TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

Dissertation

Information-Theoretic Analysis of Noncoherent Block-Fading Channels and Singular Random Variables

Ausgeführt zum Zwecke der Erlangung des akademischen Grades eines
Doktors der technischen Wissenschaften

unter der Leitung von
Ao. Univ.-Prof. Dr. Franz Hlawatsch
Institute of Telecommunications

von
Dipl.-Ing. Günther Koliander
Landgutgasse 36/6/13, 1100 Wien, Österreich

Wien, im April 2015 ______________________

Abstract

Characterizing the capacity, i.e., the maximally possible throughput, of a given communication channel is an important problem in information theory. In this dissertation, we study the capacity of block-fading channels in the noncoherent setting where neither the transmitter nor the receiver has *a priori* channel state information but both are aware of the channel statistics. We show that the number of degrees of freedom—which characterizes the channel capacity at high signal-to-noise ratio—conforms to an intuitive "dimension counting" argument based on the noiseless receive vectors, and we analyze two special settings in more detail.

First, extending the well-established constant block-fading model, we consider the class of *generic* multiple-input multiple-output (MIMO) Rayleigh block-fading channels. In these channels, we allow the fading to vary within each block with a temporal correlation that is "generic" in the sense used in the interference-alignment literature. We show that the number of degrees of freedom of a generic MIMO Rayleigh block-fading channel with T transmit antennas and block length N is given by $T(1 - 1/N)$ provided that $T < N$ and the number of receive antennas is at least $T(N-1)/(N-T)$. A comparison with the constant block-fading channel (where the fading is constant within each block) shows that, for large block lengths, generic correlation increases the number of degrees of freedom by a factor of up to four.

Furthermore, we consider an oversampled continuous-time, time-selective, Rayleigh block-fading channel. Here, we show that sampling the filtered channel output at twice the symbol rate results in a significant increase in the number of degrees of freedom.

The noiseless receive vectors of noncoherent block-fading channels and certain random variables arising in other applications are singular random variables, i.e., neither discrete nor continuous. This fact motivates the consideration of information-theoretic properties of integer-dimensional singular random variables in the second part of the thesis. For these random variables, no satisfactory definition of entropy is available. We provide a definition of entropy and show that it is a promising and useful extension of the established concepts of entropy and differential entropy. As possible applications of the proposed entropy definition, we present two new results in source coding. We show that the minimal expected binary codeword length of a quantized integer-dimensional singular random variable can be characterized by the proposed entropy to within an accuracy of one bit. Furthermore, we present a lower bound on the rate-distortion function of an integer-dimensional singular source; this bound depends on the source only via the entropy of the source.

Acknowledgments

First of all, I want to thank my advisers Prof. Franz Hlawatsch and Dr. Erwin Riegler. The two of them turned out to be the perfect mix for supporting me in my research activities. Erwin pointed me to mathematically challenging problems and always provided new ideas when I had a question. Franz found weak spots in my proofs and played devil's advocate against my arguments, and thus helped me to improve my work significantly and to think about completely new approaches.

I also want to thank my coauthors for their collaboration, in particular Prof. Giuseppe Durisi. Giuseppe had as much influence on my work as Franz and Erwin, and made my visits to Chalmers not only a productive but also a pleasant experience.

Even with great supervisors my work at the institute would not have been enjoyable without my fellow PhD students. I will always remember the party discussions with Andreas, the mathematical discussions with Georg, the weird discussions with Gregor, the gossip (with cake) from Veronika, and the discussions with Florian about our shared burden.

Finally, I am indebted to four people making all this possible: My parents who brought me up to be a researcher and always support me. My brother who showed me the beauty of mathematics and teaches me innumerable pieces of wisdom. Stephi who endures my flaws and ensures that my life will be perfect no matter where I go from here.

Contents

APPENDICES

Chapter 1

Introduction

1.1 Background and Motivation

Information theory is a fundamental mathematical framework for describing and analyzing telecommunication scenarios. For a specific channel model, i.e., a mathematical description of the relation between transmit signal and receive signal, we can use information-theoretic tools to design optimal transmission schemes or calculate the maximally possible throughputs (as characterized by the channel capacity). However, the "specific channel model" we consider is always a compromise between realistic assumptions on the one hand and a tractable mathematical model on the other hand. An example of this dilemma is the question whether channel state information (CSI) is assumed to be available. Under the assumption of perfect CSI, a classical result in information theory states that the throughput achievable in multiple-input multiple-output (MIMO) wireless systems grows linearly in the number of antennas [Telatar, 1999]. In practice, though, the assumption of perfect CSI is a strong simplification, and the actually achievable throughput is decreased by the need to acquire CSI [Marzetta and Hochwald, 1999, Zheng and Tse, 2002, Liang and Veeravalli, 2004, Schuster et al., 2009, Moser, 2009, Adhikary et al., 2013]. Especially in large networks or large-MIMO settings, the rate penalty due to channel estimation is an important factor. A fundamental way to assess this rate penalty is to study capacity in the *noncoherent setting* where neither the transmitter nor the receiver has *a priori* CSI but both are aware of the channel statistics. Unfortunately, in this more realistic setting, the mathematical analysis becomes much more difficult or even intractable. Thus, in most cases, a closed-form expression of the capacity in the noncoherent setting is unavailable. However, the capacity at high signal-to-noise ratio (SNR) has been characterized for several different noncoherent channel models, as summarized in what follows.

Stationary Channel Models

One of the most prominent settings is a Rayleigh fading channel model where the fading gains are modeled as a discrete-time stationary process. Focusing on this scenario, [Lapidoth and Moser, 2003] proved that capacity grows double-logarithmically with the SNR if the fading process is *regular*, i.e., if the present fading state cannot be inferred exactly from the past fading states. To obtain a more detailed understanding of the high-SNR capacity for the case of regular fading, one has to study the second-order term in the high-SNR capacity expansion, the so-called *fading number*. Recently, the fading number has been characterized for several stationary discrete-time channel models [Lapidoth and Moser, 2003, Lapidoth and Moser, 2006, Moser, 2009, Koch, 2009]. For the case of *nonregular* Rayleigh fading, where the present fading state can be inferred exactly from the past fading states, the high-SNR capacity behavior depends on the total bandwidth of the fading process, i.e., on the support of its power spectral density. Specifically, the *number of degrees of freedom* (i.e., the first-order term in the high-SNR capacity expansion, or equivalently the asymptotic ratio between capacity and the logarithm of the SNR as the SNR grows large, also referred to as *capacity pre-log*) is given by the measure of the set of frequencies at which the power spectral density vanishes [Lapidoth, 2005]. Moreover, it is shown in [Koch, 2009] that Rayleigh fading yields the smallest number of degrees of freedom among all stationary and ergodic fading processes whose distribution has no mass point at zero.

Continuous-Time Channel Models

The setting of continuous-time channels was recently analyzed in [Durisi et al., 2012] and [Ghozlan and Kramer, 2013]. In [Durisi et al., 2012], the capacity of continuous-time Rayleigh-fading time-frequency selective channels was considered. More specifically, upper and lower bounds on the high-SNR capacity were derived, which imply that the high-SNR capacity of a continuous-time fading channel is close to that of a nonfading channel with the same SNR and bandwidth.

In [Ghozlan and Kramer, 2013], a continuous-time channel model with very specific assumptions on the channel statistics was considered, namely the phase-noise channel model. Rather than using a matched filter and sampling its output at the symbol rate, the effect of using a different filter whose output is sampled multiple times per symbol period was studied. Specifically, for the case of Wiener phase noise, it is shown that, by sampling the filter output signal a number of times per symbol period that grows with the square-root of the SNR, one can achieve data rates that grow logarithmically with the SNR and a number of degrees of freedom of at least $1/2$. In contrast, using a matched filter whose output is sampled at the symbol rate yields only a double-logarithmic growth of capacity with the SNR.

Block-Fading Channel Models

Another prominent setting of Rayleigh fading channels are MIMO block-fading channels. Here, the statistics of the fading gains are defined for blocks of N channel inputs. Between blocks, the fading changes to independent realizations, i.e., the channel is block-memoryless. The simplest model conforming to these assumptions is the MIMO Rayleigh-fading *constant block-fading channel model* [Marzetta and Hochwald, 1999]. In this model, the fading process takes on independent realizations across blocks of N channel uses, and within each block the fading coefficients stay constant. Thus, the N-dimensional vector of channel gains between transmit antenna

$t \in \{1, \dots, T\}$ and receive antenna $r \in \{1, \dots, R\}$ within a block of length N is

$$\mathbf{h}_{r,t} = \mathsf{s}_{r,t} \mathbf{1}_{N \times 1} \,. \tag{1.1}$$

Here, $\mathbf{1}_{N \times 1}$ denotes the N-dimensional all-one vector and $\mathsf{s}_{r,t}$ are $\mathcal{CN}(0, 1)$ random variables that are independent across the receive antennas r and the transmit antennas t. This results in the input-output relation

$$\mathbf{y}_r = \sum_{t=1}^{T} \mathsf{s}_{r,t} \mathbf{x}_t + \mathbf{w}_r, \quad r \in \{1, \dots, R\} \tag{1.2}$$

where $\mathbf{y}_r \in \mathbb{R}^N$ is the channel output at receiver r, $\mathbf{x}_t$ is the channel input at transmitter t, and $\mathbf{w}_r$ is additive white Gaussian noise. Once again, only a high-SNR characterization of the capacity is available. [Zheng and Tse, 2002] proved that the number of degrees of freedom for the MIMO constant block-fading model with T transmit antennas and R receive antennas is given by

$$\chi_{\text{const}} = M\left(1 - \frac{M}{N}\right), \text{ with } M = \min\left\{T, R, \left\lfloor \frac{N}{2} \right\rfloor\right\}. \tag{1.3}$$

For the case $R + T \leq N$, they also provided a high-SNR capacity expansion that is accurate up to a $o(1)$ term (i.e., a term that vanishes as the SNR grows). This expansion was recently extended in [Yang et al., 2013] to the "large-MIMO" setting $R + T > N$.

In [Liang and Veeravalli, 2004], the high-SNR capacity of a continuous-time, time-selective, frequency-flat, Rayleigh block-fading channel is studied. The corresponding discrete-time channel model, which is sometimes referred to as *correlated block-fading* model [Morgenshtern et al., 2013], is a generalization of the standard block-fading model as it allows the fading process to change within each block. In [Liang and Veeravalli, 2004], a lower bound on the number of degrees of freedom was derived. However, this lower bound is tight only for the single-antenna case: In a single-input multiple-output (SIMO) setting, the number of degrees of freedom can be increased [Morgenshtern et al., 2010, Riegler et al., 2011, Morgenshtern et al., 2013]. This is in contrast to the coherent setting, where increasing the number of receive antennas beyond the number of transmit antennas does not provide a degrees-of-freedom gain. So far, no result on the number of degrees of freedom in the correlated block-fading MIMO setting has been available.

The results reviewed above suggest that in block-fading channels, the number of degrees of freedom characterizes the channel capacity in a regime where the noise can "effectively" be ignored. The intuitive argumentation in [Morgenshtern et al., 2013, Section III] suggests that the number of degrees of freedom can be obtained as the number of entries of the transmitted vector $\mathbf{x}$ that can be deduced from the corresponding receive vector $\mathbf{y}$ in the absence of noise, divided by the block length N.

Beyond Degrees of Freedom

Although the number of degrees of freedom is a first step towards characterizing the high-SNR capacity of a channel, it only provides a rough estimate and is very sensitive to the fine details of the channel model. Thus, a high-SNR expansion of the channel capacity up to a constant term (similar to the one available for the constant block-fading channel model [Zheng and Tse, 2002, Yang et al., 2013]) would be valuable. Whereas the number of degrees of freedom can be determined by counting the entries of the transmitted vector $\mathbf{x}$ that can be deduced from the noiseless receive vector, we expect that a more detailed information-theoretic analysis of the relation between $\mathbf{x}$ and the noiseless receive vector will result in a more precise high-SNR expansion of the channel capacity. An analysis of the information-theoretic properties of the noiseless receive vectors appears to be a promising first step towards understanding this relation.

Entropy for Singular Random Variables

The most basic information-theoretic quantity that characterizes a random variable is entropy. The classical definition of entropy for discrete random variables and its interpretation as information content go back to [Shannon, 1948] and were analyzed thoroughly from axiomatic [Csiszár, 2008] and operational [Shannon, 1948] viewpoints. A similar definition for continuous random variables, differential entropy, was also introduced by Shannon [Shannon, 1948], but its interpretation as information content is controversial [Kolmogorov, 1956]. Nonetheless, information-theoretic derivations involving undisputed quantities like Kullback-Leibler divergence or mutual information between continuous random variables can often be simplified using differential entropy. Furthermore, differential entropy arises in asymptotic expansions of the entropy of ever finer quantizations of a continuous random variable [Kolmogorov, 1956, Sec. IV]. Thus, (differential) entropy seems to be an appealing starting-point for our analysis of noiseless receive vectors.

Unfortunately, in many interesting cases, the noiseless receive vectors are neither discrete nor continuous and classical (differential) entropy cannot be used to characterize them. Such *singular* random variables arise not only in the study of noiseless receive vectors. Indeed, two other information-theoretic problems involving singular random variables have been described recently: For the vector interference channel, a singular input distribution has to be used to achieve full degrees of freedom [Stotz and Bölcskei, 2012]; and in a probabilistic formulation of compressed sensing, the underlying source distribution is singular [Wu and Verdú, 2010]. Thus, a suitable generalization of (differential) entropy to singular random variables has the potential to simplify theoretical work in these areas and to provide valuable insights.

Another field where singular random variables appear is source coding. In many high-dimensional problems, deterministic dependencies reduce the intrinsic dimension of a source. Thus, the random variable describing the source cannot be continuous but often is not discrete either. A basic example is a random variable $\mathbf{x} \in \mathbb{R}^2$ supported on the unit circle, i.e., exhibiting the deterministic dependence $x_1^2 + x_2^2 = 1$. Although $\mathbf{x}$ is defined on $\mathbb{R}^2$ and both components x_1, x_2 are continuous random variables, $\mathbf{x}$ itself is intrinsically only one-dimensional. The differential entropy of $\mathbf{x}$ is not defined and, in fact, classical information theory does not provide a rigorous definition of entropy for this random variable.

The case of arbitrary probability distributions is very hard to handle, and due to its generality even the mere definition of a meaningful entropy seems impossible. The two existing approaches to defining (differential) entropy for more general distributions are based on quantizations of the random variable in question. Usually, the entropy of these discretizations converges to infinity and, thus, a normalization has to be employed to obtain a useful result. In [Rényi, 1959], this approach is adopted for very specific quantizations of a random variable. Unfortunately, this does not always result in a well-defined entropy and sometimes even fails for continuous random variables of finite differential entropy [Rényi, 1959, pp. 197f]. Moreover, the quantization process seems analytically difficult to deal with and no theory was built based on this definition of entropy.[1] A similar approach is to consider arbitrary quantizations that are constrained by some measure of fineness to enable a limit operation. In [Kolmogorov, 1956] and [Posner and Rodemich, 1971], the authors introduce ε-entropy as the minimal entropy of all quantizations consisting of sets of diameter less than ε. However, to specify a diameter, they have to define a distortion function. Since all basic information-theoretic quantities (e.g., mutual information or Kullback-Leibler divergence) do not depend on a specific distortion, it is hardly possible to embed ε-entropy into a general information-theoretic framework. Furthermore, once again the quantization process seems to be difficult to deal with analytically.

1.2 Contributions

Extending the MIMO Constant Block-Fading Model (Chapter 2)

We extend the constant block-fading model to a much broader class of block-based input-output relations. More specifically, we only keep the assumptions that (i) in one block of length N, the fading is Gaussian and (ii) the channel outputs y_i are bilinear combinations of the channel inputs x_k and the fading coefficients s_l plus additive noise, i.e.,

$$\mathsf{y}_i = \sum_{k,l} z_{i,k,l} \mathsf{x}_k \mathsf{s}_l + \mathsf{w}_i \,. \tag{1.4}$$

Here, the deterministic coefficients $z_{i,k,l}$ determine the specific channel model and are known to transmitter and receiver. The index sets of i, k, and l can vary depending on the specific application, e.g., in a model with R receive antennas the index i will take on $N_{\mathsf{y}} = RN$ different values. We prove a lower bound on the number of degrees of freedom that is in agreement with the intuitive "dimension-counting" idea described in Section 1.1.

Generic MIMO Block-Fading Channel (Chapter 3)

We consider two specific versions of the channel model (1.4) in more detail. The first is a generic MIMO Rayleigh-fading block-fading system with T transmit and R receive antennas. As in the constant block-fading model (1.1), we assume that the fading vectors $\mathbf{h}_{r,t}$ are Gaussian

[1] This entropy should not be confused with the *information dimension* defined in the same paper [Rényi, 1959], which is indeed a very useful and widely used tool.

random vectors that are independent across the receive antennas $r \in \{1, \ldots, R\}$ and the transmit antennas $t \in \{1, \ldots, T\}$. However, we consider a more general correlation in time. More specifically, we assume that in each block the correlation is described by $Q \geq 1$ independent Gaussian random variables according to

$$\mathbf{h}_{r,t} = \boldsymbol{Z}_{r,t} \mathbf{s}_{r,t} \tag{1.5}$$

where $\boldsymbol{Z}_{r,t} \in \mathbb{C}^{N \times Q}$ with $Q \leq N$ is a deterministic matrix and $\mathbf{s}_{r,t} \in \mathbb{C}^{Q}$ contains independent $\mathcal{CN}(0,1)$ entries, which are also independent across the receive antennas r and the transmit antennas t.

By applying our degrees-of-freedom lower bound to MIMO block-fading channels modeled according to (1.5), we show that when the deterministic matrices $\boldsymbol{Z}_{r,t}$ are *generic*, the number of degrees of freedom can be larger than in the constant block-fading case as given in (1.3). Coarsely speaking, we can think of generic $\boldsymbol{Z}_{r,t}$ as being generated from an underlying joint probability density function.[2] We shall refer to (1.5) with generic $\boldsymbol{Z}_{r,t}$ as *generic block-fading model*. Our specific contribution is as follows: we show that for almost all matrices $\boldsymbol{Z}_{r,t}$, the number of degrees of freedom is given by

$$\chi_{\text{gen}} = T\left(1 - \frac{1}{N}\right) \tag{1.6}$$

provided that $T < N/Q$ and $R \geq T(N-1)/(N-TQ)$. We note that this result does not encompass the case where all matrices $\boldsymbol{Z}_{r,t}$ are *exactly* equal, and thus we do not know whether (1.6) holds in that case. Therefore, the specific setting where all matrices $\boldsymbol{Z}_{r,t}$ are exactly equal remains an open problem. We also provide an upper bound and a lower bound on χ_{gen} for the complementary case $R < T(N-1)/(N-TQ)$.

The results of Chapter 3 have been published in [Koliander et al., 2014].

Oversampled Block-Fading Channel (Chapter 4)

As a second example of the channel model (1.4), we consider a continuous-time, time-selective, frequency-flat, Rayleigh block-fading single-input single-output (SISO) channel. Focusing on the channel model introduced in [Liang and Veeravalli, 2004], we investigate whether using a matched filter whose output is sampled at the symbol rate is optimal from a degrees-of-freedom point of view or using a different filter and more frequent sampling yields a higher throughput at high SNR. We show that the latter is true. Specifically, we prove that the discrete-time channel obtained by sampling the appropriately filtered channel output signal at twice the symbol rate has at least $1 - 1/N$ degrees of freedom provided that $2M + 1 < N$. Here, M reflects the "bandwidth" of the channel fading process (note that the fading process cannot be strictly bandlimited as we consider only finite-duration processes). In contrast, the discrete-time channel obtained with symbol-rate sampling has only $1 - (2M + 1)/N$ degrees of freedom [Liang and Veer-

[2] We use the term "generic" in the same sense as it is used in the interference-alignment literature [Jafar, 2011].

avalli, 2004, Th. 1]. Hence, we conclude that the underlying continuous-time, time-selective, frequency-flat, Rayleigh block-fading channel has at least $1 - 1/N$ degrees of freedom, and that matched filtering and symbol-rate sampling is not capacity-achieving at high SNR.

The results of Chapter 4 have been published in [Dörpinghaus et al., 2014].

Entropy of Integer-Dimensional Random Variables (Chapter 6)

Motivated by the fact that the number of degrees of freedom can be obtained via a heuristic "dimension-counting" analysis of the noiseless input-output relation, we want to investigate the information-theoretic properties of the noiseless receive vectors $\bar{\mathsf{y}}$, whose components are given by (cf. (1.4))

$$\bar{\mathsf{y}}_i = \sum_{k,l} z_{i,k,l} \mathsf{x}_k \mathsf{s}_l \,. \tag{1.7}$$

Because in many interesting cases $\bar{\mathsf{y}}$ is singular, i.e., neither a discrete nor a continuous random variable, even the definition of an entropy turns out to be a demanding task. As mentioned in Section 1.1, the existing approaches to generalizing (differential) entropy are not true generalizations and depend on arbitrary assumptions, e.g., a specific distortion function. Thus, we propose a generalization of entropy and differential entropy to a broader class of random variables. In our approach, we first consider the probability mass functions of quantizations of the random variable and define a density function as a normalized limit of these probability mass functions. (In the special case of a continuous random variable, this results precisely in the probability density function due to Lebesgue's differentiation theorem.) Then we take the expectation of the logarithm of the resulting density function. Due to a fundamental result in geometric measure theory, this approach can be used only for a specific class of random variables, since otherwise the density function does not exist [De Lellis, 2008, Th. 3.1]. In fact, the existence of the density function implies that the random variable is distributed according to a *rectifiable measure* [De Lellis, 2008, Th. 1.1]. Thus, the random variables considered in the second part of this thesis are rectifiable random variables on Euclidean space. Coarsely speaking, these random variables are "integer-dimensional," i.e., they are concentrated on a subset of lower (integer) dimension. Although this is still far from the generality of arbitrary probability distributions, it covers numerous interesting cases and gives valuable insights.

In addition to proposing a definition of entropy for integer-dimensional random variables, we demonstrate connections to classical entropy and differential entropy and we prove a transformation property and invariance under unitary transformations.

The results of Chapter 6 will be submitted for publication [Koliander et al., 2015].

Joint and Conditional Integer-Dimensional Entropy (Chapter 7)

Whereas the definition of joint (differential) entropy for discrete or continuous random variables is straightforward, care has to be exercised in extending this concept to the case of integer-dimensional random variables. In particular, even the question whether a component x of a rectifiable random variable (x, y) is again rectifiable is not trivial. We introduce joint entropy for integer-dimensional random variables and discuss the connections between the entropy of the components and the entropy of the joint random variable for various settings.

Extending the concept of conditional entropy is even more challenging. Based on the definition of a regular conditional probability, we present a rigorous definition of conditional entropy for integer-dimensional random variables. Our analysis of the fine properties of conditional entropy results in a generalization of the classical chain rule for entropy and shows interesting connections between entropy and geometric properties of the random variables. Furthermore, we demonstrate relations of our entropy with the mutual information of integer-dimensional random variables. Finally, we show that our entropy satisfies an asymptotic equipartition property.

The results of Chapter 7 will be submitted for publication [Koliander et al., 2015].

Integer-Dimensional Source Coding (Chapter 8)

Based on the results obtained in Chapters 6 and 7, we demonstrate two applications of our entropy definition to source coding. First, we study the minimal expected codeword length of quantized integer-dimensional sources. More specifically, we consider partitions of the integer-dimensional support set of the source random variable such that each set in the partition has a Hausdorff measure not exceeding a predefined fineness, and we show that the minimal expected binary codeword length of the quantized random variable defined by these partitions can be characterized by our entropy to within an accuracy of one bit.

The second application concerns the rate-distortion (RD) function of integer-dimensional random sources. Based on the characterization of the RD function in [Csiszár, 1974], we prove a lower bound on the RD function that depends on the source only via the entropy of the source. We apply our lower bound to the specific setting of a uniform distribution on the unit circle in $\mathbb{R}^2$ and provide an upper bound that is within 0.2 nats of the lower bound.

The results of Chapter 8 will be submitted for publication [Koliander et al., 2015].

1.3 Notation

Sets are denoted by calligraphic letters (e.g., $\mathcal{I}$), and $|\mathcal{I}|$ denotes the cardinality of the set $\mathcal{I}$. The indicator function of a set $\mathcal{I}$ is denoted by $\mathbb{1}_{\mathcal{I}}$. Sets of sets are denoted by fraktur letters (e.g., $\mathfrak{M}$). The set of natural numbers (including zero) $\{0, 1, 2, \dots\}$ is denoted as $\mathbb{N}$. We use the notation $[M : N]$ to indicate the set $\{n \in \mathbb{N} : M \le n \le N\}$ for $M, N \in \mathbb{N}$. The open sphere with center $\boldsymbol{x} \in \mathbb{R}^M$ and radius $r > 0$ is denoted by $\mathcal{B}_r(\boldsymbol{x})$, i.e., $\mathcal{B}_r(\boldsymbol{x}) \triangleq \{\boldsymbol{y} \in \mathbb{R}^M : \|\boldsymbol{y} - \boldsymbol{x}\| < r\}$. The constant $\omega(M)$ denotes the volume of the M-dimensional unit sphere, i.e., $\omega(M) = \pi^{M/2}/\Gamma(1 + M/2)$ where Γ is the Gamma function. Boldface uppercase and lowercase letters denote matrices and vectors, respectively. Sans serif letters denote random quantities, e.g., A is a random matrix, x is a random vector, and s is a random scalar ($\boldsymbol{A}, \boldsymbol{x}$, and s denote the deterministic counterparts). For a continuous random variable x, $h(\mathsf{x})$ denotes differential entropy [Cover and Thomas, 2006, Ch. 8]. Similarly, for a discrete random variable

x, $H(\mathsf{x})$ denotes entropy [Cover and Thomas, 2006, Ch. 2]. The superscripts $^{\mathrm{T}}$ and $^{\mathrm{H}}$ stand for transposition and Hermitian transposition, respectively. The all-zero matrix of size $N \times M$ is written as $\mathbf{0}_{N\times M}$, and the $M \times M$ identity matrix as $\mathbf{I}_M$. The entry in the ith row and jth column of a matrix $\boldsymbol{A}$ is denoted by $A_{i,j}$, and the ith entry of a vector $\boldsymbol{x}$ by x_i. For an $M \times N$ matrix $\boldsymbol{A}$, we denote by $[\boldsymbol{A}]_{\mathcal{I}}^{\mathcal{J}}$, where $\mathcal{I} \subseteq [1:M]$ and $\mathcal{J} \subseteq [1:N]$, the $|\mathcal{I}| \times |\mathcal{J}|$ submatrix of $\boldsymbol{A}$ containing the entries $A_{i,j}$ with $i \in \mathcal{I}$ and $j \in \mathcal{J}$; furthermore, we let $[\boldsymbol{A}]_{\mathcal{I}} \triangleq [\boldsymbol{A}]_{\mathcal{I}}^{[1:N]}$ and $[\boldsymbol{A}]^{\mathcal{J}} \triangleq [\boldsymbol{A}]_{[1:M]}^{\mathcal{J}}$. We indicate by $[\boldsymbol{x}]_{\mathcal{I}} \in \mathbb{C}^{|\mathcal{I}|}$ the subvector of $\boldsymbol{x}$ containing the entries x_i with $i \in \mathcal{I}$. The diagonal matrix with the entries of $\boldsymbol{x}$ in its main diagonal is denoted by $\operatorname{diag}(\boldsymbol{x})$. We let $\operatorname{diag}(\boldsymbol{A}_1, \dots, \boldsymbol{A}_K)$ be the block-diagonal matrix having the matrices $\boldsymbol{A}_1, \dots, \boldsymbol{A}_K$ on the main block diagonal. By $\det(\boldsymbol{A})$, we denote the determinant of $\boldsymbol{A}$, and by $|\boldsymbol{A}|$, we denote the absolute value of $\det(\boldsymbol{A})$. For $x \in \mathbb{R}$, we define $\lfloor x \rfloor \triangleq \max\{m \in \mathbb{Z} : m \leq x\}$ and $\lceil x \rceil \triangleq \min\{m \in \mathbb{Z} : m \geq x\}$. We write $\mathbb{E}_{\mathsf{x}}[\cdot]$ for the expectation operator with respect to the random variable x. If the random variable x is clear from the context, we simply write $\mathbb{E}[\cdot]$. $\Pr\{\mathsf{x} \in \mathcal{A}\}$ denotes the probability that $\mathsf{x} \in \mathcal{A}$, and $\mathsf{x} \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{\Sigma})$ indicates that x is a circularly symmetric complex Gaussian random vector with covariance matrix $\boldsymbol{\Sigma}$. The Jacobian matrix of a Lipschitz function[3] ϕ is written as $\boldsymbol{J}_\phi$, and the Jacobian determinant is denoted $\mathscr{J}_\phi$. For a function ϕ with domain $\mathcal{D}$ and a subset $\widetilde{\mathcal{D}} \subseteq \mathcal{D}$, we denote by $\phi|_{\widetilde{\mathcal{D}}}$ the restriction of ϕ to the domain $\widetilde{\mathcal{D}}$. We use the Landau notation $f(\rho) = \mathcal{O}(g(\rho))$ to indicate that there exist constants $c_1, c_2 > 0$ such that $|f(\rho)| \leq c_1\, |g(\rho)|$ for $\rho > c_2$. Similarly, we use $f(\rho) = o(g(\rho))$ to indicate that for every $\varepsilon > 0$ there exists a constant $c_3 > 0$ such that $|f(\rho)| \leq \varepsilon\, |g(\rho)|$ for $\rho > c_3$. The function $\operatorname{sinc}(\cdot)$ is defined as

$$\operatorname{sinc}(x) = \begin{cases} \sin(\pi x)/(\pi x), & \text{if } x \neq 0 \\ 1, & \text{if } x = 0. \end{cases} \tag{1.8}$$

For $\boldsymbol{x} \in \mathbb{R}^{M_1}$ and $\boldsymbol{y} \in \mathbb{R}^{M_2}$, we denote by $\mathfrak{p}_{\mathsf{x}} \colon \mathbb{R}^{M_1+M_2} \to \mathbb{R}^{M_1}$, $\mathfrak{p}_{\mathsf{x}}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}$, the projection of $\mathbb{R}^{M_1+M_2}$ to the first M_1 components. Similarly, $\mathfrak{p}_{\mathsf{y}} \colon \mathbb{R}^{M_1+M_2} \to \mathbb{R}^{M_2}$, $\mathfrak{p}_{\mathsf{y}}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{y}$, denotes the projection of $\mathbb{R}^{M_1+M_2}$ to the last M_2 components. $\mathscr{H}^m$ denotes the m-dimensional Hausdorff measure.[4] $\mathscr{L}^M$ denotes the M-dimensional Lebesgue measure, and $\mathfrak{B}_M$ denotes the Borel σ-algebra on $\mathbb{R}^M$. For a measure μ and a measurable function f, the induced measure is given as $\mu f^{-1}(\mathcal{A}) \triangleq \mu(f^{-1}(\mathcal{A}))$. For two measures μ and ν on the same measurable space, we indicate by $\mu \ll \nu$ that μ is absolutely continuous with respect to ν (i.e., for any measurable set $\mathcal{A}$, $\nu(\mathcal{A}) = 0$ implies $\mu(\mathcal{A}) = 0$). For a measure μ and a measurable set $\mathcal{E}$, the measure $\mu|_{\mathcal{E}}$ is the restriction of μ to $\mathcal{E}$, i.e., $\mu|_{\mathcal{E}}(\mathcal{A}) = \mu(\mathcal{A} \cap \mathcal{E})$. The logarithm to the base e is denoted $\log$, and the logarithm to the base 2 is denoted ld.

---

[3] By Rademacher's theorem [Ambrosio et al., 2000, Th. 2.14], a Lipschitz function is differentiable almost everywhere and, thus, the Jacobian matrix and determinant are well defined almost everywhere.

[4] Readers unfamiliar with this concept may think of it as a measure of an m-dimensional area in a higher-dimensional space (e.g., surfaces in $\mathbb{R}^3$).

Part I

Degrees of Freedom of Noncoherent Block-Fading Channels

Chapter 2

Noncoherent Block-Fading Channel

The channel model considered in this chapter can be seen as a blueprint for various different settings. Thus, at this point, we do not provide an interpretation in terms of antennas or time slots. We give an input-output relation for a block of length N that obtains a physical meaning only when narrowed down to a specific setting (as we will do in the subsequent Chapters 3 and 4).

For the proposed channel model, we prove a lower bound on the number of degrees of freedom that conforms to the intuitive "dimension counting" argument in [Morgenshtern et al., 2013, Section III]. According to this argument, the number of degrees of freedom can be obtained as the number of entries of the transmitted vector x that can be deduced from the corresponding receive vector y in the absence of noise, divided by the block length N.

2.1 System Model

For a channel input vector $\mathsf{x} \in \mathbb{C}^{N_\mathsf{x}}$, a channel output vector $\mathsf{y} \in \mathbb{C}^{N_\mathsf{y}}$, a fading vector $\mathsf{s} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_\mathsf{s}})$, and an additive noise vector $\mathsf{w} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_\mathsf{y}})$, the input-output relation in one block of block length N is given by

$$\mathsf{y}_i = \sqrt{\rho}\,\mathsf{x}^{\mathrm{T}} \boldsymbol{Z}_i \mathsf{s} + \mathsf{w}_i \tag{2.1}$$

for all $i \in [1 : N_\mathsf{y}]$. Here $\boldsymbol{Z}_i \in \mathbb{C}^{N_\mathsf{x} \times N_\mathsf{s}}$ specifies which entries of x and s contribute to y_i. In many models, the matrices $\boldsymbol{Z}_i$ will be sparse (e.g., if y_1 depends only on x_1 and not on $[\mathsf{x}]_{[2:N_\mathsf{x}]}$ then all rows of $\boldsymbol{Z}_1$ but the first one are zero). The vectors s and w are assumed to be mutually independent and to change in an independent fashion from block to block ("block-memoryless" assumption). The block length N depends on the specific physical interpretation of the input-output relation (2.1). In a single antenna system, N will be the number of input symbols N_x that we can transmit in one block. However, in multiple antenna systems, the number of input symbols can be larger than the block length, e.g., if we use T transmit antennas then the number of transmitted symbols per block will satisfy $N_\mathsf{x} = TN$.

We can rewrite (2.1) in vector notation as

$$\mathsf{y} = \sqrt{\rho}\,\bar{\mathsf{y}} + \mathsf{w} \tag{2.2}$$

where $\bar{\mathsf{y}} = \mathsf{B}\mathsf{s}$ with

$$\mathsf{B} \triangleq \begin{pmatrix} \mathsf{x}^{\mathrm{T}} \boldsymbol{Z}_1 \\ \mathsf{x}^{\mathrm{T}} \boldsymbol{Z}_2 \\ \vdots \\ \mathsf{x}^{\mathrm{T}} \boldsymbol{Z}_{N_{\mathsf{y}}} \end{pmatrix}. \tag{2.3}$$

2.2 Dimension Counting

The most convenient way to analyze the structure of the matrices $\boldsymbol{Z}_i$ is to consider the mapping $\phi\colon \mathbb{C}^{N_{\mathsf{x}}+N_{\mathsf{s}}} \to \mathbb{C}^{N_{\mathsf{y}}}$ defined by $\phi(\boldsymbol{s}, \boldsymbol{x}) = \bar{\boldsymbol{y}}$. More specifically, the rank of the Jacobian matrix

$$\boldsymbol{J}_{\phi}(\boldsymbol{s}, \boldsymbol{x}) = \begin{pmatrix} \frac{\partial \phi_1}{\partial s_1} & \cdots & \frac{\partial \phi_1}{\partial s_{N_{\mathsf{s}}}} & \frac{\partial \phi_1}{\partial x_1} & \cdots & \frac{\partial \phi_1}{\partial x_{N_{\mathsf{x}}}} \\ \vdots & & \vdots & \vdots & & \vdots \\ \frac{\partial \phi_{N_{\mathsf{y}}}}{\partial s_1} & \cdots & \frac{\partial \phi_{N_{\mathsf{y}}}}{\partial s_{N_{\mathsf{s}}}} & \frac{\partial \phi_{N_{\mathsf{y}}}}{\partial x_1} & \cdots & \frac{\partial \phi_{N_{\mathsf{y}}}}{\partial x_{N_{\mathsf{x}}}} \end{pmatrix} \in \mathbb{C}^{N_{\mathsf{y}} \times (N_{\mathsf{x}}+N_{\mathsf{s}})} \tag{2.4}$$

plays a pivotal role in our capacity analysis. We first provide an intuitive "dimension counting" argument which results in the same number of degrees of freedom as the rigorous lower bound established in the subsequent sections. Assume that $\boldsymbol{J}_{\phi}(\boldsymbol{s}, \boldsymbol{x})$ has rank ℓ at some fixed point $(\boldsymbol{s}, \boldsymbol{x})$, i.e., there exist ℓ linearly independent columns. For simplicity, we assume that

$$[\boldsymbol{J}_{\phi}(\boldsymbol{s}, \boldsymbol{x})]^{[1:\ell]} = \begin{pmatrix} \frac{\partial \phi_1}{\partial s_1} & \cdots & \frac{\partial \phi_1}{\partial s_{N_{\mathsf{s}}}} & \frac{\partial \phi_1}{\partial x_1} & \cdots & \frac{\partial \phi_1}{\partial x_{\ell-N_{\mathsf{s}}}} \\ \vdots & & \vdots & \vdots & & \vdots \\ \frac{\partial \phi_{N_{\mathsf{y}}}}{\partial s_1} & \cdots & \frac{\partial \phi_{N_{\mathsf{y}}}}{\partial s_{N_{\mathsf{s}}}} & \frac{\partial \phi_{N_{\mathsf{y}}}}{\partial x_1} & \cdots & \frac{\partial \phi_{N_{\mathsf{y}}}}{\partial x_{\ell-N_{\mathsf{s}}}} \end{pmatrix} \in \mathbb{C}^{N_{\mathsf{y}} \times \ell}$$

has full rank ℓ. The matrix $[\boldsymbol{J}_{\phi}(\boldsymbol{s}, \boldsymbol{x})]^{[1:\ell]}$ is the Jacobian matrix of the function ϕ when considered as a function of $(\boldsymbol{s}, x_1, \dots, x_{\ell-N_{\mathsf{s}}})$ and the remaining variables $(x_{\ell-N_{\mathsf{s}}+1}, \dots, x_{N_{\mathsf{x}}})$ are treated as fixed parameters. From a communication viewpoint, we can interpret the variables $(x_{\ell-N_{\mathsf{s}}+1}, \dots, x_{N_{\mathsf{x}}})$ as pilot symbols that are known to transmitter and receiver. By the inverse function theorem, ϕ is one-to-one in a neighborhood of $(\boldsymbol{s}, x_1, \dots, x_{\ell-N_{\mathsf{s}}})$. Thus, from the observations ϕ and the parameters $(x_{\ell-N_{\mathsf{s}}+1}, \dots, x_{N_{\mathsf{x}}})$, we can reconstruct $(\boldsymbol{s}, x_1, \dots, x_{\ell-N_{\mathsf{s}}})$ locally, i.e., we can use $\ell - N_{\mathsf{s}}$ dimensions to transmit information. In the subsequent sections, we will show that $\ell - N_{\mathsf{s}}$ degrees of freedom can indeed be achieved.

Note that in general also the number of rows N_{y} is larger than the rank ℓ of the matrix. This implies that the noiseless receive vector $\bar{\boldsymbol{y}}$ belongs to an ℓ dimensional subset and the random variable $\bar{\mathsf{y}}$ is not continuous. Hence, in our proof, we will reduce the receive vector to exactly

ℓ entries, such that $\boldsymbol{J}_\phi(\boldsymbol{s}, \boldsymbol{x})$ contains only ℓ linearly independent rows. We expect that this does not reduce the number of degrees of freedom but a more detailed capacity analysis should consider the entire vector $\bar{\mathbf{y}}$.

2.3 Capacity and Degrees of Freedom

We want to characterize the capacity of the channel (2.1) as ρ goes to infinity. Usually we are interested in a normalized capacity and not a capacity per block. Thus, we will divide the capacity per block by the block length N to obtain normalized values. Because of the block-memoryless assumption, the coding theorem in [Gallager, 1968, Sec. 7.3] implies that the capacity of the channel (2.1) is given by

$$C(\rho) = \frac{1}{N} \sup I(\mathbf{x}; \mathbf{y}) \,. \tag{2.5}$$

Here, $I(\cdot;\cdot)$ denotes mutual information [Cover and Thomas, 2006, p. 251] and the supremum is taken over all probability distributions of $\mathbf{x}$ that satisfy the average-power constraint

$$\mathbb{E}\big[\|\mathbf{x}\|^2\big] \leq N \,. \tag{2.6}$$

The number of degrees of freedom χ is defined as

$$\chi \triangleq \lim_{\rho\to\infty} \frac{C(\rho)}{\log \rho} \tag{2.7}$$

which corresponds to the expansion

$$C(\rho) = \chi \log \rho + o(\log \rho) \,. \tag{2.8}$$

Our main result is stated in the following theorem.

Theorem 2.1 The number of degrees of freedom χ of the channel (2.1) is lower bounded by

$$\chi \geq \frac{1}{N}(\ell - N_{\mathsf{s}}) \tag{2.9}$$

if there exist $\boldsymbol{s} \in \mathbb{C}^{N_{\mathsf{s}}}$ and $\boldsymbol{x} \in \mathbb{C}^{N_{\mathsf{x}}}$ such that the rank of the Jacobian matrix $\boldsymbol{J}_\phi(\boldsymbol{s}, \boldsymbol{x})$ (see (2.4)) is larger than or equal to ℓ.

The remainder of this chapter is a structured proof of Theorem 2.1. Because Theorem 2.1 is only of interest for $\ell > N_{\mathsf{s}}$ (otherwise the lower bound is trivial), we will tacitly assume that $\ell > N_{\mathsf{s}}$ in what follows.

2.4 Bounding $I(\mathbf{x}; \mathbf{y})$

By (2.5), the capacity $C(\rho)$ and, hence, χ (cf. (2.7)) can be lower-bounded by evaluating $I(\mathbf{x}; \mathbf{y})$ for any specific input distribution that satisfies the power constraint (2.6). In particular, in what follows, we will assume $\mathbf{x} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_\mathsf{x}})$. As

$$I(\mathbf{x};\mathbf{y}) = h(\mathbf{y}) - h(\mathbf{y}|\mathbf{x}) \tag{2.10}$$

we can lower-bound $I(\mathbf{x};\mathbf{y})$ by upper-bounding $h(\mathbf{y}|\mathbf{x})$ and lower-bounding $h(\mathbf{y})$.

Upper Bound on $h(\mathbf{y}|\mathbf{x})$

It follows from (2.2) and (2.3) together with $\mathbf{s} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_\mathsf{s}})$ and $\mathbf{w}_r \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_\mathsf{y}})$ that $\mathbf{y}$ is conditionally Gaussian given $\mathbf{x}$, with conditional covariance matrix $\rho\,\mathbf{B}\mathbf{B}^{\mathrm{H}} + \mathbf{I}_{N_\mathsf{y}}$ (note that $\mathbf{B} = \mathbf{B}(\mathbf{x})$). Hence, $h(\mathbf{y}|\mathbf{x}) = \mathbb{E}_{\mathbf{x}}\big[\log\big((\pi e)^{N_\mathsf{y}}\,|\rho\,\mathbf{B}\mathbf{B}^{\mathrm{H}} + \mathbf{I}_{N_\mathsf{y}}|\big)\big]$ according to [Neeser and Massey, 1993, Th. 2]. By [Horn and Johnson, 1985, Th. 1.3.20], $\big|\rho\,\mathbf{B}\mathbf{B}^{\mathrm{H}}+\mathbf{I}_{N_\mathsf{y}}\big| = \big|\rho\,\mathbf{B}^{\mathrm{H}}\mathbf{B}+\mathbf{I}_{N_\mathsf{s}}\big|$. Furthermore, assuming without loss of generality that $\rho > 1$ (note that we are only interested in the asymptotic regime $\rho \to \infty$), we have $\big|\rho\,\mathbf{B}^{\mathrm{H}}\mathbf{B}+\mathbf{I}_{N_\mathsf{s}}\big| \le \big|\rho\big(\mathbf{B}^{\mathrm{H}}\mathbf{B}+\mathbf{I}_{N_\mathsf{s}}\big)\big| = \rho^{N_\mathsf{s}}\big|\mathbf{B}^{\mathrm{H}}\mathbf{B}+\mathbf{I}_{N_\mathsf{s}}\big|$. Thus,

$$\begin{aligned} h(\mathbf{y}|\mathbf{x}) &\le \mathbb{E}_{\mathbf{x}}\big[\log\big((\pi e)^{N_\mathsf{y}}\rho^{N_\mathsf{s}}\,\big|\mathbf{B}^{\mathrm{H}}\mathbf{B}+\mathbf{I}_{N_\mathsf{s}}\big|\big)\big] \\ &= N_\mathsf{s}\log\rho + \mathbb{E}_{\mathbf{x}}\big[\log\big|\mathbf{B}^{\mathrm{H}}\mathbf{B}+\mathbf{I}_{N_\mathsf{s}}\big|\big] + \mathcal{O}(1). \end{aligned} \tag{2.11}$$

By using Jensen's inequality for the concave function $\log(\cdot)$, we obtain

$$\mathbb{E}_{\mathbf{x}}\big[\log\big|\mathbf{B}^{\mathrm{H}}\mathbf{B}+\mathbf{I}_{N_\mathsf{s}}\big|\big] \le \log\mathbb{E}_{\mathbf{x}}\big[\big|\mathbf{B}^{\mathrm{H}}\mathbf{B}+\mathbf{I}_{N_\mathsf{s}}\big|\big]\,. \tag{2.12}$$

The right-hand side in (2.12) is independent of ρ and the determinant $\big|\mathbf{B}^{\mathrm{H}}\mathbf{B}+\mathbf{I}_{N_\mathsf{s}}\big|$ is some polynomial in the entries of $\mathbf{x}$ and $\mathbf{x}^{\mathrm{H}}$ (cf. (2.3)). Since $\mathbf{x} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_\mathsf{x}})$, all moments of $\mathbf{x}$, and, hence, the expectation $\mathbb{E}_{\mathbf{x}}\big[\big|\mathbf{B}^{\mathrm{H}}\mathbf{B}+\mathbf{I}_{N_\mathsf{s}}\big|\big]$, are finite. Thus, the right-hand side in (2.12) is a finite constant with respect to ρ. Hence, (2.11) together with (2.12) implies

$$h(\mathbf{y}|\mathbf{x}) \le N_\mathsf{s}\log\rho + \mathcal{O}(1)\,. \tag{2.13}$$

Lower Bound on $h(\mathbf{y})$

By assumption, we know that there exist $\boldsymbol{s} \in \mathbb{C}^{N_\mathsf{s}}$ and $\boldsymbol{x} \in \mathbb{C}^{N_\mathsf{x}}$ such that $\boldsymbol{J}_\phi(\boldsymbol{x}, \boldsymbol{s})$ has a nonsingular $\ell \times \ell$ submatrix. We denote by $\mathcal{I}$ the set of row indices and by $\mathcal{D}$ the set of column indices specifying the nonsingular submatrix, i.e.,

$$\det\big([\boldsymbol{J}_\phi(\boldsymbol{x}, \boldsymbol{s})]_{\mathcal{I}}^{\mathcal{D}}\big) \ne 0 \tag{2.14}$$

with $|\mathcal{I}| = |\mathcal{D}| = \ell$. The sets $\mathcal{I}^c \triangleq [1 : N_\mathsf{x} + N_\mathsf{s}] \setminus \mathcal{I}$ and $\mathcal{D}^c \triangleq [1 : N_\mathsf{y}] \setminus \mathcal{D}$ give the remaining indices. The main idea behind the proof of the lower bound is that we only have to take care of the observations indicated by $\mathcal{I}$ and the ones in $\mathcal{I}^c$ can be ignored. To do so, it is convenient to separate the N_y receive variables into a "useful" part, which is represented by $[\mathbf{y}]_{\mathcal{I}}$ and a "redundant" part $[\mathbf{y}]_{\mathcal{I}^c}$.

We can now lower-bound $h(\mathbf{y})$ as follows:

$$\begin{aligned}
h(\mathbf{y}) &= h([\mathbf{y}]_{\mathcal{I}}, [\mathbf{y}]_{\mathcal{I}^c}) \\
&\overset{(a)}{=} h([\mathbf{y}]_{\mathcal{I}}) + h\big([\mathbf{y}]_{\mathcal{I}^c} \,\big|\, [\mathbf{y}]_{\mathcal{I}}\big) \\
&\overset{(b)}{\geq} h\big(\sqrt{\rho}\,[\bar{\mathbf{y}}]_{\mathcal{I}} + [\mathbf{w}]_{\mathcal{I}} \,\big|\, [\mathbf{w}]_{\mathcal{I}}\big) + h\big([\mathbf{y}]_{\mathcal{I}^c} \,\big|\, \mathbf{s}, \mathbf{x}, [\mathbf{y}]_{\mathcal{I}}\big) \\
&\overset{(c)}{=} h\big(\sqrt{\rho}\,[\bar{\mathbf{y}}]_{\mathcal{I}}\big) + \mathcal{O}(1) \\
&\overset{(d)}{=} \log\big(\sqrt{\rho}\big)^{2\ell} + h([\bar{\mathbf{y}}]_{\mathcal{I}}) + \mathcal{O}(1) \\
&= \ell \log \rho + h([\bar{\mathbf{y}}]_{\mathcal{I}}) + \mathcal{O}(1)\,.
\end{aligned} \tag{2.15}$$

Here, (a) holds by the chain rule for differential entropy [Cover and Thomas, 2006, Th. 8.6.2], in (b) we used (2.2) and the fact that conditioning reduces differential entropy, (c) holds since $h\big([\mathbf{y}]_{\mathcal{I}^c} \,\big|\, \mathbf{s}, \mathbf{x}, [\mathbf{y}]_{\mathcal{I}}\big) = h\big([\mathbf{w}]_{\mathcal{I}^c}\big)$ is a finite constant, and (d) holds by the transformation property of differential entropy [Cover and Thomas, 2006, eq. (8.71)]. Using (2.13) and (2.15) in (2.10), we obtain

$$I(\mathbf{x}; \mathbf{y}) \geq (\ell - N_{\mathbf{s}}) \log \rho + h([\bar{\mathbf{y}}]_{\mathcal{I}}) + \mathcal{O}(1)\,. \tag{2.16}$$

The degrees of freedom lower bound (2.9) follows by inserting (2.16) into (2.5):

$$C(\rho) \geq \frac{1}{N} I(\mathbf{x}; \mathbf{y}) \geq \frac{1}{N}(\ell - N_{\mathbf{s}}) \log \rho + \frac{h([\bar{\mathbf{y}}]_{\mathcal{I}})}{N} + \mathcal{O}(1)$$

whence, by (2.7) and because $h([\bar{\mathbf{y}}]_{\mathcal{I}})$ does not depend on ρ,

$$\chi \geq \lim_{\rho \to \infty} \frac{\frac{1}{N}(\ell - N_{\mathbf{s}}) \log \rho + \mathcal{O}(1)}{\log \rho} = \frac{1}{N}(\ell - N_{\mathbf{s}})$$

provided that $h([\bar{\mathbf{y}}]_{\mathcal{I}}) > -\infty$. To conclude the proof, we will next show that $h([\bar{\mathbf{y}}]_{\mathcal{I}}) > -\infty$. This is the most technical part of the proof.

2.5 Proof that $h([\bar{\mathbf{y}}]_{\mathcal{I}}) > -\infty$

As $[\bar{\mathbf{y}}]_{\mathcal{I}}$ is a function of $\mathbf{s}$ and $\mathbf{x}$ (see (2.2) and (2.3)), the idea behind our proof is to relate $h([\bar{\mathbf{y}}]_{\mathcal{I}})$, which we are not able to calculate directly, to $h(\mathbf{s}, \mathbf{x})$, which can be calculated trivially. The underlying intuition is that the image of a random variable of finite differential entropy, such as $(\mathbf{s}, \mathbf{x})$, under a "well-behaved" mapping, such as $(\mathbf{s}, \mathbf{x}) \mapsto [\bar{\mathbf{y}}]_{\mathcal{I}}$, cannot have an infinite differential entropy. At the heart of the proof is the bounding of differential entropy under finite-to-one mappings, to be established in Lemma 2.3 below.

To simplify notation we introduce the shorthand $\overline{\mathbf{sx}} \triangleq (\mathbf{s}^{\mathrm{T}}\, \mathbf{x}^{\mathrm{T}})^{\mathrm{T}}$. We first need to characterize the mapping between $\overline{\mathbf{sx}}$ and $[\bar{\mathbf{y}}]_{\mathcal{I}}$. To equalize the dimensions—note that $[\bar{\mathbf{y}}]_{\mathcal{I}} \in \mathbb{C}^{\ell}$ and $\overline{\mathbf{sx}} \in \mathbb{C}^{N_{\mathbf{s}}+N_{\mathbf{x}}}$—we condition on $N_{\mathbf{s}} + N_{\mathbf{x}} - \ell$ entries of $\overline{\mathbf{sx}}$, defined by $\mathcal{D}^c$ (recall that $\mathcal{D}$ was defined in (2.14)). This results in

$$h([\bar{\mathbf{y}}]_{\mathcal{I}}) \geq h([\bar{\mathbf{y}}]_{\mathcal{I}} \,|\, [\overline{\mathbf{sx}}]_{\mathcal{D}^c}) \,. \tag{2.17}$$

The remaining entries are given by $[\overline{\mathbf{sx}}]_{\mathcal{D}}$.

Because of (2.17), it suffices to show that

$$h([\bar{\mathbf{y}}]_{\mathcal{I}} \,\big|\, [\overline{\mathbf{sx}}]_{\mathcal{D}^c}) > -\infty \,. \tag{2.18}$$

This will be done by relating $h([\bar{\mathbf{y}}]_{\mathcal{I}} \,\big|\, [\overline{\mathbf{sx}}]_{\mathcal{D}^c})$ to $h([\overline{\mathbf{sx}}]_{\mathcal{D}})$. Before doing so, we have to understand the connection between the variables $[\bar{\mathbf{y}}]_{\mathcal{I}}$ and $[\overline{\mathbf{sx}}]_{\mathcal{D}}$. This leads us to the following program:

(i) Define the polynomial mapping $\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}$ relating $[\overline{\mathbf{sx}}]_{\mathcal{D}}$ and $[\bar{\mathbf{y}}]_{\mathcal{I}}$.

(ii) Prove that $\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}$ satisfies the following two properties:

  a) Its Jacobian matrix is nonsingular almost everywhere (a.e.) for almost all (a.a.) $[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}$.

  b) It is finite-to-one[1] a.e. for a.a. $[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}$.

(iii) Apply a novel result on the change in differential entropy that occurs when a random variable undergoes a finite-to-one mapping to relate $h([\bar{\mathbf{y}}]_{\mathcal{I}} \,\big|\, [\overline{\mathbf{sx}}]_{\mathcal{D}^c})$ to $h(\mathbf{s}, [\overline{\mathbf{sx}}]_{\mathcal{D}})$.

(iv) Bound the terms resulting from this change in differential entropy.

Step (i): We consider the $[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}$-parametrized mapping

$$\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}} : \mathbb{C}^{\ell} \to \mathbb{C}^{\ell};\; [\overline{\boldsymbol{sx}}]_{\mathcal{D}} \mapsto\; [\phi(\overline{\boldsymbol{sx}})]_{\mathcal{I}} \tag{2.19}$$

where $\phi(\overline{\boldsymbol{sx}}) = \bar{\boldsymbol{y}}$ is defined in (2.2) and (2.3), i.e., $\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}$ is the mapping ϕ defined in Section 2.2 projected to the indices given by $\mathcal{I}$ and with fixed values $[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}$. The components of the vector-valued mapping $\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}$ are multivariate polynomials of degree 2 in the entries of $[\overline{\boldsymbol{sx}}]_{\mathcal{D}}$. The Jacobian matrix $\boldsymbol{J}_{\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}}$ of $\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}$ is equal to the submatrix specified by $\mathcal{I}$ and $\mathcal{D}$ of the Jacobian $\boldsymbol{J}_{\phi}$ given in (2.4) with fixed variables $[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}$, i.e., $\boldsymbol{J}_{\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}}([\overline{\boldsymbol{sx}}]_{\mathcal{D}}) = \big[\boldsymbol{J}_{\phi}(\overline{\boldsymbol{sx}})\big]_{\mathcal{I}}^{\mathcal{D}}$.

Step (ii-a): We have to show that $\boldsymbol{J}_{\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}}$ is nonsingular (i.e., $|\boldsymbol{J}_{\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}}| \neq 0$) a.e. for a.a. $[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}$. The determinant of $\boldsymbol{J}_{\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}}$ is a polynomial $p(\overline{\boldsymbol{sx}})$. By assumption, we know that $p(\overline{\boldsymbol{sx}})$ does not vanish identically. Since a polynomial vanishes either identically or on a set of measure zero [Gunning and Rossi, 1965, Cor. 10], we conclude that $p(\overline{\boldsymbol{sx}})$ does not vanish a.e. In other words, the matrix $\boldsymbol{J}_{\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}}$ is nonsingular a.e. for a.a. $[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}$.

Step (ii-b): We will invoke Bézout's theorem [van den Essen, 2000, Prop. B.2.7] to show that the mapping $\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}$ is finite-to-one a.e. for a.a. $[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}$. In what follows, note that for a given $[\bar{\boldsymbol{y}}]_{\mathcal{I}}$ in the codomain of $\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}$, the quantity $\phi^{-1}_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}([\bar{\boldsymbol{y}}]_{\mathcal{I}})$ is the preimage

$$\phi^{-1}_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}([\bar{\boldsymbol{y}}]_{\mathcal{I}}) = \{[\overline{\boldsymbol{sx}}]_{\mathcal{D}} : \phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}([\overline{\boldsymbol{sx}}]_{\mathcal{D}}) = [\bar{\boldsymbol{y}}]_{\mathcal{I}}\}$$

[1] A mapping is called finite-to-one if every element in the codomain has a preimage of finite cardinality.

and not the function value of the inverse function (which does not exist in most cases). Furthermore, for a given $[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}$, we denote by $\widetilde{\mathcal{M}} \subseteq \mathbb{C}^\ell$ the set of all $[\overline{\boldsymbol{sx}}]_{\mathcal{D}}$ for which $\boldsymbol{J}_{\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}}([\overline{\boldsymbol{sx}}]_{\mathcal{D}})$ is nonsingular, i.e.,

$$\widetilde{\mathcal{M}} \triangleq \left\{[\overline{\boldsymbol{sx}}]_{\mathcal{D}} \in \mathbb{C}^\ell : |\boldsymbol{J}_{\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}}([\overline{\boldsymbol{sx}}]_{\mathcal{D}})| \neq 0\right\}. \tag{2.20}$$

Lemma 2.2 For a given $[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}$, let $\widetilde{\mathcal{M}}$ be defined as above. Then for all $[\bar{\boldsymbol{y}}]_{\mathcal{I}} \in \phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}(\widetilde{\mathcal{M}})$,

$$\left|\phi^{-1}_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}([\bar{\boldsymbol{y}}]_{\mathcal{I}}) \cap \widetilde{\mathcal{M}}\right| \leq \widetilde{m} \triangleq 2^\ell .$$

Proof. Let $[\bar{\boldsymbol{y}}]_{\mathcal{I}} \in \phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}(\widetilde{\mathcal{M}})$. The set $\phi^{-1}_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}([\bar{\boldsymbol{y}}]_{\mathcal{I}})$ contains all points $[\overline{\boldsymbol{sx}}]_{\mathcal{D}} \in \mathbb{C}^\ell$ such that $\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}([\overline{\boldsymbol{sx}}]_{\mathcal{D}}) = [\bar{\boldsymbol{y}}]_{\mathcal{I}}$. Thus, these points are the zeros of the vector-valued mapping

$$[\overline{\boldsymbol{sx}}]_{\mathcal{D}} \mapsto \phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}([\overline{\boldsymbol{sx}}]_{\mathcal{D}}) - [\bar{\boldsymbol{y}}]_{\mathcal{I}} . \tag{2.21}$$

It follows from (2.19) that each component of the vector-valued mapping (2.21) is a polynomial of degree 2. Hence, the zeros of the mapping (2.21) are the common zeros of ℓ polynomials of degree 2. By a weak version of Bézout's theorem [van den Essen, 2000, Prop. B.2.7], the number of isolated zeros (i.e., with no other zeros in some neighborhood) cannot exceed $\widetilde{m} = 2^\ell$. Since $\boldsymbol{J}_{\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}}$ is nonsingular on $\widetilde{\mathcal{M}}$, the function $\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}$ restricted to $\widetilde{\mathcal{M}}$ is locally one-to-one [Rudin, 1976, Th. 9.24] and, hence, each zero of $\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}} - [\bar{\boldsymbol{y}}]_{\mathcal{I}}$ on $\widetilde{\mathcal{M}}$ has to be an isolated zero. Therefore, the number of points $[\overline{\boldsymbol{sx}}]_{\mathcal{D}} \in \widetilde{\mathcal{M}}$ such that $\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}([\overline{\boldsymbol{sx}}]_{\mathcal{D}}) = [\bar{\boldsymbol{y}}]_{\mathcal{I}}$ cannot exceed $\widetilde{m}$. □

By Lemma 2.2, the function $\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}$ for a given $[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}$ is finite-to-one on the set $\widetilde{\mathcal{M}}$. Because by Step (ii-a) the matrix $\boldsymbol{J}_{\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}}([\overline{\boldsymbol{sx}}]_{\mathcal{D}})$ is nonsingular a.e. for a.a. $[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}$, and because $\widetilde{\mathcal{M}} \subseteq \mathbb{C}^\ell$ is the set of all $[\overline{\boldsymbol{sx}}]_{\mathcal{D}}$ for which $\boldsymbol{J}_{\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}}([\overline{\boldsymbol{sx}}]_{\mathcal{D}})$ is nonsingular, we conclude that $\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}$ is finite-to-one a.e. for a.a. $[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}$.

