Measure Concentration of Strongly Mixing Processes with Applications

Leonid Kontorovich

May 2007
CMU-ML-07-104

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
John Lafferty, CMU (Chair)
Kavita Ramanan, CMU
Larry Wasserman, CMU
Gideon Schechtman, Weizmann Institute of Science

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

This work was partially supported by NSF ITR grant IIS-0205456 and DMS-0323668-0000000965.

The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

Keywords: concentration of measure, Lipschitz function, stochastic process, mixing

Abstract

The concentration of measure phenomenon was first discovered in the 1930's by Paul Lévy and has been investigated since then, with increasing intensity in recent decades. The probability-theoretic results have been gradually percolating throughout the mathematical community, finding applications in Banach space geometry, analysis of algorithms, statistics and machine learning.

There are several approaches to proving concentration of measure results; we shall offer a brief survey of these. The principal contribution of this thesis is a functional norm inequality, which immediately implies a concentration inequality for nonproduct measures. The inequality is proved by elementary means, yet enables one, with minimal effort, to recover and generalize the best current results for Markov chains, as well as to obtain new results for hidden Markov chains and Markov trees.

As an application of our inequalities, we give a strong law of large numbers for a broad class of non-independent processes. In particular, this allows one to analyze the convergence of inhomogeneous Markov Chain Monte Carlo algorithms. We also give some partial results on extending the Rademacher-type generalization bounds to processes with arbitrary dependence.

We end the thesis with some conjectures and open problems.

Acknowledgements

ἀνδρες μοι ἐννεπε, μοῦσα
(with apologies to Homer)

The first order of business is to thank my parents, who gave me life. Though that would've been quite sufficient, they actually gave me much more: uninterrupted unconditional support, a thirst for knowledge, and a set of values. I hope this brings you *naches*.

Throughout my career, I've had the distinctive good fortune of falling upon exceptional mentors. Daniel D. Lee took me in as an undergraduate summer intern at Bell Labs; I entered not knowing the difference between *cognitive science* and *machine learning* and emerged a sworn connectionist-empiricist, with a set of tools that I use to this day. (Thanks to Sebastian Seung for introducing us.) Yoram Singer put up with my inability to program (not to mention my broken Hebrew) and encouraged me to come up with original (and often half-baked) ideas, which he would then subject to much tough love.

I will never forget the warm encouragement of Yakov Sinai. The terror induced by his soft-spoken "No, I disagree" can only be compared to the joy of hearing him say "Yes, this looks right"; I've been on the receiving end of both. Many thanks to John Hopfield for teaching a memorable class on neural networks and agreeing to supervise my senior thesis. I will always be indebted to Corinna Cortes and Mehryar Mohri, who recognized the potential of learning languages by embeddings at a time when few people would take me seriously (our work is not included in this thesis but has led to a separate fruitful research direction). I thank Anthony Brockwell for providing the first nontrivial real-world application of my work, and Cosma Shalizi for introducing us, and also for teaching an excellent class on stochastic processes (which I'd take several times over if I could), and our many valuable conversations.

I owe a special thanks to my primary thesis advisor, John Lafferty. I thank John for giving me all the creative freedom I needed, for teaching me how to debug ideas, and of course for suggesting that I look at generalization bounds for non-iid samples. I can proudly say that I learned an intricate and beautiful art, first-hand from a master; none of this work would have been possible without his guidance.

I also extend my gratitude to Larry Wasserman and Gideon Schechtman – and not just for agreeing to serve on my Committee. Larry's inspiring course on nonparametric statistics gave me an early taste for functional analysis techniques in empirical process theory. In the short time that I've known Gideon, he has opened my eyes to the iceberg whose tip I'd only begun to scrape; I am looking forward to our collaboration at Weizmann. I thank Kavita Ramanan for agreeing to advise and support me so late into the program and for her encouragement throughout the ups and downs of the research process.

Ricardo Silva saved me countless hours by generously providing the LaTeX style file from his

thesis. Alla Gil, Vladimir Kontorovich, John Lafferty, Steven J. Miller, Kavita Ramanan and Gideon Schechtman made valuable comments on an earlier draft. Our program manager Diane Stidle heroically kept me on task through the length of the PhD program.

Contents

Chapter 1

Introduction

1.1 Background

The study of measure concentration in general metric spaces was initiated in the 1970's by Vitali Milman, who in turn drew inspiration from Paul Lévy's work (see [65] for a brief historical exposition). Since then, various deep insights have been gained into the concentration of measure phenomenon [40], with a particular surge of activity in the last decade.

The words "measure" and "concentration" suggest an interplay of analytic and geometric aspects. Indeed, there are two essential ingredients in proving a concentration result: the random variable must be continuous in a strong (Lipschitz) sense[1], and the random process must be mixing in some strong sense. We will give simple examples to illustrate how, in general, the failure of either of these conditions to hold can prevent a random variable from being concentrated.

A common way of summarizing the phenomenon is to say that in a high-dimensional space, almost all of the probability is concentrated around any set whose measure is at least $\frac{1}{2}$. Another way is to say that any "sufficiently continuous" function is tightly concentrated about its mean. To state this more formally (but still somewhat imprecisely), let $(X_i)_{1 \le i \le n}$, $X_i \in \Omega$, be the random process defined on the probability space $(\Omega^n, \mathcal{F}, \mathbf{P})$, and $f : \Omega^n \to \mathbb{R}$ be a function satisfying some Lipschitz condition – and possibly others, such as convexity. For our purposes, a concentration of measure result is an inequality of the form

$$\mathbf{P}\{|f(X) - \mathbf{E}f(X)| > t\} \quad \le \quad c \exp(-Kt^2) \tag{1.1}$$

where $c > 0$ is a small constant (typically, $c = 2$) and $K > 0$ depends on the strong mixing properties of $\mathbf{P}$ as well as the underlying metric. It is crucial that neither c nor K depend on f.[2]

A few celebrated milestones that naturally fall into the paradigm of (1.1) include Lévy's original isoperimetric inequality on the sphere (see the notes and references in [41]), McDiarmid's bounded differences inequality [51], and Marton's generalization of the latter for contracting Markov chains [45]. (Talagrand's no-less celebrated series of results [65] does not easily lend itself to such a compact description.)

Building on the work of Azuma [3] and Hoeffding [27], McDiarmid showed that if $f : \Omega^n \to \mathbb{R}$ has $\|f\|_{\text{Lip}} \le 1$ under the normalized Hamming metric and $\mathbf{P}$ is a product measure on Ω^n, we have

$$\mathbf{P}\{|f - \mathbf{E}f| > t\} \quad \le \quad 2 \exp(-2nt^2) \tag{1.2}$$

[1] But see [70] for some recent results on concentration of non-Lipschitz functions.

[2] See [40] for a much more general notion of concentration.

(he actually proved this for the more general class of weighted Hamming metrics). Using coupling and information-theoretic inequalities, Marton showed that if the conditions on $f : \Omega^n \to \mathbb{R}$ are as above and $\mathbf{P}$ is a contracting Markov measure on Ω^n with Doeblin coefficient $\theta < 1$,

$$\mathbf{P}\{|f - M_f| > t\} \quad \leq \quad 2\exp\left[-2n\left(t(1-\theta) - \sqrt{\frac{\log 2}{2n}}\right)^2\right], \tag{1.3}$$

where M_f is a $\mathbf{P}$-median of f. Since product measures are degenerate cases of Markov measures (with $\theta = 0$), Marton's result is a powerful generalization of (1.2).

Two natural directions for extending results of type (1.2) are to derive such inequalities for various measures and metrics. Talagrand's paper [65] is a tour de force in proving concentration for various (not necessarily metric) notions of distance, but it deals exclusively with product measures. Since the publication of Marton's concentration inequality in 1996 – to our knowledge, the first of its kind for a nonproduct, non-Haar measure – several authors proceeded to generalize her information-theoretic approach [15, 16], and offer alternative approaches based on the entropy method [38, 60] or martingale techniques [37, 13]. Talagrand in [65] discusses strengths and weaknesses of the martingale method, observing that "while in principle the martingale method has a wider range of applications, in many situations the [isoperimetric] inequalities [are] more powerful." Bearing out his first point we use martingales [37] to derive a general strong mixing condition for concentration, applying it to weakly contracting Markov chains. Following up, we extend the technique to hidden Markov [33] and Markov tree [34] measures.

Although a detailed survey of measure concentration literature is not our intent here, we remark that many of the results mentioned above may be described as working to extend inequalities of type (1.1) to wider classes of measures and metrics by imposing different strong mixing and Lipschitz continuity conditions. Already in [45], Marton gives a (rather stringent) mixing condition sufficient for concentration. Later, Marton [47, 48] and Samson [60] prove concentration for general classes of processes in terms of various mixing coefficients; Samson applies this to Markov chains and ϕ-mixing processes while Marton's application concerns lattice random fields.

1.2 Main results

The main result of this thesis is a concentration of measure inequality for arbitrarily dependent random variables. Postponing the technical details until the coming chapters, the inequality states that for any (nonproduct) measure $\mathbf{P}$ on Ω^n and any function $f : \Omega^n \to \mathbb{R}$, we have

$$\mathbf{P}\{|f - \mathbf{E}f| > t\} \quad \leq \quad 2\exp\left(-\frac{t^2}{2\left\|f\right\|_{\mathrm{Lip},w}^2 \left\|\Delta_n w\right\|_2^2}\right) \tag{1.4}$$

where Δ_n is the "η-mixing" matrix defined in Chapter 3.2 and $\left\|f\right\|_{\mathrm{Lip},w}$ is the Lipschitz constant of f with respect to the weighted Hamming metric d_w:

$$d_w(x,y) \quad = \quad \sum_{i=1}^{n} w_i \mathbb{1}_{\{x_i \neq y_i\}} \tag{1.5}$$

for some vector $w \in \mathbb{R}_+^n$.

It must be pointed out that (1.4) is not as novel as the author had hoped when proving it. A thorough literature search as well as discussions with experts suggest that Marton's [46, Theorem 2] should be considered the first result of comparable generality. Almost contemporaneously with our proving (1.4), Chazottes et al. published a very similar result [13]. Our main contribution is the proof technique – we offer the first (to our knowledge) non-coupling proof of (1.4), via Theorem 3.1.5, which may be of independent interest, as well as some novel applications.

Since $\|\Delta_n w\|_2 \leq \|\Delta_n\|_2 \|w\|_2$, the utility of (1.4) will depend on our ability to control the quantities $\|f\|_{\mathrm{Lip},w}$, $\|w\|_2$, and $\|\Delta_n\|_2$. In typical applications, we will have $w_i \equiv n^{-1}$ (corresponding to the usual normalized Hamming metric) and $\|f\|_{\mathrm{Lip},w} \leq 1$ (though it would be interesting to find a natural application that exploits the generality of weighted Hamming metrics[3]). The applications we have considered admit a somewhat cruder version of (1.4), via the bound

$$\|\Delta_n w\|_2^2 \quad \leq \quad n \max_{1 \leq i \leq n} (\Delta_n w)_i^2. \tag{1.6}$$

Having fixed the metric at normalized Hamming and the Lipschitz constant of f at 1, our bound on the deviation probability becomes

$$\mathbf{P}\{|f - \mathbf{E}f| > t\} \quad \leq \quad 2 \exp\left(-\frac{nt^2}{2 \|\Delta_n\|_\infty^2} \right). \tag{1.7}$$

Again, one hopes to find scenarios where the full sharpness of (1.4) would be used.

Once we establish (1.7), we will derive concentration bounds for various measures of interest by bounding $\|\Delta_n\|_\infty$. The construction of Δ_n in Chapter 3.2 implies $1 \leq \|\Delta_n\|_\infty \leq n$ but as (1.7) makes clear, this trivial bound is useless. We would need $\|\Delta_n\|_\infty = O(\sqrt{n})$ for a meaningful bound; for a number of interesting measures, we will actually have $\|\Delta_n\|_\infty = O(1)$.

Though a full statement of our concentration results for various measures will have to be deferred to Chapter 4, we will attempt a brief summary. For Markov chains with contraction (Doeblin) coefficient $\theta < 1$, we have

$$\|\Delta_n\|_\infty \quad \leq \quad 1/(1-\theta), \tag{1.8}$$

which recovers Marton's result (1.3) up to small constants and vanishing terms. Our bound is actually a strict generalization of Marton's since it is sensitive to the contraction coefficients θ_i at different time steps $1 \leq i < n$ and even allows then to achieve unity.

Expanding our class of measures to include the hidden Markov chains, we observe a surprising phenomenon: $\|\Delta_n\|_\infty$ can be controlled by the contraction coefficients of the underlying Markov chain. Thus a hidden Markov process is "at least as concentrated" as its underlying Markov process; however, this property fails for general hidden/observed process pairs.

Our concentration result for Markov trees requires a bit of overhead to state so we defer it to Chapter 4. It is stated in terms of the tree width and the contraction coefficients at each node and reduces to our Markov chain bounds in the degenerate case of a single-path tree.

[3] Since product measures (as we shall see) satisfy $\|\Delta_n\|_2 = 1$, (1.4) recovers McDiarmid's inequality (1.2) with a slightly worse constant. In its full generality, the latter reads

$$\mathbf{P}\{|f - \mathbf{E}f| > t\} \leq 2 \exp(-2t^2 / \|w\|_2^2)$$

for any f with $\|f\|_{\mathrm{Lip},w} \leq 1$.

The concentration results for various processes in turn yield new tools in statistics and machine learning. One of these is a law of large numbers for strongly mixing processes, which has already found applications in statistics [11] . We also show how (1.7) can be used to generalize the classical PAC results to non-independent samples.

1.3 Applications of concentration

The word *applications* means different things to different people. Computer scientists are typically concerned with algorithms and their performance analysis while statisticians often need a handle on the asymptotics of estimators and samplers, and mathematicians are always looking for new theorem-proving techniques – in short, each specialist wants to know what a given tool will contribute to his field.

The remarkable thing about concentration of measure is that its uses span the wide gamut from something as practical as decoding neural signals [10, 11] to esoteric topics such as analyzing convex bodies in Banach spaces [4].

Whole books and monographs have been devoted to the different applications of concentration. Toward the applied end of the spectrum, there is the forthcoming book of Dubhashi and Panconesi on analysis of algorithms [20], while the more theoretical consequences of concentration (in particular, in Banach spaces and groups) are described in Schechman's paper [61] and his book with Milman [54].

A few celebrated applications of concentration include

- Milman's proof [53] of Dvoretzky's theorem [21] on sections of convex bodies
- a widely cited lemma of Johnson and Lindenstrauss [29] concerning low-distortion dimensionality reduction in $\mathbb{R}^n$ by random projections
- Shamir and Spencer's work [62] on the concentration of a random graph's chromatic number
- statistics and empirical processes [49] and machine learning [5].

Rather than reproduce these results here, we will focus on the areas in which our methods are most readily applicable – viz., the last item. This will be covered in some detail in Chapter 4.3.

1.4 Thesis overview

This thesis is organized as follows. In Chapter 2 we define the notational conventions used throughout the thesis (Chapter 2.1.2), prove some preliminary results concerning the total variation norm (Chapter 2.2), review the main existing techniques for proving concentration (Chapter 2.3), and prove our main inequality (Chapter 2.3) as well as the concentration bound following from it (Chapter 3.3).

Chapter 4 will deal with applications of the main concentration result. First, we proceed to apply the general inequality to various processes: Markov, hidden Markov, and Markov tree. In the next application, we obtain a law of large numbers for strongly mixing processes, which in particular yields an analysis of an inhomogeneous Markov Chain Monte Carlo algorithm; this is joint work with Anthony Brockwell. Finally, we exhibit some applications of our techniques to empirical process theory and machine learning.

In Chapter 5, we collect some miscellaneous results, such as proving that $\|\cdot\|_\Phi$ and $\|\cdot\|_\Psi$ are valid norms (Chapter 5.2), giving examples where concentration fails if $\mathbf{P}$ is not mixing or f not Lipschitz (Chapter 5.5), discuss the measure-theoretic nuances of conditioning on measure-zero events (Chapter 5.4), extend our results to countable and continuous spaces (Chapter 5.1), as well as the ℓ_p metrics (Chapter 5.3). We also construct some illustrative examples of measures with specific mixing coefficients (Chapter 5.7).

Finally, Chapter 6 discusses some open problems, conjectures, and future research directions.

Chapter 2

Methods

2.1 Preliminaries

2.1.1 Measure theory

When dealing with abstract measures, one must typically take care to ensure that all the objects are in fact measurable, the conditional distributions well-defined, and so forth. On the other hand, our main contribution is not measure-theoretic in nature; indeed, the $\Omega = \{0, 1\}$ case captures most of the interesting phenomena. Furthermore, since our natural metric is a discrete one (Hamming), it seems reasonable to restrict most of our attention on the case of countable Ω. In fact, we will take Ω to be finite until Chapter 5.1, where we extend our results to countable and continuous Ω without too much effort.

Thus, until further notice, Ω is a finite set and questions of measurability need not concern us (no finiteness assumptions are made when we use the generic symbol $\mathcal{X}$).

2.1.2 Notational conventions

Random variables are capitalized (X), specified sequences (vectors) are written in lowercase ($x \in \Omega^n$), the shorthand $X_i^j = (X_i, \ldots, X_j)$ is used for all sequences, and brackets denote sequence concatenation: $[x_i^j \, x_{j+1}^k] = x_i^k$. Often, for readability, we abbreviate $[y \, w]$ as yw.

We use the indicator variable $\mathbb{1}_{\{\cdot\}}$ to assign 0-1 truth values to the predicate in $\{\cdot\}$. The sign function is defined by $\mathrm{sgn}(z) = \mathbb{1}_{\{z \geq 0\}} - \mathbb{1}_{\{z<0\}}$. The ramp function is defined by $(z)_+ = z\mathbb{1}_{\{z>0\}}$. We denote the set $(0, \infty)$ by $\mathbb{R}_+$.

The probability $\mathbf{P}$ and expectation $\mathbf{E}$ operators are defined with respect the measure space specified in context. To any probability space $(\Omega^n, \mathcal{F}, \mathbf{P})$, we associate the canonical random process $X = X_1^n$, $X_i \in \Omega$, satisfying

$$\mathbf{P}\{X \in A\} \quad = \quad \mathbf{P}(A)$$

for any $A \in \mathcal{F}$. If we wish to make the measure explicit, we will write $\mu(A)$ for probabilities $\mathbf{P}_\mu(A)$ and μf for expectations $\mathbf{E}_\mu f$.

If $(\mathcal{X}, \mathcal{F}, \mu)$ is a (positive) measure space, we write $L_p(\mathcal{X}, \mu)$ for the usual space of μ-measurable functions $f : \mathcal{X} \to \mathbb{R}$, whose L_p norm

$$\|f\|_{L_p(\mathcal{X},\mu)} \quad = \quad \left(\int_{\mathcal{X}} |f|^p d\mu \right)^{1/p}$$

is finite. We will write $\|\cdot\|_{L_p(\mathcal{X},\mu)}$ as $\|\cdot\|_{L_p(\mu)}$ or just $\|\cdot\|_{L_p}$ if there is no ambiguity; when μ is the counting measure on a discrete space, we write this as $\|\cdot\|_p$.

Likewise, the L_∞ norm, $\|f\|_{L_\infty} = \operatorname{ess\,sup} |f|$ is defined via the essential supremum:

$$\operatorname*{ess\,sup}_{x\in\mathcal{X}} f(x) \quad = \quad \inf\{a \in [-\infty,\infty] : \mu\{f(x) > a\} = 0\}.$$

A *weighted Hamming* metric on a product space Ω^n is a weighted sum of the discrete metrics on Ω:

$$d_w(x,y) \quad = \quad \sum_{i=1}^n w_i \mathbb{1}_{\{x_i \neq y_i\}}$$

for $x, y \in \Omega^n$ and some fixed $w \in \mathbb{R}_+$. We will write $\bar{d}$ for the (frequent) special case $w_i \equiv n^{-1}$.

2.2 Total variation norm: properties and characterizations

Remark 2.2.1. The results proved in this section were discovered independently by the author. It came as no surprise that they are not new; even where concrete references are not available, these inequalities are well-known in probability-theoretic folklore. We offer simple computational proofs, in contrast to the typical coupling arguments used to obtain such results (see, for example, David Pollard's book-in-progress *Asymptopia*, or his online notes[1]). It is hoped that the technique used here might some day facilitate a proof where the coupling method is less forthcoming. $\diamond$

If μ is a positive Borel measure on $(\mathcal{X},\mathcal{F})$ and τ is a signed measure on $(\mathcal{X},\mathcal{F})$, we define the *total variation* of τ by

$$2\|\tau\|_{\mathrm{TV}} \quad = \quad \sup \sum_{i=1}^\infty |\tau(E_i)|\,, \tag{2.1}$$

where the supremum is over all the countable partitions E_i of $\mathcal{X}$ (this quantity is necessarily finite, by Theorem 6.4 of [59]).[2] It is a consequence of the Lebesgue-Radon-Nikodým theorem ([59], Theorem 6.12) that if $d\tau = h d\mu$, we have

$$2\|\tau\|_{\mathrm{TV}} \quad = \quad \int_{\mathcal{X}} |h| d\mu.$$

Additionally, if τ is *balanced*, meaning that $\tau(\mathcal{X}) = 0$, we have

$$\|\tau\|_{\mathrm{TV}} \quad = \quad \int_{\mathcal{X}} (h)_+ \, d\mu; \tag{2.2}$$

this follows via the Hahn decomposition ([59], Theorem 6.14).

If p and q are two probability measures on a (finite) set $\mathcal{X}$, there are many different notions of "distance" between p and q; we refer the reader to the excellent survey [24], which summarizes

[1] `http://www.stat.yale.edu/~pollard/607.spring05/handouts/Totalvar.pdf`

[2] Note the factor of 2 in (2.1), which typically does not appear in analysis texts but is standard in probability theory, when τ is the difference of two probability measures.

the relationships among the various definitions. A central role is occupied by the *total variation* distance; it will also be of key importance in this work.

The discussion above implies the identities

$$\|p-q\|_{\mathrm{TV}} \;=\; \tfrac{1}{2}\|p-q\|_1 \;=\; \big\|(p-q)_+\big\|_1\,, \tag{2.3}$$

which we will invoke without further justification throughout this work. Another widely used relation is

$$\|p-q\|_{\mathrm{TV}} \;=\; \sup_{\text{Borel } A\subset\mathcal{X}} |p(A)-q(A)|\,; \tag{2.4}$$

its proof is elementary and is traditionally left to the reader (see [9], p. 126 for a proof).

Given the measures p and q on $\mathcal{X}$, define

$$\mathscr{M}(p,q) \;=\; \left\{u\in[0,1]^{\mathcal{X}\times\mathcal{X}} : \int_{\mathcal{X}} u(dx,\cdot)=q(\cdot), \int_{\mathcal{X}} u(\cdot,dy)=p(\cdot)\right\} \tag{2.5}$$

to be the set of all *couplings* of p and q – i.e., all joint distributions on $\mathcal{X}\times\mathcal{X}$ whose marginals are p and q, respectively. It is a basic fact (see, for example, [42, Theorem 5.2]) that

$$\|p-q\|_{\mathrm{TV}} \;=\; \min_{u\in\mathscr{M}(p,q)} \int_{\mathcal{X}\times\mathcal{X}} \mathbb{1}_{\{x\neq y\}}\,du(x,y); \tag{2.6}$$

the latter has the interpretation of $\mathbf{P}\{X\neq Y\}$ minimized over all joint measures u on $\mathcal{X}\times\mathcal{X}$ with $X\sim p=\sum_y u(\cdot,y)$ and $Y\sim q=\sum_x u(x,\cdot)$. We will give a (possibly new) elementary proof of this fact, in the case of finite $\mathcal{X}$.

Let us define the (unnormalized) measure $p\wedge q$ as the pointwise minimum of p and q:

$$(p\wedge q)(x) \;=\; \min\{p(x),q(x)\}\,.$$

The quantity $\|p\wedge q\|_1$ is called the *affinity* between p and q and satisfies

Lemma 2.2.2. *If p,p' are probability measures on $\mathcal{X}$ and $\tilde{q}=p\wedge p'$, then*

$$\big\|p-p'\big\|_{\mathrm{TV}} \;=\; 1-\|\tilde{q}\|_1\,. \tag{2.7}$$

Proof. Define the function

$$F(u,v) \;=\; 1-\sum_{x\in\mathcal{X}}\min\{u_x,v_x\}-\tfrac{1}{2}\sum_{x\in\mathcal{X}}|u_x-v_x|$$

over the convex polytope $U\subset\mathbb{R}^{\mathcal{X}}\times\mathbb{R}^{\mathcal{X}}$,

$$U=\left\{(u,v): u_x,v_x\geq 0, \sum u_x=\sum v_x=1\right\};$$

note that proving (2.7) is equivalent to showing that $F\equiv 0$ on U.

For any $\boldsymbol{\sigma}\in\{-1,+1\}^{\mathcal{X}}$, let

$$U_{\boldsymbol{\sigma}}=\{(u,v)\in U : \mathrm{sgn}(u_x-v_x)=\sigma_x\};$$

note that $U_{\boldsymbol{\sigma}}$ is a convex polytope and that $U = \bigcup_{\boldsymbol{\sigma}\in\{-1,+1\}^{\mathcal{X}}} U_{\boldsymbol{\sigma}}$.[3] Fix some $\boldsymbol{\sigma} \in \{-1,+1\}^{\mathcal{X}}$. Observing that for $u, v \in U_{\boldsymbol{\sigma}}$,

$$\min\{u_x, v_x\} \;=\; u_x \mathbb{1}_{\{\sigma_x<0\}} + v_x \mathbb{1}_{\{\sigma_x>0\}}$$

and

$$|u_x - v_x| \;=\; \sigma_x(u_x - v_x),$$

we define the function

$$F_{\boldsymbol{\sigma}}(u,v) \;=\; 1 - \sum_{x\in\mathcal{X}} (u_x \mathbb{1}_{\{\sigma_x<0\}} + v_x \mathbb{1}_{\{\sigma_x>0\}}) - \tfrac{1}{2}\sum_{x\in\mathcal{X}} \sigma_x(u_x - v_x)$$

over $U_{\boldsymbol{\sigma}}$; note that $F_{\boldsymbol{\sigma}}$ agrees with F on this domain.

Observe that $F_{\boldsymbol{\sigma}}$ is affine in its arguments (u, v) and recall that an affine function is determined by its values on the extreme points of a convex domain. Thus to verify that $F_{\boldsymbol{\sigma}} \equiv 0$ on $U_{\boldsymbol{\sigma}}$, we need only check the value of $F_{\boldsymbol{\sigma}}$ on the extreme points of $U_{\boldsymbol{\sigma}}$. The extreme points of $U_{\boldsymbol{\sigma}}$ are pairs (u, v) such that, for some $x', x'' \in \mathcal{X}$, $u = \delta(x')$ and $v = \delta(x'')$, where $\delta(z) \in \mathbb{R}^{\mathcal{X}}$ is given by $[\delta(z)]_x = \mathbb{1}_{\{x=z\}}$.

Let $(\hat{u}, \hat{v})$ be an extreme point of $U_{\boldsymbol{\sigma}}$. The case $\hat{u} = \hat{v}$ trivial, so assume $\hat{u} \neq \hat{v}$. In this case,

$$\min\{\hat{u}_x, \hat{v}_x\} \equiv 0$$

and

$$\sum_{x\in\mathcal{X}} |u_x - v_x| = 2.$$

This shows that $F_{\boldsymbol{\sigma}}$ vanishes on $U_{\boldsymbol{\sigma}}$ and proves the claim. □

An easy consequence of this lemma is the following minorization property of the total variation distance:

Lemma 2.2.3. *Let p, p', q be probability distributions on $\mathcal{X}$ satisfying*

$$p(x), p'(x) \;\geq\; \varepsilon q(x), \qquad x \in \mathcal{X} \tag{2.8}$$

for some $\varepsilon > 0$. Then

$$\|p - p'\|_{\mathrm{TV}} \;\leq\; 1 - \varepsilon. \tag{2.9}$$

Proof. Condition (2.8) implies

$$(p \wedge p')(x) \;\geq\; \varepsilon q(x), \qquad x \in \mathcal{X}.$$

Summing over $x \in \mathcal{X}$, we have $\varepsilon \leq \|p \wedge p'\|_1$, which implies (2.9) via Lemma 2.2.2. □

Another consequence of Lemma 2.2.2 is a very simple proof of (2.6):

[3] Note that the constraint $\sum_{x\in\mathcal{X}} u_x = \sum_{x\in\mathcal{X}} v_x = 1$ forces $U_{\boldsymbol{\sigma}} = \{(u, v) \in U : u_x = v_x\}$ when $\boldsymbol{\sigma} \equiv +1$ and $U_{\boldsymbol{\sigma}} = \emptyset$ when $\boldsymbol{\sigma} \equiv -1$. Both of these cases are trivial.

