \left| \ln \frac{P_{Y|X}(y|x)}{P_{Y|X}(y'|x)} \right| = \tfrac{1}{2}\, c(T), \tag{3.F.9}$$

since both left sides of (3.F.8) and (3.F.9) refer to the same conditional probability distribution $P_{Y|X}$. Combining (3.F.8) and (3.F.9) yields (3.6.102).

Acknowledgments

It is a pleasure to thank several individuals, who have carefully read parts of the manuscript in various stages and provided constructive comments, suggestions, and corrections. These include Tim van Erven, Ronen Eshel, Peter Harremoës, Eran Hof, Nicholas Kalouptsidis, Leor Kehaty, Aryeh Kontorovich, Ioannis Kontoyannis, Mokshay Madiman, Daniel Paulin, Yury Polyanskiy, Boaz Shuval, Emre Telatar, Sergio Verdú, Yihong Wu and Kostis Xenoulis. Among these people, Leor Kehaty is gratefully acknowledged for a very detailed report on the initial draft of this manuscript, and Boaz Shuval is acknowledged for some helpful comments on the first edition. The authors are thankful to the three anonymous reviewers and the Editor in Chief, Sergio Verdú, for very constructive and detailed suggestions, which contributed a lot to the presentation of the first edition of this manuscript. 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, the Swiss Federal Institute of Technology in Lausanne, during the summer of 2011 is acknowledged by I. Sason. We would like to thank the organizers of the *Information Theory and Applications Workshop* in San-Diego, California; our collaboration in

this project was initiated during this successful workshop in Feb. 2012. Finally, we are grateful to the publishers of the *Foundations and Trends (FnT) in Communications and Information Theory*: Mike Casey, James Finlay and Alet Heezemans for their assistance in both the first and second editions of this monograph (dated: Oct. 2013 and Sept. 2014, respectively, with another revision in Dec. 2016).

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. Available at `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. Available at `http://www.math.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, June 2005. Available at `http://arxiv.org/abs/math/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] S. Ghosh and L. Goldstein. Concentration of measure via size-bias coupling. *Probability Theory and Related Fields*, 149:271–278, February 2011.

[25] S. Ghosh and L. Goldstein. Applications of size-biased couplings for concentration of measures. *Electronic Communications in Probability*, 16:70–83, January 2011.

[26] L. Goldstein and U. Işlak. Concentration inequalities via zero bias coupling. *Statistics and Probability Letters*, 86:17–23, January 2014.

[27] L. Mackey, M. I. Jordan, R. Y. Chen, B. Farrell, and J. A. Tropp. Matrix concentration inequalities via the method of exchangeable pairs. *Annals of Probability*, 10(2):906–945, 2014.

[28] D. Paulin. The convex distance inequality for dependent random variables, with applications to the stochastic travelling salesman and other problems. *Electronic Journal of Probability*, 19(68):1–34, August 2014.

[29] E. Abbe and A. Montanari. On the concentration of the number of solutions of random satisfiability formulas. *Random Structures and Algorithms*, 45(3):362–382, October 2014.

[30] 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.

[31] 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.

[32] S. Kudekar. *Statistical Physics Methods for Sparse Graph Codes*. PhD thesis, EPFL - Swiss Federal Institute of Technology, Lausanne, Switzeland, July 2009.

[33] 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.

[34] 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.

[35] A. Montanari. Tight bounds for LDPC and LDGM codes under MAP decoding. *IEEE Trans. on Information Theory*, 51(9):3247–3261, September 2005.

[36] M. Talagrand. *Mean Field Models for Spin Glasses*. Springer-Verlag, 2010.

[37] S. Bobkov and M. Madiman. Concentration of the information in data with log-concave distributions. *Annals of Probability*, 39(4):1528–1543, 2011.

[38] 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.

[39] E. Shamir and J. Spencer. Sharp concentration of the chromatic number on random graphs. *Combinatorica*, 7(1):121–129, 1987.

[40] 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.

[41] 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.

[42] M. Sipser and D. A. Spielman. Expander codes. *IEEE Trans. on Information Theory*, 42(6):1710–1722, November 1996.