Step (iii): We will use the following novel result bounding the change in differential entropy that occurs when a random variable undergoes a finite-to-one mapping.

Lemma 2.3 Let $\mathbf{u} \in \mathbb{C}^n$ be a random vector with probability density function $f_{\mathbf{u}}$. Consider a continuously differentiable mapping $\kappa\colon \mathbb{C}^n \to \mathbb{C}^n$ with Jacobian matrix $\boldsymbol{J}_\kappa$. Assume that $\boldsymbol{J}_\kappa$ is nonsingular a.e. and let $\mathcal{M} \triangleq \{\boldsymbol{u} \in \mathbb{C}^n : |\boldsymbol{J}_\kappa(\boldsymbol{u})| \neq 0\}$ (thus, $\mathbb{C}^n \setminus \mathcal{M}$ has Lebesgue measure zero). Furthermore, let $\mathbf{v} \triangleq \kappa(\mathbf{u})$, and assume that for all $\boldsymbol{v} \in \mathbb{C}^n$, the cardinality of the set $\kappa^{-1}(\boldsymbol{v}) \cap \mathcal{M}$ satisfies $|\kappa^{-1}(\boldsymbol{v}) \cap \mathcal{M}| \leq m < \infty$, for some $m \in \mathbb{N}$ (i.e., $\kappa\big|_{\mathcal{M}}$ is finite-to-one). Then:

1. There exist disjoint measurable sets $\{\mathcal{U}_k\}_{k\in[1:m]}$ such that $\kappa\big|_{\mathcal{U}_k}$ is one-to-one for each $k \in [1\colon m]$ and $\bigcup_{k\in[1:m]} \mathcal{U}_k$ covers almost all of $\mathcal{M}$.

2. For every choice of such sets $\{\mathcal{U}_k\}_{k\in[1:m]}$,

$$h(\mathbf{v}) \geq h(\mathbf{u}) + \int_{\mathbb{C}^n} f_{\mathbf{u}}(\boldsymbol{u}) \log(|\boldsymbol{J}_\kappa(\boldsymbol{u})|^2)\, d\boldsymbol{u} - H(\mathsf{k}) \tag{2.22}$$

where k is a discrete random variable that takes on the value k when $\mathbf{u} \in \mathcal{U}_k$ and H denotes entropy.

Proof. See Appendix A.1. □

Since, by Step (ii-b), the mappings $\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}\big|_{\widetilde{\mathcal{M}}}$ are finite-to-one for a.a. $[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}$, we can use Lemma 2.3 with $\mathbf{u} = [\overline{\boldsymbol{sx}}]_{\mathcal{D}}$, $\kappa = \phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}$, $n = \ell$, $m = \widetilde{m}$, and $\mathcal{M} = \widetilde{\mathcal{M}}$ and obtain

$$h\big(\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}([\overline{\mathbf{sx}}]_{\mathcal{D}})\big) \\ \geq h([\overline{\mathbf{sx}}]_{\mathcal{D}}) + \int_{\mathbb{C}^\ell} f_{[\overline{\mathbf{sx}}]_{\mathcal{D}}}([\overline{\boldsymbol{sx}}]_{\mathcal{D}}) \log\left(|\boldsymbol{J}_{\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}}([\overline{\boldsymbol{sx}}]_{\mathcal{D}})|^2\right) d([\overline{\boldsymbol{sx}}]_{\mathcal{D}}) - H(\mathsf{k}_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}) \tag{2.23}$$

where $\mathsf{k}_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}$ corresponds to the random variable k from Lemma 2.3 (since $\kappa = \phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}$, we have a different k for each $[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}$). Because of $[\bar{\mathbf{y}}]_{\mathcal{I}} = \phi_{[\overline{\mathbf{sx}}]_{\mathcal{D}^c}}([\overline{\mathbf{sx}}]_{\mathcal{D}})$, we have $h\big([\bar{\mathbf{y}}]_{\mathcal{I}}\,\big|\,[\overline{\mathbf{sx}}]_{\mathcal{D}^c} = [\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}\big) = h\big(\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}([\overline{\mathbf{sx}}]_{\mathcal{D}})\big)$. Thus, (2.23) entails

$$h\big([\bar{\mathbf{y}}]_{\mathcal{I}}\,\big|\,[\overline{\mathbf{sx}}]_{\mathcal{D}^c}\big) \\ = \mathbb{E}_{[\overline{\mathbf{sx}}]_{\mathcal{D}^c}}\left[h\big(\phi_{[\overline{\mathbf{sx}}]_{\mathcal{D}^c}}([\overline{\mathbf{sx}}]_{\mathcal{D}})\big)\right] \\ \geq\ h([\overline{\mathbf{sx}}]_{\mathcal{D}}) + \mathbb{E}_{[\overline{\mathbf{sx}}]_{\mathcal{D}^c}}\left[\int_{\mathbb{C}^\ell} f_{[\overline{\mathbf{sx}}]_{\mathcal{D}}}([\overline{\boldsymbol{sx}}]_{\mathcal{D}}) \log\left(\left|\boldsymbol{J}_{\phi_{[\overline{\mathbf{sx}}]_{\mathcal{D}^c}}}([\overline{\boldsymbol{sx}}]_{\mathcal{D}})\right|^2\right) d([\overline{\boldsymbol{sx}}]_{\mathcal{D}}) - H\big(\mathsf{k}_{[\overline{\mathbf{sx}}]_{\mathcal{D}^c}}\big)\right]. \tag{2.24}$$

Step (iv): We show now that the right-hand side of (2.24) is lower-bounded by a finite constant. The differential entropy $h([\overline{\mathbf{sx}}]_{\mathcal{D}})$ is the differential entropy of a standard multivariate Gaussian random vector and thus a finite constant. The entropy $H\big(\mathsf{k}_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}\big)$ for a.a. $[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}$ does not exceed $\log(\widetilde{m})$, where $\widetilde{m} = 2^\ell$. Hence, it remains to lower-bound

$$\mathbb{E}_{[\overline{\mathbf{sx}}]_{\mathcal{D}^c}}\left[\int_{\mathbb{C}^\ell} f_{[\overline{\mathbf{sx}}]_{\mathcal{D}}}([\overline{\boldsymbol{sx}}]_{\mathcal{D}}) \log\left(|\boldsymbol{J}_{\phi_{[\overline{\mathbf{sx}}]_{\mathcal{D}^c}}}([\overline{\boldsymbol{sx}}]_{\mathcal{D}})|^2\right) d([\overline{\boldsymbol{sx}}]_{\mathcal{D}})\right] \\ = \int_{\mathbb{C}^{N_\mathsf{s}+N_\mathsf{x}-\ell}} \int_{\mathbb{C}^\ell} f_{[\overline{\mathbf{sx}}]_{\mathcal{D}^c}}([\overline{\boldsymbol{sx}}]_{\mathcal{D}^c})\, f_{[\overline{\mathbf{sx}}]_{\mathcal{D}}}([\overline{\boldsymbol{sx}}]_{\mathcal{D}}) \log\left(|\boldsymbol{J}_{\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}}([\overline{\boldsymbol{sx}}]_{\mathcal{D}})|^2\right) d([\overline{\boldsymbol{sx}}]_{\mathcal{D}})\, \mathrm{d}([\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}) \\ \stackrel{(a)}{=} \int_{\mathbb{C}^{N_\mathsf{s}+N_\mathsf{x}}} f_{\overline{\mathbf{sx}}}(\overline{\boldsymbol{sx}}) \log\left(|\boldsymbol{J}_{\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}}([\overline{\boldsymbol{sx}}]_{\mathcal{D}})|^2\right) d(\overline{\boldsymbol{sx}}) \tag{2.25}$$

where (a) holds because $[\overline{\mathbf{sx}}]_{\mathcal{D}}$ and $[\overline{\mathbf{sx}}]_{\mathcal{D}^c}$ are independent. A similar problem was solved in [Morgenshtern et al., 2013] using Hironaka's theorem on the resolution of singularities. Here, we take a much simpler approach, which relies on the fact that $\det\big(\boldsymbol{J}_{\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}}\big)$ in (2.25) is an analytic function [Rudin, 1987, Ch. 10] that does not vanish identically, and on a property of subharmonic functions[2] (see [Azarin, 2009, Th. 2.6.2.1]).

Lemma 2.4 Let f be an analytic function on $\mathbb{C}^n$ that is not identically zero. Then

$$I_1 \triangleq \int_{\mathbb{C}^n} \exp(-\|\boldsymbol{\xi}\|^2) \log(|f(\boldsymbol{\xi})|)\, d\boldsymbol{\xi} > -\infty\,. \tag{2.26}$$

[2] See [Azarin, 2009, Ch. 2.6] for a definition of subharmonic functions.

Proof. See Appendix A.2. □

The function $f_{\overline{\mathsf{sx}}}$ is the probability density function of a standard multivariate Gaussian random vector. Furthermore, since the function $\det(\boldsymbol{J}_{\phi_{[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c}}}([\overline{\boldsymbol{sx}}]_{\mathcal{D}}))$ is a complex polynomial that is nonzero a.e. (see Step (ii-a)), it is an analytic function that is not identically zero. Hence, by Lemma 2.4, the integral in (2.25) is finite. Thus, with (2.24), we obtain $h\big([\bar{\mathsf{y}}]_{\mathcal{I}}\,\big|\,[\overline{\mathsf{sx}}]_{\mathcal{D}^c}\big) > -\infty$ and, because of (2.17), that $h([\bar{\mathsf{y}}]_{\mathcal{I}}) > -\infty$. This concludes the proof.

Chapter 3

Generic MIMO Block-Fading Channel

As a special setting of the general model (2.1), we consider a MIMO block-fading channel with T transmit antennas, R receive antennas, and block length N. We assume that the fading of different transmit-receive antenna pairs is independent and that the fading vector for each transmit-receive antenna pair depends only on $Q < N$ independent random variables. Furthermore, the statistics of the fading coefficients are assumed to be "generic," i.e., they can be thought of as being generated from an underlying continuous probability distribution.

In this setting, we can use the lower bound on the number of degrees of freedom in Theorem 2.1. In addition, we prove an upper bound that matches our lower bound for a wide range of choices of T, R, Q, and N.

3.1 System Model

The discrete-time fading process associated with each transmit-receive antenna pair conforms to the following channel input-output relations within a given block of N channel uses:

$$\mathsf{y}_r = \sqrt{\rho} \sum_{t\in[1:T]} \operatorname{diag}(\mathsf{h}_{r,t})\, \mathsf{x}_t + \mathsf{w}_r\,, \quad r \in [1\!:\!R]\,. \tag{3.1}$$

Here, $\mathsf{x}_t \in \mathbb{C}^N$ is the signal vector originating from the tth transmit antenna; $\mathsf{y}_r \in \mathbb{C}^N$ is the signal vector at the rth receive antenna; $\mathsf{h}_{r,t} \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma}_{r,t})$ is the vector of N channel coefficients between the tth transmit antenna and the rth receive antenna; $\mathsf{w}_r \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_N)$ is the noise vector at the rth receive antenna; and $\rho \in \mathbb{R}^+$ is the SNR. The vectors $\mathsf{h}_{r,t}$ and w_r are assumed to be mutually independent and independent across $r \in [1:R]$ and $t \in [1:T]$, and to change in an independent fashion from block to block ("block-memoryless" assumption). The transmitted signal vectors x_t are assumed to be independent of the vectors $\mathsf{h}_{r,t}$ and w_r. We consider the noncoherent setting, where transmitter and receiver know the covariance matrix $\mathbf{\Sigma}_{r,t}$ of $\mathsf{h}_{r,t}$ but have no *a priori* knowledge of the realization of $\mathsf{h}_{r,t}$.

Because the covariance matrix $\mathbf{\Sigma}_{r,t}$ is positive-semidefinite, it can be factorized as $\mathbf{\Sigma}_{r,t} = \boldsymbol{Z}_{r,t}\boldsymbol{Z}_{r,t}^{\mathrm{H}}$ with $\boldsymbol{Z}_{r,t} \in \mathbb{C}^{N\times Q}$ and $Q = \operatorname{rank}(\mathbf{\Sigma}_{r,t}) = \operatorname{rank}(\boldsymbol{Z}_{r,t})$. We can then rewrite the channel coefficient vectors $\mathbf{h}_{r,t}$ in terms of $\boldsymbol{Z}_{r,t}$, i.e.,

$$\mathbf{h}_{r,t} = \boldsymbol{Z}_{r,t}\mathbf{s}_{r,t} \tag{3.2}$$

where $\mathbf{s}_{r,t} \in \mathbb{C}^Q$, $\mathbf{s}_{r,t} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_Q)$. Using (3.2), the R input-output relations (3.1) can be rewritten as

$$\mathbf{y}_r = \sqrt{\rho} \sum_{t\in[1:T]} \operatorname{diag}(\boldsymbol{Z}_{r,t}\mathbf{s}_{r,t})\, \mathbf{x}_t + \mathbf{w}_r\,, \quad r \in [1\!:\!R] \tag{3.3}$$

or in stacked form as

$$\mathbf{y} = \sqrt{\rho}\bar{\mathbf{y}} + \mathbf{w}, \quad \text{with } \bar{\mathbf{y}} \triangleq \mathbf{B}\mathbf{s} \tag{3.4}$$

where $\mathbf{y} \triangleq (\mathbf{y}_1^{\mathrm{T}} \cdots \mathbf{y}_R^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{C}^{RN}$, $\mathbf{w} \triangleq (\mathbf{w}_1^{\mathrm{T}} \cdots \mathbf{w}_R^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{C}^{RN}$, $\mathbf{s} \triangleq (\mathbf{s}_1^{\mathrm{T}} \cdots \mathbf{s}_R^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{C}^{RTQ}$ with $\mathbf{s}_r \triangleq (\mathbf{s}_{r,1}^{\mathrm{T}} \cdots\, \mathbf{s}_{r,T}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{C}^{TQ}$, and

$$\mathbf{B} \triangleq \begin{pmatrix} \mathbf{B}_1 & & \\ & \ddots & \\ & & \mathbf{B}_R \end{pmatrix} \in \mathbb{C}^{RN\times RTQ}, \quad \text{with } \mathbf{B}_r \triangleq (\mathbf{X}_1\boldsymbol{Z}_{r,1} \cdots \mathbf{X}_T\boldsymbol{Z}_{r,T}) \in \mathbb{C}^{N\times TQ} \tag{3.5}$$

where $\mathbf{X}_t \triangleq \operatorname{diag}(\mathbf{x}_t) \in \mathbb{C}^{N\times N}$. For later use, we also define $\mathbf{x} \triangleq (\mathbf{x}_1^{\mathrm{T}} \cdots \mathbf{x}_T^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{C}^{TN}$ and

$$\boldsymbol{Z} \triangleq \begin{pmatrix} \boldsymbol{Z}_{1,1} & \cdots & \boldsymbol{Z}_{1,T} \\ \vdots & & \vdots \\ \boldsymbol{Z}_{R,1} & \cdots & \boldsymbol{Z}_{R,T} \end{pmatrix} \in \mathbb{C}^{RN\times TQ}.$$

The matrix $\boldsymbol{Z}$ contains all information about the correlation of the channel coefficients $\mathbf{h}_{r,t}$ (recall that $\boldsymbol{\Sigma}_{r,t} = \boldsymbol{Z}_{r,t}\boldsymbol{Z}_{r,t}^{\mathrm{H}}$). We will refer to $\boldsymbol{Z}$ as *coloring matrix* and use the phrase "for a generic coloring matrix $\boldsymbol{Z}$" to indicate that a property holds for *almost every* matrix $\boldsymbol{Z}$. Here, "almost every" is understood in the precise mathematical sense that the set of all matrices $\boldsymbol{Z}$ for which the property does *not* hold has Lebesgue measure zero.

In the special (nongeneric) case where $Q = 1$ and each $\boldsymbol{Z}_{r,t} \in \mathbb{C}^{N\times 1}$ is the all-one vector, (3.3) reduces to the input-output relation of the MIMO constant block-fading model given by (cf. (1.1))

$$\mathbf{y}_r = \sqrt{\frac{\rho}{T}} \sum_{t\in[1:T]} \mathsf{s}_{r,t}\, \mathbf{x}_t + \mathbf{w}_r\,, \quad r \in [1\!:\!R]\,. \tag{3.6}$$

Comparing (3.3) and (2.1), we see that the generic MIMO block-fading model is a special case of (2.1). More precisely, we can rewrite (3.3) for each receive symbol $\mathsf{y}_{r,i}$ as

$$\begin{aligned} \mathsf{y}_{r,i} &= \sqrt{\rho} \sum_{t\in[1:T]} [\boldsymbol{Z}_{r,t}]_{\{i\}}\mathbf{s}_{r,t}\, \mathsf{x}_{t,i} + \mathsf{w}_{r,i} \\ &= \sqrt{\rho}\,(\mathsf{x}_{1,i}\ \cdots \mathsf{x}_{T,i}) \begin{pmatrix} [\boldsymbol{Z}_{r,1}]_{\{i\}} & & \\ & \ddots & \\ & & [\boldsymbol{Z}_{r,T}]_{\{i\}} \end{pmatrix} \begin{pmatrix} \mathbf{s}_{r,1} \\ \vdots \\ \mathbf{s}_{r,T} \end{pmatrix} + \mathsf{w}_{r,i} \end{aligned} \tag{3.7}$$

for $r \in [1:R]$ and $i \in [1:N]$. Note that $\mathsf{y}_{r,i}$ does not depend on all values of $\mathbf{x}$ and $\mathbf{s}$. Thus, a representation conforming to (2.1) would include many columns and rows containing only zeros. In the representation (3.7) all these rows and columns are omitted and only the relevant entries are stated.

3.2 Characterization of the Number of Degrees of Freedom

3.2.1 Main Result

Recall (see (2.5)) that the capacity of the channel (3.3) is given by

$$C(\rho) = \frac{1}{N} \sup I(\mathbf{x};\mathbf{y}) \tag{3.8}$$

where, the supremum is taken over all probability distributions of $\mathbf{x}$ that satisfy the average-power constraint

$$\mathbb{E}[\|\mathbf{x}\|^2] \leq N\,. \tag{3.9}$$

The number of degrees of freedom is given by

$$\chi = \lim_{\rho \to \infty} \frac{C(\rho)}{\log \rho}\,. \tag{3.10}$$

Our main result is stated in the following theorem.

Theorem 3.1 Let $T < N/Q$ and $R \geq T(N-1)/(N-TQ)$. For a channel conforming to the generic block-fading model, i.e., the channel (3.3) with generic coloring matrix $\boldsymbol{Z}$, the number of degrees of freedom is given by

$$\chi_{\text{gen}} = T\left(1 - \frac{1}{N}\right). \tag{3.11}$$

Proof. In Section 3.3, we will show that χ_{gen} is upper-bounded by $T(1-1/N)$ for all choices of T, R, N, Q, and $\boldsymbol{Z}$. In Section 3.4, we will apply Theorem 2.1 and show that this upper bound is achievable when $T < N/Q$, $R \geq T(N-1)/(N-TQ)$, and $\boldsymbol{Z}$ is generic (see Corollary 3.5). □

3.2.2 Comparison with the Constant Block-Fading Model

Recall that the number of degrees of freedom in the constant block-fading model (3.6) is given by

$$\chi_{\text{const}} = M\left(1 - \frac{M}{N}\right), \quad \text{with } M = \min\left\{T, R, \left\lfloor \frac{N}{2} \right\rfloor\right\}. \tag{3.12}$$

Let us compare the maximal values of χ_{const} and χ_{gen} for a fixed N, which are obtained for optimal choices of T and R. For the constant block-fading model with block length N, it can be easily verified that the number of degrees of freedom χ_{const} given in (3.12) is maximized for

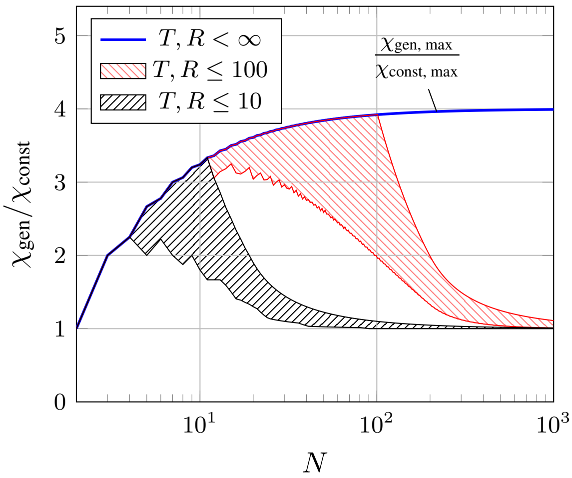

Figure 3.1: Ratio between the maximal value of χ_{gen} (for $Q = 1$) and the maximal value of χ_{const} as a function of N, with and without a constraint on the maximal number of antennas. The shaded areas indicate the regions of $\chi_{\text{gen}}/\chi_{\text{const}}$ delimited by the upper bound (3.13) and lower bound (3.22) on χ_{gen}.

$M = \lfloor N/2 \rfloor$. Setting $T = R = \lfloor N/2 \rfloor$ to obtain $M = \lfloor N/2 \rfloor$, we conclude that the maximal χ_{const} is given by

$$\chi_{\text{const,max}} = \left\lfloor \frac{N}{2} \right\rfloor \left(1 - \frac{\lfloor \frac{N}{2} \rfloor}{N} \right).$$

This can be easily shown to be upper-bounded by $N/4$. For the generic block-fading model with $Q = 1$ and $T < N$, it follows from (3.11) that the number of degrees of freedom is maximized for $T = N - 1$ and $R = (N-1)^2$, which results in

$$\chi_{\text{gen,max}} = \frac{(N-1)^2}{N}.$$

Fig. 3.1 shows the ratio between the maximal value of χ_{gen} (for the case $Q = 1$) and the maximal value of χ_{const} as a function of N. Because for the generic block-fading model the optimal number of receive antennas grows quadratically with N, which may yield an unreasonably large number of antennas for practically relevant values of N (e.g., 1000 symbols or more), in Fig. 3.1 we also show the ratio between the maximal values of χ_{gen} and χ_{const} under a constraint on the maximal number of antennas. For the case $R < T(N-1)/(N-T)$, which is relevant in the constrained setting, our upper and lower bounds on χ_{gen} (see (3.13) and (3.22) below) do not match. The degrees-of-freedom region delimited by the two bounds is represented in Fig. 3.1 by shaded areas. One can see from Fig. 3.1 that $\chi_{\text{gen, max}}$ is about four times $\chi_{\text{const, max}}$ when N grows large. However, when the maximal number of transmit and receive antennas is constrained, the ratio $\chi_{\text{gen}}/\chi_{\text{const}}$ converges to 1.

We emphasize that the only difference between the channel models (3.3) for $Q = 1$ and (3.6) is that the generic (but deterministic) vectors $\boldsymbol{Z}_{r,t}$ of (3.3) are replaced by the all-one vector in (3.6). It is important to note that the generic vectors $\boldsymbol{Z}_{r,t}$ for which (3.11) holds include vectors that are arbitrarily close to the all-one vector. Hence, *arbitrarily small perturbations of the constant block-fading model may result in a significant increase in the number of degrees of freedom.* As we will demonstrate, the potential increase in the number of the degrees of freedom obtained when going from (3.6) to (3.3) is due to the fact that, under the generic block-fading model (3.3), the receive signal vectors in the absence of noise span a subspace of higher dimension than under the constant block-fading model (3.6). We conclude that the commonly used constant block-fading model results in largely pessimistic capacity estimates at high SNR.

3.2.3 Degrees of Freedom Gain

As discussed in Section 3.2.2, (3.11) implies that the maximal achieveable number of degrees of freedom in the generic block-fading model can be about four times as large as the number of degrees of freedom in the constant block-fading model (3.12). We will now provide some intuition regarding this gain. For concreteness, we consider the case $T=2, R=3, Q=1, N=4$. In this case, (3.12) and (3.11) give $\chi_{\text{const}} = 1$ and $\chi_{\text{gen}} = 3/2$, respectively.

The number of degrees of freedom characterizes the channel capacity in a regime where the noise can "effectively" be ignored. Thus, according to the intuitive argumentation in [Morgenshtern et al., 2013, Sec. III], the number of degrees of freedom should be equal to the number of entries of $\mathsf{x} \in \mathbb{C}^8$ that can be deduced from the corresponding receive vector $\mathsf{y} \in \mathbb{C}^{12}$ in the absence of noise, divided by the block length $N=4$.

In the constant block-fading model (3.6), the noiseless receive vectors $\bar{\mathsf{y}}_r = \mathsf{s}_{r,1}\mathsf{x}_1 + \mathsf{s}_{r,2}\mathsf{x}_2$, $r = 1,2,3$ belong to the two-dimensional subspace spanned by $\{\mathsf{x}_1, \mathsf{x}_2\}$. Hence, the receive vectors $\bar{\mathsf{y}}_1, \bar{\mathsf{y}}_2, \bar{\mathsf{y}}_3$ are linearly dependent, and two of them contain all the information available about x. From two of the receive vectors, we obtain $2 \cdot 4$ scalar equations in $8+4$ scalar variables $(\mathsf{x}, \mathsf{s}_{1,1}, \mathsf{s}_{1,2}, \mathsf{s}_{2,1}, \mathsf{s}_{2,2})$. Since we do not have control of the variables $\mathsf{s}_{r,t}$, one way to reconstruct x is to fix four of its entries (or, equivalently, to transmit four pilot symbols) to obtain eight equations in eight variables. By solving this system of equations, we obtain the remaining four entries of x. Hence, we can deduce four entries of x from $\bar{\mathsf{y}}$. We conclude that the number of degrees of freedom is $4/4 = 1$, which is in agreement with (3.12).

In the generic block-fading model (3.3), on the other hand, the receive vectors without noise

$$\bar{\mathsf{y}}_r = \operatorname{diag}(\boldsymbol{Z}_{r,1}\mathsf{s}_{r,1})\mathsf{x}_1 + \operatorname{diag}(\boldsymbol{Z}_{r,2}\mathsf{s}_{r,2})\mathsf{x}_2, \quad r = 1,2,3$$

span a three-dimensional subspace almost surely. Hence, we obtain a system of $3 \cdot 4$ equations in $8+6$ variables $(\mathsf{x}, \mathsf{s}_{1,1}, \mathsf{s}_{1,2}, \mathsf{s}_{2,1}, \mathsf{s}_{2,2}, \mathsf{s}_{3,1}, \mathsf{s}_{3,2})$. Fixing two entries of x, we are able to recover the remaining six entries. Hence, the number of degrees of freedom is $6/4 = 3/2$, which is in agreement with (3.11).

This argument suggests that the reason why the generic block-fading model yields a larger number of degrees of freedom than the constant block-fading model is that the noiseless receive vectors span a subset of $\mathbb{C}^N$ of higher dimension.

3.3 Upper Bound

The results in this section provide the first part of the proof of Theorem 3.11. However, the following upper bound on the number of degrees of freedom of the channel (3.3) holds for every T, R, Q, N, and $\boldsymbol{Z}$. The assumption of a generic coloring matrix $\boldsymbol{Z}$ required in Theorem 3.11 is not necessary.

Theorem 3.2 The number of degrees of freedom of the channel (3.3) satisfies

$$\chi_{\text{gen}} \leq T\left(1-\frac{1}{N}\right). \tag{3.13}$$

Proof. We will show that the number of degrees of freedom is upper-bounded by T times the number of degrees of freedom of a constant block-fading SIMO channel; the result then follows from (3.12). To this end, we will rewrite each output vector $\mathbf{y}_r$ as the sum of the output vectors of T SIMO systems with RQ receive antennas each. This will be achieved by splitting the additive noise variables appropriately.

From (3.3), the ith entry of the receive vector $\mathbf{y}_r$ is given by

$$[\mathbf{y}_r]_i = \sqrt{\rho} \sum_{t\in[1:T]} \sum_{q\in[1:Q]} [\boldsymbol{Z}_{r,t}]_i^q \, [\mathbf{s}_{r,t}]_q \, [\mathbf{x}_t]_i + [\mathbf{w}_r]_i \tag{3.14}$$

for $r \in [1:R]$. We first decompose the noise variables according to

$$[\mathbf{w}_r]_i = \sum_{t\in[1:T]} \sum_{q\in[1:Q]} \frac{[\boldsymbol{Z}_{r,t}]_{i,q}}{\sqrt{K}} [\tilde{\mathbf{w}}_{q,r,t}]_i + [\mathbf{w}'_r]_i \,. \tag{3.15}$$

Here, all $[\tilde{\mathbf{w}}_{q,r,t}]_i$ and $[\mathbf{w}'_r]_i$ are mutually independent and independent of all $\mathbf{x}_t$ and $\mathbf{s}_{r,t}$. Furthermore, $[\tilde{\mathbf{w}}_{q,r,t}]_i \sim \mathcal{CN}(0,1)$,

$$[\mathbf{w}'_r]_i \sim \mathcal{CN}\left(0, 1 - \sum_{t\in[1:T]} \sum_{q\in[1:Q]} \frac{|[\boldsymbol{Z}_{r,t}]_{i,q}|^2}{K}\right),$$

and K is a finite constant satisfying[1]

$$K > \max_{r\in[1:R],\, i\in[1:N]} \sum_{t\in[1:T]} \sum_{q\in[1:Q]} |[\boldsymbol{Z}_{r,t}]_{i,q}|^2 \,.$$

We next define T "virtual" constant block-fading SIMO channels with RQ receive antennas each:

$$[\tilde{\mathbf{y}}_{q,r,t}]_i = \sqrt{K\rho}\, [\mathbf{s}_{r,t}]_q \, [\mathbf{x}_t]_i + [\tilde{\mathbf{w}}_{q,r,t}]_i \,, \quad \text{with } i\in[1:N], r\in[1:R], q\in[1:Q] \tag{3.16}$$

for $t \in [1:T]$. Inserting (3.15) into (3.14) and using (3.16), it can be verified that (3.14) can be

[1]This condition on K is required to ensure that the variance of all random variables $[\mathbf{w}'_r]_i$ is positive.

rewritten as

$$[\mathbf{y}_r]_i = \frac{1}{\sqrt{K}} \sum_{t\in[1:T]} \sum_{q\in[1:Q]} [\boldsymbol{Z}_{r,t}]_{i,q} [\tilde{\mathbf{y}}_{q,r,t}]_i + [\mathbf{w}'_r]_i \,. \tag{3.17}$$

Let $\tilde{\mathbf{y}}_t \triangleq (\tilde{\mathbf{y}}_{1,1,t}^{\mathrm{T}} \cdots \tilde{\mathbf{y}}_{Q,R,t}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{C}^{QRN}$. By (3.17), the random variable $\mathbf{y}$ depends on $\mathbf{x}$ only via the random variables $\{\tilde{\mathbf{y}}_t\}_{t\in[1:T]}$. Hence, the data-processing inequality [Gallager, 1968, eq. (2.3.19)] yields

$$I(\mathbf{x};\mathbf{y}) \le I(\mathbf{x};\tilde{\mathbf{y}}_1,\ldots,\tilde{\mathbf{y}}_T)\,. \tag{3.18}$$

The right-hand side of (3.18) can be upper-bounded as follows:

$$\begin{aligned} I(\mathbf{x};\tilde{\mathbf{y}}_1,\ldots,\tilde{\mathbf{y}}_T) &= h(\tilde{\mathbf{y}}_1,\ldots,\tilde{\mathbf{y}}_T) - h(\tilde{\mathbf{y}}_1,\ldots,\tilde{\mathbf{y}}_T|\mathbf{x}) \\ &\overset{(a)}{=} h(\tilde{\mathbf{y}}_1,\ldots,\tilde{\mathbf{y}}_T) - \sum_{t\in[1:T]} h(\tilde{\mathbf{y}}_t|\mathbf{x}_t) \\ &\overset{(b)}{\le} \sum_{t\in[1:T]} \big[h(\tilde{\mathbf{y}}_t) - h(\tilde{\mathbf{y}}_t|\mathbf{x}_t)\big] \\ &= \sum_{t\in[1:T]} I(\mathbf{x}_t;\tilde{\mathbf{y}}_t)\,. \end{aligned} \tag{3.19}$$

Here, (a) holds because $\tilde{\mathbf{y}}_1,\ldots,\tilde{\mathbf{y}}_T$ are conditionally independent given $\mathbf{x}$, and (b) follows from the chain rule for differential entropy and because conditioning does not increase differential entropy. Since (by assumption) the input vector $\mathbf{x}$ satisfies the power constraint (3.9), we conclude that, trivially, also each subvector $\mathbf{x}_t$ satisfies the individual power constraint $\mathbb{E}[\|\mathbf{x}_t\|^2] \le N$. Thus, the SNR (i.e., the expected power of the noiseless receive signal divided by the noise power) of each "virtual" constant block-fading SIMO channel (3.16) is given by

$$\frac{\mathbb{E}[\|\sqrt{K\rho}\,[\mathbf{s}_{r,t}]_q \mathbf{x}_t\|^2]}{\mathbb{E}[\|\tilde{\mathbf{w}}_{q,r,t}\|^2]} = \frac{K\rho\,\mathbb{E}[|[\mathbf{s}_{r,t}]_q|^2]\,\mathbb{E}[\|\mathbf{x}_t\|^2]}{\mathbb{E}[\|\tilde{\mathbf{w}}_{q,r,t}\|^2]} \le \frac{K\rho\,N}{N} = K\rho\,.$$

By (3.12) and (2.8), the capacity of a constant block-fading SIMO channel of SNR $K\rho$ is of the form[2] $(1-1/N)\log(K\rho)+o(\log\rho)$. Since, by (3.8), the capacity is the supremum of the mutual information divided by the block length, we can upper-bound each mutual information $I(\mathbf{x}_t;\tilde{\mathbf{y}}_t)$, $t\in[1\!:\!T]$ by N times the capacity. This results in

$$I(\mathbf{x}_t;\tilde{\mathbf{y}}_t) \le N\bigg(\Big(1-\frac{1}{N}\Big)\log(K\rho) + o(\log\rho)\bigg) = (N-1)\log(K\rho) + o(\log\rho)\,.$$

Hence, continuing (3.18) and (3.19), we obtain

$$\begin{aligned} I(\mathbf{x};\mathbf{y}) &\le \sum_{t\in[1:T]} I(\mathbf{x}_t;\tilde{\mathbf{y}}_t) \\ &\le T(N-1)\log(K\rho) + o(\log\rho) \end{aligned}$$

[2]Since the number of transmit antennas is one for a SIMO channel, we have $M=1$ in (1.3).

$$\stackrel{(a)}{=} T(N-1)\log\rho + o(\log\rho) \tag{3.20}$$

where (a) holds because $\log(K\rho) = \log\rho + \log K$. Thus, the mutual information $I(\mathbf{x};\mathbf{y})$ with $\mathbf{x}$ satisfying the power constraint (3.9) is upper-bounded by (3.20). Inserting (3.20) into (3.8) yields

$$C(\rho) \le T\frac{N-1}{N}\log\rho + o(\log\rho)$$

from which (3.13) follows via (3.10). □

3.4 Lower Bound

A special case of the lower bound presented in this section concludes the proof of Theorem 3.11 (see Property (ii) in Corollary 3.5).

We first derive a lower bound on χ_{gen} assuming that $\widetilde{T} \le \min\{T,R\}$ transmit antennas are effectively used (i.e., $\mathbf{x}_{\widetilde{T}+1},\dots,\mathbf{x}_T$ are set to zero). Then we maximize the lower bound by identifying the optimal number $\widetilde{T}$ of transmit antennas to use.

Proposition 3.3 The number of degrees of freedom of the channel (3.3) for a generic coloring matrix $\boldsymbol{Z}$ is lower-bounded by

$$\chi_{\text{gen}} \;\ge\; \chi_{\text{low}}(\widetilde{T}) \;\triangleq\; \min\left\{\widetilde{T}\left(1-\frac{1}{N}\right), R\left(1-\frac{\widetilde{T}Q}{N}\right)\right\} \tag{3.21}$$

for all $\widetilde{T} \le \min\{T,R\}$.

Proof. The proof is an application of the general lower bound Theorem 2.1. The details will be presented in Section 3.5. □

The minimum in (3.21) is given by $\chi_{\text{low}}(\widetilde{T}) = \widetilde{T}(1-1/N)$ when the number R of receive antennas is large enough (i.e., $R \ge \widetilde{T}(N-1)/(N-\widetilde{T}Q)$). In contrast, $\chi_{\text{low}}(\widetilde{T}) = R(1-\widetilde{T}Q/N)$ when the number of degrees of freedom is constrained by the limited number of receive antennas (i.e., $R < \widetilde{T}(N-1)/(N-\widetilde{T}Q)$).

The main result of this section is stated in the following theorem.

Theorem 3.4 The number of degrees of freedom of the channel (3.3) for a generic coloring matrix $\boldsymbol{Z}$ is lower-bounded by

$$\chi_{\text{gen}} \ge \chi^*_{\text{low}} \triangleq \max_{\widetilde{T}\le\min\{T,R\}} \chi_{\text{low}}(\widetilde{T}) = \begin{cases} T\left(1-\dfrac{1}{N}\right), & \text{if } T \le T_{\text{opt}} \\ \eta, & \text{if } T > T_{\text{opt}} \end{cases} \tag{3.22}$$

where

$$T_{\text{opt}} \triangleq \frac{RN}{N+RQ-1} \tag{3.23}$$

and

$$\eta \triangleq \max\left\{R\left(1-\frac{\lceil T_{\text{opt}}\rceil Q}{N}\right), \lfloor T_{\text{opt}}\rfloor\left(1-\frac{1}{N}\right)\right\}. \tag{3.24}$$

Proof. The idea behind the bound χ^*_{low} in (3.22) is to obtain the tightest (i.e., largest) of the lower bounds $\chi_{\text{low}}(\widetilde{T})$ in (3.21) for T transmit antennas by maximizing $\chi_{\text{low}}(\widetilde{T})$ with respect to the number of effectively used transmit antennas $\widetilde{T} \leq \min\{T, R\}$. According to (3.21), $\chi_{\text{low}}(\widetilde{T})$ is the minimum of two quantities where the first, $\widetilde{T}(1-1/N)$, is monotonically increasing in $\widetilde{T}$ and the second, $R(1-\widetilde{T}Q/N)$, is monotonically decreasing in $\widetilde{T}$. Hence, $\chi_{\text{low}}(\widetilde{T})$ attains its maximum at the intersection point T_{opt} defined in (3.23). If $T \leq T_{\text{opt}}$, we are for all $\widetilde{T} \leq \min\{T, R\}$ in the regime where $\chi_{\text{low}}(\widetilde{T})$ is monotonically increasing, and thus the best choice is to use $\widetilde{T} = T$ transmit antennas (note that because $T \leq T_{\text{opt}} \overset{(3.23)}{\leq} RN/N = R$, the choice $\widetilde{T} = T$ in Proposition 3.3 is possible). Thus, in this case we have $\chi^*_{\text{low}} = \chi_{\text{low}}(T) = T(1-1/N)$, which yields the first case in (3.22). If $T > T_{\text{opt}}$, we would like to use T_{opt} transmit antennas, but we have to take into account that T_{opt} may be noninteger. Thus, we take the maximum of the bounds $\chi_{\text{low}}(\widetilde{T})$ resulting from the closest integers, $\chi_{\text{low}}(\lfloor T_{\text{opt}}\rfloor)$ and $\chi_{\text{low}}(\lceil T_{\text{opt}}\rceil)$, which yields η in (3.24). This concludes the proof. □

Remark 3.1 For $N \geq 2$, the optimal number of transmit antennas T_{opt} is upper-bounded as follows:

$$T_{\text{opt}} < \frac{N}{Q}. \tag{3.25}$$

In fact, $T_{\text{opt}} = RN/(N+RQ-1) < RN/(RQ) = N/Q$.

Remark 3.2 For $N = Q \geq 2$, we have by (3.25) that $T_{\text{opt}} < 1$. Hence, $T > T_{\text{opt}}$ and thus, by (3.22) and (3.24), $\chi^*_{\text{low}} = \eta = \max\{R(1-Q/N), 0\} = 0$. Similarly, we obtain for $N = 1$ that $\chi_{\text{low}}(\widetilde{T}) \leq 0$ for all $\widetilde{T}$, which yields $\chi^*_{\text{low}} \leq 0$. Hence, our lower bound χ^*_{low} is trivial. In these scenarios, the capacity grows double-logarithmically in the SNR ρ [Lapidoth and Moser, 2003, Durisi and Bölcskei, 2011].

Remark 3.3 The lower bound χ^*_{low} in (3.22) can be equivalently expressed as

$$\chi^*_{\text{low}} = \min\left\{T\left(1-\frac{1}{N}\right), \eta\right\}.$$

Corollary 3.5 Let $N \geq 2$. For the lower bound χ^*_{low} in Theorem 3.4, the following properties hold:

(i) For $T \geq N/Q$, we have $T > T_{\text{opt}}$ and $\chi^*_{\text{low}} = \eta$.

(ii) For $T < N/Q$ and $R \geq T(N-1)/(N-TQ)$, we have $T \leq T_{\text{opt}}$ and $\chi^*_{\text{low}} = T(1-1/N)$.

(iii) For $T < N/Q$ and $R < T(N-1)/(N-TQ)$, we have $T > T_{\text{opt}}$ and $\chi^*_{\text{low}} = \eta$.

(iv) For fixed N and Q, the lower bound χ^*_{low} attains its maximal value for $T = \lfloor (N-1)/Q\rfloor$ transmit antennas and $R = \lceil (N-1)^2/Q\rceil$ receive antennas; this maximal value of χ^*_{low} equals $\lfloor (N-1)/Q\rfloor(1-1/N)$.

Proof. By (3.25), the inequality $T \geq N/Q$ implies $T > T_{\text{opt}}$, from which Property (i) follows by (3.22). For $T < N/Q$, the following equivalence holds:

$$T \leq T_{\text{opt}} \overset{(3.23)}{=} \frac{RN}{N+RQ-1} \quad \Leftrightarrow \quad T\frac{N-1}{N-TQ} \leq R\,. \tag{3.26}$$

Thus, the conditions in Properties (ii) and (iii) imply $T \leq T_{\text{opt}}$ and $T > T_{\text{opt}}$, respectively, and the expressions of χ^*_{low} given in Properties (ii) and (iii) follow immediately from the case distinction in (3.22).

To prove Property (iv), we first show that $\chi^*_{\text{low}} \leq \lfloor (N-1)/Q \rfloor (1-1/N)$ for arbitrary T and R. Subsequently, we will show that this upper bound is achievable for the proposed number of antennas. We first note that for each $\widetilde{T} \leq N/Q$, the lower bound $\chi_{\text{low}}(\widetilde{T})$ in (3.21) is monotonically nondecreasing in R. Furthermore, for $\widetilde{T} > N/Q$, $\chi_{\text{low}}(\widetilde{T})$ is negative and can be ignored in the maximization process, i.e., we have $\chi^*_{\text{low}} = \max_{\widetilde{T} \leq \min\{T,R,N/Q\}} \chi_{\text{low}}(\widetilde{T})$. This implies that χ^*_{low} is—as a maximum of nondecreasing functions—also monotonically nondecreasing in R. Hence, to obtain an upper bound on χ^*_{low}, we can assume R arbitrarily large without loss of generality. We choose $R > (N-1)^2/Q$. Simple algebraic manipulations yield the equivalence

$$R > \frac{(N-1)^2}{Q} \quad \Leftrightarrow \quad T_{\text{opt}} = \frac{RN}{N+RQ-1} > \frac{N-1}{Q}\,. \tag{3.27}$$

This implies $\lceil T_{\text{opt}} \rceil Q \geq N-1$ and further, because both sides of this strict inequality are integers, that $\lceil T_{\text{opt}} \rceil Q > N$. Thus, the first argument of the maximum defining η in (3.24) satisfies

$$R\left(1 - \frac{\lceil T_{\text{opt}} \rceil Q}{N}\right) \leq R(1-1) = 0$$

and, hence, η reduces to $\eta = \lfloor T_{\text{opt}} \rfloor (1-1/N)$. By (3.22), we have that χ^*_{low} is either equal to $T(1-1/N)$ (for $T \leq T_{\text{opt}}$) or equal to $\eta = \lfloor T_{\text{opt}} \rfloor (1-1/N)$ (for $T > T_{\text{opt}}$). In both cases we have $\chi^*_{\text{low}} \leq \lfloor T_{\text{opt}} \rfloor (1-1/N)$. Since $\lfloor T_{\text{opt}} \rfloor \leq \lfloor (N-1)/Q \rfloor$ by[3] (3.25), this implies $\chi^*_{\text{low}} \leq \lfloor (N-1)/Q \rfloor (1-1/N)$.

It remains to be shown that this upper bound is achievable. For $R = \lceil (N-1)^2/Q \rceil \geq (N-1)^2/Q$, we obtain (see (3.27) with "$>$" replaced by "$\geq$") that $T_{\text{opt}} \geq (N-1)/Q$. Hence, for $T = \lfloor (N-1)/Q \rfloor \leq T_{\text{opt}}$, the lower bound (3.22) simplifies to $\chi^*_{\text{low}} = T(1-1/N) = \lfloor (N-1)/Q \rfloor (1-1/N)$. Thus, we have shown that χ^*_{low} is maximized for $T = \lfloor (N-1)/Q \rfloor$ and $R = \lceil (N-1)^2/Q \rceil$ and its maximum equals $\lfloor (N-1)/Q \rfloor (1-1/N)$. □

Remark 3.4 Property (ii) in Corollary 3.5 shows that for a fixed $T < N/Q$, we can achieve $\chi^*_{\text{low}} = T(1-1/N)$ by using a sufficiently large number of receive antennas R. This coincides with the upper bound presented in Section 3.3. Thus, in this regime, the number of degrees of freedom grows linearly in the number of transmit antennas.

[3]By (3.25), $\lfloor T_{\text{opt}} \rfloor < N/Q$ and thus $Q\lfloor T_{\text{opt}} \rfloor < N$. Since both sides of this strict inequality are integers, we have $Q\lfloor T_{\text{opt}} \rfloor \leq N-1$ and hence $\lfloor T_{\text{opt}} \rfloor \leq (N-1)/Q$, which in turn implies $\lfloor T_{\text{opt}} \rfloor \leq \lfloor (N-1)/Q \rfloor$.

3.5 Proof of Proposition 3.3

In this section, we establish the lower bound (3.21). For $N \leq \widetilde{T}Q$, the inequality in (3.21) is trivially true, because in this case $R(1 - \widetilde{T}Q/N) \leq 0$ and hence $\chi_{\text{low}} \leq 0$. Therefore, we focus on the case

$$N > \widetilde{T}Q$$

which will thus be assumed in the remainder of this section. Furthermore, recall that we assumed in Proposition 3.3 that $\widetilde{T} \leq \min\{T, R\}$. Thus, setting $\mathbf{x}_{\widetilde{T}+1}, \ldots, \mathbf{x}_T$ to zero, we can replace T by $\widetilde{T}$ in the input-output relation (3.4). Finally, we shall assume that

$$R \leq \left\lceil \frac{\widetilde{T}(N-1)}{N - \widetilde{T}Q} \right\rceil.$$

If more receive antennas are available, we simply turn them off. The following dimension counting argument provides some intuition on why the use of more than $\lceil \widetilde{T}(N-1)/(N - \widetilde{T}Q) \rceil$ receive antennas is not beneficial.

3.5.1 Dimension Counting

The noiseless receive vector $\bar{\mathbf{y}} = \mathbf{B}\mathbf{s} \in \mathbb{C}^{RN}$ in (3.4) corresponds to RN polynomial equations. The unknown variables of these equations are the entries of the vectors $\mathbf{s}_{r,t} \in \mathbb{C}^Q$, $r \in [1:R]$, $t \in [1:\widetilde{T}]$ ($R\widetilde{T}Q$ unknown variables) and of the transmitted signal vectors $\mathbf{x}_t \in \mathbb{C}^N$, $t \in [1:\widetilde{T}]$ ($\widetilde{T}N$ unknown variables). Consider now a pair $(\mathbf{x}_t, \mathbf{s}_{r,t})$, consisting of a transmitted signal vector $\mathbf{x}_t$ and a fading vector $\mathbf{s}_{r,t}$ that is a solution of $\bar{\mathbf{y}} = \mathbf{B}\mathbf{s}$. Then the pair $(c_t\mathbf{x}_t, \mathbf{s}_{r,t}/c_t)$, where c_t is an arbitrary nonzero constant, is also a solution of $\bar{\mathbf{y}} = \mathbf{B}\mathbf{s}$. This implies that each $\mathbf{x}_t$ can be recovered from $\bar{\mathbf{y}}$ only up to a scaling factor. To resolve this ambiguity, we fix one entry in each $\mathbf{x}_t$. Hence, the total number of unknown variables becomes $R\widetilde{T}Q + \widetilde{T}N - \widetilde{T}$. As long as the number of equations is larger than or equal to the number of unknown variables, i.e., $RN \geq R\widetilde{T}Q + \widetilde{T}N - \widetilde{T}$, we are able to recover[4] the $N-1$ unknown entries of each $\mathbf{x}_t$. The above condition is equivalent to $R \geq \widetilde{T}(N-1)/(N - \widetilde{T}Q)$. Hence, it is reasonable to consider only the case $R \leq \lceil \widetilde{T}(N-1)/(N - \widetilde{T}Q) \rceil$, as the receive vectors resulting from the use of additional receive antennas would not help us gain more information about the transmit vectors $\{\mathbf{x}_t\}_{t \in [1:\widetilde{T}]}$.

3.5.2 Application of Theorem 2.1

Our proof is based on the general lower bound Theorem 2.1. In order to apply Theorem 2.1, we have to show that the Jacobian matrix of the polynomial mapping $\phi\colon \mathbb{C}^{R\widetilde{T}Q+\widetilde{T}N} \to \mathbb{C}^{RN}$ defined by $\phi(\boldsymbol{s}, \boldsymbol{x}) = \bar{\boldsymbol{y}}$ has a nonsingular submatrix for a specific choice of $\boldsymbol{s}$ and $\boldsymbol{x}$.

[4] Strictly speaking, this argument is true for linear equations. In our case, because we have polynomial rather than linear equations, we obtain in general a finite number of solutions for the variables $\mathbf{x}$ and not a unique solution.

A nonsingular submatrix of rank ℓ gives us the lower bound $\chi_{\text{gen}} \geq (\ell - R\widetilde{T}Q)/N$ because the dimension of the vector $\mathbf{s}$ is $N_{\mathbf{s}} = R\widetilde{T}Q$. Thus, to obtain the lower bound presented in Proposition 3.3 we have to choose

$$\ell \triangleq \min\{RN, R\widetilde{T}Q + \widetilde{T}N - \widetilde{T}\}\,. \tag{3.28}$$

If $\ell = RN$ we have to find a full rank submatrix of $\boldsymbol{J}_\phi$, i.e., we only reduce the number of columns, but keep all rows. Otherwise, i.e., if $\ell = R\widetilde{T}Q + \widetilde{T}N - \widetilde{T}$ we can get rid of $RN - (R\widetilde{T}Q + \widetilde{T}N - \widetilde{T})$ rows and find a smaller nonsingular submatrix. Since we assumed that $R \leq \lceil \widetilde{T}(N-1)/(N - \widetilde{T}Q) \rceil$ and $N > \widetilde{T}Q$,

$$\begin{aligned} RN - (R\widetilde{T}Q + \widetilde{T}N - \widetilde{T}) &= R(N - \widetilde{T}Q) - \widetilde{T}(N-1) \\ &\leq \left\lceil \frac{\widetilde{T}(N-1)}{N - \widetilde{T}Q} \right\rceil (N - \widetilde{T}Q) - \widetilde{T}(N-1) \\ &= \underbrace{\left(\left\lceil \frac{\widetilde{T}(N-1)}{N - \widetilde{T}Q} \right\rceil - \frac{\widetilde{T}(N-1)}{N - \widetilde{T}Q} \right)}_{<1} (N - \widetilde{T}Q) \\ &< N - \widetilde{T}Q\,. \end{aligned} \tag{3.29}$$

we remove at most $N - \widetilde{T}Q - 1$ rows. We specify the rows we keep by $\mathcal{I} \triangleq [1:\ell]$.

Our goal is to show that there exists a set of indices $\mathcal{D}$, and vectors $\boldsymbol{s} \in \mathbb{C}^{R\widetilde{T}Q}$ and $\boldsymbol{x} \in \mathbb{C}^{\widetilde{T}N}$ such that the submatrix $[\boldsymbol{J}_\phi(\boldsymbol{s}, \boldsymbol{x})]_{\mathcal{I}}^{\mathcal{D}}$ is nonsingular for almost all coloring matrices $\boldsymbol{Z}$. To this end, we first look at the matrix $[\boldsymbol{J}_\phi(\boldsymbol{s}, \boldsymbol{x})]_{\mathcal{I}}^{\mathcal{D}}$ in more detail. Recall that

$$\phi(\boldsymbol{s}, \boldsymbol{x}) = \bar{\boldsymbol{y}} = \boldsymbol{B}\boldsymbol{s}, \quad \text{with } \boldsymbol{B} = \begin{pmatrix} \boldsymbol{B}_1 & & \\ & \ddots & \\ & & \boldsymbol{B}_R \end{pmatrix} \tag{3.30}$$

where

$$\boldsymbol{B}_r = (\boldsymbol{X}_1 \boldsymbol{Z}_{r,1} \cdots \boldsymbol{X}_{\widetilde{T}} \boldsymbol{Z}_{r,\widetilde{T}}), \quad \text{with } \boldsymbol{X}_t = \operatorname{diag}(\boldsymbol{x}_t)\,. \tag{3.31}$$

The Jacobian matrix $[\boldsymbol{J}_\phi(\boldsymbol{s}, \boldsymbol{x})]_{\mathcal{I}}^{\mathcal{D}}$ is equal to

$$[\boldsymbol{J}_\phi(\boldsymbol{s}, \boldsymbol{x})]_{\mathcal{I}}^{\mathcal{D}} = [(\boldsymbol{B}\ \ \boldsymbol{A})]_{\mathcal{I}}^{\mathcal{D}} \in \mathbb{C}^{\ell \times \ell}, \quad \text{with } \boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{1,1} & \cdots & \boldsymbol{A}_{1,\widetilde{T}} \\ \vdots & & \vdots \\ \boldsymbol{A}_{R,1} & \cdots & \boldsymbol{A}_{R,\widetilde{T}} \end{pmatrix} \in \mathbb{C}^{RN \times \widetilde{T}N} \tag{3.32}$$

where

$$\boldsymbol{A}_{r,t} \triangleq \operatorname{diag}(\boldsymbol{a}_{r,t}), \quad t \in [1:\widetilde{T}],\, r \in [1:R]\,, \quad \text{with } \boldsymbol{a}_{r,t} \triangleq \boldsymbol{Z}_{r,t}\boldsymbol{s}_{r,t}\,. \tag{3.33}$$

We have to find $(\boldsymbol{s}, \boldsymbol{x})$ such that $[\boldsymbol{J}_\phi(\boldsymbol{s}, \boldsymbol{x})]_{\mathcal{I}}^{\mathcal{D}}$ is nonsingular (i.e., $|[\boldsymbol{J}_\phi(\boldsymbol{s}, \boldsymbol{x})]_{\mathcal{I}}^{\mathcal{D}}| \neq 0$) for a generic coloring matrix $\boldsymbol{Z}$. For fixed values $(\boldsymbol{s}, \boldsymbol{x})$, the determinant of $[\boldsymbol{J}_\phi(\boldsymbol{s}, \boldsymbol{x})]_{\mathcal{I}}^{\mathcal{D}}$ is a polynomial $p(\boldsymbol{Z})$ (i.e., a polynomial in all the entries of $\boldsymbol{Z}$). We will show that $p(\boldsymbol{Z})$ does not vanish at a specific point $(\tilde{\boldsymbol{Z}})$, i.e., $p(\tilde{\boldsymbol{Z}}) \neq 0$. This implies that $p(\boldsymbol{Z})$ is not identically zero. Since a polynomial vanishes either identically or on a set of measure zero [Gunning and Rossi, 1965, Cor. 10], we conclude that $p(\boldsymbol{Z}) \neq 0$ for $\boldsymbol{Z} \in \mathcal{Z}$, where $\mathcal{Z}$ is a set with a complement of Lebesgue measure zero. In other words, for a generic coloring matrix $\boldsymbol{Z}$, the matrix $[\boldsymbol{J}_\phi(\boldsymbol{s}, \boldsymbol{x})]_{\mathcal{I}}^{\mathcal{D}}$ is nonsingular.

It remains to find values $(\boldsymbol{s}, \boldsymbol{x})$ and a realization $\tilde{\boldsymbol{Z}}$ such that $p(\tilde{\boldsymbol{Z}}) \neq 0$. This, in turn, requires to find a specific set $\mathcal{D}$. This is done in the proof of the following lemma.

Lemma 3.6 Let $R \geq \widetilde{T}$, $N > \widetilde{T}Q$, and $R \leq \lceil \widetilde{T}(N-1)/(N-\widetilde{T}Q) \rceil$. Then there exists a triple $(\boldsymbol{Z}, \boldsymbol{s}, \boldsymbol{x})$ and a choice of $\mathcal{D}$ for which the determinant of the Jacobian matrix $[\boldsymbol{J}_\phi(\boldsymbol{s}, \boldsymbol{x})]_{\mathcal{I}}^{\mathcal{D}}$ in (3.32) is nonzero.

Proof. See Appendix A.3. □

This concludes the proof of Proposition 3.3.

Chapter 4

Oversampled SISO Block-Fading Channel

We consider the continuous-time, time-selective, frequency-flat, Rayleigh block-fading single-input single-output (SISO) channel introduced in [Liang and Veeravalli, 2004]. A discretization of the channel using a matched filter and sampling at the symbol rate results in a well-known discrete-time block-fading channel with a known number of degrees of freedom. To investigate whether this approach is optimal, we consider a different filter and a higher sampling rate. This results in a discrete-time block-fading channel that can be interpreted as a special case of the general model (2.1). Thus, we can use Theorem 2.1 to obtain a lower bound on the number of degrees of freedom. This bound shows that the number of degrees of freedom is higher than for the standard matched-filter symbol-rate-sampling approach.

4.1 System Model

We consider the continuous-time, time-selective, Rayleigh-fading channel

$$\mathsf{y}(t) = \mathsf{h}(t)\mathsf{x}(t) + \mathsf{w}(t). \tag{4.1}$$

Here, $\mathsf{h}(t)$ is the channel fading process, $\mathsf{x}(t)$ is the transmit signal, $\mathsf{w}(t)$ is additive white Gaussian noise, and $\mathsf{y}(t)$ denotes the channel output. All these random quantities are complex.

We restrict ourselves to transmit signals of the form

$$\mathsf{x}(t) = \sum_{i=1}^{\infty} \sqrt{\rho}\, \mathsf{x}_i\, p(t-(i-1)T_S) \tag{4.2}$$

where in (4.2) the pulse $p(t)$ has unit energy and $p(t) = 0$ if $t \notin (0, T_S)$, with T_S being the symbol duration. For simplicity, in the following we will assume that $p(t)$ is a rectangular pulse, i.e.,

$$p(t) = \frac{1}{\sqrt{T_S}} \mathbb{1}_{[0,T_S)}(t).$$

Recall that our aim is to establish an achievability result, i.e., a lower bound on the number of degrees of freedom. Hence, we are allowed to select a specific pulse shape. The choice of a rectangular pulse is convenient because it yields simpler mathematical expressions. In practice, pulses with lower side lobes in frequency are preferable. Although our proof can be generalized to a larger family of pulse shapes, we decided to omit this extension because it is rather technical and may obfuscate the actual contribution.