Lemma 2.2.4. *If p and q are probability measures on $\mathcal{X}$, then*

$$\|p-q\|_{\mathrm{TV}} \;=\; \min_{u\in\mathscr{M}(p,q)} \sum_{x,y\in\mathcal{X}} u(x,y)\mathbb{1}_{\{x\neq y\}}.$$

Proof. For a given $u\in\mathscr{M}(p,q)$, the r.h.s. becomes

$$\begin{aligned}\sum_{x,y\in\mathcal{X}} u(x,y)\mathbb{1}_{\{x\neq y\}} &= \sum_x\sum_{y\neq x} u(x,y)\\ &= \sum_x [p(x)-u(x,x)]\\ &= 1-\sum_x u(x,x).\end{aligned}$$

The constraint $u\in\mathscr{M}(p,q)$ implies

$$0\le u(x,x)\le \min\{p(x),q(x)\}$$

and minimizing the r.h.s. means taking $u=p\wedge q$. □

Another useful property of total variation is the following "tensorizing" inequality; the proof below first appeared in [34]:

Lemma 2.2.5. *Consider two finite sets $\mathcal{X},\mathcal{Y}$, with probability measures p,p' on $\mathcal{X}$ and q,q' on $\mathcal{Y}$. Then*

$$\left\|p\otimes q-p'\otimes q'\right\|_{\mathrm{TV}} \;\le\; \left\|p-p'\right\|_{\mathrm{TV}}+\left\|q-q'\right\|_{\mathrm{TV}}-\left\|p-p'\right\|_{\mathrm{TV}}\left\|q-q'\right\|_{\mathrm{TV}}. \tag{2.10}$$

Remark 2.2.6. Note that $p\otimes q$ is a 2-tensor in $\mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$ and a probability measure on $\mathcal{X}\times\mathcal{Y}$.

Proof. Fix q,q' and define the function

$$F(u,v)=\sum_{x\in\mathcal{X}}|u_x-v_x|+\left\|q-q'\right\|_{\mathrm{TV}}\left(2-\sum_{x\in\mathcal{X}}|u_x-v_x|\right)-\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\left|u_xq_y-v_xq'_y\right|$$

over the convex polytope $U\subset\mathbb{R}^{\mathcal{X}}\times\mathbb{R}^{\mathcal{X}}$,

$$U=\left\{(u,v): u_x,v_x\ge 0, \sum u_x=\sum v_x=1\right\};$$

note that proving the claim is equivalent to showing that $F\ge 0$ on U.

For any $\boldsymbol{\sigma}\in\{-1,+1\}^{\mathcal{X}}$, let

$$U_{\boldsymbol{\sigma}}=\{(u,v)\in U:\mathrm{sgn}(u_x-v_x)=\sigma_x\};$$

note that $U_{\boldsymbol{\sigma}}$ is a convex polytope and that $U=\bigcup_{\boldsymbol{\sigma}\in\{-1,+1\}^{\mathcal{X}}}U_{\boldsymbol{\sigma}}$.

Pick an arbitrary $\boldsymbol{\tau}\in\{-1,+1\}^{\mathcal{X}\times\mathcal{Y}}$ and define

$$F_{\boldsymbol{\sigma}}(u,v)=\sum_x\sigma_x(u_x-v_x)+\left\|q-q'\right\|_{\mathrm{TV}}\left(2-\sum_x\sigma_x(u_x-v_x)\right)-\sum_{x,y}\tau_{xy}(u_xq_y-v_xq'_y)$$

over $U_{\boldsymbol{\sigma}}$. Since $\sigma_x(u_x - v_x) = |u_x - v_x|$ and $\boldsymbol{\tau}$ can be chosen (for any given u, v, q, q') so that $\tau_{xy}(u_x q_y - v_x q'_y) = \left|u_x q_y - v_x q'_y\right|$, the claim that $F \geq 0$ on U will follow if we can show that $F_{\boldsymbol{\sigma}} \geq 0$ on $U_{\boldsymbol{\sigma}}$.

Observe that $F_{\boldsymbol{\sigma}}$ is affine in its arguments (u, v) and recall that an affine function achieves its extreme values on the extreme points of a convex domain. Thus to verify that $F_{\boldsymbol{\sigma}} \geq 0$ on $U_{\boldsymbol{\sigma}}$, we need only check the value of $F_{\boldsymbol{\sigma}}$ on the extreme points of $U_{\boldsymbol{\sigma}}$. The extreme points of $U_{\boldsymbol{\sigma}}$ are pairs (u, v) such that, for some $x', x'' \in \mathcal{X}$, $u = \delta(x')$ and $v = \delta(x'')$, where $\delta(x_0) \in \mathbb{R}^{\mathcal{X}}$ is given by $[\delta(x_0)]_x = \mathbb{1}_{\{x=x_0\}}$. Let $(\hat{u}, \hat{v})$ be an extreme point of $U_{\boldsymbol{\sigma}}$. The case $\hat{u} = \hat{v}$ is trivial, so assume $\hat{u} \neq \hat{v}$. In this case, $\sum_{x\in\mathcal{X}} \sigma_x(\hat{u}_x - \hat{v}_x) = 2$ and

$$\begin{aligned} \Big|\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} \tau_{xy}(\hat{u}_x q_y - \hat{v}_x q'_y)\Big| &\leq \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} \left|\hat{u}_x q_y - \hat{v}_x q'_y\right| \\ &\leq 2. \end{aligned}$$

This shows that $F_{\boldsymbol{\sigma}} \geq 0$ on $U_{\boldsymbol{\sigma}}$ and completes the proof. □

Remark 2.2.7. Lemma 2.2.5 has a simple coupling proof and appears to be folklore knowledge among probability theorists (we were not able to locate it in the literature). Our proof technique, consisting of converting the inequality into an affine function taking nonnegative values over a convex polytope, appears to be novel. Aside from being an alternative to coupling, this technique can lead to natural insights and generalizations. For instance, our proof of Lemma 2.2.5 admits an immediate generalization to the case of Markov kernels.

Let p_0 be a probability measure on $\mathcal{X}$, and $p_1(\cdot\,|\,x)$, $x \in \mathcal{X}$, a (conditional probability) kernel from $\mathcal{X}$ to $\mathcal{Y}$, and write $\mu = p_0 \otimes p_1$ for the measure on $\mathcal{X} \times \mathcal{Y}$ defined by

$$\mu(x, y) \;=\; p_0(x)p_1(y\,|\,x), \qquad x, y \in \mathcal{X} \times \mathcal{Y}.$$

Similarly, let q_0 be a measure on $\mathcal{X}$ and q_1 a kernel from $\mathcal{X}$ to $\mathcal{Y}$; define $\nu = q_0 \otimes q_1$.

Then a straightforward modification of the proof of Lemma 2.2.5 yields

$$\|\mu - \nu\|_{\mathrm{TV}} \;\leq\; d_0 + d_1 - d_0 d_1, \tag{2.11}$$

where $d_0 = \|p_0 - q_0\|_{\mathrm{TV}}$ and

$$d_1 \;=\; \max_{x\in\mathcal{X}} \|p_1(\cdot\,|\,x) - q_1(\cdot\,|\,x)\|_{\mathrm{TV}}\,.$$

There may well be a simple coupling proof of (2.11), though it seems to us that a bit of work would be required. Our point is that our "affine-function" technique *suggested* the generalization and offered a proof, with minimal effort. ◇

2.3 Survey of concentration techniques

Given the excellent survey papers and monographs dealing with concentration of measure (in particular, [40], [61], and [43]), we will confine ourselves to briefly mentioning the main techniques and refer the reader to the cited works for details and proofs.

2.3.1 Lévy families and concentration in metric spaces

A natural language for discussing measure concentration in general metric spaces is that of Lévy families. This definition is taken, with minor variations, from Chapter 6 of [54]. Let $(\mathcal{X}, \rho, \mu)$ be a metric probability space (that is, a Borel probability space whose topology is induced by the metric ρ). Whenever we write $A \subset \mathcal{X}$, it is implicit that A is a Borel subset of $\mathcal{X}$. For $t > 0$, define the *t-enlargement* of $A \subset \mathcal{X}$:

$$A_t = \{x \in \mathcal{X} : \rho(x, A) < t\}.$$

The *concentration function* $\alpha(\cdot) = \alpha_{\mathcal{X},\rho,\mu}(\cdot)$ is defined by:

$$\alpha(t) \quad = \quad 1 - \inf\{\mu(A_t) : A \subset \mathcal{X},\ \mu(A) \geq \tfrac{1}{2}\}.$$

Let $(\mathcal{X}_n, \rho_n, \mu_n)_{n\geq 1}$ be a family of metric probability spaces with $\mathrm{diam}_{\rho_n}(\mathcal{X}_n) < \infty$, where

$$\mathrm{diam}_{\rho_n}(\mathcal{X}_n) \quad = \quad \sup_{x,y\in\mathcal{X}_n} \rho_n(x, y). \tag{2.12}$$

This family is called a *normal Lévy* family if there are constants $c_1, c_2 > 0$ such that

$$\alpha_{\mathcal{X}_n,\rho_n,\mu_n}(t) \quad \leq \quad c_1 \exp(-c_2 n t^2)$$

for each $t > 0$ and $n \geq 1$.

The condition of being a normal Lévy family implies strong concentration of a Lipschitz $f : \mathcal{X}_n \to \mathbb{R}$ about its median (and mean); this connection is explored in-depth in [40]. In particular, if $(\mathcal{X}, \rho, \mu)$ is a metric probability space and $f : \mathcal{X} \to \mathbb{R}$ is measurable, define its *modulus of continuity* by

$$\omega_f(\delta) \quad = \quad \sup\{|f(x) - f(y)| : \rho(x, y) < \delta\}. \tag{2.13}$$

A number $M_f \in \mathbb{R}$ is called a *μ-median* of f if

$$\mu\{f \leq M_f\} \geq \tfrac{1}{2} \qquad \text{and} \qquad \mu\{f \geq M_f\} \geq \tfrac{1}{2}$$

(a median need not be unique). These definitions immediately imply the deviation inequality [40](1.9)

$$\mu\{|f - M_f| > \omega_f(\delta)\} \quad \leq \quad 2\alpha_{\mathcal{X},\rho,\mu}(\delta),$$

which in turn yields [40](1.13)

$$\mu\{|f - M_f| > t\} \quad \leq \quad 2\alpha_{\mathcal{X},\rho,\mu}(t/\|f\|_{\mathrm{Lip}}), \tag{2.14}$$

where the Lipschitz constant $\|f\|_{\mathrm{Lip}}$ is the smallest constant C for which $\omega_f(\delta) \leq C\delta$, for all $\delta > 0$. In particular, (2.14) lets us take $\|f\|_{\mathrm{Lip}} = 1$ without loss of generality, which we shall do below. The following result lets us convert concentration about a median to concentration about any constant:

Theorem (Thm. 1.8 in [40]). *Let f be a measurable function on a probability space $(\mathcal{X}, \mathcal{A}, \mathbf{P})$. Assume that for some $a \in \mathbb{R}$ and a non-negative function α on $\mathbb{R}_+$ such that $\lim_{r\to\infty} \alpha(r) = 0$,*

$$\mathbf{P}\{|f - a| \geq r\} \quad \leq \quad \alpha(r)$$

for all $r > 0$. Then

$$\mathbf{P}\{|f - M_f| \geq r + r_0\} \quad \leq \quad \alpha(r), \quad r > 0,$$

where M_f is a $\mathbf{P}$-median of f and where $r_0 > 0$ is such that $\alpha(r_0) < \frac{1}{2}$. If moreover $\bar{\alpha} = \int_0^\infty \alpha(r)dr < \infty$ then f is integrable, $|a - \mathbf{E}f| \leq \bar{\alpha}$, and for every $r > 0$,

$$\mathbf{P}\{|f - \mathbf{E}f| \geq r + \bar{\alpha}\} \quad \leq \quad \alpha(r).$$

Thus, for a normal Lévy family, deviation inequalities for the mean and median are equivalent up to the constants c_1, c_2. Theorem 1.7 in [40] is a converse to (2.14), showing that if Lipschitz functions on a metric probability space $(\mathcal{X}, \rho, \mu)$ are tightly concentrated about their means, this implies a rapid decay of $\alpha_{\mathcal{X},\rho,\mu}(\cdot)$. We remark that for typical applications it is usually most convenient to bound the deviation of a variable about its mean.

2.3.2 Martingales

Let $(\mathcal{X}, \mathcal{F}, \mathbf{P})$ be a probability space and consider some filtration

$$\{\emptyset, \mathcal{X}\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \ldots \subset \mathcal{F}_n = \mathcal{F}. \tag{2.15}$$

For $i = 1, \ldots, n$ and $f \in L_1(\mathcal{X}, \mathbf{P})$, define the martingale[4] difference

$$V_i \quad = \quad \mathbf{E}[f \,|\, \mathcal{F}_i] - \mathbf{E}[f \,|\, \mathcal{F}_{i-1}]. \tag{2.16}$$

It is a classical result,[5] going back to Azuma [3] and Hoeffding [27] in the 1960's, that

$$\mathbf{P}\{|f - \mathbf{E}f| > t\} \quad \leq \quad 2\exp(-t^2/2D^2) \tag{2.17}$$

where $D^2 \geq \sum_{i=1}^n \|V_i\|_\infty^2$ (the meaning of $\|V_i\|_\infty$ will be made explicit later). Thus, the problem of obtaining deviation inequalities of type (1.1) is reduced to the problem of bounding D^2. This will be the approach we take in this thesis, where our ability to control D^2 will depend on the continuity properties of f and the mixing properties of the measure $\mathbf{P}$.

The most natural application of Azuma's inequality is to a product measure $\mathbf{P}$ on $\mathcal{X} = \Omega^n$ and $f : \Omega^n \to \mathbb{R}$ with $\|f\|_{\mathrm{Lip},w} \leq 1$. In this case, it is straightforward to verify that $\|V_i\|_\infty \leq w_i$, and so[6]

$$\mathbf{P}\{|f - \mathbf{E}f| > t\} \quad \leq \quad 2\exp(-2t^2/\|w\|_2^2);$$

this is McDiarmid's inequality [51], which has found particularly fruitful applications in computer science and combinatorics. Kim and Vu [31] have recently used martingales to obtain a concentration result for a class of non-Lipschitz functions, again with combinatorial applications in mind.

Besides the present work, we are only aware of one systematic application of the martingale method to nonproduct measures, namely that of Chazottes et al. [13]. The authors define a coupling matrix D^σ, analogous to our Δ_n, and use Chernoff exponential bounding together with Markov's inequality to obtain an inequality of a similar flavor to (1.4), applying it to random fields. The main

[4] The sequence $\{\mathbf{E}[f \,|\, \mathcal{F}_i]\}_{i=0}^n$ is a martingale with respect to $\{\mathcal{F}_i\}_{i=0}^n$.

[5] See [40] for a modern presentation and a short proof of (2.17).

[6] The improvement by a factor of 4 is obtained by observing that in fact $\sup V_i - \inf V_i \leq w_i$.

result of Chazottes et al. is essentially identical (modulo small constants) to our Theorem 3.3.4; their bound was obtained contemporaneously with ours via a totally different method (coupling).

We cannot resist mentioning a rather clever application of the martingale method. Let (Ω, ρ) be any finite metric space. For $a \in \mathbb{R}_+^n$, define an a-partition sequence of Ω to be a sequence

$$\{\Omega\} = \mathcal{A}_0, \mathcal{A}_1, \ldots, \mathcal{A}_n = \{\{x\}\}_{x\in\Omega}$$

of partitions of Ω such that $\mathcal{A}_{i+1}$ refines $\mathcal{A}_i$ and whenever $A \in \mathcal{A}_{k-1}, B, C \subset A$ and $B, C \in \mathcal{A}_k$, there is a bijection $h : B \to C$ such that $\rho(x, h(x)) \le a_k$ for all $x \in B$. The *length* $\ell = \|a\|_2$ of (Ω, ρ) is defined to be the infimum over all a-partition sequences. Then we have (see [54], [61] or [40])

$$\alpha_{\Omega,\rho,\mu}(r) \quad \le \quad \exp(-r^2/8\ell^2)$$

where μ is the normalized counting measure on Ω. In the case of the symmetric group of permutations $\mathcal{S}_n$ with the (normalized Hamming) metric

$$\rho(\sigma, \pi) = n^{-1} \sum_{i=1}^{n} \mathbb{1}_{\{\sigma(i)\neq\pi(i)\}},$$

we may bound its length by $\ell \le \sqrt{2/n}$ and thus its concentration function by $\alpha(r) \le \exp(-nr^2/32)$, recovering Maurey's theorem [50].[7] More results of this sort have been obtained on general groups; see [54].

2.3.3 Isoperimetry

Classical isoperimetric results relate the measure of a set's boundary to its full measure. Following Ledoux [40], we endow a metric space $(\mathcal{X}, \rho)$ with a positive Borel measure μ, and define the *boundary measure* (Minkowski content) of a Borel $A \subset \mathcal{X}$ to be

$$\mu^+(A) \quad = \quad \liminf_{r\to 0} \mu(A_r \setminus A)$$

where A_r is the r-enlargement of A defined in §2.3.1.

The *isoperimetric function* $I_\mu : [0, \mu(A)] \to \mathbb{R}_+$ is defined, for each Borel $A \subset \mathcal{X}$ with $\mu(A) < \infty$, to be the maximal value that satisfies

$$\mu^+(A) \quad \ge \quad I_\mu(\mu(A)). \tag{2.18}$$

Any set B achieving equality in (2.18) minimizes the boundary measure $\mu^+(A)$ among all sets A with $\mu(A) = \mu(B)$ and is called an *extremal set*. One may obtain concentration from isoperimetry by imposing mild conditions on μ and assuming the existence of a strictly increasing differentiable function $v : \mathbb{R} \to [0, \mu(\mathcal{X})]$ for which $I_\mu \ge v' \circ v^{-1}$. Under these assumptions, we can bound the concentration function [40, Corollary 2.2]:

$$\alpha_{\mathcal{X},\rho,\mu}(r) \quad \le \quad 1 - v(v^{-1}(\tfrac{1}{2}) + r).$$

The isoperimetric function is notoriously difficult to compute in general, but admits a few benign special cases, among them the unit sphere $\mathbb{S}^n \subset \mathbb{R}^n$ endowed with the uniform probability measure σ^n and geodesic distance ρ. In this case, we have [40, Theorem 2.3]

$$\alpha_{\mathbb{S},\rho,\sigma^n}(r) \quad \le \quad \exp(-(n-1)r^2/2),$$

[7] The latter also has an elementary proof; see §5.8.

which is essentially Lévy's inequality.

A more modern approach, pioneered by Talagrand [65], dispenses with the isoperimetric function and works directly with enlargements. Endow any product probability space $(\Omega^n, \mathbf{P})$ with the weighted Hamming metric d_w, $w \in \mathbb{R}_+^n$ defined in (1.5) and define the *convex distance*

$$D_A(x) \quad = \quad \sup_{\|w\|_2 \le 1} d_w(x, A)$$

for Borel $A \subset \Omega^n$.

Then Talagrand's famous inequality [40, Theorem 4.6] reads

$$\mathbf{P}\{D_A \ge t\} \quad \le \quad \mathbf{P}(A)^{-1} \exp(-t^2/4). \tag{2.19}$$

Though at first glance not much different from McDiarmid's inequality, (2.19) is actually quite a bit more powerful, with numerous applications given in [65].

2.3.4 Logarithmic Sobolev and Poincaré inequalities

Let $(\mathcal{X}, \rho, \mu)$ be a metric probability space. In this case, $|\nabla f|$ may be defined as

$$|\nabla f(x)| \quad = \quad \limsup_{y \to x} \frac{|f(x) - f(y)|}{\rho(x, y)} \tag{2.20}$$

without assigning independent meaning to ∇f. Define the following three functionals mapping $f \in \mathbb{R}^{\mathcal{X}}$ to $\mathbb{R}_+$: *entropy*,

$$\mathrm{Ent}_\mu(f) \quad = \quad \int_{\mathcal{X}} f \log f d\mu - \int_{\mathcal{X}} f d\mu \log \int_{\mathcal{X}} f d\mu \tag{2.21}$$

variance,

$$\mathrm{Var}_\mu(f) \quad = \quad \int_{\mathcal{X}} f^2 d\mu - \left(\int_{\mathcal{X}} f d\mu \right)^2, \tag{2.22}$$

and *energy*,

$$\mathcal{E}_\mu(f) \quad = \quad \int_{\mathcal{X}} |\nabla f(x)|^2 \, d\mu. \tag{2.23}$$

(in each case we make the necessary assumptions for the quantities to be well-defined and finite). The measure μ is said to satisfy a logarithmic Sobolev inequality with constant C if

$$\mathrm{Ent}_\mu(f^2) \quad \le \quad 2C\mathcal{E}_\mu(f) \tag{2.24}$$

and a Poincaré (or spectral gap) inequality with constant C if

$$\mathrm{Var}_\mu(f) \quad \le \quad C\mathcal{E}_\mu(f); \tag{2.25}$$

in each case the inequality is asserted to hold for all f with $\|f\|_{\mathrm{Lip}} \le 1$ (with respect to ρ).

If (2.24) holds, we have [40, Theorem 5.3]

$$\alpha_{\mathcal{X},\rho,\mu}(r) \quad \le \quad \exp(-r^2/8C);$$

if (2.25) holds, we have [40, Corollary 3.2]

$$\alpha_{\mathcal{X},\rho,\mu}(r) \quad \le \quad \exp(-r/3\sqrt{C}).$$

These inequalities are proved in [61], [40] and [39], the latter an encyclopedic source on the subject. The reader is referred to [6] for recent results and literature surveys.

2.3.5 Transportation

The technique of using *transportation* (alternatively: *information*) inequalities to prove concentration was pioneered by Marton in her widely cited paper on contracting Markov chains [45]. Since the publication of Marton's paper, several authors proceeded to generalize the information-theoretic approach [16, 15, 17, 25]. These techniques are also at the heart of Samson's result [60], which we shall discuss in greater detail below. The material in this section is taken from [40]; a comprehensive treatment is given in [69].

For a metric space $(\mathcal{X}, \rho)$ and two Borel probability measures μ, ν on $\mathcal{X}$, define the *transportation cost distance* (also referred to as Wasserstein 1, Monge-Kantorovich, or earthmover distance [55]) between μ and ν:

$$T_\rho(\mu,\nu) \;=\; \inf_{\pi\in\mathscr{M}(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} \rho(x,y)d\pi(x,y), \tag{2.26}$$

where $\mathscr{M}(\mu,\nu)$ is defined in (2.5). Define also the *relative entropy* (or Kullback-Leibler divergence) of ν with respect to μ as

$$H(\nu\,|\,\mu) \;=\; \mathrm{Ent}_\mu\left(\frac{d\nu}{d\mu}\right) \;=\; \int_{\mathcal{X}} \log\frac{d\nu}{d\mu}d\nu \tag{2.27}$$

whenever $\nu \ll \mu$ with Radon-Nikodým derivative $\frac{d\nu}{d\mu}$.

The measure μ is said to satisfy a transportation inequality with constant C if

$$T_\rho(\mu,\nu) \;\le\; \sqrt{2CH(\nu\,|\,\mu)} \tag{2.28}$$

for every ν. This condition implies concentration for μ:

$$\alpha_{\mathcal{X},\rho,\mu}(r) \;\le\; \exp(-r^2/8C), \qquad r \ge 2\sqrt{2C\log 2}.$$

Note that computing T_ρ (for finite $\mathcal{X}$) amounts to solving a linear program. A consequence of our main inequality in Theorem 3.1.2 is a simple bound on T_ρ; see Chapter 5.6.

The transportation cost distance T_ρ can be vastly generalized by replacing ρ in (2.26) with a (possibly non-metric) $\tilde{c} : \mathcal{X}\times\mathcal{X} \to \mathbb{R}_+$ and numerous concentration results can be obtained via these methods; see the references in [40] and [69] for details.

2.3.6 Exchangeable pairs

A novel technique for proving concentration was presented in Sourav Chatterjee's 2005 PhD thesis [12]. It is based on Stein's method for exchangeable pairs, which is explained in Chatterjee's dissertation.

Let μ be a probability measure on Ω^n, and $X = (X_1 \dots X_n) \in \Omega^n$ be a random variable with law μ. For any $x \in \Omega^n$, let

$$\bar{x}^i = [x_1^{i-1}\, x_{i+1}^n]$$

denote the element of Ω^{n-1} obtained by omitting the i^{th} symbol in x. For each $1 \le i \le n$ and $x \in \Omega^n$, let $\mu_i(\cdot\,|\,\bar{x}^i)$ be the conditional law of X_i given $\bar{X}^i = \bar{x}^i$. One of the main results in Chatterjee's thesis is the following elegant inequality:

Theorem (Thm. 4.3 of [12]). *Suppose $A = (a_{ij}) \in \mathbb{R}_+^{n\times n}$ satisfies $a_{ii} = 0$ and*

$$\left\|\mu_i(\cdot \,|\, \bar{x}^i) - \mu_i(\cdot \,|\, \bar{y}^i)\right\|_{\mathrm{TV}} \;\leq\; \sum_{j=1}^{n} a_{ij} \mathbb{1}_{\{x_j \neq y_j\}},$$

for all $x, y \in \Omega^n$ and $1 \leq i \leq n$. Then, if $f : \Omega^n \to \mathbb{R}$ satisfies $\|f\|_{\mathrm{Lip},w} \leq 1$ for some $w \in \mathbb{R}_+^n$ and $\|A\|_2 < 1$, we have

$$\mathbf{P}\{|f - \mathbf{E}f| > t\} \;\leq\; 2\exp(-(1 - \|A\|_2)t^2 / \|w\|_2^2),$$

where $\|A\|_2$ is the ℓ_2 operator norm of A.

Chatterjee applies his technique, among other things, to spectra of random matrices.

Chapter 3

Linear programming inequality and concentration

3.1 The inequality

In this section, taken almost verbatim from [32], we prove the main inequality of this thesis, which is at the core of all of our concentration results. We extend the subsequence notation defined in §2.1.2 to vectors $w \in \mathbb{R}^n$: for $1 \le k \le \ell \le n$, we write $w_k^\ell = (w_k, \ldots, w_\ell) \in \mathbb{R}^{k-\ell+1}$. Fixing a finite set Ω, $n > 0$ and $w \in \mathbb{R}^n_+$, we make the following definitions.

1. K_n denotes the set of all functions $\kappa : \Omega^n \to \mathbb{R}$ (and $K_0 = \mathbb{R}$)
2. the *weighted Hamming metric* d_w on $\Omega^n \times \Omega^n$ is defined as in (1.5)
3. for $\varphi \in K_n$, its *Lipschitz constant* with respect to d_w, denoted by $\|\varphi\|_{\mathrm{Lip},w}$, is defined to be the smallest c for which
$$|\varphi(x) - \varphi(y)| \quad \le \quad c d_w(x, y)$$
for all $x, y \in \Omega^n$; any φ with $\|\varphi\|_{\mathrm{Lip},w} \le c$ is called c-Lipschitz
4. for $v \in [0, \infty)$, define $\Phi_{w,n}^{+v} \subset K_n$ to be the set of all φ such that $\|\varphi\|_{\mathrm{Lip},w} \le 1$ and
$$0 \ \le \ \varphi(x) \ \le \ \|w\|_1 + v, \qquad x \in \Omega^n;$$
we omit the $+v$ superscript when $v = 0$, writing simply $\Phi_{w,n}$
5. the *projection operator* $(\cdot)'$ takes $\kappa \in K_n$ to $\kappa' \in K_{n-1}$ by
$$\kappa'(y) \quad = \quad \sum_{x_1 \in \Omega} \kappa(x_1 y), \qquad y \in \Omega^{n-1};$$
for $n = 1$, κ' is the scalar $\kappa' = \sum_{x_1 \in \Omega} \kappa(x_1)$
6. for $y \in \Omega$, the *y-section* operator $(\cdot)_y$ takes $\kappa \in K_n$ to $\kappa_y \in K_{n-1}$ by
$$\kappa_y(x) \quad = \quad \kappa(xy), \qquad x \in \Omega^{n-1};$$
for $n = 1$, $\kappa_y(\cdot)$ is the scalar $\kappa(y)$

7. the functional $\Psi_{w,n} : K_n \to \mathbb{R}$ is defined by $\Psi_{w,0}(\cdot) = 0$ and

$$\Psi_{w,n}(\kappa) \quad = \quad w_1 \sum_{x \in \Omega^n} (\kappa(x))_+ + \Psi_{w_2^n, n-1}(\kappa'); \tag{3.1}$$

   when $w_i \equiv 1$ we omit it from the subscript, writing simply Ψ_n

8. the finite-dimensional vector space K_n is equipped with the inner product

$$\langle \kappa, \lambda \rangle \quad = \quad \sum_{x \in \Omega^n} \kappa(x)\lambda(x)$$

9. two norms are defined on $\kappa \in K_n$: the Φ_w-norm,

$$\|\kappa\|_{\Phi,w} \quad = \quad \sup_{\varphi \in \Phi_{w,n}} |\langle \kappa, \varphi \rangle| \tag{3.2}$$

   and the Ψ_w-norm,

$$\|\kappa\|_{\Psi,w} \quad = \quad \max_{s=\pm 1} \Psi_{w,n}(s\kappa). \tag{3.3}$$

Remark 3.1.1. For the special case $w_i \equiv 1$, d_w is the unweighted Hamming metric used in [37]. It is straightforward to verify that Φ_w-norm and Ψ_w-norm satisfy the vector-space norm axioms for any $w \in \mathbb{R}^n_+$; this is done in [37] for $w_i \equiv 1$. Since we will not be appealing to any norm properties of these functionals, we defer the proof to Chapter 5.2. Note that for any $y \in \Omega$, the projection and y-section operators commute; in other words, for $\kappa \in K_{n+2}$, we have $(\kappa')_y = (\kappa_y)' \in K_n$ and so we can denote this common value by $\kappa'_y \in K_n$:

$$\kappa'_y(z) = \sum_{x_1 \in \Omega} \kappa_y(x_1 z) = \sum_{x_1 \in \Omega} \kappa(x_1 z y), \qquad z \in \Omega^n.$$

Finally, recall that a norm $\|\cdot\|$ is called *absolute* if $\|x\| = \||x|\|$, where $|\cdot|$ is applied componentwise and *monotone* if $\|x\| \leq \|y\|$ whenever $|x| \leq |y|$ componentwise. Norms having these properties are also called *Riesz* norms; the two conditions are equivalent for finite-dimensional spaces [28]. Neither Φ_w-norm nor Ψ_w-norm is a Riesz norm; these should be thought of as measuring *oscillation* as opposed to *magnitude*. Indeed, for nonnegative κ, we trivially have

$$\|\kappa\|_{\Phi,w} = \|\kappa\|_{\Psi,w} = \|w\|_1 \|\kappa\|_1 ,$$

so the inequality is only interesting for κ with oscillating signs. $\diamond$

The main result of this section is

Theorem 3.1.2. *For all $w \in \mathbb{R}^n_+$ and all $\kappa \in K_n$, we have*

$$\|\kappa\|_{\Phi,w} \quad \leq \quad \|\kappa\|_{\Psi,w} . \tag{3.4}$$

Remark 3.1.3. We refer to (3.4) – more properly, to (3.8), from which the former immediately follows – as a *linear programming inequality* for the reason that $F(\cdot) = \langle \kappa, \cdot \rangle$ is a linear function being maximized over the finitely generated, compact, convex polytope $\Phi_{w,n} \subset \mathbb{R}^{\Omega^n}$. We make no use of this simple fact and therefore forgo its proof, but see [37, Lemma 4.4] for a proof of a closely related claim. The term "linear programming" is a bit of a red herring since no actual LP techniques are being used; for lack of an obvious natural name, we have alternatively referred to precursors of (3.4) in previous papers and talks as the "Φ-norm bound" or the "Φ-Ψ inequality." $\diamond$

The key technical lemma is a decomposition of $\Psi_{w,n}(\cdot)$ in terms of y-sections, proved in [37] for the case $w_i \equiv 1$:

Lemma 3.1.4. *For all $n \geq 1$, $w \in \mathbb{R}^n_+$ and $\kappa \in K_n$, we have*

$$\Psi_{w,n}(\kappa) \;=\; \sum_{y\in\Omega}\left[\Psi_{w_1^{n-1},n-1}(\kappa_y) + w_n\left(\sum_{x\in\Omega^{n-1}}\kappa_y(x)\right)_+\right]. \tag{3.5}$$

Proof. We proceed by induction on n. To prove the $n=1$ case, recall that Ω^0 is the set containing a single (null) word and that for $\kappa \in K_1$, $\kappa_y \in K_0$ is the scalar $\kappa(y)$. Thus, by definition of $\Psi_{w,1}(\cdot)$, we have

$$\Psi_{w,1}(\kappa) \;=\; w_1\sum_{y\in\Omega}\kappa(y),$$

which proves (3.5) for $n=1$.