[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] M. H. M. Costa. A new entropy power inequality. *IEEE Trans. on Information Theory*, 31(6):751–760, November 1985.

[48] A. Dembo, T. M. Cover, and J. A. Thomas. Information theoretic inequalities. *IEEE Trans. on Information Theory*, 37(6):1501–1518, November 1991.

[49] C. Villani. A short proof of the 'concavity of entropy power'. *IEEE Trans. on Information Theory*, 46(4):1695–1696, July 2000.

[50] G. Toscani. An information-theoretic proof of Nash's inequality. *Rendiconti Lincei: Matematica e Applicazioni*, 24(1):83–93, 2013.

[51] A. Guionnet and B. Zegarlinski. Lectures on logarithmic Sobolev inequalities. *Séminaire de probabilités (Strasbourg)*, 36:1–134, 2002.

[52] 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.

[53] 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.

[54] S. G. Bobkov and F. Götze. Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. *Journal of Functional Analysis*, 163:1–28, 1999.

[55] 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.

[56] S. G. Bobkov and P. Tetali. Modified logarithmic Sobolev inequalities in discrete settings. *Journal of Theoretical Probability*, 19(2):289–336, 2006.

[57] D. Chafaï. Entropies, convexity, and functional inequalities: Φ-entropies and Φ-Sobolev inequalities. *J. Math. Kyoto University*, 44(2):325–363, 2004.

[58] C. P. Kitsos and N. K. Tavoularis. Logarithmic Sobolev inequalities for information measures. *IEEE Trans. on Information Theory*, 55(6):2554–2561, June 2009.

[59] K. Marton. Bounding $\bar{d}$-distance by informational divergence: a method to prove measure concentration. *Annals of Probability*, 24(2):857–866, 1996.

[60] K. Marton. Distance-divergence inequalities. *IEEE Information Theory Society Newsletter*, 64(1):9–13, March 2014.

[61] 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.

[62] 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.

[63] Y. Polyanskiy and S. Verdú. Empirical distribution of good channel codes with non-vanishing error probability. *IEEE Trans. on Information Theory*, 60(1):5–21, January 2014.

[64] Y. Polyanskiy and Y. Wu. Wasserstein continuity of entropy and outer bounds for interference channels. *IEEE Trans. on Information Theory*, 62(7):3992–4002, July 2016.

[65] Y. Steinberg and S. Verdú. Simulation of random processes and rate-distortion theory. *IEEE Trans. on Information Theory*, 42(1):63–86, January 1996.

[66] C. Villani. *Topics in Optimal Transportation*. American Mathematical Society, Providence, RI, 2003.

[67] C. Villani. *Optimal Transport: Old and New*. Springer, 2009.

[68] P. Cattiaux and A. Guillin. On quadratic transportation cost inequalities. *Journal de Matématiques Pures et Appliquées*, 86:342–361, 2006.

[69] A. Dembo and O. Zeitouni. Transportation approach to some concentration inequalities in product spaces. *Electronic Communications in Probability*, 1:83–90, 1996.

[70] 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.

[71] N. Gozlan. A characterization of dimension free concentration in terms of transportation inequalities. *Annals of Probability*, 37(6):2480–2498, 2009.

[72] E. Milman. Properties of isoperimetric, functional and transport-entropy inequalities via concentration. *Probability Theory and Related Fields*, 152:475–507, 2012.

[73] 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.

[74] 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.

[75] K. Marton. A simple proof of the blowing-up lemma. *IEEE Trans. on Information Theory*, 32(3):445–446, May 1986.

[76] Y. Altuğ and A. B. Wagner. Refinement of the sphere-packing bound: asymmetric channels. *IEEE Trans. on Information Theory*, 60(3):1592–1614, March 2014.

[77] 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.

[78] 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.

[79] Y. Kontoyiannis and S. Verdú. Optimal lossless data compression: non-asymptotics and asymptotics. *IEEE Trans. on Information Theory*, 60(2):777–795, February 2014.

[80] 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.

[81] W. Matthews. A linear program for the finite block length converse of Polyanskiy-Poor-Verdú via nonsignaling codes. *IEEE Trans. on Information Theory*, 59(12):7036–7044, December 2012.

[82] 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.

[83] 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.