We assume that the additive noise process $\mathsf{w}(t)$ is white zero-mean proper complex Gaussian. The channel fading process $\mathsf{h}(t)$ is also assumed zero-mean proper complex Gaussian. Furthermore, we consider a block-fading setting, i.e., we assume that the fading changes independently between blocks of a time duration $T = NT_S$. On each block (for simplicity we consider $t \in [0, T]$) we can decompose the fading process according to

$$\mathsf{h}(t) = \sum_{m=-\infty}^{\infty} \mathsf{h}_m e^{j2\pi m \frac{t}{T}}, \quad t \in [0, T]$$

with zero mean complex Gaussian coefficients h_m. We assume that only a finite number of coefficients h_m is random and all others vanish identically. This corresponds to a "band-limitation" of the fading process.[1] Thus, we obtain

$$\mathsf{h}(t) = \sum_{m=-M}^{M} \mathsf{h}_m e^{j2\pi m \frac{t}{T}}, \quad t \in [0, T] \, .$$

Here, we assume $\mathbf{h} \triangleq (\mathsf{h}_{-M} \ \cdots \ \mathsf{h}_M)^{\mathrm{T}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma})$ with a nonsingular covariance matrix $\mathbf{\Sigma} \in \mathbb{C}^{(2M+1)\times(2M+1)}$. We can rewrite $\mathbf{h} = \boldsymbol{Z}\mathbf{s}$ with $\mathbf{s} = (\mathsf{s}_{-M} \ \cdots \ \mathsf{s}_M)^{\mathrm{T}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{2M+1})$ and a nonsingular matrix $\boldsymbol{Z}$. Consequentially,

$$\mathsf{h}(t) = \boldsymbol{f}(t)\boldsymbol{Z}\mathbf{s}, \quad t \in [0, T] \tag{4.3}$$

where $\boldsymbol{f}(t) \triangleq (e^{-j2\pi M \frac{t}{T}} \ \cdots \ e^{j2\pi M \frac{t}{T}})$. Inserting (4.2) and (4.3) into (4.1), we thus obtain

$$\mathsf{y}(t) = \frac{\sqrt{\rho}}{\sqrt{T_S}} \boldsymbol{f}(t)\boldsymbol{Z}\mathbf{s} \sum_{i=1}^{N} \mathsf{x}_i \mathbb{1}_{[(i-1)T_S, iT_S)}(t) + \mathsf{w}(t), \quad t \in [0, T] \, .$$

4.2 Matched Filter Approach

Filtering the receive signal $\mathsf{y}(t)$ in (4.1) with $p^*(-t)$ and then sampling at symbol rate $1/T_S$ yields for each fading block the following discrete-time input-output relation:

$$\mathsf{y}_i = \int_{-\infty}^{\infty} \mathsf{y}(\tau) p^*(\tau - (i-1)T_S) \, \mathrm{d}\tau$$

[1] Note that the process cannot be strictly band-limited because it has only a finite time duration.

$$
\begin{aligned}
&= \frac{1}{\sqrt{T_S}} \int_{(i-1)T_S}^{iT_S} \mathsf{y}(\tau)\, \mathrm{d}\tau \\
&= \frac{\sqrt{\rho}}{T_S} \int_{(i-1)T_S}^{iT_S} \boldsymbol{f}(\tau)\boldsymbol{Z}\mathsf{s}\, \mathsf{x}_i + \mathsf{w}(\tau)\, \mathrm{d}\tau \\
&= \sqrt{\rho}\, \mathsf{x}_i \boldsymbol{r}_i \boldsymbol{Z}\mathsf{s} + \mathsf{w}_i
\end{aligned}
\tag{4.4}
$$

where

$$
\boldsymbol{r}_i \triangleq \left(\mathrm{sinc}\left(-\frac{M}{N}\right) e^{-j2\pi\frac{M(i-1)}{N}} \;\; \cdots \;\; \mathrm{sinc}\left(\frac{M}{N}\right) e^{j2\pi\frac{M(i-1)}{N}} \right) \in \mathbb{C}^{1\times(2M+1)}
$$

and

$$
\mathsf{w}_i = \frac{1}{\sqrt{T_S}} \int_{(i-1)T_S}^{iT_S} \mathsf{w}(\tau)\, \mathrm{d}\tau\, .
$$

The additive noise random variables $\{\mathsf{w}_i\}$ are i.i.d. zero-mean proper complex Gaussian. To keep the notation simple and without loss of generality, we assume that the input-output relation is normalized and that the $\{\mathsf{w}_i\}$ have unit variance. Then ρ in (4.4) can be thought of as the SNR. We can stack (4.4) for $i \in \{1, \ldots, N\}$ and obtain the vector input-output relation

$$
\mathbf{y} = \sqrt{\rho}\, \mathrm{diag}(\mathbf{x}) \boldsymbol{R}\boldsymbol{Z}\mathbf{s} + \mathbf{w} \tag{4.5}
$$

where $\mathbf{y} \triangleq (\mathsf{y}_1 \;\cdots\; \mathsf{y}_N)^{\mathrm{T}}$, $\mathbf{x} \triangleq (\mathsf{x}_1 \;\cdots\; \mathsf{x}_N)^{\mathrm{T}}$, $\boldsymbol{R} \triangleq (\boldsymbol{r}_1^{\mathrm{T}} \;\cdots\; \boldsymbol{r}_N^{\mathrm{T}})^{\mathrm{T}}$, and $\mathbf{w} \triangleq (\mathsf{w}_1 \;\cdots\; \mathsf{w}_N)^{\mathrm{T}}$.

Recall that the capacity of the discrete-time channel (4.5) is given by $C(\rho) = 1/N \sup I(\mathbf{x};\mathbf{y})$ where the supremum is over all probability measures on $\mathbf{x}$ that satisfy the average-power constraint $\mathbb{E}[\|\mathbf{x}\|^2] \leq N$. No closed-form expressions for $C(\rho)$ are known for the case $N > 1$. Recall that the number of degrees of freedom χ is defined as

$$
\chi = \lim_{\rho\to\infty} \frac{C(\rho)}{\log\rho}. \tag{4.6}
$$

It follows from [Liang and Veeravalli, 2004, Th. 1] that

$$
\chi = 1 - \frac{2M+1}{N} \tag{4.7}
$$

provided that $2M + 1 < N$. The intuition behind this result is as follows [Durisi and Bölcskei, 2011]: $2M + 1$ out of the N available symbols per block need to be sacrificed to learn the channels. This can be done for example by transmitting $2M + 1$ pilot symbols per block. The remaining $N - 2M - 1$ symbols can be used to communicate information. Hence, the number of degrees of freedom, which can be thought of as the number of "dimensions" per channel use available for communication is $(N - 2M - 1)/N = 1 - (2M + 1)/N$.

Note that (4.7) provides a lower bound on the number of degrees of freedom of the underlying continuous-time channel (4.1) because (4.7) is obtained i) by constraining the input signal to be of the form (4.2) and ii) by using a matched filter and sampling at the symbol rate at the receiver side. Both choices may be suboptimal.

4.3 The Oversampled Input-Output Relation

We show in this section that the matched filter approach reviewed in Section 4.2 is suboptimal. Specifically, we prove that by oversampling the filter output by a factor of two, $1 - 1/N$ degrees of freedom can be achieved.

Our oversampled, discrete-time, input-output relation is obtained as follows: The receive signal $\mathsf{y}(t)$ is filtered using a rectangular pulse whose width is half the symbol time (i.e., half the width of the transmit pulse $p(t)$). The resulting filtered output signal is then sampled at twice the symbol rate. Thus, we obtain

$$\mathsf{y}_i = \sqrt{\frac{2}{T_S}} \int_{(i-1)\frac{T_S}{2}}^{i\frac{T_S}{2}} \mathsf{y}(\tau)\,\mathrm{d}\tau$$

for $i \in \{1, \dots, 2N\}$. It is convenient to separately calculate y_i for even and odd i, hence, we define $\mathbf{y}^{(\mathrm{o})} = (\mathsf{y}_1\ \mathsf{y}_3\ \cdots\ \mathsf{y}_{2N-1})^{\mathrm{T}}$ and $\mathbf{y}^{(\mathrm{e})} = (\mathsf{y}_2\ \mathsf{y}_4\ \cdots\ \mathsf{y}_{2N})^{\mathrm{T}}$. For $\mathbf{y}^{(\mathrm{o})}$ we obtain

$$\begin{aligned} \mathsf{y}_i^{(\mathrm{o})} &= \sqrt{\frac{2}{T_S}} \int_{(i-1)T_S}^{(i-\frac{1}{2})T_S} \mathsf{y}(\tau)\,\mathrm{d}\tau \\ &= \frac{\sqrt{\rho}}{T_S} \int_{(i-1)T_S}^{(i-\frac{1}{2})T_S} \boldsymbol{f}(\tau)\boldsymbol{Z}\mathbf{s}\,\mathsf{x}_i + \mathsf{w}(\tau)\,\mathrm{d}\tau \\ &= \sqrt{\rho}\,\mathsf{x}_i \boldsymbol{r}_i^{(\mathrm{o})} \boldsymbol{Z}\mathbf{s} + \mathsf{w}_i^{(\mathrm{o})} \end{aligned} \tag{4.8}$$

where

$$\boldsymbol{r}_i^{(\mathrm{o})} \triangleq \left(\mathrm{sinc}\left(-\frac{M}{2N}\right) e^{-j\pi\frac{M}{N}(2i-\frac{3}{2})} \ \cdots\ \mathrm{sinc}\left(\frac{M}{2N}\right) e^{j\pi\frac{M}{N}(2i-\frac{3}{2})} \right) \in \mathbb{C}^{1\times(2M+1)} \tag{4.9}$$

and

$$\mathsf{w}_i^{(\mathrm{o})} = \sqrt{\frac{2}{T_S}} \int_{(i-1)T_S}^{(i-\frac{1}{2})T_S} \mathsf{w}(\tau)\,\mathrm{d}\tau\,.$$

Similarly, we obtain for $\mathbf{y}^{(\mathrm{e})}$

$$\begin{aligned} \mathsf{y}_i^{(\mathrm{e})} &= \sqrt{\frac{2}{T_S}} \int_{(i-\frac{1}{2})T_S}^{iT_S} \mathsf{y}(\tau)\,\mathrm{d}\tau \\ &= \sqrt{\rho}\,\mathsf{x}_i \boldsymbol{r}_i^{(\mathrm{e})} \boldsymbol{Z}\mathbf{s} + \mathsf{w}_i^{(\mathrm{e})} \end{aligned} \tag{4.10}$$

where

$$\boldsymbol{r}_i^{(\mathrm{e})} \triangleq \left(\mathrm{sinc}\left(-\frac{M}{2N}\right) e^{-j\pi\frac{M}{N}(2i-\frac{1}{2})} \ \cdots\ \mathrm{sinc}\left(\frac{M}{2N}\right) e^{j\pi\frac{M}{N}(2i-\frac{1}{2})} \right) \in \mathbb{C}^{1\times(2M+1)} \tag{4.11}$$

and

$$\mathsf{w}_i^{(\mathrm{e})} = \sqrt{\frac{2}{T_S}} \int_{(i-\frac{1}{2})T_S}^{iT_S} \mathsf{w}(\tau)\,\mathrm{d}\tau\,.$$

Combining (4.8) and (4.10), the vector input-output relation in the oversampled case is given by

$$\begin{pmatrix} \mathbf{y}^{(o)} \\ \mathbf{y}^{(e)} \end{pmatrix} = \sqrt{\rho} \begin{pmatrix} \operatorname{diag}(\mathbf{x})\boldsymbol{R}^{(o)} \\ \operatorname{diag}(\mathbf{x})\boldsymbol{R}^{(e)} \end{pmatrix} \boldsymbol{Z}\mathbf{s} + \begin{pmatrix} \mathbf{w}^{(o)} \\ \mathbf{w}^{(e)} \end{pmatrix} \tag{4.12}$$

where $\mathbf{x} \triangleq (\mathsf{x}_1 \ \cdots \ \mathsf{x}_N)^{\mathrm{T}}$, $\mathbf{w}^{(e)} \triangleq \left(\mathsf{w}_1^{(e)} \ \cdots \ \mathsf{w}_N^{(e)}\right)^{\mathrm{T}}$, $\mathbf{w}^{(o)} \triangleq \left(\mathsf{w}_1^{(o)} \ \cdots \ \mathsf{w}_N^{(o)}\right)^{\mathrm{T}}$, and

$$\boldsymbol{R}^{(e)} \triangleq \begin{pmatrix} \boldsymbol{r}_1^{(e)} \\ \vdots \\ \boldsymbol{r}_N^{(e)} \end{pmatrix}, \qquad \boldsymbol{R}^{(o)} \triangleq \begin{pmatrix} \boldsymbol{r}_1^{(o)} \\ \vdots \\ \boldsymbol{r}_N^{(o)} \end{pmatrix}.$$

Note that the input-output relations (4.8) and (4.10) are a special case of the channel model (2.1). As in (3.7) we again omit all rows and columns with zero entries and we see that the vector $\mathbf{x}^{\mathrm{T}}$ reduces to a single variable x_i for all receive symbols $\mathsf{y}_i^{(e)}$ and $\mathsf{y}_i^{(o)}$.

4.4 Degrees-of-Freedom Analysis

The capacity of the oversampled discrete-time channel (4.12) is given by

$$C(\rho) = \frac{1}{N} \sup I(\mathbf{x}; \mathbf{y}^{(e)}, \mathbf{y}^{(o)})$$

where the supremum is taken over all input distributions that satisfy the average power constraint $\mathbb{E}[\|\mathbf{x}\|^2] \leq N$. The number of degrees of freedom is defined as in (4.6). Our main result is given in the following theorem.

Theorem 4.1 The number of degrees of freedom of the channel (4.12) is lower-bounded as

$$\chi \geq 1 - \frac{1}{N}.$$

In order to apply Theorem 2.1, we have to consider the function ϕ: $\mathbb{C}^{2M+1+N} \to \mathbb{C}^{2N}$ which maps $\boldsymbol{x}$ and $\boldsymbol{s}$ onto the noiseless receive vector, i.e.,

$$\phi(\boldsymbol{s}, \boldsymbol{x}) = \begin{pmatrix} \operatorname{diag}(\boldsymbol{x})\boldsymbol{R}^{(o)} \\ \operatorname{diag}(\boldsymbol{x})\boldsymbol{R}^{(e)} \end{pmatrix} \boldsymbol{Z}\boldsymbol{s}.$$

More specifically, if we can show that the Jacobian determinant of ϕ has rank ℓ for some choice of $\boldsymbol{s}$ and $\boldsymbol{x}$, then the number of degrees of freedom is lower-bounded as

$$\chi \geq \frac{1}{N}(\ell - (2M+1)). \tag{4.13}$$

The Jacobian matrix $\boldsymbol{J}_\phi(\boldsymbol{s}, \boldsymbol{x})$ is given by

$$\boldsymbol{J}_\phi(\boldsymbol{s}, \boldsymbol{x}) = \begin{pmatrix} \operatorname{diag}(\boldsymbol{x})\boldsymbol{R}^{(o)}\boldsymbol{Z} & \operatorname{diag}(\boldsymbol{R}^{(o)}\boldsymbol{Z}\boldsymbol{s}) \\ \operatorname{diag}(\boldsymbol{x})\boldsymbol{R}^{(e)}\boldsymbol{Z} & \operatorname{diag}(\boldsymbol{R}^{(e)}\boldsymbol{Z}\boldsymbol{s}) \end{pmatrix} \in \mathbb{C}^{2N \times (2M+1+N)}.$$

We choose $\boldsymbol{x} = (1 \ \cdots \ 1)^{\mathrm{T}}$, i.e., $\mathrm{diag}(\boldsymbol{x}) = \mathbf{I}_N$. Furthermore, we choose $\boldsymbol{s} \neq \mathbf{0}$ such that $\boldsymbol{Z}\boldsymbol{s}$ is orthogonal to the first $2M$ rows of $\boldsymbol{R}^{(\mathrm{o})}$. This is possible because we assumed that $\boldsymbol{Z}$ is nonsingular and $\boldsymbol{s} \in \mathbb{C}^{2M+1}$. By the following result, $\boldsymbol{Z}\boldsymbol{s}$ is not orthogonal to all other rows of $\boldsymbol{R}^{(\mathrm{o})}$ and all rows of $\boldsymbol{R}^{(\mathrm{e})}$.

Lemma 4.2 The matrix $\left(\boldsymbol{R}^{(\mathrm{o})\mathrm{T}} \ \boldsymbol{R}^{(\mathrm{e})\mathrm{T}}\right)^{\mathrm{T}}$ is *full spark* [Alexeev et al., 2012, Def. 1], i.e., every set of $2M+1$ rows is linearly independent.

Proof. By (4.9), we can rewrite

$$\boldsymbol{r}_i^{(\mathrm{o})} = e^{-j\pi\frac{M}{N}\frac{1}{2}}\left(e^{-j\pi\frac{M}{N}(2i-2)} \ \cdots \ e^{j\pi\frac{M}{N}(2i-2)}\right)\begin{pmatrix}\mathrm{sinc}\left(-\frac{M}{2N}\right) & & \\ & \ddots & \\ & & \mathrm{sinc}\left(\frac{M}{2N}\right)\end{pmatrix}$$

for $i \in \{1,\ldots,N\}$. Similarly, by (4.11), we obtain

$$\boldsymbol{r}_i^{(\mathrm{e})} = e^{-j\pi\frac{M}{N}\frac{1}{2}}\left(e^{-j\pi\frac{M}{N}(2i-1)} \ \cdots \ e^{j\pi\frac{M}{N}(2i-1)}\right)\begin{pmatrix}\mathrm{sinc}\left(-\frac{M}{2N}\right) & & \\ & \ddots & \\ & & \mathrm{sinc}\left(\frac{M}{2N}\right)\end{pmatrix}$$

for $i \in \{1,\ldots,N\}$. Thus, stacking the rows $\boldsymbol{r}_i^{(\mathrm{e})}$ and $\boldsymbol{r}_i^{(\mathrm{o})}$ alternately, we obtain

$$\begin{pmatrix}\boldsymbol{r}_1^{(\mathrm{o})} \\ \boldsymbol{r}_1^{(\mathrm{e})} \\ \vdots \\ \boldsymbol{r}_N^{(\mathrm{o})} \\ \boldsymbol{r}_N^{(\mathrm{e})}\end{pmatrix} = e^{-j\pi\frac{M}{N}\frac{1}{2}}\,\boldsymbol{V}\begin{pmatrix}\mathrm{sinc}\left(-\frac{M}{2N}\right) & & \\ & \ddots & \\ & & \mathrm{sinc}\left(\frac{M}{2N}\right)\end{pmatrix} \tag{4.14}$$

where

$$\boldsymbol{V} \triangleq \begin{pmatrix} e^{-j\pi\frac{M}{N}0} & \cdots & e^{j\pi\frac{M}{N}0} \\ e^{-j\pi\frac{M}{N}1} & \cdots & e^{j\pi\frac{M}{N}1} \\ \vdots & & \vdots \\ e^{-j\pi\frac{M}{N}(2N-1)} & \cdots & e^{j\pi\frac{M}{N}(2N-1)}\end{pmatrix}.$$

The matrix $\boldsymbol{V}$ is full spark because it is a Vandermonde matrix with nonequal columns [Alexeev et al., 2012, Lem. 2]. Thus, also the matrix in (4.14) is full spark and interchanging the columns we finally obtain that $\left(\boldsymbol{R}^{(\mathrm{o})\mathrm{T}} \ \boldsymbol{R}^{(\mathrm{e})\mathrm{T}}\right)^{\mathrm{T}}$ is full spark. □

By Lemma 4.2, all elements in $\boldsymbol{R}^{(\mathrm{e})}\boldsymbol{Z}\boldsymbol{s}$ are nonzero (otherwise $\boldsymbol{Z}\boldsymbol{s}$ would be orthogonal to $2M+1$ linearly independent vectors and thus be zero). Hence, we can rewrite $\boldsymbol{J}_\phi(\boldsymbol{s},\boldsymbol{x})$ for our choice of $\boldsymbol{s}$ and $\boldsymbol{x}$ as

$$\boldsymbol{J}_\phi(\boldsymbol{s},\boldsymbol{x}) = \begin{pmatrix}[\boldsymbol{R}^{(\mathrm{o})}\boldsymbol{Z}]_{[1:2M]} & \mathbf{0}_{2M\times N} \\ [\boldsymbol{R}^{(\mathrm{o})}\boldsymbol{Z}]_{[2M+1:N]} & [\mathrm{diag}(\boldsymbol{R}^{(\mathrm{o})}\boldsymbol{Z}\boldsymbol{s})]_{[2M+1:N]} \\ \boldsymbol{R}^{(\mathrm{e})}\boldsymbol{Z} & \mathrm{diag}(\boldsymbol{R}^{(\mathrm{e})}\boldsymbol{Z}\boldsymbol{s})\end{pmatrix}$$

where $\mathrm{diag}(\boldsymbol{R}^{(\mathrm{e})}\boldsymbol{Z}\boldsymbol{s})$ is a nonsingular diagonal matrix. To show that $\boldsymbol{J}_\phi(\boldsymbol{s},\boldsymbol{x})$ has rank $2M+N$, we consider the square submatrix consisting of the rows in $\mathcal{I} \triangleq [1:2M] \cup [N+1:2N]$ and the columns in $\mathcal{D} \triangleq [1:2M+1] \cup [2M+3:2M+1+N]$. We obtain the submatrix

$$[\boldsymbol{J}_\phi(\boldsymbol{s},\boldsymbol{x})]_{\mathcal{I}}^{\mathcal{D}} = \begin{pmatrix} [\boldsymbol{R}^{(\mathrm{o})}\boldsymbol{Z}]_{[1:2M]} & \mathbf{0}_{2M\times(N-1)} \\ [\boldsymbol{R}^{(\mathrm{e})}\boldsymbol{Z}]_{\{1\}} & \mathbf{0}_{1\times(N-1)} \\ [\boldsymbol{R}^{(\mathrm{e})}\boldsymbol{Z}]_{[2:N]} & [\mathrm{diag}(\boldsymbol{R}^{(\mathrm{e})}\boldsymbol{Z}\boldsymbol{s})]_{[2:N]}^{[2:N]} \end{pmatrix} \tag{4.15}$$

The matrix $[\boldsymbol{J}_\phi(\boldsymbol{s},\boldsymbol{x})]_{\mathcal{I}}^{\mathcal{D}}$ in (4.15) is nonsingular if and only if the two matrices $\begin{pmatrix} [\boldsymbol{R}^{(\mathrm{o})}\boldsymbol{Z}]_{[1:2M]} \\ [\boldsymbol{R}^{(\mathrm{e})}\boldsymbol{Z}]_{\{1\}} \end{pmatrix}$ and $[\mathrm{diag}(\boldsymbol{R}^{(\mathrm{e})}\boldsymbol{Z}\boldsymbol{s})]_{[2:N]}^{[2:N]}$ are nonsingular. We already saw that $\mathrm{diag}(\boldsymbol{R}^{(\mathrm{e})}\boldsymbol{Z}\boldsymbol{s})$ is nonsingular and, by Lemma 4.2, $\begin{pmatrix} [\boldsymbol{R}^{(\mathrm{o})}\boldsymbol{Z}]_{[1:2M]} \\ [\boldsymbol{R}^{(\mathrm{e})}\boldsymbol{Z}]_{\{1\}} \end{pmatrix}$ is also nonsingular. Therefore, the rank of $\boldsymbol{J}_\phi(\boldsymbol{s},\boldsymbol{x})$ is at least $2M+N$. Inserting into (4.13) results in

$$\chi \geq \frac{1}{N}(N-1)$$

which concludes the proof.

Chapter 5

Conclusion of Part I

We characterized the number of degrees of freedom of block-fading channels in the noncoherent setting. In particular, we proved a lower bound on the number of degrees of freedom that conforms to an intuitive "dimension-counting" argument. Our channel model encompasses correlated block-fading MIMO channels as special cases, and it also allows the analysis of continuous-time channels.

We considered two channel models in more detail. The first was a generic block-fading MIMO channel model. Although this model seems to be just a minor variation of the classically used constant block-fading model, our result shows that the assumption of generic correlation may strongly affect the number of degrees of freedom. In fact, we showed that the (potentially small) perturbation in the channel model that results from making the coloring matrix $\boldsymbol{Z}$ generic may yield a significant increase in the number of degrees of freedom. This suggests once more (see also [Lapidoth and Moser, 2003, Durisi et al., 2012]) that care must be exercised in using this asymptotic quantity as a performance measure.

The highest gain in terms of the number of degrees of freedom is obtained for a sufficiently large number of receive antennas. In this case, the number of degrees of freedom is equal to T times the number of degrees of freedom in the SIMO case, as long as the number T of transmit antennas satisfies $T < N/Q$. This may be of interest for the uplink of massive-MIMO systems [Rusek et al., 2013].

It may appear questionable to assume that the coloring matrix $\boldsymbol{Z}$ is generic—an assumption that is needed for our result to hold. In particular, the special case where all matrices $\boldsymbol{Z}_{r,t}$ are *exactly* equal is nongeneric, and thus still an open problem. However, it should be noted that any nonzero perturbation of the model with exactly equal $\boldsymbol{Z}_{r,t}$—be it arbitrarily small—yields the generic model we considered. One may then argue that the assumption of exactly equal $\boldsymbol{Z}_{r,t}$ is an idealization that may be convenient in theoretical analyses but will not be satisfied in practical systems. An important conclusion to be drawn from our analysis is the fact that, as far as the number of degrees of freedom is concerned, the model with exactly equal $\boldsymbol{Z}_{r,t}$ is highly nonrobust, since arbitrarily small perturbations yield a potentially large change in the number of degrees of freedom.

An open problem is a characterization of the capacity of generic block-fading MIMO channels beyond the number of degrees of freedom. Such a characterization would help understand whether the sensitivity of the number of degrees of freedom discussed above is an indication of a similar sensitivity of the capacity that occurs already at moderate SNR, or merely an asymptotic peculiarity. Furthermore, it would be interesting to develop a capacity characterization that is nonasymptotic in the SNR for asymptotic block length. This might enable a capacity analysis of, e.g., stationary channel models.

The second model we considered was a continuous-time, time-selective, Rayleigh block-fading channel. We showed that in this scenario the number of degrees of freedom is lower-bounded by $1 - 1/N$. This number, which can be achieved by using a nonstandard receive filter and sampling the filter output signal at twice the symbol rate, is independent of the rank $2M + 1$ of the covariance matrix characterizing the temporal correlation of the fading inside each block. In contrast, the standard approach of matched filtering and sampling at the symbol rate leads to the looser lower bound $1 - (2M + 1)/N$.

Coarsely speaking, oversampling yields an increase of the dimension of the output space spanned by the receive samples. This increased dimension can be used to acquire knowledge about the fading channel at the receiver. Indeed, as demonstrated in Section 4.4, one pilot symbol per fading block is sufficient for the case of oversampling, whereas $2M + 1$ pilot symbols are required for the case of matched filtering and sampling at the symbol rate. This explains the degrees-of-freedom gain resulting from oversampling. The processing needed to acquire this additional channel knowledge is nonlinear, which is the reason why certain parts of the proof are somewhat technical.

A generalization of our results to more general (nonrectangular) pulse shapes and different fading statistics seems possible and constitutes an interesting line of future research. Furthermore, the combination of multiple antennas and oversampling could be analyzed using similar techniques.

Part II

Information Theory of Integer-Dimensional Singular Distributions

Chapter 6

Integer-Dimensional Entropy

Motivated by the fact that the number of degrees of freedom can be obtained via a heuristic "dimension-counting" analysis of the noiseless input-output relation, we want to investigate the information-theoretic properties of random variables whose distribution is similar to that of the noiseless receive vectors $\bar{\mathsf{y}}$ in Section 2.1. As mentioned in Section 2.2, in many interesting cases, $\bar{\mathsf{y}}$ is a singular random variable, i.e., neither discrete nor continuous. Up to now, even (differential) entropy—the most basic information-theoretic description of a random variable—has not been available for singular random variables.

In this chapter, we extend the concept of (differential) entropy to a broad class of singular random variables. More precisely, we consider integer-dimensional random variables and provide a definition of entropy that encompasses the classical entropy and differential entropy as special cases. Although the proofs are partly technical, the final results conform to intuition and have the potential to simplify information-theoretic analyses involving singular random variables. We show that the proposed entropy transforms in a natural manner under Lipschitz mappings, and we derive a transformation property that relates our entropy to differential entropy.

6.1 Previous Work and Motivation

We first recall the definitions of entropy for discrete random variables [Cover and Thomas, 2006, Ch. 2] and differential entropy for continuous random variables [Cover and Thomas, 2006, Ch. 8]. Let x be a discrete random variable with probability mass function $p_{\mathsf{x}}(\boldsymbol{x}_i) = \Pr\{\mathsf{x} = \boldsymbol{x}_i\}$, $i \in \mathcal{I}$, where $\mathcal{I}$ is the finite or countably infinite set specifying all possible realizations $\boldsymbol{x}_i$ of x. The entropy of x is

$$H(\mathsf{x}) \triangleq -\mathbb{E}_{\mathsf{x}}[\log p_{\mathsf{x}}(\mathsf{x})] = -\sum_{i\in\mathcal{I}} p_{\mathsf{x}}(\boldsymbol{x}_i) \log p_{\mathsf{x}}(\boldsymbol{x}_i)\,. \tag{6.1}$$

For a continuous random variable x on $\mathbb{R}^M$ with probability density function f_{x}, the differential entropy is

$$h(\mathsf{x}) \triangleq -\mathbb{E}_{\mathsf{x}}[\log f_{\mathsf{x}}(\mathsf{x})] = -\int_{\mathbb{R}^M} f_{\mathsf{x}}(\boldsymbol{x}) \log f_{\mathsf{x}}(\boldsymbol{x})\, \mathrm{d}\mathscr{L}^M(\boldsymbol{x})\,. \tag{6.2}$$

6.1.1 Rényi Entropy and ε Entropy

There exist two previously proposed generalizations of (differential) entropy to a larger set of probability distributions. The first generalization is based on quantizations of the random variable to ever finer cubes [Rényi, 1959]. More specifically, for a (possibly singular) random variable $\mathsf{x} \in \mathbb{R}^M$, the *Rényi information dimension* of x is

$$d(\mathsf{x}) \triangleq \lim_{n\to\infty} \frac{H\left(\frac{\lfloor n\mathsf{x}\rfloor}{n}\right)}{\log n} \tag{6.3}$$

and the *Rényi entropy of dimension* $d(\mathsf{x})$ of x is defined as

$$h^{\mathrm{R}}_{d(\mathsf{x})}(\mathsf{x}) \triangleq \lim_{n\to\infty} \left(H\left(\frac{\lfloor n\mathsf{x}\rfloor}{n}\right) - d(\mathsf{x})\log n \right) \tag{6.4}$$

provided the limits in (6.3) and (6.4) exist.

The definition of Rényi entropy corresponds to the following procedure:

1. Quantize the random variable x, by partitioning $\mathbb{R}^M$ into the cubes $\prod_{i=1}^{M}\left[\frac{k_i}{n}, \frac{k_i+1}{n}\right)$, where $\boldsymbol{k} \triangleq (k_1 \cdots k_M)^{\mathrm{T}} \in \mathbb{Z}^M$, i.e., consider the discrete random variable with probabilities $p_{\boldsymbol{k}} = \Pr\left\{\mathsf{x} \in \prod_{i=1}^{M}\left[\frac{k_i}{n}, \frac{k_i+1}{n}\right)\right\}$.

2. Calculate the entropy of the quantized random variable, i.e., calculate the negative expectation of the logarithm of the probability mass function $p_{\boldsymbol{k}}$.

3. Add the correction term $-d(\mathsf{x})\log n$ to account for the dimension of the random variable x.

4. Take the limit $n \to \infty$.

Although this approach seems reasonable, there are several issues. First, the definition of $h^{\mathrm{R}}_{d(\mathsf{x})}(\mathsf{x})$ seems to be difficult to handle analytically, and connections to major information-theoretic concepts such as mutual information are not available. Furthermore, the quantization used is just one of many possible—we might, e.g., also consider a shifted version of the set of cubes $\prod_{i=1}^{M}\left[\frac{k_i}{n}, \frac{k_i+1}{n}\right)$.

An approach that overcomes the latter issue is the concept of ε entropy [Kolmogorov, 1956, Posner and Rodemich, 1971]. The definition of ε entropy does not use a specific quantization but takes the infimum of the entropy over all possible (countable) quantizations under a constraint on the diameter of the quantization sets. This is motivated by data compression: the quantization should be such that an error of maximally ε is made (thus, the quantization sets have maximal diameter ε) and at the same time the minimal possible number of bits should be used to encode the data (thus, the entropy is minimized over all possible quantizations). More specifically, for a random variable $\mathsf{x} \in \mathbb{R}^M$, let $\mathfrak{P}_\varepsilon$ denote the set of all countable partitions of $\mathbb{R}^M$ into mutually disjoint, measurable sets of diameter at most ε. Furthermore, for a partition $\mathfrak{Q} = \{\mathcal{A}_i : i \in$

$\mathbb{N}\} \in \mathfrak{P}_\varepsilon$, the quantization $[\mathsf{x}]_{\mathfrak{Q}}$ is the discrete random variable defined by $p_i = \Pr\{[\mathsf{x}]_{\mathfrak{Q}} = i\} = \Pr\{\mathsf{x} \in \mathcal{A}_i\}$ for $i \in \mathbb{N}$. Then the ε *entropy of* x is defined as

$$H_\varepsilon(\mathsf{x}) \triangleq \inf_{\mathfrak{Q} \in \mathfrak{P}_\varepsilon} H([\mathsf{x}]_{\mathfrak{Q}}) \,. \tag{6.5}$$

Here, a problem is that $H_\varepsilon(\mathsf{x})$ is only defined for a fixed $\varepsilon > 0$ and the limit $\varepsilon \to 0$ converges to ∞ for nondiscrete distributions. However, as in the case of Rényi entropy, a correction term can be obtained using the following seemingly new definition of information dimension:

$$d^*(\mathsf{x}) \triangleq \lim_{\varepsilon \to 0} \frac{H_\varepsilon(\mathsf{x})}{\log \frac{1}{\varepsilon}} \,.$$

By [Kawabata and Dembo, 1994, Prop. 3.3], the definitions of information dimension using Rényi's approach and the ε entropy approach coincide, i.e., $d^*(\mathsf{x}) = d(\mathsf{x})$. This suggests the following new definition of a $d(\mathsf{x})$-dimensional entropy.

Definition 6.1 Let $\mathsf{x} \in \mathbb{R}^M$ be a random variable with existing information dimension $d(\mathsf{x})$. Then the *asymptotic* ε *entropy of dimension* $d(\mathsf{x})$ is defined as

$$h^*_{d(\mathsf{x})}(\mathsf{x}) \triangleq \lim_{\varepsilon \to 0} \big(H_\varepsilon(\mathsf{x}) + d(\mathsf{x}) \log \varepsilon \big) \,.$$

This definition corresponds to the following procedure:

1. Quantize the random variable x using an entropy-minimizing quantization[1] $\mathfrak{Q}$ given a diameter constraint ε, i.e., consider the discrete random variable $[\mathsf{x}]_{\mathfrak{Q}}$ with probabilities $p_i = \Pr\{[\mathsf{x}]_{\mathfrak{Q}} = i\} = \Pr\{\mathsf{x} \in \mathcal{A}_i\}$ for $\mathcal{A}_i \in \mathfrak{Q}$, where the diameter of each $\mathcal{A}_i$ is upper bounded by ε.

2. Calculate the entropy of the quantized random variable $[\mathsf{x}]_{\mathfrak{Q}}$, i.e., calculate the negative expectation of the logarithm of the probability mass function p_i.

3. Add the correction term $d(\mathsf{x}) \log \varepsilon$ to account for the dimension of the random variable x.

4. Take the limit $\varepsilon \to 0$.

Although this entropy is more general than Rényi entropy, the fundamental problems persist: we are still restricted to the choice of specific sets of small diameter (this is of course useful if we consider maximal distance as a measure of distortion but can yield unnecessarily many quantization points for areas of almost zero probability), and the definition still seems to be difficult to handle analytically and lacks connections to established information-theoretic quantities such as mutual information.

[1] We assume for simplicity that an entropy-minimizing quantization exists although in general the infimum in (6.5) may not be attained.

6.1.2 An Alternative Approach

Here, we propose a different approach, which is motivated by the definition of differential entropy. The basic idea is to perform the entropy calculation at the end. Assuming $\mathsf{x} \in \mathbb{R}^M$, this results in the following procedure:

1. For some $\boldsymbol{x} \in \mathbb{R}^M$, divide the probability $\Pr\{\mathsf{x} \in \mathcal{B}_\varepsilon(\boldsymbol{x})\}$ by the correction factor $\omega(d(\mathsf{x}))\, \varepsilon^{d(\mathsf{x})}$ (recall that $\omega(d(\mathsf{x}))$ is the volume of the $d(\mathsf{x})$-dimensional unit sphere).[2]

2. Take the limit $\varepsilon \to 0$.

3. Calculate the entropy as the negative expectation of the logarithm of the resulting density function.

More specifically, steps 1–2 yield the density function[3]

$$\theta_{\mathsf{x}}(\boldsymbol{x}) \triangleq \lim_{\varepsilon \to 0} \frac{\Pr\{\mathsf{x} \in \mathcal{B}_\varepsilon(\boldsymbol{x})\}}{\omega(d(\mathsf{x}))\, \varepsilon^{d(\mathsf{x})}} \tag{6.6}$$

and the entropy in step 3 is thus given by

$$\mathfrak{h}^{d(\mathsf{x})}(\mathsf{x}) \triangleq -\mathbb{E}_{\mathsf{x}}[\log \theta_{\mathsf{x}}(\mathsf{x})]\,. \tag{6.7}$$

We will show that this definition of entropy will lead to definitions of joint and conditional entropy, various useful relations, connections to mutual information, an asymptotic equipartition property, and bounds relevant to source coding. However, it does have one limitation: as pointed out in [Wu and Verdú, 2010, Sec. VII-A], the existence of the limit in (6.6) for almost every $\boldsymbol{x} \in \mathbb{R}^M$ is a much stronger assumption than the existence of the Rényi information dimension (6.3). Loosely speaking, the existence of the limit in (6.6) requires that the random variable x is almost everywhere $d(\mathsf{x})$-dimensional whereas the existence of the Rényi information dimension merely requires that the random variable is "on average" of dimension $d(\mathsf{x})$. By Preiss' Theorem [Preiss, 1987, Th. 5.6], convergence in (6.6) even implies that the probability measure induced by the random variable x is rectifiable (see Definition 6.4 below), which means that our definition does not apply to, e.g., self-similar fractal distributions. However, we are not aware of any application or calculation of Rényi entropy (or the asymptotic version of ε entropy) for fractal distributions, and it does not seem clear whether Rényi entropy is well defined in that case (although the information dimension (6.3) exists). An extension of our theory to mixtures of rectifiable measures of different dimensions may provide an interesting direction for future work.

Motivated by the entropy expression in (6.7), a formal definition of the entropy of an integer-dimensional random variable will be given in Section 6.3.1, based on the mathematical theory of rectifiable measures discussed next.

[2]The constant factor $\omega(d(\mathsf{x}))$ is used to obtain equality with differential entropy in the special case $d(\mathsf{x}) = M$. A different factor would result in an additive constant in the entropy definition.

[3]A mathematically rigorous definition will be provided in Section 6.2.2.

6.2 Rectifiable Random Variables

As mentioned in Section 6.1.2, the existence of a $d(\mathbf{x})$-dimensional density implies that the random variable $\mathbf{x}$ is rectifiable. In this section, we recall the definitions of rectifiable sets and measures and introduce rectifiable random variables as a straightforward extension. Furthermore, we present some basic properties that will be used in subsequent sections. For the convenience of readers who prefer to skip the mathematical details, we summarize the most important facts in Corollary 6.10.

6.2.1 Rectifiable Sets

Our basic geometric objects of interest are rectifiable sets [Federer, 1969, Sec. 3.2.14]. As the definition of rectifiable sets is not consistent in the literature, we provide the definition most convenient for our purpose. We recall that $\mathscr{H}^m$ denotes the m-dimensional Hausdorff measure.

Definition 6.2 ([Ambrosio et al., 2000, Def. 2.57]) For $m \in \mathbb{N}$, an $\mathscr{H}^m$-measurable set $\mathcal{E} \subseteq \mathbb{R}^M$ $(m \leq M)$ is called *m-rectifiable*[4] if there exist bounded sets $\mathcal{A}_k \subseteq \mathbb{R}^m$ and Lipschitz functions $f_k\colon \mathcal{A}_k \to \mathbb{R}^M$, both for[5] $k \in \mathbb{N}$, such that $\mathscr{H}^m\big(\mathcal{E} \setminus \bigcup_{k\in\mathbb{N}} f_k(\mathcal{A}_k)\big) = 0$. A set $\mathcal{E} \subseteq \mathbb{R}^M$ is called 0-rectifiable if it is finite or countably infinite.

Remark 6.1 Hereafter, we will often consider the setting of m-rectifiable sets in $\mathbb{R}^M$ and tacitly assume $m \in \mathbb{N}$ and $m \leq M$.

Rectifiable sets satisfy the following basic properties.

Lemma 6.3 Let $\mathcal{E}$ be an m-rectifiable subset of $\mathbb{R}^M$.

1. Any subset $\mathcal{D} \subseteq \mathcal{E}$ is also m-rectifiable.
2. The measure $\mathscr{H}^m|_{\mathcal{E}}$ is σ-finite.
3. Let $\phi\colon \mathbb{R}^M \to \mathbb{R}^N$ with $N \geq m$ be a Lipschitz function. If $\phi(\mathcal{E})$ is $\mathscr{H}^m$-measurable then it is m-rectifiable.
4. For $n > m$, we have $\mathscr{H}^n(\mathcal{E}) = 0$.
5. Let $\mathcal{E}_i$ for $i \in \mathbb{N}$ be m-rectifiable sets. Then $\bigcup_{i\in\mathbb{N}} \mathcal{E}_i$ is m-rectifiable.
6. $\mathbb{R}^m$ is m-rectifiable.

Proof. Properties 1–6 are well known; however, their proofs are not always provided in the literature. Therefore, for the reader's convenience, we provide proofs in Appendix B.1. □

Examples of rectifiable sets include affine subspaces, algebraic varieties, differentiable manifolds, and graphs of Lipschitz functions. As the countable union of rectifiable sets is again rectifiable, further examples are countable unions of any of the aforementioned sets.

[4]In [Ambrosio et al., 2000, Def. 2.57] these sets are called *countably $\mathscr{H}^m$-rectifiable*.

[5]Note that this definition also encompasses finite index sets $k \in \{1, \ldots, K\}$; it suffices to set $\mathcal{A}_k = \emptyset$ for $k > K$.

6.2.2 Rectifiable Measures

Loosely speaking, rectifiable measures are measures that are concentrated on a rectifiable set. The most convenient way to define "concentrated on" mathematically is in terms of absolute continuity with respect to a specific Hausdorff measure.

Definition 6.4 ([Ambrosio et al., 2000, Def. 2.59]) A Borel measure μ on $\mathbb{R}^M$ is called *m-rectifiable* if there exists an m-rectifiable set $\mathcal{E} \subseteq \mathbb{R}^M$ such that $\mu \ll \mathscr{H}^m|_{\mathcal{E}}$.

Remark 6.2 Let μ be an m-rectifiable measure, i.e., $\mu \ll \mathscr{H}^m|_{\mathcal{E}}$ for an m-rectifiable set $\mathcal{E} \subseteq \mathbb{R}^M$. By Property 2 in Lemma 6.3, $\mathscr{H}^m|_{\mathcal{E}}$ is σ-finite, and thus, by the Radon-Nikodym theorem [Ambrosio et al., 2000, Th. 1.28], $\mathrm{d}\mu = f \,\mathrm{d}\mathscr{H}^m|_{\mathcal{E}}$ with f being the Radon-Nikodym derivative $\frac{\mathrm{d}\mu}{\mathrm{d}\mathscr{H}^m|_{\mathcal{E}}}$.

To avoid the nuisance of taking care whether $\frac{\mathrm{d}\mu}{\mathrm{d}\mathscr{H}^m|_{\mathcal{E}}} = 0$, we construct a rectifiable set $\widetilde{\mathcal{E}}$ such that $\frac{\mathrm{d}\mu}{\mathrm{d}\mathscr{H}^m|_{\widetilde{\mathcal{E}}}} > 0$ almost everywhere.

Lemma 6.5 Let μ be an m-rectifiable measure on $\mathbb{R}^M$, i.e., $\mu \ll \mathscr{H}^m|_{\mathcal{E}}$ for an m-rectifiable set $\mathcal{E} \subseteq \mathbb{R}^M$. Then there exists an m-rectifiable set $\widetilde{\mathcal{E}} \subseteq \mathcal{E}$ such that

1. $\mu \ll \mathscr{H}^m|_{\widetilde{\mathcal{E}}}$,
2. $\frac{\mathrm{d}\mu}{\mathrm{d}\mathscr{H}^m|_{\widetilde{\mathcal{E}}}} = \frac{\mathrm{d}\mu}{\mathrm{d}\mathscr{H}^m|_{\mathcal{E}}}$ $\mathscr{H}^m|_{\widetilde{\mathcal{E}}}$-almost everywhere,
3. $\frac{\mathrm{d}\mu}{\mathrm{d}\mathscr{H}^m|_{\widetilde{\mathcal{E}}}} > 0$ $\mathscr{H}^m|_{\widetilde{\mathcal{E}}}$-almost everywhere.

Proof. See Appendix B.2. □

Definition 6.6 For an m-rectifiable measure μ, an m-rectifiable set $\mathcal{E}$ is called *a support of* μ if $\mu \ll \mathscr{H}^m|_{\mathcal{E}}$ and $\frac{\mathrm{d}\mu}{\mathrm{d}\mathscr{H}^m|_{\mathcal{E}}} > 0$ $\mathscr{H}^m|_{\mathcal{E}}$-almost everywhere.

Note that, by Lemma 6.5, a support exists for every m-rectifiable measure. However, a support is not unique and only defined up to a set of $\mathscr{H}^m$-measure zero.

For m-rectifiable measures, it is possible to interpret the Radon-Nikodym derivative as a measure of "local probability per area." This interpretation is based on the definition of the Hausdorff density.

Definition 6.7 ([Ambrosio et al., 2000, Def. 2.55]) Let μ be a Borel measure. The *m-dimensional Hausdorff density of μ* is defined as

$$\theta^m_\mu(\boldsymbol{x}) \triangleq \lim_{r \to 0} \frac{\mu(\mathcal{B}_r(\boldsymbol{x}))}{\omega(m) r^m}$$

provided the limit exists.

If μ is an m-rectifiable measure with $\mu \ll \mathscr{H}^m|_{\mathcal{E}}$, then the following result shows that θ^m_μ is indeed the Radon-Nikodym derivative $\frac{\mathrm{d}\mu}{\mathrm{d}\mathscr{H}^m|_{\mathcal{E}}}$ and vanishes for $\mathscr{H}^m$-almost all points not in $\mathcal{E}$.

Lemma 6.8 ([Ambrosio et al., 2000, Th. 2.83 and eq. (2.42)]) Let μ be an m-rectifiable measure, i.e., $\mu \ll \mathscr{H}^m|_\mathcal{E}$ for an m-rectifiable set $\mathcal{E}$. Then the m-dimensional Hausdorff density θ_μ^m exists and coincides with the Radon-Nikodym derivative $\frac{\mathrm{d}\mu}{\mathrm{d}\mathscr{H}^m|_\mathcal{E}}$ $\mathscr{H}^m|_\mathcal{E}$-almost everywhere. Furthermore, θ_μ^m is zero $\mathscr{H}^m$-almost everywhere on $\mathcal{E}^c$.

6.2.3 Rectifiable Random Variables

As we are only interested in probability measures and because information theory is often formulated for random variables, we define m-rectifiable random variables. In what follows, we consider a random variable $\mathsf{x}\colon (\Omega, \mathfrak{S}) \to (\mathbb{R}^M, \mathfrak{B}_M)$ on a probability space $(\Omega, \mathfrak{S}, \mu)$, i.e., Ω is a set, $\mathfrak{S}$ is a σ-algebra on Ω, and μ is a probability measure on $(\Omega, \mathfrak{S})$. The probability measure induced by the random variable x is denoted by $\mu\mathsf{x}^{-1}$. For $\mathcal{A} \in \mathfrak{B}_M$, $\mu\mathsf{x}^{-1}(\mathcal{A})$ equals the probability that $\mathsf{x} \in \mathcal{A}$, i.e.,

$$\mu\mathsf{x}^{-1}(\mathcal{A}) = \mu(\mathsf{x}^{-1}(\mathcal{A})) = \Pr\{\mathsf{x} \in \mathcal{A}\}\,. \tag{6.8}$$

Definition 6.9 A random variable $\mathsf{x}\colon (\Omega, \mathfrak{S}) \to (\mathbb{R}^M, \mathfrak{B}_M)$ on a probability space $(\Omega, \mathfrak{S}, \mu)$ is called *m-rectifiable* if the induced probability measure $\mu\mathsf{x}^{-1}$ on $\mathbb{R}^M$ is m-rectifiable, i.e., there exists an m-rectifiable set $\mathcal{E} \subseteq \mathbb{R}^M$ such that $\mu\mathsf{x}^{-1} \ll \mathscr{H}^m|_\mathcal{E}$. The m-dimensional Hausdorff density of an m-rectifiable random variable x is defined as

$$\theta_\mathsf{x}^m(\boldsymbol{x}) \triangleq \theta_{\mu\mathsf{x}^{-1}}^m(\boldsymbol{x}) = \lim_{r\to 0} \frac{\Pr\{\mathsf{x} \in \mathcal{B}_r(\boldsymbol{x})\}}{\omega(m) r^m}\,. \tag{6.9}$$

Furthermore, a support of the measure $\mu\mathsf{x}^{-1}$ is called a *support of* x, i.e., $\mathcal{E}$ is a support of x if $\mu\mathsf{x}^{-1} \ll \mathscr{H}^m|_\mathcal{E}$ and $\frac{\mathrm{d}\mu\mathsf{x}^{-1}}{\mathrm{d}\mathscr{H}^m|_\mathcal{E}}(\boldsymbol{x}) > 0$ $\mathscr{H}^m|_\mathcal{E}$-almost everywhere.

One can think of θ_x^m as an m-dimensional probability density function of the random variable x.

Based on the results of Section 6.2.2, we can find a characterization of m-rectifiable random variables that resembles well-known properties of continuous random variables. This characterization is stated in the next corollary. Note, however, that although everything seems to be similar to the continuous case, Hausdorff measures lack substantial properties of the Lebesgue measure, e.g., the product measure will not always be again a Hausdorff measure.

Corollary 6.10 Let x be an m-rectifiable random variable on $\mathbb{R}^M$, i.e., $\mu\mathsf{x}^{-1} \ll \mathscr{H}^m|_\mathcal{E}$ for an m-rectifiable set $\mathcal{E} \subseteq \mathbb{R}^M$. Then there exists the m-dimensional Hausdorff density θ_x^m and the following properties hold.

1. The m-dimensional Hausdorff density θ_x^m coincides with the Radon-Nikodym derivative $\frac{\mathrm{d}\mu\mathsf{x}^{-1}}{\mathrm{d}\mathscr{H}^m|_\mathcal{E}}$ $\mathscr{H}^m|_\mathcal{E}$-almost everywhere, i.e.,

$$\theta_\mathsf{x}^m(\boldsymbol{x}) = \frac{\mathrm{d}\mu\mathsf{x}^{-1}}{\mathrm{d}\mathscr{H}^m|_\mathcal{E}}(\boldsymbol{x}) \quad \mathscr{H}^m|_\mathcal{E}\text{-almost everywhere.} \tag{6.10}$$

2. The probability $\Pr\{\mathsf{x} \in \mathcal{A}\}$ for a measurable set $\mathcal{A} \subseteq \mathbb{R}^M$ can be calculated as the integral of θ^m_{x} over $\mathcal{A}$ with respect to the m-dimensional Hausdorff measure restricted to $\mathcal{E}$, i.e.,
$$\Pr\{\mathsf{x} \in \mathcal{A}\} = \mu\mathsf{x}^{-1}(\mathcal{A}) = \int_{\mathcal{A}} \theta^m_{\mathsf{x}}(\boldsymbol{x}) \, \mathrm{d}\mathscr{H}^m|_{\mathcal{E}}(\boldsymbol{x}) \,. \tag{6.11}$$
3. The expectation of a measurable function $f\colon \mathbb{R}^M \to \mathbb{R}$ with respect to the random variable x can be expressed as
$$\mathbb{E}_{\mathsf{x}}[f(\mathsf{x})] = \int_{\mathbb{R}^M} f(\boldsymbol{x}) \, \theta^m_{\mathsf{x}}(\boldsymbol{x}) \, \mathrm{d}\mathscr{H}^m|_{\mathcal{E}}(\boldsymbol{x}) \,. \tag{6.12}$$
4. The Hausdorff density θ^m_{x} is zero $\mathscr{H}^m$-almost everywhere on $\mathcal{E}^c$.
5. The random variable x is in $\mathcal{E}$ with probability one, i.e.,
$$\Pr\{\mathsf{x} \in \mathcal{E}\} = \mu\mathsf{x}^{-1}(\mathcal{E}) = \int_{\mathcal{E}} \theta^m_{\mathsf{x}}(\boldsymbol{x}) \, \mathrm{d}\mathscr{H}^m|_{\mathcal{E}}(\boldsymbol{x}) = 1 \,. \tag{6.13}$$
6. There exists a support $\widetilde{\mathcal{E}} \subseteq \mathcal{E}$ of x.
7. $\mathcal{E}$ is a support of x if and only if the Hausdorff density θ^m_{x} is positive $\mathscr{H}^m$-almost everywhere on $\mathcal{E}$.

Proof. By Definition 6.9, $\theta^m_{\mathsf{x}} = \theta^m_{\mu\mathsf{x}^{-1}}$ and, by Lemma 6.8, $\theta^m_{\mu\mathsf{x}^{-1}}$ exists and is equal to the Radon-Nikodym derivative $\frac{\mathrm{d}\mu\mathsf{x}^{-1}}{\mathrm{d}\mathscr{H}^m|_{\mathcal{E}}}$ $\mathscr{H}^m|_{\mathcal{E}}$-almost everywhere. Thus, we obtain (6.10).

Furthermore, we have for any measurable set $\mathcal{A} \subseteq \mathbb{R}^M$
$$\begin{aligned}
\Pr\{\mathsf{x} \in \mathcal{A}\} &\overset{(6.8)}{=} \mu\mathsf{x}^{-1}(\mathcal{A}) \\
&= \int_{\mathcal{A}} \mathrm{d}\mu\mathsf{x}^{-1}(\boldsymbol{x}) \\
&= \int_{\mathcal{A}} \frac{\mathrm{d}\mu\mathsf{x}^{-1}}{\mathrm{d}\mathscr{H}^m|_{\mathcal{E}}}(\boldsymbol{x}) \, \mathrm{d}\mathscr{H}^m|_{\mathcal{E}}(\boldsymbol{x}) \\
&\overset{(6.10)}{=} \int_{\mathcal{A}} \theta^m_{\mathsf{x}}(\boldsymbol{x}) \, \mathrm{d}\mathscr{H}^m|_{\mathcal{E}}(\boldsymbol{x})
\end{aligned}$$
i.e., (6.11) holds.

For a measurable function $f\colon \mathbb{R}^M \to \mathbb{R}$, we have
$$\begin{aligned}
\mathbb{E}_{\mathsf{x}}[f(\mathsf{x})] &= \int_{\mathbb{R}^M} f(\boldsymbol{x}) \, \mathrm{d}\mu\mathsf{x}^{-1}(\boldsymbol{x}) \\
&= \int_{\mathbb{R}^M} f(\boldsymbol{x}) \, \frac{\mathrm{d}\mu\mathsf{x}^{-1}}{\mathrm{d}\mathscr{H}^m|_{\mathcal{E}}}(\boldsymbol{x}) \, \mathrm{d}\mathscr{H}^m|_{\mathcal{E}}(\boldsymbol{x}) \\
&\overset{(6.10)}{=} \int_{\mathbb{R}^M} f(\boldsymbol{x}) \, \theta^m_{\mathsf{x}}(\boldsymbol{x}) \, \mathrm{d}\mathscr{H}^m|_{\mathcal{E}}(\boldsymbol{x})
\end{aligned} \tag{6.14}$$

i.e., (6.12) holds.

By Lemma 6.8, $\theta^m_{\mu\mathsf{x}^{-1}} = 0$ $\mathscr{H}^m$-almost everywhere on $\mathcal{E}^c$. Thus, since $\theta^m_{\mathsf{x}} = \theta^m_{\mu\mathsf{x}^{-1}}$, Property 4 holds.

We have

$$\Pr\{\mathsf{x} \in \mathcal{E}\} \overset{(6.8)}{=} \mu\mathsf{x}^{-1}(\mathcal{E}) \overset{(a)}{=} \mu\mathsf{x}^{-1}(\mathcal{E}) + \underbrace{\mu\mathsf{x}^{-1}(\mathcal{E}^c)}_{=0} \overset{(b)}{=} \mu\mathsf{x}^{-1}(\mathcal{E} \cup \mathcal{E}^c)$$
$$= \mu\mathsf{x}^{-1}(\mathbb{R}^M) = \mu(\mathsf{x}^{-1}(\mathbb{R}^M)) \overset{(c)}{=} \mu(\Omega) = 1$$

where (a) holds because $\mathscr{H}^m|_{\mathcal{E}}(\mathcal{E}^c) = 0$ and $\mu\mathsf{x}^{-1} \ll \mathscr{H}^m|_{\mathcal{E}}$, (b) holds because of the additivity of the measure $\mu\mathsf{x}^{-1}$, and (c) holds because $\mathsf{x}^{-1}(\mathbb{R}^M) = \Omega$. Thus, (6.13) holds.

Because $\mu\mathsf{x}^{-1} \ll \mathscr{H}^m|_{\mathcal{E}}$ and by Lemma 6.5, there exists a support $\widetilde{\mathcal{E}} \subseteq \mathcal{E}$. Thus, Property 6 holds.

By Definition 6.9, a set $\mathcal{E}$ satisfying $\mu\mathsf{x}^{-1} \ll \mathscr{H}^m|_{\mathcal{E}}$ is a support of x if and only if $\frac{\mathrm{d}\mu\mathsf{x}^{-1}}{\mathrm{d}\mathscr{H}^m|_{\mathcal{E}}} > 0$ $\mathscr{H}^m$-almost everywhere on $\mathcal{E}$. By (6.10), this is equivalent to $\theta^m_{\mathsf{x}} > 0$ $\mathscr{H}^m$-almost everywhere on $\mathcal{E}$. Thus, Property 7 holds. □

A trivial but noteworthy fact is that the special cases $m = 0$ and $m = M$ reduce to well-known concepts.

Theorem 6.11 Let x be a random variable on $\mathbb{R}^M$. Then:

1. x is 0-rectifiable if and only if it is a discrete random variable, i.e., there exists a probability mass function $p_{\mathsf{x}}(\boldsymbol{x}_i) = \Pr\{\mathsf{x} = \boldsymbol{x}_i\} > 0$, $i \in \mathcal{I}$, where $\mathcal{I}$ is a finite or countably infinite index set specifying all possible different realizations $\boldsymbol{x}_i$ of x. In this case, $\theta^0_{\mathsf{x}} = p_{\mathsf{x}}$ and $\mathcal{E} = \{\boldsymbol{x}_i : i \in \mathcal{I}\}$ is a support of x.

2. x is M-rectifiable if and only if it is a continuous random variable, i.e., there exists a probability density function f_{x} such that $\Pr\{\mathsf{x} \in \mathcal{A}\} = \int_{\mathcal{A}} f_{\mathsf{x}}(\boldsymbol{x})\, \mathrm{d}\mathscr{L}^M(\boldsymbol{x})$. In this case, $\theta^M_{\mathsf{x}} = f_{\mathsf{x}}$ $\mathscr{L}^M$-almost everywhere.

Proof. See Appendix B.3. □

The following theorem introduces a nontrivial class of m-rectifiable random variables.

Theorem 6.12 Let x be a continuous random variable on $\mathbb{R}^m$. Furthermore, let $\phi\colon \mathbb{R}^m \to \mathbb{R}^M$ with $M > m$ be a Lipschitz mapping whose m-dimensional Jacobian determinant[6] $\mathscr{J}_\phi(\boldsymbol{x})$ is nonzero $\mathscr{L}^m$-almost everywhere and assume that $\phi(\mathbb{R}^m)$ is $\mathscr{H}^m$-measurable. Then $\mathsf{y} \triangleq \phi(\mathsf{x})$ is an m-rectifiable random variable on $\mathbb{R}^M$.