Suppose the claim holds for some $n=\ell\geq 1$. Pick any $w\in\mathbb{R}^{\ell+1}_+$ and $\kappa\in K_{\ell+1}$ and examine

$$\begin{aligned}
&\sum_{y\in\Omega}\left[\Psi_{w_1^\ell,\ell}(\kappa_y) + w_{\ell+1}\left(\sum_{x\in\Omega^\ell}\kappa_y(x)\right)_+\right]\\
&\qquad= \sum_{y\in\Omega}\left[\left(w_1\sum_{x\in\Omega^\ell}(\kappa_y(x))_+ + \Psi_{w_2^\ell,\ell-1}(\kappa'_y)\right) + w_{\ell+1}\left(\sum_{x\in\Omega^\ell}\kappa_y(x)\right)_+\right]\\
&\qquad= \sum_{y\in\Omega}\left[\Psi_{w_2^\ell,\ell-1}(\kappa'_y) + w_{\ell+1}\left(\sum_{u\in\Omega^{\ell-1}}\kappa'_y(u)\right)_+\right] + w_1\sum_{z\in\Omega^{\ell+1}}(\kappa(z))_+
\end{aligned} \tag{3.6}$$

where the first equality follows from the definition of $\Psi_{w_1^\ell,\ell}$ in (3.1) and the second one from the easy identities

$$\sum_{y\in\Omega}\sum_{x\in\Omega^\ell}(\kappa_y(x))_+ \;=\; \sum_{z\in\Omega^{\ell+1}}(\kappa(z))_+$$

and

$$\sum_{x\in\Omega^\ell}\kappa_y(x) \;=\; \sum_{u\in\Omega^{\ell-1}}\kappa'_y(u).$$

On the other hand, by definition we have

$$\Psi_{w,\ell+1}(\kappa) \;=\; w_1\sum_{z\in\Omega^{\ell+1}}(\kappa(z))_+ + \Psi_{w_2^{\ell+1},\ell}(\kappa'). \tag{3.7}$$

To compare the r.h.s. of (3.6) with the r.h.s. of (3.7), note that the $w_1\sum_{z\in\Omega^{\ell+1}}(\kappa(z))_+$ term is common to both and

$$\sum_{y\in\Omega}\left[\Psi_{w_2^\ell,\ell-1}(\kappa'_y) + w_{\ell+1}\left(\sum_{u\in\Omega^{\ell-1}}\kappa'_y(u)\right)_+\right] \;=\; \Psi_{w_2^{\ell+1},\ell}(\kappa')$$

by the inductive hypothesis. This establishes (3.5) for $n=\ell+1$ and proves the claim. □

Our main result, Theorem 3.1.2, is an immediate consequence of

Theorem 3.1.5. *For all $n \geq 1$, $w \in \mathbb{R}_+^n$, $v \in [0,\infty)$ and $\kappa \in K_n$, we have*

$$\sup_{\varphi \in \Phi_{w,n}^{+v}} \langle \kappa, \varphi \rangle \quad \leq \quad \Psi_{w,n}(\kappa) + v \left(\sum_{x \in \Omega^n} \kappa(x) \right)_+ . \tag{3.8}$$

Proof. We will prove the claim by induction on n. For $n = 1$, pick any $w_1 \in \mathbb{R}_+$, $v \in [0,\infty)$ and $\kappa \in K_1$. Since by construction any $\varphi \in \Phi_{w_1,1}^{+v}$ is w_1-Lipschitz with respect to the discrete metric on Ω, φ must be of the form

$$\varphi(x) \quad = \quad \tilde{\varphi}(x) + \tilde{v}, \qquad x \in \Omega,$$

where $\tilde{\varphi} : \Omega \to [0, w_1]$ and $0 \leq \tilde{v} \leq v$ (in fact, we have the explicit value $\tilde{v} = (\max_{x \in \Omega} \varphi(x) - w_1)_+$). Therefore,

$$\langle \kappa, \varphi \rangle \quad = \quad \langle \kappa, \tilde{\varphi} \rangle + \tilde{v} \sum_{x \in \Omega} \kappa(x). \tag{3.9}$$

The first term in the r.h.s. of (3.9) is clearly maximized when $\tilde{\varphi}(x) = w_1 \mathbb{1}_{\{\kappa(x) > 0\}}$ for all $x \in \Omega$, which shows that it is bounded by $\Psi_{w_1,1}(\kappa)$. Since the second term in the r.h.s. of (3.9) is bounded by $v \left(\sum_{x \in \Omega} \kappa(x) \right)_+$, we have established (3.8) for $n = 1$.

Now suppose the claim holds for $n = \ell$, and pick any $w \in \mathbb{R}_+^{\ell+1}$, $v \in [0,\infty)$ and $\kappa \in K_{\ell+1}$. By the reasoning given above (i.e., using the fact that $0 \leq \varphi \leq v + \sum_{i=1}^{\ell+1} w_i$ and that φ is 1-Lipschitz with respect to d_w), any $\varphi \in \Phi_{w,\ell+1}^{+v}$, must be of the form $\varphi = \tilde{\varphi} + \tilde{v}$, where $\tilde{\varphi} \in \Phi_{w,\ell+1}$ and $0 \leq \tilde{v} \leq v$. Thus we write $\langle \kappa, \varphi \rangle = \langle \kappa, \tilde{\varphi} \rangle + \tilde{v} \sum_{x \in \Omega^{\ell+1}} \kappa(x)$ and decompose

$$\langle \kappa, \tilde{\varphi} \rangle \quad = \quad \sum_{y \in \Omega} \langle \kappa_y, \tilde{\varphi}_y \rangle , \tag{3.10}$$

making the obvious but crucial observation that

$$\tilde{\varphi} \in \Phi_{w,\ell+1} \quad \Longrightarrow \quad \tilde{\varphi}_y \in \Phi_{w_1^\ell,\ell}^{+w_{\ell+1}}.$$

Then it follows by the inductive hypothesis that

$$\langle \kappa_y, \tilde{\varphi}_y \rangle \quad \leq \quad \Psi_{w_1^\ell,\ell}(\kappa_y) + w_{\ell+1} \left(\sum_{x \in \Omega^\ell} \kappa_y(x) \right)_+ . \tag{3.11}$$

Applying Lemma 3.1.4 to (3.11), we have

$$\sum_{y \in \Omega} \langle \kappa_y, \tilde{\varphi}_y \rangle \ \leq \ \sum_{y \in \Omega} \left[\Psi_{w_1^\ell,\ell}(\kappa_y) + w_{\ell+1} \left(\sum_{x \in \Omega^\ell} \kappa_y(x) \right)_+ \right] \ = \ \Psi_{w,\ell+1}(\kappa). \tag{3.12}$$

This, combined with (3.10) and the trivial bound

$$\tilde{v} \sum_{x \in \Omega^{\ell+1}} \kappa(x) \quad \leq \quad v \left(\sum_{x \in \Omega^{\ell+1}} \kappa(x) \right)_+$$

proves the claim for $n = \ell + 1$ and hence for all n. □

Remark 3.1.6. The power of Theorem 3.1.2 comes from its bound of a natural but not readily computable (in closed form) quantity by a less intuitive but easily computed quantity. Although our main application of this inequality is to bound the martingale difference in Theorem 3.3.4, one hopes that it will find other applications. One such possibility is a bound on the transportation cost distance, via Kantorovich's duality theorem; see §5.6. $\diamond$

3.2 η-mixing

The notion of mixing we define here is by no means new; it can be traced (at least implicitly) to Marton's work [46] and is quite explicit in Samson [60] and Chazottes et al. [13]. We are not aware of a standardized term for this type of mixing, and have referred to it as *η-mixing* in previous work [37]. That choice of terminology is perhaps suboptimal in light of the unrelated notion of *η-dependence* of Doukhan et al. [19], but the sufficiently distinct contexts should help avoid confusion. We will observe a few simple facts about η-mixing coefficients.

3.2.1 Definition

Let $(\Omega^n, \mathcal{F}, \mathbf{P})$ be a probability space and $(X_i)_{1 \le i \le n}$ its associated random process. For $1 \le i < j \le n$ and $x \in \Omega^i$, let

$$\mathcal{L}(X_j^n \,|\, X_1^i = x)$$

be the law (distribution) of X_j^n conditioned on $X_1^i = x$. For $y \in \Omega^{i-1}$ and $w, w' \in \Omega$, define

$$\eta_{ij}(y, w, w') \;=\; \left\| \mathcal{L}(X_j^n \,|\, X_1^i = yw) - \mathcal{L}(X_j^n \,|\, X_1^i = yw') \right\|_{\mathrm{TV}}, \tag{3.13}$$

where

$$\bar{\eta}_{ij} \;=\; \max_{y \in \Omega^{i-1}} \max_{w, w' \in \Omega} \eta_{ij}(y, w, w'). \tag{3.14}$$

Let $\Delta_n = \Delta_n(\mathbf{P})$ be the upper-triangular $n \times n$ matrix defined by $(\Delta_n)_{ii} = 1$ and

$$(\Delta_n)_{ij} = \bar{\eta}_{ij}. \tag{3.15}$$

for $1 \le i < j \le n$. Recall that the ℓ_∞ operator norm is given by

$$\|\Delta_n\|_\infty \;=\; \max_{1 \le i < n} (1 + \bar{\eta}_{i,i+1} + \ldots + \bar{\eta}_{i,n}). \tag{3.16}$$

We say that the process X on $(\Omega^{\mathbb{N}}, \mathcal{F}, \mathbf{P})$ is *η-mixing* if

$$\sup_{n \ge 1} \|\Delta_n(\mathbf{P})\|_\infty < \infty. \tag{3.17}$$

Let us collect some simple observations about $\|\Delta_n(\cdot)\|_\infty$:

Lemma 3.2.1. *Let μ be a probability measure on Ω^n. Then*

(a) $1 \le \|\Delta_n(\mu)\|_\infty \le n$

(b) $\|\Delta_n(\mu)\|_\infty = 1$ *iff μ is a product measure*

(c) if ν is a measure on Ω^m then

$$\|\Delta_{n+m}(\mu \otimes \nu)\|_\infty \quad \leq \quad \max\left\{\|\Delta_n(\mu)\|_\infty, \|\Delta_m(\nu)\|_\infty\right\}.$$

Proof. (a) is immediate from $\bar{\eta}_{ij} \leq 1$; (b) the "if" direction is trivial; "only if" is established by proving the (straightforward) $n = 2$ case and applying induction; (c) follows by observing that $\Delta_{m+n}(\mu \otimes \nu)$ is a block-diagonal matrix. □

Remark 3.2.2. A careful reader will note that $\bar{\eta}_{ij}$ may also depend on the sequence length n; thus any meaningful bound on this quantity must either take this dependence into account or be dimension-free. The bounds we derive below are of the latter type. ◇

3.2.2 Connection to ϕ-mixing

Samson [60], using techniques quite different from those here, showed that if $\Omega = [0,1]$, and $f : [0,1]^n \to \mathbb{R}$ is convex with $\|f\|_{\text{Lip}} \leq 1$ (in the ℓ_2 metric), then

$$\mathbf{P}\{|f(X) - \mathbf{E}f(X)| > t\} \quad \leq \quad 2\exp\left(-\frac{t^2}{2\|\Gamma_n\|_2^2}\right) \tag{3.18}$$

where $\|\Gamma_n\|_2$ is the ℓ_2 operator norm of the matrix

$$(\Gamma_n)_{ij} = \sqrt{(\Delta_n)_{ij}}, \tag{3.19}$$

where $\sqrt{\cdot}$ is applied to Γ_n component-wise. Following Bradley [7], for the random process $(X_i)_{i\in\mathbb{Z}}$ on $(\Omega^{\mathbb{Z}}, \mathcal{F}, \mathbf{P})$, we define the ϕ-mixing coefficient

$$\phi(k) \quad = \quad \sup_{j\in\mathbb{Z}} \phi(\mathcal{F}_{-\infty}^j, \mathcal{F}_{j+k}^\infty), \tag{3.20}$$

where $\mathcal{F}_i^j \subset \mathcal{F}$ is the σ-algebra generated by the X_i^j, and for the σ-algebras $\mathcal{A}, \mathcal{B} \subset \mathcal{F}$, $\phi(\mathcal{A}, \mathcal{B})$ is defined by

$$\phi(\mathcal{A}, \mathcal{B}) \quad = \quad \sup\{|\mathbf{P}(B \,|\, A) - P(B)| : A \in \mathcal{A},\ B \in \mathcal{B}, \mathbf{P}(A) > 0\}. \tag{3.21}$$

Samson observes that

$$\bar{\eta}_{ij} \quad \leq \quad 2\phi_{j-i}, \tag{3.22}$$

which follows from

$$\begin{aligned} \left\|\mathcal{L}(X_j^n \,|\, X_1^i = y_1^{i-1}w) - \mathcal{L}(X_j^n \,|\, X_1^i = y_1^{i-1}w')\right\|_{\text{TV}} \quad &\leq \quad \left\|\mathcal{L}(X_j^n \,|\, X_1^i = y_1^{i-1}w) - \mathcal{L}(X_j^n)\right\|_{\text{TV}} \\ &+ \quad \left\|\mathcal{L}(X_j^n \,|\, X_1^i = y_1^{i-1}w') - \mathcal{L}(X_j^n)\right\|_{\text{TV}}. \end{aligned}$$

This observation, together with (3.16), implies a sufficient condition for η-mixing:

$$\sum_{k=1}^\infty \phi_k \quad < \quad \infty; \tag{3.23}$$

this certainly holds if (ϕ_k) admits a geometric decay, as assumed in [60].

Although η-mixing seems to be a stronger condition than ϕ-mixing (the latter only requires $\phi_k \to 0$), we are presently unable to obtain any nontrivial implications (or non-implications) between η-mixing and either ϕ-mixing or any of the other strong mixing conditions discussed in [7]. A fuller discussion of mixing is deferred until Chapter 6.5.1.

3.3 Concentration inequality

The main probability-theoretic inequality of this thesis is the following:

Theorem 3.3.1. *Let Ω be a finite set and $\mathbf{P}$ a measure on Ω^n, for $n \geq 1$. For any $w \in \mathbb{R}_+^n$ and $f : \Omega^n \to \mathbb{R}$, we have*

$$\mathbf{P}\{|f - \mathbf{E}f| > t\} \quad \leq \quad 2\exp\left(-\frac{t^2}{2\,\|f\|_{\mathrm{Lip},w}^2\,\|\Delta_n w\|_2^2}\right)$$

where $\|\cdot\|_{\mathrm{Lip},w}$ is defined in §3.1 and Δ_n is the η-mixing matrix defined in (3.15).

It is proved by bounding the martingale difference and appealing to Azuma's inequality. The first order of business is to bound the martingale difference by a slightly more tractable quantity.

For $f : \Omega^n \to \mathbb{R}$, $1 \leq i \leq n$ and $y_1^i \in \Omega^i$, define

$$V_i(f; y_1^i) \quad = \quad \mathbf{E}[f(X) \,|\, X_1^i = y_1^i] - \mathbf{E}[f(X) \,|\, X_1^{i-1} = y_1^{i-1}]; \tag{3.24}$$

this is just the martingale difference. It will be more convenient to work with the modified martingale difference:

$$\hat{V}_i(f; y_1^{i-1}, z_i, z_i') \quad = \quad \mathbf{E}[f(X) \,|\, X_1^i = y_1^{i-1} z_i] - \mathbf{E}[f(X) \,|\, X_1^i = y_1^{i-1} z_i'], \tag{3.25}$$

where $z_i, z_i' \in \Omega$. These two quantities have a simple relationship, which may be stated symbolically as $\|V_i(f;\cdot)\|_\infty \leq \|\hat{V}_i(f;\cdot)\|_\infty$ and is proved in the following lemma, adapted from [35]:

Lemma 3.3.2. *Suppose $f : \Omega^n \to \mathbb{R}$ and $y_1^i \in \Omega^i$. Then there are $z_i, z_i' \in \Omega$ such that*

$$\left|V_i(f; y_1^i)\right| \quad \leq \quad \left|\hat{V}_i(f; y_1^{i-1}, z_i, z_i')\right|. \tag{3.26}$$

Remark 3.3.3. Here and below $p(x_j^n \,|\, y_1^i)$ will occasionally be used in place of $\mathbf{P}\left\{X_j^n = x_j^n \,|\, X_1^i = y_1^i\right\}$; no ambiguity should arise.

Proof. Let

$$a = \mathbf{E}[f(X) \,|\, X_1^i = y_1^i] = \sum_{x_{i+1}^n \in \Omega^{n-i}} p(x_{i+1}^n \,|\, y_1^i) f(y_1^i x_{i+1}^n);$$

then

$$\begin{aligned} V_i(f; y_1^i) \quad &= \quad a - \sum_{x_i^n \in \Omega^{n-i+1}} p(x_i^n \,|\, y_1^{i-1}) f(y_1^{i-1} x_i^n) \\ &= \quad a - \sum_{z \in \Omega} p(z \,|\, y_1^{i-1}) \left(\sum_{x_{i+1}^n \in \Omega^{n-i}} p(x_{i+1}^n \,|\, y_1^{i-1} z) f(y_1^{i-1} z x_{i+1}^n) \right) \end{aligned}$$

We use the simple fact that for $g, h : \Omega \to \mathbb{R}_+$

$$\min h(z) \sum g(z) \;\leq\; \sum g(z)h(z) \;\leq\; \max h(z) \sum g(z),$$

together with $\sum_{z\in\Omega} p(z \,|\, y_1^{i-1}) = 1$, to deduce the existence of a $z_i' \in \Omega$ such that

$$\left|V_i(f; y_1^i)\right| \;\leq\; \left| a - \sum_{x_{i+1}^n \in \Omega^{n-i}} p(x_{i+1}^n \,|\, y_1^{i-1} z_i') f(y_1^{i-1}\, z_i' x_{i+1}^n) \right|.$$

Taking $z_i = y_i$, this proves the claim. □

The next step is to notice that $\hat{V}_i(\cdot; y_1^{i-1}, z_i, z_i')$, as a functional on K_n (see §3.1), is linear; in fact, it is given by

$$\hat{V}_i(f; y_1^{i-1}, z_i, z_i') \;=\; \sum_{x\in\Omega^n} f(x)\hat{g}(x) \;=\; \langle f, \hat{g}\rangle\,, \tag{3.27}$$

where

$$\hat{g}(x) \;=\; \mathbb{1}_{\{x_1^i = y_1^{i-1} z_i\}} p(x_{i+1}^n \,|\, y_1^{i-1} z_i) - \mathbb{1}_{\{x_1^i = y_1^{i-1} z_i'\}} p(x_{i+1}^n \,|\, y_1^{i-1} z_i'). \tag{3.28}$$

An application of Theorem 3.1.2 to $\hat{g}$ yields a bound on the martingale difference.

Theorem 3.3.4. *Let Ω be a finite set, and let $(X_i)_{1\leq i\leq n}$, $X_i \in \Omega$ be the random process associated with the measure $\mathbf{P}$ on Ω^n. Let Δ_n be the upper-triangular $n \times n$ matrix defined in (3.15). Then, for all $w \in \mathbb{R}_+^n$ and $f : \Omega^n \to \mathbb{R}$, we have*

$$\sum_{i=1}^n \bar{V}_i^2(f) \;\leq\; \|f\|_{\mathrm{Lip},w}^2 \, \|\Delta_n w\|_2^2 \tag{3.29}$$

where

$$\bar{V}_i(f) \;=\; \max_{y_1^i \in \Omega^i} \left|V_i(f; y_1^i)\right|. \tag{3.30}$$

Remark 3.3.5. Since $\bar{V}_i(f)$ and $\|f\|_{\mathrm{Lip},w}$ are both homogeneous functionals of f (in the sense that $T(af) = |a|T(f)$ for $a \in \mathbb{R}$), there is no loss of generality in taking $\|f\|_{\mathrm{Lip},w} = 1$. Additionally, since $V_i(f; y)$ is translation-invariant (in the sense that $V_i(f; y) = V_i(f + a; y)$ for all $a \in \mathbb{R}$), there is no loss of generality in restricting the range of f to $[0, \mathrm{diam}_{d_w}(\Omega^n)]$. In other words, it suffices to consider $f \in \Phi_{w,n}$. Since essentially this result (for $w_i \equiv 1$) is proved in [37] in some detail, we only give a proof sketch here, highlighting the changes needed for general w.

Proof. It was shown in Section 5 of [37] that if d_w is the unweighted Hamming metric (that is, $w_i \equiv 1$) and $f : \Omega^n \to \mathbb{R}$ is 1-Lipschitz with respect to d_w, then

$$\bar{V}_i(f) \;\leq\; 1 + \sum_{j=i+1}^n \bar{\eta}_{ij}. \tag{3.31}$$

This is seen by combining Lemma 3.3.2 with (3.27) to conclude that for $1 \leq i \leq n$ and $y \in \Omega^i$, there are $z_i, z_i' \in \Omega$ and $\hat{g}_i : \Omega^n \to \mathbb{R}$ (whose explicit construction, depending on y, z_i, z_i' and $\mathbf{P}$, is given in (3.28)), such that for all $f : \Omega^n \to \mathbb{R}$, we have

$$|V_i(f; y)| \;\leq\; |\langle \hat{g}_i, f\rangle|\,. \tag{3.32}$$

It is likewise easily verified (as done in [37, Theorem 5.1]) that

$$\langle \hat{g}_i, f \rangle \quad = \quad \langle T_y \hat{g}_i, T_y f \rangle \,,$$

where the operator $T_y : K_n \to K_{n-i+1}$ is defined by

$$(T_y h)(x) \quad = \quad h(yx), \qquad \text{for all } x \in \Omega^{n-i+1}.$$

Appealing to Theorem 3.1.5 with $w_i \equiv 1$, we get

$$\langle T_y \hat{g}_i, T_y f \rangle \quad \leq \quad \Psi_n(T_y \hat{g}_i). \tag{3.33}$$

It is now a simple matter to apply the definition of Ψ_n and recall a characterization of $\|\cdot\|_{\mathrm{TV}}$ (namely, (2.3)), to obtain

$$\Psi_n(T_y \hat{g}_i) \quad \leq \quad 1 + \sum_{j=i+1}^{n} \bar{\eta}_{ij}, \tag{3.34}$$

establishing (3.31). To generalize (3.31) to $w_i \not\equiv 1$, we use the fact that if $f \in K_n$ is 1-Lipschitz with respect to d_w, then $T_y f \in K_{n-i+1}$ is 1-Lipschitz with respect to $d_{w_i^n}$. Thus, applying Theorem 3.1.5, we get

$$\langle T_y \hat{g}_i, f \rangle \quad \leq \quad \Psi_{w_i^n, n-i+1}(T_y \hat{g}_i). \tag{3.35}$$

It follows directly from the definition of $\Psi_{w,n}$ and the calculations above that

$$\bar{V}_i(f) \quad \leq \quad w_i + \sum_{j=i+1}^{n} w_j \bar{\eta}_{ij} \tag{3.36}$$

$$= \quad \sum_{j=1}^{n} (\Delta_n)_{ij} w_j = (\Delta_n w)_i. \tag{3.37}$$

Squaring and summing over i, we obtain (3.29). □

Proof of Theorem 3.3.4. Since by definition of the ℓ_2 operator norm, $\|\Delta_n w\|_2 \leq \|\Delta_n\|_2 \, \|w\|_2$, the claim follows immediately via (2.17) and (3.29). □

Chapter 4

Applications

We will have three sections dealing with applications. First, we proceed to apply the general inequality to various processes: Markov, hidden Markov, and Markov tree. The next application deals with laws of large numbers for strongly mixing processes and yields an analysis of an inhomogeneous Markov Chain Monte Carlo algorithm; this is joint work with Anthony Brockwell. Finally, we exhibit some applications of our techniques to empirical process theory and machine learning.

4.1 Bounding $\bar{\eta}_{ij}$ for various processes

4.1.1 Notational conventions

Sums will range over the entire space of the summation variable; thus $\sum_{x_i^j} f(x_i^j)$ stands for

$$\sum_{x_i^j \in \Omega^{j-i+1}} f(x_i^j).$$

By convention, when $i > j$, we define

$$\sum_{x_i^j} f(x_i^j) \equiv f(\varepsilon)$$

where ε is the null sequence.

4.1.2 Markov chains

Although technically this section might be considered superfluous – its results are strictly generalized in both §4.1.4 and §4.1.5, it is instructive to work out the simple Markov case as it provides the cleanest illustration of our techniques. This was, in fact, the motivating example that prompted the investigation of the more general case, culminating in Theorem 3.1.2.

Let μ be an inhomogeneous Markov measure on Ω^n, induced by the kernels p_0 and $p_i(\cdot \,|\, \cdot)$, $1 \leq i < n$. Thus,

$$\mu(x) \quad = \quad p_0(x_1) \prod_{i=1}^{n-1} p_i(x_{i+1} \,|\, x_i).$$

Define the i^{th} *contraction coefficient*:

$$\theta_i \;=\; \max_{y,y' \in \Omega} \left\| p_i(\cdot \,|\, y) - p_i(\cdot \,|\, y') \right\|_{\text{TV}};$$

this quantity turns out to control the η-mixing coefficients for μ:

Theorem.

$$\bar{\eta}_{ij} \;\leq\; \theta_i \theta_{i+1} \ldots \theta_{j-1}.$$

This fact is proved in [60] using coupling. We will take a different route, via the Markov contraction lemma:

Lemma 4.1.1. *Let $P : \mathbb{R}^\Omega \to \mathbb{R}^\Omega$ be a Markov operator:*

$$(P\nu)(x) = \sum_{y \in \Omega} P(x \,|\, y) \nu(y),$$

where $P(x \,|\, y) \geq 0$ and $\sum_{x \in \Omega} P(x \,|\, y) = 1$. Define the contraction coefficient of P as above:

$$\theta \;=\; \max_{y,y' \in \Omega} \left\| P(\cdot \,|\, y) - P(\cdot \,|\, y') \right\|_{\text{TV}}.$$

Then

$$\left\| P\nu \right\|_{\text{TV}} \;\leq\; \theta \left\| \nu \right\|_{\text{TV}}$$

for any balanced signed measure ν on Ω (i.e., $\nu \in \mathbb{R}^\Omega$ with $\sum_{x \in \Omega} \nu(x) = 0$).

This result is sometimes credited to Dobrushin [18]; in fact, θ is also known as *Dobrushin's ergodicity coefficient*. However, the inequality seems to go as far back as Markov himself [44]; see [37] for a proof.