[84] J. S. Rosenthal. *A First Look at Rigorous Probability Theory*. World Scientific, second edition, 2006.

[85] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Springer, second edition, 1997.

[86] B. Efron and C. Stein. The jackknife estimate of variance. *Annals of Statistics*, 9:586–596, 1981.

[87] J. M. Steele. An Efron–Stein inequality for nonsymmetric statistics. *Annals of Statistics*, 14:753–758, 1986.

[88] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer, 2001.

[89] 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.

[90] S. N. Bernstein. *The Theory of Probability*. Gos. Izdat., Moscow/Leningrad, 1927. in Russian.

[91] S. Verdú. *Multiuser Detection*. Cambridge University Press, 1998.

[92] C. McDiarmid. Centering sequences with bounded differences. *Combinatorics, Probability and Computing*, 6(1):79–86, March 1997.

[93] C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, volume 141, pages 148–188. Cambridge University Press, 1989.

[94] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

[95] E. Rio. On Mcdiarmid's concentration inequality. *Electronic Communications in Probability*, 18(44):1–11, 2013.

[96] J. Dedecker and X. Fan. Deviation inequalities for separately Lipschitz functionals of iterated random variables. *Stochastic Processes and their Applications*, accepted in August 2014.
Available at `http://dx.doi.org/10.1016/j.spa.2014.08.001`.

[97] 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.

[98] D. Berend and A. Kontorovich. On the concentration of the missing mass. *Electronic Communications in Probability*, 18(3):1–7, January 2013.

[99] S. G. From and A. W. Swift. A refinement of Hoeffding's inequality. *Journal of Statistical Computation and Simulation*, pages 1–7, December 2011.

[100] X. Fan, I. Grama, and Q. Liu. Hoeffding's inequality for supermartingales. *Stochastic Processes and their Applications*, 122(10):3545–3559, October 2012.

[101] X. Fan, I. Grama, and Q. Liu. Large deviation exponential inequalities for supermartingales. *Electronic Communications in Probability*, 17(59):1–8, December 2012.

[102] P. Billingsley. *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics, 3rd edition, 1995.

[103] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. Oxford University Press, third edition, 2001.

[104] 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.

[105] A. Barg and G. D. Forney. Random codes: minimum distances and error exponents. *IEEE Trans. on Information Theory*, 48(9):2568–2573, September 2002.

[106] M. Breiling. A logarithmic upper bound on the minimum distance of turbo codes. *IEEE Trans. on Information Theory*, 50(8):1692–1710, August 2004.

[107] A. F. Molisch. *Wireless Communications*. John Wiley and Sons, 2005.

[108] 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. *IEEE Signal Processing Magazine*, 30(6):130–144, November 2013.

[109] S. Litsyn and G. Wunder. Generalized bounds on the crest-factor distribution of OFDM signals with applications to code design. *IEEE Trans. on Information Theory*, 52(3):992–1006, March 2006.

[110] R. Salem and A. Zygmund. Some properties of trigonometric series whose terms have random signs. *Acta Mathematica*, 91(1):245–301, 1954.

[111] 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.

[112] I. Sason. On the concentration of the crest factor for OFDM signals. In *Proceedings of the 8th International Symposium on Wireless Communication Systems*, pages 784–788, Aachen, Germany, November 2011.

[113] R. G. Gallager. *Low-Density Parity-Check Codes*. PhD thesis, MIT, Cambridge, MA, USA, 1963.

[114] 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.

[115] I. Sason. On universal properties of capacity-approaching LDPC code ensembles. *IEEE Trans. on Information Theory*, 55(7):2956–2990, July 2009.

[116] I. Sason and R. Eshel. On concentration of measures for LDPC code ensembles. In *Proceedings of the 2011 IEEE International Symposium on Information Theory*, pages 1273–1277, Saint Petersburg, Russia, August 2011.

[117] 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.

[118] 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.

[119] R. Eshel. *Aspects of Convex Optimization and Concentration in Coding*. Technion - Israel Institute of Technology, Haifa, Israel, February 2012.

[120] 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.

[121] 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.

[122] A. Shokrollahi. Capacity-achieving sequences. In *Volume in Mathematics and its Applications*, volume 123, pages 153–166, 2000.