Proof. According to Definition 6.9, we have to show that $\mu\mathsf{y}^{-1} \ll \mathscr{H}^m|_{\mathcal{E}}$ for an m-rectifiable set $\mathcal{E} \subseteq \mathbb{R}^M$. By Properties 3 and 6 in Lemma 6.3, the set $\mathcal{E} \triangleq \phi(\mathbb{R}^m)$ is m-rectifiable. Thus,

[6] The m-dimensional Jacobian determinant is defined as $\mathscr{J}_\phi(\boldsymbol{x}) = \sqrt{\det(D\phi^{\mathrm{T}}(\boldsymbol{x})\, D\phi(\boldsymbol{x}))}$, where $D\phi(\boldsymbol{x}) \in \mathbb{R}^{M \times m}$ denotes the (almost everywhere existing) Jacobian matrix of ϕ. Note in particular that $\mathscr{J}_\phi(\boldsymbol{x})$ is nonnegative.

it suffices to show that $\mu\mathsf{y}^{-1} \ll \mathscr{H}^m|_{\phi(\mathbb{R}^m)}$, i.e., that for any $\mathscr{H}^m$-measurable set $\mathcal{A} \subseteq \mathbb{R}^M$, $\mathscr{H}^m|_{\phi(\mathbb{R}^m)}(\mathcal{A}) = 0$ implies $\mu\mathsf{y}^{-1}(\mathcal{A}) = 0$. To this end, assume that $\mathscr{H}^m|_{\phi(\mathbb{R}^m)}(\mathcal{A}) = 0$ for an $\mathscr{H}^m$-measurable set $\mathcal{A} \subseteq \mathbb{R}^M$. Let f denote the probability density function of x. By the generalized change of variables formula [Ambrosio et al., 2000, eq. (2.47)], we have

$$\begin{aligned}\int_{\phi^{-1}(\mathcal{A})} f(\boldsymbol{x}) \mathscr{J}_\phi(\boldsymbol{x}) \,\mathrm{d}\mathscr{L}^m(\boldsymbol{x}) &= \int_{\phi(\phi^{-1}(\mathcal{A}))} \sum_{\boldsymbol{x}\in\phi^{-1}(\mathcal{A})\cap\phi^{-1}(\{\boldsymbol{y}\})} f(\boldsymbol{x}) \,\mathrm{d}\mathscr{H}^m(\boldsymbol{y}) \\ &= \int_{\mathcal{A}\cap\phi(\mathbb{R}^m)} \sum_{\boldsymbol{x}\in\phi^{-1}(\mathcal{A})\cap\phi^{-1}(\{\boldsymbol{y}\})} f(\boldsymbol{x}) \,\mathrm{d}\mathscr{H}^m(\boldsymbol{y}) \\ &\overset{(a)}{=} 0 \end{aligned} \tag{6.15}$$

where (a) holds because $\mathscr{H}^m(\mathcal{A} \cap \phi(\mathbb{R}^m)) = \mathscr{H}^m|_{\phi(\mathbb{R}^m)}(\mathcal{A}) = 0$. Because $\mathscr{J}_\phi(\boldsymbol{x}) > 0$ $\mathscr{L}^m$-almost everywhere, (6.15) implies $\int_{\phi^{-1}(\mathcal{A})} f(\boldsymbol{x}) \,\mathrm{d}\mathscr{L}^m(\boldsymbol{x}) = 0$. Thus, we have

$$\mu\mathsf{y}^{-1}(\mathcal{A}) = \mu\mathsf{x}^{-1}(\phi^{-1}(\mathcal{A})) = \int_{\phi^{-1}(\mathcal{A})} f(\boldsymbol{x}) \,\mathrm{d}\mathscr{L}^m(\boldsymbol{x}) = 0$$

□

6.3 Entropy of Rectifiable Random Variables

6.3.1 Definition

The m-rectifiable random variables introduced in Definition 6.9 will be the objects considered in our entropy definition. Due to the existence of the m-dimensional Hausdorff density θ^m_{x} for these random variables (see (6.9)), the heuristic approach described in Section 6.1.2 (see (6.6) and (6.7)) can be made rigorous.

Definition 6.13 Let x be an m-rectifiable random variable on $\mathbb{R}^M$. The *m-dimensional entropy of* x is defined as

$$\mathfrak{h}^m(\mathsf{x}) \triangleq -\mathbb{E}_{\mathsf{x}}\big[\log \theta^m_{\mathsf{x}}(\mathsf{x})\big] = -\int_{\mathbb{R}^M} \log \theta^m_{\mathsf{x}}(\boldsymbol{x}) \,\mathrm{d}\mu\mathsf{x}^{-1}(\boldsymbol{x})\,. \tag{6.16}$$

By (6.12), we obtain

$$\mathfrak{h}^m(\mathsf{x}) = -\int_{\mathbb{R}^M} \theta^m_{\mathsf{x}}(\boldsymbol{x}) \log \theta^m_{\mathsf{x}}(\boldsymbol{x}) \,\mathrm{d}\mathscr{H}^m|_{\mathcal{E}}(\boldsymbol{x}) = -\int_{\mathcal{E}} \theta^m_{\mathsf{x}}(\boldsymbol{x}) \log \theta^m_{\mathsf{x}}(\boldsymbol{x}) \,\mathrm{d}\mathscr{H}^m(\boldsymbol{x}) \tag{6.17}$$

where $\mathcal{E} \subseteq \mathbb{R}^M$ is an arbitrary m-rectifiable set satisfying $\mu\mathsf{x}^{-1} \ll \mathscr{H}^m|_{\mathcal{E}}$.

6.3.2 Relation to Entropy and Differential Entropy

In the special cases $m = 0$ and $m = M$, our entropy definition reduces to classical entropy (6.1) and differential entropy (6.2), respectively.

Theorem 6.14 Let x be a random variable on $\mathbb{R}^M$. If x is a 0-rectifiable (i.e., discrete) random variable, then the 0-dimensional entropy of x coincides with the classical entropy, i.e., $\mathfrak{h}^0(\mathsf{x}) = H(\mathsf{x})$. If x is an M-rectifiable (i.e., continuous) random variable, then the M-dimensional entropy of x coincides with the differential entropy, i.e., $\mathfrak{h}^M(\mathsf{x}) = h(\mathsf{x})$.

Proof. Let x be a 0-rectifiable random variable. By Theorem 6.11, x is a discrete random variable with possible realizations $\boldsymbol{x}_i$, $i \in \mathcal{I}$, the 0-dimensional Hausdorff density θ^0_{x} is the probability mass function of x, and a support is given by $\mathcal{E} = \{\boldsymbol{x}_i : i \in \mathcal{I}\}$. Thus, (6.17) yields

$$\mathfrak{h}^0(\mathsf{x}) = -\int_{\mathcal{E}} \theta^0_{\mathsf{x}}(\boldsymbol{x}) \log \theta^0_{\mathsf{x}}(\boldsymbol{x}) \, \mathrm{d}\mathscr{H}^0(\boldsymbol{x}) \stackrel{(a)}{=} -\sum_{i\in\mathcal{I}} \Pr\{\mathsf{x} = \boldsymbol{x}_i\} \log \Pr\{\mathsf{x} = \boldsymbol{x}_i\} = H(\mathsf{x})$$

where (a) holds because $\mathscr{H}^0$ is the counting measure.

Let x be an M-rectifiable random variable. By Theorem 6.11, x is a continuous random variable and the M-dimensional Hausdorff density θ^M_{x} is equal to the probability density function f_{x}. Thus, (6.16) yields

$$\mathfrak{h}^M(\mathsf{x}) = -\mathbb{E}_{\mathsf{x}}\big[\log \theta^M_{\mathsf{x}}(\mathsf{x})\big] = -\mathbb{E}_{\mathsf{x}}[\log f_{\mathsf{x}}(\mathsf{x})] \stackrel{(6.2)}{=} h(\mathsf{x}) \, .$$

□

To get an idea of the m-dimensional entropy of random variables in between the discrete and continuous cases, we can use Theorem 6.12 to construct m-rectifiable random variables. More specifically, we consider a continuous random variable x on $\mathbb{R}^m$ and a one-to-one Lipschitz mapping $\phi\colon \mathbb{R}^m \to \mathbb{R}^M$ $(M \geq m)$ whose Jacobian determinant $\mathscr{J}_\phi$ is nonzero $\mathscr{L}^m$-almost everywhere. Intuitively, we should see a connection between the differential entropy of x and the m-dimensional entropy of $\mathsf{y} \triangleq \phi(\mathsf{x})$. By Theorem 6.12, the random variable y is m-rectifiable and, because ϕ is one-to-one, we can indeed calculate the m-dimensional entropy.

Corollary 6.15 Let x be a continuous random variable on $\mathbb{R}^m$ with finite differential entropy $h(\mathsf{x})$ and probability density function f. Furthermore, let $\phi\colon \mathbb{R}^m \to \mathbb{R}^M$ $(M \geq m)$ be a one-to-one Lipschitz mapping with $\mathscr{J}_\phi$ nonzero $\mathscr{L}^m$-almost everywhere and such that $\mathbb{E}_{\mathsf{x}}[\log \mathscr{J}_\phi(\mathsf{x})]$ exists and is finite. Then the m-dimensional entropy of the m-rectifiable random variable $\mathsf{y} \triangleq \phi(\mathsf{x})$ is

$$\mathfrak{h}^m(\mathsf{y}) = h(\mathsf{x}) + \mathbb{E}_{\mathsf{x}}[\log \mathscr{J}_\phi(\mathsf{x})] \, .$$

For the special case of an embedding $\phi\colon \mathbb{R}^m \to \mathbb{R}^M$, $\phi(x_1, \dots, x_m) = (x_1 \cdots x_m \; 0 \cdots 0)^{\mathrm{T}}$, this results in

$$\mathfrak{h}^m(\mathsf{x}_1, \dots, \mathsf{x}_m, 0, \dots, 0) = h(\mathsf{x}) \, . \tag{6.18}$$

Proof. The first part is the special case $N = m$ and $\mathcal{E} = \mathbb{R}^m$ of the more general result in Theorem 6.16, which will be proved in Appendix B.4. The result (6.18) then follows from the fact that, for the considered embedding, $\mathscr{J}_\phi(\boldsymbol{x})$ is identically 1. □

6.3.3 Transformation Property

One important property of differential entropy is its invariance under unitary transformations. A similar result holds for m-dimensional entropy. We can even give a more general result for arbitrary one-to-one Lipschitz mappings.

Theorem 6.16 Let x be an m-rectifiable random variable on $\mathbb{R}^N$ with $1 \leq m \leq N$, finite m-dimensional entropy $\mathfrak{h}^m(\mathsf{x})$, support $\mathcal{E}$, and m-dimensional Hausdorff density θ^m_{x}. Furthermore, let $\phi\colon \mathbb{R}^N \to \mathbb{R}^M$ with $M \geq m$ be a Lipschitz mapping with[7] $\mathscr{J}^{\mathcal{E}}_{\phi} \neq 0$ $\mathscr{H}^m|_{\mathcal{E}}$-almost everywhere, $\phi(\mathcal{E})$ $\mathscr{H}^m$-measurable, and such that $\mathbb{E}_{\mathsf{x}}[\log \mathscr{J}^{\mathcal{E}}_{\phi}(\mathsf{x})]$ exists and is finite. If the restriction of ϕ to $\mathcal{E}$ is one-to-one, then $\mathsf{y} \triangleq \phi(\mathsf{x})$ is an m-rectifiable random variable and its m-dimensional entropy is

$$\mathfrak{h}^m(\mathsf{y}) = \mathfrak{h}^m(\mathsf{x}) + \mathbb{E}_{\mathsf{x}}[\log \mathscr{J}^{\mathcal{E}}_{\phi}(\mathsf{x})]\,.$$

Proof. See Appendix B.4. □

Remark 6.3 Theorem 6.16 shows that for the special case of a unitary transformation ϕ (e.g., a translation),

$$\mathfrak{h}^m(\phi(\mathsf{x})) = \mathfrak{h}^m(\mathsf{x})$$

because $\mathscr{J}^{\mathcal{E}}_{\phi}(\mathsf{x})$ is identically one in that case.

Remark 6.4 In general, no result resembling Theorem 6.16 holds for Lipschitz functions $\phi\colon \mathbb{R}^m \to \mathbb{R}^M$ $(M \geq m)$ that are not one-to-one on $\mathcal{E}$. We can argue as in the proof of Theorem 6.16 and obtain that $\mathsf{y} = \phi(\mathsf{x})$ is m-rectifiable and that the m-dimensional Hausdorff density is

$$\theta^m_{\mathsf{y}}(\boldsymbol{y}) = \sum_{\boldsymbol{x} \in \phi^{-1}(\{\boldsymbol{y}\})} \frac{\theta^m_{\mathsf{x}}(\boldsymbol{x})}{\mathscr{J}^{\mathcal{E}}_{\phi}(\boldsymbol{x})}$$

$\mathscr{H}^m|_{\phi(\mathcal{E})}$-almost everywhere. We then obtain for the m-dimensional entropy

$$\begin{aligned}\mathfrak{h}^m(\mathsf{y}) &= -\int_{\phi(\mathcal{E})} \left(\sum_{\boldsymbol{x} \in \phi^{-1}(\{\boldsymbol{y}\})} \frac{\theta^m_{\mathsf{x}}(\boldsymbol{x})}{\mathscr{J}^{\mathcal{E}}_{\phi}(\boldsymbol{x})} \right) \log \left(\sum_{\boldsymbol{x} \in \phi^{-1}(\{\boldsymbol{y}\})} \frac{\theta^m_{\mathsf{x}}(\boldsymbol{x})}{\mathscr{J}^{\mathcal{E}}_{\phi}(\boldsymbol{x})} \right) \mathrm{d}\mathscr{H}^m(\boldsymbol{y}) \\ &\overset{(a)}{=} -\int_{\mathcal{E}} \theta^m_{\mathsf{x}}(\boldsymbol{x}) \log \left(\sum_{\boldsymbol{x}' \in \phi^{-1}(\{\phi(\boldsymbol{x})\})} \frac{\theta^m_{\mathsf{x}}(\boldsymbol{x}')}{\mathscr{J}^{\mathcal{E}}_{\phi}(\boldsymbol{x}')} \right) \mathrm{d}\mathscr{H}^m(\boldsymbol{x})\end{aligned}$$

where (a) holds because of the generalized area formula [Ambrosio et al., 2000, Th. 2.91]. However, this cannot be easily expressed in terms of a differential entropy due to the sum in the logarithm.

[7] Here $\mathscr{J}^{\mathcal{E}}_{\phi}$ denotes the Jacobian determinant of the tangential differential of ϕ in $\mathcal{E}$. For details see [Ambrosio et al., 2000, Def. 2.89].

Chapter 7

Joint and Conditional Integer-Dimensional Entropy

An important concept in information theory is the amount of information that two random variables share. This concept is formalized by the mutual information, which constitutes the basis for many fundamental information-theoretic results. To find connections between the mutual information of integer-dimensional random variables and their entropy, we first define and study joint and conditional entropy for integer-dimensional singular random variables. We then derive expressions of the mutual information between integer-dimensional singular random variables in terms of joint and conditional entropy. We also extend several classical results for joint (differential) entropy, such as the chain rule and the asymptotic equipartition property. Finally, we show that the classical rule that conditioning does not increase entropy may be violated if the dimensions of the involved random variables do not match in a specific sense.

7.1 Joint Entropy

Joint entropy is a widely used concept although it can in fact be covered by the general concept of higher-dimensional entropy, because a pair of random variables (x, y) with $\mathsf{x} \in \mathbb{R}^{M_1}$ and $\mathsf{y} \in \mathbb{R}^{M_2}$ can also be interpreted as a single random variable on $\mathbb{R}^{M_1+M_2}$. Thus, our concept of entropy automatically generalizes to more than one random variable. Using this interpretation, we obtain from (6.16) for an m-rectifiable pair of random variables (x, y) (i.e., $\mu(\mathsf{x}, \mathsf{y})^{-1} \ll \mathscr{H}^m|_{\mathcal{E}}$ for an m-rectifiable set $\mathcal{E}$)

$$\mathfrak{h}^m(\mathsf{x}, \mathsf{y}) \triangleq -\mathbb{E}_{(\mathsf{x},\mathsf{y})}\big[\log \theta^m_{(\mathsf{x},\mathsf{y})}(\mathsf{x}, \mathsf{y})\big] \tag{7.1}$$

$$= -\int_{\mathbb{R}^M} \log \theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}\mu(\mathsf{x}, \mathsf{y})^{-1}(\boldsymbol{x}, \boldsymbol{y})$$

$$= -\int_{\mathbb{R}^M} \theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x}, \boldsymbol{y}) \log \theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}\mathscr{H}^m|_{\mathcal{E}}(\boldsymbol{x}, \boldsymbol{y}) \tag{7.2}$$

with $M = M_1 + M_2$. However, there are still some questions to answer:

- Suppose we have an m_1-rectifiable random variable $\mathbf{x}$ and an m_2-rectifiable random variable $\mathbf{y}$ on the same probability space. Can we conclude that $(\mathbf{x}, \mathbf{y})$ is (m_1+m_2)-rectifiable?
- Conversely, suppose we have an m-rectifiable random variable $(\mathbf{x}, \mathbf{y})$. Can we conclude that $\mathbf{x}$ and $\mathbf{y}$ are rectifiable?
- Assuming that $\mathbf{x}$, $\mathbf{y}$, and $(\mathbf{x}, \mathbf{y})$ are m_1-, m_2-, and m-rectifiable, respectively, is there a relationship between the quantities $\mathfrak{h}^{m_1}(\mathbf{x})$, $\mathfrak{h}^{m_2}(\mathbf{y})$, and $\mathfrak{h}^{m}(\mathbf{x}, \mathbf{y})$ provided they exist?

Whereas no answers to these questions exist in complete generality, we will provide answers under appropriate conditions on the involved random variables.

7.1.1 Product-Compatible Sets

One important shortcoming of Hausdorff measures (in contrast to, e.g., the Lebesgue measure) is that the product of two Hausdorff measures is in general not again a Hausdorff measure. However, restricting the Hausdorff measures to specific rectifiable sets will guarantee that the product is again a Hausdorff measure (see Lemma 7.2 below). As this property will be used throughout this section, we provide the following definition.

Definition 7.1 Let $\mathcal{E}_1 \subseteq \mathbb{R}^{M_1}$ be an m_1-rectifiable set and let $\mathcal{E}_2 \subseteq \mathbb{R}^{M_2}$ be an m_2-rectifiable set. We call $\mathcal{E}_1$ and $\mathcal{E}_2$ *product-compatible* if

$$\mathscr{H}^{m_1+m_2}|_{\mathcal{E}_1 \times \mathcal{E}_2} = \mathscr{H}^{m_1}|_{\mathcal{E}_1} \times \mathscr{H}^{m_2}|_{\mathcal{E}_2} . \tag{7.3}$$

The equality of measures in (7.3) is understood on the product σ-algebra of the σ-algebra of all $\mathscr{H}^{m_1}$-measurable sets in $\mathbb{R}^{M_1}$ and the σ-algebra of all $\mathscr{H}^{m_2}$-measurable sets in $\mathbb{R}^{M_2}$. Thus, (7.3) is equivalent to

$$\mathscr{H}^{m_1+m_2}(\mathcal{B}_1 \times \mathcal{B}_2) = \mathscr{H}^{m_1}(\mathcal{B}_1)\mathscr{H}^{m_2}(\mathcal{B}_2) \tag{7.4}$$

for all $\mathcal{B}_1 \subseteq \mathcal{E}_1$ $\mathscr{H}^{m_1}$-measurable and all $\mathcal{B}_2 \subseteq \mathcal{E}_2$ $\mathscr{H}^{m_2}$-measurable.

We can give some sufficient conditions for product compatibility. The result under Condition 1 is due to [Federer, 1969, Th. 3.2.23].

Lemma 7.2 Let $\mathcal{E}_1 \subseteq \mathbb{R}^{M_1}$ be an m_1-rectifiable Borel set, and let $\mathcal{E}_2 \subseteq \mathbb{R}^{M_2}$ be an m_2-rectifiable Borel set. Each of the following conditions implies that $\mathcal{E}_1$ and $\mathcal{E}_2$ are product-compatible.

1. The set $\mathcal{E}_1$ satisfies $\mathscr{H}^{m_1}(\mathcal{E}_1) < \infty$. Furthermore, $\mathcal{E}_2 = \phi(\mathcal{A})$ with some Lipschitz function $\phi\colon \mathbb{R}^{m_2} \to \mathbb{R}^{M_2}$ and some bounded set $\mathcal{A} \subseteq \mathbb{R}^{m_2}$.
2. The sets $\mathcal{E}_1$ and $\mathcal{E}_2$ can be decomposed as $\mathcal{E}_1 = \bigcup_{k\in\mathbb{N}} \mathcal{E}_1^{(k)}$ and $\mathcal{E}_2 = \bigcup_{\ell\in\mathbb{N}} \mathcal{E}_2^{(\ell)}$, where $\mathcal{E}_1^{(k)}$ and $\mathcal{E}_2^{(\ell)}$ are Borel sets such that $\mathscr{H}^{m_1}(\mathcal{E}_1^{(k)}) < \infty$ and $\mathcal{E}_2^{(\ell)} = \phi_\ell(\mathcal{A}_\ell)$ with Lipschitz functions $\phi_\ell\colon \mathbb{R}^{m_2} \to \mathbb{R}^{M_2}$ and bounded sets $\mathcal{A}_\ell \subseteq \mathbb{R}^{m_2}$.

3. We have $\mathcal{E}_1 = \bigcup_{k\in\mathbb{N}} \mathcal{E}_1^{(k)}$, where $\mathcal{E}_1^{(k)}$ are Borel sets such that $\mathscr{H}^{m_1}(\mathcal{E}_1^{(k)}) < \infty$. Furthermore, $m_2 = M_2$.

4. We have $m_1 = 0$, i.e., the set $\mathcal{E}_1$ is countable.

Proof. See Appendix B.5. □

One important property of product-compatible sets is that rectifiability of the sets implies that the product is again rectifiable.

Lemma 7.3 Let $\mathcal{E}_1 \subseteq \mathbb{R}^{M_1}$ be m_1-rectifiable, $\mathcal{E}_2 \subseteq \mathbb{R}^{M_2}$ be m_2-rectifiable, and $\mathcal{E}_1$ and $\mathcal{E}_2$ be product-compatible. Then $\mathcal{E}_1 \times \mathcal{E}_2$ is $(m_1 + m_2)$-rectifiable.

Proof. See Appendix B.6. □

7.1.2 Joint Entropy for Independent Random Variables

We start our investigation of joint entropy with independent random variables on product-compatible supports. In this case, it turns out that the m-dimensional entropy is additive.

Theorem 7.4 Let $\mathsf{x}\colon \Omega \to \mathbb{R}^{M_1}$ and $\mathsf{y}\colon \Omega \to \mathbb{R}^{M_2}$ be independent random variables on a probability space $(\Omega, \mathfrak{S}, \mu)$. Furthermore, let x be m_1-rectifiable with support $\mathcal{E}_1$ and let y be m_2-rectifiable with support $\mathcal{E}_2$, where $\mathcal{E}_1$ and $\mathcal{E}_2$ are product-compatible. Then the following properties hold:

1. The random variable $(\mathsf{x}, \mathsf{y})\colon \Omega \to \mathbb{R}^{M_1+M_2}$ is $(m_1 + m_2)$-rectifiable.

2. The $(m_1 + m_2)$-dimensional Hausdorff density of (x, y) satisfies

$$\theta^{m_1+m_2}_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x}, \boldsymbol{y}) = \theta^{m_1}_{\mathsf{x}}(\boldsymbol{x})\, \theta^{m_2}_{\mathsf{y}}(\boldsymbol{y}) \tag{7.5}$$

   $\mathscr{H}^{m_1+m_2}$-almost everywhere.

3. $\mathcal{E}_1 \times \mathcal{E}_2$ is a support of (x, y).

4. If $\mathfrak{h}^{m_1}(\mathsf{x})$ and $\mathfrak{h}^{m_2}(\mathsf{y})$ are finite, then the $(m_1 + m_2)$-dimensional entropy of the random variable (x, y) is given by

$$\mathfrak{h}^{m_1+m_2}(\mathsf{x}, \mathsf{y}) = \mathfrak{h}^{m_1}(\mathsf{x}) + \mathfrak{h}^{m_2}(\mathsf{y})\,.$$

Proof. See Appendix B.7. □

A corollary of Theorem 7.4 is a result for sequences of independent random variables. This setting will be important for our discussion of typical sets in Section 7.4.

Corollary 7.5 Let $\mathsf{x}_{1:n} \triangleq (\mathsf{x}_1, \ldots, \mathsf{x}_n)$ be a sequence of independent random variables, where $\mathsf{x}_i \in \mathbb{R}^{M_i}$, $i \in \{1, \ldots, n\}$ is m_i-rectifiable with support $\mathcal{E}_i$ and m_i-dimensional Hausdorff density $\theta^{m_i}_{\mathsf{x}_i}$. Furthermore, let $\mathcal{E}_1 \times \mathcal{E}_2 \times \cdots \times \mathcal{E}_{i-1}$ and $\mathcal{E}_i$ be product-compatible for all $i \in \{2, \ldots, n\}$. Then $\mathsf{x}_{1:n}$ is an m-rectifiable random variable on $\mathbb{R}^M$, where $m = \sum_{i=1}^n m_i$ and $M = \sum_{i=1}^n M_i$, and a support of $\mathsf{x}_{1:n}$ is given by $\mathcal{E} = \mathcal{E}_1 \times \mathcal{E}_2 \times \cdots \times \mathcal{E}_n$. Moreover, the m-dimensional Hausdorff density of $\mathsf{x}_{1:n}$ is given by

$$\theta^m_{\mathsf{x}_{1:n}}(\boldsymbol{x}_{1:n}) = \prod_{i=1}^n \theta^{m_i}_{\mathsf{x}_i}(\boldsymbol{x}_i) . \tag{7.6}$$

Finally, if $\mathfrak{h}^{m_i}(\mathsf{x}_i)$ is finite for $i \in \{1, \ldots, n\}$, then

$$\mathfrak{h}^m(\mathsf{x}_{1:n}) = \sum_{i=1}^n \mathfrak{h}^{m_i}(\mathsf{x}_i) . \tag{7.7}$$

Proof. The corollary follows by inductively applying Theorem 7.4 to the two random variables $(\mathsf{x}_1, \ldots, \mathsf{x}_{i-1})$ and x_i. □

7.1.3 Dependent Random Variables

The case of dependent random variables is more involved. The rectifiability of x and y does not necessarily imply the rectifiability of (x, y) (which is expected, since the marginal distributions carry only a small part of the information carried by the joint distribution). In general, even for continuous random variables x and y, we cannot calculate the joint differential entropy $h(\mathsf{x}, \mathsf{y})$ from the knowledge of the differential entropies $h(\mathsf{x})$ and $h(\mathsf{y})$. However, it is always possible to bound the differential entropy according to [Cover and Thomas, 2006, eq. (8.63)]

$$h(\mathsf{x}, \mathsf{y}) \leq h(\mathsf{x}) + h(\mathsf{y}) . \tag{7.8}$$

In general, no bound resembling (7.8) holds for our entropy definition. Indeed, the following simple setting provides a counterexample.

Example 7.1 Let $(\mathrm{x}, \mathrm{y}) \in \mathbb{R}^2$ be distributed according to a uniform distribution on the unit circle $\mathcal{S}_1$, i.e., for any Borel set $\mathcal{A} \subseteq \mathbb{R}^2$

$$\begin{aligned} \mu(\mathrm{x}, \mathrm{y})^{-1}(\mathcal{A}) = \Pr\{(\mathrm{x}, \mathrm{y}) \in \mathcal{A}\} &= \int_{\mathcal{A}} \frac{1}{\mathscr{H}^1(\mathcal{S}_1)} \, \mathrm{d}\mathscr{H}^1|_{\mathcal{S}_1}(x, y) \\ &= \int_{\mathcal{A}} \frac{1}{2\pi} \, \mathrm{d}\mathscr{H}^1|_{\mathcal{S}_1}(x, y) = \frac{1}{2\pi} \mathscr{H}^1|_{\mathcal{S}_1}(\mathcal{A}) . \end{aligned} \tag{7.9}$$

Choosing $f_1 \colon [-\pi, \pi] \to \mathbb{R}^2$, $f_1(t) = (\cos t \;\; \sin t)^{\mathrm{T}}$ results in $\mathcal{S}_1 \setminus f_1([-\pi, \pi]) = \emptyset$. Thus, according to Definition 6.2, the set $\mathcal{S}_1$ is 1-rectifiable. By (7.9), we have $\mu(\mathrm{x}, \mathrm{y})^{-1} \ll \mathscr{H}^1|_{\mathcal{S}_1}$, which implies that the random variable (x, y) is 1-rectifiable (see Definition 6.9). Again by (7.9), $\frac{\mathrm{d}\mu(\mathrm{x},\mathrm{y})^{-1}}{\mathrm{d}\mathscr{H}^1|_{\mathcal{S}_1}}(x, y) = 1/(2\pi)$. By (6.10), $\theta^1_{(\mathrm{x},\mathrm{y})}(x, y) = \frac{\mathrm{d}\mu(\mathrm{x},\mathrm{y})^{-1}}{\mathrm{d}\mathscr{H}^1|_{\mathcal{S}_1}}(x, y)$ and thus the 1-dimensional

Hausdorff density of (x,y) is given by

$$\theta^1_{(\mathsf{x},\mathsf{y})}(x,y) = \frac{1}{2\pi}$$

$\mathscr{H}^1$-almost everywhere on $\mathcal{S}_1$. Using (7.2), we obtain

$$\begin{aligned}\mathfrak{h}^1(\mathsf{x},\mathsf{y}) &= -\int_{\mathcal{S}_1} \theta^1_{(\mathsf{x},\mathsf{y})}(x,y) \log \theta^1_{(\mathsf{x},\mathsf{y})}(x,y)\, \mathrm{d}\mathscr{H}^1(x,y) \\ &= -\int_{\mathcal{S}_1} \frac{1}{2\pi} \log\left(\frac{1}{2\pi}\right) \mathrm{d}\mathscr{H}^1(x,y) \\ &= \frac{\log(2\pi)}{2\pi} \mathscr{H}^1(\mathcal{S}_1) \\ &= \log(2\pi)\,. \end{aligned} \tag{7.10}$$

One can easily see that x is a continuous random variable and its probability density function is given by $f_{\mathsf{x}}(x) = 1/(\pi\sqrt{1-x^2})$. By symmetry, the same holds for y, i.e., $f_{\mathsf{y}}(y) = 1/(\pi\sqrt{1-y^2})$. Basic calculus then yields for the differential entropy of x and y

$$h(\mathsf{x}) = h(\mathsf{y}) = \log\left(\frac{\pi}{2}\right). \tag{7.11}$$

Since x and y are continuous random variables, it follows from Theorem 6.14 that, $\mathfrak{h}^1(\mathsf{x}) = h(\mathsf{x})$ and $\mathfrak{h}^1(\mathsf{y}) = h(\mathsf{y})$. Thus,

$$\mathfrak{h}^1(\mathsf{x}) + \mathfrak{h}^1(\mathsf{y}) = 2\log\left(\frac{\pi}{2}\right) < \log(2\pi)\,.$$

Comparing with (7.10), we see that $\mathfrak{h}^1(\mathsf{x},\mathsf{y}) > \mathfrak{h}^1(\mathsf{x}) + \mathfrak{h}^1(\mathsf{y})$.

The reason for this seemingly unintuitive behavior of our entropy are the geometric properties of the projection $\mathfrak{p}_{\mathbf{y}} \colon \mathbb{R}^{M_1+M_2} \to \mathbb{R}^{M_2}$, $\mathfrak{p}_{\mathbf{y}}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{y}$, i.e., the projection of $\mathbb{R}^{M_1+M_2}$ to the last M_2 components. Although $\mathfrak{p}_{\mathbf{y}}$ is linear and has a classical Jacobian determinant $\mathscr{J}_{\mathfrak{p}_{\mathbf{y}}}$ of 1 everywhere on $\mathbb{R}^{M_1+M_2}$, things get more involved once we consider $\mathfrak{p}_{\mathbf{y}}$ as a mapping between rectifiable sets and want to calculate the Jacobian determinant $\mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_{\mathbf{y}}}$ of the tangential differential of $\mathfrak{p}_{\mathbf{y}}$ which maps an m-rectifiable set $\mathcal{E} \subseteq \mathbb{R}^{M_1+M_2}$ to an m_2-rectifiable set $\mathcal{E}_2 \subseteq \mathbb{R}^{M_2}$. In this setting, $\mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_{\mathbf{y}}}$ is not necessarily constant and may also become zero. Thus, the marginalization of an m-dimensional Hausdorff density is not as easy as the marginalization of a probability density function. The following theorem shows how to marginalize Hausdorff densities and describes the implications on m-dimensional entropy.

Theorem 7.6 Let $(\mathbf{x},\mathbf{y}) \in \mathbb{R}^{M_1+M_2}$ be an m-rectifiable random variable with m-dimensional Hausdorff density $\theta^m_{(\mathbf{x},\mathbf{y})}$ and support $\mathcal{E}$. Furthermore, let $\widetilde{\mathcal{E}}_2 \triangleq \mathfrak{p}_{\mathbf{y}}(\mathcal{E}) \subseteq \mathbb{R}^{M_2}$ be m_2-rectifiable ($m_2 \le m$), $\mathscr{H}^{m_2}(\widetilde{\mathcal{E}}_2) < \infty$, and $\mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_{\mathbf{y}}} \neq 0$ $\mathscr{H}^m|_{\mathcal{E}}$-almost everywhere. Then the following properties hold:

1. The random variable y is m_2-rectifiable.

2. There exists a support $\mathcal{E}_2 \subseteq \widetilde{\mathcal{E}}_2$ of y.

3. The m_2-dimensional Hausdorff density of y is given by

$$\theta_{\mathsf{y}}^{m_2}(\boldsymbol{y}) = \int_{\mathcal{E}^{(\boldsymbol{y})}} \frac{\theta_{(\mathsf{x},\mathsf{y})}^{m}(\boldsymbol{x},\boldsymbol{y})}{\mathscr{J}_{\mathfrak{p}_{\mathsf{y}}}^{\mathcal{E}}(\boldsymbol{x},\boldsymbol{y})} \,\mathrm{d}\mathscr{H}^{m-m_2}(\boldsymbol{x}) \tag{7.12}$$

   $\mathscr{H}^{m_2}$-almost everywhere, where $\mathcal{E}^{(\boldsymbol{y})} \triangleq \{\boldsymbol{x} \in \mathbb{R}^{M_1} : (\boldsymbol{x},\boldsymbol{y}) \in \mathcal{E}\}$.

4. An expression of the m_2-dimensional entropy of y is

$$\mathfrak{h}^{m_2}(\mathsf{y}) = -\int_{\mathcal{E}} \theta_{(\mathsf{x},\mathsf{y})}^{m}(\boldsymbol{x},\boldsymbol{y}) \log\left(\int_{\mathcal{E}^{(\boldsymbol{y})}} \frac{\theta_{(\mathsf{x},\mathsf{y})}^{m}(\tilde{\boldsymbol{x}},\boldsymbol{y})}{\mathscr{J}_{\mathfrak{p}_{\mathsf{y}}}^{\mathcal{E}}(\tilde{\boldsymbol{x}},\boldsymbol{y})} \,\mathrm{d}\mathscr{H}^{m-m_2}(\tilde{\boldsymbol{x}})\right) \mathrm{d}\mathscr{H}^{m}(\boldsymbol{x},\boldsymbol{y}) \tag{7.13}$$

   provided the integral on the right-hand side exists and is finite.

Under the assumptions that $\widetilde{\mathcal{E}}_1 \triangleq \mathfrak{p}_{\mathsf{x}}(\mathcal{E})$ is m_1-rectifiable ($m_1 \leq m$), $\mathscr{H}^{m_1}(\widetilde{\mathcal{E}}_1) < \infty$, and $\mathscr{J}_{\mathfrak{p}_{\mathsf{x}}}^{\mathcal{E}} \neq 0$ $\mathscr{H}^{m}|_{\mathcal{E}}$-almost everywhere, a symmetric result holds for x.

Proof. See Appendix B.8. □

Remark 7.1 Some assumptions we made in Theorem 7.6 might not be necessary. In particular, we need the assumption $\mathscr{H}^{m_2}(\mathcal{E}_2) < \infty$ only to be able to apply a modified version of the coarea formula (see Theorem B.1 in Appendix B.8).

We will illustrate the main findings of Theorem 7.6 in the setting of Example 7.1.

Example 7.2 As in Example 7.1, we consider $(\mathsf{x},\mathsf{y}) \in \mathbb{R}^2$ uniformly distributed on the unit circle $\mathcal{S}_1$, i.e., $\theta_{(\mathsf{x},\mathsf{y})}^{1}(x,y) = 1/(2\pi)$ $\mathscr{H}^1$-almost everywhere on $\mathcal{S}_1$. In Example 7.1, we already obtained $\mathfrak{h}^1(\mathsf{y}) = \log(\pi/2)$ (there, we used the fact that y is a continuous random variable and that, by Theorem 6.14, $\mathfrak{h}^1(\mathsf{y}) = h(\mathsf{y})$). Let us now calculate $\mathfrak{h}^1(\mathsf{y})$ using Theorem 7.6. Note first that $\mathfrak{p}_{\mathsf{y}}(\mathcal{S}_1) = [-1,1]$ which is 1-rectifiable and satisfies $\mathscr{H}^1([-1,1]) = 2 < \infty$. Next, we calculate the Jacobian determinant $\mathscr{J}_{\mathfrak{p}_{\mathsf{y}}}^{\mathcal{S}_1}(x,y)$. Consider a point on the unit circle, $\big(\pm\sqrt{1-y^2}, \pm y\big)$ with $y \in [0,1]$. At that point, the projection $\mathfrak{p}_{\mathsf{y}}$ restricted to the tangent space of $\mathcal{S}_1$ can be shown to amount to a multiplication by the factor $\sqrt{1-y^2}$. Thus, $\mathscr{J}_{\mathfrak{p}_{\mathsf{y}}}^{\mathcal{S}_1}\big(\pm\sqrt{1-y^2}, \pm y\big) = \sqrt{1-y^2}$. Hence, we obtain from (7.13)

$$\begin{aligned}
\mathfrak{h}^1(\mathsf{y}) &= -\int_{\mathcal{S}_1} \theta_{(\mathsf{x},\mathsf{y})}^{1}(x,y) \log\left(\int_{\mathcal{S}_1^{(y)}} \frac{\theta_{(\mathsf{x},\mathsf{y})}^{1}(\tilde{x},y)}{\mathscr{J}_{\mathfrak{p}_{\mathsf{y}}}^{\mathcal{S}_1}(\tilde{x},y)} \,\mathrm{d}\mathscr{H}^{1-1}(\tilde{x})\right) \mathrm{d}\mathscr{H}^1(x,y) \\
&= -\int_{\mathcal{S}_1} \frac{1}{2\pi} \log\left(\int_{\mathcal{S}_1^{(y)}} \frac{\frac{1}{2\pi}}{\sqrt{1-y^2}} \,\mathrm{d}\mathscr{H}^0(\tilde{x})\right) \mathrm{d}\mathscr{H}^1(x,y) \\
&\overset{(a)}{=} -\frac{1}{2\pi} \int_{\mathcal{S}_1} \log\left(\sum_{\tilde{x} \in \mathcal{S}_1^{(y)}} \frac{\frac{1}{2\pi}}{\sqrt{1-y^2}}\right) \mathrm{d}\mathscr{H}^1(x,y)
\end{aligned}$$

$$\stackrel{(b)}{=} -\frac{1}{2\pi}\int_{\mathcal{S}_1}\log\left(2\frac{\frac{1}{2\pi}}{\sqrt{1-y^2}}\right)\mathrm{d}\mathscr{H}^1(x,y)$$
$$= -\frac{1}{2\pi}\int_{\mathcal{S}_1}\log\left(\frac{1}{\pi\sqrt{1-y^2}}\right)\mathrm{d}\mathscr{H}^1(x,y)$$
$$= -\frac{1}{2\pi}\int_0^{2\pi}\log\left(\frac{1}{\pi|\cos(\phi)|}\right)\mathrm{d}\phi$$
$$= \log\left(\frac{\pi}{2}\right) \tag{7.14}$$

where (a) holds because $\mathscr{H}^0$ is the counting measure and (b) holds because $\mathcal{S}_1^{(y)} = \{x \in \mathbb{R} : (x,y) \in \mathcal{S}_1\} = \{\sqrt{1-y^2}, -\sqrt{1-y^2}\}$ contains two points for all $y \in (-1,1)$. Note that our above result for $\mathfrak{h}^1(\mathsf{y})$ coincides with the result previously obtained in Example 7.1.

7.1.4 Product-Compatible Random Variables

There are special settings in which m-dimensional entropy more closely matches the behavior we know from (differential) entropy. In these cases, the three random variables $\mathbf{x}$, $\mathbf{y}$, and $(\mathbf{x}, \mathbf{y})$ are rectifiable with "matching" dimensions, and we will see that, an inequality similar to (7.8) holds.

Definition 7.7 Let $\mathbf{x}$ be an m_1-rectifiable random variable on $\mathbb{R}^{M_1}$ with support $\mathcal{E}_1$, and let $\mathbf{y}$ be an m_2-rectifiable random variable on $\mathbb{R}^{M_2}$ with support $\mathcal{E}_2$. The random variables $\mathbf{x}$ and $\mathbf{y}$ are called *product-compatible* if $\mathcal{E}_1$ and $\mathcal{E}_2$ are product-compatible and $(\mathbf{x}, \mathbf{y})$ is an $(m_1 + m_2)$-rectifiable random variable on $\mathbb{R}^{M_1+M_2}$ with support $\mathcal{E} \subseteq \mathcal{E}_1 \times \mathcal{E}_2$.

The most important part of Definition 7.7 is that the dimensions of $\mathbf{x}$ and $\mathbf{y}$ add up to the joint dimension of $(\mathbf{x}, \mathbf{y})$. Note that this was not the case in Example 7.2, where x and y "shared" the dimension $m = 1$ of (x, y). A simple example of product-compatible random variables is the case of an m_1-rectifiable random variable $\mathbf{x}$ and an independent m_2-rectifiable random variable $\mathbf{y}$ with product-compatible supports $\mathcal{E}_1$ and $\mathcal{E}_2$. Indeed, by Theorem 7.4, $(\mathbf{x}, \mathbf{y})$ is $m_1 + m_2$-rectifiable with support $\mathcal{E}_1 \times \mathcal{E}_2$.

Another example of product-compatible random variables can be deduced from Theorem 7.6: Let $(\mathbf{x}, \mathbf{y})$ be $(m_1 + m_2)$-rectifiable. Assume that $\widetilde{\mathcal{E}}_2 \triangleq \mathfrak{p}_{\mathbf{y}}(\mathcal{E}) \subseteq \mathbb{R}^{M_2}$ is m_2-rectifiable, $\mathscr{H}^{m_2}(\widetilde{\mathcal{E}}_2) < \infty$, and $\mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_{\mathbf{y}}} \neq 0$ $\mathscr{H}^m|_{\mathcal{E}}$-almost everywhere. Furthermore, assume that $\widetilde{\mathcal{E}}_1 \triangleq \mathfrak{p}_{\mathbf{x}}(\mathcal{E})$ is m_1-rectifiable, $\mathscr{H}^{m_1}(\widetilde{\mathcal{E}}_1) < \infty$, and $\mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_{\mathbf{x}}} \neq 0$ $\mathscr{H}^m|_{\mathcal{E}}$-almost everywhere. By Theorem 7.6, $\mathbf{x}$ is m_1-rectifiable and $\mathbf{y}$ is m_2-rectifiable. Thus, if in addition $\widetilde{\mathcal{E}}_1$ and $\widetilde{\mathcal{E}}_2$ are product-compatible then $\mathbf{x}$ and $\mathbf{y}$ are product-compatible.

The setting of product-compatible random variables will be especially important for our discussion of mutual information in Section 7.3. However, already for joint entropy, we obtain some useful results.

Theorem 7.8 Let $\mathbf{x}$ be an m_1-rectifiable random variable on $\mathbb{R}^{M_1}$ with support $\mathcal{E}_1$, and let $\mathbf{y}$ be an m_2-rectifiable random variable on $\mathbb{R}^{M_2}$ with support $\mathcal{E}_2$. Furthermore, let $\mathbf{x}$ and $\mathbf{y}$ be product-compatible. Denote by $\theta^{m_1+m_2}_{(\mathsf{x},\mathsf{y})}$ the (m_1+m_2)-dimensional Hausdorff density of (x,y) and by $\mathcal{E} \subseteq \mathcal{E}_1 \times \mathcal{E}_2$ a support of (x,y). Then the following properties hold:

1. The m_2-dimensional Hausdorff density of y is given by

$$\theta^{m_2}_{\mathsf{y}}(\boldsymbol{y}) = \int_{\mathcal{E}_1} \theta^{m_1+m_2}_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y}) \,\mathrm{d}\mathscr{H}^{m_1}(\boldsymbol{x}) \tag{7.15}$$

   $\mathscr{H}^{m_2}$-almost everywhere.

2. An expression of the m_2-dimensional entropy of y is

$$\mathfrak{h}^{m_2}(\mathsf{y}) = -\int_{\mathcal{E}} \theta^{m_1+m_2}_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y}) \log\left(\int_{\mathcal{E}_1} \theta^{m_1+m_2}_{(\mathsf{x},\mathsf{y})}(\tilde{\boldsymbol{x}},\boldsymbol{y}) \,\mathrm{d}\mathscr{H}^{m_1}(\tilde{\boldsymbol{x}})\right) \mathrm{d}\mathscr{H}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y}) \tag{7.16}$$

   provided the integral on the right-hand side exists and is finite.

3. The inequality

$$\mathfrak{h}^{m_1+m_2}(\mathsf{x},\mathsf{y}) \leq \mathfrak{h}^{m_1}(\mathsf{x}) + \mathfrak{h}^{m_2}(\mathsf{y}) \tag{7.17}$$

   holds, provided $\mathfrak{h}^{m_1+m_2}(\mathsf{x},\mathsf{y})$ exists and is finite.

Due to symmetry, equivalent properties hold for $\theta^{m_1}_{\mathsf{x}}$ and $\mathfrak{h}^{m_1}(\mathsf{x})$.

Proof. We present a proof of Properties 1 and 2 in Appendix B.9. Although a direct proof of inequality (7.17) is possible, this task will be much easier once we considered the mutual information between rectifiable random variables. Thus, we postpone the proof of Property 3 to Corollary 7.19 in Section 7.3. □

7.2 Conditional Entropy

In contrast to joint entropy, conditional entropy is a nontrivial extension of entropy. We would like to define the entropy for a random variable x on $\mathbb{R}^{M_1}$ under the condition that a dependent variable y on $\mathbb{R}^{M_2}$ takes on a specific value $\boldsymbol{y}$. This "conditional" random variable is usually denoted as $(\mathsf{x} \,|\, \mathsf{y} = \boldsymbol{y})$. For discrete and—under appropriate assumptions—for continuous random variables, the random variable $(\mathsf{x} \,|\, \mathsf{y} = \boldsymbol{y})$ is well defined and so is the associated entropy $H(\mathsf{x} \,|\, \mathsf{y} = \boldsymbol{y})$ or differential entropy $h(\mathsf{x} \,|\, \mathsf{y} = \boldsymbol{y})$. Averaging over all $\boldsymbol{y}$ results in the well-known definitions of conditional entropy $H(\mathsf{x} \,|\, \mathsf{y})$, involving only the probability mass functions $p_{(\mathsf{x},\mathsf{y})}$ and p_{y}, or of conditional differential entropy $h(\mathsf{x} \,|\, \mathsf{y})$, involving only the probability density functions $f_{(\mathsf{x},\mathsf{y})}$ and f_{y}. Indeed, if x and y are discrete random variables, we have

$$\begin{aligned} H(\mathsf{x} \,|\, \mathsf{y}) &= \sum_{j\in\mathbb{N}} p_{\mathsf{y}}(\boldsymbol{y}_j)\, H(\mathsf{x} \,|\, \mathsf{y} = \boldsymbol{y}_j) \\ &= -\sum_{i,j\in\mathbb{N}} p_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x}_i,\boldsymbol{y}_j) \log\left(\frac{p_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x}_i,\boldsymbol{y}_j)}{p_{\mathsf{y}}(\boldsymbol{y}_j)}\right) \end{aligned} \tag{7.18}$$

$$= -\mathbb{E}_{(\mathsf{x},\mathsf{y})}\left[\log\left(\frac{p_{(\mathsf{x},\mathsf{y})}(\mathsf{x},\mathsf{y})}{p_{\mathsf{y}}(\mathsf{y})}\right)\right] \tag{7.19}$$

and, if x and y are continuous random variables, we have

$$h(\mathsf{x}\,|\,\mathsf{y}) = \int_{\mathbb{R}^{M_2}} f_{\mathsf{y}}(\boldsymbol{y})\, h(\mathsf{x}\,|\,\mathsf{y}=\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \tag{7.20}$$

$$= -\int_{\mathbb{R}^{M_1+M_2}} f_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y}) \log\left(\frac{f_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y})}{f_{\mathsf{y}}(\boldsymbol{y})}\right) \mathrm{d}(\boldsymbol{x},\boldsymbol{y})$$

$$= -\mathbb{E}_{(\mathsf{x},\mathsf{y})}\left[\log\left(\frac{f_{(\mathsf{x},\mathsf{y})}(\mathsf{x},\mathsf{y})}{f_{\mathsf{y}}(\mathsf{y})}\right)\right]. \tag{7.21}$$

A straightforward generalization to rectifiable measures would be to mimic the right-hand sides in (7.19) and (7.21) using Hausdorff densities. However, it will turn out that this naive approach is only partly correct: due to the geometric subtleties of the projection discussed in Section 7.1.3, we have to include a correction term that reflects the geometry of the conditioning process.

7.2.1 Conditional Probability

For general random variables x and y, a unique definition of a random variable $(\mathsf{x}\,|\,\mathsf{y}=\boldsymbol{y})$ is not possible and we have to resort to the concept of conditional probabilities, which can be summarized as follows (a detailed account can be found in [Gray, 2010, Ch. 5]): For a pair of random variables (x,y) on $\mathbb{R}^{M_1+M_2}$, there exists a regular conditional probability $\Pr\{\mathsf{x}\in\mathcal{A}\,|\,\mathsf{y}=\boldsymbol{y}\}$, i.e., for each measurable set $\mathcal{A}$, the function $\boldsymbol{y}\mapsto\Pr\{\mathsf{x}\in\mathcal{A}\,|\,\mathsf{y}=\boldsymbol{y}\}$ is measurable and $\Pr\{\mathsf{x}\in\mathcal{A}\,|\,\mathsf{y}=\boldsymbol{y}\}$ defines a probability measure for each $\boldsymbol{y}\in\mathbb{R}^{M_2}$. Furthermore, the regular conditional probability $\Pr\{\mathsf{x}\in\mathcal{A}\,|\,\mathsf{y}=\boldsymbol{y}\}$ satisfies

$$\Pr\{(\mathsf{x},\mathsf{y})\in\mathcal{A}_1\times\mathcal{A}_2\} = \int_{\mathcal{A}_2} \Pr\{\mathsf{x}\in\mathcal{A}_1\,|\,\mathsf{y}=\boldsymbol{y}\}\,\mathrm{d}\mu\mathsf{y}^{-1}(\boldsymbol{y}). \tag{7.22}$$

However, the regular conditional probability $\Pr\{\mathsf{x}\in\mathcal{A}\,|\,\mathsf{y}=\boldsymbol{y}\}$ involved in (7.22) is not unique. Nevertheless, we can still use (7.22) for a definition of conditional entropy because any version of the regular conditional probability satisfies (7.22). For the remainder of this section, we consider a fixed version of the regular conditional probability and denote a random variable distributed according to $\Pr\{\mathsf{x}\in\mathcal{A}\,|\,\mathsf{y}=\boldsymbol{y}\}$ as $(\mathsf{x}\,|\,\mathsf{y}=\boldsymbol{y})$.

7.2.2 Definition of Conditional Entropy

In order to be able to calculate the entropy of a random variable $(\mathsf{x}\,|\,\mathsf{y}=\boldsymbol{y})$, we first have to show that $(\mathsf{x}\,|\,\mathsf{y}=\boldsymbol{y})$ is rectifiable. The next theorem establishes sufficient conditions such that $(\mathsf{x}\,|\,\mathsf{y}=\boldsymbol{y})$ is rectifiable for almost every $\boldsymbol{y}$. As before, we denote by $\mathfrak{p}_{\mathsf{y}}\colon\mathbb{R}^{M_1+M_2}\to\mathbb{R}^{M_2}$ the projection of $\mathbb{R}^{M_1+M_2}$ to the last M_2 components, i.e., $\mathfrak{p}_{\mathsf{y}}(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{y}$

Theorem 7.9 Let (x,y) be an m-rectifiable random variable on $\mathbb{R}^{M_1+M_2}$ with m-dimensional Hausdorff density $\theta^m_{(\mathsf{x},\mathsf{y})}$ and support $\mathcal{E}$. Furthermore, let $\widetilde{\mathcal{E}}_2\triangleq\mathfrak{p}_{\mathsf{y}}(\mathcal{E})\subseteq\mathbb{R}^{M_2}$ be m_2-rectifiable

($m_2 \leq m$), $\mathscr{H}^{m_2}(\widetilde{\mathcal{E}}_2) < \infty$, and $\mathscr{J}^{\mathcal{E}}_{\mathsf{p_y}} \neq 0$ $\mathscr{H}^{m}|_{\mathcal{E}}$-almost everywhere. Then the following properties hold:

1. The random variable $(\mathsf{x} \,|\, \mathsf{y} = \boldsymbol{y})$ is $(m - m_2)$-rectifiable for $\mathscr{H}^{m_2}|_{\mathcal{E}_2}$-almost every $\boldsymbol{y} \in \mathbb{R}^{M_2}$, where $\mathcal{E}_2 \subseteq \widetilde{\mathcal{E}}_2$ is a support[1] of y.

2. The $(m - m_2)$-dimensional Hausdorff density of $(\mathsf{x} \,|\, \mathsf{y} = \boldsymbol{y})$ is given by

$$\theta^{m-m_2}_{(\mathsf{x}\,|\,\mathsf{y}=\boldsymbol{y})}(\boldsymbol{x}) = \frac{\theta^{m}_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x}, \boldsymbol{y})}{\mathscr{J}^{\mathcal{E}}_{\mathsf{p_y}}(\boldsymbol{x}, \boldsymbol{y})\, \theta^{m_2}_{\mathsf{y}}(\boldsymbol{y})} \tag{7.23}$$

   $\mathscr{H}^{m-m_2}|_{\mathcal{E}^{(\boldsymbol{y})}}$-almost everywhere, for $\mathscr{H}^{m_2}|_{\mathcal{E}_2}$-almost every $\boldsymbol{y} \in \mathbb{R}^{M_2}$. Here, as before, $\mathcal{E}^{(\boldsymbol{y})} \triangleq \{\boldsymbol{x} \in \mathbb{R}^{M_1} : (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{E}\}$.

3. For $\mathscr{H}^{m_2}|_{\mathcal{E}_2}$-almost every $\boldsymbol{y} \in \mathbb{R}^{M_2}$, the $(m - m_2)$-dimensional entropy of $(\mathsf{x} \,|\, \mathsf{y} = \boldsymbol{y})$ is given by

$$\mathfrak{h}^{m-m_2}(\mathsf{x} \,|\, \mathsf{y} = \boldsymbol{y}) = -\int_{\mathcal{E}^{(\boldsymbol{y})}} \frac{\theta^{m}_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x}, \boldsymbol{y})}{\mathscr{J}^{\mathcal{E}}_{\mathsf{p_y}}(\boldsymbol{x}, \boldsymbol{y})\, \theta^{m_2}_{\mathsf{y}}(\boldsymbol{y})} \log\left(\frac{\theta^{m}_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x}, \boldsymbol{y})}{\mathscr{J}^{\mathcal{E}}_{\mathsf{p_y}}(\boldsymbol{x}, \boldsymbol{y})\, \theta^{m_2}_{\mathsf{y}}(\boldsymbol{y})}\right) \mathrm{d}\mathscr{H}^{m-m_2}(\boldsymbol{x})\,. \tag{7.24}$$

Proof. See Appendix B.10. □

As for joint entropy, the case of product-compatible random variables (see Definition 7.7) is of special interest and results in a more intuitive characterization of the entropy of $(\mathsf{x} \,|\, \mathsf{y} = \boldsymbol{y})$.

Theorem 7.10 Let x be an m_1-rectifiable random variable on $\mathbb{R}^{M_1}$ with support $\mathcal{E}_1$, and let y be an m_2-rectifiable random variable on $\mathbb{R}^{M_2}$ with support $\mathcal{E}_2$. Furthermore, let x and y be product-compatible. Then the following properties hold:

1. The random variable $(\mathsf{x} \,|\, \mathsf{y} = \boldsymbol{y})$ is m_1-rectifiable for $\mathscr{H}^{m_2}|_{\mathcal{E}_2}$-almost every $\boldsymbol{y} \in \mathbb{R}^{M_2}$.

2. The m_1-dimensional Hausdorff density of $(\mathsf{x} \,|\, \mathsf{y} = \boldsymbol{y})$ is given by

$$\theta^{m_1}_{(\mathsf{x}\,|\,\mathsf{y}=\boldsymbol{y})}(\boldsymbol{x}) = \frac{\theta^{m_1+m_2}_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x}, \boldsymbol{y})}{\theta^{m_2}_{\mathsf{y}}(\boldsymbol{y})} \tag{7.25}$$

   $\mathscr{H}^{m_1}|_{\mathcal{E}_1}$-almost everywhere, for $\mathscr{H}^{m_2}|_{\mathcal{E}_2}$-almost every $\boldsymbol{y} \in \mathbb{R}^{M_2}$.

3. For $\mathscr{H}^{m_2}|_{\mathcal{E}_2}$-almost every $\boldsymbol{y} \in \mathbb{R}^{M_2}$, the m_1-dimensional entropy of $(\mathsf{x} \,|\, \mathsf{y} = \boldsymbol{y})$ is given by

$$\mathfrak{h}^{m_1}(\mathsf{x} \,|\, \mathsf{y} = \boldsymbol{y}) = -\int_{\mathcal{E}_1} \frac{\theta^{m_1+m_2}_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x}, \boldsymbol{y})}{\theta^{m_2}_{\mathsf{y}}(\boldsymbol{y})} \log\left(\frac{\theta^{m_1+m_2}_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x}, \boldsymbol{y})}{\theta^{m_2}_{\mathsf{y}}(\boldsymbol{y})}\right) \mathrm{d}\mathscr{H}^{m_1}(\boldsymbol{x})\,. \tag{7.26}$$

[1] By Theorem 7.6, the random variable y is m_2-rectifiable with Hausdorff density $\theta^{m_2}_{\mathsf{y}}$ (given by (7.12)) and some support $\mathcal{E}_2 \subseteq \widetilde{\mathcal{E}}_2$.

Proof. See Appendix B.11. □

Note that Theorems 7.9 and 7.10 hold for any version of the regular conditional probability $\Pr\{\mathsf{x} \in \mathcal{A} \,|\, \mathsf{y} = \boldsymbol{y}\}$. However, for different versions, the statement "for $\mathscr{H}^{m_2}|_{\mathcal{E}_2}$-almost every $\boldsymbol{y} \in \mathbb{R}^{M_2}$" may refer to different sets of $\mathscr{H}^{m_2}|_{\mathcal{E}_2}$-measure zero, e.g., (7.23) may hold for different $\boldsymbol{y} \in \mathbb{R}^{M_2}$. Thus, results that are independent of the version of the regular conditional probability can only be obtained if we can avoid these "almost everywhere"-statements. To this end, we will calculate the expectation of $\mathfrak{h}^{m-m_2}(\mathsf{x} \,|\, \mathsf{y} = \boldsymbol{y})$. The resulting expression will no longer depend on the specific version of the regular conditional probability. Anticipating this independence (cf. Theorem 7.12) and motivated by (7.18) and (7.20), we define conditional entropy for rectifiable random variables.

Definition 7.11 Let (x, y) be an m-rectifiable random variable on $\mathbb{R}^{M_1+M_2}$ and let y be m_2-rectifiable with m_2-dimensional Hausdorff density $\theta^{m_2}_{\mathsf{y}}$ and support $\mathcal{E}_2$. The *conditional entropy of* $(\mathsf{x} \,|\, \mathsf{y})$ is defined as

$$\mathfrak{h}^{m-m_2}(\mathsf{x} \,|\, \mathsf{y}) \triangleq \int_{\mathcal{E}_2} \theta^{m_2}_{\mathsf{y}}(\boldsymbol{y})\, \mathfrak{h}^{m-m_2}(\mathsf{x} \,|\, \mathsf{y} = \boldsymbol{y})\, \mathrm{d}\mathscr{H}^{m_2}(\boldsymbol{y}) \tag{7.27}$$

provided the right-hand side in (7.27) exists and coincides for all versions of the regular conditional probability $\Pr\{\mathsf{x} \in \mathcal{A} \,|\, \mathsf{y} = \boldsymbol{y}\}$.