Proof of Theorem 4.1.2. Fix $1 \leq i < j \leq n$ and $y_1^{i-1} \in \Omega^{i-1}$, $w_i, w_i' \in \Omega$. Then

$$\begin{aligned} \eta_{ij}(y, w, w') \;&=\; \tfrac{1}{2} \sum_{x_j^n} \left| \mu(x_j^n \,|\, y_1^{i-1} w_i) - \mu(x_j^n \,|\, y_1^{i-1} w_i') \right| \\ &=\; \tfrac{1}{2} \sum_{x_j^n} \pi(x_j^n) \left| \zeta(x_j) \right| \end{aligned}$$

where

$$\pi(u_k^\ell) \;=\; \prod_{t=k}^{\ell-1} p_t(u_{t+1} \,|\, u_t)$$

and

$$\zeta(x_j) \;=\; \begin{cases} \displaystyle\sum_{z_{i+1}^{j-1}} p_{j-1}(x_j \,|\, z_{j-1}) \pi(z_{i+1}^{j-1}) \left(p_i(z_{i+1} \,|\, w_i) - p_i(z_{i+1} \,|\, w_i') \right), & j - i > 1 \\ p_i(x_j \,|\, w_i) - p_i(x_j \,|\, w_i'), & j - i = 1. \end{cases} \tag{4.1}$$

Define $\mathbf{h} \in \mathbb{R}^\Omega$ by $\mathbf{h}_v = p_i(v \,|\, w_i) - p_i(v \,|\, w_i')$ and $P^{(k)} \in \mathbb{R}^{\Omega\times\Omega}$ by $P^{(k)}_{u,v} = p_k(u \,|\, v)$. Likewise, define $\mathbf{z} \in \mathbb{R}^\Omega$ by $\mathbf{z}_v = \zeta(v)$. It follows that

$$\mathbf{z} = P^{(j-1)} P^{(j-2)} \dots P^{(i+2)} P^{(i+1)} \mathbf{h}.$$

Therefore,

$$\begin{aligned} \eta_{ij}(y, w, w') &= \tfrac{1}{2} \sum_{x_j^n} \pi(x_j^n) \left|\mathbf{z}_{x_j}\right| \\ &= \tfrac{1}{2} \sum_{x_j} \left|\mathbf{z}_{x_j}\right| \sum_{x_{j+1}^n} \pi(x_j^n) \\ &= \tfrac{1}{2} \sum_{x_j} \left|\mathbf{z}_{x_j}\right| = \|\mathbf{z}\|_{\mathrm{TV}}. \end{aligned}$$

The claim follows by the contraction lemma. □

4.1.3 Undirected Markov chains

For any graph $G = (V, E)$, where $|V| = n$ and the maximal cliques have size 2 (are edges), we can define a measure on Ω^n as follows

$$\mu(x) \equiv \mathbf{P}\{X = x\} = \frac{\prod_{(i,j)\in E} \psi_{ij}(x_i, x_j)}{\sum_{x'\in\Omega^n} \prod_{(i,j)\in E} \psi_{ij}(x_i', x_j')} \equiv \frac{\prod_{(i,j)\in E} \psi_{ij}(x_i, x_j)}{Z_G}$$

for some $\psi_{ij} \geq 0$.

Consider the very simple case of chain graphs; any such measure is a Markov measure on Ω^n. We can relate the induced Markov transition kernel $p_i(\cdot \,|\, \cdot)$ to the random field measure μ as follows:

$$p_i(x \,|\, y) = \frac{\sum_{v_1^{i-1}} \sum_{z_{i+2}^n} \mu[v\, y\, x\, z]}{\sum_{x'\in\Omega} \sum_{v'^{i-1}_1} \sum_{z'^n_{i+2}} \mu[v'\, y\, x'\, z']}, \qquad x, y \in \Omega.$$

Our goal is to bound the i^{th} contraction coefficient θ_i of the Markov chain:

$$\theta_i = \max_{y,y'\in\Omega} \frac{1}{2} \sum_{x\in\Omega} \left|p_i(x \,|\, y) - p_i(x \,|\, y')\right|.$$

in terms of ψ_{ij}. We claim a simple relationship between θ_i and ψ_{ij}:

Theorem 4.1.2.

$$\theta_i \leq \frac{R_i - r_i}{R_i + r_i} \tag{4.2}$$

where

$$R_i = \max_{x,y\in\Omega} \psi_{i,i+1}(x, y)$$

and

$$r_i = \min_{x,y\in\Omega} \psi_{i,i+1}(x, y).$$

First we prove a simple lemma:

Lemma 4.1.3. *Let $\alpha, \beta, \gamma \in \mathbb{R}_+^{k+1}$ and $r, R \in \mathbb{R}$ be such that $0 \leq r \leq \alpha_i, \beta_i \leq R$, for $1 \leq i \leq k+1$. Then*

$$\frac{1}{2}\sum_{i=1}^{k+1}\left|\frac{\alpha_i\gamma_i}{\sum_{j=1}^{k+1}\alpha_j\gamma_j} - \frac{\beta_i\gamma_i}{\sum_{j=1}^{k+1}\beta_j\gamma_j}\right| \leq \frac{R-r}{R+r}. \tag{4.3}$$

Proof. When $p, q \in \mathbb{R}_+^{k+1}$ are two distributions satisfying $0 < r \leq p_i, q_i$, it is straightforward to verify that $\|p-q\|_1$ may be maximized, with value d, by choosing $a \in [r, (1-d)/k]$, $b = a + d/k$ and setting $p_i = a, q_i = b$ for $1 \leq i \leq k$ and $p_{k+1} = 1 - ka$, $q_{k+1} = 1 - kb$. Applying this principle to (4.2), we obtain

$$\begin{aligned}\sum_{i=1}^{k+1}\left|\frac{\alpha_i\gamma_i}{\sum_{j=1}^{k+1}\alpha_j\gamma_j} - \frac{\beta_i\gamma_i}{\sum_{j=1}^{k+1}\beta_j\gamma_j}\right| &\leq \frac{gkR - g'r}{gkR + g'r} - \frac{g'R - gkr}{g'R + gkr} \\ &= \frac{2g''k(R^2 - r^2)}{(R + g''kr)(g''kR + r)}\end{aligned}$$

where $g = \sum_{i=1}^{k}\gamma_i$, $g' = \gamma_{k+1}$ and $g'' = g/g'$.

Define $f : \mathbb{R}_+ \to \mathbb{R}$ by

$$f(x) = \frac{2(R^2 - r^2)x}{(R + rx)(Rx + r)};$$

elementary calculus verifies that f is maximized at $x = 1$. □

Proof of Theorem 4.1.2. Let us define the shorthand notation:

$$\pi(u_k^l) = \prod_{t=k}^{l-1}\psi_{t,t+1}(u_t, u_{t+1})$$

Then we expand

$$\begin{aligned}p_i(x\,|\,y) &= \frac{\sum_{v_1^{i-1}}\sum_{z_{i+2}^n}\pi(v_1^{i-2})\psi_{i-1,i}(v_{i-1}, y)\psi_{i,i+1}(y, x)\psi_{i+1,i+2}(x, z_{i+2})\pi(z_{i+2}^n)}{\sum_{x'\in\Omega}\sum_{{v'}_1^{i-1}}\sum_{{z'}_{i+2}^n}\pi({v'}_1^{i-2})\psi_{i-1,i}(v'_{i-1}, y)\psi_{i,i+1}(y, x')\psi_{i+1,i+2}(x', z'_{i+2})\pi({z'}_{i+2}^n)} \\ &= \frac{\psi_{i,i+1}(y, x)a_{yx}}{\sum_{x'\in\Omega}\psi_{i,i+1}(y, x')a_{yx'}}\end{aligned}$$

where

$$a_{yx} = \sum_{v_1^{i-1}}\sum_{z_{i+2}^n}\pi(v_1^{i-2})\psi_{i-1,i}(v_{i-1}, y)\psi_{i+1,i+2}(x, z_{i+2})\pi(z_{i+2}^n)$$

(we take the natural convention that $\psi_{i,j}(\cdot\,|\,\cdot) = 1$ whenever $(i,j) \notin E$).

Fix $y, y' \in \Omega$. Define the quantities, for each $x \in \Omega$:

$$\begin{aligned}\alpha_x &= \psi_{i,i+1}(y, x) \\ \beta_x &= \psi_{i,i+1}(y', x) \\ \gamma_x &= a_{yx} \\ \gamma'_x &= a_{y'x}.\end{aligned}$$

Then

$$\sum_{x\in\Omega} \left|p_i(x\,|\,y) - p_i(x\,|\,y')\right| = \sum_{x\in\Omega} \left| \frac{\alpha_x\gamma_x}{\sum_{x'\in\Omega}\alpha_{x'}\gamma_{x'}} - \frac{\beta_x\gamma'_x}{\sum_{x'\in\Omega}\beta_{x'}\gamma'_{x'}} \right| \tag{4.4}$$

$$= \sum_{x\in\Omega} \left| \frac{\alpha_x\gamma_x}{\sum_{x'\in\Omega}\alpha_{x'}\gamma_{x'}} - \frac{\beta_x\gamma_x}{\sum_{x'\in\Omega}\beta_{x'}\gamma_{x'}} \right|; \tag{4.5}$$

the last equality follows since $\gamma'_x = c\gamma_x$, where $c = \frac{\psi_{i-1,i}(v_{i-1},y')}{\psi_{i-1,i}(v_{i-1},y)}$. Now Lemma 4.1.3 can be applied to establish the claim. □

Since all the inequalities invoked are tight, so is the bound in Theorem 4.1.2.

4.1.4 Hidden Markov chains

The material in this section is taken almost entirely from [33]. Consider two finite sets, $\hat{\Omega}$ (the "hidden state" space) and Ω (the "observed state" space). Let $(\hat{\Omega}^n, \mu)$ be a probability space, where μ is a Markov measure with transition kernels $p_i(\cdot\,|\,\cdot)$. Thus for $\hat{x} \in \hat{\Omega}^n$, we have

$$\mu(\hat{x}) = p_0(\hat{x}_1)\prod_{k=1}^{n-1} p_k(\hat{x}_{k+1}\,|\,\hat{x}_k).$$

Suppose $(\hat{\Omega}^n \times \Omega^n, \nu)$ is a probability space whose measure ν is defined by

$$\nu(\hat{x}, x) = \mu(\hat{x})\prod_{\ell=1}^{n} q_\ell(x_\ell\,|\,\hat{x}_\ell), \tag{4.6}$$

where $q_\ell(\cdot\,|\,\hat{x})$ is a probability measure on Ω for each $\hat{x} \in \hat{\Omega}$ and $1 \le \ell \le n$. On this product space we define the random process $(\hat{X}_i, X_i)_{1\le i\le n}$, which is clearly Markov since

$$\begin{aligned}\mathbf{P}\Big\{(\hat{X}_{i+1}, X_{i+1}) = (\hat{x}, x)\,|\,(\hat{X}_1^i, X_1^i) = (\hat{y}, y)\Big\} &= p_i(\hat{x}\,|\,\hat{y}_i)q_{i+1}(x\,|\,\hat{x}) \\ &= \mathbf{P}\Big\{(\hat{X}_{i+1}, X_{i+1}) = (\hat{x}, x)\,|\,(\hat{X}_i, X_i) = (\hat{y}_i, y_i)\Big\}.\end{aligned}$$

The (marginal) projection of $(\hat{X}_i, X_i)$ onto X_i results in a random process on the probability space (Ω^n, ρ), where

$$\rho(x) = \mathbf{P}\{X = x\} = \sum_{\hat{x}\in\hat{\Omega}^n} \nu(\hat{x}, x). \tag{4.7}$$

The random process $(X_i)_{1\le i\le n}$ (or measure ρ) on Ω^n is called a *hidden Markov* process (resp., measure); it is well known that (X_i) need not be Markov to any order[1]. We will refer to $(\hat{X}_i)$ as the *underlying* process; it is Markov by construction.

[1] One can easily construct a hidden Markov process over $\hat{\Omega} = \{0, 1, 2\}$ and $\Omega = \{a, b\}$ where, with probability 1, consecutive runs of b will have even length. Such a process cannot be Markov.

Theorem 4.1.4. *Let $(X_i)_{1\le i\le n}$ be a hidden Markov process, whose underlying process $(\hat{X}_i)_{1\le i\le n}$ is defined by the transition kernels $p_i(\cdot\,|\,\cdot)$. Define the k^{th}* contraction coefficient *θ_k by*

$$\theta_k \;=\; \sup_{\hat{x},\hat{x}'\in\hat{\Omega}} \big\|p_k(\cdot\,|\,\hat{x}) - p_k(\cdot\,|\,\hat{x}')\big\|_{\text{TV}}\,.$$

Then for the hidden Markov process X, we have

$$\bar{\eta}_{ij} \;\le\; \theta_i\theta_{i+1}\cdots\theta_{j-1},$$

for $1\le i<j\le n$.

Since the calculation is notationally intensive, we emphasize readability, sometimes at the slight expense of formalistic precision. We will consistently distinguish between hidden and observed state sequences, indicating the former with a $\hat{}$.

As usual, sums range over the entire space of the summation variable; thus $\sum_{x_i^j} f(x_i^j)$ stands for $\sum_{x_i^j\in\Omega^{j-i+1}} f(x_i^j)$ with an analogous convention for $\sum_{\hat{x}_i^j} f(\hat{x}_i^j)$.

The probability operator $\mathbf{P}\{\cdot\}$ is defined with respect to (Ω^n,ρ) whose measure ρ is given in (4.7). Lastly, we use the shorthand

$$\begin{aligned}
\mu(\hat{u}_k^\ell) &= p_0(\hat{u}_k)^{\mathbb{1}_{\{k=1\}}}\prod_{t=k}^{\ell-1} p_t(\hat{u}_{t+1}\,|\,\hat{u}_t)\\
\nu(u_k^\ell\,|\,\hat{u}_k^\ell) &= \prod_{t=k}^{\ell} q_t(u_t\,|\,\hat{u}_t)\\
\rho(u_k^\ell) &= \mathbf{P}\Big\{X_k^\ell = u_k^\ell\Big\}.
\end{aligned}$$

The proof of Theorem 4.1.4 is elementary – it basically amounts to careful bookkeeping of summation indices, rearrangement of sums, and probabilities marginalizing to 1. As in the ordinary Markov case in §4.1.2, the Markov contraction Lemma (4.1.1) plays a central role.

Proof of Theorem 4.1.4. For $1\le i<j\le n$, $y_1^{i-1}\in\Omega^{i-1}$ and $w_i,w_i'\in\Omega$, we expand

$$\begin{aligned}
\eta_{ij}(y_1^{i-1},w_i,w_i') &= \tfrac{1}{2}\sum_{x_j^n}\big|\mathbf{P}\big\{X_j^n = x_j^n\,|\,X_1^i = [y_1^{i-1}\,w_i]\big\} - \mathbf{P}\big\{X_j^n = x_j^n\,|\,X_1^i = [y_1^{i-1}\,w_i']\big\}\big|\\
&= \tfrac{1}{2}\sum_{x_j^n}\Bigg|\sum_{z_{i+1}^{j-1}}\Big(\mathbf{P}\Big\{X_{i+1}^n = [z_{i+1}^{j-1}\,x_j^n]\,|\,X_1^i = [y_1^{i-1}\,w_i]\Big\}\\
&\qquad\qquad -\mathbf{P}\Big\{X_{i+1}^n = [z_{i+1}^{j-1}\,x_j^n]\,|\,X_1^i = [y_1^{i-1}\,w_i']\Big\}\Big)\Bigg|\\
&= \tfrac{1}{2}\sum_{x_j^n}\Bigg|\sum_{z_{i+1}^{j-1}}\sum_{\hat{s}_1^n}\mu(\hat{s}_1^n)\left(\frac{\nu([y_1^{i-1}\,w_i\,z_{i+1}^{j-1}\,x_j^n]\,|\,\hat{s}_1^n)}{\rho([y_1^{i-1}\,w_i])} - \frac{\nu([y_1^{i-1}\,w_i'\,z_{i+1}^{j-1}\,x_j^n]\,|\,\hat{s}_1^n)}{\rho([y_1^{i-1}\,w_i'])}\right)\Bigg|\\
&= \tfrac{1}{2}\sum_{x_j^n}\Bigg|\sum_{z_{i+1}^{j-1}}\sum_{\hat{y}_1^i}\sum_{\hat{z}_{i+1}^{j-1}}\sum_{\hat{x}_j^n}\mu([\hat{y}_1^i\,\hat{z}_{i+1}^{j-1}\,\hat{x}_j^n])\nu(x_j^n\,|\,\hat{x}_j^n)\nu(z_{i+1}^{j-1}\,|\,\hat{z}_{i+1}^{j-1})\nu(y_1^{i-1}\,|\,\hat{y}_1^{i-1})\delta(\hat{y}_i)\Bigg|,
\end{aligned}$$

where

$$\delta(\hat{y}_i) = \frac{q_i(w_i \,|\, \hat{y}_i)}{\rho([y_1^{i-1}\, w_i])} - \frac{q_i(w_i' \,|\, \hat{y}_i)}{\rho([y_1^{i-1}\, w_i'])}.$$

Since $\bigl|\sum_{ij} a_i b_j\bigr| \le \sum_i a_i \bigl|\sum_j b_j\bigr|$ for $a_i \ge 0$ and $b_i \in \mathbb{R}$, we may bound

$$\begin{aligned} \eta_{ij}(y_1^{i-1}, w_i, w_i') &\le \tfrac{1}{2} \sum_{\hat{x}_j^n} \sum_{x_j^n} \mu(\hat{x}_j^n) \nu(x_j^n \,|\, \hat{x}_j^n) \bigl|\zeta(\hat{x}_j)\bigr| \qquad (4.8) \\ &= \tfrac{1}{2} \sum_{\hat{x}_j^n} \mu(\hat{x}_j^n) \bigl|\zeta(\hat{x}_j)\bigr|, \qquad (4.9) \end{aligned}$$

where

$$\begin{aligned} \zeta(\hat{x}_j) &= \sum_{z_{i+1}^{j-1}} \sum_{\hat{z}_{i+1}^{j-1}} \sum_{\hat{y}_1^i} \mu([\hat{y}_1^i\, \hat{z}_{i+1}^{j-1}\, \hat{x}_j]) \nu(y_1^{i-1} \,|\, \hat{y}_1^{i-1}) \nu(z_{i+1}^{j-1} \,|\, \hat{z}_{i+1}^{j-1}) \delta(\hat{y}_i) \\ &= \sum_{\hat{z}_{i+1}^{j-1}} \sum_{\hat{y}_1^i} \mu([\hat{y}_1^i\, \hat{z}_{i+1}^{j-1}\, \hat{x}_j]) \nu(y_1^{i-1} \,|\, \hat{y}_1^{i-1}) \delta(\hat{y}_i). \end{aligned}$$

Define the vector $\mathbf{h} \in \mathbb{R}^{\hat{\Omega}}$ by

$$\mathbf{h}_{\hat{v}} = \delta(\hat{v}) \sum_{\hat{y}_1^{i-1}} \mu([\hat{y}_1^{i-1}\, \hat{v}])\, \nu(y_1^{i-1} \,|\, \hat{y}_1^{i-1}). \qquad (4.10)$$

Then

$$\zeta(\hat{x}_j) = \sum_{\hat{z}_{i+1}^{j-1}} \sum_{\hat{y}_i} \mu([\hat{y}_i\, \hat{z}_{i+1}^{j-1}\, \hat{x}_j]) \mathbf{h}_{\hat{y}_i}.$$

Define the matrix $A^{(k)} \in \mathbb{R}^{\hat{\Omega} \times \hat{\Omega}}$ by $A^{(k)}_{\hat{u},\hat{v}} = p_k(\hat{u} \,|\, \hat{v})$, for $1 \le k < n$. With this notation, we have $\zeta(\hat{x}_j) = \mathbf{z}_{\hat{x}_j}$, where $\mathbf{z} \in \mathbb{R}^{\hat{\Omega}}$ is given by

$$\mathbf{z} = A^{(j-1)} A^{(j-2)} \cdots A^{(i+1)} A^{(i)} \mathbf{h}. \qquad (4.11)$$

In order to apply Lemma 4.1.1 to (4.11), we must verify that

$$\sum_{\hat{v} \in \hat{\Omega}} \mathbf{h}_{\hat{v}} = 0, \qquad \|\mathbf{h}\|_{\mathrm{TV}} \le 1. \qquad (4.12)$$

From (4.10) we have

$$\mathbf{h}_{\hat{v}} = \left(\frac{q_i(w_i \,|\, \hat{v})}{\rho([y_1^{i-1}\, w_i])} - \frac{q_i(w_i' \,|\, \hat{v})}{\rho([y_1^{i-1}\, w_i'])} \right) \sum_{\hat{y}_1^{i-1}} \mu([\hat{y}_1^{i-1}\, \hat{v}])\, \nu(y_1^{i-1} \,|\, \hat{y}_1^{i-1}).$$

Summing over $\hat{v}$, we get

$$\begin{aligned} \sum_{\hat{v} \in \hat{\Omega}} \left(\frac{q_i(w_i \,|\, \hat{v})}{\rho([y_1^{i-1}\, w_i])} \right) \sum_{\hat{y}_1^{i-1}} \mu([\hat{y}_1^{i-1}\, \hat{v}])\, \nu(y_1^{i-1} \,|\, \hat{y}_1^{i-1}) &= \frac{1}{\mathbf{P}\{X_1^i = [y_1^{i-1}\, w_i]\}} \sum_{\hat{y}_1^i} \mu(\hat{y}_1^i)\, \nu([y_1^{i-1}\, w_i] \,|\, \hat{y}_1^i) \\ &= 1; \end{aligned}$$

an analogous identity holds for the $\frac{q_i(w_i' \,|\, \hat{y}_i)}{\rho([y_1^{i-1}\, w_i'])}$ term, which proves (4.12).

Therefore, combining (4.9), (4.11), and Lemma 4.1.1, we have

$$\begin{aligned}
\eta_{ij}(y_1^{i-1}, w_i, w_i') &\leq \tfrac{1}{2} \sum_{\hat{x}_j^n} \mu(\hat{x}_j^n) \big|\mathbf{z}_{\hat{x}_j}\big| \\
&= \tfrac{1}{2} \sum_{\hat{x}_j} \big|\mathbf{z}_{\hat{x}_j}\big| \sum_{\hat{x}_{j+1}^n} \mu(\hat{x}_j^n) \\
&= \|\mathbf{z}\|_{\mathrm{TV}} \\
&\leq \theta_i \theta_{i+1} \cdots \theta_{j-1}.
\end{aligned}$$

□

Observe that the η-mixing coefficients of a hidden Markov chain are bounded by those of the underlying Markov one. One might thus be tempted to pronounce Theorem 4.1.4 as "obvious" in retrospect, based on the intuition that the observed sequence X_i is an independent process conditioned the hidden sequence $\hat{X}_i$. Thus, the reasoning might go, all the dependence structure is contained in $\hat{X}_i$, and it is not surprising that the underlying process alone suffices to bound $\bar{\eta}_{ij}$ – which, after all, is a measure of the dependence in the process.

Such an intuition, however, would be wrong, as it fails to carry over to the case where the underlying process is not Markov. As a numerical example, take $n = 4$, $\hat{\Omega} = \Omega = \{0, 1\}$ and define the probability measure μ on $\hat{\Omega}^4$ as given in Figure 4.1. Define the conditional probability

$$q(x \,|\, \hat{x}) = \tfrac{1}{4} \mathbb{1}_{\{x=\hat{x}\}} + \tfrac{3}{4} \mathbb{1}_{\{x\neq\hat{x}\}}.$$

Together, μ and q define the measure ρ on Ω^4:

$$\rho(x) = \sum_{\hat{x} \in \hat{\Omega}^4} \mu(\hat{x}) \prod_{\ell=1}^{4} q(x_\ell \,|\, \hat{x}_\ell).$$

Associate to $(\hat{\Omega}^4, \mu)$ the "hidden" process $(\hat{X}_i)_1^4$ and to (Ω^4, ρ) the "observed" process $(X_i)_1^4$. A straightforward numerical computation (whose explicit steps are given in the proof of Theorem 4.1.4) shows that the values of μ can be chosen so that $\bar{\eta}_{24}(X) > 0.06$ while $\bar{\eta}_{24}(\hat{X})$ is arbitrarily small.

Thus one cannot, in general, bound $\bar{\eta}_{ij}(X)$ by $c\bar{\eta}_{ij}(\hat{X})$ for some universal constant c; we were rather fortunate to be able to do so in the hidden Markov case.

4.1.5 Markov trees

The material in this section is taken almost entirely from [34]. We begin by defining some notation specific to this section. A collection of variables may be indexed by subset: if $I = \{i_1, i_2, \ldots, i_m\}$ then we write $x_I \equiv x[I] = \{x_{i_1}, x_{i_2}, \ldots, x_{i_m}\}$; we will write x_I and $x[I]$ interchangeably, as dictated by convenience. To avoid cumbersome subscripts, we will also occasionally use the bracket notation for vector components. Thus, $\mathbf{u} \in \mathbb{R}^{\Omega^I}$, then

$$\mathbf{u}_{x_I} \equiv \mathbf{u}_{x[I]} \equiv \mathbf{u}[x[I]] = \mathbf{u}_{(x_{i_1}, x_{i_2}, \ldots, x_{i_m})} \in \mathbb{R}$$

$\hat{x}_1^n$	$\mu(\hat{x}_1^n)$
0000	0.000000
0001	0.000000
0010	0.288413
0011	0.000000
0100	0.000000
0101	0.000000
0110	0.176290
0111	0.000000
1000	0.000000
1001	0.010514
1010	0.000000
1011	0.139447
1100	0.000000
1101	0.024783
1110	0.000000
1111	0.360553

Figure 4.1: The numerical values of μ on $\hat{\Omega}^4$

for each $x[I] \in \Omega^I$. A similar bracket notation will apply for matrices. If A is a matrix then $A_{*,j} = A[*, j]$ will denote its j^{th} column.

We will use $|\cdot|$ to denote set cardinalities. Sums will range over the entire space of the summation variable; thus $\sum_{x_i^j} f(x_i^j)$ stands for $\sum_{x_i^j \in \Omega^{j-i+1}} f(x_i^j)$, and $\sum_{x[I]} f(x[I])$ is shorthand for $\sum_{x[I] \in \Omega^I} f(x[I])$.

We will write $[n]$ for the set $\{1, \ldots, n\}$. Anytime $\|\cdot\|$ appears in this section without a subscript, it will always denote the total variation norm $\|\cdot\|_{\text{TV}}$.

If $G = (V, E)$ is a graph, we will frequently abuse notation and write $u \in G$ instead of $u \in V$, blurring the distinction between a graph and its vertex set. This notation will carry over to set-theoretic operations ($G = G_1 \cap G_2$) and indexing of variables (e.g., X_G).

Graph-theoretic preliminaries

Consider a directed acyclic graph $G = (V, E)$, and define a partial order $\prec_G$ on G by the transitive closure of the relation

$$u \prec_G v \qquad \text{if} \qquad (u, v) \in E.$$

We define the *parents* and *children* of $v \in V$ in the natural way:

$$\text{parents}(v) = \{u \in V : (u, v) \in E\}$$

and

$$\text{children}(v) = \{w \in V : (v, w) \in E\}.$$

If G is connected and each $v \in V$ has at most one parent, G is called a *(directed) tree*. In a tree, whenever $u \prec_G v$ there is a unique directed path from u to v. A tree T always has a unique

minimal (w.r.t. $\prec_T$) element $r_0 \in V$, called its *root*. Thus, for every $v \in V$ there is a unique directed path $r_0 \prec_T r_1 \prec_T \ldots \prec_T r_d = v$; define the *depth* of v, $\mathrm{dep}_T(v) = d$, to be the length (i.e., number of edges) of this path. Note that $\mathrm{dep}_T(r_0) = 0$. We define the depth of the tree by $\mathrm{dep}(T) = \sup_{v \in T} \mathrm{dep}_T(v)$.

For $d = 0, 1, \ldots$ define the d^{th} *level* of the tree T by

$$\mathrm{lev}_T(d) = \{v \in V : \mathrm{dep}_T(v) = d\};$$

note that the levels induce a disjoint partition on V:

$$V = \bigcup_{d=1}^{\mathrm{dep}(T)} \mathrm{lev}_T(d).$$

We define the *width*[2] of a tree as the greatest number of nodes in any level:

$$\mathrm{wid}(T) = \sup_{1 \le d \le \mathrm{dep}(T)} |\mathrm{lev}_T(d)|. \tag{4.13}$$

We will consistently take $|V| = n$ for finite V. An ordering $J : V \to \mathbb{N}$ of the nodes is said to be *breadth-first* if

$$\mathrm{dep}_T(u) < \mathrm{dep}_T(v) \Longrightarrow J(u) < J(v). \tag{4.14}$$

Since every finite directed tree $T = (V, E)$ has some breadth-first ordering,[3] we will henceforth blur the distinction between $v \in V$ and $J(v)$, simply taking $V = [n]$ (or $V = \mathbb{N}$) and assuming that $\mathrm{dep}_T(u) < \mathrm{dep}_T(v) \Rightarrow u < v$ holds. This will allow us to write Ω^V simply as Ω^n for any set Ω.

Note that we have two orders on V: the partial order $\prec_T$, induced by the tree topology, and the total order $<$, given by the breadth-first enumeration. Observe that $i \prec_T j$ implies $i < j$ but not vice versa.