[123] 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.

[124] 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.

[125] A. P. Godbole and P. Hitczenko. Beyond the method of bounded differences. In *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, volume 41, pages 43–58. American Mathematical Society, 1998.

[126] 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.

[127] S. Verdú and T. Weissman. The information lost in erasures. *IEEE Trans. on Information Theory*, 54(11):5030–5058, November 2008.

[128] E. A. Carlen. Superadditivity of Fisher's information and logarithmic Sobolev inequalities. *Journal of Functional Analysis*, 101:194–211, 1991.

[129] 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.

[130] G. Blower. *Random Matrices: High Dimensional Phenomena*. London Mathematical Society Lecture Notes. Cambridge University Press, Cambridge, U.K., 2009.

[131] O. Johnson. *Information Theory and the Central Limit Theorem*. Imperial College Press, London, 2004.

[132] E. H. Lieb and M. Loss. *Analysis*. American Mathematical Society, Providence, RI, 2nd edition, 2001.

[133] 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.

[134] P. J. Huber and E. M. Ronchetti. *Robust Statistics*. Wiley Series in Probability and Statistics, second edition, 2009.

[135] O. Johnson and A. Barron. Fisher information inequalities and the central limit theorem. *Probability Theory and Related Fields*, 129:391–409, 2004.

[136] S. Verdú. Mismatched estimation and relative entropy. *IEEE Trans. on Information Theory*, 56(8):3712–3720, August 2010.

[137] H. L. van Trees. *Detection, Estimation and Modulation Theory, Part I*. Wiley, 1968.

[138] L. C. Evans and R. F. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, 1992.

[139] M. C. Mackey. *Time's Arrow: The Origins of Thermodynamic Behavior*. Springer, New York, 1992.

[140] B. Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, Berlin, 5 edition, 1998.

[141] I. Karatzas and S. Shreve. *Brownian Motion and Stochastic Calculus*. Springer, second edition, 1988.

[142] F. C. Klebaner. *Introduction to Stochastic Calculus with Applications*. Imperial College Press, second edition, 2005.

[143] T. van Erven and P. Harremoës. Rényi divergence and Kullback-Leibler divergence. *IEEE Trans. on Information Theory*, 60(7):3797–3820, July 2014.

[144] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley and Sons, second edition, 2006.

[145] A. Maurer. Thermodynamics and concentration. *Bernoulli*, 18(2):434–454, 2012.

[146] N. Merhav. *Statistical Physics and Information Theory*, volume 6 of Foundations and Trends in Communications and Information Theory. Now Publishers, Delft, the Netherlands, 2009.

[147] S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities using the entropy method. *Annals of Probability*, 31(3):1583–1614, 2003.

[148] I. Kontoyiannis and M. Madiman. Measure concentration for compound Poisson distributions. *Electronic Communications in Probability*, 11:45–57, 2006.

[149] M. Gromov. *Metric Structures for Riemannian and Non-Riemannian Spaces*. Birkhäuser, 2001.

[150] S. Bobkov. A functional form of the isoperimetric inequality for the Gaussian measure. *Journal of Functional Analysis*, 135:39–49, 1996.

[151] L. V. Kantorovich. On the translocation of masses. *Journal of Mathematical Sciences*, 133(4):1381–1382, 2006.

[152] E. Ordentlich and M. Weinberger. A distribution dependent refinement of Pinsker's inequality. *IEEE Trans. on Information Theory*, 51(5):1836–1840, May 2005.

[153] 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.

[154] I. Csiszár. Sanov property, generalized I-projection and a conditional limit theorem. *Annals of Probability*, 12(3):768–793, 1984.

[155] 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.

[156] M. Talagrand. Transportation cost for Gaussian and other product measures. *Geometry and Functional Analysis*, 6(3):587–600, 1996.

[157] R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2004.

[158] F. Otto and C. Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, June 2000.

[159] Y. Wu. A simple transportation-information inequality with applications to HWI inequalities, and predictive density estimation. Technical Report, September 2011.

[160] D. Cordero-Erausquin. Some applications of mass transport to Gaussian-type inequalities. *Archive for Rational Mechanics and Analysis*, 161(3):257–269, February 2002.