The following theorem gives a characterization of conditional entropy and sufficient conditions for (7.27) to be well-defined in the sense that the right-hand side in (7.27) coincides for all versions of the regular conditional probability $\Pr\{\mathsf{x} \in \mathcal{A} \,|\, \mathsf{y} = \boldsymbol{y}\}$.

Theorem 7.12 Let (x, y) be an m-rectifiable random variable on $\mathbb{R}^{M_1+M_2}$ with m-dimensional Hausdorff density $\theta^{m}_{(\mathsf{x},\mathsf{y})}$ and support $\mathcal{E}$. Furthermore, let $\mathcal{E}_2 \triangleq \mathfrak{p}_{\mathsf{y}}(\mathcal{E})$ be m_2-rectifiable, $\mathscr{H}^{m_2}(\mathcal{E}_2) < \infty$, and $\mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_{\mathsf{y}}} \neq 0$ $\mathscr{H}^{m}|_{\mathcal{E}}$-almost everywhere. Then

$$\mathfrak{h}^{m-m_2}(\mathsf{x} \,|\, \mathsf{y}) = -\mathbb{E}_{(\mathsf{x},\mathsf{y})}\left[\log\left(\frac{\theta^{m}_{(\mathsf{x},\mathsf{y})}(\mathsf{x}, \mathsf{y})}{\theta^{m_2}_{\mathsf{y}}(\mathsf{y})}\right)\right] + \mathbb{E}_{(\mathsf{x},\mathsf{y})}\left[\log \mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_{\mathsf{y}}}(\mathsf{x}, \mathsf{y})\right] \tag{7.28}$$

provided the right-hand side in (7.28) exists and is finite.

Proof. See Appendix B.12. □

Note the difference between (7.28) and the expressions (7.19) and (7.21) of $H(\mathsf{x} \,|\, \mathsf{y})$ and $h(\mathsf{x} \,|\, \mathsf{y})$, respectively: in the case of rectifiable random variables, we have to take the geometric correction term $\mathbb{E}_{(\mathsf{x},\mathsf{y})}\left[\log \mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_{\mathsf{y}}}(\mathsf{x}, \mathsf{y})\right]$ into account. However, in the case of product-compatible rectifiable random variables, this correction term does not appear.

Theorem 7.13 Let the m_1-rectifiable random variable x on $\mathbb{R}^{M_1}$ and the m_2-rectifiable random variable y on $\mathbb{R}^{M_2}$ be product-compatible. Then

$$\mathfrak{h}^{m_1}(\mathsf{x} \,|\, \mathsf{y}) = -\mathbb{E}_{(\mathsf{x},\mathsf{y})}\left[\log\left(\frac{\theta^{m_1+m_2}_{(\mathsf{x},\mathsf{y})}(\mathsf{x}, \mathsf{y})}{\theta^{m_2}_{\mathsf{y}}(\mathsf{y})}\right)\right] \tag{7.29}$$

provided the right-hand side in (7.29) exists and is finite.

Proof. See Appendix B.13. □

7.2.3 Chain Rule for Rectifiable Random Variables

As in the case of entropy and differential entropy, we can give a chain rule for m-dimensional entropy.

Theorem 7.14 Let (x, y) be an m-rectifiable random variable on $\mathbb{R}^{M_1+M_2}$ with m-dimensional Hausdorff density $\theta^m_{(\mathsf{x},\mathsf{y})}$ and support $\mathcal{E}$. Furthermore, let $\mathcal{E}_2 \triangleq \mathfrak{p}_{\mathsf{y}}(\mathcal{E})$ be m_2-rectifiable, $\mathscr{H}^{m_2}(\mathcal{E}_2) < \infty$, and $\mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_{\mathsf{y}}} \neq 0$ $\mathscr{H}^m|_{\mathcal{E}}$-almost everywhere. Then

$$\mathfrak{h}^m(\mathsf{x}, \mathsf{y}) = \mathfrak{h}^{m_2}(\mathsf{y}) + \mathfrak{h}^{m-m_2}(\mathsf{x} \,|\, \mathsf{y}) - \mathbb{E}_{(\mathsf{x},\mathsf{y})}\big[\log \mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_{\mathsf{y}}}(\mathsf{x}, \mathsf{y})\big] \tag{7.30}$$

provided the corresponding integrals exist and are finite.

Proof. By the definition of $\mathfrak{h}^m(\mathsf{x}, \mathsf{y})$ in (7.1) and the definition of $\mathfrak{h}^{m_2}(\mathsf{y})$ in (6.16), we have

$$\begin{aligned} &\mathfrak{h}^m(\mathsf{x}, \mathsf{y}) - \mathfrak{h}^{m_2}(\mathsf{y}) + \mathbb{E}_{(\mathsf{x},\mathsf{y})}\big[\log \mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_{\mathsf{y}}}(\mathsf{x}, \mathsf{y})\big] \\ &\qquad = -\mathbb{E}_{(\mathsf{x},\mathsf{y})}\big[\log \theta^m_{(\mathsf{x},\mathsf{y})}(\mathsf{x}, \mathsf{y})\big] + \mathbb{E}_{\mathsf{y}}\big[\log \theta^{m_2}_{\mathsf{y}}(\mathsf{y})\big] + \mathbb{E}_{(\mathsf{x},\mathsf{y})}\big[\log \mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_{\mathsf{y}}}(\mathsf{x}, \mathsf{y})\big] \\ &\qquad = -\mathbb{E}_{(\mathsf{x},\mathsf{y})}\left[\log\left(\frac{\theta^m_{(\mathsf{x},\mathsf{y})}(\mathsf{x}, \mathsf{y})}{\theta^{m_2}_{\mathsf{y}}(\mathsf{y})}\right)\right] + \mathbb{E}_{(\mathsf{x},\mathsf{y})}\big[\log \mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_{\mathsf{y}}}(\mathsf{x}, \mathsf{y})\big]. \end{aligned} \tag{7.31}$$

Because we assumed in the theorem that the integrals corresponding to the terms on the left-hand side in (7.31) are finite, the right-hand side in (7.31) is also finite. By (7.28), the right-hand side in (7.31) equals $\mathfrak{h}^{m-m_2}(\mathsf{x} \,|\, \mathsf{y})$. Thus, (7.30) holds. □

Next, we continue Examples 7.1 and 7.2 from Section 7.1 where we will see that the geometric correction term in the chain rule, $\mathbb{E}_{(\mathsf{x},\mathsf{y})}\big[\log \mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_{\mathsf{y}}}(\mathsf{x}, \mathsf{y})\big]$, is indeed necessary.

Example 7.3 As in Examples 7.1 and 7.2, we consider $(\mathrm{x}, \mathrm{y}) \in \mathbb{R}^2$ uniformly distributed on the unit circle $\mathcal{S}_1$, i.e., $\theta^1_{(\mathrm{x},\mathrm{y})}(x, y) = 1/(2\pi)$ $\mathscr{H}^1$-almost everywhere on $\mathcal{S}_1$. According to (7.14),

$$\mathfrak{h}^1(\mathrm{y}) = \log\left(\frac{\pi}{2}\right) \tag{7.32}$$

and, according to (7.10),

$$\mathfrak{h}^1(\mathrm{x}, \mathrm{y}) = \log(2\pi). \tag{7.33}$$

To calculate the conditional entropy $\mathfrak{h}^0(\mathrm{x} \,|\, \mathrm{y})$ (note that $m - m_2 = 1 - 1 = 0$), we consider the regular conditional probability $\Pr\{\mathrm{x} \in \mathcal{A} \,|\, \mathrm{y} = y\}$. It is easy to see that one possible version of $\Pr\{\mathrm{x} \in \mathcal{A} \,|\, \mathrm{y} = y\}$ is the following: for $y \in (-1, 1)$, the random variable $(\mathrm{x} \,|\, \mathrm{y} = y)$ is a binary random variable taking on the values $\pm\sqrt{1 - y^2}$ with equal probability $1/2$. The random variables $(\mathsf{x}\,|\,\mathsf{y}=y)$ for $|y|\geq 1$ are irrelevant because $\Pr\{\mathsf{y}\notin(-1,1)\}=0$. Hence, the entropy $\mathfrak{h}^0(\mathsf{x}\,|\,\mathsf{y}=y)$ is the binary entropy function at $1/2$, i.e., $\log 2$, for all $y\in(-1,1)$. Thus, also the expectation with respect to y is the same, i.e.,

$$\mathfrak{h}^0(\mathsf{x}\,|\,\mathsf{y}) = \log 2\,. \tag{7.34}$$

This is different from $\mathfrak{h}^1(\mathsf{x},\mathsf{y}) - \mathfrak{h}^1(\mathsf{y}) = \log(2\pi) - \log(\pi/2)$, and therefore the conjecture that a chain rule holds without a correction term is wrong. To calculate the correction term, which is given by $\mathbb{E}_{(\mathsf{x},\mathsf{y})}\big[\log \mathscr{J}^{\mathcal{S}_1}_{\mathfrak{p}_\mathsf{y}}(\mathsf{x},\mathsf{y})\big]$ according to (7.30), we recall from Example 7.2 that $\mathscr{J}^{\mathcal{S}_1}_{\mathfrak{p}_\mathsf{y}}\big(\pm\sqrt{1-y^2},\pm y\big)=\sqrt{1-y^2}$ or, more conveniently, $\mathscr{J}^{\mathcal{S}_1}_{\mathfrak{p}_\mathsf{y}}(\cos\phi,\sin\phi)=|\cos\phi|$. Thus, we obtain

$$\begin{aligned}
\mathbb{E}_{(\mathsf{x},\mathsf{y})}\big[\log \mathscr{J}^{\mathcal{S}_1}_{\mathfrak{p}_\mathsf{y}}(\mathsf{x},\mathsf{y})\big] &= \int_{\mathcal{S}_1} \frac{1}{2\pi} \log \mathscr{J}^{\mathcal{S}_1}_{\mathfrak{p}_\mathsf{y}}(x,y)\,\mathrm{d}\mathscr{H}^1(x,y) \\
&= \int_0^{2\pi} \frac{1}{2\pi} \log|\cos\phi|\,\mathrm{d}\phi \\
&= -\log 2\,.
\end{aligned} \tag{7.35}$$

We finally verify that (7.35) is consistent with the chain rule (7.30). Starting from (7.33), we obtain

$$\begin{aligned}
\mathfrak{h}^1(\mathsf{x},\mathsf{y}) &= \log(2\pi) \\
&= \log\left(\frac{\pi}{2}\right) + \log 2 - (-\log 2) \\
&= \mathfrak{h}^1(\mathsf{y}) + \mathfrak{h}^0(\mathsf{x}\,|\,\mathsf{y}) - \mathbb{E}_{(\mathsf{x},\mathsf{y})}\big[\log \mathscr{J}^{\mathcal{S}_1}_{\mathfrak{p}_\mathsf{y}}(\mathsf{x},\mathsf{y})\big]
\end{aligned}$$

where the final expansion is obtained by using (7.32), (7.34), and (7.35).

Example 7.3 also provides a counterexample to the rule "conditioning does not increase entropy," which holds for the entropy of discrete random variables and the differential entropy of continuous random variables. Indeed, comparing (7.11) and (7.34), we see that for the components of a uniform distribution on the unit circle, we have $\mathfrak{h}^1(\mathsf{x}) < \mathfrak{h}^0(\mathsf{x}\,|\,\mathsf{y})$. However, as we will see in Corollary 7.19 below, this is only due to a "reduction of dimensions": if $\mathbf{x}$ and $\mathbf{y}$ are product-compatible, which implies that $\mathfrak{h}^{m_1}(\mathbf{x})$ and $\mathfrak{h}^{m-m_2}(\mathbf{x}\,|\,\mathbf{y})$ have the same dimension $m_1 = m - m_2$, conditioning will indeed reduce entropy, i.e., $\mathfrak{h}^{m_1}(\mathbf{x}\,|\,\mathbf{y}) \leq \mathfrak{h}^{m_1}(\mathbf{x})$ (see (7.48) below). Also the chain rule (7.30) reduces to its traditional form in the case of product-compatible random variables, as stated next.

Theorem 7.15 Let the m_1-rectifiable random variable $\mathbf{x}$ on $\mathbb{R}^{M_1}$ and the m_2-rectifiable random variable $\mathbf{y}$ on $\mathbb{R}^{M_2}$ be product-compatible. Then

$$\mathfrak{h}^{m_1+m_2}(\mathbf{x},\mathbf{y}) = \mathfrak{h}^{m_2}(\mathbf{y}) + \mathfrak{h}^{m_1}(\mathbf{x}\,|\,\mathbf{y}) \tag{7.36}$$

provided the entropies $\mathfrak{h}^{m_1+m_2}(\mathsf{x},\mathsf{y})$ and $\mathfrak{h}^{m_2}(\mathsf{y})$ exist and are finite.

Proof. By the definition of $\mathfrak{h}^{m_1+m_2}(\mathsf{x},\mathsf{y})$ in (7.1) and the definition of $\mathfrak{h}^{m_2}(\mathsf{y})$ in (6.16), we have

$$\begin{aligned}\mathfrak{h}^{m_1+m_2}(\mathsf{x},\mathsf{y}) - \mathfrak{h}^{m_2}(\mathsf{y}) &= -\mathbb{E}_{(\mathsf{x},\mathsf{y})}\big[\log\theta^{m_1+m_2}_{(\mathsf{x},\mathsf{y})}(\mathsf{x},\mathsf{y})\big] + \mathbb{E}_{\mathsf{y}}\big[\log\theta^{m_2}_{\mathsf{y}}(\mathsf{y})\big] \\ &= -\mathbb{E}_{(\mathsf{x},\mathsf{y})}\left[\log\left(\frac{\theta^{m_1+m_2}_{(\mathsf{x},\mathsf{y})}(\mathsf{x},\mathsf{y})}{\theta^{m_2}_{\mathsf{y}}(\mathsf{y})}\right)\right]. \end{aligned} \tag{7.37}$$

By (7.29), the right-hand side in (7.37) coincides with $\mathfrak{h}^{m_1}(\mathsf{x}\,|\,\mathsf{y})$. Thus, (7.36) holds. □

We can extend the chain rule (7.36) to a sequence of random variables.

Corollary 7.16 Let $\mathsf{x}_{1:n} \triangleq (\mathsf{x}_1,\dots,\mathsf{x}_n)$ be a sequence of random variables where each $\mathsf{x}_i \in \mathbb{R}^{M_i}$ is m_i-rectifiable. Assume that $\mathsf{x}_{1:i-1}$ and x_i are product-compatible for $i \in \{2,\dots,n\}$. Then

$$\mathfrak{h}^{m}(\mathsf{x}_{1:n}) = \mathfrak{h}^{m_1}(\mathsf{x}_1) + \sum_{i=2}^{n}\mathfrak{h}^{m_i}(\mathsf{x}_i\,|\,\mathsf{x}_{1:i-1}) \tag{7.38}$$

with $m \triangleq \sum_{i=1}^{n} m_i$, provided the corresponding integrals exist and are finite.

Proof. We prove (7.38) by induction. For $n = 2$, (7.38) reduces to (7.36). Thus, we only have to show the inductive step. Assume that (7.38) holds for $n-1$ random variables, i.e.,

$$\mathfrak{h}^{\widetilde{m}}(\tilde{\mathsf{x}}_{1:n-1}) = \mathfrak{h}^{\widetilde{m}_1}(\tilde{\mathsf{x}}_1) + \sum_{i=2}^{n-1}\mathfrak{h}^{\widetilde{m}_i}(\tilde{\mathsf{x}}_i\,|\,\tilde{\mathsf{x}}_{1:i-1}) \tag{7.39}$$

for $\tilde{\mathsf{x}}_{1:n-1}$ such that each $\tilde{\mathsf{x}}_i \in \mathbb{R}^{\widetilde{M}_i}$ is $\widetilde{m}_i$-rectifiable, and $\tilde{\mathsf{x}}_{1:i-1}$ and $\tilde{\mathsf{x}}_i$ are product-compatible for $i \in \{2,\dots,n\}$. Choosing $\tilde{\mathsf{x}}_1 \triangleq (\mathsf{x}_1,\mathsf{x}_2)$ and $\tilde{\mathsf{x}}_i \triangleq \mathsf{x}_{i+1}$ for $i \in \{2,\dots,n-1\}$, (7.39) implies

$$\mathfrak{h}^{m}(\mathsf{x}_{1:n}) = \mathfrak{h}^{m_1+m_2}(\mathsf{x}_1,\mathsf{x}_2) + \sum_{i=3}^{n}\mathfrak{h}^{m_i}(\mathsf{x}_i\,|\,\mathsf{x}_{1:i-1}). \tag{7.40}$$

By (7.36), we also have

$$\mathfrak{h}^{m_1+m_2}(\mathsf{x}_1,\mathsf{x}_2) = \mathfrak{h}^{m_1}(\mathsf{x}_1) + \mathfrak{h}^{m_2}(\mathsf{x}_2\,|\,\mathsf{x}_1). \tag{7.41}$$

Combining (7.40) and (7.41), we obtain (7.38). □

7.3 Mutual Information

The basic definition of mutual information is for discrete random variables x and y with probability mass functions $p_{\mathsf{x}}(\boldsymbol{x}_i)$ and $p_{\mathsf{y}}(\boldsymbol{y}_j)$, and joint probability mass function $p_{\mathsf{x},\mathsf{y}}(\boldsymbol{x}_i,\boldsymbol{y}_j)$. The mutual information between x and y is given by [Cover and Thomas, 2006, eq. (2.28)]

$$I(\mathsf{x};\mathsf{y}) \triangleq \sum_{i,j} p_{\mathsf{x},\mathsf{y}}(\boldsymbol{x}_i,\boldsymbol{y}_j)\log\left(\frac{p_{\mathsf{x},\mathsf{y}}(\boldsymbol{x}_i,\boldsymbol{y}_j)}{p_{\mathsf{x}}(\boldsymbol{x}_i)p_{\mathsf{y}}(\boldsymbol{y}_j)}\right). \tag{7.42}$$

However, mutual information is also defined between arbitrary random variables x and y on a common probability space. This definition is based on (7.42) and quantizations $[\mathsf{x}]_{\mathfrak{Q}}$ and $[\mathsf{y}]_{\mathfrak{R}}$ [Cover and Thomas, 2006, eq. (8.54)]. We recall from Section 6.1.1 that for a measurable, finite partition $\mathfrak{Q} = \{\mathcal{A}_1, \dots, \mathcal{A}_N\}$ of $\mathbb{R}^{M_1}$ (i.e., $\mathbb{R}^{M_1} = \bigcup_{i=1}^{N} \mathcal{A}_i$ with $\mathcal{A}_i \in \mathfrak{Q}$ mutually disjoint and measurable), the quantization $[\mathsf{x}]_{\mathfrak{Q}}$ is defined as the discrete random variable with probability distribution $\Pr\{[\mathsf{x}]_{\mathfrak{Q}} = i\} = \Pr\{\mathsf{x} \in \mathcal{A}_i\}$ for $i \in \{1, \dots, N\}$.

Definition 7.17 ([Cover and Thomas, 2006, eq. (8.54)]) Let $\mathsf{x}\colon \Omega \to \mathbb{R}^{M_1}$ and $\mathsf{y}\colon \Omega \to \mathbb{R}^{M_2}$ be random variables on a common probability space $(\Omega, \mathfrak{S}, \mu)$. The mutual information between x and y is defined as

$$I(\mathsf{x}; \mathsf{y}) \triangleq \sup_{\mathfrak{Q}, \mathfrak{R}} I([\mathsf{x}]_{\mathfrak{Q}}; [\mathsf{y}]_{\mathfrak{R}})$$

where the supremum is taken over all measurable, finite partitions $\mathfrak{Q}$ of $\mathbb{R}^{M_1}$ and $\mathfrak{R}$ of $\mathbb{R}^{M_2}$.

By the Gelfand-Yaglom-Perez theorem [Gray, 1990, Lem. 5.2.3], mutual information can also be expressed in terms of Radon-Nikodym derivatives: For random variables $\mathsf{x}\colon \Omega \to \mathbb{R}^{M_1}$ and $\mathsf{y}\colon \Omega \to \mathbb{R}^{M_2}$ on a common probability space $(\Omega, \mathfrak{S}, \mu)$,

$$I(\mathsf{x}; \mathsf{y}) = \begin{cases} \displaystyle\int_{\mathbb{R}^{M_1+M_2}} \log\left(\frac{\mathrm{d}\mu(\mathsf{x},\mathsf{y})^{-1}}{\mathrm{d}\big(\mu\mathsf{x}^{-1} \times \mu\mathsf{y}^{-1}\big)}(\boldsymbol{x},\boldsymbol{y})\right) \mathrm{d}\mu(\mathsf{x},\mathsf{y})^{-1}(\boldsymbol{x},\boldsymbol{y}) & \\ & \text{if } \mu(\mathsf{x},\mathsf{y})^{-1} \ll \mu\mathsf{x}^{-1} \times \mu\mathsf{y}^{-1} \\ \infty & \text{else.} \end{cases} \tag{7.43}$$

For the special cases of discrete and continuous random variables, there exists an expression of mutual information in terms of entropy and differential entropy, respectively. We will extend this expression to the case of rectifiable random variables. The resulting generalization will involve the entropies $\mathfrak{h}^{m_1}(\mathsf{x})$, $\mathfrak{h}^{m_2}(\mathsf{y})$, and $\mathfrak{h}^{m}(\mathsf{x},\mathsf{y})$.

Theorem 7.18 Let x be an m_1-rectifiable random variable with support $\mathcal{E}_1 \subseteq \mathbb{R}^{M_1}$, let y be an m_2-rectifiable random variable with support $\mathcal{E}_2 \subseteq \mathbb{R}^{M_2}$, let (x,y) be m-rectifiable with support $\mathcal{E} \subseteq \mathcal{E}_1 \times \mathcal{E}_2$, and assume that $\mathcal{E}_1$ and $\mathcal{E}_2$ are product-compatible. The mutual information satisfies:

1. If x and y are product-compatible (i.e., $m = m_1 + m_2$),

$$I(\mathsf{x}; \mathsf{y}) = \int_{\mathcal{E}} \theta^{m}_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y}) \log\left(\frac{\theta^{m}_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y})}{\theta^{m_1}_{\mathsf{x}}(\boldsymbol{x})\theta^{m_2}_{\mathsf{y}}(\boldsymbol{y})}\right) \mathrm{d}\mathscr{H}^{m}(\boldsymbol{x},\boldsymbol{y})\,. \tag{7.44}$$

Furthermore,

$$I(\mathsf{x}; \mathsf{y}) = \mathfrak{h}^{m_1}(\mathsf{x}) + \mathfrak{h}^{m_2}(\mathsf{y}) - \mathfrak{h}^{m}(\mathsf{x},\mathsf{y}) \tag{7.45}$$

and

$$I(\mathbf{x};\mathbf{y}) = \mathfrak{h}^{m_1}(\mathbf{x}) - \mathfrak{h}^{m_1}(\mathbf{x}\,|\,\mathbf{y}) = \mathfrak{h}^{m_2}(\mathbf{y}) - \mathfrak{h}^{m_2}(\mathbf{y}\,|\,\mathbf{x}) \tag{7.46}$$

provided the entropies $\mathfrak{h}^{m_1}(\mathbf{x})$, $\mathfrak{h}^{m_2}(\mathbf{y})$, and $\mathfrak{h}^{m}(\mathbf{x},\mathbf{y})$ exist and are finite.

2. If $m < m_1 + m_2$ then $I(\mathbf{x};\mathbf{y}) = \infty$.

Proof. See Appendix B.14. □

In Theorem 7.18, the case $m < m_1 + m_2$ can be interpreted as $\mathbf{x}$ and $\mathbf{y}$ "sharing" at least one dimension. In a communication scenario, this would imply that it is possible to reconstruct an at least one-dimensional component of $\mathbf{x}$ from $\mathbf{y}$ (and, also, to reconstruct an at least one-dimensional component of $\mathbf{y}$ from $\mathbf{x}$). Thus, an infinite amount of information could be transmitted over a channel $\mathbf{x} \longrightarrow \mathbf{y}$ (or $\mathbf{y} \longrightarrow \mathbf{x}$). This fits our result that $I(\mathbf{x};\mathbf{y}) = \infty$ in the case $m < m_1 + m_2$.

A corollary of Theorem 7.18 is that for product-compatible random variables, we can upper-bound the joint entropy by the sum of the single entropies and prove that conditioning does not increase entropy.

Corollary 7.19 Let the m_1-rectifiable random variable $\mathbf{x}$ on $\mathbb{R}^{M_1}$ and the m_2-rectifiable random variable $\mathbf{y}$ on $\mathbb{R}^{M_2}$ be product-compatible. Then

$$\mathfrak{h}^{m_1+m_2}(\mathbf{x},\mathbf{y}) \le \mathfrak{h}^{m_1}(\mathbf{x}) + \mathfrak{h}^{m_2}(\mathbf{y}) \tag{7.47}$$

and

$$\mathfrak{h}^{m_1}(\mathbf{x}\,|\,\mathbf{y}) \le \mathfrak{h}^{m_1}(\mathbf{x}) \tag{7.48}$$

provided the entropies $\mathfrak{h}^{m_1}(\mathbf{x})$, $\mathfrak{h}^{m_2}(\mathbf{y})$, and $\mathfrak{h}^{m_1+m_2}(\mathbf{x},\mathbf{y})$ exist and are finite.

Proof. The inequality (7.47) follows from (7.45) and the nonnegativity of mutual information. Similarly, (7.48) follows from (7.46) and the nonnegativity of mutual information. □

7.4 Asymptotic Equipartition Property

Similar to classical entropy and differential entropy, the m-dimensional entropy $\mathfrak{h}^m(\mathbf{x})$ satisfies an asymptotic equipartition property (AEP). Let us consider a sequence $\mathbf{x}_{1:n} \triangleq (\mathbf{x}_1, \dots, \mathbf{x}_n)$ of independent and identically distributed (i.i.d.) random variables $\mathbf{x}_i$. Our main findings are similar to the discrete and continuous cases: based on $\mathfrak{h}^m(\mathbf{x})$, we define sets $\mathcal{A}_\varepsilon^{(n)}$ of typical sequences $\boldsymbol{x}_{1:n}$ and show that, for sufficiently large n, a random sequence $\mathbf{x}_{1:n}$ belongs to $\mathcal{A}_\varepsilon^{(n)}$ with probability arbitrarily close to one. Furthermore, we obtain upper and lower bounds on the size of $\mathcal{A}_\varepsilon^{(n)}$ given by $e^{n(\mathfrak{h}^m(\mathbf{x})+\varepsilon)}$ and $(1-\delta)e^{n(\mathfrak{h}^m(\mathbf{x})-\varepsilon)}$, respectively. In the case of classical entropy and differential entropy these properties are useful in the proof of various coding theorems as they allow us to consider only typical sequences.

Our analysis follows the steps in [Cover and Thomas, 2006, Ch. 8.2]. However, whereas in the discrete case the size of a set of sequences $\boldsymbol{x}_{1:n}$ is measured by its cardinality and in the continuous case by its Lebesgue measure, in the present case of m-rectifiable random variables x_i, we resort to the Hausdorff measure.

Lemma 7.20 Let $\mathsf{x}_{1:n} = (\mathsf{x}_1, \ldots, \mathsf{x}_n)$ be a sequence of i.i.d. m-rectifiable random variables x_i on $\mathbb{R}^M$, where each x_i has the same m-dimensional Hausdorff density θ^m_{x} and m-dimensional entropy $\mathfrak{h}^m(\mathsf{x})$. Then the random variable $-(1/n)\sum_{i=1}^n \log \theta^m_{\mathsf{x}}(\mathsf{x}_i)$ converges to $\mathfrak{h}^m(\mathsf{x})$ in probability, i.e., for any $\varepsilon > 0$

$$\lim_{n\to\infty} \Pr\left\{\left| -\frac{1}{n}\sum_{i=1}^n \log \theta^m_{\mathsf{x}}(\mathsf{x}_i) - \mathfrak{h}^m(\mathsf{x})\right| > \varepsilon\right\} = 0\,. \tag{7.49}$$

Proof. By (6.16), we have $\mathfrak{h}^m(\mathsf{x}) = -\mathbb{E}_{\mathsf{x}}\big[\log \theta^m_{\mathsf{x}}(\mathsf{x})\big]$, and by the weak law of large numbers, the sample mean $-(1/n)\sum_{i=1}^n \log \theta^m_{\mathsf{x}}(\mathsf{x}_i)$ converges to the expectation $-\mathbb{E}_{\mathsf{x}}\big[\log \theta^m_{\mathsf{x}}(\mathsf{x})\big]$ in probability. □

We can define typical sets in the usual way [Cover and Thomas, 2006, Ch. 8.2].

Definition 7.21 Let x be an m-rectifiable random variable on $\mathbb{R}^M$ with support $\mathcal{E}$ and m-dimensional Hausdorff density θ^m_{x}. For $\varepsilon > 0$ and $n \in \mathbb{N}$, the ε-typical set $\mathcal{A}^{(n)}_\varepsilon \subseteq \mathbb{R}^{nM}$ is defined as

$$\mathcal{A}^{(n)}_\varepsilon \triangleq \left\{\boldsymbol{x}_{1:n} \in \mathcal{E}^n : \left| -\frac{1}{n}\sum_{i=1}^n \log \theta^m_{\mathsf{x}}(\boldsymbol{x}_i) - \mathfrak{h}^m(\mathsf{x})\right| \le \varepsilon\right\}. \tag{7.50}$$

Note that $\mathcal{A}^{(n)}_\varepsilon \subseteq \mathcal{E}^n$. The assumption $\boldsymbol{x}_{1:n} \in \mathcal{E}^n$ simplifies working with $\mathcal{A}^{(n)}_\varepsilon$. This is not a strong restriction because, by Property 4 in Corollary 6.10, $\theta^m_{\mathsf{x}}(\boldsymbol{x}) = 0$ $\mathscr{H}^m$-almost everywhere on $\mathcal{E}^c$.

The AEP for sequences of m-rectifiable random variables is expressed by the following central result.

Theorem 7.22 Let $\mathsf{x}_{1:n} = (\mathsf{x}_1, \ldots, \mathsf{x}_n)$ be a sequence of i.i.d. m-rectifiable random variables x_i on $\mathbb{R}^M$, where each x_i has the same m-dimensional Hausdorff density θ^m_{x}, support $\mathcal{E}$, and m-dimensional entropy $\mathfrak{h}^m(\mathsf{x})$. Furthermore, let $\mathcal{E}^n$ and $\mathcal{E}$ be product-compatible for all $n \in \mathbb{N}$. Then the typical set $\mathcal{A}^{(n)}_\varepsilon$ satisfies the following properties.

1. For $\delta > 0$ and n sufficiently large,

$$\Pr\{\mathsf{x}_{1:n} \in \mathcal{A}^{(n)}_\varepsilon\} > 1 - \delta\,.$$

2. For all $n \in \mathbb{N}$,

$$\mathscr{H}^{nm}(\mathcal{A}^{(n)}_\varepsilon) \le e^{n(\mathfrak{h}^m(\mathsf{x})+\varepsilon)}\,. \tag{7.51}$$

3. For $\delta > 0$ and n sufficiently large,

$$\mathscr{H}^{nm}(\mathcal{A}_\varepsilon^{(n)}) > (1-\delta)e^{n(\mathfrak{h}^m(\mathsf{x})-\varepsilon)}\,. \tag{7.52}$$

Proof. See Appendix B.15. □

Chapter 8

Source Coding for Integer-Dimensional Singular Random Variables

Based on the definitions and results provided in the preceding Chapters 6 and 7, we present two applications of integer-dimensional entropy to the field of source coding. First, we extend the classical result that the entropy of a discrete random variable provides a bound on the expected codeword length to the much wider class of integer-dimensional random variables. Second, we present a lower bound on the rate-distortion function of integer-dimensional sources. This bound provides an extension of the continuous Shannon lower bound.

8.1 Entropy Bounds on Expected Codeword Length

A well-known result for discrete random variables is a connection between the minimal expected codeword length of a lossless source code and the entropy of the random variable [Cover and Thomas, 2006, Th. 5.4.1]. More specifically, let x be a discrete random variable on $\mathbb{R}^M$ with possible realizations $\{\boldsymbol{x}_i : i \in \mathcal{I}\}$. In variable-length lossless source coding, a one-to-one function $f\colon \{\boldsymbol{x}_i : i \in \mathcal{I}\} \to \{0,1\}^*$, where $\{0,1\}^*$ denotes the set of all finite-length binary sequences, is used to represent each realization $\boldsymbol{x}_i$ by a finite-length binary sequence $\boldsymbol{s}_i = f(\boldsymbol{x}_i)$. The expected codeword length is defined as

$$L_f(\mathsf{x}) \triangleq \mathbb{E}_{\mathsf{x}}[\ell(f(\mathsf{x}))]$$

where $\ell(\boldsymbol{s})$ denotes the length of a binary sequence $\boldsymbol{s} \in \{0,1\}^*$. The minimal expected binary codeword length $L^*(\mathsf{x})$ is defined as the minimum of $L_f(\mathsf{x})$ over the set of all possible one-to-one functions f. By [Cover and Thomas, 2006, Th. 5.4.1], $L^*(\mathsf{x})$ satisfies[1]

$$H(\mathsf{x}) \operatorname{ld} e \leq L^*(\mathsf{x}) < H(\mathsf{x}) \operatorname{ld} e + 1\,. \tag{8.1}$$

[1] The factor $\operatorname{ld} e$ appears because we defined entropy using the natural logarithm.

For a nondiscrete m-rectifiable random variable x (i.e., $m \geq 1$), a lossless code of finite codeword length does not exist. However, quantizations of x can be encoded using finite-length binary sequences. We will present results for the minimal expected codeword length of constrained quantizations of x.

Definition 8.1 Let $\mathcal{E} \subseteq \mathbb{R}^M$ be an m-rectifiable set. Furthermore, let $\mathfrak{Q} = \{\mathcal{A}_1, \ldots, \mathcal{A}_N\}$ be a finite $\mathscr{H}^m$-measurable partition of $\mathcal{E}$, i.e., all sets $\mathcal{A}_i$ are mutually disjoint and $\mathscr{H}^m$-measurable, and $\bigcup_{i=1}^N \mathcal{A}_i = \mathcal{E}$. The partition $\mathfrak{Q}$ is said to be an (m,δ)*-partition of* $\mathcal{E}$ if $\mathscr{H}^m(\mathcal{A}_i) \leq \delta$ for all $i \in \{1, \ldots, N\}$. The set of all (m,δ)-partitions of $\mathcal{E}$ is denoted $\mathfrak{P}_{m,\delta}^{(\mathcal{E})}$.

Note that the definition of an (m,δ)-partition of an m-rectifiable set $\mathcal{E}$ does not involve a distortion function. On the one hand, this is convenient because we do not have to argue about a good distortion measure. On the other hand, the points in a set $\mathcal{A}_i$ of a partition $\mathfrak{Q} \in \mathfrak{P}_{m,\delta}^{(\mathcal{E})}$ are not necessarily "close" to each other; in fact, $\mathcal{A}_i$ is not even necessarily connected. Thus, although the partitions in $\mathfrak{P}_{m,\delta}^{(\mathcal{E})}$ contain measure-theoretically small sets, these sets might be considered large in terms of specific distortion measures.

In what follows, we will consider the quantized random variable $[\mathsf{x}]_{\mathfrak{Q}}$ for $\mathfrak{Q} \in \mathfrak{P}_{m,\delta}^{(\mathcal{E})}$. We recall that $[\mathsf{x}]_{\mathfrak{Q}}$ is the discrete random variable such that $\Pr\{[\mathsf{x}]_{\mathfrak{Q}} = i\} = \Pr\{\mathsf{x} \in \mathcal{A}_i\}$ for $i \in \{1, \ldots, N\}$. We first prove an expression of the m-dimensional entropy of an m-rectifiable random variable x as the infimum of the entropy of quantizations $[\mathsf{x}]_{\mathfrak{Q}}$. This expression will be used in the proof of Theorem 8.3.

Lemma 8.2 Let x be an m-rectifiable random variable with $m \geq 1$ and support $\mathcal{E}$ satisfying $\mathscr{H}^m(\mathcal{E}) < \infty$. Let $\mathfrak{P}_{m,\infty}^{(\mathcal{E})}$ denote the set of all finite, measurable partitions of $\mathcal{E}$. Then

$$\mathfrak{h}^m(\mathsf{x}) = \inf_{\mathfrak{Q} \in \mathfrak{P}_{m,\infty}^{(\mathcal{E})}} \left(- \sum_{\mathcal{A} \in \mathfrak{Q}} \mu \mathsf{x}^{-1}(\mathcal{A}) \log \left(\frac{\mu \mathsf{x}^{-1}(\mathcal{A})}{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A})} \right) \right) \tag{8.2}$$

$$= \inf_{\mathfrak{Q} \in \mathfrak{P}_{m,\infty}^{(\mathcal{E})}} \left(H([\mathsf{x}]_{\mathfrak{Q}}) + \sum_{\mathcal{A} \in \mathfrak{Q}} \mu \mathsf{x}^{-1}(\mathcal{A}) \log \mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}) \right). \tag{8.3}$$

Proof. See Appendix B.16. □

The terms in (8.3) give an interesting interpretation of m-dimensional entropy. Looking for a quantization that minimizes $H([\mathsf{x}]_{\mathfrak{Q}})$ corresponds to minimizing the amount of data required to represent this quantization. Of course, the minimum is simply obtained with the partition $\mathfrak{Q} = \{\mathcal{E}\}$, which gives $H([\mathsf{x}]_{\mathfrak{Q}}) = 0$. But in (8.3), we also have a term that penalizes a bad "resolution" of the quantization: If the quantized random variable $[\mathsf{x}]_{\mathfrak{Q}}$ is with high probability—corresponding to $\mu \mathsf{x}^{-1}(\mathcal{A})$ being large—in a large quantization set $\mathcal{A}$, then this is penalized by the term $\mu \mathsf{x}^{-1}(\mathcal{A}) \log \mathscr{H}^m|_{\mathcal{E}}(\mathcal{A})$. Thus, (8.3) shows that m-dimensional entropy can be interpreted in terms of a tradeoff between fine resolution and efficient representation.

We now turn to a generalization of (8.1) to rectifiable random variables.

Theorem 8.3 Let x be an m-rectifiable random variable with $m \geq 1$ and support $\mathcal{E}$ satisfying $\mathscr{H}^m(\mathcal{E}) < \infty$. For any $\mathfrak{Q} \in \mathfrak{P}_{m,\delta}^{(\mathcal{E})}$, the minimal expected binary codeword length of the quantized random variable $[\mathsf{x}]_{\mathfrak{Q}}$ satisfies

$$L^*([\mathsf{x}]_{\mathfrak{Q}}) \geq \mathfrak{h}^m(\mathsf{x}) \operatorname{ld} e - \operatorname{ld} \delta \,. \tag{8.4}$$

Furthermore, for each $\varepsilon > 0$, there exists $\delta_\varepsilon > 0$ such that the following holds: for each $\delta \in (0, \delta_\varepsilon)$, there exists a partition $\mathfrak{Q}_\delta \in \mathfrak{P}_{m,\delta}^{(\mathcal{E})}$ such that

$$L^*([\mathsf{x}]_{\mathfrak{Q}_\delta}) < \mathfrak{h}^m(\mathsf{x}) \operatorname{ld} e - \operatorname{ld} \delta + 1 + \varepsilon \,. \tag{8.5}$$

Proof. See Appendix B.17. We note that the proof is based on (8.1) and the representation of $\mathfrak{h}^m(\mathsf{x})$ given in Lemma 8.2. □

The bound (8.4) shows the following: if we want a quantization $\mathfrak{Q}$ of x with good resolution (in the sense that $\mathscr{H}^m(\mathcal{A}) \leq \delta$ for all $\mathcal{A} \in \mathfrak{Q}$), then we have to use *at least* $\mathfrak{h}^m(\mathsf{x}) \operatorname{ld} e - \operatorname{ld} \delta$ bits to represent this quantized random variable. However, by (8.5), we know that for a sufficiently fine resolution (i.e., $\delta < \delta_\varepsilon$), we have to use *at most* $1 + \varepsilon$ additional bits (in addition to the lower bound $\mathfrak{h}^m(\mathsf{x}) \operatorname{ld} e - \operatorname{ld} \delta$) to achieve the desired resolution.

We will now apply Theorem 8.3 to sequences of i.i.d. random variables. To this end, we consider quantizations of an entire sequence, $[\mathsf{x}_{1:n}]_{\mathfrak{Q}} = [(\mathsf{x}_1, \dots, \mathsf{x}_n)]_{\mathfrak{Q}}$ with $\mathfrak{Q} \in \mathfrak{P}_{nm,\delta^n}^{(\mathcal{E}^n)}$. We denote by $L_n^*([\mathsf{x}_{1:n}]_{\mathfrak{Q}}) \triangleq L^*([\mathsf{x}_{1:n}]_{\mathfrak{Q}})/n$ the minimal expected binary codeword length per source symbol.

Corollary 8.4 Let $\mathsf{x}_{1:n} = (\mathsf{x}_1, \dots, \mathsf{x}_n)$ be a sequence of i.i.d. m-rectifiable random variables ($m \geq 1$) on $\mathbb{R}^M$ with m-dimensional entropy $\mathfrak{h}^m(\mathsf{x})$ and support $\mathcal{E}$ satisfying $\mathscr{H}^m(\mathcal{E}) < \infty$. Furthermore, assume that $\mathcal{E}^i$ and $\mathcal{E}$ are product-compatible for $i \in \{1, \dots, n-1\}$. Then, for each $\varepsilon > 0$, there exists $\delta_\varepsilon > 0$ such that the following holds: for each $\delta \in (0, \delta_\varepsilon)$, there exists a partition $\mathfrak{Q} \in \mathfrak{P}_{nm,\delta^n}^{(\mathcal{E}^n)}$ such that the minimal expected binary codeword length per source symbol satisfies

$$\mathfrak{h}^m(\mathsf{x}) \operatorname{ld} e - \operatorname{ld} \delta \leq L_n^*([\mathsf{x}_{1:n}]_{\mathfrak{Q}}) \leq \mathfrak{h}^m(\mathsf{x}) \operatorname{ld} e - \operatorname{ld} \delta + \frac{1 + \varepsilon}{n} \,. \tag{8.6}$$

Proof. By Corollary 7.5, the random variable $\mathsf{x}_{1:n}$ is nm-rectifiable with support $\mathcal{E}^n$ and nm-dimensional entropy $\mathfrak{h}^{nm}(\mathsf{x}_{1:n}) = n\mathfrak{h}^m(\mathsf{x})$. Thus, by Theorem 8.3, there exists $\hat{\delta}_\varepsilon > 0$ such that the following holds:

($*$) For all $\hat{\delta} \in (0, \hat{\delta}_\varepsilon)$ there exists a partition $\mathfrak{Q} \in \mathfrak{P}_{nm,\hat{\delta}}^{(\mathcal{E}^n)}$ satisfying

$$n\mathfrak{h}^m(\mathsf{x}) \operatorname{ld} e - \operatorname{ld} \hat{\delta} \leq L^*([\mathsf{x}_{1:n}]_{\mathfrak{Q}}) < n\mathfrak{h}^m(\mathsf{x}) \operatorname{ld} e - \operatorname{ld} \hat{\delta} + 1 + \varepsilon \,.$$

Choosing $\delta_\varepsilon \triangleq \hat{\delta}_\varepsilon^{1/n}$, we have for $\delta \in (0, \delta_\varepsilon)$ that $\delta^n \in (0, \hat{\delta}_\varepsilon)$. Thus, by ($*$), there exists a partition $\mathfrak{Q} \in \mathfrak{P}_{nm,\delta^n}^{(\mathcal{E}^n)}$ satisfying

$$n\mathfrak{h}^m(\mathsf{x}) \operatorname{ld} e - \operatorname{ld} \delta^n \leq L^*([\mathsf{x}_{1:n}]_{\mathfrak{Q}}) < n\mathfrak{h}^m(\mathsf{x}) \operatorname{ld} e - \operatorname{ld} \delta^n + 1 + \varepsilon\,.$$

Dividing by n gives (8.6). □

Corollary 8.4 shows that the upper bound on the expected codeword length per source symbol becomes closer to the lower bound $\mathfrak{h}^m(\mathsf{x}) \operatorname{ld} e - \operatorname{ld} \delta$ if we are allowed to quantize and code entire sequences. However, note that using the quantization $\mathfrak{Q} \in \mathfrak{P}_{nm,\delta^n}^{(\mathcal{E}^n)}$ of the joint random variable $\mathsf{x}_{1:n}$, we cannot always reconstruct each x_i to within a set $\mathcal{A}_i$ satisfying $\mathscr{H}^m(\mathcal{A}_i) \leq \delta$. All we know is that each $\mathcal{A} \in \mathfrak{Q}$ satisfies $\mathscr{H}^{nm}(\mathcal{A}) \leq \delta^n$, i.e., the overall resolution of the sequence is good, but the resolution of each individual source symbol is not guaranteed to be also good.

8.2 Shannon Lower Bound for Integer-Dimensional Sources

The problem considered in rate-distortion (RD) theory is to represent a given random variable x using as few values as possible while keeping the expected distortion below some threshold [Gray, 1990, Ch. 4]. We will consider throughout this section a distance distortion function $d(\cdot,\cdot)$ on $\mathbb{R}^M \times \mathbb{R}^M$, i.e., $d(\boldsymbol{x},\boldsymbol{y}) = d(\boldsymbol{x}-\boldsymbol{y},\boldsymbol{0})$. Furthermore, we assume that $d(\cdot,\cdot)$ satisfies $\inf_{\boldsymbol{y}\in\mathbb{R}^M} d(\boldsymbol{x},\boldsymbol{y}) = 0$ for each $\boldsymbol{x} \in \mathbb{R}^M$. The RD function is then defined as [Gray, 1990, eq. (4.1.3)]

$$R(D) \triangleq \inf_{\mathbb{E}_{(\mathsf{x},\mathsf{y})}[d(\mathsf{x},\mathsf{y})]\leq D} I(\mathsf{x};\mathsf{y})$$

for $D \geq 0$. Here, the above constrained infimum is taken over all joint probability distributions (x,y) with the fixed probability distribution of x as the first marginal. We assume that there exist $D \geq 0$ such that $R(D)$ is finite, and we denote by D_0 the infimum of these D. Furthermore, we assume that there exists a finite set $\mathcal{B} \subseteq \mathbb{R}^M$ such that

$$\mathbb{E}_{\mathsf{x}}\Big[\min_{\boldsymbol{y}\in\mathcal{B}} d(\mathsf{x},\boldsymbol{y})\Big] < \infty\,.$$

This assumption guarantees that there exists a finite quantization of x with bounded expected distortion. Under these standard assumptions, we have the following characterization of the RD function [Csiszár, 1974, Th. 2.3]: For each $D > D_0$,

$$R(D) = \max_{s\geq 0} \max_{\alpha_s(\boldsymbol{x})} \big(-sD + \mathbb{E}_{\mathsf{x}}[\log \alpha_s(\mathsf{x})]\big) \tag{8.7}$$

where the second maximization is with respect to all functions[2] $\alpha_s \colon \mathbb{R}^M \to (0,\infty)$ satisfying for each $\boldsymbol{y} \in \mathbb{R}^M$

$$\mathbb{E}_{\mathsf{x}}\big[\alpha_s(\mathsf{x}) e^{-sd(\mathsf{x},\boldsymbol{y})}\big] \leq 1\,. \tag{8.8}$$

[2]Although in [Csiszár, 1974, Th. 2.3] $\alpha_s(\boldsymbol{x}) \geq 1$ is assumed, (8.7) also holds for $\alpha_s(\boldsymbol{x}) > 0$ because of [Csiszár, 1974, Lem. 1.2 and Lem. 1.4].

8.2.1 Shannon Lower Bound

The most common form of the traditional Shannon lower bound [Gray, 1990, Ch. 4.3] for a *discrete* source x is the following inequality

$$R(D) \geq H(\mathsf{x}) - \max H(\mathsf{w}) \tag{8.9}$$

where the maximum is taken over all random variables w whose expected distortion relative to $\mathbf{0}$ is equal to D, i.e., $\mathbb{E}_{\mathsf{w}}\big[d(\mathsf{w}, \mathbf{0})\big] = D$. An important part of the bound (8.9) is that the contribution of the source x and the contribution of the distortion D and distortion function $d(\cdot, \cdot)$ become separated. For a fixed distortion function and a given distortion, we can calculate $\max H(\mathsf{w})$ and then use the bound (8.9) for different sources x simply by calculating their entropy $H(\mathsf{x})$.

For a *continuous* random variable x on $\mathbb{R}^M$, a bound similar to (8.9) can also be derived under certain assumptions. However, it is more convenient to state the continuous Shannon lower bound in the following parametric form [Gray, 1990, Ch. 4.6]

$$R(D) \geq h(\mathsf{x}) - sD - \log \widetilde{\gamma}(s) \tag{8.10}$$

where

$$\widetilde{\gamma}(s) \triangleq \int_{\mathbb{R}^M} e^{-sd(\boldsymbol{x}, \mathbf{0})} \, \mathrm{d}\mathscr{L}^M(\boldsymbol{x}) \tag{8.11}$$

and (8.10) holds for all $s \geq 0$. The right-hand side of (8.10) can be maximized with respect to s and it turns out that [Gray, 1990, Lem. 4.6.2]

$$\min_{s \geq 0} \big(sD + \log \widetilde{\gamma}(s)\big) = \max h(\mathsf{w})$$

where the maximum is taken over all continuous random variables w whose expected distortion relative to $\mathbf{0}$ is equal to D, i.e., $\mathbb{E}_{\mathsf{w}}\big[d(\mathsf{w}, \mathbf{0})\big] = D$. This results again in the simple formula (cf. (8.9))

$$R(D) \geq h(\mathsf{x}) - \max h(\mathsf{w}) \,.$$

Because the parametric bound (8.10) is more convenient in most cases and already allows us to separate the source from the distortion, we will concentrate on a generalization of (8.10) to rectifiable random variables. To this end, we will use the characterization of the RD function in (8.7) with a specific choice of the function α_s.

Theorem 8.5 Let x be an m-rectifiable random variable with support $\mathcal{E}$, and let

$$\gamma(s) \triangleq \sup_{\boldsymbol{y} \in \mathbb{R}^M} \int_{\mathcal{E}} e^{-sd(\boldsymbol{x}, \boldsymbol{y})} \, \mathrm{d}\mathscr{H}^m(\boldsymbol{x}), \quad \text{for } s \geq 0 \,. \tag{8.12}$$

Then for each $s \geq 0$ the RD function is lower bounded by

$$R(D) \geq R_{\mathrm{SLB}}(D, s) \triangleq \mathfrak{h}^m(\mathsf{x}) - sD - \log \gamma(s) \,. \tag{8.13}$$

Proof. We start by noting that (8.12) implies

$$\int_{\mathcal{E}} e^{-sd(\boldsymbol{x},\boldsymbol{y})}\, \mathrm{d}\mathscr{H}^m(\boldsymbol{x}) \leq \gamma(s) \tag{8.14}$$

for all $\boldsymbol{y} \in \mathbb{R}^M$. Let $s \geq 0$ be fixed. By (8.7),

$$R(D) \geq -sD + \mathbb{E}_{\mathbf{x}}[\log \alpha_s(\mathbf{x})] \tag{8.15}$$

for every function α_s satisfying (8.8). We have

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}}\left[\frac{1}{\theta^m_{\mathbf{x}}(\mathbf{x})\gamma(s)} e^{-sd(\mathbf{x},\boldsymbol{y})}\right] &= \int_{\mathcal{E}} \theta^m_{\mathbf{x}}(\boldsymbol{x}) \frac{1}{\theta^m_{\mathbf{x}}(\boldsymbol{x})\gamma(s)} e^{-sd(\boldsymbol{x},\boldsymbol{y})}\, \mathrm{d}\mathscr{H}^m(\boldsymbol{x}) \\
&= \frac{1}{\gamma(s)} \int_{\mathcal{E}} e^{-sd(\boldsymbol{x},\boldsymbol{y})}\, \mathrm{d}\mathscr{H}^m(\boldsymbol{x}) \\
&\overset{(8.14)}{\leq} \frac{\gamma(s)}{\gamma(s)} \\
&= 1
\end{aligned}$$

for all $\boldsymbol{y} \in \mathbb{R}^M$. Therefore, the choice $\alpha_s(\boldsymbol{x}) \triangleq \frac{1}{\theta^m_{\mathbf{x}}(\boldsymbol{x})\gamma(s)}$ satisfies (8.8). Inserting $\alpha_s(\boldsymbol{x}) = \frac{1}{\theta^m_{\mathbf{x}}(\boldsymbol{x})\gamma(s)}$ into (8.15), we obtain

$$\begin{aligned}
R(D) &\geq -sD + \mathbb{E}_{\mathbf{x}}\left[\log \frac{1}{\theta^m_{\mathbf{x}}(\mathbf{x})\gamma(s)}\right] \\
&= -\mathbb{E}_{\mathbf{x}}[\log \theta^m_{\mathbf{x}}(\mathbf{x})] - sD - \mathbb{E}_{\mathbf{x}}[\log \gamma(s)] \\
&= \mathfrak{h}^m(\mathbf{x}) - sD - \log \gamma(s)\,.
\end{aligned}$$

□

Note that for continuous random variables with positive probability density function almost everywhere (i.e., M-rectifiable with support $\mathbb{R}^M$), the definitions of $\widetilde{\gamma}(s)$ in (8.11) and $\gamma(s)$ in (8.12) coincide. Indeed, because $d(\boldsymbol{x},\boldsymbol{y}) = d(\boldsymbol{x}-\boldsymbol{y},\mathbf{0})$ and a translation by $\boldsymbol{y}$ does not change the value of the integral over $\mathbb{R}^M$, (8.11) becomes (recall that $\mathscr{H}^M = \mathscr{L}^M$)

$$\int_{\mathbb{R}^M} e^{-sd(\boldsymbol{x},\mathbf{0})}\, \mathrm{d}\mathscr{L}^M(\boldsymbol{x}) = \int_{\mathbb{R}^M} e^{-sd(\boldsymbol{x},\boldsymbol{y})}\, \mathrm{d}\mathscr{H}^M(\boldsymbol{x}) \tag{8.16}$$

for any $\boldsymbol{y} \in \mathbb{R}^M$. Because the left-hand side in (8.16) does not depend on $\boldsymbol{y}$, taking the supremum over $\boldsymbol{y} \in \mathbb{R}^M$ in (8.16) results in

$$\int_{\mathbb{R}^M} e^{-sd(\boldsymbol{x},\mathbf{0})}\, \mathrm{d}\mathscr{L}^M(\boldsymbol{x}) = \sup_{\boldsymbol{y}\in\mathbb{R}^M} \int_{\mathbb{R}^M} e^{-sd(\boldsymbol{x},\boldsymbol{y})}\, \mathrm{d}\mathscr{H}^M(\boldsymbol{x})$$

which is (8.12). Thus, in this case, the Shannon lower bounds (8.10) and (8.13) coincide. However, for continuous random variables with a smaller support $\mathcal{E} \subsetneq \mathbb{R}^M$, the Shannon lower bound

(8.13) is tighter (i.e., larger) than (8.10). This is due to the fact that (8.13) incorporates the additional information that the random variable is restricted to $\mathcal{E}$.

The optimal choice of s in (8.13) depends on D and is in general hard to find. In fact, we do not even know whether the optimal (i.e., largest) lower bound $R_{\text{SLB}}(D, s)$ is achieved for a finite s or $R_{\text{SLB}}(D, s)$ increases as s goes to ∞. The following lemma shows that the latter alternative is not possible.

Lemma 8.6 Let $\mathbf{x}$ be an m-rectifiable random variable with support $\mathcal{E}$ and finite m-dimensional entropy $\mathfrak{h}^m(\mathbf{x})$. Then for $D > D_0$ the lower bound $R_{\text{SLB}}(D, s)$ in (8.13) satisfies

$$\lim_{s\to\infty} R_{\text{SLB}}(D, s) = -\infty\,. \tag{8.17}$$

Proof. See Appendix B.18. □

If $R_{\text{SLB}}(D, s)$ is continuous, Lemma 8.6 implies that the global maximum of $R_{\text{SLB}}(D, s)$ for a fixed $D > D_0$ exists and is either a local maximum or the boundary point $s = 0$. If $\gamma(s)$ is differentiable with respect to s, we can specify the local maxima of $R_{\text{SLB}}(D, s)$.

Corollary 8.7 Let $\mathbf{x}$ be an m-rectifiable random variable with support $\mathcal{E}$, and let $\gamma(s)$ (see (8.12)) be differentiable with respect to s. Then for $D > D_0$ the lower bound $R_{\text{SLB}}(D, s)$ in (8.13) is maximized either for $s = 0$ or for some $s > 0$ satisfying

$$D = \widetilde{D}(s) \triangleq -\frac{\gamma'(s)}{\gamma(s)}\,. \tag{8.18}$$

Proof. Because $\gamma(s)$ is differentiable, we can differentiate $R_{\text{SLB}}(D, s)$ with respect to s and then set the result to zero to obtain a necessary condition for a local maximum. Solving the resulting equation for D yields (8.18). Thus, for a given $D > D_0$, $R_{\text{SLB}}(D, s)$ can only have a local maximum $s \in (0, \infty)$ for s satisfying (8.18). By Lemma 8.6, the global maximum can either be a local maximum or is achieved for $s = 0$, which concludes the proof. □

If $\gamma(s)$ is differentiable, Corollary 8.7 provides a "parametrization" of the graph of the largest bound $R_{\text{SLB}}(D, s)$, i.e., we can characterize all pairs $(D, \sup_{s\geq 0} R_{\text{SLB}}(D, s))$ for $D > D_0$.

Corollary 8.8 Let $\mathbf{x}$ be an m-rectifiable random variable with support $\mathcal{E}$, and let $\gamma(s)$ (see (8.12)) be differentiable with respect to s. Then

$$\Big\{\Big(D, \sup_{s\geq 0} R_{\text{SLB}}(D, s)\Big) \in \mathbb{R}^2 : D > D_0\Big\} \\ \subseteq \big\{\big(\widetilde{D}(s), R_{\text{SLB}}(\widetilde{D}(s), s)\big) : s > 0\big\} \cup \big\{\big(D, \mathfrak{h}^m(\mathbf{x}) - \log \mathscr{H}^m(\mathcal{E})\big) : D > D_0\big\}\,. \tag{8.19}$$

Proof. Let $(D, R_{\text{SLB}}(D, s))$ belong to the set on the left-hand side in (8.19). By Corollary 8.7, this implies $s = 0$ or $D = \widetilde{D}(s)$.

Case $s = 0$: In this case, we have

$$\begin{aligned}(D, R_{\text{SLB}}(D,s)) &= (D, R_{\text{SLB}}(D,0)) \\ &\overset{(8.13)}{=} (D, \mathfrak{h}^m(\mathsf{x}) - \log\gamma(0)) \\ &\overset{(a)}{=} (D, \mathfrak{h}^m(\mathsf{x}) - \log\mathscr{H}^m(\mathcal{E}))\end{aligned} \tag{8.20}$$

where (a) holds because $\gamma(0) \overset{(8.12)}{=} \int_{\mathcal{E}} 1 \,\mathrm{d}\mathscr{H}^m(\boldsymbol{x}) = \mathscr{H}^m(\mathcal{E})$. By (8.20), $(D, R_{\text{SLB}}(D,s))$ belongs to the second set on the right-hand side in (8.19).

Case $D = \widetilde{D}(s)$: We have $(D, R_{\text{SLB}}(D,s)) = \big(\widetilde{D}(s), R_{\text{SLB}}(\widetilde{D}(s), s)\big)$, which belongs to the first set on the right-hand side in (8.19).

In either case $(D, R_{\text{SLB}}(D,s))$ belongs to the right-hand side in (8.19), which concludes the proof. □

Corollary 8.8 implies that we can construct the graph of the best Shannon lower bound by constructing the pairs $\big(\widetilde{D}(s), R_{\text{SLB}}(\widetilde{D}(s), s)\big)$ for all $s > 0$ and the pairs $(D, \mathfrak{h}^m(\mathsf{x}) - \log\mathscr{H}^m(\mathcal{E}))$ for all $D > D_0$ and taking the upper envelope of the resulting set. This idea results in the following program:

(P1) For $s > 0$, calculate $\widetilde{D}(s)$.

(P2) Plot the s-parametrized curve $\big(\widetilde{D}(s), R_{\text{SLB}}(\widetilde{D}(s), s)\big)$.

(P3) Plot the horizontal line $\big(D, \mathfrak{h}^m(\mathsf{x}) - \log\mathscr{H}^m(\mathcal{E})\big)$ for $D > D_0$.

(P4) The upper envelope of the resulting curve is the best Shannon lower bound.