If $T = (V, E)$ is a tree and $u \in V$, we define the *subtree* induced by u, $T_u = (V_u, E_u)$ by $V_u = \{v \in V : u \preceq_T v\}$, $E_u = \{(v, w) \in E : v, w \in V_u\}$.

Markov tree measure

If Ω is a finite set, a *Markov tree measure* μ is defined on Ω^n by a tree $T = (V, E)$ and transition kernels p_0, $\{p_{ij}(\cdot \,|\, \cdot)\}_{(i,j) \in E}$. Continuing our convention above, we have a breadth-first order $<$ and the total order $\prec_T$ on V, and take $V = \{1, \ldots, n\}$. Together, the topology of T and the transition kernels determine the measure μ on Ω^n:

$$\mu(x) = p_0(x_1) \prod_{(i,j) \in E} p_{ij}(x_j \,|\, x_i). \tag{4.15}$$

A measure on Ω^n satisfying (4.15) for some T and $\{p_{ij}\}$ is said to be *compatible* with tree T; a measure is a Markov tree measure if it is compatible with some tree.

[2] Note that this definition is nonstandard.

[3] One can easily construct a breadth-first ordering on a given tree by ordering the nodes arbitrarily within each level and listing the levels in ascending order: $\mathrm{lev}_T(1), \mathrm{lev}_T(2), \ldots$.

Suppose Ω is a finite set and $(X_i)_{i\in\mathbb{N}}$, $X_i \in \Omega$ is a random process defined on $(\Omega^{\mathbb{N}}, \mathbf{P})$. If for each $n > 0$ there is a tree $T^{(n)} = ([n], E^{(n)})$ and a Markov tree measure μ_n compatible with $T^{(n)}$ such that for all $x \in \Omega^n$ we have

$$\mathbf{P}\{X_1^n = x\} = \mu_n(x)$$

then we call X a *Markov tree process*. The trees $\{T^{(n)}\}$ are easily seen to be consistent in the sense that $T^{(n)}$ is an induced subgraph of $T^{(n+1)}$. So corresponding to any Markov tree process is the unique infinite tree $T = (\mathbb{N}, E)$. The uniqueness of T is easy to see, since for $v > 1$, the parent of v is the smallest $u \in \mathbb{N}$ such that

$$\mathbf{P}\{X_v = x_v \,|\, X_1^u \;=\; x_1^u\} = \mathbf{P}\{X_v = x_v \,|\, X_u = x_u\}\,;$$

thus $\mathbf{P}$ determines the topology of T.

It is straightforward to verify that a Markov tree process $\{X_v\}_{v\in T}$ compatible with tree T has the following *Markov property*: if v and v' are children of u in T, then

$$\mathbf{P}\big\{X_{T_v} = x, X_{T_{v'}} = x' \,|\, X_u = y\big\} = \mathbf{P}\{X_{T_v} = x \,|\, X_u = y\}\,\mathbf{P}\big\{X_{T_{v'}} = x' \,|\, X_u = y\big\}\,.$$

In other words, the subtrees induced by the children are conditionally independent given the parent; this follows directly from the definition of the Markov tree measure in (4.15).

Statement of result

Theorem 4.1.5. *Let Ω be a finite set and let $(X_i)_{1\le i\le n}$, $X_i \in \Omega$ be a Markov tree process, defined by a tree $T = (V, E)$ and transition kernels p_0, $\{p_{uv}(\cdot\,|\,\cdot)\}_{(u,v)\in E}$. Define the (u, v)-* contraction coefficient *θ_{uv} by*

$$\theta_{uv} \;=\; \max_{y,y'\in\Omega} \big\|p_{uv}(\cdot\,|\,y) - p_{uv}(\cdot\,|\,y')\big\|_{\mathrm{TV}}\,. \tag{4.16}$$

Suppose $\max_{(u,v)\in E}\theta_{uv} \le \theta < 1$ for some θ and $\mathrm{wid}(T) \le L$. Then for the Markov tree process X we have

$$\bar{\eta}_{ij} \;\le\; \big(1 - (1-\theta)^L\big)^{\lfloor (j-i)/L\rfloor} \tag{4.17}$$

for $1 \le i < j \le n$.

To cast (4.17) in more usable form, we first note that for $L \in \mathbb{N}$ and $k \in \mathbb{N}$, if $k \ge L$ then

$$\left\lfloor \frac{k}{L} \right\rfloor \ge \frac{k}{2L-1} \tag{4.18}$$

(we omit the elementary number-theoretic proof). Using (4.18), we have

$$\bar{\eta}_{ij} \;\le\; \tilde{\theta}^{j-i}, \qquad \text{for } j \ge i + L \tag{4.19}$$

where

$$\tilde{\theta} = (1 - (1-\theta)^L)^{1/(2L-1)}.$$

In the (degenerate) case where the Markov tree is a chain, we have $L = 1$ and therefore $\tilde{\theta} = \theta$; thus we recover the Markov chain concentration results in [37, 45, 60].

Proof of Theorem 4.1.5

The proof of Theorem 4.1.5 is combination of elementary graph theory and tensor algebra. We start with a graph-theoretic lemma:

Lemma 4.1.6. *Let $T = ([n], E)$ be a tree and fix $1 \le i < j \le n$. Suppose $(X_i)_{1\le i\le n}$ is a Markov tree process whose law $\mathbf{P}$ on Ω^n is compatible with T (this notion is defined above). Define the set*

$$T_i^j = T_i \cap \{j, j+1, \ldots, n\},$$

consisting of those nodes in the subtree T_i whose breadth-first numbering does not precede j. Then, for $y \in \Omega^{i-1}$ and $w, w' \in \Omega$, we have

$$\eta_{ij}(y, w, w') \;=\; \begin{cases} 0 & T_i^j = \emptyset \\ \eta_{ij_0}(y, w, w') & \text{otherwise,} \end{cases} \tag{4.20}$$

where j_0 is the minimum (with respect to $<$) element of T_i^j.

Remark 4.1.7. This lemma tells us that when computing η_{ij} it is sufficient to restrict our attention to the subtree induced by i.

Proof. The case $j \in T_i$ implies $j_0 = j$ and is trivial; thus we assume $j \notin T_i$. In this case, the subtrees T_i and T_j are disjoint. Putting $\bar{T}_i = T_i \setminus \{i\}$, we have by the Markov property,

$$\mathbf{P}\big\{X_{\bar{T}_i} = x_{\bar{T}_i}, X_{T_j} = x_{T_j} \,|\, X_1^i = yw\big\} \;=\; \mathbf{P}\big\{X_{\bar{T}_i} = x_{\bar{T}_i} \,|\, X_i = w\big\}\, \mathbf{P}\big\{X_{T_j} = x_{T_j} \,|\, X_1^{i-1} = y\big\}.$$

Then from (3.13) and (2.1), and by marginalizing out the X_{T_j}, we have

$$\begin{aligned} \eta_{ij}(y, w, w') &= \tfrac{1}{2} \sum_{x_j^n} \big|\mathbf{P}\big\{X_j^n = x_j^n \,|\, X_1^i = yw\big\} - \mathbf{P}\big\{X_j^n = x_j^n \,|\, X_1^i = yw'\big\}\big| \\ &= \tfrac{1}{2} \sum_{x_{T_i^j}} \Big|\mathbf{P}\Big\{X_{T_i^j} = x_{T_i^j} \,|\, X_i = w\Big\} - \mathbf{P}\Big\{X_{T_i^j} = x_{T_i^j} \,|\, X_i = w'\Big\}\Big|. \end{aligned}$$

If $T_i^j = \emptyset$ then obviously $\eta_{ij} = 0$; otherwise, $\eta_{ij} = \eta_{ij_0}$, since j_0 is the "first" element of T_i^j. □

Next we develop some basic results for tensor norms; recall that unless specified otherwise, the norm used in this paper is the total variation norm. If $\mathbf{A}$ is an $M \times N$ column-stochastic matrix: ($\mathbf{A}_{ij} \ge 0$ for $1 \le i \le M$, $1 \le j \le N$ and $\sum_{i=1}^M \mathbf{A}_{ij} = 1$ for all $1 \le j \le N$) and $\mathbf{u} \in \mathbb{R}^N$ is *balanced* in the sense that $\sum_{j=1}^N \mathbf{u}_j = 0$, we have, by Lemma 4.1.1

$$\|\mathbf{A}\mathbf{u}\| \;\le\; \|\mathbf{A}\|\, \|\mathbf{u}\|, \tag{4.21}$$

where

$$\|\mathbf{A}\| \;=\; \max_{1\le j,j'\le N} \big\|\mathbf{A}_{*,j} - \mathbf{A}_{*,j'}\big\|, \tag{4.22}$$

and $\mathbf{A}_{*,j} \equiv \mathbf{A}[\cdot, j]$ denotes the j^{th} column of $\mathbf{A}$. An immediate consequence of (4.21) is that $\|\cdot\|$ satisfies

$$\|\mathbf{A}\mathbf{B}\| \;\le\; \|\mathbf{A}\|\, \|\mathbf{B}\| \tag{4.23}$$

for column-stochastic matrices $\mathbf{A} \in \mathbb{R}^{M\times N}$ and $\mathbf{B} \in \mathbb{R}^{N\times P}$.

Remark 4.1.8. Note that if $\mathbf{A}$ is a column-stochastic matrix then $\|\mathbf{A}\| \leq 1$, and if additionally $\mathbf{u}$ is balanced then $\mathbf{Au}$ is also balanced. $\diamond$

If $\mathbf{u} \in \mathbb{R}^M$ and $\mathbf{v} \in \mathbb{R}^N$, define their tensor product $\mathbf{w} = \mathbf{v} \otimes \mathbf{u}$ by

$$\mathbf{w}_{(i,j)} \quad = \quad \mathbf{u}_i \mathbf{v}_j,$$

where the notation $(\mathbf{v} \otimes \mathbf{u})_{(i,j)}$ is used to distinguish the 2-tensor $\mathbf{w}$ from an $M \times N$ matrix. The tensor $\mathbf{w}$ is a vector in $\mathbb{R}^{MN}$ indexed by pairs $(i,j) \in [M] \times [N]$; its norm is naturally defined to be

$$\|\mathbf{w}\| = \tfrac{1}{2} \sum_{(i,j) \in [M] \times [N]} \left| \mathbf{w}_{(i,j)} \right|. \tag{4.24}$$

To develop a convenient tensor notation, we will fix the index set $V = \{1, \ldots, n\}$. For $I \subset V$, a tensor indexed by I is a vector $\mathbf{u} \in \mathbb{R}^{\Omega^I}$. A special case of such an I-tensor is the product $\mathbf{u} = \bigotimes_{i \in I} \mathbf{v}^{(i)}$, where $\mathbf{v}^{(i)} \in \mathbb{R}^{\Omega}$ and

$$\mathbf{u}[x_I] \quad = \quad \prod_{i \in I} \mathbf{v}^{(i)}[x_i]$$

for each $x_I \in \Omega^I$. To gain more familiarity with the notation, let us write the total variation norm of an I-tensor:

$$\|\mathbf{u}\| \quad = \quad \tfrac{1}{2} \sum_{x_I \in \Omega^I} |\mathbf{u}[x_I]|. \tag{4.25}$$

In order to extend Lemma 2.2.5 to product tensors, we will need to define the function $\alpha_k : \mathbb{R}^k \to \mathbb{R}$ and state some of its properties:

Lemma 4.1.9. *Define $\alpha_k : \mathbb{R}^k \to \mathbb{R}$ recursively as $\alpha_1(x) = x$ and*

$$\alpha_{k+1}(x_1, x_2, \ldots, x_{k+1}) = x_{k+1} + (1 - x_{k+1}) \alpha_k(x_1, x_2, \ldots, x_k). \tag{4.26}$$

Then

(a) *α_k is symmetric in its k arguments, so it is well-defined as a mapping*

$$\alpha : \{x_i : 1 \leq i \leq k\} \mapsto \mathbb{R}$$

from finite real sets to the reals

(b) *α_k takes $[0,1]^k$ to $[0,1]$ and is monotonically increasing in each argument on $[0,1]^k$*

(c) *If $B \subset C \subset [0,1]$ are finite sets then $\alpha(B) \leq \alpha(C)$*

(d) *$\alpha_k(x, x, \ldots, x) = 1 - (1-x)^k$*

(e) *if B is finite and $1 \in B \subset [0,1]$ then $\alpha(B) = 1$.*

(f) *if $B \subset [0,1]$ is a finite set then $\alpha(B) \leq \sum_{x \in B} x$.*

Remark 4.1.10. In light of (a), we will use the notation $\alpha_k(x_1, x_2, \ldots, x_k)$ and $\alpha(\{x_i : 1 \le i \le k\})$ interchangeably, as dictated by convenience.

Proof. Claims (a), (b), (e), (f) are straightforward to verify from the recursive definition of α and induction. Claim (c) follows from (b) since

$$\alpha_{k+1}(x_1, x_2, \ldots, x_k, 0) = \alpha_k(x_1, x_2, \ldots, x_k)$$

and (d) is easily derived from the binomial expansion of $(1-x)^k$. □

The function α_k is the natural generalization of $\alpha_2(x_1, x_2) = x_1 + x_2 - x_1 x_2$ to k variables, and it is what we need for the analog of Lemma 2.2.5 for a product of k tensors:

Corollary 4.1.11. *Let $\{\mathbf{u}^{(i)}\}_{i\in I}$ and $\{\mathbf{v}^{(i)}\}_{i\in I}$ be two sets of tensors and assume that each of $\mathbf{u}^{(i)}, \mathbf{v}^{(i)}$ is a probability measure on Ω. Then we have*

$$\left\| \bigotimes_{i\in I} \mathbf{u}^{(i)} - \bigotimes_{i\in I} \mathbf{v}^{(i)} \right\| \;\le\; \alpha\left\{ \left\| \mathbf{u}^{(i)} - \mathbf{v}^{(i)} \right\| : i \in I \right\}. \tag{4.27}$$

Proof. Pick an $i_0 \in I$ and let $\mathbf{p} = \mathbf{u}^{(i_0)}$, $\mathbf{q} = \mathbf{v}^{(i_0)}$,

$$\mathbf{p}' = \bigotimes_{i_0 \neq i \in I} \mathbf{u}^{(i)}, \qquad \mathbf{q}' = \bigotimes_{i_0 \neq i \in I} \mathbf{v}^{(i)}.$$

Apply Lemma 2.2.5 to $\|\mathbf{p}\otimes\mathbf{q} - \mathbf{p}'\otimes\mathbf{q}'\|$ and proceed by induction. □

Our final generalization concerns linear operators over I-tensors. An I, J-matrix $\mathbf{A}$ has dimensions $|\Omega^J| \times |\Omega^I|$ and takes an I-tensor $\mathbf{u}$ to a J-tensor $\mathbf{v}$: for each $y_J \in \Omega^J$, we have

$$\mathbf{v}[y_J] \;=\; \sum_{x_I \in \Omega^I} \mathbf{A}[y_J, x_I]\mathbf{u}[x_I], \tag{4.28}$$

which we write as $\mathbf{Au} = \mathbf{v}$. If $\mathbf{A}$ is an I, J-matrix and $\mathbf{B}$ is a J, K-matrix, the matrix product $\mathbf{BA}$ is defined analogously to (4.28).

As a special case, an I, J-matrix might factorize as a tensor product of $|\Omega| \times |\Omega|$ matrices $\mathbf{A}^{(i,j)} \in \mathbb{R}^{\Omega\times\Omega}$. We will write such a factorization in terms of a bipartite graph[4] $G = (I + J, E)$, where $E \subset I \times J$ and the factors $\mathbf{A}^{(i,j)}$ are indexed by $(i, j) \in E$:

$$\mathbf{A} = \bigotimes_{(i,j)\in E} \mathbf{A}^{(i,j)}, \tag{4.29}$$

where

$$\mathbf{A}[y_J, x_I] = \prod_{(i,j)\in E} \mathbf{A}^{(i,j)}_{y_j, x_i}$$

for all $x_I \in \Omega^I$ and $y_J \in \Omega^J$. The norm of an I, J-matrix is a natural generalization of the matrix norm defined in (4.22):

$$\|\mathbf{A}\| = \max_{x_I, x'_I \in \Omega^I} \left\| \mathbf{A}[\cdot, x_I] - \mathbf{A}[\cdot, x'_I] \right\| \tag{4.30}$$

[4] Our notation for bipartite graphs is standard; it is equivalent to $G = (I \cup J, E)$ where I and J are always assumed to be disjoint.

where $\mathbf{u} = \mathbf{A}[\cdot, x_I]$ is the J-tensor given by

$$\mathbf{u}[y_J] = \mathbf{A}[y_J, x_I];$$

(4.30) is well-defined via the tensor norm in (4.25). Since I, J matrices act on I-tensors by ordinary matrix multiplication, $\|\mathbf{Au}\| \leq \|\mathbf{A}\| \, \|\mathbf{u}\|$ continues to hold when $\mathbf{A}$ is a column-stochastic I, J-matrix and $\mathbf{u}$ is a balanced I-tensor; if, additionally, $\mathbf{B}$ is a column-stochastic J, K-matrix, $\|\mathbf{BA}\| \leq \|\mathbf{B}\| \, \|\mathbf{A}\|$ also holds. Likewise, since another way of writing (4.29) is

$$\mathbf{A}[\cdot, x_I] = \bigotimes_{(i,j)\in E} \mathbf{A}^{(i,j)}[\cdot, x_i],$$

Corollary 4.1.11 extends to tensor products of matrices:

Lemma 4.1.12. *Fix index sets I, J and a bipartite graph $(I + J, E)$. Let $\left\{\mathbf{A}^{(i,j)}\right\}_{(i,j)\in E}$ be a collection of column-stochastic $|\Omega| \times |\Omega|$ matrices, whose tensor product is the I, J matrix*

$$\mathbf{A} = \bigotimes_{(i,j)\in E} \mathbf{A}^{(i,j)}.$$

Then

$$\|\mathbf{A}\| \quad \leq \quad \alpha\left\{\left\|\mathbf{A}^{(i,j)}\right\| : (i,j) \in E\right\}.$$

We are now in a position to state the main technical lemma, from which Theorem 4.1.5 will follow straightforwardly:

Lemma 4.1.13. *Let Ω be a finite set and let $(X_i)_{1\leq i\leq n}$, $X_i \in \Omega$ be a Markov tree process, defined by a tree $T = (V, E)$ and transition kernels p_0, $\{p_{uv}(\cdot \,|\, \cdot)\}_{(u,v)\in E}$. Let the (u, v)-contraction coefficient θ_{uv} be as defined in (4.16).*

Fix $1 \leq i < j \leq n$ and let $j_0 = j_0(i, j)$ be as defined in Lemma 4.1.6 (we are assuming its existence, for otherwise $\bar{\eta}_{ij} = 0$). Then we have

$$\bar{\eta}_{ij} \quad \leq \quad \prod_{d=\mathrm{dep}_T(i)+1}^{\mathrm{dep}_T(j_0)} \alpha\left\{\theta_{uv} : v \in \mathrm{lev}_T(d)\right\}. \tag{4.31}$$

Proof. For $y \in \Omega^{i-1}$ and $w, w' \in \Omega$, we have

$$\eta_{ij}(y, w, w') \quad = \quad \tfrac{1}{2}\sum_{x_j^n} \left|\mathbf{P}\left\{X_j^n = x_j^n \,|\, X_1^i = yw\right\} - \mathbf{P}\left\{X_j^n = x_j^n \,|\, X_1^i = yw'\right\}\right| \tag{4.32}$$

$$= \quad \tfrac{1}{2}\sum_{x_j^n} \left| \sum_{z_{i+1}^{j-1}} \Big(\mathbf{P}\Big\{X_{i+1}^n = z_{i+1}^{j-1} x_j^n \,|\, X_1^i = yw\Big\} \right.$$

$$\left. -\mathbf{P}\Big\{X_{i+1}^n = [z_{i+1}^{j-1}\, x_j^n] \,|\, X_1^i = yw'\Big\} \Big) \right|. \tag{4.33}$$

Let T_i be the subtree induced by i and

$$Z = T_i \cap \{i+1, \ldots, j_0 - 1\} \qquad \text{and} \qquad C = \{v \in T_i : (u,v) \in E, u < j_0, v \geq j_0\}. \tag{4.34}$$

Then by Lemma 4.1.6 and the Markov property, we get

$$\eta_{ij}(y, w, w') \;=\; \tfrac{1}{2} \sum_{x[C]} \left| \sum_{x[Z]} \Big(\mathbf{P}\{X[C \cup Z] = x[C \cup Z] \,|\, X_i = w\} - \mathbf{P}\big\{X[C \cup Z] = x[C \cup Z] \,|\, X_i = w'\big\} \Big) \right| \tag{4.35}$$

(the sum indexed by $\{j_0, \ldots, n\} \setminus C$ marginalizes out).

Define $D = \{d_k : k = 0, \ldots, |D|\}$ with $d_0 = \mathrm{dep}_T(i)$, $d_{|D|} = \mathrm{dep}_T(j_0)$ and $d_{k+1} = d_k + 1$ for $0 \leq k < |D|$. For $d \in D$, let $I_d = T_i \cap \mathrm{lev}_T(d)$ and $G_d = (I_{d-1} + I_d, E_d)$ be the bipartite graph consisting of the nodes in I_{d-1} and I_d, and the edges in E joining them (note that $I_{d_0} = \{i\}$).

For $(u,v) \in E$, let $\mathbf{A}^{(u,v)}$ be the $|\Omega| \times |\Omega|$ matrix given by

$$\mathbf{A}^{(u,v)}_{x,x'} = p_{uv}(x \,|\, x')$$

and note that $\left\|\mathbf{A}^{(u,v)}\right\| = \theta_{uv}$. Then by the Markov property, for each $z[I_d] \in \Omega^{I_d}$ and $x[I_{d-1}] \in \Omega^{I_{d-1}}$, $d \in D \setminus \{d_0\}$, we have

$$\mathbf{P}\big\{X_{I_d} = z_{I_d} \,|\, X_{I_{d-1}} = x_{I_{d-1}}\big\} \;=\; \mathbf{A}^{(d)}[z_{I_d}, x_{I_{d-1}}],$$

where

$$\mathbf{A}^{(d)} \;=\; \bigotimes_{(u,v) \in E_d} \mathbf{A}^{(u,v)}.$$

Likewise, for $d \in D \setminus \{d_0\}$,

$$\begin{aligned} \mathbf{P}\{X_{I_d} = x_{I_d} \,|\, X_i = w\} \;&=\; \sum_{x'_{I_1}} \sum_{x''_{I_2}} \cdots \sum_{x^{(d-1)}_{I_{d-1}}} \\ &\quad \mathbf{P}\big\{X_{I_1} = x'_{I_1} \,|\, X_i = w\big\}\, \mathbf{P}\big\{X_{I_2} = x''_{I_2} \,|\, X_{I_1} = x'_{I_1}\big\} \cdots \\ &\quad \mathbf{P}\Big\{X_{I_d} = x_{I_d} \,|\, X_{I_{d-1}} = x^{(d-1)}_{I_{d-1}}\Big\} \\ &=\; (\mathbf{A}^{(d)} \mathbf{A}^{(d-1)} \cdots \mathbf{A}^{(d_1)})[x_{I_d}, w]. \end{aligned} \tag{4.36}$$

Define the (balanced) I_{d_1}-tensor

$$\mathbf{h} = \mathbf{A}^{(d_1)}[\cdot, w] - \mathbf{A}^{(d_1)}[\cdot, w'], \tag{4.37}$$

the $I_{d_{|D|}}$-tensor

$$\mathbf{f} = \mathbf{A}^{(d_{|D|})} \mathbf{A}^{(d_{|D|-1})} \cdots \mathbf{A}^{(d_2)} \mathbf{h}, \tag{4.38}$$

and $C_0, C_1, Z_0 \subset \{1, \ldots, n\}$:

$$C_0 = C \cap I_{\mathrm{dep}_T(j_0)}, \qquad C_1 = C \setminus C_0, \qquad Z_0 = I_{\mathrm{dep}_T(j_0)} \setminus C_0, \tag{4.39}$$

where C and Z are defined in (4.34). For readability we will write $\mathbf{P}(x_U \,|\, \cdot)$ instead of $\mathbf{P}\{X_U = x_U \,|\, \cdot\}$ below; no ambiguity should arise. Combining (4.35) and (4.36), we have

$$\begin{aligned}
\eta_{ij}(y, w, w') &= \tfrac{1}{2} \sum_{x_C} \Big| \sum_{x_Z} \big(\mathbf{P}(x[C \cup Z] \,|\, X_i = w) - \mathbf{P}(x[C \cup Z] \,|\, X_i = w') \big) \Big| && (4.40) \\
&= \tfrac{1}{2} \sum_{x_{C_0}} \sum_{x_{C_1}} \left| \sum_{x_{Z_0}} \mathbf{P}(x[C_1] \,|\, x[Z_0]) \mathbf{f}[C_0 \cup Z_0] \right| && (4.41) \\
&= \|\mathbf{B}\mathbf{f}\| && (4.42)
\end{aligned}$$

where $\mathbf{B}$ is the $|\Omega^{C_0 \cup C_1}| \times |\Omega^{C_0 \cup Z_0}|$ column-stochastic matrix given by

$$\mathbf{B}[x_{C_0} \cup x_{C_1}, x'_{C_0} \cup x_{Z_0}] = \mathbb{1}_{\{x_{C_0} = x'_{C_0}\}} \mathbf{P}(x_{C_1} \,|\, x_{Z_0})$$

with the convention that $\mathbf{P}(x_{C_1} \,|\, x_{Z_0}) = 1$ if either of $\{Z_0, C_1\}$ is empty. The claim now follows by reading off the results previously obtained:

$$\begin{aligned}
\|\mathbf{B}\mathbf{f}\| &\leq \|\mathbf{B}\| \, \|\mathbf{f}\| && \text{Eq. (2.1)} \\
&\leq \|\mathbf{f}\| && \text{Remark 4.1.8} \\
&\leq \|\mathbf{h}\| \prod_{k=2}^{|D|} \left\| \mathbf{A}^{(d_k)} \right\| && \text{Eqs. (4.23,4.38)} \\
&\leq \prod_{k=1}^{|D|} \alpha\{ \left\| \mathbf{A}^{(u,v)} \right\| : (u, v) \in E_{d_k} \} && \text{Lemma 4.1.12.}
\end{aligned}$$

□

Proof of Theorem 4.1.5. We will borrow the definitions from the proof of Lemma 4.1.13. To upper-bound $\bar{\eta}_{ij}$ we first bound $\alpha\{ \left\| \mathbf{A}^{(u,v)} \right\| : (u, v) \in E_{d_k} \}$. Since

$$|E_{d_k}| \leq \mathrm{wid}(T) \leq L$$

(because every node in I_{d_k} has exactly one parent in $I_{d_{k-1}}$) and

$$\left\| \mathbf{A}^{(u,v)} \right\| = \theta_{uv} \leq \theta < 1,$$

we appeal to Lemma 4.1.9 to obtain

$$\alpha\{ \left\| \mathbf{A}^{(u,v)} \right\| : (u, v) \in E_{d_k} \} \leq 1 - (1 - \theta)^L. \tag{4.43}$$

Now we must lower-bound the quantity $h = \mathrm{dep}_T(j_0) - \mathrm{dep}_T(i)$. Since every level can have up to L nodes, we have

$$j_0 - i \leq hL$$

and so $h \geq \lfloor (j_0 - i)/L \rfloor \geq \lfloor (j - i)/L \rfloor$. □

The calculations in Lemma 4.1.13 yield considerably more information than the simple bound in (4.17). For example, suppose the tree T has levels $\{I_d : d = 0, 1, \ldots\}$ with the property that the levels are growing at most linearly:

$$|I_d| \leq cd$$

for some $c > 0$. Let $d_i = \mathrm{dep}_T(i)$, $d_j = \mathrm{dep}_T(j_0)$, and $h = d_j - d_i$. Then

$$\begin{aligned} j - i \leq j_0 - i &\leq c \sum_{d_i+1}^{d_j} k \\ &= \frac{c}{2}(d_j(d_j+1) - d_i(d_i+1)) \\ &< \frac{c}{2}((d_j+1)^2 - d_i^2) \\ &< \frac{c}{2}(d_i + h + 1)^2 \end{aligned}$$

so

$$h > \sqrt{2(j-i)/c} - d_i - 1,$$

which yields the bound, via Lemma 4.1.9(f),

$$\bar{\eta}_{ij} \leq \prod_{k=1}^{h} \sum_{(u,v)\in E_k} \theta_{uv}. \tag{4.44}$$

Let $\theta_k = \max\{\theta_{uv} : (u,v) \in E_k\}$; then if $ck\theta_k \leq \beta$ holds for some $\beta \in \mathbb{R}$, this becomes

$$\begin{aligned} \bar{\eta}_{ij} &\leq \prod_{k=1}^{h} (ck\theta_k) \\ &< \prod_{k=1}^{\sqrt{2(j-i)/c}-d_i-1} (ck\theta_k) \\ &\leq \beta^{\sqrt{2(j-i)/c}-d_i-1}. \end{aligned} \tag{4.45}$$

This is a non-trivial bound for trees with linearly growing levels: recall that to bound $\|\Delta\|_\infty$ (3.16), we must bound the series

$$\sum_{j=i+1}^{\infty} \bar{\eta}_{ij}.$$

By the limit comparison test with the series $\sum_{j=1}^{\infty} 1/j^2$, we have that

$$\sum_{j=i+1}^{\infty} \beta^{\sqrt{2(j-i)/c}-d_i-1}$$

converges for $\beta < 1$. Similar techniques may be applied when the level growth is bounded by other slowly increasing functions. It is hoped that these techniques will be extended to obtain concentration bounds for larger classes of directed acyclic graphical models.