[161] D. Bakry and M. Emery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX*, volume 1123 of *Lecture Notes in Mathematics*, pages 177–206. Springer, 1985.

[162] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistics Society*, 28(1):131–142, 1966.

[163] I. Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf den bewis der ergodizität von markhoffschen ketten. *Publ. Math. Inst. Hungar. Acad. Sci.*, 8(8):85–108, 1963.

[164] I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, January 1967.

[165] M. Zakai and J. Ziv. A generalization of the rate-distortion theory and applications. In G. Longo, editor, *Information Theory - New Trends and Open Problems*, pages 87–123. Springer, 1975.

[166] I. Sason and S. Verdú. f-divergence inequalities. *IEEE Trans. on Information Theory*, 62(11):5973–6006, November 2016.

[167] P.-M. Samson. Concentration of measure inequalities for Markov chains and ϕ-mixing processes. *Annals of Probability*, 28(1):416–461, 2000.

[168] 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.

[169] K. Marton. Measure concentration for Euclidean distance in the case of dependent random variables. *Annals of Probability*, 32(3B):2526–2544, 2004.

[170] K. Marton. Correction to 'Measure concentration for Euclidean distance in the case of dependent random variables'. *Annals of Probability*, 38(1):439–442, 2010.

[171] R. L. Dobrushin and S. B. Shlosman. Completely analytical Gibbs fields. In *Statistical Physics and Dynamical Systems*, pages 371–403. Springer, 1985.

[172] K. Marton. An inequality for relative entropy and logarithmic Sobolev inequalities in Euclidean spaces. *Journal of Functional Analysis*, 264(1):34–61, January 2013.

[173] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2nd edition, 2011.

[174] G. Margulis. Probabilistic characteristics of graphs with large connectivity. *Problems of Information Transmission*, 10(2):174–179, 1974.

[175] A. El Gamal and Y.-H. Kim. *Network Information Theory*. Cambridge University Press, 2011.

[176] 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.

[177] F. M. J. Willems. The maximal-error and average-error capacity regions of the broadcast channel are identical: a direct proof. *Problems of Control and Information Theory*, 19(4):339–347, 1990.

[178] T. M. Cover. Broadcast channels. *IEEE Trans. on Information Theory*, 18(1):2–14, January 1972.

[179] P. P. Bergmans. Random coding theorem for broadcast channels with degraded components. *IEEE Trans. on Information Theory*, 19(2):197–207, March 1973.

[180] A. D. Wyner. A theorem on the entropy of certain binary sequences and applications: Part II. *IEEE Trans. on Information Theory*, 19(6):772–777, March 1973.

[181] R. G. Gallager. Capacity and coding for degraded broadcast channels. *Problems of Information Transmission*, 10(3):3–14, July-September 1974.

[182] 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.

[183] S. Shamai and S. Verdú. The empirical distribution of good codes. *IEEE Trans. on Information Theory*, 43(3):836–846, May 1997.

[184] T. S. Han and S. Verdú. Approximation theory of output statistics. *IEEE Trans. on Information Theory*, 39(3):752–772, May 1993.

[185] M. Raginsky and I. Sason. Refined bounds on the empirical distribution of good channel codes via concentration inequalities. In *Proceedings of the 2013 IEEE International Workshop on Information Theory*, pages 221–225, Istanbul, Turkey, July 2013.

[186] F. Topsøe. An information theoretical identity and a problem involving capacity. *Studia Scientiarum Mathematicarum Hungarica*, 2:291–292, 1967.

[187] J. H. B. Kemperman. On the Shannon capacity of an arbitrary channel. *Indagationes Mathematicae*, 77(2):101–115, 1974.

[188] U. Augustin. Gedächtnisfreie Kanäle für diskrete Zeit. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 6:10–61, 1966.

[189] 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.

[190] Y. Kontoyiannis. Sphere-covering, measure concentration, and source coding. *IEEE Trans. on Information Theory*, 47(4):1544–1552, May 2001.

[191] Y. H. Kim, A. Sutivong, and T. M. Cover. State amplification. *IEEE Trans. on Information Theory*, 54(5):1850–1859, May 2008.