Remark 8.1 One can show that, under certain smoothness conditions, the supremum in (8.12) is in fact a maximum, and $\widetilde{D}(s)$ can be rewritten as

$$\widetilde{D}(s) = D^*(s) \triangleq \frac{1}{\gamma(s)} \int_{\mathcal{E}} d(\boldsymbol{x}, \tilde{\boldsymbol{y}}(s)) e^{-sd(\boldsymbol{x}, \tilde{\boldsymbol{y}}(s))} \,\mathrm{d}\mathscr{H}^m(\boldsymbol{x})$$

where $\tilde{\boldsymbol{y}}(s)$ is the maximizing value in the definition of $\gamma(s)$ (see (8.12)):

$$\tilde{\boldsymbol{y}}(s) \triangleq \operatorname*{arg\,max}_{\boldsymbol{y} \in \mathbb{R}^M} \int_{\mathcal{E}} e^{-sd(\boldsymbol{x}, \boldsymbol{y})} \,\mathrm{d}\mathscr{H}^m(\boldsymbol{x})\,. \tag{8.21}$$

(Thus, $\gamma(s) = \int_{\mathcal{E}} e^{-sd(\boldsymbol{x}, \tilde{\boldsymbol{y}}(s))} \,\mathrm{d}\mathscr{H}^m(\boldsymbol{x})$.) Therefore, we can also apply the program (P1)–(P4) with $D^*(s)$ in place of $\widetilde{D}(s)$. In fact, even if $\gamma(s)$ is not differentiable, we can use the program (P1)–(P4) with $D^*(s)$ to obtain a lower bound on the RD function (although, we do not know the optimality of the resulting bound). Indeed, by Theorem 8.5, the points calculated in steps (P2) and (P3) indicate lower bounds on the RD function.

8.2.2 Shannon Lower Bound on the Unit Circle

To demonstrate the practical relevance of Theorem 8.5, we apply it to the simple example given by $\mathcal{E} = \mathcal{S}_1$, i.e., the unit circle in $\mathbb{R}^2$, and squared error distortion, i.e., $d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|^2$. In order to calculate $\gamma(s)$, we first show that $\tilde{\boldsymbol{y}}(s)$ in (8.21) exists, i.e., that $\gamma(s) = \max_{\boldsymbol{y}\in\mathbb{R}^2} \int_{\mathcal{S}_1} e^{-s\|\boldsymbol{x}-\boldsymbol{y}\|^2}\, \mathrm{d}\mathscr{H}^1(\boldsymbol{x})$ for all $s > 0$. Let $s > 0$ be arbitrary but fixed. Note that we can restrict to $\boldsymbol{y} = (y_1\ 0)^{\mathrm{T}}$, with $y_1 \geq 0$, because the problem is invariant under rotations. Thus,

$$\int_{\mathcal{S}_1} e^{-s\|\boldsymbol{x}-\boldsymbol{y}\|^2}\, \mathrm{d}\mathscr{H}^1(\boldsymbol{x}) = \int_{\mathcal{S}_1} e^{-s((x_1-y_1)^2+x_2^2)}\, \mathrm{d}\mathscr{H}^1(\boldsymbol{x})$$

and therefore we have to maximize the function

$$f_s(y_1) \triangleq \int_{\mathcal{S}_1} e^{-s((x_1-y_1)^2+x_2^2)}\, \mathrm{d}\mathscr{H}^1(\boldsymbol{x})$$

on $[0, \infty)$. To this end, we consider the derivative f_s'. Because $\mathscr{H}^1|_{\mathcal{S}_1}$ is a finite measure and $e^{-s((x_1-y_1)^2+x_2^2)} \leq 1$ for $(x_1\ x_2)^{\mathrm{T}} \in \mathcal{S}_1$, we can change the order of differentiation and integration. This results in the expression

$$f_s'(y_1) = \int_{\mathcal{S}_1} 2s(x_1 - y_1)\, e^{-s((x_1-y_1)^2+x_2^2)}\, \mathrm{d}\mathscr{H}^1(\boldsymbol{x})\,. \tag{8.22}$$

Because $x_1 \leq 1$ for $\boldsymbol{x} \in \mathcal{S}_1$, we have $f_s'(y_1) < 0$ for $y_1 > 1$, i.e., f_s is monotonically decreasing on $(1, \infty)$. Thus, the function f_s can only attain its maximum in the compact interval $[0, 1]$. Because f_s is a continuous function, we conclude that $\gamma(s) = \max_{\boldsymbol{y}\in\mathbb{R}^2} \int_{\mathcal{S}_1} e^{-s\|\boldsymbol{x}-\boldsymbol{y}\|^2}\, \mathrm{d}\mathscr{H}^1(\boldsymbol{x})$ exists for each $s > 0$.

To characterize $\gamma(s)$ in more detail, we consider the equation $f_s'(y_1) = 0$ to find local maxima. By (8.22) and because $x_1^2 + x_2^2 = 1$ for $\boldsymbol{x} \in \mathcal{S}_1$, $f_s'(y_1) = 0$ is equivalent to

$$2se^{-s(1+y_1^2)} \int_{\mathcal{S}_1} (x_1 - y_1)\, e^{2sx_1y_1}\, \mathrm{d}\mathscr{H}^1(\boldsymbol{x}) = 0\,. \tag{8.23}$$

Furthermore, because $2se^{-s(1+y_1^2)} > 0$ and using the transformation $x_1 = \cos\phi$, $x_2 = \sin\phi$, we obtain that (8.23) is equivalent to

$$\int_0^{2\pi} (\cos\phi - y_1)\, e^{2sy_1\cos\phi}\, \mathrm{d}\phi = 0\,. \tag{8.24}$$

Because we know that the function f_s' can only have zeros on $[0, 1]$, we can solve (8.24) numerically for any fixed $s > 0$ and compare the values of f_s at the different solutions and at the boundary points 0 and 1 to find $\gamma(s)$. In Fig. 8.1, the values of $\gamma(s)$ are depicted for $s \in [0.01, 5000]$. We now have all the ingredients to calculate the parametric lower bound $R_{\mathrm{SLB}}(D, s)$ in (8.13) for any given distortion D and an arbitrary source x on $\mathcal{S}_1$. In particular, let us consider the case of a uniform distribution of x on $\mathcal{S}_1$, in which $\mathfrak{h}^1(\mathsf{x}) = \log(2\pi)$ (see (7.10)). In Fig. 8.2, we show

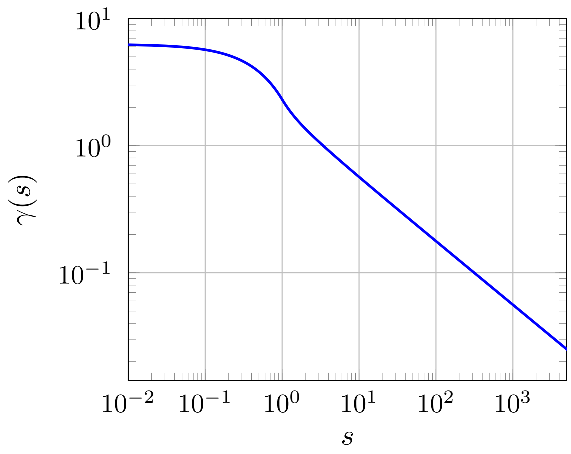

Figure 8.1: Graph of $\gamma(s)$ for different values of s.

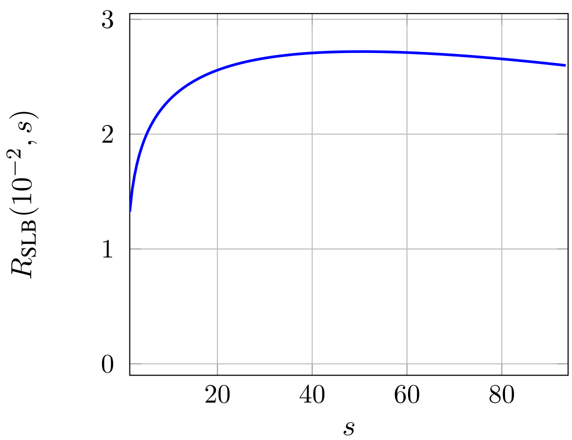

Figure 8.2: Shannon lower bound $R_{\text{SLB}}(10^{-2}, s)$ for $s \in [1, 94]$.

the lower bound $R_{\text{SLB}}(D, s)$ for $s \in [1, 94]$ and distortion $D = 10^{-2}$. It can be seen that the maximal lower bound $R_{\text{SLB}}(10^{-2}, s)$ is obtained for $s \approx 50$. To plot Fig. 8.2, we had to calculate $\gamma(s)$ for many different values of s. We also used "trial and error" to find the region of s where the maximal lower bound $R_{\text{SLB}}(10^{-2}, s)$ arises. To avoid this tedious optimization procedure for all values of D under consideration, we can simply use the program (P1)–(P4). In Fig. 8.3, we show the resulting graph for $s \in [1, 10^5]$ corresponding to bounds on $R(D)$ for $D \in [5 \cdot 10^{-4}, 1]$. We also show in Fig. 8.3 an upper bound using the following result.

Theorem 8.9 Let the random variable x on $\mathbb{R}^2$ be uniformly distributed on the unit circle. For any $n \in \mathbb{N}$,

$$R(\bar{D}(n)) \leq \log n \tag{8.25}$$

where

$$\bar{D}(n) = 1 - \left(\frac{n}{\pi} \sin \frac{\pi}{n}\right)^2. \tag{8.26}$$

Proof. See Appendix B.19. □

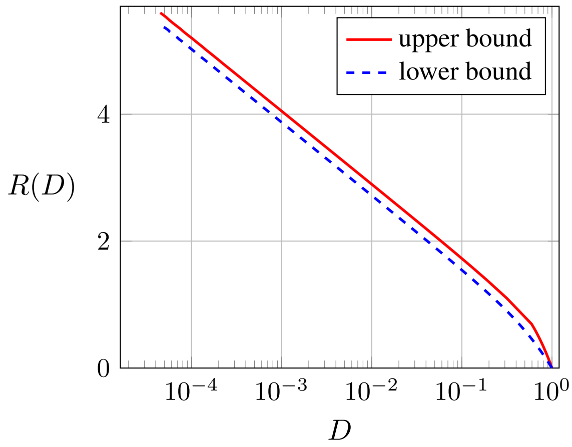

Figure 8.3: Parametrized Shannon lower bound constructed by (P1)–(P4) and upper bound (8.25) on the RD function for a source $\mathbf{x}$ on $\mathbb{R}^2$ uniformly distributed on the unit circle and squared error distortion.

The upper bound depicted in Fig. 8.3 was obtained by linearly interpolating the upper bounds (8.25) corresponding to different values of n (and, hence, $\bar{D}(n)$). This is justified by the convexity of the RD function [Cover and Thomas, 2006, Lem, 10.4.1].

Chapter 9

Conclusion of Part II

We presented a generalization of entropy to singular random variables supported on integer-dimensional subsets of Euclidean space. More specifically, we considered random variables distributed according to a rectifiable measure. Similar to continuous random variables, these rectifiable random variables can be described by a density. However, in contrast to continuous random variables, the density is nonzero only on a lower-dimensional subset and has to be integrated with respect to a Hausdorff measure to calculate probabilities. Our entropy definition is based on this Hausdorff density but otherwise resembles the usual definition of differential entropy. However, this formal similarity has to be interpreted with caution because Hausdorff measures and projections of the rectifiable sets do not always conform to intuition. We thus emphasized mathematical rigor and carefully stated all the assumptions underlying our results.

We showed that for the special cases of rectifiable random variables given by discrete and continuous random variables, our entropy definition reduces to the classical entropy and the differential entropy, respectively. Furthermore, we established a connection between our entropy and differential entropy for a rectifiable random variable that is obtained from a continuous random variable through a one-to-one transformation. For joint and conditional entropy, our analysis showed that the geometry of the support sets of the random variables plays an important role. This role is evidenced by the facts that the chain rule may contain a geometric correction term and conditioning may increase entropy.

Random variables that are neither discrete nor continuous are not only of theoretical interest. Continuity of a random variable cannot be assumed if there are deterministic dependencies reducing the intrinsic dimension of the random variable, which is especially likely to occur in higher-dimensional problems. As a basic example, we considered a random variable $\mathsf{x} \in \mathbb{R}^2$ supported on the unit circle, which is intrinsically only one-dimensional. Here, the differential entropy of x is not defined and, in fact, classical information theory does not provide a rigorous definition of entropy for this random variable.

As an application of our entropy definition to source coding, we provided a characterization of the minimal codeword length of quantizations of integer-dimensional sources. Furthermore, we presented a result in rate-distortion theory that generalizes the Shannon lower bound for discrete and continuous random variables to the larger class of rectifiable random variables. The usefulness of this bound was demonstrated by the example of a uniform source on the unit circle. The resulting bound appears to be the first rigorous lower bound on the rate-distortion function for that distribution.

Possible directions for future work include the extension of our entropy definition to distributions mixing different dimensions (e.g., discrete-continuous mixtures). The extension to noninteger-dimensional singular distributions seems to be possible only in terms of upper and lower entropies, which could be defined based on the upper and lower Hausdorff densities[1] [Ambrosio et al., 2000, Def. 2.55]. Furthermore, our entropy can be extended to infinite-length sequences of rectifiable random variables, which leads to the definition of an entropy rate generalizing the (differential) entropy rate of a sequence of discrete or continuous random variables. Finally, applications of our entropy to source coding and channel coding problems involving integer-dimensional singular random variables are largely unexplored.

[1]The upper and lower Hausdorff densities exist for arbitrary distributions, whereas, by Preiss' Theorem [Preiss, 1987, Th. 5.6], the existence of the Hausdorff density implies that the measure is rectifiable.

Appendices

Appendix A

Proofs of Part I

A.1 Proof of Lemma 2.3

A.1.1 Proof of Statement 1

To prove that almost all of $\mathcal{M}$ can be covered by the union of disjoint measurable subsets $\mathcal{U}_k$, we will use the following lemma, which is an application of the result reported in [Federer, 1969, Cor. 3.2.4].

Lemma A.1 Let $\mathcal{A} \subseteq \mathbb{C}^n$ be a Lebesgue measurable set and $\kappa\colon \mathbb{C}^n \to \mathbb{C}^n$ a continuously differentiable mapping (e.g., the mapping in Lemma 2.3). Then there exists a Lebesgue measurable set $\mathcal{B} \subseteq \mathcal{A} \cap \{\boldsymbol{u} \in \mathbb{C}^n : |\boldsymbol{J}_\kappa(\boldsymbol{u})| \neq 0\}$ such that $\kappa\big|_{\mathcal{B}}$ is one-to-one and $\kappa(\mathcal{A}) \setminus \kappa(\mathcal{B}) = \mathcal{N}$, where $\mathcal{N}$ is a set of Lebesgue measure zero.

We will use Lemma A.1 repeatedly to construct the disjoint sets $\{\mathcal{U}_j\}_{j\in[1:m]}$.

Lemma A.2 Let κ and $\mathcal{M}$ be as in Lemma 2.3, i.e., $\kappa\colon \mathbb{C}^n \to \mathbb{C}^n$ is a continuously differentiable mapping with Jacobian matrix $\boldsymbol{J}_\kappa$ such that $\boldsymbol{J}_\kappa(\boldsymbol{u})$ is nonsingular a.e. and $\mathcal{M} \triangleq \{\boldsymbol{u} \in \mathbb{C}^n : |\boldsymbol{J}_\kappa(\boldsymbol{u})| \neq 0\}$. Assume that for all $\boldsymbol{v} \in \mathbb{C}^n$, the cardinality of the set $\kappa^{-1}(\boldsymbol{v}) \cap \mathcal{M}$ satisfies $|\kappa^{-1}(\boldsymbol{v}) \cap \mathcal{M}| \leq m < \infty$, for some $m \in \mathbb{N}$ (i.e., $\kappa\big|_{\mathcal{M}}$ is finite-to-one). Then, for $k \in [1:m]$, there exist disjoint Lebesgue measurable sets $\{\mathcal{U}_j\}_{j\in[1:k]}$ with $\mathcal{U}_j \subseteq \mathcal{M}$ such that $\kappa\big|_{\mathcal{U}_j}$ is one-to-one for $j \in [1:k]$. Furthermore, there exists a set $\mathcal{N}_k$ of Lebesgue measure zero such that

$$\left|\kappa^{-1}(\boldsymbol{v}) \cap \left(\mathcal{M} \setminus \bigcup_{j\in[1:k]} \mathcal{U}_j\right)\right| \leq m-k\,, \quad \text{for all } \boldsymbol{v} \in \kappa\left(\mathcal{M} \setminus \bigcup_{j\in[1:k-1]} \mathcal{U}_j\right) \setminus \mathcal{N}_k\,. \tag{A.1}$$

Proof. We prove Lemma A.2 by induction over k.

Base case (proof for $k = 1$): By Lemma A.1 with $\mathcal{A} = \mathcal{M}$, we obtain a set $\mathcal{B} \subseteq \mathcal{M}$ (recall that $\mathcal{M} = \{\boldsymbol{u} \in \mathbb{C}^n : |\boldsymbol{J}_\kappa(\boldsymbol{u})| \neq 0\}$ and thus $\mathcal{M} \cap \{\boldsymbol{u} \in \mathbb{C}^n : |\boldsymbol{J}_\kappa(\boldsymbol{u})| \neq 0\} = \mathcal{M}$) such that $\kappa\big|_{\mathcal{B}}$ is one-to-one. Furthermore, $\kappa(\mathcal{M}) \setminus \kappa(\mathcal{B}) = \mathcal{N}_1$ for a set $\mathcal{N}_1$ of Lebesgue measure zero. Because $\kappa(\mathcal{B}) \subseteq \kappa(\mathcal{M})$, this implies $\kappa(\mathcal{M}) \setminus \mathcal{N}_1 = \kappa(\mathcal{B})$. Thus, for each $\boldsymbol{v} \in \kappa(\mathcal{M}) \setminus \mathcal{N}_1$, there exists $\boldsymbol{u} \in \mathcal{B}$ such that $\kappa(\boldsymbol{u}) = \boldsymbol{v}$. Equivalently, $\kappa^{-1}(\boldsymbol{v}) \cap \mathcal{B} \neq \emptyset$. Hence, for $\boldsymbol{v} \in \kappa(\mathcal{M}) \setminus \mathcal{N}_1$,

$$\begin{aligned} |\kappa^{-1}(\boldsymbol{v}) \cap (\mathcal{M} \setminus \mathcal{B})| &= |(\kappa^{-1}(\boldsymbol{v}) \cap \mathcal{M}) \setminus (\kappa^{-1}(\boldsymbol{v}) \cap \mathcal{B})| \\ &\overset{(a)}{=} |\kappa^{-1}(\boldsymbol{v}) \cap \mathcal{M}| - |\kappa^{-1}(\boldsymbol{v}) \cap \mathcal{B}| \\ &\overset{(b)}{=} |\kappa^{-1}(\boldsymbol{v}) \cap \mathcal{M}| - 1 \\ &\overset{(c)}{\leq} m-1 \end{aligned} \tag{A.2}$$

where (a) holds because $\kappa^{-1}(\boldsymbol{v}) \cap \mathcal{B} \subseteq \kappa^{-1}(\boldsymbol{v}) \cap \mathcal{M}$, (b) holds because $\kappa^{-1}(\boldsymbol{v}) \cap \mathcal{B}$ is nonempty and contains at most one element since $\kappa\big|_{\mathcal{B}}$ is one-to-one, and (c) holds because we assumed that $|\kappa^{-1}(\boldsymbol{v}) \cap \mathcal{M}| \leq m$. We set $\mathcal{U}_1 \triangleq \mathcal{B}$ and, by (A.2), the property (A.1) is satisfied for $k = 1$. Furthermore, $\kappa\big|_{\mathcal{U}_1} = \kappa\big|_{\mathcal{B}}$ is one-to-one, which concludes the proof for the base case.

Inductive step (transition from k to $k+1$): Suppose we already constructed the k disjoint measureable sets $\{\mathcal{U}_j\}_{j \in [1:k]}$ and the set $\mathcal{N}_k$ satisfying (A.1). To simplify notation, define

$$\mathcal{U}^{[k]} \triangleq \bigcup_{j \in [1:k]} \mathcal{U}_j \,.$$

Note that (A.1) can now be written as

$$\big|\kappa^{-1}(\boldsymbol{v}) \cap \big(\mathcal{M} \setminus \mathcal{U}^{[k]}\big)\big| \leq m - k \,, \quad \text{for all } \boldsymbol{v} \in \kappa\big(\mathcal{M} \setminus \mathcal{U}^{[k-1]}\big) \setminus \mathcal{N}_k \,. \tag{A.3}$$

By Lemma A.1 with $\mathcal{A} = \mathcal{M} \setminus \mathcal{U}^{[k]}$, we obtain a set $\mathcal{B}$ such that $\kappa\big|_{\mathcal{B}}$ is one-to-one and

$$\mathcal{B} \subseteq \mathcal{M} \setminus \mathcal{U}^{[k]} \,. \tag{A.4}$$

Furthermore, $\kappa\big(\mathcal{M} \setminus \mathcal{U}^{[k]}\big) \setminus \kappa(\mathcal{B}) = \widetilde{\mathcal{N}}_{k+1}$ for a set $\widetilde{\mathcal{N}}_{k+1}$ of Lebesgue measure zero. Because $\kappa(\mathcal{B}) \subseteq \kappa\big(\mathcal{M} \setminus \mathcal{U}^{[k]}\big)$, this implies $\kappa\big(\mathcal{M} \setminus \mathcal{U}^{[k]}\big) \setminus \widetilde{\mathcal{N}}_{k+1} = \kappa(\mathcal{B})$. Hence, for $\boldsymbol{v} \in \kappa\big(\mathcal{M} \setminus \mathcal{U}^{[k]}\big) \setminus \widetilde{\mathcal{N}}_{k+1}$, there exists $\boldsymbol{u} \in \mathcal{B}$ such that $\kappa(\boldsymbol{u}) = \boldsymbol{v}$, or equivalently, $\kappa^{-1}(\boldsymbol{v}) \cap \mathcal{B} \neq \emptyset$. Thus, similarly to (A.2), we obtain for $\boldsymbol{v} \in \kappa\big(\mathcal{M} \setminus \mathcal{U}^{[k]}\big) \setminus \big(\widetilde{\mathcal{N}}_{k+1} \cup \mathcal{N}_k\big)$

$$\begin{aligned} \big|\kappa^{-1}(\boldsymbol{v}) \cap \big(\big(\mathcal{M} \setminus \mathcal{U}^{[k]}\big) \setminus \mathcal{B}\big)\big| &= \big|\big(\kappa^{-1}(\boldsymbol{v}) \cap \big(\mathcal{M} \setminus \mathcal{U}^{[k]}\big)\big) \setminus (\kappa^{-1}(\boldsymbol{v}) \cap \mathcal{B})\big| \\ &\overset{(a)}{=} \big|\kappa^{-1}(\boldsymbol{v}) \cap \big(\mathcal{M} \setminus \mathcal{U}^{[k]}\big)\big| - |\kappa^{-1}(\boldsymbol{v}) \cap \mathcal{B}| \\ &\overset{(b)}{=} \big|\kappa^{-1}(\boldsymbol{v}) \cap \big(\mathcal{M} \setminus \mathcal{U}^{[k]}\big)\big| - 1 \\ &\overset{(c)}{\leq} m - k - 1 \end{aligned} \tag{A.5}$$

where (a) holds because $\kappa^{-1}(\boldsymbol{v}) \cap \mathcal{B} \subseteq \kappa^{-1}(\boldsymbol{v}) \cap \big(\mathcal{M} \setminus \mathcal{U}^{[k]}\big)$, (b) holds because $\kappa^{-1}(\boldsymbol{v}) \cap \mathcal{B}$ is nonempty and contains at most one element since $\kappa\big|_{\mathcal{B}}$ is one-to-one, and (c) holds because of our induction hypothesis (A.3). Setting $\mathcal{U}_{k+1} \triangleq \mathcal{B}$, the left-hand side in (A.5) is equal to $\big|\kappa^{-1}(\boldsymbol{v}) \cap \big(\big(\mathcal{M} \setminus \mathcal{U}^{[k]}\big) \setminus \mathcal{U}_{k+1}\big)\big| = \big|\kappa^{-1}(\boldsymbol{v}) \cap \big(\mathcal{M} \setminus \mathcal{U}^{[k+1]}\big)\big|$, so that (A.5) becomes $\big|\kappa^{-1}(\boldsymbol{v}) \cap \big(\mathcal{M} \setminus \mathcal{U}^{[k+1]}\big)\big| \leq m - k - 1$ for all $\boldsymbol{v} \in \kappa\big(\mathcal{M} \setminus \mathcal{U}^{[k]}\big) \setminus \big(\widetilde{\mathcal{N}}_{k+1} \cup \mathcal{N}_k\big)$. This is exactly the property (A.3) with k replaced by $k+1$ and $\mathcal{N}_k$ replaced by $\mathcal{N}_{k+1} \triangleq \widetilde{\mathcal{N}}_{k+1} \cup \mathcal{N}_k$. Furthermore, we have by (A.4) that $\mathcal{U}_{k+1} = \mathcal{B} \subseteq \mathcal{M} \setminus \mathcal{U}^{[k]}$ and thus $\mathcal{U}_{k+1} \cap \mathcal{U}_j = \emptyset$ for $j \in [1:k]$. Finally, $\kappa\big|_{\mathcal{U}_{k+1}} = \kappa\big|_{\mathcal{B}}$ is one-to-one, which concludes the proof. □

The sets $\{\mathcal{U}_j\}_{j\in[1:m]}$ constructed in Lemma A.2 are disjoint and $\kappa\big|_{\mathcal{U}_j}$ is one-to-one for all $j \in [1:m]$. It remains to be shown that $\mathcal{U}^{[m]} = \bigcup_{j\in[1:m]} \mathcal{U}_j$ covers almost all of $\mathcal{M}$. To this end, we first show that $\kappa\big(\mathcal{M}\setminus\mathcal{U}^{[m]}\big)\setminus\mathcal{N}_m$ is empty. Assume by contradiction that $\boldsymbol{v} \in \kappa\big(\mathcal{M}\setminus\mathcal{U}^{[m]}\big)\setminus\mathcal{N}_m$. By (A.1) with $k = m$, we have that for all $\boldsymbol{v} \in \kappa\big(\mathcal{M} \setminus \mathcal{U}^{[m-1]}\big) \setminus \mathcal{N}_m$

$$\big|\kappa^{-1}(\boldsymbol{v}) \cap \big(\mathcal{M} \setminus \mathcal{U}^{[m]}\big)\big| \leq m - m = 0$$

i.e., there exists no $\boldsymbol{u} \in \mathcal{M} \setminus \mathcal{U}^{[m]}$ such that $\kappa(\boldsymbol{u}) = \boldsymbol{v}$. This is a contradiction to the assumption $\boldsymbol{v} \in \kappa\big(\mathcal{M} \setminus \mathcal{U}^{[m]}\big) \setminus \mathcal{N}_m$, and thus we conclude that there is no $\boldsymbol{v} \in \kappa\big(\mathcal{M} \setminus \mathcal{U}^{[m]}\big) \setminus \mathcal{N}_m$, i.e., $\kappa\big(\mathcal{M} \setminus \mathcal{U}^{[m]}\big) \setminus \mathcal{N}_m = \emptyset$. Hence, we have

$$\kappa\big(\mathcal{M} \setminus \mathcal{U}^{[m]}\big) \subseteq \mathcal{N}_m \,. \tag{A.6}$$

We next use the integral transformation reported in [Federer, 1969, Th. 3.2.3] to obtain

$$\int_{\mathcal{M}\setminus\mathcal{U}^{[m]}} |\boldsymbol{J}_\kappa(\boldsymbol{u})|^2 \,\mathrm{d}\boldsymbol{u} \leq m \int_{\kappa(\mathcal{M}\setminus\mathcal{U}^{[m]})} \mathrm{d}\boldsymbol{v} \overset{\text{(A.6)}}{\leq} m \int_{\mathcal{N}_m} \mathrm{d}\boldsymbol{v} = 0 \,.$$

Because the function $|\boldsymbol{J}_\kappa(\boldsymbol{u})|$ is positive on $\mathcal{M}$, it follows that the Lebesgue measure of the set $\mathcal{M} \setminus \mathcal{U}^{[m]}$ has to be zero, i.e., $\mathcal{U}^{[m]}$ covers almost all of $\mathcal{M}$. This concludes the proof of part 1.

A.1.2 Proof of Statement 2

To establish statement 2, i.e., the bound (2.22), we first note that

$$h(\mathsf{v}) \geq h(\mathsf{v}\,|\,\mathsf{k}) = \sum_{k\in[1:m]} h(\mathsf{v}\,|\,\mathsf{k}{=}k)\, p_k \tag{A.7}$$

where k is the discrete random variable that takes on the value k when $\mathsf{u} \in \mathcal{U}_k$, and $p_k \triangleq \Pr\{\mathsf{u} \in \mathcal{U}_k\} = \int_{\mathcal{U}_k} f_{\mathsf{u}}(\boldsymbol{u})\, d\boldsymbol{u}$. We assume without loss of generality[1] that $p_k \neq 0$, $k \in [1:m]$. Since $\kappa\big|_{\mathcal{U}_k}$ is one-to-one, we can use the transformation rule for one-to-one mappings (see, e.g., [Morgenshtern et al., 2013, Lemma 3]) to relate $h(\mathsf{v}\,|\,\mathsf{k}{=}k)$ to $h(\mathsf{u}\,|\,\mathsf{k}{=}k)$:

$$h(\mathsf{v}\,|\,\mathsf{k}{=}k) = h(\mathsf{u}\,|\,\mathsf{k}{=}k) + \int_{\mathbb{C}^n} f_{\mathsf{u}|\mathsf{k}=k}(\boldsymbol{u}) \log(|\boldsymbol{J}_\kappa(\boldsymbol{u})|^2) \,\mathrm{d}\boldsymbol{u}. \tag{A.8}$$

[1] If $p_k = 0$ for some k, we simply omit the corresponding term in (A.7).

The conditional probability density function of u given $\mathsf{k}=k$ is $f_{\mathsf{u}|\mathsf{k}=k}(\boldsymbol{u}) = \mathbb{1}_{\mathcal{U}_k}(\boldsymbol{u})\, f_{\mathsf{u}}(\boldsymbol{u})/p_k$. Thus, $h(\mathsf{u}\,|\,\mathsf{k}=k) = -\int_{\mathcal{U}_k}\big(f_{\mathsf{u}}(\boldsymbol{u})/p_k\big)\log\big(f_{\mathsf{u}}(\boldsymbol{u})/p_k\big)\, d\boldsymbol{u}$, and (A.8) becomes

$$
\begin{aligned}
h(\mathsf{v}\,|\,\mathsf{k}=k) &= \frac{1}{p_k}\bigg(-\int_{\mathcal{U}_k} f_{\mathsf{u}}(\boldsymbol{u})\log\bigg(\frac{f_{\mathsf{u}}(\boldsymbol{u})}{p_k}\bigg)\,\mathrm{d}\boldsymbol{u} + \int_{\mathcal{U}_k} f_{\mathsf{u}}(\boldsymbol{u})\log(|\boldsymbol{J}_\kappa(\boldsymbol{u})|^2)\,\mathrm{d}\boldsymbol{u}\bigg) \\
&= \frac{1}{p_k}\bigg(-\int_{\mathcal{U}_k} f_{\mathsf{u}}(\boldsymbol{u})\log f_{\mathsf{u}}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} + \int_{\mathcal{U}_k} f_{\mathsf{u}}(\boldsymbol{u})\log(|\boldsymbol{J}_\kappa(\boldsymbol{u})|^2)\,\mathrm{d}\boldsymbol{u} + p_k\log p_k\bigg)\,.
\end{aligned}
$$

Inserting this expression into (A.7), and recalling that the sets $\mathcal{U}_k$ are disjoint, that $\mathcal{U}^{[m]} = \bigcup_{k\in[1:m]}\mathcal{U}_k$ covers almost all of $\mathcal{M}$, and that $\mathbb{C}^n \setminus \mathcal{M}$ has Lebesgue measure zero, we obtain

$$
\begin{aligned}
h(\mathsf{v}) &\geq \sum_{k\in[1:m]}\bigg(-\int_{\mathcal{U}_k} f_{\mathsf{u}}(\boldsymbol{u})\log f_{\mathsf{u}}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} + \int_{\mathcal{U}_k} f_{\mathsf{u}}(\boldsymbol{u})\log(|\boldsymbol{J}_\kappa(\boldsymbol{u})|^2)\,\mathrm{d}\boldsymbol{u} + p_k\log p_k\bigg) \\
&= -\int_{\mathcal{U}^{[m]}} f_{\mathsf{u}}(\boldsymbol{u})\log f_{\mathsf{u}}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} + \int_{\mathcal{U}^{[m]}} f_{\mathsf{u}}(\boldsymbol{u})\log(|\boldsymbol{J}_\kappa(\boldsymbol{u})|^2)\,\mathrm{d}\boldsymbol{u} + \underbrace{\sum_{k\in[1:m]} p_k\log p_k}_{-H(\mathsf{k})} \\
&= -\int_{\mathbb{C}^n} f_{\mathsf{u}}(\boldsymbol{u})\log f_{\mathsf{u}}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} + \int_{\mathbb{C}^n} f_{\mathsf{u}}(\boldsymbol{u})\log(|\boldsymbol{J}_\kappa(\boldsymbol{u})|^2)\,\mathrm{d}\boldsymbol{u} - H(\mathsf{k}) \\
&= h(\mathsf{u}) + \int_{\mathbb{C}^n} f_{\mathsf{u}}(\boldsymbol{u})\log(|\boldsymbol{J}_\kappa(\boldsymbol{u})|^2)\,\mathrm{d}\boldsymbol{u} - H(\mathsf{k})\,.
\end{aligned}
$$

A.2 Proof of Lemma 2.4

Since f is not identically zero, there exists a $\boldsymbol{\xi}_0 \in \mathbb{C}^n$ such that $f(\boldsymbol{\xi}_0) \neq 0$. The function $g(\boldsymbol{\xi}) \triangleq f(\boldsymbol{\xi} + \boldsymbol{\xi}_0)$ is an analytic function that satisfies $g(\mathbf{0}) \neq 0$. By performing the change of variables $\boldsymbol{\xi} \mapsto \boldsymbol{\xi} + \boldsymbol{\xi}_0$, we can rewrite I_1 in (2.26) in the following more convenient form:

$$
I_1 = \int_{\mathbb{C}^n} \exp(-\|\boldsymbol{\xi} + \boldsymbol{\xi}_0\|^2)\log(|g(\boldsymbol{\xi})|)\,\mathrm{d}\boldsymbol{\xi}\,.
$$

We have

$$
\begin{aligned}
\|\boldsymbol{\xi} + \boldsymbol{\xi}_0\|^2 &\leq (\|\boldsymbol{\xi}\| + \|\boldsymbol{\xi}_0\|)^2 \\
&= \|\boldsymbol{\xi}\|^2 + 2\|\boldsymbol{\xi}\|\|\boldsymbol{\xi}_0\| + \|\boldsymbol{\xi}_0\|^2 \\
&\leq \|\boldsymbol{\xi}\|^2 + 2\max\{\|\boldsymbol{\xi}\|^2, \|\boldsymbol{\xi}_0\|^2\} + \|\boldsymbol{\xi}_0\|^2 \\
&\leq 3\|\boldsymbol{\xi}\|^2 + 3\|\boldsymbol{\xi}_0\|^2\,. \qquad \text{(A.9)}
\end{aligned}
$$

Using (A.9), we lower-bound I_1 as follows:

$$
I_1 \geq c\int_{\mathbb{C}^n} \exp(-3\|\boldsymbol{\xi}\|^2)\log(|g(\boldsymbol{\xi})|)\,\mathrm{d}\boldsymbol{\xi} \triangleq I_2 \qquad \text{(A.10)}
$$

where $c \triangleq \exp(-3\|\boldsymbol{\xi}_0\|^2)$. We next define the mapping $\varphi\colon \mathbb{R}^{2n} \to \mathbb{C}^n$; $\boldsymbol{x} \mapsto \big([\boldsymbol{x}]_{[1:n]} + i[\boldsymbol{x}]_{[n+1:2n]}\big)$, and rewrite I_2 in (A.10) as

$$I_2 = c \int_{\mathbb{R}^{2n}} \exp(-3\|\boldsymbol{x}\|^2)\, u(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \tag{A.11}$$

with $u(\boldsymbol{x}) \triangleq \log(|g(\varphi(\boldsymbol{x}))|)$. Since $g(\mathbf{0}) \neq 0$, we have that $u(\mathbf{0}) > -\infty$. By [Azarin, 2009, Example 2.6.1.3], $u(\boldsymbol{x})$ is a *subharmonic* function. We shall use the following property of subharmonic functions, which is a special case of the more general result reported in [Azarin, 2009, Th. 2.6.2.1].

Lemma A.3 Let u be a subharmonic function on $\mathcal{W} \subseteq \mathbb{R}^{2n}$. If $\{\boldsymbol{x} \in \mathbb{R}^{2n} \colon \|\boldsymbol{x}\| \leq r\} \subseteq \mathcal{W}$ for some $r > 0$, then

$$u(\mathbf{0}) \leq \frac{1}{\omega(2n)\, r^{2n-1}} \int_{\mathcal{S}_r} u(\boldsymbol{x})\, \mathrm{d}\mathscr{H}^{2n-1}(\boldsymbol{x})$$

where $\mathcal{S}_r \triangleq \{\boldsymbol{x} \in \mathbb{R}^{2n} \colon \|\boldsymbol{x}\| = r\}$ and the constant $\omega(2n)$ denotes the area of the unit sphere in $\mathbb{R}^{2n}$.

Using a well-known measure-theoretic result (see, e.g., [Federer, 1969, Th. 3.2.12]), we have for $u(\boldsymbol{x}) = \log(|g(\varphi(\boldsymbol{x}))|)$

$$\int_{\mathbb{R}^{2n}} \exp(-3\|\boldsymbol{x}\|^2)\, u(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = \int_0^\infty \left(\int_{\mathcal{S}_r} u(\boldsymbol{x})\, \mathrm{d}s(\boldsymbol{x}) \right) \exp(-3r^2)\, \mathrm{d}r\,. \tag{A.12}$$

Inserting (A.12) in (A.11), we obtain

$$\begin{aligned} I_2 &= c \int_0^\infty \left(\int_{\mathcal{S}_r} u(\boldsymbol{x})\, \mathrm{d}s(\boldsymbol{x}) \right) \exp(-3r^2)\, \mathrm{d}r \\ &\overset{(a)}{\geq} c\, \sigma_{2n}\, u(\mathbf{0}) \int_0^\infty \exp(-3r^2)\, r^{2n-1}\, \mathrm{d}r \\ &\overset{(b)}{>} -\infty\,. \end{aligned}$$

Here, (a) is due to Lemma A.3 and (b) holds because $u(\mathbf{0}) > -\infty$ and

$$0 < \int_0^\infty \exp(-3r^2)\, r^{2n-1}\, \mathrm{d}r < \infty\,.$$

Using (A.10), we conclude that $I_1 > -\infty$.

A.3 Proof of Lemma 3.6

Since the proof of Lemma 3.6 is quite technical, we shall first (in Section A.3.1) illustrate its key steps by focusing on the special case $\widetilde{T} = 2$, $R = 3$, $N = 4$, and $Q = 1$. The proof for arbitrary $\widetilde{T}$, R, N, and Q will be provided in Section A.3.2.

A.3.1 Special Case $\widetilde{T} = 2$, $R = 3$, $N = 4$, $Q = 1$

By (3.28), we have $\ell = 12$ and, thus, $\mathcal{I} = [1:12]$. Furthermore, we choose $\mathcal{D}^c = \{7, 12\}$. Hence, recalling that $\overline{\boldsymbol{sx}} = (s_{1,1}\ s_{2,1}\ s_{3,1}\ s_{1,2}\ s_{2,2}\ s_{3,2}\ \boldsymbol{x}_1^{\mathrm{T}}\ \boldsymbol{x}_2^{\mathrm{T}})^{\mathrm{T}}$, we have

$$[\overline{\boldsymbol{sx}}]_{\mathcal{D}^c} = ([\boldsymbol{x}_1]_1\ [\boldsymbol{x}_2]_2)^{\mathrm{T}}$$

and

$$[\overline{\boldsymbol{sx}}]_{\mathcal{D}} = (s_{1,1}\ s_{2,1}\ s_{3,1}\ s_{1,2}\ s_{2,2}\ s_{3,2}\ [\boldsymbol{x}_1]_2\ [\boldsymbol{x}_1]_3\ [\boldsymbol{x}_1]_4\ [\boldsymbol{x}_2]_1\ [\boldsymbol{x}_2]_3\ [\boldsymbol{x}_2]_4)^{\mathrm{T}}\,.$$

We also choose $\boldsymbol{x}$ as the all-one vector. For these choices, the Jacobian $[\boldsymbol{J}_\phi(\boldsymbol{s}, \boldsymbol{x})]_{\mathcal{I}}^{\mathcal{D}}$ in (3.32) is equal to the following matrix

$$\begin{pmatrix}
[\boldsymbol{Z}_{1,1}]_1 & [\boldsymbol{Z}_{1,2}]_1 & & & & & 0 & & & [\boldsymbol{Z}_{1,2}]_1 s_{1,2} & & \\
[\boldsymbol{Z}_{1,1}]_2 & [\boldsymbol{Z}_{1,2}]_2 & & & & & [\boldsymbol{Z}_{1,1}]_2 s_{1,1} & & & & 0 & \\
[\boldsymbol{Z}_{1,1}]_3 & [\boldsymbol{Z}_{1,2}]_3 & & & & & & [\boldsymbol{Z}_{1,1}]_3 s_{1,1} & & & [\boldsymbol{Z}_{1,2}]_3 s_{1,2} & \\
[\boldsymbol{Z}_{1,1}]_4 & [\boldsymbol{Z}_{1,2}]_4 & & & & & & & [\boldsymbol{Z}_{1,1}]_4 s_{1,1} & & & [\boldsymbol{Z}_{1,2}]_4 s_{1,2} \\
& & [\boldsymbol{Z}_{2,1}]_1 & [\boldsymbol{Z}_{2,2}]_1 & & & 0 & & & [\boldsymbol{Z}_{2,2}]_1 s_{2,2} & & \\
& & [\boldsymbol{Z}_{2,1}]_2 & [\boldsymbol{Z}_{2,2}]_2 & & & [\boldsymbol{Z}_{2,1}]_2 s_{2,1} & & & & 0 & \\
& & [\boldsymbol{Z}_{2,1}]_3 & [\boldsymbol{Z}_{2,2}]_3 & & & & [\boldsymbol{Z}_{2,1}]_3 s_{2,1} & & & [\boldsymbol{Z}_{2,2}]_3 s_{2,2} & \\
& & [\boldsymbol{Z}_{2,1}]_4 & [\boldsymbol{Z}_{2,2}]_4 & & & & & [\boldsymbol{Z}_{2,1}]_4 s_{2,1} & & & [\boldsymbol{Z}_{2,2}]_4 s_{2,2} \\
& & & & [\boldsymbol{Z}_{3,1}]_1 & \colorbox{lightgray}{$[\boldsymbol{Z}_{3,2}]_1$} & 0 & & & \colorbox{lightgray}{$[\boldsymbol{Z}_{3,2}]_1 s_{3,2}$} & & \\
& & & & \colorbox{lightgray}{$[\boldsymbol{Z}_{3,1}]_2$} & [\boldsymbol{Z}_{3,2}]_2 & \colorbox{lightgray}{$[\boldsymbol{Z}_{3,1}]_2 s_{3,1}$} & & & & 0 & \\
& & & & [\boldsymbol{Z}_{3,1}]_3 & \colorbox{lightgray}{$[\boldsymbol{Z}_{3,2}]_3$} & & [\boldsymbol{Z}_{3,1}]_3 s_{3,1} & & & \colorbox{lightgray}{$[\boldsymbol{Z}_{3,2}]_3 s_{3,2}$} & \\
& & & & \colorbox{lightgray}{$[\boldsymbol{Z}_{3,1}]_4$} & [\boldsymbol{Z}_{3,2}]_4 & & & \colorbox{lightgray}{$[\boldsymbol{Z}_{3,1}]_4 s_{3,1}$} & & & [\boldsymbol{Z}_{3,2}]_4 s_{3,2}
\end{pmatrix}. \tag{A.13}$$

We have to find $\boldsymbol{Z}$ and $\boldsymbol{s}$ such that the determinant of this matrix is nonzero. Setting $[\boldsymbol{Z}_{3,2}]_1 = [\boldsymbol{Z}_{3,1}]_2 = [\boldsymbol{Z}_{3,2}]_3 = [\boldsymbol{Z}_{3,1}]_4 = 0$, the entries highlighted in gray in (A.13) become zero. Furthermore, choosing nonzero $[\boldsymbol{Z}_{3,1}]_1$, $[\boldsymbol{Z}_{3,1}]_3$, $[\boldsymbol{Z}_{3,2}]_2$, $[\boldsymbol{Z}_{3,2}]_4$, $s_{3,1}$, and $s_{3,2}$ and operating a Laplace expansion on the last four rows in (A.13), it is seen that the determinant of the matrix in (A.13) is nonzero if and only if the determinant of the following matrix is nonzero:

$$\begin{pmatrix}
[\boldsymbol{Z}_{1,1}]_1 & [\boldsymbol{Z}_{1,2}]_1 & & & 0 & & \colorbox{lightgray}{$[\boldsymbol{Z}_{1,2}]_1 s_{1,2}$} & \\
[\boldsymbol{Z}_{1,1}]_2 & [\boldsymbol{Z}_{1,2}]_2 & & & [\boldsymbol{Z}_{1,1}]_2 s_{1,1} & & & 0 \\
[\boldsymbol{Z}_{1,1}]_3 & [\boldsymbol{Z}_{1,2}]_3 & & & & 0 & & \colorbox{lightgray}{$[\boldsymbol{Z}_{1,2}]_3 s_{1,2}$} \\
[\boldsymbol{Z}_{1,1}]_4 & [\boldsymbol{Z}_{1,2}]_4 & & & & [\boldsymbol{Z}_{1,1}]_4 s_{1,1} & & 0 \\
& & [\boldsymbol{Z}_{2,1}]_1 & [\boldsymbol{Z}_{2,2}]_1 & 0 & & [\boldsymbol{Z}_{2,2}]_1 s_{2,2} & \\
& & [\boldsymbol{Z}_{2,1}]_2 & [\boldsymbol{Z}_{2,2}]_2 & \colorbox{lightgray}{$[\boldsymbol{Z}_{2,1}]_2 s_{2,1}$} & & & 0 \\
& & [\boldsymbol{Z}_{2,1}]_3 & [\boldsymbol{Z}_{2,2}]_3 & & 0 & & [\boldsymbol{Z}_{2,2}]_3 s_{2,2} \\
& & [\boldsymbol{Z}_{2,1}]_4 & [\boldsymbol{Z}_{2,2}]_4 & & \colorbox{lightgray}{$[\boldsymbol{Z}_{2,1}]_4 s_{2,1}$} & & 0
\end{pmatrix}. \tag{A.14}$$

The matrix in (A.14) is the Jacobian matrix corresponding to the case $\widetilde{T} = 2, R = 2, N = 4$, and $Q = 1$. In other words, by performing the matrix manipulations just described, we reduced the case $R = 3$ to the case $R = 2$. A similar idea will be used in the proof for the general case provided in Section A.3.2, where we will reduce R inductively until $R = \widetilde{T}$. Setting $s_{1,2} = s_{2,1} = 0$, the entries highlighted in gray in (A.14) become zero. By choosing nonzero $[\boldsymbol{Z}_{1,1}]_2, [\boldsymbol{Z}_{1,1}]_4, [\boldsymbol{Z}_{2,2}]_1, [\boldsymbol{Z}_{2,2}]_3, s_{1,1}$, and $s_{2,2}$ and operating a Laplace expansion on the last four

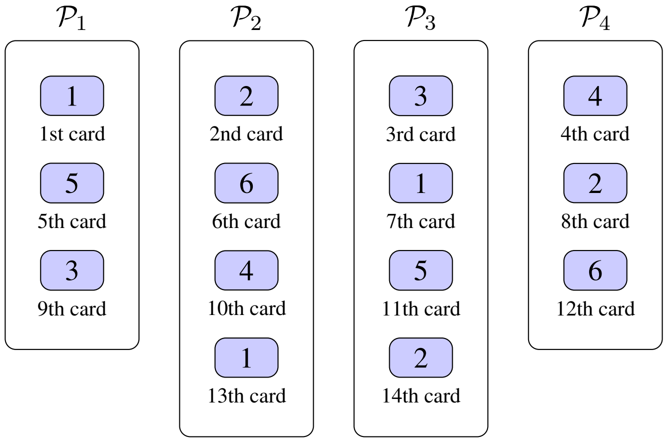

Figure A.1: Construction of the sets $\mathcal{P}_t$ for $\widetilde{T} = 4$, $N = 6$, and $\vartheta_R = 14$.

columns, it is seen that it is sufficient to show that the determinant of the following matrix is nonzero:

$$\begin{pmatrix} [\boldsymbol{Z}_{1,1}]_1 & [\boldsymbol{Z}_{1,2}]_1 & & \\ [\boldsymbol{Z}_{1,1}]_3 & [\boldsymbol{Z}_{1,2}]_3 & & \\ & & [\boldsymbol{Z}_{2,1}]_2 & [\boldsymbol{Z}_{2,2}]_2 \\ & & [\boldsymbol{Z}_{2,1}]_4 & [\boldsymbol{Z}_{2,2}]_4 \end{pmatrix}.$$

This can be achieved, e.g., by setting all off-diagonal entries (i.e., $[\boldsymbol{Z}_{1,1}]_3$, $[\boldsymbol{Z}_{1,2}]_1$, $[\boldsymbol{Z}_{2,1}]_4$, and $[\boldsymbol{Z}_{2,2}]_2$) to zero and choosing all diagonal entries (i.e., $[\boldsymbol{Z}_{1,1}]_1$, $[\boldsymbol{Z}_{1,2}]_3$, $[\boldsymbol{Z}_{2,1}]_2$, and $[\boldsymbol{Z}_{2,2}]_4$) nonzero.

A.3.2 Proof for the General Case

We have to find $\boldsymbol{Z}$, $\boldsymbol{s}$, $\boldsymbol{x}$, and $\mathcal{D}$ such that $\left|[\boldsymbol{J}_\phi(\boldsymbol{s},\boldsymbol{x})]_{\mathcal{I}}^{\mathcal{D}}\right| \neq 0$.

Construction of $\mathcal{D}$

We start by constructing the set $\mathcal{D}$. As in the special case above the set $\mathcal{D}$ includes all indices pointing to variables in $\boldsymbol{s}$ (i.e., $[1 : R\widetilde{T}Q] \subseteq \mathcal{D}$). For the indices of $\boldsymbol{x}$ we will define sets $\mathcal{P}_t \subseteq [1 : N]$ that specify the indices of every $\boldsymbol{x}_t$ not belonging to $\mathcal{D}$ (we can think of $\mathcal{P}_t$ as specifying the positions of the pilot symbols at transmit antenna t), i.e.,

$$\mathcal{D}^c = \{R\widetilde{T}Q + i + (t-1)N : i \in \mathcal{P}_t, t \in [1:\widetilde{T}]\}. \tag{A.15}$$

Because we need $|\mathcal{D}| = \ell$, the sets $\{\mathcal{P}_t\}_{t\in[1:\widetilde{T}]}$ have to satisfy

$$\begin{aligned} \sum_{t\in[1:\widetilde{T}]} |\mathcal{P}_t| &= |\mathcal{D}^c| \\ &= R\widetilde{T}Q + \widetilde{T}N - \ell \\ &\overset{(3.28)}{=} R\widetilde{T}Q + \widetilde{T}N - \min\{RN, R\widetilde{T}Q + \widetilde{T}N - \widetilde{T}\} \\ &= \max\{\widetilde{T}, R\widetilde{T}Q - (R-\widetilde{T})N\} \\ &\triangleq \vartheta_R. \end{aligned} \tag{A.16}$$

(We use the subscript R in ϑ_R because the dependence on R will be important later.) To provide intuition about our choice of the sets $\mathcal{P}_t$, we use a card game metaphor. Consider a deck of $\widetilde{T}N$ cards showing numbers from 1 to N sorted as follows: $1, 2, \ldots, N, \ldots, 1, 2, \ldots, N$ (i.e., the sequence $1, 2, \ldots, N$ repeated $\widetilde{T}$ times). The idea is to choose the ϑ_R positions of the pilot symbols by assigning the indices $i \in [1\!:\!N]$ to the sets $\mathcal{P}_t$ in the same way as the first ϑ_R cards are distributed to $\widetilde{T}$ players (in Fig. A.1, we give an example of the algorithm for $\vartheta_R = 14$, $N = 6$, and $\widetilde{T} = 4$): The first card shows 1 and goes to $\mathcal{P}_1$, i.e., $1 \in \mathcal{P}_1$, and in the same way we proceed with $2 \in \mathcal{P}_2, \ldots, \widetilde{T} \in \mathcal{P}_{\widetilde{T}}$ (this corresponds to the 1st to 4th card in Fig. A.1). When we run out of sets (players), we start with the first set (player) again: $\widetilde{T}+1 \in \mathcal{P}_1$, $\widetilde{T}+2 \in \mathcal{P}_2$, etc. After the card showing index N (recall that $\mathcal{P}_t \subseteq [1\!:\!N]$), the next card starts with index 1 again (in Fig. A.1, the 6th card shows $N = 6$ and goes to $\mathcal{P}_2$ and the 7th card shows 1 and goes to $\mathcal{P}_3$). This scheme works as long as we avoid assigning an index to a set $\mathcal{P}_t$ to which that index was already assigned in a previous round. (In Fig. A.1, this would happen after the 12th card. The 13th card shows 1 and the algorithm would set $1 \in \mathcal{P}_1$, which was already assigned to $\mathcal{P}_1$ in the first round.) To avoid this issue, we introduce an offset and skip one set (resulting in the 13th card going to $\mathcal{P}_2$ in Fig. A.1) and proceed as before. The algorithm stops when ϑ_R indices (cards) have been assigned to the sets (players) $\mathcal{P}_t$.

We now present a mathematical formulation of the algorithm we just outlined. Let the function $\boldsymbol{\beta}\colon [1\!:\!\widetilde{T}N] \to [1\!:\!\widetilde{T}] \times [1\!:\!N]$ be defined as

$$\boldsymbol{\beta}(j) = \begin{pmatrix} \beta_1(j) \\ \beta_2(j) \end{pmatrix} \triangleq \begin{pmatrix} \left(j + \left\lfloor \frac{j-1}{\mathrm{lcm}(\widetilde{T},N)} \right\rfloor\right) \bmod^* \widetilde{T} \\ j \bmod^* N \end{pmatrix}, \quad j \in [1\!:\!\widetilde{T}N]\,. \tag{A.17}$$

Here $\mathrm{lcm}(\cdot,\cdot)$ denotes the least common multiple and $a \bmod^* b \triangleq a - b\lfloor (a-1)/b \rfloor$ denotes the residuum of a divided by b in $[1:b]$ (and not in $[0:b-1]$ as commonly done). We use the function $\boldsymbol{\beta}$ to assign up to $\widetilde{T}N$ elements (note that $\vartheta_R \le \widetilde{T}N$) to the sets $\mathcal{P}_t$ as follows: for $j \in [1\!:\!\vartheta_R]$, the function $\beta_1(j)$ specifies $t \in [1\!:\!\widetilde{T}]$ (equivalently, one of the sets $\mathcal{P}_t$, $t \in [1\!:\!\widetilde{T}]$), and the function $\beta_2(j)$ specifies the index $i \in [1\!:\!N]$ that is assigned to $\mathcal{P}_t$ (again invoking our card game metaphor, the jth card shows the index $\beta_2(j)$ and is assigned to player $\mathcal{P}_{\beta_1(j)}$). Using $\beta_1(j)$ and $\beta_2(j)$, we can compactly describe each set $\mathcal{P}_t$ as follows:

$$\mathcal{P}_t \triangleq \beta_2\big(\beta_1^{-1}(t) \cap [1\!:\!\vartheta_R]\big), \quad t \in [1\!:\!\widetilde{T}]. \tag{A.18}$$

Here, the set $\beta_1^{-1}(t)$ consists of all values $j \in [1\!:\!\widetilde{T}N]$ that correspond to an assignment of an index i to the set $\mathcal{P}_t$. Since we only want to assign a total of ϑ_R indices, we take the intersection with $[1\!:\!\vartheta_R]$. For each $j \in \beta_1^{-1}(t) \cap [1\!:\!\vartheta_R]$, the function β_2 now chooses an index $i \in [1\!:\!N]$, and we obtain the definition (A.18).

The sets $\mathcal{P}_t$ in (A.18) satisfy the properties listed in the following lemma.

Lemma A.4 Suppose that $R \ge \widetilde{T}$, $N > \widetilde{T}Q$, and $R \le \lceil \widetilde{T}(N-1)/(N-\widetilde{T}Q) \rceil$. Let the sets $\{\mathcal{P}_t\}_{t\in[1:\widetilde{T}]}$ be defined as in (A.18). Then the following properties hold:

(i) $\sum_{t\in[1:\widetilde{T}]}|\mathcal{P}_t| = \vartheta_R$;

(ii) $|\mathcal{P}_t| \le \widetilde{T}Q$;

If $R > \widetilde{T}$, let $\{\widetilde{\mathcal{P}}_t\}_{t\in[1:\widetilde{T}]}$ be the corresponding sets for the case of $R-1$ receive antennas, i.e.,

$$\widetilde{\mathcal{P}}_t \triangleq \beta_2\big(\beta_1^{-1}(t) \cap [1\!:\!\vartheta_{R-1}]\big)\,. \tag{A.19}$$

Furthermore, we set

$$\mathcal{L}_t \triangleq \widetilde{\mathcal{P}}_t \setminus \mathcal{P}_t \tag{A.20}$$

and

$$\widetilde{\mathcal{L}} \triangleq \bigcup_{t\in[1:\widetilde{T}]} \mathcal{L}_t\,. \tag{A.21}$$

Then the following properties hold:

(iii) $\mathcal{L}_t \cap \mathcal{L}_{t'} = \emptyset$ for $t \neq t'$;

(iv) $\mathcal{L}_t \subseteq [1\!:\!N-\gamma]$, where

$$\gamma \triangleq RN - \ell\,; \tag{A.22}$$

(v) There exist sets $\mathcal{G}_t \subseteq [1\!:\!N-\gamma]$, $t \in [1\!:\!\widetilde{T}]$ satisfying

a) $|\mathcal{G}_t| = Q$,

b) $\mathcal{G}_t \cap \mathcal{G}_{t'} = \emptyset$ for $t \neq t'$,

c) $\mathcal{G}_t \cap \mathcal{P}_t \neq \emptyset$,

d) $\bigcup_{t\in[1:\widetilde{T}]} \mathcal{G}_t = \mathcal{G} \triangleq [1\!:\!N-\gamma] \setminus \widetilde{\mathcal{L}}$.

Proof. See Appendix A.4. □

Remark A.1 Property (i) states that the sets $\mathcal{P}_t$ have the correct size (see (A.16)). Properties (iii), (iv), and (v) state that we can partition the set $[1:N-\gamma]$ into $2\widetilde{T}$ disjoint sets $\mathcal{L}_t$ and $\mathcal{G}_{t'}$, $t,t' \in [1\!:\!\widetilde{T}]$ (i.e., $\mathcal{G}_t \cap \mathcal{G}_{t'} = \emptyset$, $\mathcal{G}_t \cap \mathcal{L}_{t'} = \emptyset$, and $\mathcal{L}_t \cap \mathcal{L}_{t'} = \emptyset$ for $t \neq t'$) such that in each $\mathcal{G}_{t'}$ there is a point $g_{t'} \in \mathcal{P}_{t'}$.

Construction of $\boldsymbol{Z}$, $\boldsymbol{s}$, and $\boldsymbol{x}$

It remains to find a triple $(\boldsymbol{Z}, \boldsymbol{s}, \boldsymbol{x})$ such that $\det\big([\boldsymbol{J}_\phi(\boldsymbol{s},\boldsymbol{x})]_{\mathcal{I}}^{\mathcal{D}}\big)$ is nonzero (with the choice of $\mathcal{D}$ described above). This will be done by an induction argument over $R \ge \widetilde{T}$. For this purpose, it is convenient to define the sets

$$\mathcal{D}_t \triangleq [1\!:\!N] \setminus \mathcal{P}_t\,. \tag{A.23}$$

Note that by (A.15) and because $\mathcal{D} = [1\!:\!R\widetilde{T}Q + \widetilde{T}N] \setminus \mathcal{D}^c$, we have that

$$\begin{aligned}\mathcal{D} &= [1\!:\!R\widetilde{T}Q + \widetilde{T}N] \setminus \mathcal{D}^c \\ &= [1\!:\!R\widetilde{T}Q + \widetilde{T}N] \setminus \{R\widetilde{T}Q + i + (t-1)N : i \in \mathcal{P}_t, t \in [1\!:\!\widetilde{T}]\}\end{aligned}$$

$$= [1\!:\!R\widetilde{T}Q] \cup \{R\widetilde{T}Q + i + (t-1)N : i \in \mathcal{D}_t, t \in [1\!:\!\widetilde{T}]\} \tag{A.24}$$

i.e., $\mathcal{D}_t \subseteq [1\!:\!N]$ specifies the positions of the symbols in the vector $\boldsymbol{x}_t$ specified by $\mathcal{D}$. We will make repeated use of the next result, which follows from [Horn and Johnson, 1985, Sec. 0.8.5].