4.2 Law of large numbers

The material in this section is taken, with minor modifications, from a paper in progress with Anthony Brockwell [36].

Roughly speaking, laws of large numbers assert the convergence of empirical averages to true expectations, and, under appropriate assumptions, ensure that inferences about persistent world phenomena become increasingly more valid as data accumulates. When this convergence is in probability, we have a *weak* law of large numbers (LLN); when the convergence is almost sure we have a *strong* LLN, and the Borel-Cantelli lemma [63] allows one to convert a sufficiently rapidly converging weak LLN into a strong LLN.

We give a brief survey of strong LLNs in [36], with a special emphasis on results for non-independent processes. Even this specialized field has produced a formidable body of work – to do each result justice could easily require a book. We refer the reader to
`http://www.stats.org.uk/law-of-large-numbers/`
in the hope that the list of papers is both comprehensive and regularly updated. Keeping in mind the necessarily limited nature of any such endeavor (i.e., confined to a single thesis chapter), we nevertheless attempt a rough summary of the state of affairs in non-independent strong LLNs.

From Birkhoff's ergodic theorem we get a law of large numbers for ergodic processes; this has been strengthened by Breiman [8] to cover the case where the stationary distribution is singular with respect to the Lebesgue measure. Assumptions of ergodicity are typically too weak to provide a convergence rate – this requires a stronger mixing condition. A classical (and perhaps first of its kind) example of the latter is the paper by Blum, Hanson and Koopmans [26], which proves a strong law of large numbers under a mixing condition known in a modern form as ψ-mixing [7]. This mixing condition guarantees exponentially fast convergence, but the proof does not directly yield rate constants.

For many practical applications, one actually wants to know for how many steps to run an algorithm to achieve a specified accuracy at a given confidence level – and this is precisely the problem we wish to address. Our strong LLN provides a finite-sample bound with readily computable (at least in principle) rate constants. The downside is that we must assume a stronger mixing condition, though it turns out to be quite realistic in many applications [11].

We state our LLN for a real-valued random process. Though all of our results so far have been for finite Ω, they readily generalize to the continuous case, as shown in Chapter 5.1.

Theorem 4.2.1. *Let $(\Omega, \mathscr{B}, \mu)$ be a (positive) Borel measure space. Define the random process X_1^∞ on the measure space $(\Omega^{\mathbb{N}}, \mathscr{B}^{\mathbb{N}}, \mathbf{P})$, and assume that for all $n \geq 1$ we have $\mathbf{P}_n \ll \mu^n$, where $\mathbf{P}_n$ is the marginal distribution on X_1^n and μ^n is the corresponding product measure on $(\Omega^n, \mathscr{B}^n)$. Suppose further that for any measurable $A \subset \Omega$, its empirical measure defined by*

$$\hat{P}_n(A) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \in A\}}$$

has uniformly converging expectation:

$$\lim_{n \to \infty} \left\| \mathbf{E}\hat{P}_n(\cdot) - \nu(\cdot) \right\|_{\mathrm{TV}} \quad \to \quad 0;$$

define $n_0 = n_0(\varepsilon)$ to be such that $\left\| \mathbf{E}\hat{P}_n(\cdot) - \nu(\cdot) \right\|_{\mathrm{TV}} < \varepsilon$ for all $n > n_0(\varepsilon)$.

Then $\hat{P}_n(A)$ converges to $\nu(A)$ almost surely, exponentially fast:

$$\mathbf{P}_n\left\{\left|\hat{P}_n(A) - \nu(A)\right| > t + \varepsilon\right\} \quad \leq \quad 2\exp(-nt^2/2\,\|\Delta_n\|_\infty^2)$$

for all $n > n_0(\varepsilon)$, where Δ_n is the η-mixing matrix defined in (3.15).

Modulo questions regarding the generalization from finite Ω to $\mathbb{R}$ (which are addressed in Chapter 5.1), this result follows directly from Theorem 3.3.1, by observing that the function $\varphi_A : X_1^n \mapsto \mathbb{R}$ defined by $\varphi_A(X_1^n) = \hat{P}_n(A)$ has Lipschitz constant $1/n$. For a recent application of this result in statistics, see Brockwell's forthcoming paper [11] on a Monte Carlo estimator for a particle filter.

4.3 Empirical processes and machine learning

Empirical process theory is concerned with establishing the almost sure convergence of path functionals to their expected values, uniformly over classes of permissible functionals. A classical reference is Pollard's excellent book [57]; for a more recent treatment, focusing on non-iid processes, see [14].

In the simplest setting, we have a set $\mathcal{X}$ (*sample space*) and collection of subsets $\mathcal{C} \subset 2^{\mathcal{X}}$ (*concept class*). If $\mathbf{P} = \bigotimes_{i=1}^{\infty} P$ is a product measure on $\mathcal{X}^{\mathbb{N}}$, we say that $\mathcal{C}$ is a Glivenko-Cantelli class if for all $\varepsilon, \delta > 0$ there is an $m_0 = m_0(\varepsilon, \delta)$ such that for all distributions P on $\mathcal{X}$, we have

$$\sup_{n \geq m_0} P^n \left\{\sup_{A \in \mathcal{C}} \left|\hat{P}_n(A) - P(A)\right| > \varepsilon\right\} \quad < \quad \delta, \tag{4.46}$$

where $\hat{P}_n(A)$ is the empirical measure defined by

$$\hat{P}_n(A) \quad = \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \in A\}}.$$

In general, measurability issues may arise concerning $A \subset \mathcal{X}$ and – more subtly – the possibly uncountable supremum over $\mathcal{C}$ in (4.46); these are addressed in detail in [57], and following Pollard, we call $\mathcal{C}$ *permissible* if it avoids such pathologies.

Machine-learning theorists define *learnability* in essentially the same way, though they use the language of oracles and learners [30]. An *oracle* labels examples $x \in \mathcal{X}$ as "positive" or "negative," depending on their membership in some target $C \in \mathcal{C}$. Given a random P-iid labeled sample X_1^n, $X_i \in \mathcal{X}$, the learner produces a hypothesis $H = H(X_1^n)$, whose *empirical error* $\hat{E}(H)$ is the normalized count of the examples it mislabels and whose true error $E(H) = P(C \Delta H)$ is the probability of misclassifying a P-random example[5]. The concept class $\mathcal{C}$ is said to be Probably Approximately Correct (PAC) learnable if for all distributions P and all $\varepsilon, \delta > 0$ there is an $m_0 = m_0(\varepsilon, \delta)$ such that for all samples of size $n \geq m_0$, with P^n-probability at least $1 - \delta$, we have $E(H) \leq \hat{E}(H) + \varepsilon$. The main difference between the empirical process and machine learning approaches is that the latter imposes the additional constraint that the problem of finding an $H \in \mathcal{C}$ with low empirical error be efficiently solvable, while the former is mainly concerned with characterizations and rates of convergence.

[5] $C \Delta H = (C \setminus H) \cup (H \setminus C)$ is the symmetric set difference.

The necessary and sufficient conditions for $\mathcal{C}$ to be a Glivenko-Cantelli class (and therefore also PAC-learnable) are stated in terms of a combinatorial property known as the VC-dimension (see, for example [30] or [67]) and have been generalized to real-valued (as opposed to binary) concepts [2].

In this work, we are concerned with relaxing the independence assumption in Glivenko-Cantelli laws. Extensions of uniform laws of large numbers to non-iid processes are fairly recent. Nobel and Dembo [56] show that if $\mathcal{C}$ satisfies (4.46) for an iid process, then the statement also holds for a β-mixing process having identical marginals (see [7]). Following up, Yu [71] gave asymptotic rates for β- and ϕ-mixing stationary processes. For the more specialized setting of homogeneous Markov chains, Gamarnik [23] gives a finite sample PAC-type bound in terms of the spectral gap.

We shall depart from the methods above, emphasizing uniform laws of large numbers as consequences of measure concentration results. Our goal is to obtain finite-sample (as opposed to asymptotic), possibly data-dependent bounds for arbitrary processes having identical marginals. Our result hinges on a decoupling conjecture, which though still open, has accumulated much numerical evidence.

Boucheron, Bousquet and Lugosi [5] present a powerful and aesthetic technique for deriving generalization bounds from concentration inequalities. We begin with a condensed summary of empirical risk minimization, taken from [5], pp. 3-7. In this section, $\Omega = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the *instance space* and $\mathcal{Y} = \{-1, +1\}$ are the labels. Our random process (sample) is a sequence of labeled instances, $(Z_i, Y_i)_{1 \le i \le n}$, and we take it to be iid for now.

Let $\mathcal{C}$ be a collection of classifiers $g : \mathcal{X} \to \{-1, +1\}$ and take the loss function

$$L(g) = \mathbf{P}\{g(Z) \neq Y\}$$

which we estimate by the *empirical error*

$$L_n(g) = \frac{1}{n} \sum_i \mathbb{1}_{\{g(x_i) \neq y_i\}}.$$

Let g_n^* be such that

$$L_n(g_n^*) \le L_n(g) \quad \text{for all } g \in \mathcal{C}.$$

We wish to control the amount by which the empirical error $L_n(g_n^*)$ can differ from the true error $L(g_n^*)$, so we are interested in bounding the quantity

$$\sup_{g \in \mathcal{C}} |L_n(g) - L(g)|,$$

called the *uniform deviation*.

We investigate this problem in a general setting. Let $X = (X_1, \ldots, X_n)$ consist of iid random variables taking values in Ω and let $\mathcal{F}$ be a class of bounded functions $f : \Omega \to [-1, 1]$. We use Pollard's convention of writing $Pf = \mathbf{E}f(X_1)$ and $P_n f = \frac{1}{n} \sum_i f(X_i)$. The quantity of interest is $\sup_{f \in \mathcal{F}} |Pf - P_n f|$, which is a random variable of the sample:

$$\varphi(X_1^n) = \sup_{f \in \mathcal{F}} \left| \sum_{x \in \Omega} P(x) f(x) - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right|. \tag{4.47}$$

Boucheron et al. begin by observing that φ is $2/n$-Lipschitz with respect to the ($w_i \equiv 1$) Hamming metric, and so McDiarmid's inequality applies:

$$\mathbf{P}\{\varphi(X_1^n) - \mathbf{E}\varphi(X_1^n) > t\} \leq \exp\left(-nt^2/2\right),$$

meaning that with probability at least $1-\delta$, we have

$$\sup_{f\in\mathcal{F}} |Pf - P_n f| \;\leq\; \mathbf{E}\left[\sup_{f\in\mathcal{F}} |Pf - P_n f|\right] + \sqrt{\frac{2\log(1/\delta)}{n}}. \tag{4.48}$$

It remains to bound the quantity $\mathbf{E}\left[\sup_{f\in\mathcal{F}} |Pf - P_n f|\right]$, which Boucheron et al. do in terms of Rademacher averages. Define the Rademacher sequence of iid variables $\{\sigma_i\}_{1\leq i\leq n}$, with σ_i taking on the values ± 1 with equal probability. Then (via Jensen's inequality and some calculations) one obtains

$$\mathbf{E}_{P^n}\left[\sup_{f\in\mathcal{F}} |Pf - P_n f|\right] \;\leq\; 2\mathbf{E}_{P^n,\sigma}\sup_{f\in\mathcal{F}}\sum_{i=1}^n \sigma_i f(X_i) \;=\; 2\mathbf{E}_{P^n} R_n(\mathcal{F}(X_1^n)), \tag{4.49}$$

where

$$\mathcal{F}(X_1^n) = \{f(X_i) : 1 \leq i \leq n, f \in \mathcal{F}\}$$

is the projection of $\mathcal{F}$ onto the sample, and the *Rademacher average* R_n is defined for any bounded $A \subset \mathbb{R}^n$ by

$$R_n(A) \;=\; \mathbf{E}_\sigma \sup_{a\in A} \frac{1}{n}\left|\sum_{i=1}^n \sigma_i a_i\right|.$$

Now $R_n(\mathcal{F}(X_1^n))$ is once again a $1/n$-Lipschitz function of the sample, and so will be very close to its mean with a high probability. Thus one obtains

Theorem 4.3.1 (Thm. 3.2 of [5]). *With probability at least $1-\delta$, we have*

$$\sup_{f\in\mathcal{F}} |Pf - P_n f| \;\leq\; 2\mathbf{E}R_n(\mathcal{F}(X_1^n)) + \sqrt{\frac{2\log(1/\delta)}{n}} \tag{4.50}$$

and

$$\sup_{f\in\mathcal{F}} |Pf - P_n f| \;\leq\; 2R_n(\mathcal{F}(X_1^n)) + \sqrt{\frac{2\log(2/\delta)}{n}}. \tag{4.51}$$

The inequality in (4.50) reduces the problem of bounding the generalization error to one of bounding $\mathbf{E}R_n(\mathcal{F}(X_1^n))$; the latter is a property of the function class $\mathcal{F}$ alone, and may be bounded (for $\{0,1\}$-valued f) by $c\sqrt{d_{\text{VC}}/n}$ where d_{VC} is the VC-dimension of $\mathcal{F}$ and c is a universal constant. The bound in (4.51) has the useful property of being *data-dependent*: its value will vary depending on the observed sequence and thus it has the potential of being significantly sharper when the learner gets a particularly "friendly" sample[6].

[6] It may not be obvious how to compute $R_n(\mathcal{F}(X_1^n))$ for a given sample; Boucheron et al. suggest Monte Carlo integration as one method.

We would like to extend this technique to non-iid processes. The analog of (4.48) comes essentially for free. Indeed, let $\mathbf{P}$ be a measure on $\Omega^{\mathbb{N}}$ having identical marginals P. Then it follows directly from Theorem 3.3.1 that

$$\mathbf{P}\{\varphi(X_1^n) - \mathbf{E}\varphi(X_1^n) > t\} \leq \exp\left(-nt^2/4\left\|\Delta_n\right\|_\infty^2\right),$$

where φ is the uniform deviation (4.47), from which we get that

$$\sup_{f\in\mathcal{F}}|Pf - P_nf| \leq \mathbf{E}\left[\sup_{f\in\mathcal{F}}|Pf - P_nf|\right] + 2\left\|\Delta_n\right\|_\infty\sqrt{\frac{\log(1/\delta)}{n}}$$

holds with $\mathbf{P}$-probability at least $\geq 1-\delta$. Observe also that the bounds in (4.50) and (4.51) continue to hold for non-iid processes, with the confidence term modified as above.

The crux of the matter is to generalize (4.49) to non-iid processes. We attempted to modify the definition of a Rademacher average, but this approach yielded no fruit. Instead, we shall pursue a path of further abstraction. Let us now state a conjecture, for which there is strongly compelling numerical evidence; for partial results, see Chapter 6.

Conjecture 4.3.2. *Let μ be an arbitrary probability measure on Ω^n. Let $\tilde{\mu}$ be the unique product measure on Ω^n having the same marginals as μ. Let $\varphi : \Omega^n \to [0,\infty)$ be 1-Lipschitz with respect to the unnormalized Hamming metric. Then*

$$\sum_{x\in\Omega^n}\mu(x)\varphi(x) \quad\leq\quad 1 + \left\|\Delta_n(\mu)\right\|_\infty\sum_{x\in\Omega^n}\tilde{\mu}(x)\varphi(x).$$

If this conjecture were true, we would have

$$\mathbf{E}_{\mathbf{P}}\left[R_n(\mathcal{F}(X_1^n))\right] \quad\leq\quad \left\|\Delta_n(\mathbf{P})\right\|_\infty\mathbf{E}_{P^n}\left[R_n(\mathcal{F}(X_1^n))\right] + \frac{1}{n} \tag{4.52}$$

and the latter can be bounded by the numerous classical methods for bounding Rademacher complexities of function classes (see the amply documented references in [5]).

Chapter 5

Examples and extensions

5.1 Countable and continuous state space

This section borrows heavily from [35]. Our concentration results extend quite naturally to the countable case $\Omega = \mathbb{N}$ and the continuous case $\Omega = \mathbb{R}$. We need to clear three potential hurdles:

(1) check that the martingale difference is well-defined and Azuma's inequality (2.17) continues to hold

(2) check that Φ_w-norm and Ψ_w-norm are well-defined

(3) check that the various inequalities – (3.26), (3.4), (3.29) – continue to hold.

In the $\Omega = \mathbb{N}$ case, no measurability issues arise, so we only need to verify (2). Let $\ell_1(\Omega^n)$ be the set of all summable $\kappa : \Omega^n \to \mathbb{R}$. The Ψ_w-norm continues to be well-defined by (3.1) and is finite since

$$\|\kappa\|_{\Psi,w} \quad \leq \quad n \|f\|_1 , \tag{5.1}$$

as shown in Theorem 5.2.2 below. The definition of Φ_w-norm in (5.12) is likewise unchanged, and again a trivial bound holds by Hölder's inequality:

$$\|\kappa\|_{\Phi,w} \quad \leq \quad \|w\|_1 \|\kappa\|_1 . \tag{5.2}$$

The requisite inequality follows from Sec. 8.1 of [35]:

Theorem 5.1.1. *For $\Omega = \mathbb{N}$ and $\kappa \in \ell_1(\Omega^n)$, we have*

$$\|\kappa\|_{\Phi,w} \quad \leq \quad \|\kappa\|_{\Psi,w} .$$

Proof. Pick any φ, κ in $\ell_1(\Omega^n)$, with the additional constraint that $\varphi : \Omega^n \to [0, \|w\|_1]$ have $\|\varphi\|_{\mathrm{Lip},w} \leq 1$. For $m \geq 1$, let $\Omega_m = \{k \in \Omega : k \leq m\}$ and define the *m-truncation* of κ to be the following function in $\ell_1(\Omega^n)$:

$$\kappa_m(x) \quad = \quad \mathbb{1}_{\{x \in \Omega_m^n\}} \kappa(x).$$

Then we have, by Theorem 3.1.2,

$$\langle \kappa_m, \varphi \rangle \quad \leq \quad \Psi_{w,n}(\kappa_m)$$

for all $m \geq 1$, and $\lim_{m\to\infty} \kappa_m(x) = \kappa(x)$ for all $x \in \Omega^n$. Let $h_m(x) = \kappa_m(x)\varphi(x)$ and note that $|h_m(x)| \leq \|w\|_1 |k(x)|$, the latter in $\ell_1(\Omega^n)$. Thus by Lebesgue's Dominated Convergence theorem, we have $\langle \kappa_m, \varphi\rangle \to \langle \kappa, \varphi\rangle$. A similar dominated convergence argument shows that $\Psi_{w,n}(\kappa_m) \to \Psi_{w,n}(\kappa)$, which proves the claim. □

The continuous version of Theorem 3.1.2 likewise follows by a straightforward approximation argument. We consider the function space $K_n = L_1(\mathbb{R}^n, \mathscr{B}^n, \mu^n)$, where μ^n is the Lebesgue measure. Sums are replaced with integrals over $\mathbb{R}^n$ with respect to μ^n (see §5.2 for formal details); the finiteness of the corresponding Φ_w and Ψ_w norms is easily established (Theorems 5.2.3 and 5.2.2).

Theorem 5.1.2 (Thm. 3.3 of [36]). *Let μ^n be the Lebesgue measure on $\mathbb{R}^n$. Then, for all $f, g \in L_1(\mathbb{R}^n, \mu^n)$ with $g : \mathbb{R}^n \to [0, \|w\|_1]$ and $\|g\|_{\mathrm{Lip},w} \leq 1$, we have*

$$\int_{\mathbb{R}^n} f(x)g(x)d\mu^n(x) \quad \leq \quad \Psi_{w,n}(f). \tag{5.3}$$

Proof. For readability we take $w_i \equiv 1$ and suppress it from the subscripts; this incurs no loss of generality. To avoid ambiguity we will indicate explicitly when inner products, Φ-norms and Ψ-norms are computed over $\mathbb{R}^n$ using the notation $\Phi_{\mathbb{R}}$, $\Psi_{\mathbb{R}}(\cdot)$, and $\langle\cdot,\cdot\rangle_{\mathbb{R}}$.

Let C_c denote the space of continuous functions $f : \mathbb{R}^n \to \mathbb{R}$ with compact support; it follows from [59, Theorem 3.14 of] that C_c is dense in $L_1(\mathbb{R}^n, \mu^n)$, in the topology induced by $\|\cdot\|_{L_1}$. This implies that for any $f \in L_1(\mathbb{R}^n, \mu^n)$ and $\varepsilon > 0$, there is a $g \in C_c$ such that $\|f - g\|_{L_1} < \varepsilon/n$ and therefore (via (5.6) and (5.9)),

$$\|f - g\|_\Phi < \varepsilon \qquad \text{and} \qquad \|f - g\|_\Psi < \varepsilon$$

so it suffices to prove (5.3) for $f \in C_c$.

For $m \in \mathbb{N}$, define $Q_m \subset \mathbb{Q}$ to be the rational numbers with denominator m:

$$Q_m \quad = \quad \{p/r \in \mathbb{Q} : r = m\}.$$

Define the map $\gamma_m : \mathbb{R} \to Q_m$ by

$$\gamma_m(x) \quad = \quad \max\{q \in Q_m : q \leq x\}$$

and extend it to $\gamma_m : \mathbb{R}^n \to Q_m^n$ by defining $[\gamma_m(x)]_i = \gamma_m(x_i)$. The set $Q_m^n \subset \mathbb{R}^n$ will be referred to as the *m-grid points*.

We say that $g \in L_1(\mathbb{R}^n, \mu^n)$ is a *grid-constant function* if there is an $m > 1$ such that $g(x) = g(y)$ whenever $\gamma_m(x) = \gamma_m(y)$; thus a grid-constant function is constant on the grid cells induced by Q_m. Let G_c be the space of the grid-constant functions with compact support; note that $G_c \subset L_1(\mathbb{R}^n, \mu^n)$. It is easy to see that G_c is dense in C_c. Indeed, pick any $f \in C_c$ and let $M \in \mathbb{N}$ be such that $\mathrm{supp}(f) \subset [-M, M]^n$. Now a continuous function is uniformly continuous on a compact set, and so for any $\varepsilon > 0$, there is a $\delta > 0$ such that $\omega_f(\delta) < \varepsilon/(2M)^n$, where ω_f is the ℓ_∞ modulus of continuity of f. Take $m = \lceil 1/\delta \rceil$ and let $g \in G_c$ be such that $\mathrm{supp}(g) \subset [-M, M]^n$ and g agrees with f on the m-grid points. Then we have

$$\|f - g\|_{L_1} \leq (2M)^n \|f - g\|_{L_\infty} < \varepsilon.$$

Thus we need only prove (5.3) for $f \in G_c$, $g \in G_c \cap \Phi_{\mathbb{R}}$.

Let $f \in G_c$ and $g \in G_c \cap \Phi_\mathbb{R}$ be fixed, and let $m > 1$ be such that f and g are m-grid-constant functions. Let $\bar{\kappa}, \bar{\varphi} : Q_m^n \to \mathbb{R}$ be such that $\bar{\kappa}(\gamma_m(x)) = f(x)$ and $\bar{\varphi}(\gamma_m(x)) = g(x)$ for all $x \in \mathbb{R}^n$. As above, choose $M \in \mathbb{N}$ so that $\mathrm{supp}(f) \cup \mathrm{supp}(g) \subset [-M, M]^n$. Thus we have

$$\langle f, g \rangle_\mathbb{R} \quad = \quad \left(\frac{2M}{m}\right)^n \langle \bar{\kappa}, \bar{\varphi} \rangle$$

and

$$\Psi_\mathbb{R}(f) \quad = \quad \left(\frac{2M}{m}\right)^n \Psi_n(\bar{\kappa}).$$

Now Q_m is finite and by construction, $\bar{\varphi} \in \Phi_n$, so Theorem 3.1.5 applies. This shows $\langle f, g \rangle_\mathbb{R} \leq \Psi_\mathbb{R}(f)$ and completes the proof. □

The next order of business is item (1) above. The simplest way to ensure that Azuma's inequality continues to apply is to assume that $\mathbf{P} \ll \mu^n$, where μ^n is the Lebsegue measure on $\mathbb{R}^n$. This allows us replace $\|\cdot\|_\infty$ on V_i with ess sup with respect to μ^n; Azuma's inequality only requires that the martingale difference be bounded almost surely. Similarly, the max in the definition of $\bar{\eta}_{ij}$ (3.14) gets replaced by ess sup with respect to μ^n. The subtleties of conditioning on measure-zero events (and indeed, the existence of conditional distributions) are addressed in §5.4.

5.2 Norm properties of $\|\cdot\|_\Phi$ and $\|\cdot\|_\Psi$

In this section, taken almost entirely from [35], we take $w_i \equiv 1$ without loss of generality and suppress it in the subscripts. It was proved in [37] that $\|\cdot\|_\Phi$ and $\|\cdot\|_\Psi$ are valid norms when Ω is finite. We now do this in a significantly more general setting, and examine the strength of the topologies induced by these norms. We begin with a formal definition of the two norms in abstract metric spaces.

Let $(\mathcal{X}, \rho)$ be a metric space and define $\mathrm{Lip}(\mathcal{X}, \rho)$ to be the set of all $f : \mathcal{X} \to [0, \mathrm{diam}_\rho(\mathcal{X})]$ such that

$$\sup_{x \neq y \in \mathcal{X}} \frac{|f(x) - f(y)|}{\rho(x, y)} \quad \leq \quad 1; \tag{5.4}$$

note that Lipschitz-continuity does not guarantee measurability (indeed, no measure has been specified as of yet).

Let μ be a positive Borel measure on Ω and let $F_n = L_1(\Omega^n, \mu^n)$ be equipped with the inner product

$$\langle f, g \rangle \quad = \quad \int_{\Omega^n} f(x) g(x) d\mu^n(x). \tag{5.5}$$

Since $f, g \in F_n$ might not be in $L_2(\Omega^n, \mu^n)$, the expression in (5.5) in general might not be finite. However, for $g \in \mathrm{Lip}(\Omega^n, \rho)$, we have

$$|\langle f, g \rangle| \quad \leq \quad \mathrm{diam}_\rho(\Omega^n) \, \|f\|_{L_1(\mu^n)}. \tag{5.6}$$

The continuous analog of the *projection* operator $\pi : F_n \to F_{n-1}$ is defined as follows. If $f : \Omega^n \to \mathbb{R}$ then $(\pi f) : \Omega^{n-1} \to \mathbb{R}$ is given by

$$(\pi f)(x_2, \ldots, x_n) \quad = \quad \int_\Omega f(x_1, x_2, \ldots, x_n) d\mu(x_1). \tag{5.7}$$

Note that by Fubini's theorem (Thm. 8.8(c) in [59]), $\pi f \in L_1(\Omega^{n-1}, \mu^{n-1})$. Define the functional $\Psi_n : F_n \to \mathbb{R}$ recursively: $\Psi_0 = 0$ and

$$\Psi_n(f) \quad = \quad \int_{\Omega^n} (f(x))_+ \, d\mu^n(x) + \Psi_{n-1}(\pi f) \tag{5.8}$$

for $n \geq 1$. The latter is finite since

$$\Psi_n(f) \quad \leq \quad n \left\| f \right\|_{L_1(\mu)}, \tag{5.9}$$

as shown in Theorem 5.2.2 below.

We say that the metric space (Ω^n, ρ) is Ψ-*dominated* with respect to a positive Borel measure μ on Ω if the inequality

$$\sup_{g \in \mathrm{Lip}(\Omega^n, \rho)} \langle f, g \rangle \quad \leq \quad \Psi_n(f) \tag{5.10}$$

holds for all $f \in L_1(\Omega^n, \mu^n)$.

Lemma 5.2.1. *Suppose (Ω^n, ρ) is a Ψ-dominated metric space with respect to some (positive Borel) measure μ and (Ω^n, τ) is another metric space, with τ dominated by ρ, in the sense that*

$$\tau(x, y) \quad \leq \quad \rho(x, y), \qquad x, y \in \Omega^n. \tag{5.11}$$

Then (Ω^n, τ) is also Ψ-dominated with respect to μ.

Proof. By (5.11), we have

$$\mathrm{Lip}(\Omega^n, \tau) \quad \subset \quad \mathrm{Lip}(\Omega^n, \rho),$$

which in turn implies

$$\sup_{g \in \mathrm{Lip}(\Omega^n, \tau)} |\langle f, g \rangle| \quad \leq \quad \sup_{g \in \mathrm{Lip}(\Omega^n, \rho)} |\langle f, g \rangle| \leq \Psi_n(f).$$

□

The Φ-norm and Ψ-norm are defined as before:

$$\left\| f \right\|_\Phi \quad = \quad \sup_{g \in \mathrm{Lip}(\Omega^n, \rho)} |\langle f, g \rangle| \tag{5.12}$$

and

$$\left\| f \right\|_\Psi \quad = \quad \max_{s = \pm 1} \Psi_n(sf); \tag{5.13}$$

note that both depend on the measure μ and Φ-norm also depends on the metric.