Lemma A.5 Let $\boldsymbol{M} \in \mathbb{C}^{n\times n}$, and let $\mathcal{E}, \mathcal{F} \subseteq [1:n]$ with $|\mathcal{E}| = |\mathcal{F}|$. If $[\boldsymbol{M}]_{[1:n]\setminus\mathcal{E}}^{\mathcal{F}} = \mathbf{0}$ or $[\boldsymbol{M}]_{\mathcal{E}}^{[1:n]\setminus\mathcal{F}} = \mathbf{0}$, and if $[\boldsymbol{M}]_{\mathcal{E}}^{\mathcal{F}}$ is nonsingular, then $\det(\boldsymbol{M}) \neq 0$ if and only if $\det\Big([\boldsymbol{M}]_{[1:n]\setminus\mathcal{E}}^{[1:n]\setminus\mathcal{F}}\Big) \neq 0$.

Remark A.2 Lemma A.5 is just an abstract way to describe a situation where given a matrix $\boldsymbol{M}$, one is able to perform row and column interchanges that yield a new matrix of the form $\left(\begin{smallmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \mathbf{0} & \boldsymbol{C}\end{smallmatrix}\right)$, where $\boldsymbol{A}$ and $\boldsymbol{C}$ are square matrices. In this case, a basic result in linear algebra states that the determinant of $\boldsymbol{M}$ equals the product of the determinants of $\boldsymbol{A}$ and $\boldsymbol{C}$, and hence, assuming that $\boldsymbol{C}$ is nonsingular, $\det(\boldsymbol{M}) \neq 0$ if and only if $\det(\boldsymbol{A}) \neq 0$.

We will now present the inductive construction of $\boldsymbol{Z}$, $\boldsymbol{s}$, and $\boldsymbol{x}$.

Induction hypothesis: For $\widetilde{T} \leq R \leq \lceil \widetilde{T}(N-1)/(N-\widetilde{T}Q)\rceil$, $\widetilde{T}Q < N$ (as assumed throughout the proof), and $\{\mathcal{P}_t\}_{t\in[1:\widetilde{T}]}$ as in (A.18), there exists a triple $(\boldsymbol{Z}, \boldsymbol{s}, \boldsymbol{x})$ with $\boldsymbol{x} = (1\cdots 1)^{\mathrm{T}}$ such that $\det\big([\boldsymbol{J}_\phi(\boldsymbol{s},\boldsymbol{x})]_{\mathcal{I}}^{\mathcal{D}}\big)$ is nonzero.

Base case (proof for $R = \widetilde{T}$): When $R = \widetilde{T}$, (A.16) reduces to $\sum_{t\in[1:\widetilde{T}]}|\mathcal{P}_t| = \widetilde{T}^2 Q$. Using Property (ii) in Lemma A.4, this implies that $|\mathcal{P}_t| = \widetilde{T}Q$. Furthermore, $\ell = RN$ (see (3.28)), resulting in $\mathcal{I} = [1:RN]$. To establish the desired result, we first choose $\boldsymbol{s}_{r,t} = \mathbf{0}$ for $r \neq t$. With this choice, the matrix $[\boldsymbol{J}_\phi(\boldsymbol{s},\boldsymbol{x})]_{\mathcal{I}}^{\mathcal{D}}$ in (3.32) looks as follows:

$$[\boldsymbol{J}_\phi(\boldsymbol{s},\boldsymbol{x})]_{\mathcal{I}}^{\mathcal{D}} = \left[\left(\boldsymbol{B}\begin{pmatrix}\boldsymbol{A}_{1,1} & \cdots & \boldsymbol{A}_{1,\widetilde{T}}\\ \vdots & & \vdots\\ \boldsymbol{A}_{\widetilde{T},1} & \cdots & \boldsymbol{A}_{\widetilde{T},\widetilde{T}}\end{pmatrix}\right)\right]_{\mathcal{I}}^{\mathcal{D}} = \begin{pmatrix}\boldsymbol{B}_1 & & & [\boldsymbol{A}_{1,1}]^{\mathcal{D}_1} & & \\ & \ddots & & & \ddots & \\ & & \boldsymbol{B}_{\widetilde{T}} & & & [\boldsymbol{A}_{\widetilde{T},\widetilde{T}}]^{\mathcal{D}_{\widetilde{T}}}\end{pmatrix} \tag{A.25}$$

where we used the sets $\{\mathcal{D}_t\}_{t\in[1:\widetilde{T}]}$ given in (A.23), and where (cf. (3.31)) $\boldsymbol{B}_r = (\boldsymbol{Z}_{r,1}\cdots \boldsymbol{Z}_{r,\widetilde{T}})$, for $r \in [1\!:\!\widetilde{T}]$ and (cf. (3.33))

$$\boldsymbol{A}_{t,t} = \operatorname{diag}(\boldsymbol{a}_{t,t}), \quad t \in [1\!:\!\widetilde{T}], \quad \text{with } \boldsymbol{a}_{t,t} \triangleq \boldsymbol{Z}_{t,t}\boldsymbol{s}_{t,t}\,. \tag{A.26}$$

We choose[2] $[\boldsymbol{Z}_{r,t}]_{\mathcal{P}_r} \in \mathbb{C}^{\widetilde{T}Q\times Q}$ such that the square matrices $[\boldsymbol{B}_r]_{\mathcal{P}_r} = \big[\big(\boldsymbol{Z}_{r,1}\ \cdots\ \boldsymbol{Z}_{r,\widetilde{T}}\big)\big]_{\mathcal{P}_r} \in \mathbb{C}^{\widetilde{T}Q\times\widetilde{T}Q}$ are nonsingular. Furthermore, we have that $[\boldsymbol{A}_{t,t}]_{\mathcal{P}_t}^{\mathcal{D}_t} = \mathbf{0}$ (by (A.26), $\boldsymbol{A}_{t,t}$ is a diagonal matrix, and because $\mathcal{P}_t \cap \mathcal{D}_t = \emptyset$, the matrix $[\boldsymbol{A}_{t,t}]_{\mathcal{P}_t}^{\mathcal{D}_t}$ contains only off-diagonal entries). We will use Lemma A.5 with $\boldsymbol{M} = [\boldsymbol{J}_\phi(\boldsymbol{s},\boldsymbol{x})]_{\mathcal{I}}^{\mathcal{D}}$ given by (A.25), $n = \widetilde{T}^2Q + |\mathcal{D}|$, $\mathcal{E} = \mathcal{P}$ (i.e., the rows where $[\boldsymbol{A}_{t,t}]^{\mathcal{D}_t}$ is zero), and $\mathcal{F} = [1:\widetilde{T}^2Q]$ (i.e., the columns of all $\boldsymbol{B}_r$, $r \in [1\!:\!\widetilde{T}]$). This choice yields $[\boldsymbol{M}]_{\mathcal{E}}^{\mathcal{F}} = \operatorname{diag}\big([\boldsymbol{B}_1]_{\mathcal{P}_1}, \ldots, [\boldsymbol{B}_{\widetilde{T}}]_{\mathcal{P}_{\widetilde{T}}}\big)$, which is nonsingular because it is a block diagonal

[2] Note that so far we used the index t in the sets $\mathcal{P}_t$. Now we consider the matrix $[\boldsymbol{B}_r]_{\mathcal{P}_t}$ for $t = r$. Thus, it is convenient to use only the index r.

matrix where each block on the diagonal, $[\boldsymbol{B}_r]_{\mathcal{P}_r}$, was chosen nonsingular. Furthermore, we have that $[\boldsymbol{M}]_{\mathcal{E}}^{[1:n]\setminus\mathcal{F}} = \mathrm{diag}\left([\boldsymbol{A}_{1,1}]_{\mathcal{P}_1}^{\mathcal{D}_1}, \dots, [\boldsymbol{A}_{\widetilde{T},\widetilde{T}}]_{\mathcal{P}_{\widetilde{T}}}^{\mathcal{D}_{\widetilde{T}}}\right) = \mathbf{0}$. Thus, the requirements of Lemma A.5 are met and, hence, $\det(\boldsymbol{M}) = \det\left([\boldsymbol{J}_\phi(\boldsymbol{s},\boldsymbol{x})]_{\mathcal{I}}^{\mathcal{D}}\right) \neq 0$ if and only if the determinant of the following matrix is nonzero:

$$[\boldsymbol{M}]_{[1:n]\setminus\mathcal{E}}^{[1:n]\setminus\mathcal{F}} = \begin{pmatrix} [\boldsymbol{A}_{1,1}]_{\mathcal{D}_1}^{\mathcal{D}_1} & & \\ & \ddots & \\ & & [\boldsymbol{A}_{\widetilde{T},\widetilde{T}}]_{\mathcal{D}_{\widetilde{T}}}^{\mathcal{D}_{\widetilde{T}}} \end{pmatrix}. \tag{A.27}$$

Because of (A.26), we have $[\boldsymbol{A}_{t,t}]_{\mathcal{D}_t}^{\mathcal{D}_t} = [\mathrm{diag}(\boldsymbol{a}_{t,t})]_{\mathcal{D}_t}^{\mathcal{D}_t}$. Hence, the matrix in (A.27) is a diagonal matrix and can be chosen to have nonzero diagonal entries by choosing $[\boldsymbol{Z}_{t,t}]_{\mathcal{D}_t}$ and $\boldsymbol{s}_{t,t}$ such that $[\boldsymbol{a}_{t,t}]_i = [\boldsymbol{Z}_{t,t}]_{\{i\}}\boldsymbol{s}_{t,t} \neq 0$ for all $i \in \mathcal{D}_t$ (again see (A.26)). Thus, its determinant is nonzero and, in turn, $\det(\boldsymbol{M}) \neq 0$.

Inductive step (transition from $R-1$ to R): Assuming that $\boldsymbol{Z}_{r,t}$ and $\boldsymbol{s}_{r,t}$ for $t \in [1:\widetilde{T}]$, $r \in [1:R-1]$ have already been chosen such that the determinant of $[\boldsymbol{J}_\phi(\boldsymbol{s},\boldsymbol{x})]_{\mathcal{I}}^{\mathcal{D}}$ is nonzero in the $R-1$ setting, we want to show that there exist $\boldsymbol{Z}_{R,t}$ and $\boldsymbol{s}_{R,t}$, $t \in [1:\widetilde{T}]$ for which the determinant of the matrix $[\boldsymbol{J}_\phi(\boldsymbol{s},\boldsymbol{x})]_{\mathcal{I}}^{\mathcal{D}}$ in (3.32) is nonzero. To facilitate the exposition, we rewrite the matrices involved in a more convenient form. For the case of R receive antennas, denoted by the superscript $[R]$, we rewrite the Jacobian matrix $[\boldsymbol{J}_\phi(\boldsymbol{s},\boldsymbol{x})]_{\mathcal{I}}^{\mathcal{D}}$ in (3.32) as

$$[\boldsymbol{J}_\phi^{[R]}(\boldsymbol{s},\boldsymbol{x})]_{\mathcal{I}}^{\mathcal{D}} = \begin{pmatrix} \boldsymbol{B}_1 & & & [\boldsymbol{A}_{1,1}]^{\mathcal{D}_1} & \cdots & [\boldsymbol{A}_{1,\widetilde{T}}]^{\mathcal{D}_{\widetilde{T}}} \\ & \ddots & & \vdots & & \vdots \\ & & \boldsymbol{B}_{R-1} & [\boldsymbol{A}_{R-1,1}]^{\mathcal{D}_1} & \cdots & [\boldsymbol{A}_{R-1,\widetilde{T}}]^{\mathcal{D}_{\widetilde{T}}} \\ & & [\boldsymbol{B}_R]_{[1:N-\gamma]} & [\boldsymbol{A}_{R,1}]_{[1:N-\gamma]}^{\mathcal{D}_1} & \cdots & [\boldsymbol{A}_{R,\widetilde{T}}]_{[1:N-\gamma]}^{\mathcal{D}_{\widetilde{T}}} \end{pmatrix} \tag{A.28}$$

where we used (A.24) and that $\mathcal{I} = [1:RN-\gamma]$. For the $R-1$ case, the Jacobian matrix is given by

$$[\boldsymbol{J}_\phi^{[R-1]}(\boldsymbol{s},\boldsymbol{x})]_{\widetilde{\mathcal{I}}}^{\widetilde{\mathcal{D}}} = \begin{pmatrix} \boldsymbol{B}_1 & & [\boldsymbol{A}_{1,1}]^{\widetilde{\mathcal{D}}_1} & \cdots & [\boldsymbol{A}_{1,\widetilde{T}}]^{\widetilde{\mathcal{D}}_{\widetilde{T}}} \\ & \ddots & \vdots & & \vdots \\ & \boldsymbol{B}_{R-1} & [\boldsymbol{A}_{R-1,1}]^{\widetilde{\mathcal{D}}_1} & \cdots & [\boldsymbol{A}_{R-1,\widetilde{T}}]^{\widetilde{\mathcal{D}}_{\widetilde{T}}} \end{pmatrix} \tag{A.29}$$

where

$$\widetilde{\mathcal{D}}_t \triangleq [1:N] \setminus \widetilde{\mathcal{P}}_t \tag{A.30}$$

with the sets $\widetilde{\mathcal{P}}_t$ introduced in Lemma A.4. Note that in (A.29), we do not need to truncate the matrix when selecting the rows in the set $\widetilde{\mathcal{I}}$ as required by (3.32). This follows because $\ell = (R-1)N$ for $R-1 \leq \widetilde{T}(N-1)/(N-\widetilde{T}Q)$ (which holds because $R \leq \lceil \widetilde{T}(N-1)/(N-\widetilde{T}Q) \rceil$) and, hence, $\widetilde{\mathcal{I}} = [1:(R-1)N]$.

Let $\mathcal{G}$, $\mathcal{G}_t$, and $\mathcal{L}_t$ be defined as in Lemma A.4. Set $[\boldsymbol{Z}_{R,t}]_{\mathcal{G}\setminus\mathcal{G}_t} = \mathbf{0}$ for all $t \in [1:\widetilde{T}]$, and choose $[\boldsymbol{Z}_{R,t}]_{\mathcal{G}_t} \in \mathbb{C}^{Q\times Q}$ nonsingular for all $t \in [1:\widetilde{T}]$. With these choices, and recalling that we set $\boldsymbol{x} = (1 \cdots 1)^{\mathrm{T}}$ in the induction hypothesis (whence $\boldsymbol{X}_t = \mathbf{I}_N$), it follows from (3.31) that $[\boldsymbol{B}_R]_{\mathcal{G}} = \big([\boldsymbol{Z}_{R,1}]_{\mathcal{G}} \cdots [\boldsymbol{Z}_{R,\widetilde{T}}]_{\mathcal{G}}\big)$ is nonsingular. Next, for each $t \in [1:\widetilde{T}]$, select an index g_t in the set $\mathcal{G}_t \cap \mathcal{P}_t$ (note that this set is non-empty due to Property (v-c) in Lemma A.4). Furthermore, choose $\boldsymbol{s}_{R,t}$ to be orthogonal to the rows of $[\boldsymbol{Z}_{R,t}]_{\mathcal{G}_t \setminus \{g_t\}} \in \mathbb{C}^{(Q-1)\times Q}$ and to satisfy $[\boldsymbol{Z}_{R,t}]_{\{g_t\}} \boldsymbol{s}_{R,t} \neq 0$ (note that since $\boldsymbol{s}_{r,t} \in \mathbb{C}^Q$, it is always possible to choose $\boldsymbol{s}_{r,t}$ such that it is orthogonal to $Q-1$ vectors of a set of Q linearly independent vectors and not orthogonal to the last one). Recalling (3.33), we have

$$[\boldsymbol{A}_{R,t}]_{\mathcal{G}} = \big[\operatorname{diag}(\boldsymbol{a}_{R,t})\big]_{\mathcal{G}}, \quad t \in [1:\widetilde{T}]$$

where $[\boldsymbol{a}_{R,t}]_i = [\boldsymbol{Z}_{R,t}]_{\{i\}} \boldsymbol{s}_{R,t} = 0$ for $i \in \mathcal{G} \setminus \mathcal{G}_t$ by our choice $[\boldsymbol{Z}_{R,t}]_{\mathcal{G} \setminus \mathcal{G}_t} = \mathbf{0}$, and for $i \in \mathcal{G}_t \setminus \{g_t\}$ because we chose $\boldsymbol{s}_{R,t}$ to be orthogonal to the rows of $[\boldsymbol{Z}_{R,t}]_{\mathcal{G}_t \setminus \{g_t\}}$. Thus, $[\boldsymbol{A}_{R,t}]_{\mathcal{G}}$ has only one nonzero entry $[\boldsymbol{a}_{R,t}]_{g_t}$, which is in the g_tth column. But since $g_t \in \mathcal{P}_t$ and $\mathcal{P}_t \cap \mathcal{D}_t = \emptyset$, taking only the columns indexed by $\mathcal{D}_t$ results in $[\boldsymbol{A}_{R,t}]_{\mathcal{G}}^{\mathcal{D}_t} = \mathbf{0}$. We will use Lemma A.5 with $\boldsymbol{M} = [\boldsymbol{J}_{\phi}^{[R]}(\boldsymbol{s}, \boldsymbol{x})]_{\mathcal{I}}^{\mathcal{D}}$ given in (A.28), $n = |\mathcal{D}|$, $\mathcal{E} = \{i + (R-1)N : i \in \mathcal{G}\}$ (i.e., the rows of $[\boldsymbol{B}_R]_{[1:N-\gamma]}$ specified by $\mathcal{G}$), and $\mathcal{F} = [(R-1)\widetilde{T}Q+1 : R\widetilde{T}Q]$ (i.e., the columns of $[\boldsymbol{B}_R]_{[1:N-\gamma]}$). This choice yields

$$[\boldsymbol{M}]_{\mathcal{E}}^{\mathcal{F}} = [\boldsymbol{B}_R]_{\mathcal{G}} = \big([\boldsymbol{Z}_{R,1}]_{\mathcal{G}} \cdots [\boldsymbol{Z}_{R,\widetilde{T}}]_{\mathcal{G}}\big)$$

which is nonsingular as noted above. Furthermore, we have

$$[\boldsymbol{M}]_{\mathcal{E}}^{[1:n]\setminus\mathcal{F}} = (\mathbf{0} \;\; [\boldsymbol{A}_{R,1}]_{\mathcal{G}}^{\mathcal{D}_1} \cdots [\boldsymbol{A}_{R,\widetilde{T}}]_{\mathcal{G}}^{\mathcal{D}_{\widetilde{T}}}) = \mathbf{0}\,.$$

Hence, the requirements of Lemma A.5 are satisfied. We obtain that the determinant of $\boldsymbol{M} = [\boldsymbol{J}_{\phi}^{[R]}(\boldsymbol{s}, \boldsymbol{x})]_{\mathcal{I}}^{\mathcal{D}}$ in (A.28) is nonzero if and only if the determinant of the following matrix is nonzero:

$$\boldsymbol{K} \triangleq [\boldsymbol{M}]_{[1:n]\setminus\mathcal{E}}^{[1:n]\setminus\mathcal{F}} = \begin{pmatrix} \boldsymbol{B}_1 & & & [\boldsymbol{A}_{1,1}]^{\mathcal{D}_1} & \cdots & [\boldsymbol{A}_{1,\widetilde{T}}]^{\mathcal{D}_{\widetilde{T}}} \\ & \ddots & & \vdots & & \vdots \\ & & \boldsymbol{B}_{R-1} & [\boldsymbol{A}_{R-1,1}]^{\mathcal{D}_1} & \cdots & [\boldsymbol{A}_{R-1,\widetilde{T}}]^{\mathcal{D}_{\widetilde{T}}} \\ & \mathbf{0} & & [\boldsymbol{A}_{R,1}]_{\widetilde{\mathcal{L}}}^{\mathcal{D}_1} & \cdots & [\boldsymbol{A}_{R,\widetilde{T}}]_{\widetilde{\mathcal{L}}}^{\mathcal{D}_{\widetilde{T}}} \end{pmatrix}. \tag{A.31}$$

Here, we used Property (v-d) in Lemma A.4, i.e., that $[1 : N-\gamma] \setminus \mathcal{G} = \widetilde{\mathcal{L}}$.

So far, we specified only the rows $[\boldsymbol{Z}_{R,t}]_{\mathcal{G}}$. Because $\mathcal{G} \cap \widetilde{\mathcal{L}} = \emptyset$ by Property (v-d) in Lemma A.4, we can still freely choose the remaining rows $[\boldsymbol{Z}_{R,t}]_{\widetilde{\mathcal{L}}}$. We first choose the rows indexed by $\mathcal{L}_t$ such that $[\boldsymbol{Z}_{R,t}]_{\mathcal{L}_t} \boldsymbol{s}_{R,t}$ does not have zero entries (e.g., $[\boldsymbol{Z}_{R,t}]_{\{i\}} = \boldsymbol{s}_{R,t}^{\mathrm{H}}$ for $i \in \mathcal{L}_t$, resulting in $[\boldsymbol{Z}_{R,t}]_{\{i\}} \boldsymbol{s}_{R,t} = \|\boldsymbol{s}_{R,t}\|^2 \neq 0$). Next, we choose the remaining rows, indexed by $\widetilde{\mathcal{L}} \setminus \mathcal{L}_t$, to be zero, i.e., $[\boldsymbol{Z}_{R,t}]_{\widetilde{\mathcal{L}} \setminus \mathcal{L}_t} = \mathbf{0}$. With these choices and using (3.33), we obtain $[\boldsymbol{A}_{R,t}]_{\widetilde{\mathcal{L}} \setminus \mathcal{L}_t}^{\mathcal{D}_t} = \mathbf{0}$ and $\det\big([\boldsymbol{A}_{R,t}]_{\mathcal{L}_t}^{\mathcal{L}_t}\big) \neq 0$. We apply Lemma A.5 once again, with $\boldsymbol{M} = \boldsymbol{K}$ given in (A.31), $n = (R-1)\widetilde{T}Q + |\mathcal{D}|$, $\mathcal{E} = [(R-1)N + 1 : (R-1)\widetilde{T}Q + |\mathcal{D}|]$ (i.e., all rows below

$\boldsymbol{B}_{R-1}$), and

$$\mathcal{F} = \bigcup_{t\in[1:\widetilde{T}]} \Big\{ i + (R-1)\widetilde{T}Q + \sum_{t'\in[1:t-1]} |\mathcal{D}_{t'}| : i \in \mathcal{L}_t \Big\}$$

(i.e., the columns of $[\boldsymbol{A}_{R,t}]_{\mathcal{L}_t}^{\mathcal{D}_t}$ for all $t \in [1:\widetilde{T}]$). This choice results in

$$[\boldsymbol{M}]_{\mathcal{E}}^{\mathcal{F}} = \operatorname{diag}\left([\boldsymbol{A}_{R,1}]_{\mathcal{L}_1}^{\mathcal{L}_1}, \dots, [\boldsymbol{A}_{R,\widetilde{T}}]_{\mathcal{L}_{\widetilde{T}}}^{\mathcal{L}_{\widetilde{T}}}\right)$$

which is nonsingular because $\det\left([\boldsymbol{A}_{R,t}]_{\mathcal{L}_t}^{\mathcal{L}_t}\right) \neq 0$. Furthermore, we have

$$[\boldsymbol{M}]_{\mathcal{E}}^{[1:n]\setminus\mathcal{F}} = \Big(\boldsymbol{0} \;\; [\boldsymbol{A}_{R,1}]_{\widetilde{\mathcal{L}}}^{\mathcal{D}_1\setminus\mathcal{L}_1} \;\cdots\; [\boldsymbol{A}_{R,\widetilde{T}}]_{\widetilde{\mathcal{L}}}^{\mathcal{D}_{\widetilde{T}}\setminus\mathcal{L}_{\widetilde{T}}}\Big) = \boldsymbol{0}\,.$$

Thus, the requirements of Lemma A.5 are satisfied. We obtain that the determinant of $\boldsymbol{K}$ in (A.31) is nonzero if and only if the determinant of the following matrix is nonzero:

$$[\boldsymbol{M}]_{[1:n]\setminus\mathcal{E}}^{[1:n]\setminus\mathcal{F}} = \begin{pmatrix} \boldsymbol{B}_1 & & & [\boldsymbol{A}_{1,1}]^{\mathcal{D}_1\setminus\mathcal{L}_1} & \dots & [\boldsymbol{A}_{1,\widetilde{T}}]^{\mathcal{D}_{\widetilde{T}}\setminus\mathcal{L}_{\widetilde{T}}} \\ & \ddots & & \vdots & & \vdots \\ & & \boldsymbol{B}_{R-1} & [\boldsymbol{A}_{R-1,1}]^{\mathcal{D}_1\setminus\mathcal{L}_1} & \dots & [\boldsymbol{A}_{R-1,\widetilde{T}}]^{\mathcal{D}_{\widetilde{T}}\setminus\mathcal{L}_{\widetilde{T}}} \end{pmatrix}. \tag{A.32}$$

By the definitions $\mathcal{L}_t = \widetilde{\mathcal{P}}_t \setminus \mathcal{P}_t$, $\mathcal{D}_t = [1:N] \setminus \mathcal{P}_t$, and $\widetilde{\mathcal{D}}_t = [1:N] \setminus \widetilde{\mathcal{P}}_t$ (see (A.20), (A.23), and (A.30)), we obtain $\mathcal{D}_t \setminus \mathcal{L}_t = ([1:N] \setminus \mathcal{P}_t) \setminus (\widetilde{\mathcal{P}}_t \setminus \mathcal{P}_t) \overset{(a)}{=} [1:N] \setminus \widetilde{\mathcal{P}}_t = \widetilde{\mathcal{D}}_t$ for all $t \in [1:\widetilde{T}]$, where (a) holds because $\mathcal{P}_t \subseteq \widetilde{\mathcal{P}}_t$. Thus, $[\boldsymbol{M}]_{[1:n]\setminus\mathcal{E}}^{[1:n]\setminus\mathcal{F}}$ in (A.32) is equal to $[\boldsymbol{J}_{\phi}^{[R-1]}(\boldsymbol{s},\boldsymbol{x})]_{\widetilde{\mathcal{I}}}^{\widetilde{\mathcal{D}}}$ in (A.29). Altogether, we obtain that the determinant of $[\boldsymbol{J}_{\phi}^{[R]}(\boldsymbol{s},\boldsymbol{x})]_{\mathcal{I}}^{\mathcal{D}}$ in (A.28) is nonzero if and only if the determinant of $[\boldsymbol{M}]_{[1:n]\setminus\mathcal{E}}^{[1:n]\setminus\mathcal{F}} = [\boldsymbol{J}_{\phi}^{[R-1]}(\boldsymbol{s},\boldsymbol{x})]_{\widetilde{\mathcal{I}}}^{\widetilde{\mathcal{D}}}$ in (A.29) is nonzero. But the determinant of $[\boldsymbol{J}_{\phi}^{[R-1]}(\boldsymbol{s},\boldsymbol{x})]_{\widetilde{\mathcal{I}}}^{\widetilde{\mathcal{D}}}$ is nonzero by the induction hypothesis.

A.4 Proof of Lemma A.4

A.4.1 Bijectivity of $\boldsymbol{\beta}$

In order to prove Lemma A.4, we will use the following property of the function $\boldsymbol{\beta}$ in (A.17).

Lemma A.6 The function $\boldsymbol{\beta}$ defined in (A.17) is bijective.

Proof. To facilitate the exposition, we introduce the notation

$$L \triangleq \operatorname{lcm}(\widetilde{T}, N)\,.$$

Recall that $\boldsymbol{\beta}(j) = \big(\beta_1(j)\;\beta_2(j)\big)^{\mathrm{T}}$ with $\beta_1(j) = \big(j + \lfloor (j-1)/L \rfloor\big) \bmod^* \widetilde{T} \in [1:\widetilde{T}]$ and $\beta_2(j) = j \bmod^* N \in [1:N]$, for $j \in [1:\widetilde{T}N]$. We start by proving that $\boldsymbol{\beta}$ is one-to-one. Assume that there exist $j_1, j_2 \in [1:\widetilde{T}N]$ with $j_1 \leq j_2$ such that $\boldsymbol{\beta}(j_1) = \boldsymbol{\beta}(j_2)$. From $\beta_2(j_1) = \beta_2(j_2)$, it follows that $j_1 \bmod^* N = j_2 \bmod^* N$ and, hence,[3] $j_2 = j_1 + nN$ for some $n \in [0:\widetilde{T}-1]$. Similarly, $\beta_1(j_1) = \beta_1(j_2)$ implies that

$$j_1 + \left\lfloor \frac{j_1 - 1}{L} \right\rfloor = j_2 + \left\lfloor \frac{j_2 - 1}{L} \right\rfloor - m\widetilde{T}$$

for some $m \in \mathbb{N}$, and thus

$$j_1 + \left\lfloor \frac{j_1 - 1}{L} \right\rfloor = j_1 + nN + \left\lfloor \frac{j_1 + nN - 1}{L} \right\rfloor - m\widetilde{T}$$

or, equivalently,

$$m\widetilde{T} - nN = \left\lfloor \frac{j_1 + nN - 1}{L} \right\rfloor - \left\lfloor \frac{j_1 - 1}{L} \right\rfloor . \tag{A.33}$$

We can write $j_1 = kL + \tilde{\jmath}_1$ with some $k \in \mathbb{N}$ and $\tilde{\jmath}_1 \in [1:L]$ and simplify (A.33) as follows:

$$\begin{aligned} m\widetilde{T} - nN &= \left\lfloor \frac{kL + \tilde{\jmath}_1 + nN - 1}{L} \right\rfloor - \left\lfloor \frac{kL + \tilde{\jmath}_1 - 1}{L} \right\rfloor \\ &= k + \left\lfloor \frac{\tilde{\jmath}_1 + nN - 1}{L} \right\rfloor - k - \left\lfloor \frac{\tilde{\jmath}_1 - 1}{L} \right\rfloor \\ &\overset{(a)}{=} \left\lfloor \frac{\tilde{\jmath}_1 + nN - 1}{L} \right\rfloor . \end{aligned} \tag{A.34}$$

Here, (a) holds because $\tilde{\jmath}_1 - 1 < L$ and thus $\lfloor (\tilde{\jmath}_1 - 1)/L \rfloor = 0$. We will next show that the right-hand side of (A.34) is zero, by establishing the following chain of inequalities:

$$0 \le \left\lfloor \frac{\tilde{\jmath}_1 + nN - 1}{L} \right\rfloor \overset{(a)}{\le} \left\lfloor \frac{j_1 + nN - 1}{L} \right\rfloor \overset{(b)}{\le} \left\lfloor \frac{\widetilde{T}N - 1}{L} \right\rfloor \overset{(c)}{=} \left\lfloor \gcd(\widetilde{T}, N) - \frac{1}{L} \right\rfloor = \gcd(\widetilde{T}, N) - 1 . \tag{A.35}$$

Here, (a) holds because $\tilde{\jmath}_1 \le j_1$, (b) holds because $j_1 + nN = j_2 \le \widetilde{T}N$, and (c) holds because $\widetilde{T}N = \gcd(\widetilde{T}, N)L$ [Hardy and Wright, 1975, Th. 52] (here, $\gcd(\cdot,\cdot)$ denotes the greatest common divisor). Note now that $\gcd(\widetilde{T}, N)$ divides the left-hand side of (A.34) and, hence, also the right-hand side. But by (A.35), the right-hand side of (A.34) is an element of $[0:\gcd(\widetilde{T}, N) - 1]$. Hence, it must be zero, and thus (A.34) becomes

$$m\widetilde{T} - nN = \left\lfloor \frac{\tilde{\jmath}_1 + nN - 1}{L} \right\rfloor = 0 . \tag{A.36}$$

Therefore, $\tilde{\jmath}_1 + nN - 1 < L$. Since $nN \le \tilde{\jmath}_1 + nN - 1$, we obtain $nN < L$. Furthermore, by (A.36), we have that $m\widetilde{T} = nN$. Thus, nN is a common multiple of $\widetilde{T}$ and N that is less than the least (positive) common multiple. Therefore, $n = 0$ and, hence, $j_1 = j_1 + nN = j_2$. We have thus shown that $\boldsymbol{\beta}(j_1) = \boldsymbol{\beta}(j_2)$ implies $j_1 = j_2$, which means that $\boldsymbol{\beta}$ is one-to-one. Since the domain of $\boldsymbol{\beta}$, $[1:\widetilde{T}N]$, and its codomain, $[1:\widetilde{T}] \times [1:N]$, are finite and of the same

[3]Recall that we defined $a \bmod^* b \triangleq a - b\lfloor (a-1)/b \rfloor$ to be the residuum of a divided by b in $[1:b]$ (and not in $[0:b-1]$ as commonly done).

cardinality (namely, $\widetilde{T}N$), we conclude that $\boldsymbol{\beta}$ is also bijective. □

We will now prove the individual properties stated in Lemma A.4.

A.4.2 Proof of Property (i)

We first show that $\beta_2\big|_{\beta_1^{-1}(t)}$ is one-to-one, i.e., if $\beta_2(j_1) = \beta_2(j_2)$ for $j_1, j_2 \in \beta_1^{-1}(t)$ then $j_1 = j_2$. To this end, let $j_1, j_2 \in \beta_1^{-1}(t)$ (i.e., $\beta_1(j_1) = \beta_1(j_2) = t$) and assume that $\beta_2(j_1) = \beta_2(j_2) = i$. Then $\boldsymbol{\beta}(j_1) = \boldsymbol{\beta}(j_2) = (t\ i)^{\mathrm{T}}$. Since $\boldsymbol{\beta}$ is one-to-one by Lemma A.6, we conclude that $j_1 = j_2$. Hence, $\beta_2\big|_{\beta_1^{-1}(t)}$ is one-to-one. Furthermore, since $\beta_1^{-1}(t) \cap [1\!:\!\vartheta_R] \subseteq \beta_1^{-1}(t)$, we have (cf. (A.18))

$$|\mathcal{P}_t| = \big|\beta_2\big(\beta_1^{-1}(t) \cap [1\!:\!\vartheta_R]\big)\big| = \big|\beta_1^{-1}(t) \cap [1\!:\!\vartheta_R]\big| \tag{A.37}$$

for $t \in [1\!:\!\widetilde{T}]$. To conclude the proof, we will use the following basic lemma.

Lemma A.7 The sets $\{\beta_1^{-1}(t)\}_{t\in[1:\widetilde{T}]}$ form a partition of the domain $[1\!:\!\widetilde{T}N]$ of β_1, i.e.,

$$\beta_1^{-1}(t) \cap \beta_1^{-1}(t') = \emptyset, \quad \text{for } t, t' \in [1\!:\!\widetilde{T}] \text{ with } t \neq t' \tag{A.38}$$

and

$$\bigcup_{t\in[1:\widetilde{T}]} \beta_1^{-1}(t) = [1\!:\!\widetilde{T}N]\,. \tag{A.39}$$

Proof. This lemma follows from the definition of a function, i.e., the fact that β_1 maps every element in the domain to exactly one element in the codomain. □

By Lemma A.7, we obtain

$$\begin{aligned}
\sum_{t\in[1:\widetilde{T}]} |\mathcal{P}_t| &\overset{\text{(A.37)}}{=} \sum_{t\in[1:\widetilde{T}]} \big|\beta_1^{-1}(t) \cap [1\!:\!\vartheta_R]\big| \\
&\overset{\text{(A.38)}}{=} \left|\left(\bigcup_{t\in[1:\widetilde{T}]} \beta_1^{-1}(t)\right) \cap [1\!:\!\vartheta_R]\right| \\
&\overset{\text{(A.39)}}{=} \big|[1\!:\!\widetilde{T}N] \cap [1\!:\!\vartheta_R]\big| \\
&= \min\{\widetilde{T}N, \vartheta_R\}\,.
\end{aligned} \tag{A.40}$$

Since $N > \widetilde{T}Q$, we have that $\vartheta_R = \max\{\widetilde{T}, R\widetilde{T}Q - (R - \widetilde{T})N\} = \max\{\widetilde{T}, \widetilde{T}N - R(N - \widetilde{T}Q)\} < \widetilde{T}N$. Combining this with (A.40), we conclude that

$$\sum_{t\in[1:\widetilde{T}]} |\mathcal{P}_t| = \vartheta_R\,.$$

A.4.3 Proof of Property (ii)

We will make use of the following lemma.

Lemma A.8 Let $p, q \in \mathbb{N}$ with $p < q$. Then

$$\left|\left\{j \in [p+1:q] : (j+a) \operatorname{mod}^* b = c\right\}\right| \leq \left\lceil \frac{q-p}{b} \right\rceil$$

for all $a, b, c \in \mathbb{N}$ with $b \geq 2$, $c \geq 1$, and $c \leq b$.

Proof. We prove Lemma A.8 by contradiction. Assume

$$\left|\left\{j \in [p+1:q] : (j+a) \operatorname{mod}^* b = c\right\}\right| > \left\lceil \frac{q-p}{b} \right\rceil \triangleq d\,.$$

Thus, the set $\left\{j \in [p+1:q] : (j+a) \operatorname{mod}^* b = c\right\}$ contains at least $d+1$ elements $\{j_i\}_{i\in[1:d+1]}$, i.e., there exist at least $d+1$ distinct elements $j_i \in [p+1:q]$ satisfying $(j_i + a) \operatorname{mod}^* b = c$. Hence, there exist distinct $k_i \in \mathbb{N}$, $i \in [1:d+1]$ such that

$$j_i + a = c + k_i b \in [p+1:q]\,. \tag{A.41}$$

Assume, without loss of generality, that $k_i < k_{i+1}$ for $i \in [1:d]$. Because $k_i \in \mathbb{N}$, we obtain $k_i \leq k_{i+1} - 1$ and thus, iteratively, $k_1 \leq k_2 - 1 \leq k_3 - 2 \leq \cdots$, and finally

$$k_1 \leq k_{d+1} - d\,. \tag{A.42}$$

Hence,

$$j_{d+1} - j_1 \overset{\text{(A.41)}}{=} k_{d+1}b - k_1 b = (k_{d+1} - k_1)b \overset{\text{(A.42)}}{\geq} d\,b = \left\lceil \frac{q-p}{b} \right\rceil b \geq q - p$$

which contradicts $j_1, j_{d+1} \in [p+1:q]$. □

To prove Property (ii), we first establish an upper bound on ϑ_R. We have that

$$\begin{aligned} R\widetilde{T}Q - (R-\widetilde{T})N &= (R-\widetilde{T})\widetilde{T}Q - (R-\widetilde{T})N + \widetilde{T}^2 Q \\ &= \underbrace{(R-\widetilde{T})}_{\geq 0}\underbrace{(\widetilde{T}Q - N)}_{<0} + \widetilde{T}^2 Q \\ &\leq \widetilde{T}^2 Q \end{aligned}$$

and, hence,

$$\vartheta_R = \max\{\widetilde{T}, R\widetilde{T}Q - (R-\widetilde{T})N\} \leq \widetilde{T}^2 Q\,. \tag{A.43}$$

To bound the size of the sets $\mathcal{P}_t$, we use (A.37) and the definition of β_1 to conclude that

$$\begin{aligned} |\mathcal{P}_t| &= \left|\{j \in [1:\vartheta_R] : \beta_1(j) = t\}\right| \\ &= \left|\left\{j \in [1:\vartheta_R] : \left(j + \left\lfloor \frac{j-1}{L} \right\rfloor\right) \operatorname{mod}^* \widetilde{T} = t\right\}\right|. \end{aligned} \tag{A.44}$$

Choose $m \in \mathbb{N}$ such that $(m-1)L < \vartheta_R \leq mL$. We can partition the set $[1:\vartheta_R]$ as follows:

$$[1:\vartheta_R] = \left(\bigcup_{n\in[0:m-2]} \big[nL+1:(n+1)L\big] \right) \cup \big[(m-1)L+1:\vartheta_R\big]. \tag{A.45}$$

Note that the intervals $\big[nL+1:(n+1)L\big]$, $n \in [0:m-2]$ and $\big[(m-1)L+1:\vartheta_R\big]$ in (A.45) are disjoint and satisfy

$$\left\lfloor \frac{j-1}{L} \right\rfloor = \begin{cases} n, & \text{for } j \in \big[nL+1:(n+1)L\big] \\ m-1, & \text{for } j \in \big[(m-1)L+1:\vartheta_R\big]. \end{cases} \tag{A.46}$$

Thus, using (A.45) and (A.46) in (A.44), we obtain

$$\begin{aligned} |\mathcal{P}_t| = \sum_{n\in[0:m-2]} &\left|\left\{j \in \big[nL+1:(n+1)L\big] : (j+n) \bmod^* \widetilde{T} = t\right\}\right| \\ &+ \left|\left\{j \in \big[(m-1)L+1:\vartheta_R\big] : (j+m-1) \bmod^* \widetilde{T} = t\right\}\right|. \end{aligned} \tag{A.47}$$

By Lemma A.8, we have

$$\left|\left\{j \in \big[nL+1:(n+1)L\big] : (j+n) \bmod^* \widetilde{T} = t\right\}\right| \leq \left\lceil \frac{L}{\widetilde{T}} \right\rceil = \frac{L}{\widetilde{T}} \tag{A.48}$$

and

$$\left|\left\{j \in \big[(m-1)L+1:\vartheta_R\big] : (j+m-1) \bmod^* \widetilde{T} = t\right\}\right| \leq \left\lceil \frac{\vartheta_R - (m-1)L}{\widetilde{T}} \right\rceil. \tag{A.49}$$

Thus, inserting (A.48) and (A.49) into (A.47), we obtain

$$\begin{aligned} |\mathcal{P}_t| &\leq (m-1)\frac{L}{\widetilde{T}} + \left\lceil \frac{\vartheta_R - (m-1)L}{\widetilde{T}} \right\rceil \\ &\overset{(a)}{=} (m-1)\frac{L}{\widetilde{T}} + \left\lceil \frac{\vartheta_R}{\widetilde{T}} \right\rceil - (m-1)\frac{L}{\widetilde{T}} \\ &= \left\lceil \frac{\vartheta_R}{\widetilde{T}} \right\rceil \\ &\overset{\text{(A.43)}}{\leq} \left\lceil \frac{\widetilde{T}^2 Q}{\widetilde{T}} \right\rceil \\ &= \widetilde{T}Q \end{aligned}$$

where (a) holds because $L/\widetilde{T} \in \mathbb{N}$ (recall that $L = \operatorname{lcm}(\widetilde{T}, N)$).

A.4.4 Proof of Property (iii)

To prove Properties (iii)–(v), we calculate the difference $\vartheta_{R-1} - \vartheta_R$. Because we assumed that $R \leq \lceil \widetilde{T}(N-1)/(N-\widetilde{T}Q) \rceil$, we have $R-1 < \widetilde{T}(N-1)/(N-\widetilde{T}Q)$. This is easily verified to be equivalent to $(R-1)\widetilde{T}Q - (R-1-\widetilde{T})N > \widetilde{T}$. Hence, using (A.16),

$$\begin{aligned} \vartheta_{R-1} &= \max\{\widetilde{T}, (R-1)\widetilde{T}Q - (R-1-\widetilde{T})N\} \\ &= (R-1)\widetilde{T}Q - (R-1-\widetilde{T})N\,. \end{aligned} \tag{A.50}$$

Thus, we have

$$\begin{aligned} \vartheta_{R-1} - \vartheta_R &= (R-1)\widetilde{T}Q - (R-1-\widetilde{T})N - \max\{\widetilde{T}, R\widetilde{T}Q - (R-\widetilde{T})N\} \\ &= R\widetilde{T}Q - (R-\widetilde{T})N + N - \widetilde{T}Q - \max\{\widetilde{T}, R\widetilde{T}Q - (R-\widetilde{T})N\} \\ &= N - \widetilde{T}Q - \max\big\{\widetilde{T} - \big(R\widetilde{T}Q - (R-\widetilde{T})N\big), 0\big\} \\ &= N - \widetilde{T}Q - \gamma \end{aligned} \tag{A.51}$$

where γ was defined in (A.22). Furthermore, by (3.29), $\gamma < N - \widetilde{T}Q$ and thus (A.51) implies

$$\vartheta_{R-1} - \vartheta_R > 0\,. \tag{A.52}$$

We are now ready to prove Property (iii). From the definitions $\mathcal{P}_t \triangleq \beta_2\big(\beta_1^{-1}(t) \cap [1:\vartheta_R]\big)$ in (A.18) and $\widetilde{\mathcal{P}}_t \triangleq \beta_2\big(\beta_1^{-1}(t) \cap [1:\vartheta_{R-1}]\big)$ in (A.19), it follows that $\mathcal{L}_t = \widetilde{\mathcal{P}}_t \setminus \mathcal{P}_t$ (recall (A.20)) can be written as

$$\begin{aligned} \mathcal{L}_t &= \beta_2\big(\beta_1^{-1}(t) \cap [1:\vartheta_{R-1}]\big) \setminus \beta_2\big(\beta_1^{-1}(t) \cap [1:\vartheta_R]\big) \\ &\overset{(a)}{=} \beta_2\big((\beta_1^{-1}(t) \cap [1:\vartheta_{R-1}]) \setminus (\beta_1^{-1}(t) \cap [1:\vartheta_R])\big) \\ &= \beta_2\big(\beta_1^{-1}(t) \cap [\vartheta_R + 1:\vartheta_{R-1}]\big) \end{aligned} \tag{A.53}$$

where (a) holds because $\beta_2\big|_{\beta_1^{-1}(t)}$ is one-to-one (see Section A.4.2). Since $\beta_2(j) = j \bmod^* N$, the function β_2 is one-to-one on every set consisting of up to N consecutive integers. In particular, (A.52) and (A.51) imply that $\big|[\vartheta_R + 1:\vartheta_{R-1}]\big| = \vartheta_{R-1} - \vartheta_R = N - \widetilde{T}Q - \gamma$ and hence $\beta_2\big|_{[\vartheta_R+1:\vartheta_{R-1}]}$ is one-to-one. Because by Lemma A.7 the sets $\beta_1^{-1}(t)$, $t \in [1:\widetilde{T}]$ are pairwise disjoint, we conclude that the sets $\beta_1^{-1}(t) \cap [\vartheta_R + 1:\vartheta_{R-1}]$, $t \in [1:\widetilde{T}]$ are pairwise disjoint too. Hence, by (A.53) and because $\beta_2\big|_{[\vartheta_R+1:\vartheta_{R-1}]}$ is one-to-one, the sets $\mathcal{L}_t$ are pairwise disjoint.

A.4.5 Proof of Property (iv)

By (A.53), we have

$$\mathcal{L}_t = \beta_2\big(\beta_1^{-1}(t) \cap [\vartheta_R + 1:\vartheta_{R-1}]\big) \subseteq \beta_2([1:\vartheta_{R-1}])\,. \tag{A.54}$$

Hence, it remains to prove that

$$\beta_2([1:\vartheta_{R-1}]) \subseteq [1:N-\gamma]\,. \tag{A.55}$$

Recall that we assumed $R \leq \lceil \widetilde{T}(N-1)/(N-\widetilde{T}Q) \rceil$. If $R < \lceil \widetilde{T}(N-1)/(N-\widetilde{T}Q) \rceil$, then $R < \widetilde{T}(N-1)/(N-\widetilde{T}Q)$ (because $R \in \mathbb{N}$), which implies $RN-(R\widetilde{T}Q+\widetilde{T}N-\widetilde{T}) < 0$; hence, it follows from the definition of γ in (A.22) that $\gamma = 0$. In this case, it follows from the definition of β_2 in (A.17), i.e., $\beta_2(j) = j \bmod^* N$ for $j \in [1:\widetilde{T}N]$, that (A.55) is trivially true. For the complementary case $R = \lceil \widetilde{T}(N-1)/(N-\widetilde{T}Q) \rceil$, we note that $RN - (R\widetilde{T}Q + \widetilde{T}N - \widetilde{T}) \geq 0$ and hence, using the definition of γ in (A.22),

$$\begin{aligned} N-\gamma &= N-(RN-R\widetilde{T}Q-\widetilde{T}N+\widetilde{T}) \\ &= R\widetilde{T}Q-(R-1-\widetilde{T})N-\widetilde{T} \\ &\geq (R-1)\widetilde{T}Q-(R-1-\widetilde{T})N \\ &\overset{\text{(A.50)}}{=} \vartheta_{R-1}\,. \end{aligned}$$

Thus, $[1:\vartheta_{R-1}] \subseteq [1:N-\gamma]$ and, further, $\beta_2([1:\vartheta_{R-1}]) \subseteq \beta_2([1:N-\gamma]) = [1:N-\gamma]$, i.e., (A.55) is again true. Combining (A.54) and (A.55) concludes the proof that $\mathcal{L}_t \subseteq [1:N-\gamma]$.

A.4.6 Proof of Property (v)

We have

$$\begin{aligned} \widetilde{\mathcal{L}} &\overset{\text{(A.21)}}{=} \bigcup_{t\in[1:\widetilde{T}]} \mathcal{L}_t \\ &\overset{\text{(A.53)}}{=} \bigcup_{t\in[1:\widetilde{T}]} \beta_2\big(\beta_1^{-1}(t) \cap [\vartheta_R+1:\vartheta_{R-1}]\big) \\ &\overset{(a)}{=} \beta_2\Bigg(\bigcup_{t\in[1:\widetilde{T}]} \Big(\beta_1^{-1}(t) \cap [\vartheta_R+1:\vartheta_{R-1}]\Big)\Bigg) \\ &= \beta_2\Bigg(\Bigg(\bigcup_{t\in[1:\widetilde{T}]} \beta_1^{-1}(t)\Bigg) \cap [\vartheta_R+1:\vartheta_{R-1}]\Bigg) \\ &\overset{\text{(A.39)}}{=} \beta_2([\vartheta_R+1:\vartheta_{R-1}]) \end{aligned} \tag{A.56}$$

where (a) holds because β_2 is one-to-one on every set consisting of up to N consecutive integers. Thus, $\big|\widetilde{\mathcal{L}}\big| = \big|\beta_2([\vartheta_R+1:\vartheta_{R-1}])\big| = \vartheta_{R-1}-\vartheta_R \overset{\text{(A.51)}}{=} N-\widetilde{T}Q-\gamma$. Furthermore, Property (iv) implies that the set $\widetilde{\mathcal{L}}$ is a subset of $[1:N-\gamma]$, and hence we obtain for the size of $\mathcal{G} = [1:N-\gamma] \setminus \widetilde{\mathcal{L}}$

$$|\mathcal{G}| = \big|[1:N-\gamma] \setminus \widetilde{\mathcal{L}}\big| = N-\gamma-(N-\widetilde{T}Q-\gamma) = \widetilde{T}Q\,.$$

Thus, we can partition $\mathcal{G}$ as $\mathcal{G} = \bigcup_{t\in[1:\widetilde{T}]} \mathcal{G}_t$, with disjoint $\mathcal{G}_t$ of size Q each. We have thus shown the existence of sets $\mathcal{G}_t$ satisfying (v-a), (v-b), and (v-d).

It remains to show (v-c), i.e., that we can choose $\{\mathcal{G}_t\}_{t\in[1:\widetilde{T}]}$ such that each $\mathcal{G}_t$ has a nonempty intersection with $\mathcal{P}_t$. Because β_2 is one-to-one on sets of up to N consecutive integers and

$$\vartheta_{R-1} - (\vartheta_R - \widetilde{T}) \overset{\text{(A.51)}}{=} N - \widetilde{T}Q - \gamma + \widetilde{T} = N - \gamma - \widetilde{T}(Q-1) \leq N - \gamma$$

we obtain that $\beta_2\big|_{[\vartheta_R - \widetilde{T}+1:\vartheta_{R-1}]}$ is one-to-one. Thus,

$$\begin{aligned} \beta_2([\vartheta_R - \widetilde{T} + 1:\vartheta_R]) \cap \beta_2([\vartheta_R + 1:\vartheta_{R-1}]) &= \beta_2([\vartheta_R - \widetilde{T} + 1:\vartheta_R] \cap [\vartheta_R + 1:\vartheta_{R-1}]) \\ &= \beta_2(\emptyset) \\ &= \emptyset\,. \end{aligned} \tag{A.57}$$

Inserting (A.56) into (A.57), we obtain

$$\beta_2([\vartheta_R - \widetilde{T} + 1:\vartheta_R]) \cap \widetilde{\mathcal{L}} = \emptyset\,. \tag{A.58}$$

By the fact that $[\vartheta_R - \widetilde{T} + 1:\vartheta_R] \subseteq [1:\vartheta_{R-1}]$ and (A.55), we have that $\beta_2([\vartheta_R - \widetilde{T} + 1:\vartheta_R]) \subseteq \beta_2([1:\vartheta_{R-1}]) \subseteq [1:N-\gamma]$. Hence, (A.58) implies that

$$\beta_2([\vartheta_R - \widetilde{T} + 1:\vartheta_R]) \subseteq [1:N-\gamma] \setminus \widetilde{\mathcal{L}} = \mathcal{G}\,.$$

Thus, we identified $\widetilde{T}$ elements $\beta_2(\vartheta_R - \widetilde{T} + 1), \beta_2(\vartheta_R - \widetilde{T} + 2), \dots, \beta_2(\vartheta_R)$ in the set $\mathcal{G}$, which will now be used to construct the sets $\mathcal{G}_t$. We will show that we can assign a different index $t \in [1:\widetilde{T}]$ to each of these $\widetilde{T}$ elements such that the element with index t belongs to $\mathcal{P}_t$, i.e.,

$$\beta_2([\vartheta_R - \widetilde{T} + 1:\vartheta_R]) = \{g_1, \dots, g_{\widetilde{T}}\}, \quad \text{with } g_t \in \mathcal{P}_t, t \in [1:\widetilde{T}]\,. \tag{A.59}$$

The desired sets $\mathcal{G}_t$ are then obtained by assigning g_t to $\mathcal{G}_t$, for $t \in [1:\widetilde{T}]$. Thus, recalling that $|\mathcal{G}_t| = Q$, $\mathcal{G}_t$ consists of $g_t \in \mathcal{P}_t$ and $Q-1$ additional elements taken from the set $\mathcal{G} \setminus \beta_2([\vartheta_R - \widetilde{T} + 1:\vartheta_R])$.

In order to prove (A.59), we distinguish two cases.

Case $nL \notin [\vartheta_R - \widetilde{T} + 1:\vartheta_R - 1]$ for all $n \in \mathbb{N}$

In this case, there exists $m \in \mathbb{N}$ such that $mL \leq \vartheta_R - \widetilde{T}$ and $(m+1)L \geq \vartheta_R$. Thus, for all $j \in [\vartheta_R - \widetilde{T} + 1:\vartheta_R]$, we have

$$\left\lfloor \frac{j-1}{L} \right\rfloor \geq \left\lfloor \frac{\vartheta_R - \widetilde{T}}{L} \right\rfloor \geq \left\lfloor \frac{mL}{L} \right\rfloor = m \tag{A.60}$$

and

$$\left\lfloor \frac{j-1}{L} \right\rfloor < \left\lfloor \frac{\vartheta_R}{L} \right\rfloor \leq \left\lfloor \frac{(m+1)L}{L} \right\rfloor = m+1\,. \tag{A.61}$$

Combining (A.60) and (A.61), we obtain that the offset in (A.17) satisfies $\lfloor (j-1)/L \rfloor = m$ for all $j \in [\vartheta_R - \widetilde{T} + 1:\vartheta_R]$. Thus, we have $\beta_1\big|_{[\vartheta_R - \widetilde{T}+1:\vartheta_R]}(j) = (j+m) \operatorname{mod}^* \widetilde{T}$, which implies that $\beta_1([\vartheta_R - \widetilde{T} + 1 : \vartheta_R]) = [1 : \widetilde{T}]$. Hence, we can write

$$[\vartheta_R - \widetilde{T} + 1 : \vartheta_R] = \{\tilde{\jmath}_1, \dots, \tilde{\jmath}_{\widetilde{T}}\}, \quad \text{where } \tilde{\jmath}_t \in \beta_1^{-1}(t) \text{ for } t \in [1 : \widetilde{T}] \,.$$

We then obtain

$$\beta_2([\vartheta_R - \widetilde{T} + 1 : \vartheta_R]) = \{\beta_2(\tilde{\jmath}_1), \dots, \beta_2(\tilde{\jmath}_{\widetilde{T}})\}$$

and assign the indices $t \in [1 : \widetilde{T}]$ according to $g_t = \beta_2(\tilde{\jmath}_t)$. By construction, we have both $g_t = \beta_2(\tilde{\jmath}_t) \in \beta_2(\beta_1^{-1}(t))$ and $g_t = \beta_2(\tilde{\jmath}_t) \in \beta_2([\vartheta_R - \widetilde{T} + 1 : \vartheta_R]) \subseteq \beta_2([1 : \vartheta_R])$, so that we also have

$$g_t \in \beta_2(\beta_1^{-1}(t) \cap [1 : \vartheta_R]) = \mathcal{P}_t$$

(recall (A.18)). Thus, our choice of the g_t satisfies (A.59).