Establishing the norm properties of $\left\| \cdot \right\|_\Psi$ is straightforward:

Theorem 5.2.2. *Let $F_n = L_1(\Omega^n, \mu^n)$ for some positive Borel measure μ. Then*

(a) $\|\cdot\|_\Psi$ is a vector-space norm on F_n

(b) for all $f \in F_n$,

$$\tfrac{1}{2}\|f\|_{L_1} \;\le\; \|f\|_\Psi \;\le\; n\,\|f\|_{L_1}.$$

Proof. We prove (b) first. Since

$$\|f\|_{L_1} \;=\; \|(f)_+\|_{L_1} + \|(-f)_+\|_{L_1},$$

we have that $\|f\|_\Psi$ (defined in (5.8) and (5.13)) is the sum of n terms, each one at most $\|f\|_{L_1}$ and the first one at least $\frac{1}{2}\|f\|_{L_1}$; this proves (b).

To prove (a) we check the norm axioms:

Positivity: It is obvious that $\|f\|_\Psi \ge 0$ and (b) shows that $\|f\|_\Psi = 0$ and iff $f = 0$ a.e. $[\mu]$.

Homogeneity: It is immediate from (5.8) that $\Psi_n(af) = a\Psi_n(f)$ for $a \ge 0$. From (5.13) we have $\|f\|_\Psi = \|-f\|_\Psi$. Together these imply $\|af\|_\Psi = |a|\,\|f\|_\Psi$.

Subadditivity: It follows from the subadditivity of the function $h(z) = (z)_+$ and additivity of integration that $\|f+g\|_\Psi \le \|f\|_\Psi + \|g\|_\Psi$. □

Theorem 5.2.3. *Let $F_n = L_1(\Omega^n, \mu)$ for some measure space (Ω^n, μ^n). Then $\|\cdot\|_\Phi$ is a seminorm on F_n, for any metric ρ.*

Proof. Nonnegativity: $\|f\|_\Phi \ge 0$ is obvious from the definition (5.12).

Homogeneity: It is clear from the definition that $\|af\|_\Phi = |a|\,\|f\|_\Phi$ for any $a \in \mathbb{R}$.

Subadditivity: $\|f+g\|_\Phi \le \|f\|_\Phi + \|g\|_\Phi$ follows from the linearity of $\langle\cdot,\cdot\rangle$ and the triangle inequality for $|\cdot|$. □

Under mild conditions on the Borel measure space (Ω^n, μ^n), $\|\cdot\|_\Phi$ is a genuine norm. Let μ be a Borel measure on $\mathcal{X}$, whose σ-algebra is generated by some topology $\mathscr{T}$. The measure μ is called *outer regular* if

$$\mu(E) = \inf\{\mu(V) : E \subset V,\ V \text{ is } \mathscr{T}\text{-open}\}$$

for all measurable $E \subset \mathcal{X}$; μ is called non-atomic if $\mu(x) = 0$ for all $x \in \mathcal{X}$.

Theorem 5.2.4. *Let μ be a non-atomic outer regular Borel measure on $\mathcal{X}$. Then for any $f \in L_1(\mathcal{X}, \mu)$, for any metric ρ on $\mathcal{X}$, $\|f\|_\Phi = 0$ iff $f = 0$ a.e. $[\mu]$.*

Proof. Suppose $f \in L_1(\mathcal{X}, \mu)$. The case $f \le 0$ a.e. $[\mu]$ is trivial, so we assume the existence of a $\mathscr{T}$-closed Borel $E \subset \mathcal{X}$ such that

$$0 < \mu(E) < \infty, \qquad f > 0 \text{ on } E.$$

Since μ is outer regular, there is a sequence of $\mathscr{T}$-open sets $V_n \subset \mathcal{X}$ such that $E \subset V_n$ and $\lim_{n\to\infty} \mu(V_n) = \mu(E)$. Define

$$h_n(x) = \frac{\rho(x, V_n^c)}{\rho(x, E) + \rho(x, V_n^c)}$$

where $V_n^c = \mathcal{X} \setminus V_n$; assuming without loss of generality $\mathrm{diam}_\rho(\mathcal{X}) \geq 1$ it is straightforward to verify that $h_n \in \mathrm{Lip}(\mathcal{X}, \rho)$.

By non-atomicity of μ, we have

$$\lim_{n\to\infty} \langle f, h_n \rangle = \int_E f d\mu > 0,$$

which implies that $\langle f, \cdot \rangle$ cannot vanish on all of $\mathrm{Lip}(\mathcal{X}, \rho)$, and so $\|f\|_\Phi > 0$. □

Theorem 5.2.2 shows that $\|\cdot\|_\Psi$ is topologically equivalent to $\|\cdot\|_{L_1}$. The norm strength of $\|\cdot\|_\Phi$ is a more interesting matter. In the case of finite Ω, $F_n = \ell_1(\Omega^n)$ is a finite-dimensional space so all norms on F_n are trivially equivalent. Suppose Ω is a countable set (equipped with the counting measure) and ρ has the property that

$$d_0 = \inf_{x\neq y} \rho(x, y) > 0.$$

The functions $g(x) = d_0 \mathbb{1}_{\{f(x)>0\}}$ and $h(x) = d_0 \mathbb{1}_{\{f(x)<0\}}$ are both in $\mathrm{Lip}(\Omega, \rho)$, and since $d_0 \|f\|_1 = |\langle f, g\rangle| + |\langle f, h\rangle|$, we have

$$\tfrac{1}{2} d_0 \|f\|_1 \;\leq\; \|f\|_\Phi \;\leq\; \mathrm{diam}_\rho(\Omega) \|f\|_1 \tag{5.14}$$

for all $f \in F_n$, so the norms $\|\cdot\|_\Phi$ and $\|\cdot\|_1$ are equivalent in this case.

Suppose, on the other hand, that $T = \{x_1, x_2, \ldots\}$ forms a Cauchy sequence in the countable space Ω, with $\delta_i = \rho(x_i, x_{i+1})$ approaching zero. Let $f \in \ell_1(\Omega)$ be such that $f(x_{2k}) = -f(x_{2k-1})$ for $k = 1, 2, \ldots$ and $f(x) = 0$ for $x \notin T$; then

$$\|f\|_\Phi \;\leq\; \sum_{k=1}^{\infty} |f(x_{2k-1})|\, \delta_{2k-1} \;\leq\; \|f\|_1 \sum_{k=1}^{\infty} \delta_{2k-1}. \tag{5.15}$$

If $\Omega = \mathbb{Q} \cap [0, 1]$ (the rationals in $[0, 1]$) with $\rho(x, y) = |x - y|$ as the metric on Ω, the r.h.s. of (5.15) can be made arbitrarily small, so for this metric space,

$$\inf \{\|f\|_\Phi : \|f\|_1 = 1\} \;=\; 0$$

and $\|\cdot\|_\Phi$ is a strictly weaker norm than $\|\cdot\|_1$.

Similarly, when Ω is a continuous set, $\|\cdot\|_\Phi$ will be strictly weaker than $\|\cdot\|_{L_1}$ in a fairly general setting. As an example, take $n = 1$, $\Omega = [0, 1]$, μ the Lebesgue measure on $[0, 1]$, and $\rho(x, y) = |x - y|$. For $N \in \mathbb{N}$, define $\gamma_N : [0, 1] \to \mathbb{N}$ by

$$\gamma_N(x) \;=\; \max \{0 \leq k < N : k/N \leq x\}.$$

Consider the function

$$f_N(x) = (-1)^{\gamma_N(x)},$$

for $N = 2, 4, 6, \ldots$; note that f is measurable and $\|f\|_{L_1} = 1$.

For a fixed even N, define the k^{th} segment

$$I_k = \{x \in [0, 1] : k \leq \gamma_N(x) \leq k + 2\} = \left[\frac{k}{N}, \frac{k+2}{N}\right],$$

for $k = 0, 2, \ldots, N-2$. Since $\operatorname{diam} I_k = 2/N$, for any $g \in \operatorname{Lip}(\Omega, \rho)$, we have

$$\sup_{I_k} g(x) - \inf_{I_k} g(x) \leq 2/N;$$

this implies

$$\int_{I_k} f_N(x) g(x) d\mu(x) \leq 2/N^2.$$

Now $[0,1]$ is a union of $N/2$ such segments, so

$$\int_0^1 f_N(x) g(x) d\mu(x) \leq 1/N.$$

This means that $\|f\|_\Phi$ can be made arbitrarily small while $\|f\|_{L_1} = 1$, so once again and $\|\cdot\|_\Phi$ is a strictly weaker norm than $\|\cdot\|_{L_1}$.

5.3 ℓ_p and other Ψ-dominated metrics

In this section, we show how our concentration results extend to metrics other than Hamming, such as the ℓ_p metrics on $\mathbb{R}^n$. Throughout this discussion, we will take $\Omega = [0,1]$ and μ to be the Lebesgue measure. For $f : \mathbb{R}^n \to \mathbb{R}$, we define $\|f\|_{\mathrm{Lip},p}$ to be the Lipschitz constant of f with respect to the metric $d(x,y) = \|x - y\|_p$, where $1 \leq p \leq \infty$.

We begin with the simple observation that the unnormalized Hamming metric on $[0,1]^n$ dominates $\ell_1([0,1]^n)$. Recall also that for any $1 < p \leq \infty$ and any $x \in \mathbb{R}^n$, we have

$$\|x\|_p \;\leq\; \|x\|_1 \;\leq\; n^{1/q} \|x\|_p , \tag{5.16}$$

where $1/p + 1/q = 1$. The first inequality holds because the convex function $x \mapsto \|x\|_p$ is maximized on the extreme points (corners) of the convex polytope $\{x \in \mathbb{R}^n : \|x\|_1 = 1\}$. The second inequality is checked by applying Hölder's inequality to $\sum x_i y_i$, with $y \equiv 1$. Both are tight. Thus, in light of Lemma 5.2.1, the Ψ-dominance (with respect to the Lebesgue measure), of $\|\cdot\|_1$ implies the Ψ-dominance of $\|\cdot\|_p$.

We are now in a position to attempt a rough comparison between the results obtained here and the main result of Samson's 2000 paper [60]. Assume for simplicity that for a given random process X on $[0,1]^n$, the two quantities $\|\Delta_n\|_\infty$ and $\|\Gamma_n\|_2$ (defined in (3.19)) are of the same order of magnitude. For example, for the case of contracting Markov chains with contraction coefficient $\theta < 1$, we have

$$\|\Delta_n\|_\infty \leq \frac{1}{1-\theta}, \qquad \|\Gamma_n\|_2 \leq \frac{1}{1-\theta^{1/2}}$$

(as computed in §4.1.2 and [60], respectively).

Suppose $f : [0,1]^n \to \mathbb{R}$ has $\|f\|_{\mathrm{Lip},2} \leq 1$. Samson gives the deviation inequality

$$\mathbf{P}\{|f - \mathbf{E}f| > t\} \;\leq\; 2\exp\left(-\frac{t^2}{2\|\Gamma_n\|_2^2}\right)$$

with the additional requirement that f be convex. By (5.16) we have $\|f\|_{\mathrm{Lip},1} \leq 1$ and we have established above that the ℓ_1 metric is Ψ-dominated. Thus, Theorem 3.3.1 applies:

$$\mathbf{P}\big\{|f - \mathbf{E}f| > t\sqrt{n}\big\} \quad \leq \quad 2\exp\left(-\frac{t^2}{2\,\|\Delta_n\|_\infty^2}\right) \tag{5.17}$$

for any $f : [0,1]^n \to \mathbb{R}$ with $\|f\|_{\mathrm{Lip},2} \leq 1$ (convexity is not required).

To convert from the bound in Theorem 3.1.2 to Samson's bound, we start with a convex $f : [0,1]^n \to \mathbb{R}$, having $\|f\|_{\mathrm{Lip},1} \leq 1$. By (5.16), this means that $\|f\|_{\mathrm{Lip},2} \leq \sqrt{n}$, or equivalently, $\big\|n^{-1/2}f\big\|_{\mathrm{Lip},2} \leq 1$. Applying Samson's bound to $n^{-1/2}f$, we get

$$\mathbf{P}\big\{|f - \mathbf{E}f| > t\sqrt{n}\big\} \quad \leq \quad 2\exp\left(-\frac{t^2}{2\,\|\Gamma_n\|_2^2}\right), \tag{5.18}$$

while the bound provided by Theorem 3.1.2 remains as stated in (5.17).

We stress that the factor of $\sqrt{n}$ in (5.17) and (5.18) appears in the two bounds for rather different reasons. In (5.17), it is simply another way of stating Theorems 3.3.1 for $\|f\|_{\mathrm{Lip},1} \leq 1$; namely, $\mathbf{P}\{|f - \mathbf{E}f| > t\} \leq 2\exp(-t^2/2n\,\|\Delta_n\|_\infty^2)$. In (5.18), the $\sqrt{n}$ was the "conversion cost" between the ℓ_1 and the ℓ_2 metrics.

5.4 Measure-theoretic subtleties

Measurability issues typically arise in probability theory when one considers continuous-time processes [66], takes suprema over uncountable function classes [64], or considers set enlargements with respect to a metric incompatible with the topology generating the Borel σ-algebra [65]. Our martingale approach involves neither, so the only potentially sticky issue is the existence and well-behavedness of the conditional distributions we so heavily rely upon.

Let us illustrate the sort of difficulty that arises when conditioning on measure-zero events, with the following example. Let $\Omega = \{0,1\}$ and define the measure μ on Ω^3 as follows:

$$\mu(x) \quad = \quad \tfrac{1}{2}\mathbb{1}_{\{x_1=x_2=x_3\}}. \tag{5.19}$$

What value does the definition in (3.14) imply for $\bar{\eta}_{2,3}(\mu)$? One might consider slightly perturbing μ to make it strictly positive and appeal to the continuity of η_{ij}. Conditional distributions are indeed well-behaved *if they are well-defined*:

Lemma 5.4.1. *Let $\mathcal{X}$ be a measurable space and suppose the sequence of probability measures μ_n converges in $\|\cdot\|_{\mathrm{TV}}$ to some probability measures μ. If $A, B \subset \mathcal{X}$ are measurable, with $\mu(B) > 0$ and $\mu_n(B) > 0$ for all n, then*

$$\lim_{n\to\infty} \mu_n(A \,|\, B) = \mu(A \,|\, B).$$

Proof. Let the measures ν and ν' be such $\|\nu - \nu'\|_{\mathrm{TV}} < \varepsilon$. Then

$$\begin{aligned}
\left|\frac{\nu(A\cap B)}{\nu(B)} - \frac{\nu'(A\cap B)}{\nu'(B)}\right| \quad &\leq \quad \left|\frac{\nu(A\cap B)}{\nu(B)} - \frac{\nu'(A\cap B)}{\nu(B)+\varepsilon}\right| \\
&\leq \quad \left|\frac{\nu(B)(\nu(A\cap B) - \nu'(A\cap B)) - \varepsilon\nu(A\cap B)}{\nu(B)^2}\right| \\
&\leq \quad \frac{2\varepsilon}{\nu(B)^2}.
\end{aligned}$$

□

However, when conditioning on sets of measure zero, all bets are off. Define the sequence of measures $\{\mu_k\}_{k\geq 3}$ on $\{0,1\}^3$ as follows:

$$\mu_k(x) \quad = \quad \begin{cases} \frac{1}{2} - \frac{1}{k}, & \text{if } x_1 = x_2 = x_3 \\ \\ \frac{1}{3k}, & \text{else} \end{cases} .$$

It is straightforward to verify that μ_k converges in $\|\cdot\|_{\text{TV}}$ to the measure μ defined in (5.19), and that $\bar{\eta}_{2,3}(\mu_k) \to 1/2$. On the other hand, consider the homogeneous Markov measure ν_k on $\{0,1\}^3$ given by $p_0(0) = p_0(1) = 1/2$ and

$$p(a \,|\, b) \quad = \quad \mathbb{1}_{\{a=b\}}(1 - k^{-1}) + \mathbb{1}_{\{a\neq b\}} k^{-1}$$

for $a, b \in \{0,1\}$ and $k \geq 2$. Again, it is easily seen that $\nu_k \to \mu$, but this time $\bar{\eta}_{2,3}(\nu_k) \to 1$. The moral of the story is that when conditioning on sets of measure zero via a limiting process, the limit is not uniquely defined.

Fortunately, this is no cause for despair. Our overarching goal is to bound the deviation probability $\mu\{|f - \mu f| > r\}$, and this quantity is surely insensitive to small perturbations of μ. Thus we may safely approximate a measure μ on a countable set by a sequence of strictly positive measures μ_k. Though different limiting sequences will give rise to different values of $D^* = \lim_{k\to\infty} \|\Delta_n(\mu_k)\|_\infty$, we are justified in using the best (i.e., smallest) value we obtain from any limiting sequence to bound the deviation probability.

The case of $\Omega = \mathbb{R}$ is somewhat simpler – mostly due to our requirement that the probability measure μ on $\mathbb{R}^n$ have a density with respect to the Lebesgue measure. The conditional densities may be obtained by dividing the joint by the marginal; Theorem 3.12 of Pollard [58] assures that under mild conditions the ratio will be well-defined and well-behaved almost everywhere. (Pollard also gives a fascinating discussion of *disintegration* – the mild yet subtle topological conditions under which a joint measure decomposes into a kernel product.)

Most relevant to us is the observation that since D^2 in Azuma's inequality (2.17) need only bound $\sum_{i=1}^n \|V_i\|_\infty^2$ almost surely, we may define $\bar{\eta}_{ij}$ via the ess sup in the continuous case:

$$\bar{\eta}_{ij} \quad = \quad \operatorname*{ess\,sup}_{y\in\mathbb{R}^{i-1}, w, w'\in\mathbb{R}} \eta_{ij}(y, w, w'), \tag{5.20}$$

where ess sup is taken with respect to the Lebesgue measure.

5.5 Breakdown of concentration

Lipschitz continuity and strong mixing have been a prevailing theme throughout this work; let us demonstrate by example (taken from [35]) how concentration can fail if either of these conditions is dropped.

Let μ be the uniform probability measure on $\{0,1\}^n$ and $(X_i)_{1\leq i\leq n}$ be the associated (independent) process. Though different notions of mixing exist [7], X trivially satisfies them all, being an iid process. Define $f : \{0,1\}^n \to [0,1]$ by

$$f(x) = x_1 \oplus x_2 \oplus \ldots \oplus x_n,$$

where $\oplus$ is addition mod 2. Since $\mathbf{P}\{f(X)=0\} = \mathbf{P}\{f(X)=1\} = \frac{1}{2}$, f is certainly not concentrated about its mean (or any other constant). Though X is as well-behaved as can be, f is ill-behaved in the sense that flipping any single input bit causes the output to fluctuate by 1.[1]

For the second example, take $f : \{0,1\}^n \to [0,1]$ to be

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

If $(X_i)_{1\le i\le n}$ is the iid process from the previous example, it is easy to show that the martingale difference V_i is bounded by $1/n$, and so by Azuma's inequality, f is concentrated about its mean. What if we relax the independence condition? Consider the (degenerate) homogeneous Markov process: $\mathbf{P}\{X_1=0\} = \mathbf{P}\{X_1=1\} = \frac{1}{2}$ and $X_{i+1} = X_i$ with probability 1. This process trivially fails to satisfy any (reasonable) definition of mixing [7]. Our well-behaved f is no longer concentrated, since we again have $\mathbf{P}\{f(X)=0\} = \mathbf{P}\{f(X)=1\} = \frac{1}{2}$.

5.6 Using Ψ-norm to bound the transportation cost

Recall the definition of the transportation cost distance from Chapter 2.3.5. Villani [69] gives a fascinating account of the independent discovery of this distance by Monge and Kantorovich, and explains the origin of the synonym "earthmover".

Recall the transporation cost distance between two Borel probability measures μ, ν on a metric space $(\mathcal{X}, \rho)$, defined by

$$T_\rho(\mu,\nu) \;=\; \inf_{\pi} \int_{\mathcal{X}\times\mathcal{X}} \rho(x,y)\, d\pi(x,y), \tag{5.21}$$

where the infimum is taken over all couplings of μ and ν.

Our main interest in this distance is due to Marton's transportation cost inequality (2.28), but this notion extends well beyond measure concentration into areas such as mathematical physics and economics; again, Villani [69] is an encyclopedic source on the matter.

A fortuitous consequence of our linear programming inequality (Theorem 3.1.2) is a simple analytic bound on $T_\rho(\mu,\nu)$, for the case where ρ is any metric dominated[2] by the weighted Hamming metric d_w on Ω^n. This bound is made possible by the Kantorovich duality theorem [69, Thm. 5.9], which states that

$$T_\rho(\mu,\nu) \;=\; \sup_{\varphi\in \mathrm{Lip}_0(\mathcal{X},\rho)} \left(\int_{\mathcal{X}} \varphi d\mu - \int_{\mathcal{X}} \varphi d\nu \right),$$

where $\mathrm{Lip}_0(\mathcal{X},\rho)$ is the set of all $\varphi : \Omega^n \to \mathbb{R}$ such that $\|\varphi\|_{\mathrm{Lip}} \le 1$ (with respect to ρ) and $\varphi(x_0) = 0$ for some $x_0 \in \mathcal{X}$. Applying this to $\mathcal{X} = \Omega^n$ and $\rho = d_w$ (and noting the *translation invariance*: $\langle \varphi, \mu-\nu\rangle = \langle \varphi + a, \mu-\nu\rangle$ for any $a \in \mathbb{R}$), we have that

$$\begin{aligned} T_\rho(\mu,\nu) &= \|\mu-\nu\|_{\Phi,w} \\ &\le \|\mu-\nu\|_{\Psi,w}. \end{aligned}$$

[1] Without making far-reaching claims, we comment on a possible connection between the oscillatory behavior of f and the notorious difficulty of learning noisy parity functions [22]. By contrast, the problem of learning conjunctions and disjunctions under noise has been solved some time ago [30].

[2] in the sense of $\rho(x,y) \le d_w(x,y)$ for all $x, y \in \Omega^n$

Then point is that T_ρ, both in its primal and dual form, involves solving a linear program, and is not, in general, computable in closed form, while the $\Psi_{w,n}$ functional provides a simple, closed-form bound. We hope that this observation leads to new concentration results via Marton's transportation cost inequality, and perhaps finds other applications.

5.7 A "worst-case" family of measures with constant $\|\Delta_n\|_\infty$

The brief investigation we embark upon in this section was motivated in [35] by a comparison between our main indicator of mixing, $\|\Delta_n\|_\infty$, and Samson's [60] closely related quantity $\|\Gamma_n\|_2$, given in (3.18). We proved in [35] that neither is uniformly a sharper indicator of the mixing properties of a measure:

Theorem (Thm. 5.3 of [35]). *There exist families of probability spaces $(\Omega^n, \mu_n)_{n\geq 1}$ such that $R_n \to 0$ and also such that $R_n \to \infty$, where*

$$R_n \quad = \quad \frac{\|\Gamma_n(\mu_n)\|_2}{\|\Delta_n(\mu_n)\|_\infty}.$$

Since Samson's concentration result is for convex, ℓ_2-Lipschitz functions while ours is for d_w-Lipschitz ones (without the convexity requirement), it is not clear how meaningful such a comparison is in general – though we attempt one in §5.3.

A potentially interesting byproduct of this investigation is the problem of constructing families of measures μ whose mixing coefficients $\Delta_n(\mu)$ behave in some prescribed way. In particular, we construct a process $(X_i)_{1\leq i\leq n}$ that achieves a sort of "worst-case" mixing behavior while still having $\|\Delta_n\|_\infty = 2$:

Lemma (Lemma 5.1 of [35]). *There exists a family of probability spaces $(\Omega^n, \mu_n)_{n\geq 1}$ such that*

$$\bar{\eta}_{ij}(\mu_n) = 1/(n-i) \tag{5.22}$$

for $1 \leq i < j \leq n$.

Proof. Let $\Omega = \{0,1\}$ and fix an $n \in \mathbb{N}$. For $1 \leq k < n$, we will call $x \in \{0,1\}^n$ a *k-good* sequence if $x_k = x_n$ and a *k-bad* sequence otherwise. Define $A_n^{(k)} \subset \{0,1\}^n$ to be the set of the k-good sequences and $B_n^{(k)} = \{0,1\}^n \setminus A_n^{(k)}$ to be the bad sequences; note that $\left|A_n^{(k)}\right| = \left|B_n^{(k)}\right| = 2^{n-1}$. Let $\mu_n^{(0)}$ be the uniform measure on $\{0,1\}^n$:

$$\mu_n^{(0)}(x) = 2^{-n}, \qquad x \in \{0,1\}\,.$$

Now take $k = 1$ and define, for some $p_k \in [0, 1/2]$,

$$\mu_n^{(k)}(x) \quad = \quad \alpha_k \mu_n^{(k-1)}(x) \left(p_k \mathbb{1}_{\left\{x \in A_n^{(k)}\right\}} + (1-p_k)\mathbb{1}_{\left\{x \in B_n^{(k)}\right\}} \right), \tag{5.23}$$

where α_k is the normalizing constant, chosen so that $\sum_{x\in\{0,1\}^n} \mu_n^{(k)}(x) = 1$.

We will say that a probability measure μ on $\{0,1\}^n$ is *k-row homogeneous* if for all $1 \leq \ell \leq k$ we have

(a) $h_\ell(\mu) := \bar{\eta}_{\ell,\ell+1}(\mu) = \bar{\eta}_{\ell,\ell+2}(\mu) = \ldots = \bar{\eta}_{\ell,n}(\mu)$

(b) $\bar{\eta}_{ij}(\mu) = 0$ for $k < i < j$

(c) h_k is a continuous function of $p_k \in [0, 1/2]$, with $h_k(0) = 1$ and $h_k(1/2) = 0$.

It is straightforward to verify that $\mu_n^{(1)}$, as constructed in (5.23), is 1-row homogeneous.[3] Therefore, we may choose p_1 in (5.23) so that $h_1 = 1/(n-1)$. Iterating the formula in (5.23) we obtain the sequence of measures $\left\{\mu_n^{(k)} : 1 \le k < n\right\}$; each $\mu_n^{(k)}$ is easily seen to be k-row homogeneous. Another easily verified observation is that $h_\ell(\mu_n^{(k)}) = h_\ell(\mu_n^{(k+1)})$ for all $1 \le k < n-1$ and $1 \le \ell \le k$. This means that we can choose the $\{p_k\}$ so that $h_k(\mu_n^{(k)}) = 1/(n-k)$ for each $1 \le k < n$. The measure $\mu_n := \mu_n^{(n-1)}$ has the desired property (5.22). □

5.8 The significance of ordering and parametrization

An important feature of the martingale method is that it is sensitive to the ordering and the parametrization of the random process. We illustrate the first point with a simple (if not trivial) example.

Define the measure μ on $\{0,1\}^n$ as assigning equal probability to the $x \in \{0,1\}^n$ with $x_1 = x_n$ and zero probability to the rest:

$$\mu(x_1^n) \quad = \quad 2^{-n+1} \mathbb{1}_{\{x_1 = x_n\}},$$

and let $(X_i)_{1 \le i \le n}$ be the associated random process. For this measure, it is easy to see that

$$\bar{\eta}_{ij} \quad = \quad \mathbb{1}_{\{i=1\}}, \qquad 1 \le i < j \le n,$$

which forces $\|\Delta_n(\mu)\|_\infty = n$. Let π be the permutation on $\{1, \ldots, n\}$ that exchanges 2 and n, leaving the other elements fixed, and define the random process $Y = \pi(X)$ by $Y_i = X_{\pi(i)}$, $1 \le i \le n$. It is easily verified that $\|\Delta_n(Y)\|_\infty = 2$. Thus if $f : \{0,1\}^n \to \mathbb{R}$ is invariant under permutations and $\xi_1, \xi_2 \in \mathbb{R}$ are random variables defined by $\xi_1 = f(X)$, $\xi_2 = f(\pi(X))$, we have $\xi_1 = \xi_2$ with probability 1, yet the martingale technique proves much tighter concentration for ξ_2 than for ξ_1. Of course, knowing this special relationship between ξ_1 and ξ_2, we can deduce a corresponding concentration result for ξ_1; what is crucial is that the concentration for ξ_1 is obtained by re-indexing the random variables.

Our second example is perhaps more interesting. Recall from Chapter 2.3.2 that a martingale-derived method, using the notion of metric-space length, can be used to prove Maurey's theorem: if μ is the Haar measure on the symmetric group $\mathcal{S}_n$ then

$$\mu\{|f - \mu f| > r\} \quad \le \quad \exp(-nr^2/32) \tag{5.24}$$

for any $f : \mathcal{S}_n \to \mathbb{R}$ with $\|f\|_{\text{Lip}} \le 1$ with respect to the normalized Hamming metric [61].

A naive attempt to re-derive (5.24) from McDiarmid's inequality might proceed as follows. Let $\Omega = \{1, 2, \ldots, n\}$ and define μ to be the measure on Ω^n that assigns equal weight to permutations (i.e., sequences $x \in \Omega^n$ without repeating symbols) and zero weight to all other sequences. This approach is doomed to fail, since an easy computation yields $\bar{\eta}_{1j} = (n - j + 1)/(n-1)$ for this process, forcing $\|\Delta_n\|_\infty$ is to grow linearly with n.

[3] The continuity of h_k follows from Lemma 5.4.1 and the perturbation argument.

A more clever parametrization (communicated to us by Jim Pitman) does enable one to recover (5.24) via martingale bounds. Let $\Omega_k = \{1, \ldots, k\}$ and consider the independent process $(X_i)_{1 \le i \le n}$, with $X_i \in \Omega_i$ distributed uniformly. A sequence $x \in \Omega_1 \times \Omega_2 \times \ldots \times \Omega_n$ encodes a permutation on $\{1, \ldots, n\}$ by specifying the location into which the next element gets inserted. Applying McDiarmid's inequality to the independent process X_1^n, we get

$$\mu\{|f - \mu f| > r\} \quad \le \quad \exp(-2nr^2),$$

which is even an improvement over (5.24).

Chapter 6

Open problems, conjectures, future directions

One of the joys of mathematics is that the right questions have a way of asking themselves, and in the course of writing this thesis, many more fascinating problems came up than I could hope to solve within the timeframe of a doctorate. That a novice is able to stumble onto such deep problems so early in his journey indicates that the field of measure concentration has no risk of running dry in the near future, and promises to be a fertile ground for fundamental ideas for many years to come. Therefore, I end this thesis by listing some open problems, conjectures, and future directions – both as personal goals and an invitation to the readers to join in the exploration.

6.1 Further applications

The principal contribution of this thesis is the linear programming inequality of Theorem 3.1.2 and the concentration results it implies for nonproduct measures. Anthony Brockwell and I did find an application of these bounds to a concrete real-world problem, and a general application to empirical processes was sketched out in Chapter 4. However, the applications we surveyed in Chapter 1.3 make one optimistic about exploiting the nonproduct nature of our inequalities to extend the corresponding results for product measures. It would be particularly good to find a learning problem, randomized algorithm, or Banach space phenomenon where the random variables have a dependence structure that lends itself to Theorem 3.3.1. We are also hopeful about finding novel ways to apply the linear programming inequality in functional analysis.

6.2 Decoupling

Decoupling inequalities deal with bounding the expectation of a random variable under a nonproduct measure by the same expectation under a product measure; see [68] for a survey of results. As discussed in Chapter 4.3, we need just such a result in order to extend the method of Rademacher averages to non-independent processes.

We conjecture that whenever μ is a measure on Ω^n and $\tilde{\mu}$ is its *product approximation* (i.e., the

unique product measure on Ω^n having the same marginals as μ), we have

$$\sum_{x\in\Omega^n} \mu(x)\varphi(x) \;\;\leq\;\; 1 + \|\Delta_n(\mu)\|_\infty \sum_{x\in\Omega^n} \tilde{\mu}(x)\varphi(x).$$

for any $\varphi : \Omega^n \to [0,\infty)$ that is 1-Lipschitz with respect to the unnormalized Hamming metric. There is compelling numerical evidence supporting this conjecture, and if true, it will have important implications for empirical process theory. I thank Richard Bradley, Richard Dudley, Magda Peligrad and Víctor de la Peña for the helpful correspondence regarding this question.

Another intriguing decoupling possibility is the following. As above, μ and $\tilde{\mu}$ are measures on Ω^n. For $A \subset \Omega^n$ with $\mu(A) \geq 1/2$, we conjecture

$$(c_0 \|\Delta_n(\mu)\|_\infty)^{-1} \;\leq\; \frac{\tilde{\mu}(A)}{\mu(A)} \;\leq\; c_0 \|\Delta_n(\mu)\|_\infty \tag{6.1}$$

for some universal constant $c_0 \approx 2$. If true, (6.1) would provide a generalization of Talagrand's inequality (see below) for nonproduct measures; the evidence for (6.1) is currently scant, however.

6.3 Extending Talagrand's inequality to nonproduct measures

We mentioned Talagrand's powerful inequality in Chapter 2.3.3. Let d_w, $w \in \mathbb{R}^n_+$ be the weighted Hamming metric on some product probability space (Ω^n, μ), and recall the definition of the convex distance:

$$D_A(x) \;\;=\;\; \sup_{\|w\|_2\leq 1} d_w(x, A)$$

for Borel $A \subset \Omega^n$. Talagrand's inequality reads

$$\mu(\overline{A_t}) \;\;\leq\;\; \mu(A)^{-1} \exp(-t^2/4) \tag{6.2}$$

where $\overline{A_t} = \Omega^n \setminus A_t$ and $A_t = \{x \in \Omega^n : D_A(x) \geq t\}$ is the t-enlargement of A.

A natural problem, posed to us by Amir Dembo, is to extend (6.2) for nonproduct measures. Let μ be a nonproduct measure on Ω^n and $\tilde{\mu}$ its product approximation; fix $A \subset \Omega^n$ with $\mu(A) \geq 1/2$. If (6.1) holds, then we have

$$\mu(\overline{A_t}) \;\leq\; k\tilde{\mu}(\overline{A_t}) \;\leq\; k\tilde{\mu}(A)^{-1} \exp(-t^2/4) \;\leq\; k^2 \mu(A)^{-1} \exp(-t^2/4),$$

where $k = c_0 \|\Delta_n(\mu)\|_\infty$; this would provide the desired nonproduct generalization.

Alternatively, one could work with (6.2) directly. Going out on a limb, one might be tempted to conjecture the following generalization:

$$\mu(\overline{A_t}) \;\;\leq\;\; \mu(A)^{-1} \exp\left(-\frac{t^2}{4\|\Delta_n\|_2^2}\right), \tag{6.3}$$

where $\|\Delta_n\|_2$ is the ℓ_2 operator norm of the η-mixing matrix defined in (3.15). Set-measure inequalities have a disadvantage over functional inequalities in that they are much more difficult to test numerically; the evidence in favor of (6.3) is at this point rather scant.

Consider, however, the following variant. Let F be a countable (to avoid measurability issues) subset of the unit ball $B_n = \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$. Define $\varphi : [0,1]^n \to \mathbb{R}$ by

$$\varphi(x) \quad = \quad \sup_{w \in F} \sum_{i=1}^{n} w_i x_i. \tag{6.4}$$

Let us write $\bar{V}_i(\varphi)$ for the maximal i^{th} martingale difference, as in (3.30):

$$\bar{V}_i(\varphi) \quad = \quad \max_{y_1^i \in \Omega^i} \left| V_i(\varphi; y_1^i) \right|$$

(see (3.24) to recall the definition of $V_i(\varphi; \cdot)$).

Conjecture 6.3.1. *If $\varphi : [0,1] \to \mathbb{R}$ is defined as in (6.4) then we have*

$$\sum_{i=1}^{n} \bar{V}_i(\varphi)^2 \quad \leq \quad c \log(n) \, \|\Delta_n\|_2^2 \tag{6.5}$$

for some universal constant c.

Being a functional inequality, (6.5) lends itself more easily to numerical investigation and has accumulated a fair amount of evidence to lend it credence. The conjectured bound is not quite dimension-free, but is significantly stronger than the one furnished by Theorem 3.3.4:

$$\sum_{i=1}^{n} \bar{V}_i(\varphi)^2 \;\leq\; \|\Delta_n \mathbf{1}\|_2^2 \;\leq\; n \, \|\Delta_n\|_\infty^2 \, ;$$

the first place to start would be to prove Conjecture 6.3.1 for the product case.

6.4 Spectral transportation inequality

Let $A = (a_{ij})$ be a column-stochastic matrix, meaning that $a_{ij} \geq 0$ and $\sum_i a_{ij} = 1$. Compute its (complex) eigenvalues, take absolute values, sort in decreasing order (keeping multiplicities), and let λ_2 be the second value on the list[1]; this value is known as second largest eigenvalue modulus (SLEM), [9].

Define the contraction coefficient of A to be

$$\theta = \max \|A_i - A_j\|_{\text{TV}} \, ,$$

where A_i and A_j range over the columns of A; θ is alternatively referred to as Doeblin's or Dobrushin's coefficient.

Consider the metric probability space $(\Omega^n, \bar{d}, \mu)$, where Ω is a finite set, $\bar{d}$ is the normalized Hamming metric on Ω^n and μ is a homogeneous Markov measure on Ω^n.

We may represent the transition kernel of μ by a column-stochastic matrix A, and define λ_2 and θ as above. We know (either from [37] or, in a slightly different form from [45]) that the contraction coefficient θ controls the concentration of μ:

$$\mu \{|f - \mu f| > r\} \quad \leq \quad 2 \exp(-n(1-\theta)^2 r^2/2) \tag{6.6}$$

[1] The largest-modulus eigenvalue is always 1, by the Perron-Frobenius theorem [28].

for any $f : \Omega^n \to \mathbb{R}$ with $\|f\|_{\text{Lip}} \leq 1$ (with respect to $\bar{d}$).

It is also well-known (see, for example, [1] or [52]) that λ_2 controls the rate of convergence (in $\|\cdot\|_{\text{TV}}$) of the Markov chain to the stationary distribution:

$$\|P_n - P_*\|_{\text{TV}} \quad < \quad \frac{1}{2}\sqrt{N}\exp(-(1-\lambda_2)n) \tag{6.7}$$

where $N = |\Omega|$, P_n is the marginal distribution at time step n, and P_* is the stationary distribution.

Now θ can be a rather crude indicator of the mixing properties of the Markov chain. Consider, for example,

$$A \quad = \quad \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix};$$

here, $\theta = 1$ and yields a trivial bound, while $\lambda_2 = 0$, correctly indicating that the chain is actually very rapidly mixing. For 2×2 matrices, it is straightforward to verify that $\theta = \lambda_2$; in general we have

$$\lambda_2 \quad \leq \quad \theta. \tag{6.8}$$

A proof may be found in [1] or [9, Corollary 7.2] and I thank David Aldous, Pierre Brémaud, and Cristopher Moore for providing proof sketches and pointing me to the references.

In light of (6.7) and (6.8), it is tempting to conjecture a bound of the type

$$\mu\{|f - \mu f| > r\} \quad \leq \quad 2\exp(-n(1-\lambda_2)^2 r^2) \tag{6.9}$$

which would be a significant strengthening of (6.6). In the discussion following Proposition $4'$, Marton [46] discusses a similar spectral bound, but hers depends on the stationary distribution of the chain and blows up if the latter takes small values.

One approach for proving (6.6) would be via a transportation inequality (see Chapter 2.3.5):

$$T_\rho(\mu, \nu) \quad \leq \quad \frac{1}{1-\lambda_2}\sqrt{\frac{1}{n}H(\nu\,|\,\mu)} \tag{6.10}$$

where T_ρ is the transportation cost distance (with respect to $\rho = \bar{d}$) and $H(\nu\,|\,\mu)$ is the Kullback-Leibler divergence. If true, (6.10) would strengthen Marton's analog of Pinsker's inequality [45]; some preliminary numerical evidence indeed supports this conjecture. I thank Michel Ledoux for helpful correspondence.

6.5 Questions regarding η-mixing

6.5.1 Connection to other kinds of mixing

Let $(\mathcal{X}, \mathcal{F}, P)$ be a probability space and $\mathcal{A}, \mathcal{B} \subset \mathcal{F}$ be two σ-algebras. Following Bradley [7], we recall some common measures of dependence:

$$\begin{aligned}
\alpha(\mathcal{A},\mathcal{B}) &= \sup\left\{\left|P(A\cap B) - P(A)P(B)\right|, A\in\mathcal{A}, B\in\mathcal{B}\right\} && (6.11)\\
\phi(\mathcal{A},\mathcal{B}) &= \sup\left\{\left|P(B\,|\,A) - P(B)\right|, A\in\mathcal{A}, B\in\mathcal{B}, P(A)>0\right\} && (6.12)\\
\psi(\mathcal{A},\mathcal{B}) &= \sup\left\{\left|\frac{P(A\cap B)}{P(A)P(B)} - 1\right|, A\in\mathcal{A}, B\in\mathcal{B}, P(A)>0, P(B)>0\right\} && (6.13)\\
\beta(\mathcal{A},\mathcal{B}) &= \sup\frac{1}{2}\left\{\sum_{i\in I}\sum_{j\in J}\left|P(A_i\cap B_j) - P(A_i)P(B_j)\right|\right\} && (6.14)
\end{aligned}$$

where the last sup is over all pairs of finite partitions $\{A_i : i \in I\}$ and $\{B_j : j \in J\}$ of $\mathcal{X}$ such that $A_i \in \mathcal{A}$ and $B_j \in \mathcal{B}$. (See [7] for other types of mixing.) The following relations are known and elementary:

$$2\alpha(\mathcal{A},\mathcal{B}) \le \beta(\mathcal{A},\mathcal{B}) \le \phi(\mathcal{A},\mathcal{B}) \le (1/2)\psi(\mathcal{A},\mathcal{B}).$$

If $\mathcal{X} = \Omega^{\mathbb{Z}}$ and $X_{-\infty}^{\infty}$ is the associated random process, we can define the various mixing coefficients as follows:

$$\begin{aligned}
\alpha(i) &= \sup_{j\in\mathbb{Z}} \alpha(\mathcal{F}_{-\infty}^{j}, \mathcal{F}_{j+i}^{\infty});\\
\phi(i) &= \sup_{j\in\mathbb{Z}} \phi(\mathcal{F}_{-\infty}^{j}, \mathcal{F}_{j+i}^{\infty});\\
\psi(i) &= \sup_{j\in\mathbb{Z}} \psi(\mathcal{F}_{-\infty}^{j}, \mathcal{F}_{j+i}^{\infty});\\
\beta(i) &= \sup_{j\in\mathbb{Z}} \beta(\mathcal{F}_{-\infty}^{j}, \mathcal{F}_{j+i}^{\infty}),
\end{aligned}$$

where $\mathcal{F}_a^b$ denotes the σ-algebra induced by X_a^b. The process X is said to be

- *strongly* or α-mixing if $\alpha(i) \to 0$ as $i \to \infty$
- ϕ-mixing if $\phi(i) \to 0$ as $i \to \infty$
- ψ-mixing if $\psi(i) \to 0$ as $i \to \infty$
- *absolutely regular* or *weak Bernoulli* or β-mixing if $\beta(i) \to 0$ as $i \to \infty$;

each type of mixing is additionally called *geometric* if the corresponding quantity decays to zero as λ^i, for $0 < \lambda < 1$. We have shown in Chapter 3.2.2 that[2]

$$\bar{\eta}_{ij} \le 2\phi_{j-i}, \qquad (6.15)$$

which implies that for processes with summable ϕ-mixing coefficients we have $\sup_{n\ge1} \|\Delta_n\|_\infty < \infty$.

[2] This observation seems to have been first made by Samson [60].

As discussed in Chapter 3.2.2, η-mixing appears to be a stronger condition than ϕ-mixing; but it would be good to obtain some nontrivial implications (or non-implications) between η-mixing and the other types of mixing mentioned here.

We conjecture that for any measure μ on Ω^n, we have

$$\frac{1}{2}\sum_{i=1}^{n-1}\phi_i \;\le\; \|\Delta_n\|_\infty \;\le\; 2\sum_{i=1}^{n-1}\phi_i; \tag{6.16}$$

note that the second inequality is nothing else than (6.15). There is a fair amount of evidence for (6.16).

6.5.2 Local independence and mixing

A question worth investigating is the following: if a measure μ satisfies some "local" independence properties, what can be said about (any) mixing properties of μ? Throughout this section, Ω will be a finite set and $\mathcal{P}_n$ will denote the set of all probability measures on Ω^n.

To formalize the notion of *local independence*, consider some $\mathcal{I} \subset 2^{\{1,\ldots,n\}}$. For $I \in \mathcal{I}$, define the I-marginal operator $T_I : \mathcal{P}_n \to \mathcal{P}_{|I|}$ by

$$(T_I\mu)(y) \;=\; \sum_{x\in\Omega^n, x[I]=y} \mu(x), \qquad y \in \Omega^I.$$

The notation $x[I]$ is to be interpreted as follows: for $I = (i_1, i_2, \ldots, i_k)$ and $x \in \Omega^n$,

$$x[I] \;=\; (x_{i_1}, x_{i_2}, \ldots, x_{i_k});$$

also, we will henceforth write μ_I to denote $T_I\mu$. We say that a measure $\mu \in \mathcal{P}_n$ is $\mathcal{I}$-independent if for each $I \in \mathcal{I}$, μ_I is a product measure on Ω^I.

A natural indicator of how dependent the components of $X \in \Omega^n$ under measure $\mu \in \mathcal{P}_n$ is the quantity

$$\mathrm{dep}(\mu) \;=\; \|\Delta_n(\mu)\|_\infty\,;$$

recall that $\mathrm{dep}(\mu) = 1$ iff μ is a product measure.

This leads to a quantitative notion of $\mathcal{I}$-independence. For $\mathcal{I} \subset 2^{\{1,\ldots,n\}}$, define the quantity

$$R_{\mathcal{I}} \;=\; \sup_{\mu\in\mathcal{P}_n} \frac{\mathrm{dep}(\mu)}{\prod_{I\in\mathcal{I}} \mathrm{dep}(\mu_I)} \tag{6.17}$$

(the sup is actually a max since $\mathcal{P}_n$ is a compact set). We have the trivial bound $0 < R_{\mathcal{I}} \le n$. Note that if we are able to upper-bound $R_{\mathcal{I}}$ by some constant independent of n, we will have shown that any $\mathcal{I}$-independent measure is η-mixing.

The first notion of independence we will consider is a "sliding window" of width k. Formally, define $\mathcal{V}_k \subset 2^{\{1,\ldots,n\}}$ by

$$\mathcal{V}_k \;=\; \{I \subset \{1,\ldots,n\} : |I| = k, \quad \max I - \min I = k-1\}\,.$$

It appears that the sliding window is a very weak notion of independence – far too weak to say anything about the η-mixing of μ. Formally, we have

Conjecture 6.5.1.

$$R_{\mathcal{V}_k} \quad = \quad n \tag{6.18}$$

for all $1 \leq k < n$.

(Trivially, $R_{\mathcal{V}_n} = 1$.) This means that a measure μ on Ω^n can be such that all of its windows of width $n-1$ have marginal product measures, yet μ itself has the worst η-mixing constant possible. There is compelling numerical evidence to support this conjecture.

Our second notion is k-wise independence, formalized by defining

$$\mathcal{I}_k \quad = \quad \{I \subset \{1, \ldots, n\} : |I| = k\}\,.$$

Conjecture 6.5.2.

$$R_{\mathcal{I}_k} \quad = \quad n - k + 1 \tag{6.19}$$

for all $1 \leq k < n$.

(Trivially, $R_{\mathcal{I}_n} = 1$.) Again, there is compelling numerical evidence for this conjecture. This is not a strong result. It means that a k-wise independent measure μ can have $\|\Delta_n(\mu)\|_\infty = n-k+1$. To give a meaningful concentration bound, $\|\Delta_n(\mu)\|_\infty$ must be of order $O(\sqrt{n})$. To achieve this bound on the rate of growth, we have to require $k \sim n - \sqrt{n}$.

Proving the conjecture in (6.18) should not be difficult; it suffices to construct the measures $\mu \in \mathcal{P}_n$ that achieve the requisite $R_{\mathcal{V}_k}$. Proving (6.19) might be more involved, but it's not clear how worthwhile the effort would be, given its relative uninformativeness. One possible direction for the future is to define a stronger (yet still realistic) notion of local independence, which does imply nontrivial bounds on $\|\Delta_n(\mu)\|_\infty$.

6.5.3 Constructing Δ_n with given entries

By the construction in Chapter 3.2, $\Delta_n = (\bar{\eta}_{ij})$ is an upper-triangular matrix whose entries are in $[0, 1]$. It is easy to see that for all $1 \leq i < n$ and $i < j_1 < j_2 \leq n$, we have $\bar{\eta}_{i,j_1} \geq \bar{\eta}_{i,j_2}$. Do these constraints completely specify the set of the possible Δ_n – or are there other constraints that all such matrices must satisfy? We are inclined to conjecture the former, but leave this question open for now.

6.5.4 A structural decomposition of Δ_n

Let μ be a probability measure on Ω^n, and for $y \in \Omega$, define $\mu|_y$ to be the law of X_2^n conditioned on $X_1 = y$:

$$\mu|_y(x) \quad = \quad \mathbf{P}\{X_2^n = x \,|\, X_1 = y\}\,, \qquad x \in \Omega^{n-1}.$$

Define $p_1(\cdot)$ to be the marginal law of X_1. Then we conjecture that

$$\sum_{y \in \Omega} p_1(y) \left\| \Delta_{n-1}(\mu|_y) \right\|_\infty \quad \leq \quad \|\Delta_n(\mu)\|_\infty\,. \tag{6.20}$$

Though there is substantial numerical evidence for (6.20), its applications (or indeed a proof) are waiting to be discovered.

Bibliography

[1] David Aldous and James Allen Fill. *Reversible Markov Chains and Random Walks on Graphs.* 2002.

[2] Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44(4):615–631, 1997.

[3] Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Math. Journal*, 19:357–367, 1967.

[4] Keith Ball. *An elementary introduction to modern convex geometry. Flavors of Geometry, ed. by S. Levy, MSRI Publications vol. 31*, pages 1–58. Cambridge University Press, New York, 1997.

[5] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of recent advances. *ESAIM Probab. Statist.*, 9:323–375, 2005.

[6] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. Concentration inequalities using the entropy method. *Ann. Probab.*, 31(3):1583–1614, 2003.

[7] Richard C. Bradley. Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions. *Probab. Surveys*, 2:107–144, 2005.

[8] Leo Breiman. The strong law of large numbers for a class of Markov chains. *Ann. Math. Statist.*, 31:801–803, 1960.

[9] Pierre Brémaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues.* Springer-Verlag, 1999.

[10] Anthony Brockwell, Alexis Rojas, and Robert Kass. Recursive Bayesian decoding of motor cortical signals by particle filtering. *Journal of Neurophysiology*, 91:1899–1907, 2004.

[11] Anthony E. Brockwell. An Interacting-Particle Filter and Application to Likelihood Estimation for State-Space Models. 2007.

[12] Sourav Chatterjee. *Concentration inequalities with exchangeable pairs.* PhD thesis, Stanford University, 2005.

[13] Jean-René Chazottes, Pierre Collet, Christof Külske, and Frank Redig. Concentration inequalities for random fields via coupling. *Probability Theory and Related Fields*, 137(1-2):201–225, 2007.

[14] Herold Dehling, Thomas Mikosch, and Michael Sørensen. *Empirical Process Techniques for Dependent Data*. Birkhäuser, 2002.

[15] Amir Dembo. Information inequalities and concentration of measure. *Ann. Probab.*, 25:927–939, 1997.

[16] Amir Dembo and Ofer Zeitouni. Transportation approach to some concentration inequalities in product spaces. *Elect. Comm. Probab. 1*, 9:83–90, 1996.

[17] Hacène Djellout, M. Arnaud Guillin, and Li Ming Wu. Transportation cost-information inequalities for random dynamical systems and diffusions. *Ann. Probab.*, 32(3B):2702–2732, 2004.

[18] Roland Dobrushin. Central limit theorem for nonstationary Markov chains. II. *Teor. Veroyatnost. i Primenen.*, 1:365–425, 1956.

[19] Paul Doukhan and Sana Louhichi. A new weak dependence condition and applications to moment inequalities. *Stochastic Processes and their Applications*, 84(2):313–342, 1999.

[20] Devdatt Dubhashi and Alessandro Panconesi. Concentration of measure for the analysis of randomised algorithms, book draft. 1998.

[21] Aryeh Dvoretzky. Some results on convex bodies and Banach spaces. In *Proc. Internat. Sympos. Linear Spaces, Jerusalem Academic Press, Jerusalem; Pergamon, Oxford*, pages 123–160, 1960.

[22] Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. New results for learning noisy parities and halfspaces. In *47th Symposium on Foundations of Computer Science (FOCS)*, 2006.

[23] David Gamarnik. Extension of the pac framework to finite and countable markov chains. *IEEE Trans. Inform. Theory*, 49(1):338–345, 2003.

[24] Alison L. Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

[25] Nathael Gozlan. Integral criteria for transportation-cost inequalities. *Elect. Comm. in Probab.*, 11:64–77, 2006.

[26] David L. Hanson, Julius R. Blum, and Lambert H. Koopmans. On the strong law of large numbers for a class of stochastic processes. *Probability Theory and Related Fields*, 2(1):1–11, 1963.

[27] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *American Statistical Association Journal*, 58:13–30, 1963.

[28] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[29] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. Conference in modern analysis and probability (New Haven, Conn.). In *Contemp. Math., 26, Amer. Math. Soc., Providence*, pages 189–206, 1982.

[30] Micheal Kearns and Umesh Vazirani. *An Introduction to Computational Learning Theory*. The MIT Press, 1997.

[31] Jeong Han Kim and Van H. Vu. Concentration of multivariate polynomials and its applications. *Combinatorica*, 20(3):1439–6912, 2000.

[32] Leonid Kontorovich. A Linear Programming Inequality with Applications to Concentration of Measure. 2006.

[33] Leonid Kontorovich. Measure Concentration of Hidden Markov Processes. 2006.

[34] Leonid Kontorovich. Measure Concentration of Markov Tree Processes. 2006.

[35] Leonid Kontorovich. Metric and Mixing Sufficient Conditions for Concentration of Measure. 2006.

[36] Leonid Kontorovich and Anthony E. Brockwell. A Strong Law of Large Numbers for Strongly Mixing Processes. 2007.

[37] Leonid Kontorovich and Kavita Ramanan. Concentration Inequalities for Dependent Random Variables via the Martingale Method. 2006.

[38] Michel Ledoux. On Talagrand's deviation inequalities for product measure. *ESAIM Probab. Statist.*, 1:63–87, 1997.

[39] Michel Ledoux. Concentration of measure and logarithmic sobolev inequalities. *Séminaire de probabilités de Strasbourg*, 33:120–216, 1999.

[40] Michel Ledoux. *The Concentration of Measure Phenomenon. Mathematical Surveys and Monographs Vol. 89*. American Mathematical Society, 2001.

[41] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces*. Springer-Verlag, 1991.

[42] Torgny Lindvall. *Lectures on the Coupling Method*. Dover Publications, 2002.

[43] Gábor Lugosi. Concentration-of-measure inequalities, `http://www.econ.upf.es/~lugosi/anu.ps`, 2003.

[44] Andrei A. Markov. Extension of the law of large numbers to dependent quantities. *Izvestiia Fiz.-Matem. Obsch. Kazan Univ.*, 15:135–156, 1906.

[45] Katalin Marton. Bounding $\bar{d}$-distance by informational divergence: a method to prove measure concentration. *Ann. Probab.*, 24(2):857–866, 1996.

[46] Katalin Marton. Measure concentration for a class of random processes. *Probability Theory and Related Fields*, 110(3):427–439, 1998.

[47] Katalin Marton. Measure concentration and strong mixing. *Studia Scientiarum Mathematicarum Hungarica*, 19(1-2):95–113, 2003.

[48] Katalin Marton. Measure concentration for Euclidean distance in the case of dependent random variables. *Ann. Probab.*, 32(3):2526–2544, 2004.

[49] Pascal Massart. Some applications of concentration inequalities to statistics. *Annales de la faculté des sciences de Toulouse Sér. 6*, 9(2):245–303, 2000.

[50] Bernard Maurey. Construction de suites symétriques. *C. R. Acad. Sci. Paris Sér. A-B 288*, (14):A679–A681, 1979.

[51] Colin McDiarmid. On the method of bounded differences. In J. Siemons, editor, *Surveys in Combinatorics, volume 141 of LMS Lecture Notes Series*, pages 148–188. Morgan Kaufmann Publishers, San Mateo, CA, 1989.

[52] Stephan Mertens and Cristopher Moore. *The Nature of Computation*. Oxford University Press, forthcoming.

[53] Vitali D. Milman. A new proof of A. Dvoretzky's theorem on cross-sections of convex bodies (Russian). *Funkcional. Anal. i Priložen.*, 5(4):28–37, 1971.

[54] Vitali D. Milman and Gideon Schechtman. *Asymptotic Theory of finite Dimensional Normed Spaces. Lecture Notes in Math. 1200*. Springer-Verlag, 1986.

[55] Assaf Naor and Gideon Schechtman. Planar earthmover is not in l_1. In *47th Symposium on Foundations of Computer Science (FOCS)*, 2006.

[56] Andrew Nobel and Amir Dembo. A note on uniform laws of averages for dependent processes. *Statistics and Probability Letters*, 17:169–172, 1993.

[57] David Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

[58] David Pollard. *A user's guide to measure theoretic probability*. Cambridge University Press, 2002.

[59] Walter Rudin. *Real and Complex Analysis*. McGraw-Hill, 1987.

[60] Paul-Marie Samson. Concentration of measure inequalities for Markov chains and Φ-mixing processes. *Ann. Probab.*, 28(1):416–461, 2000.

[61] Gideon Schechtman. *Concentration, results and applications. Handbook of the Geometry of Banach Spaces, Volume 2*. North-Holland, 2003.

[62] Eli Shamir and Joel H. Spencer. Sharp concentration of the chromatic number of random graphs $G_{n,p}$. *Combinatorica*, 7(1):121–129, 1987.

[63] Yakov G. Sinai. *Probability Theory: An Introductory Course*. Springer, 1992.

[64] Michel Talagrand. The Glivenko-Cantelli Problem. *Ann. Probab.*, 15(3):837–870, 1987.

[65] Michel Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'IHÉS*, 81:73–205, 1995.

[66] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer, 1996.

[67] Vladimir N. Vapnik. *Estimation of dependences based on empirical data*. Springer-Verlag, 1982.

[68] Víctor de la Peña and Evarist Giné. *Decoupling: From Dependence to Independence*. Springer, 1998.

[69] Cédric Villani. Optimal transport, old and new. to appear.

[70] Van H. Vu. Concentration of non-lipschitz functions and applications. *Random Struct. Algorithms*, 20(3):262–316, 2002.

[71] Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *Ann. Probab.*, 22(1):94–116, 1994.