Case $nL \in [\vartheta_R - \widetilde{T} + 1 : \vartheta_R - 1]$ **for some** $n \in \mathbb{N}$

We first note that

$$\begin{aligned} \beta_2([\vartheta_R - \widetilde{T} + 1 : \vartheta_R]) &= \beta_2([\vartheta_R - \widetilde{T} + 1 : nL]) \cup \beta_2([nL + 1 : \vartheta_R]) \\ &\overset{(a)}{=} \beta_2([\vartheta_R - \widetilde{T} + 1 : nL]) \cup \beta_2([nL - L + 1 : \vartheta_R - L]) \\ &= \beta_2([\vartheta_R - \widetilde{T} + 1 : nL]) \cup \beta_2([(n-1)L + 1 : \vartheta_R - L]) \end{aligned} \tag{A.62}$$

where (a) holds because (recall that $L = \mathrm{lcm}(\widetilde{T}, N)$ is a multiple of N)

$$\beta_2(j) = j \bmod^* N = (j - L) \bmod^* N = \beta_2(j - L)$$

for $j > L$. We will next calculate the offset $\lfloor (j-1)/L \rfloor$ in (A.17) for j belonging to either of the intervals in the arguments in (A.62), i.e., $j \in [\vartheta_R - \widetilde{T} + 1 : nL]$ or $j \in [(n-1)L + 1 : \vartheta_R - L]$. Note that

$$nL \in [\vartheta_R - \widetilde{T} + 1 : \vartheta_R - 1] \tag{A.63}$$

and

$$L \geq \widetilde{T} \,. \tag{A.64}$$

Thus, we have

$$(n-1)L = nL - L \overset{\text{(A.63)}}{<} \vartheta_R - L \overset{\text{(A.64)}}{\leq} \vartheta_R - \widetilde{T} \tag{A.65}$$

and

$$\vartheta_R \overset{\text{(A.63)}}{<} nL + \widetilde{T} \overset{\text{(A.64)}}{\leq} (n+1)L \,. \tag{A.66}$$

For $j \in [\vartheta_R - \widetilde{T} + 1 : nL]$, we obtain that $j - 1 \geq \vartheta_R - \widetilde{T} \overset{\text{(A.65)}}{>} (n-1)L$ and $j - 1 \leq nL - 1$. Hence, $n - 1 < (j-1)/L < n$ and further

$$\left\lfloor \frac{j-1}{L} \right\rfloor = n - 1, \quad \text{for } j \in [\vartheta_R - \widetilde{T} + 1 : nL] \,. \tag{A.67}$$

Similarly, for $j \in [(n-1)L+1 \colon \vartheta_R - L]$, we obtain $j-1 \le \vartheta_R - L - 1 \overset{\text{(A.66)}}{<} (n+1)L - L - 1 = nL - 1$ and $j - 1 \ge (n-1)L$. Thus, $n - 1 \le (j-1)/L < n$ and further

$$\left\lfloor \frac{j-1}{L} \right\rfloor = n - 1, \quad \text{for } j \in [(n-1)L + 1 \colon \vartheta_R - L]\,. \tag{A.68}$$

Combining (A.67) and (A.68), we conclude that the offset in (A.17) satisfies

$$\left\lfloor \frac{j-1}{L} \right\rfloor = n - 1\,, \quad \text{for } j \in [\vartheta_R - \widetilde{T} + 1 \colon nL] \cup [(n-1)L + 1 \colon \vartheta_R - L]\,. \tag{A.69}$$

Let us next consider β_1 on the sets $[\vartheta_R - \widetilde{T} + 1 \colon nL]$ and $[(n-1)L + 1 \colon \vartheta_R - L]$. We obtain

$$\begin{aligned}
&\beta_1([\vartheta_R - \widetilde{T} + 1 \colon nL]) \\
&\quad = \big\{k = \beta_1(j) = \big(j + \lfloor (j-1)/L \rfloor\big) \bmod^* \widetilde{T} : j \in [\vartheta_R - \widetilde{T} + 1 \colon nL]\big\} \\
&\quad \overset{\text{(A.69)}}{=} \big\{k = \beta_1(j) = (j + n - 1) \bmod^* \widetilde{T} : j \in [\vartheta_R - \widetilde{T} + 1 \colon nL]\big\} \\
&\quad = \big\{k = j \bmod^* \widetilde{T} : j \in [\vartheta_R - \widetilde{T} + n \colon nL + n - 1]\big\} \\
&\quad \overset{(a)}{=} \big\{k \in [1 \colon \widetilde{T}] : \exists\, m \in \mathbb{N} \text{ such that } k + m\widetilde{T} \in [\vartheta_R - \widetilde{T} + n \colon nL + n - 1]\big\}
\end{aligned} \tag{A.70}$$

where (a) holds because $k = j \bmod^* \widetilde{T}$ is equivalent to $j = k + m\widetilde{T}$ for some $m \in \mathbb{N}$. Similarly,

$$\begin{aligned}
&\beta_1([(n-1)L + 1 \colon \vartheta_R - L]) \\
&\quad = \big\{k = \beta_1(j) = \big(j + \lfloor (j-1)/L \rfloor\big) \bmod^* \widetilde{T} : j \in [(n-1)L + 1 \colon \vartheta_R - L]\big\} \\
&\quad \overset{\text{(A.69)}}{=} \big\{k = \beta_1(j) = (j + n - 1) \bmod^* \widetilde{T} : j \in [(n-1)L + 1 \colon \vartheta_R - L]\big\} \\
&\quad = \big\{k = j \bmod^* \widetilde{T} : j \in [(n-1)L + n \colon \vartheta_R - L + n - 1]\big\} \\
&\quad = \big\{k \in [1 \colon \widetilde{T}] : \exists\, m \in \mathbb{N} \text{ such that } k + m\widetilde{T} \in [(n-1)L + n \colon \vartheta_R - L + n - 1]\big\} \\
&\quad \overset{(a)}{=} \big\{k \in [1 \colon \widetilde{T}] : \exists\, m \in \mathbb{N} \text{ such that } k + m\widetilde{T} \in [nL + n \colon \vartheta_R + n - 1]\big\}
\end{aligned} \tag{A.71}$$

where (a) holds because a shift of the interval by L (which is a multiple of $\widetilde{T}$) can be compensated by choosing a different $m \in \mathbb{N}$. Combining (A.70) and (A.71), we obtain

$$\begin{aligned}
&\beta_1\big([\vartheta_R - \widetilde{T} + 1 \colon nL] \cup [(n-1)L + 1 \colon \vartheta_R - L]\big) \\
&\qquad = \big\{k \in [1 \colon \widetilde{T}] : \exists\, m \in \mathbb{N} \text{ such that} \\
&\qquad\qquad k + m\widetilde{T} \in [\vartheta_R - \widetilde{T} + n \colon nL + n - 1] \cup [nL + n \colon \vartheta_R + n - 1]\big\} \\
&\qquad = \big\{k \in [1 \colon \widetilde{T}] : \exists\, m \in \mathbb{N} \text{ such that } k + m\widetilde{T} \in [\vartheta_R - \widetilde{T} + n \colon \vartheta_R + n - 1]\big\} \\
&\qquad \overset{(a)}{=} [1 \colon \widetilde{T}]
\end{aligned} \tag{A.72}$$

where (a) holds because $[\vartheta_R - \widetilde{T} + n \colon \vartheta_R + n - 1]$ is an interval of length $\widetilde{T}$ and thus for every $k \in [1 \colon \widetilde{T}]$ we can find an $m \in \mathbb{N}$ such that $k + m\widetilde{T} \in [\vartheta_R - \widetilde{T} + n \colon \vartheta_R + n - 1]$. Similarly to the previous case, (A.72) allows us to write

$$[\vartheta_R - \widetilde{T} + 1 \colon nL] \cup [(n-1)L + 1 \colon \vartheta_R - L] = \{\tilde{\jmath}_1, \ldots, \tilde{\jmath}_{\widetilde{T}}\}$$

where $\tilde{\jmath}_t \in \beta_1^{-1}(t)$ for $t \in [1 \colon \widetilde{T}]$. By (A.62), we then obtain

$$\begin{aligned} \beta_2([\vartheta_R - \widetilde{T} + 1 \colon \vartheta_R]) &= \beta_2([\vartheta_R - \widetilde{T} + 1 \colon nL] \cup [(n-1)L + 1 \colon \vartheta_R - L]) \\ &= \{\beta_2(\tilde{\jmath}_1), \ldots, \beta_2(\tilde{\jmath}_{\widetilde{T}})\}\,. \end{aligned}$$

By the same arguments as in the previous case, we find that assigning $g_t = \beta_2(\tilde{\jmath}_t)$ satisfies (A.59).

Appendix B

Proofs of Part II

B.1 Proof of Lemma 6.3

We first assume $m \geq 1$; the case $m = 0$ will be considered separately.

B.1.1 Case $m \geq 1$

By Definition 6.2, there exist Lipschitz functions f_k and bounded sets $\mathcal{A}_k \subseteq \mathbb{R}^m$ such that

$$\mathscr{H}^m\bigg(\mathcal{E} \setminus \bigcup_{k\in\mathbb{N}} f_k(\mathcal{A}_k)\bigg) = 0\,. \tag{B.1}$$

Proof of Property 1: We first note that $\mathcal{D} \setminus \bigcup_{k\in\mathbb{N}} f_k(\mathcal{A}_k) \subseteq \mathcal{E} \setminus \bigcup_{k\in\mathbb{N}} f_k(\mathcal{A}_k)$ and thus, by the monotonicity of measures, $\mathscr{H}^m\big(\mathcal{D} \setminus \bigcup_{k\in\mathbb{N}} f_k(\mathcal{A}_k)\big) \leq \mathscr{H}^m\big(\mathcal{E} \setminus \bigcup_{k\in\mathbb{N}} f_k(\mathcal{A}_k)\big) = 0$. Therefore, by Definition 6.2, $\mathcal{D}$ is m-rectifiable.

Proof of Property 2: We have to show that $\mathbb{R}^M$ is the countable union of sets of finite $\mathscr{H}^m|_{\mathcal{E}}$ measure. We have

$$\mathbb{R}^M = \bigcup_{k\in\mathbb{N}} f_k(\mathcal{A}_k) \cup \bigg(\mathbb{R}^M \setminus \bigcup_{k\in\mathbb{N}} f_k(\mathcal{A}_k)\bigg)\,. \tag{B.2}$$

Thus, if we can show that all sets on the right-hand side of (B.2) have finite $\mathscr{H}^m|_{\mathcal{E}}$ measure, we can conclude that $\mathscr{H}^m|_{\mathcal{E}}$ is σ-finite. To this end, we first consider the sets $f_k(\mathcal{A}_k)$. We have for all $k \in \mathbb{N}$

$$\begin{aligned}\mathscr{H}^m|_{\mathcal{E}}(f_k(\mathcal{A}_k)) &\leq \mathscr{H}^m(f_k(\mathcal{A}_k)) \\ &\overset{(a)}{\leq} (\mathrm{Lip}(f_k))^m \mathscr{L}^m(\mathcal{A}_k) \\ &\overset{(b)}{<} \infty \end{aligned} \tag{B.3}$$

where (a) holds because of [Ambrosio et al., 2000, Prop. 2.49(iv)] and (b) holds because bounded sets have finite Lebesgue measure. Furthermore,

$$\mathscr{H}^m|_{\mathcal{E}}\Big(\mathbb{R}^M \setminus \bigcup_{k\in\mathbb{N}} f_k(\mathcal{A}_k)\Big) = \mathscr{H}^m\Big(\mathcal{E} \setminus \bigcup_{k\in\mathbb{N}} f_k(\mathcal{A}_k)\Big) = 0 \tag{B.4}$$

where (B.1) was used. Hence, by (B.3) and (B.4), all sets on the right-hand side of (B.2) have finite $\mathscr{H}^m|_{\mathcal{E}}$ measure, which concludes the proof of Property 2.

Proof of Property 3: We consider the Lipschitz functions $\phi \circ f_k$. We have that $\phi(\mathcal{E}) \setminus \bigcup_{k\in\mathbb{N}} \phi(f_k(\mathcal{A}_k)) \subseteq \phi\big(\mathcal{E} \setminus \bigcup_{k\in\mathbb{N}} f_k(\mathcal{A}_k)\big)$. Thus, we obtain

$$\begin{aligned}\mathscr{H}^m\Big(\phi(\mathcal{E}) \setminus \bigcup_{k\in\mathbb{N}} \phi(f_k(\mathcal{A}_k))\Big) &\leq \mathscr{H}^m\Big(\phi\Big(\mathcal{E} \setminus \bigcup_{k\in\mathbb{N}} f_k(\mathcal{A}_k)\Big)\Big)\\ &\overset{(a)}{\leq} (\mathrm{Lip}(\phi))^m \mathscr{H}^m\Big(\mathcal{E} \setminus \bigcup_{k\in\mathbb{N}} f_k(\mathcal{A}_k)\Big)\\ &= 0\end{aligned}$$

where (a) holds because of [Ambrosio et al., 2000, Prop. 2.49(iv)]. Thus, by Definition 6.2, with the sets $\mathcal{A}_k$ and the Lipschitz functions $\phi \circ f_k$, the set $\phi(\mathcal{E})$ is m-rectifiable.

Proof of Property 4: By the σ-subadditivity of $\mathscr{H}^n$ [Ambrosio et al., 2000, Def. 1.2], we have

$$\mathscr{H}^n(\mathcal{E}) \leq \mathscr{H}^n\Big(\mathcal{E} \setminus \bigcup_{k\in\mathbb{N}} f_k(\mathcal{A}_k)\Big) + \sum_{k\in\mathbb{N}} \mathscr{H}^n(f_k(\mathcal{A}_k))\,. \tag{B.5}$$

To prove Property 4, i.e., $\mathscr{H}^n(\mathcal{E}) = 0$, it suffices to show that all the terms on the right-hand side of (B.5) are zero. We have

$$\mathscr{H}^n(f_k(\mathcal{A}_k)) \overset{(a)}{\leq} (\mathrm{Lip}(f_k))^n \mathscr{H}^n(\mathcal{A}_k) \overset{(b)}{=} 0$$

where (a) holds because of [Ambrosio et al., 2000, Prop. 2.49(iv)] and (b) holds because the Hausdorff measure $\mathscr{H}^n$ is identically zero on $\mathbb{R}^m$ [Ambrosio et al., 2000, Prop. 2.49(ii)]. Furthermore, by [Ambrosio et al., 2000, Prop. 2.49(iii)], $\mathscr{H}^m\big(\mathcal{E} \setminus \bigcup_{k\in\mathbb{N}} f_k(\mathcal{A}_k)\big) = 0$ (see (B.1)) implies $\mathscr{H}^n\big(\mathcal{E} \setminus \bigcup_{k\in\mathbb{N}} f_k(\mathcal{A}_k)\big) = 0$. Thus, all the terms on the right-hand side of (B.5) are zero and Property 4 holds.

Proof of Property 5: We have to show that there exists Lipschitz functions g_ℓ and bounded sets $\mathcal{B}_\ell$ such that

$$\mathscr{H}^m\Big(\bigcup_{i\in\mathbb{N}} \mathcal{E}_i \setminus \bigcup_{\ell\in\mathbb{N}} g_\ell(\mathcal{B}_\ell)\Big) = 0\,. \tag{B.6}$$

Now because $\mathcal{E}_i$ is m-rectifiable for each $i \in \mathbb{N}$, there exist Lipschitz functions $f_k^{(i)}$ and bounded sets $\mathcal{A}_k^{(i)} \subseteq \mathbb{R}^m$ such that

$$\mathscr{H}^m\Big(\mathcal{E}_i \setminus \bigcup_{k\in\mathbb{N}} f_k^{(i)}(\mathcal{A}_k^{(i)})\Big) = 0$$

for each $i \in \mathbb{N}$. Thus,

$$\mathscr{H}^m\bigg(\bigcup_{i\in\mathbb{N}} \mathcal{E}_i \setminus \bigcup_{i,k\in\mathbb{N}} f_k^{(i)}\big(\mathcal{A}_k^{(i)}\big)\bigg) \leq \sum_{i\in\mathbb{N}} \mathscr{H}^m\bigg(\mathcal{E}_i \setminus \bigcup_{k\in\mathbb{N}} f_k^{(i)}\big(\mathcal{A}_k^{(i)}\big)\bigg) = 0\,.$$

This implies that (B.6) is satisfied (with $g_\ell = f_k^{(i)}$ and $\mathcal{B}_\ell = \mathcal{A}_k^{(i)}$), and thus $\bigcup_{i\in\mathbb{N}} \mathcal{E}_i$ is m-rectifiable.

Proof of Property 6: We have to show that there exist Lipschitz functions f_k and bounded sets $\mathcal{A}_k$ such that

$$\mathscr{H}^m\bigg(\mathbb{R}^m \setminus \bigcup_{k\in\mathbb{N}} f_k\big(\mathcal{A}_k\big)\bigg) = 0\,. \tag{B.7}$$

We choose each $f_k = \mathrm{id}_{\mathbb{R}^m}$, i.e., equal to the identity on $\mathbb{R}^m$ (which is obviously Lipschitz continuous). Furthermore, we set $\mathcal{A}_k \triangleq \mathcal{B}_k(\mathbf{0})$. We obtain

$$\bigcup_{k\in\mathbb{N}} f_k\big(\mathcal{A}_k\big) = \bigcup_{k\in\mathbb{N}} \mathcal{B}_k(\mathbf{0}) = \mathbb{R}^m$$

and, hence, (B.7) holds.

B.1.2 Case $m = 0$

Recall that a 0-rectifiable set is countable and $\mathscr{H}^0$ is the counting measure. Property 1 follows because each subset of a countable set is countable. Property 2 holds because the counting measure of a countable set is σ-finite. Property 3 holds because the image of a countable set under any function is again countable. Property 4 can be shown as follows. By [Ambrosio et al., 2000, Prop. 2.49(iii)], every finite set $\mathcal{A}$ has Hausdorff measure $\mathscr{H}^n(\mathcal{A}) = 0$ for $n > 0$. By the σ-additivity of Hausdorff measures, this also holds for countable sets. Finally, Property 5 follows because the countable union of countable sets is again countable.

B.2 Proof of Lemma 6.5

We consider an arbitrary version of the Radon-Nikodym derivative $\frac{\mathrm{d}\mu}{\mathrm{d}\mathscr{H}^m|_{\mathcal{E}}}$ that is $\mathscr{H}^m$-measurable. Thus, the set

$$\mathcal{D} \triangleq \left(\frac{\mathrm{d}\mu}{\mathrm{d}\mathscr{H}^m|_{\mathcal{E}}}\right)^{-1}(\{0\}) = \left\{\boldsymbol{s} \in \mathbb{R}^M : \frac{\mathrm{d}\mu}{\mathrm{d}\mathscr{H}^m|_{\mathcal{E}}}(\boldsymbol{s}) = 0\right\}$$

is $\mathscr{H}^m$-measurable. We set $\widetilde{\mathcal{E}} \triangleq \mathcal{E} \setminus \mathcal{D} = \mathcal{E} \cap \mathcal{D}^c$ and note that $\widetilde{\mathcal{E}}$ is, as the intersection of the $\mathscr{H}^m$-measurable sets $\mathcal{E}$ and $\mathcal{D}^c$, again $\mathscr{H}^m$-measurable. Thus, by Property 1 in Lemma 6.3, the set $\widetilde{\mathcal{E}} \subseteq \mathcal{E}$ is m-rectifiable. We have for any μ-measurable set $\mathcal{B}$

$$\mu(\mathcal{B}) = \int_{\mathcal{B}} \frac{\mathrm{d}\mu}{\mathrm{d}\mathscr{H}^m|_{\mathcal{E}}}\,\mathrm{d}\mathscr{H}^m|_{\mathcal{E}}$$

$$
\begin{aligned}
&= \int_{\mathcal{B}\cap\mathcal{E}} \frac{\mathrm{d}\mu}{\mathrm{d}\mathscr{H}^m|_{\mathcal{E}}} \, \mathrm{d}\mathscr{H}^m \\
&\overset{(a)}{=} \int_{\mathcal{B}\cap\widetilde{\mathcal{E}}} \frac{\mathrm{d}\mu}{\mathrm{d}\mathscr{H}^m|_{\mathcal{E}}} \, \mathrm{d}\mathscr{H}^m \\
&= \int_{\mathcal{B}} \frac{\mathrm{d}\mu}{\mathrm{d}\mathscr{H}^m|_{\mathcal{E}}} \, \mathrm{d}\mathscr{H}^m|_{\widetilde{\mathcal{E}}} \qquad \text{(B.8)}
\end{aligned}
$$

where (a) holds because $\frac{\mathrm{d}\mu}{\mathrm{d}\mathscr{H}^m|_{\mathcal{E}}} = 0$ on $\mathcal{D}$ and $\mathcal{E} \setminus \widetilde{\mathcal{E}} = \mathcal{E} \setminus (\mathcal{E} \setminus \mathcal{D}) = \mathcal{E} \cap (\mathcal{E} \cap \mathcal{D}^c)^c = \mathcal{E} \cap (\mathcal{E}^c \cup \mathcal{D}) = \mathcal{E} \cap \mathcal{D} \subseteq \mathcal{D}$.

Proof of Property 1: By (B.8), $\mathscr{H}^m|_{\widetilde{\mathcal{E}}}(\mathcal{B}) = 0$ implies $\mu(\mathcal{B}) = 0$, i.e., we have $\mu \ll \mathscr{H}^m|_{\widetilde{\mathcal{E}}}$.

Proof of Property 2: Again, by (B.8), $\frac{\mathrm{d}\mu}{\mathrm{d}\mathscr{H}^m|_{\mathcal{E}}}$ is a version of the Radon Nikodym derivative $\frac{\mathrm{d}\mu}{\mathrm{d}\mathscr{H}^m|_{\widetilde{\mathcal{E}}}}$, i.e., $\frac{\mathrm{d}\mu}{\mathrm{d}\mathscr{H}^m|_{\mathcal{E}}} = \frac{\mathrm{d}\mu}{\mathrm{d}\mathscr{H}^m|_{\widetilde{\mathcal{E}}}}$ $\mathscr{H}^m|_{\widetilde{\mathcal{E}}}$-almost everywhere.

Proof of Property 3: Because $\frac{\mathrm{d}\mu}{\mathrm{d}\mathscr{H}^m|_{\mathcal{E}}} > 0$ on $\mathcal{D}^c$ and $\widetilde{\mathcal{E}} \subseteq \mathcal{D}^c$, we also have $\frac{\mathrm{d}\mu}{\mathrm{d}\mathscr{H}^m|_{\widetilde{\mathcal{E}}}} > 0$ $\mathscr{H}^m|_{\widetilde{\mathcal{E}}}$-almost everywhere.

B.3 Proof of Theorem 6.11

Proof of Statement 1: Let x be 0-rectifiable with support $\mathcal{E}$, i.e., $\mu\mathsf{x}^{-1} \ll \mathscr{H}^0|_{\mathcal{E}}$ for a 0-rectifiable set $\mathcal{E}$. Recall that a 0-rectifiable set $\mathcal{E}$ is by definition countable, i.e., $\mathcal{E} = \{\boldsymbol{x}_i : i \in \mathcal{I}\}$ for a countable index set $\mathcal{I}$. By (6.13), $\Pr\{\mathsf{x} \in \mathcal{E}\} = 1$, which implies that x is a discrete random variable. Finally,

$$
\begin{aligned}
p_{\mathsf{x}}(\boldsymbol{x}_i) = \Pr\{\mathsf{x} = \boldsymbol{x}_i\} &\overset{(6.8)}{=} \mu\mathsf{x}^{-1}(\{\boldsymbol{x}_i\}) \\
&= \int_{\{\boldsymbol{x}_i\}} \frac{\mathrm{d}\mu\mathsf{x}^{-1}}{\mathrm{d}\mathscr{H}^0|_{\mathcal{E}}}(\boldsymbol{x}) \, \mathrm{d}\mathscr{H}^0|_{\mathcal{E}}(\boldsymbol{x}) \overset{(a)}{=} \frac{\mathrm{d}\mu\mathsf{x}^{-1}}{\mathrm{d}\mathscr{H}^0|_{\mathcal{E}}}(\boldsymbol{x}_i) \overset{(6.10)}{=} \theta^0_{\mathsf{x}}(\boldsymbol{x}_i)
\end{aligned}
$$

where (a) holds because $\mathscr{H}^0$ is the counting measure.

Conversely, let x be a discrete random variable taking on the values $\boldsymbol{x}_i$, $i \in \mathcal{I}$. We set $\mathcal{E} \triangleq \{\boldsymbol{x}_i : i \in \mathcal{I}\}$, which is countable and, thus, 0-rectifiable. Because $\mathcal{E}$ includes all possible values of x, we have $\Pr\{\mathsf{x} \in \mathcal{E}^c\} = \mu\mathsf{x}^{-1}(\mathcal{E}^c) = 0$. For $\mathcal{A} \subseteq \mathbb{R}^M$, the measure $\mathscr{H}^0|_{\mathcal{E}}(\mathcal{A})$ counts the number of points in $\mathcal{A}$ that also belong to $\mathcal{E}$. Thus, for any set $\mathcal{A}$ such that $\mathscr{H}^0|_{\mathcal{E}}(\mathcal{A}) = 0$, we obtain that $\mathcal{A} \cap \mathcal{E} = \emptyset$ and hence $\mathcal{A} \subseteq \mathcal{E}^c$. This implies $\mu\mathsf{x}^{-1}(\mathcal{A}) \leq \mu\mathsf{x}^{-1}(\mathcal{E}^c) = 0$. Thus, we showed that $\mu\mathsf{x}^{-1}(\mathcal{A}) = 0$ for any set $\mathcal{A}$ with $\mathscr{H}^0|_{\mathcal{E}}(\mathcal{A}) = 0$, i.e., $\mu\mathsf{x}^{-1} \ll \mathscr{H}^0|_{\mathcal{E}}$. Hence, x is 0-rectifiable.

Proof of Statement 2: Let x be M-rectifiable on $\mathbb{R}^M$, i.e., $\mu\mathsf{x}^{-1} \ll \mathscr{H}^M|_{\mathcal{E}}$ for an M-rectifiable set $\mathcal{E}$. Because $\mathscr{H}^M$ is equal to the Lebesgue measure $\mathscr{L}^M$, we obtain $\mu\mathsf{x}^{-1} \ll \mathscr{L}^M|_{\mathcal{E}} \ll \mathscr{L}^M$. Thus, by the Radon-Nikodym theorem, there exists the Radon-Nikodym derivative $f_{\mathsf{x}} = \frac{\mathrm{d}\mu\mathsf{x}^{-1}}{\mathrm{d}\mathscr{L}^M}$ satisfying $\Pr\{\mathsf{x} \in \mathcal{A}\} = \int_{\mathcal{A}} f_{\mathsf{x}}(\boldsymbol{x}) \, \mathrm{d}\mathscr{L}^M(\boldsymbol{x})$ for any measurable $\mathcal{A} \subseteq \mathbb{R}^M$, i.e., x is a continuous random variable. By (6.10), $\theta^M_{\mathsf{x}} = f_{\mathsf{x}}$ $\mathscr{L}^M$-almost everywhere.

Conversely, let x be a continuous random variable on $\mathbb{R}^M$ with probability density function f_{x}. For a measurable set $\mathcal{A} \subseteq \mathbb{R}^M$ satisfying $\mathscr{L}^M(\mathcal{A}) = 0$, we obtain $\mu\mathsf{x}^{-1}(\mathcal{A}) = \Pr\{\mathsf{x} \in \mathcal{A}\} =$

$\int_{\mathcal{A}} f_{\mathsf{x}}(\boldsymbol{x})\,\mathrm{d}\mathscr{L}^M(\boldsymbol{x}) = 0$. Thus, we have $\mu\mathsf{x}^{-1} \ll \mathscr{L}^M$. Because $\mathscr{L}^M = \mathscr{H}^M = \mathscr{H}^M|_{\mathbb{R}^M}$, this is equivalent to $\mu\mathsf{x}^{-1} \ll \mathscr{H}^M|_{\mathbb{R}^M}$. Because, by Property 6 in Lemma 6.3 $\mathbb{R}^M$ is M-rectifiable, it follows from Definition 6.9 that x is an M-rectifiable random variable.

B.4 Proof of Theorem 6.16

We first note that the set $\phi(\mathcal{E})$ is m-rectifiable because $\mathcal{E}$ is m-rectifiable and because of Property 3 in Lemma 6.3. To prove that y is m-rectifiable, we will show that $\mu\mathsf{y}^{-1} \ll \mathscr{H}^m|_{\phi(\mathcal{E})}$. For a measurable set $\mathcal{A} \subseteq \mathbb{R}^M$, we have

$$\begin{aligned}
\mu\mathsf{y}^{-1}(\mathcal{A}) &= \Pr\{\phi(\mathsf{x}) \in \mathcal{A}\} \\
&= \Pr\{\mathsf{x} \in \phi^{-1}(\mathcal{A})\} \\
&\overset{(6.11)}{=} \int_{\phi^{-1}(\mathcal{A})} \theta^m_{\mathsf{x}}(\boldsymbol{x})\,\mathrm{d}\mathscr{H}^m|_{\mathcal{E}}(\boldsymbol{x}) \\
&= \int_{\phi^{-1}(\mathcal{A})\cap\mathcal{E}} \frac{\theta^m_{\mathsf{x}}(\boldsymbol{x})}{\mathscr{J}^{\mathcal{E}}_{\phi}(\boldsymbol{y})} \mathscr{J}^{\mathcal{E}}_{\phi}(\boldsymbol{x})\,\mathrm{d}\mathscr{H}^m(\boldsymbol{x}) \\
&\overset{(a)}{=} \int_{\mathcal{A}\cap\phi(\mathcal{E})} \frac{\theta^m_{\mathsf{x}}(\phi^{-1}(\boldsymbol{y}))}{\mathscr{J}^{\mathcal{E}}_{\phi}(\phi^{-1}(\boldsymbol{y}))}\,\mathrm{d}\mathscr{H}^m(\boldsymbol{y}) \\
&= \int_{\mathcal{A}} \frac{\theta^m_{\mathsf{x}}(\phi^{-1}(\boldsymbol{y}))}{\mathscr{J}^{\mathcal{E}}_{\phi}(\phi^{-1}(\boldsymbol{y}))}\,\mathrm{d}\mathscr{H}^m|_{\phi(\mathcal{E})}(\boldsymbol{y})\,.
\end{aligned} \tag{B.9}$$

Here, (a) holds because of the generalized area formula [Ambrosio et al., 2000, Th. 2.91], and $\phi^{-1}\colon \phi(\mathcal{E}) \to \mathcal{E}$ is well defined because ϕ is one-to-one on $\mathcal{E}$. For a measurable set $\mathcal{A} \subseteq \mathbb{R}^M$ satisfying $\mathscr{H}^m|_{\phi(\mathcal{E})}(\mathcal{A}) = 0$, (B.9) implies $\mu\mathsf{y}^{-1}(\mathcal{A}) = 0$, i.e., $\mu\mathsf{y}^{-1} \ll \mathscr{H}^m|_{\phi(\mathcal{E})}$. Thus, y is an m-rectifiable random variable.

By (B.9), $\frac{\theta^m_{\mathsf{x}}(\phi^{-1}(\boldsymbol{y}))}{\mathscr{J}^{\mathcal{E}}_{\phi}(\phi^{-1}(\boldsymbol{y}))}$ coincides with the Radon-Nikodym derivative $\frac{\mathrm{d}\mu\mathsf{y}^{-1}}{\mathrm{d}\mathscr{H}^m|_{\phi(\mathcal{E})}}$, and thus, we obtain

$$\frac{\theta^m_{\mathsf{x}}(\phi^{-1}(\boldsymbol{y}))}{\mathscr{J}^{\mathcal{E}}_{\phi}(\phi^{-1}(\boldsymbol{y}))} = \frac{\mathrm{d}\mu\mathsf{y}^{-1}}{\mathrm{d}\mathscr{H}^m|_{\phi(\mathcal{E})}}(\boldsymbol{y}) \overset{(6.10)}{=} \theta^m_{\mathsf{y}}(\boldsymbol{y}) \tag{B.10}$$

$\mathscr{H}^m|_{\phi(\mathcal{E})}$-almost everywhere. We conclude

$$\begin{aligned}
\mathfrak{h}^m(\mathsf{y}) &= -\int_{\phi(\mathcal{E})} \theta^m_{\mathsf{y}}(\boldsymbol{y}) \log \theta^m_{\mathsf{y}}(\boldsymbol{y})\,\mathrm{d}\mathscr{H}^m(\boldsymbol{y}) \\
&\overset{(B.10)}{=} -\int_{\phi(\mathcal{E})} \frac{\theta^m_{\mathsf{x}}(\phi^{-1}(\boldsymbol{y}))}{\mathscr{J}^{\mathcal{E}}_{\phi}(\phi^{-1}(\boldsymbol{y}))} \log\left(\frac{\theta^m_{\mathsf{x}}(\phi^{-1}(\boldsymbol{y}))}{\mathscr{J}^{\mathcal{E}}_{\phi}(\phi^{-1}(\boldsymbol{y}))}\right) \mathrm{d}\mathscr{H}^m(\boldsymbol{y}) \\
&\overset{(a)}{=} -\int_{\mathcal{E}} \frac{\theta^m_{\mathsf{x}}(\boldsymbol{x})}{\mathscr{J}^{\mathcal{E}}_{\phi}(\boldsymbol{x})} \log\left(\frac{\theta^m_{\mathsf{x}}(\boldsymbol{x})}{\mathscr{J}^{\mathcal{E}}_{\phi}(\boldsymbol{x})}\right) \mathscr{J}^{\mathcal{E}}_{\phi}(\boldsymbol{x})\,\mathrm{d}\mathscr{H}^m(\boldsymbol{x}) \\
&= -\int_{\mathcal{E}} \theta^m_{\mathsf{x}}(\boldsymbol{x}) \log \theta^m_{\mathsf{x}}(\boldsymbol{x})\,\mathrm{d}\mathscr{H}^m(\boldsymbol{x}) + \int_{\mathcal{E}} \theta^m_{\mathsf{x}}(\boldsymbol{x}) \log \mathscr{J}^{\mathcal{E}}_{\phi}(\boldsymbol{x})\,\mathrm{d}\mathscr{H}^m(\boldsymbol{x}) \\
&= \mathfrak{h}^m(\mathsf{x}) + \mathbb{E}_{\mathsf{x}}[\log \mathscr{J}^{\mathcal{E}}_{\phi}(\mathsf{x})]
\end{aligned}$$

where (a) holds because of the generalized area formula [Ambrosio et al., 2000, Th. 2.91].

B.5 Proof of Lemma 7.2

We first prove product-compatibility of $\mathcal{E}_1$ and $\mathcal{E}_2$ under the individual Conditions 1–3 for the case $m_1, m_2 \geq 1$. Subsequently, we will prove that $\mathcal{E}_1$ and $\mathcal{E}_2$ are always product-compatible if $m_1 = 0$, which covers all cases where at least one random variable is discrete.

By [Federer, 1969, Th. 3.2.23], Condition 1 implies that $\mathcal{E}_1$ and $\mathcal{E}_2$ are product-compatible.

To prove product-compatibility under Condition 2, we first construct a sequence of mutually disjoint sets $\widetilde{\mathcal{E}}_1^{(k)}$ whose union equals that of the sets $\mathcal{E}_1^{(k)}$, i.e., $\bigcup_{k\in\mathbb{N}} \widetilde{\mathcal{E}}_1^{(k)} = \bigcup_{k\in\mathbb{N}} \mathcal{E}_1^{(k)}$. Let $\widetilde{\mathcal{E}}_1^{(1)} \triangleq \mathcal{E}_1^{(1)}$ and define inductively $\widetilde{\mathcal{E}}_1^{(k)} \triangleq \mathcal{E}_1^{(k)} \setminus \bigcup_{i=1}^{k-1} \widetilde{\mathcal{E}}_1^{(i)}$ for $k \geq 2$. All sets $\widetilde{\mathcal{E}}_1^{(k)}$ are Borel sets and $\mathscr{H}^{m_1}(\widetilde{\mathcal{E}}_1^{(k)}) \leq \mathscr{H}^{m_1}(\mathcal{E}_1^{(k)})$. Similarly, we construct a sequence of mutually disjoint Borel sets $\widetilde{\mathcal{E}}_2^{(\ell)}$ by defining $\widetilde{\mathcal{E}}_2^{(1)} \triangleq \mathcal{E}_2^{(1)}$ and $\widetilde{\mathcal{E}}_2^{(\ell)} \triangleq \mathcal{E}_2^{(\ell)} \setminus \bigcup_{i=1}^{\ell-1} \widetilde{\mathcal{E}}_2^{(i)}$ for $\ell \geq 2$. Recall that $\mathscr{H}^{m_1}(\mathcal{E}_1^{(k)}) < \infty$, which implies

$$\mathscr{H}^{m_1}(\widetilde{\mathcal{E}}_1^{(k)}) \leq \mathscr{H}^{m_1}(\mathcal{E}_1^{(k)}) < \infty \,. \tag{B.11}$$

Furthermore, recall that $\mathcal{E}_2^{(\ell)} = \phi_\ell(\mathcal{A}_\ell)$ with Lipschitz functions $\phi_\ell \colon \mathbb{R}^{m_2} \to \mathbb{R}^{M_2}$ and bounded sets $\mathcal{A}_\ell \subseteq \mathbb{R}^{m_2}$. Thus, $\widetilde{\mathcal{E}}_2^{(\ell)} \subseteq \phi_\ell(\mathcal{A}_\ell)$, which implies $\widetilde{\mathcal{E}}_2^{(\ell)} = \phi_\ell\big(\mathcal{A}_\ell \cap \phi_\ell^{-1}(\widetilde{\mathcal{E}}_2^{(\ell)})\big)$. By (B.11) and because the set $\mathcal{A}_\ell \cap \phi_\ell^{-1}(\widetilde{\mathcal{E}}_2^{(\ell)})$ is bounded, the sets $\widetilde{\mathcal{E}}_1^{(k)}$ and $\widetilde{\mathcal{E}}_2^{(\ell)}$ satisfy the assumptions in Condition 1 and we obtain (cf. (7.3))

$$\mathscr{H}^{m_1+m_2}|_{\widetilde{\mathcal{E}}_1^{(k)} \times \widetilde{\mathcal{E}}_2^{(\ell)}} = \mathscr{H}^{m_1}|_{\widetilde{\mathcal{E}}_1^{(k)}} \times \mathscr{H}^{m_2}|_{\widetilde{\mathcal{E}}_2^{(\ell)}} \,. \tag{B.12}$$

We then have for any measurable sets $\mathcal{B}_1 \subseteq \mathbb{R}^{M_1}$ and $\mathcal{B}_2 \subseteq \mathbb{R}^{M_2}$

$$\begin{aligned}
\mathscr{H}^{m_1+m_2}|_{\mathcal{E}_1\times\mathcal{E}_2}(\mathcal{B}_1 \times \mathcal{B}_2) &= \mathscr{H}^{m_1+m_2}\big((\mathcal{B}_1 \times \mathcal{B}_2) \cap (\mathcal{E}_1 \times \mathcal{E}_2)\big) \\
&= \mathscr{H}^{m_1+m_2}\left((\mathcal{B}_1 \times \mathcal{B}_2) \cap \left(\bigcup_{k\in\mathbb{N}} \mathcal{E}_1^{(k)} \times \bigcup_{\ell\in\mathbb{N}} \mathcal{E}_2^{(\ell)}\right)\right) \\
&= \mathscr{H}^{m_1+m_2}\left((\mathcal{B}_1 \times \mathcal{B}_2) \cap \left(\bigcup_{k\in\mathbb{N}} \widetilde{\mathcal{E}}_1^{(k)} \times \bigcup_{\ell\in\mathbb{N}} \widetilde{\mathcal{E}}_2^{(\ell)}\right)\right) \\
&= \mathscr{H}^{m_1+m_2}\left((\mathcal{B}_1 \times \mathcal{B}_2) \cap \left(\bigcup_{k,\ell\in\mathbb{N}} \big(\widetilde{\mathcal{E}}_1^{(k)} \times \widetilde{\mathcal{E}}_2^{(\ell)}\big)\right)\right) \\
&= \mathscr{H}^{m_1+m_2}\left(\bigcup_{k,\ell\in\mathbb{N}} \Big((\mathcal{B}_1 \times \mathcal{B}_2) \cap \big(\widetilde{\mathcal{E}}_1^{(k)} \times \widetilde{\mathcal{E}}_2^{(\ell)}\big)\Big)\right) \\
&\overset{(a)}{=} \sum_{k,\ell\in\mathbb{N}} \mathscr{H}^{m_1+m_2}\Big((\mathcal{B}_1 \times \mathcal{B}_2) \cap \big(\widetilde{\mathcal{E}}_1^{(k)} \times \widetilde{\mathcal{E}}_2^{(\ell)}\big)\Big) \\
&= \sum_{k,\ell\in\mathbb{N}} \mathscr{H}^{m_1+m_2}|_{\widetilde{\mathcal{E}}_1^{(k)} \times \widetilde{\mathcal{E}}_2^{(\ell)}}(\mathcal{B}_1 \times \mathcal{B}_2)
\end{aligned}$$

$$
\begin{aligned}
&\overset{\text{(B.12)}}{=} \sum_{k,\ell\in\mathbb{N}} \big(\mathscr{H}^{m_1}|_{\widetilde{\mathcal{E}}_1^{(k)}} \times \mathscr{H}^{m_2}|_{\widetilde{\mathcal{E}}_2^{(\ell)}}\big)(\mathcal{B}_1 \times \mathcal{B}_2)\\
&= \sum_{k,\ell\in\mathbb{N}} \mathscr{H}^{m_1}\big(\mathcal{B}_1 \cap \widetilde{\mathcal{E}}_1^{(k)}\big)\, \mathscr{H}^{m_2}\big(\mathcal{B}_2 \cap \widetilde{\mathcal{E}}_2^{(\ell)}\big)\\
&= \Bigg(\sum_{k\in\mathbb{N}} \mathscr{H}^{m_1}\big(\mathcal{B}_1 \cap \widetilde{\mathcal{E}}_1^{(k)}\big)\Bigg)\Bigg(\sum_{\ell\in\mathbb{N}} \mathscr{H}^{m_2}\big(\mathcal{B}_2 \cap \widetilde{\mathcal{E}}_2^{(\ell)}\big)\Bigg)\\
&= \mathscr{H}^{m_1}\Bigg(\mathcal{B}_1 \cap \bigcup_{k\in\mathbb{N}} \widetilde{\mathcal{E}}_1^{(k)}\Bigg)\, \mathscr{H}^{m_2}\Bigg(\mathcal{B}_2 \cap \bigcup_{\ell\in\mathbb{N}} \widetilde{\mathcal{E}}_2^{(\ell)}\Bigg)\\
&= \mathscr{H}^{m_1}(\mathcal{B}_1 \cap \mathcal{E}_1)\, \mathscr{H}^{m_2}(\mathcal{B}_2 \cap \mathcal{E}_2)\\
&= \mathscr{H}^{m_1}|_{\mathcal{E}_1}(\mathcal{B}_1)\, \mathscr{H}^{m_2}|_{\mathcal{E}_2}(\mathcal{B}_2)
\end{aligned}
$$

where (a) holds by the σ-additivity of measures and the disjointness of the sets $\widetilde{\mathcal{E}}_1^{(k)} \times \widetilde{\mathcal{E}}_2^{(\ell)}$. By the uniqueness of the product measure of σ-finite measures, we obtain $\mathscr{H}^{m_1+m_2}|_{\mathcal{E}_1\times\mathcal{E}_2} = \mathscr{H}^{m_1}|_{\mathcal{E}_1} \times \mathscr{H}^{m_2}|_{\mathcal{E}_2}$. Hence, according to Definition 7.1, $\mathcal{E}_1$ and $\mathcal{E}_2$ are product-compatible.

Condition 3 is a special case of Condition 2. Because $m_2 = M_2$, we can simply choose the ϕ_ℓ in Condition 2 as the identity on $\mathbb{R}^{M_2}$ and $\mathcal{A}_\ell = \mathcal{E}_2 \cap \mathcal{B}_\ell(\mathbf{0})$, i.e., the intersection of $\mathcal{E}_2$ with spheres of radius ℓ centered at $\mathbf{0}$. We then obtain $\mathcal{E}_2^{(\ell)} \triangleq \phi_\ell(\mathcal{A}_\ell) = \mathcal{A}_\ell$ and, further, $\bigcup_{\ell\in\mathbb{N}} \mathcal{E}_2^{(\ell)} = \bigcup_{\ell\in\mathbb{N}} \mathcal{A}_\ell = \mathcal{E}_2$. Because each $\mathcal{A}_\ell$ is bounded and $\mathcal{E}_2^{(\ell)}$ is Borel (being the intersection of the Borel sets $\mathcal{E}_2$ and $\mathcal{B}_\ell(\mathbf{0})$), all assumptions in Condition 2 are satisfied and, thus, $\mathcal{E}_1$ and $\mathcal{E}_2$ are product-compatible.

Finally, we turn to Condition 4, i.e., the case $m_1 = 0$. Let $\mathcal{B}_1 \times \mathcal{B}_2 \subseteq \mathbb{R}^{M_1+M_2}$ be a Borel rectangle. Because for $m_1 = 0$ the set $\mathcal{E}_1$ is countable (see Definition 6.2), the set $\mathcal{B}_1 \cap \mathcal{E}_1$ is countable and a sum $\sum_{\boldsymbol{x}\in\mathcal{B}_1\cap\mathcal{E}_1} a_{\boldsymbol{x}}$ with nonnegative summands $a_{\boldsymbol{x}}$ is always well defined. Furthermore, $(\mathcal{B}_1 \times \mathcal{B}_2) \cap (\mathcal{E}_1 \times \mathcal{E}_2) = \bigcup_{\boldsymbol{x}\in\mathcal{B}_1\cap\mathcal{E}_1} \big(\{\boldsymbol{x}\} \times (\mathcal{B}_2 \cap \mathcal{E}_2)\big)$. Thus,

$$
\begin{aligned}
\mathscr{H}^{m_2}|_{\mathcal{E}_1\times\mathcal{E}_2}(\mathcal{B}_1 \times \mathcal{B}_2) &= \mathscr{H}^{m_2}\big((\mathcal{B}_1 \times \mathcal{B}_2) \cap (\mathcal{E}_1 \times \mathcal{E}_2)\big)\\
&= \mathscr{H}^{m_2}\Bigg(\bigcup_{\boldsymbol{x}\in\mathcal{B}_1\cap\mathcal{E}_1} \big(\{\boldsymbol{x}\} \times (\mathcal{B}_2 \cap \mathcal{E}_2)\big)\Bigg)\\
&\overset{(a)}{=} \sum_{\boldsymbol{x}\in\mathcal{B}_1\cap\mathcal{E}_1} \mathscr{H}^{m_2}\big(\{\boldsymbol{x}\} \times (\mathcal{B}_2 \cap \mathcal{E}_2)\big)\\
&\overset{(b)}{=} \sum_{\boldsymbol{x}\in\mathcal{B}_1\cap\mathcal{E}_1} \mathscr{H}^{m_2}(\mathcal{B}_2 \cap \mathcal{E}_2)\\
&= \sum_{\boldsymbol{x}\in\mathcal{B}_1\cap\mathcal{E}_1} \mathscr{H}^{m_2}|_{\mathcal{E}_2}(\mathcal{B}_2)\\
&\overset{(c)}{=} \mathscr{H}^{0}|_{\mathcal{E}_1}(\mathcal{B}_1)\, \mathscr{H}^{m_2}|_{\mathcal{E}_2}(\mathcal{B}_2)
\end{aligned}
$$

where (a) holds by the σ-additivity of the Hausdorff measure, (b) holds because the Hausdorff measure does not depend on the ambient space [Ambrosio et al., 2000, Remark 2.48], and (c) holds because $\mathscr{H}^0$ is the counting measure and thus $\mathscr{H}^0|_{\mathcal{E}_1}(\mathcal{B}_1)$ equals the number of elements in $\mathcal{B}_1 \cap \mathcal{E}_1$. By the uniqueness of the product measure of σ-finite measures, we obtain $\mathscr{H}^{m_2}|_{\mathcal{E}_1\times\mathcal{E}_2} = \mathscr{H}^0|_{\mathcal{E}_1} \times \mathscr{H}^{m_2}|_{\mathcal{E}_2}$, which is (7.3) for $m_1 = 0$.

B.6 Proof of Lemma 7.3

We consider the cases $m_1 = m_2 = 0$; $m_1, m_2 \geq 1$; and $m_1 = 0, m_2 \geq 1$ separately (due to symmetry, it is not necessary to consider the case $m_2 = 0, m_1 \geq 1$).

Case $m_1 = m_2 = 0$: By Definition 6.2, the sets $\mathcal{E}_1$ and $\mathcal{E}_2$ are countable. Because the product of countable sets is countable, the set $\mathcal{E}_1 \times \mathcal{E}_2$ is countable, i.e., 0-rectifiable.

Case $m_1, m_2 \geq 1$: By Definition 6.2, there exist bounded sets $\mathcal{A}_k \subseteq \mathbb{R}^{m_1}$ and $\mathcal{B}_\ell \subseteq \mathbb{R}^{m_2}$ and Lipschitz functions $\phi_k \colon \mathcal{A}_k \to \mathbb{R}^{M_1}$ and $\psi_\ell \colon \mathcal{B}_\ell \to \mathbb{R}^{M_2}$, for $k, \ell \in \mathbb{N}$, satisfying

$$\mathscr{H}^{m_1}\bigg(\mathcal{E}_1 \setminus \bigcup_{k\in\mathbb{N}} \phi_k(\mathcal{A}_k)\bigg) = 0 \tag{B.13}$$

and

$$\mathscr{H}^{m_2}\bigg(\mathcal{E}_2 \setminus \bigcup_{\ell\in\mathbb{N}} \psi_\ell(\mathcal{B}_\ell)\bigg) = 0 \tag{B.14}$$

respectively. The sets $\mathcal{A}_k \times \mathcal{B}_\ell \subseteq \mathbb{R}^{m_1+m_2}$ are again bounded and the functions $(\phi_k, \psi_\ell) \colon \mathcal{A}_k \times \mathcal{B}_\ell \to \mathbb{R}^{M_1+M_2}$ are Lipschitz. Furthermore, we have

$$\begin{aligned}
&\mathscr{H}^{m_1+m_2}\bigg((\mathcal{E}_1 \times \mathcal{E}_2) \setminus \bigcup_{k,\ell\in\mathbb{N}} (\phi_k, \psi_\ell)(\mathcal{A}_k \times \mathcal{B}_\ell)\bigg) \\
&\quad = \mathscr{H}^{m_1+m_2}\bigg((\mathcal{E}_1 \times \mathcal{E}_2) \setminus \bigg(\bigcup_{k\in\mathbb{N}} \phi_k(\mathcal{A}_k) \times \bigcup_{\ell\in\mathbb{N}} \psi_\ell(\mathcal{B}_\ell)\bigg)\bigg) \\
&\quad \overset{(a)}{=} \mathscr{H}^{m_1+m_2}\bigg(\bigg(\mathcal{E}_1 \times \bigg(\mathcal{E}_2 \setminus \bigcup_{\ell\in\mathbb{N}} \psi_\ell(\mathcal{B}_\ell)\bigg)\bigg) \cup \bigg(\bigg(\mathcal{E}_1 \setminus \bigcup_{k\in\mathbb{N}} \phi_k(\mathcal{A}_k)\bigg) \times \mathcal{E}_2\bigg)\bigg) \\
&\quad \leq \mathscr{H}^{m_1+m_2}\bigg(\mathcal{E}_1 \times \bigg(\mathcal{E}_2 \setminus \bigcup_{\ell\in\mathbb{N}} \psi_\ell(\mathcal{B}_\ell)\bigg)\bigg) + \mathscr{H}^{m_1+m_2}\bigg(\bigg(\mathcal{E}_1 \setminus \bigcup_{k\in\mathbb{N}} \phi_k(\mathcal{A}_k)\bigg) \times \mathcal{E}_2\bigg) \\
&\quad \overset{(b)}{=} \mathscr{H}^{m_1}(\mathcal{E}_1)\mathscr{H}^{m_2}\bigg(\mathcal{E}_2 \setminus \bigcup_{\ell\in\mathbb{N}} \psi_\ell(\mathcal{B}_\ell)\bigg) + \mathscr{H}^{m_1}\bigg(\mathcal{E}_1 \setminus \bigcup_{k\in\mathbb{N}} \phi_k(\mathcal{A}_k)\bigg)\mathscr{H}^{m_2}(\mathcal{E}_2) \\
&\quad \overset{(c)}{=} 0
\end{aligned}$$

where (a) holds because for arbitrary sets $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \mathcal{F}_4$

$$(\mathcal{F}_1 \times \mathcal{F}_2) \setminus (\mathcal{F}_3 \times \mathcal{F}_4) = \big(\mathcal{F}_1 \times (\mathcal{F}_2 \setminus \mathcal{F}_4)\big) \cup \big((\mathcal{F}_1 \setminus \mathcal{F}_3) \times \mathcal{F}_2\big)$$

(b) holds because $\mathcal{E}_1$ and $\mathcal{E}_2$ are product-compatible (see (7.4)), and (c) is obtained by inserting (B.13) and (B.14). Thus, according to Definition 6.2, $\mathcal{E}_1 \times \mathcal{E}_2$ is $(m_1 + m_2)$-rectifiable.

Case $m_1 = 0, m_2 \geq 1$: Because $\mathcal{E}_1$ is countable, the product $\mathcal{E}_1 \times \mathcal{E}_2$ is simply the countable union

$$\mathcal{E}_1 \times \mathcal{E}_2 = \bigcup_{\boldsymbol{x}\in\mathcal{E}_1} \{\boldsymbol{x}\} \times \mathcal{E}_2 \,.$$

Here, each set $\{\boldsymbol{x}\} \times \mathcal{E}_2$ is, by Property 3 in Lemma 6.3, m_2-rectifiable because it is the Lipschitz image of the m_2-rectifiable set $\mathcal{E}_2$ under the embedding $\phi_{\boldsymbol{x}}\colon \mathbb{R}^{M_2} \to \mathbb{R}^{M_1+M_2}$, $\phi_{\boldsymbol{x}}(\boldsymbol{y}) = (\boldsymbol{x}, \boldsymbol{y})$. Thus, $\mathcal{E}_1 \times \mathcal{E}_2$ is the countable union of m_2-rectifiable sets and, hence, by Property 5 in Lemma 6.3, again m_2-rectifiable.

B.7 Proof of Theorem 7.4

Proof of Property 1: We first show that for any $\mu(\mathsf{x},\mathsf{y})^{-1}$-measurable set $\mathcal{A} \subseteq \mathbb{R}^{M_1+M_2}$

$$\mu(\mathsf{x},\mathsf{y})^{-1}(\mathcal{A}) = \Pr\{(\mathsf{x},\mathsf{y}) \in \mathcal{A}\} = \int_{\mathcal{A}} \theta^{m_1}_{\mathsf{x}}(\boldsymbol{x})\theta^{m_2}_{\mathsf{y}}(\boldsymbol{y})\, \mathrm{d}\mathscr{H}^{m_1+m_2}|_{\mathcal{E}_1\times\mathcal{E}_2}(\boldsymbol{x},\boldsymbol{y})\,. \tag{B.15}$$

To this end, we first consider the rectangles $\mathcal{A}_1 \times \mathcal{A}_2$ with $\mathcal{A}_1 \subseteq \mathbb{R}^{M_1}$ $\mathscr{H}^{m_1}$-measurable and $\mathcal{A}_2 \subseteq \mathbb{R}^{M_2}$ $\mathscr{H}^{m_2}$-measurable. We have

$$\begin{aligned}
\Pr\{(\mathsf{x},\mathsf{y}) \in \mathcal{A}_1 \times \mathcal{A}_2\} &\overset{(a)}{=} \Pr\{\mathsf{x} \in \mathcal{A}_1\}\ \Pr\{\mathsf{y} \in \mathcal{A}_2\} \\
&\overset{(6.11)}{=} \int_{\mathcal{A}_1} \theta^{m_1}_{\mathsf{x}}(\boldsymbol{x})\, \mathrm{d}\mathscr{H}^{m_1}|_{\mathcal{E}_1}(\boldsymbol{x}) \int_{\mathcal{A}_2} \theta^{m_2}_{\mathsf{y}}(\boldsymbol{y})\, \mathrm{d}\mathscr{H}^{m_2}|_{\mathcal{E}_2}(\boldsymbol{y}) \\
&\overset{(b)}{=} \int_{\mathcal{A}_1\times\mathcal{A}_2} \theta^{m_1}_{\mathsf{x}}(\boldsymbol{x})\theta^{m_2}_{\mathsf{y}}(\boldsymbol{y})\, \mathrm{d}\big(\mathscr{H}^{m_1}|_{\mathcal{E}_1} \times \mathscr{H}^{m_2}|_{\mathcal{E}_2}\big)(\boldsymbol{x},\boldsymbol{y}) \\
&\overset{(c)}{=} \int_{\mathcal{A}_1\times\mathcal{A}_2} \theta^{m_1}_{\mathsf{x}}(\boldsymbol{x})\theta^{m_2}_{\mathsf{y}}(\boldsymbol{y})\, \mathrm{d}\mathscr{H}^{m_1+m_2}|_{\mathcal{E}_1\times\mathcal{E}_2}(\boldsymbol{x},\boldsymbol{y})
\end{aligned} \tag{B.16}$$

where (a) holds because x and y are independent, (b) holds by Fubini's theorem, and (c) holds because $\mathcal{E}_1$ and $\mathcal{E}_2$ are product-compatible (see (7.3)). Because the rectangles generate the $\mu(\mathsf{x},\mathsf{y})^{-1}$-measurable sets, (B.16) implies (B.15). For a $\mu(\mathsf{x},\mathsf{y})^{-1}$-measurable set $\mathcal{A} \subseteq \mathbb{R}^{M_1+M_2}$ satisfying $\mathscr{H}^{m_1+m_2}|_{\mathcal{E}_1\times\mathcal{E}_2}(\mathcal{A}) = 0$, (B.15) implies $\mu(\mathsf{x},\mathsf{y})^{-1}(\mathcal{A}) = 0$, and thus, $\mu(\mathsf{x},\mathsf{y})^{-1} \ll \mathscr{H}^{m_1+m_2}|_{\mathcal{E}_1\times\mathcal{E}_2}$. Furthermore, since $\mathcal{E}_1$ is m_1-rectifiable and $\mathcal{E}_2$ is m_2-rectifiable, it follows from Lemma 7.3 that $\mathcal{E}_1 \times \mathcal{E}_2$ is $(m_1 + m_2)$-rectifiable. Thus, according to Definition 6.9, (x,y) is an $(m_1 + m_2)$-rectifiable random variable.

Proof of Property 2: Again by (B.15), we see that $\theta^{m_1}_{\mathsf{x}}\theta^{m_2}_{\mathsf{y}}$ is equal to the Radon-Nikodym derivative $\frac{\mathrm{d}\mu(\mathsf{x},\mathsf{y})^{-1}}{\mathrm{d}\mathscr{H}^{m_1+m_2}|_{\mathcal{E}_1\times\mathcal{E}_2}}$. On the other hand, by (6.10), $\frac{\mathrm{d}\mu(\mathsf{x},\mathsf{y})^{-1}}{\mathrm{d}\mathscr{H}^{m_1+m_2}|_{\mathcal{E}_1\times\mathcal{E}_2}} = \theta^{m_1+m_2}_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y})$ $\mathscr{H}^{m_1+m_2}|_{\mathcal{E}_1\times\mathcal{E}_2}$-almost everywhere. Therefore, $\theta^{m_1}_{\mathsf{x}}(\boldsymbol{x})\theta^{m_2}_{\mathsf{y}}(\boldsymbol{y})$ equals the Hausdorff density $\theta^{m_1+m_2}_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y})$ $\mathscr{H}^{m_1+m_2}|_{\mathcal{E}_1\times\mathcal{E}_2}$-almost everywhere. Moreover, by Property 4 in Corollary 6.10, $\theta^{m_1}_{\mathsf{x}}(\boldsymbol{x})\theta^{m_2}_{\mathsf{y}}(\boldsymbol{y})$ and $\theta^{m_1+m_2}_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y})$ are both zero, and thus equal, $\mathscr{H}^{m_1+m_2}$-almost everywhere on $(\mathcal{E}_1 \times \mathcal{E}_2)^c$. Hence, $\theta^{m_1+m_2}_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y}) = \theta^{m_1}_{\mathsf{x}}(\boldsymbol{x})\theta^{m_2}_{\mathsf{y}}(\boldsymbol{y})$ $\mathscr{H}^{m_1+m_2}$-almost everywhere.

Proof of Property 3: According to Property 7 in Corollary 6.10, we have to show that

$\theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y}) > 0$ $\mathscr{H}^{m_1+m_2}$-almost everywhere on $\mathcal{E}_1 \times \mathcal{E}_2$, i.e., that the set

$$\mathcal{N} \triangleq \{(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{E}_1 \times \mathcal{E}_2 : \theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y}) = 0\} \tag{B.17}$$

has $\mathscr{H}^{m_1+m_2}$-measure zero.

Because $\mathcal{E}_1$ is a support of x, we have $\theta_{\mathsf{x}}^{m_1}(\boldsymbol{x}) > 0$ $\mathscr{H}^{m_1}$-almost everywhere on $\mathcal{E}_1$. Thus, the set $\mathcal{N}_1 \subseteq \mathbb{R}^{M_1}$ on which $\theta_{\mathsf{x}}^{m_1}(\boldsymbol{x}) = 0$ must have an intersection with $\mathcal{E}_1$ of $\mathscr{H}^{m_1}$-measure zero, i.e.,

$$\mathscr{H}^{m_1}|_{\mathcal{E}_1}(\mathcal{N}_1) = 0 \quad \text{with } \mathcal{N}_1 \triangleq \{\boldsymbol{x} \in \mathcal{E}_1 : \theta_{\mathsf{x}}^{m_1}(\boldsymbol{x}) = 0\}\,. \tag{B.18}$$

Along the same lines, we obtain

$$\mathscr{H}^{m_2}|_{\mathcal{E}_2}(\mathcal{N}_2) = 0 \quad \text{with } \mathcal{N}_2 \triangleq \{\boldsymbol{y} \in \mathcal{E}_2 : \theta_{\mathsf{y}}^{m_2}(\boldsymbol{y}) = 0\}\,. \tag{B.19}$$

Because, by (7.5), $\theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y}) = \theta_{\mathsf{x}}^{m_1}(\boldsymbol{x})\theta_{\mathsf{y}}^{m_2}(\boldsymbol{y})$, each $(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{N}$ satisfies either $\theta_{\mathsf{x}}^{m_1}(\boldsymbol{x}) = 0$ or $\theta_{\mathsf{y}}^{m_2}(\boldsymbol{y}) = 0$, i.e., $(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{N}_1 \times \mathcal{E}_2$ or $(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{E}_1 \times \mathcal{N}_2$. Thus, $\mathcal{N} \subseteq (\mathcal{N}_1 \times \mathcal{E}_2) \cup (\mathcal{E}_1 \times \mathcal{N}_2)$, and we conclude

$$\begin{aligned}
\mathscr{H}^{m_1+m_2}(\mathcal{N}) &\leq \mathscr{H}^{m_1+m_2}\big((\mathcal{N}_1 \times \mathcal{E}_2) \cup (\mathcal{E}_1 \times \mathcal{N}_2)\big) \\
&\overset{(a)}{\leq} \mathscr{H}^{m_1+m_2}(\mathcal{N}_1 \times \mathcal{E}_2) + \mathscr{H}^{m_1+m_2}(\mathcal{E}_1 \times \mathcal{N}_2) \\
&\overset{(b)}{=} \mathscr{H}^{m_1}|_{\mathcal{E}_1}(\mathcal{N}_1)\mathscr{H}^{m_2}|_{\mathcal{E}_2}(\mathcal{E}_2) + \mathscr{H}^{m_1}|_{\mathcal{E}_1}(\mathcal{E}_1)\mathscr{H}^{m_2}|_{\mathcal{E}_2}(\mathcal{N}_2) \\
&\overset{(c)}{=} 0
\end{aligned} \tag{B.20}$$

where (a) holds by the subadditivity of measures, (b) holds because $\mathcal{E}_1$ and $\mathcal{E}_2$ are product-compatible (see (7.4)), and (c) holds because $\mathscr{H}^{m_1}|_{\mathcal{E}_1}(\mathcal{N}_1) = 0$ and $\mathscr{H}^{m_2}|_{\mathcal{E}_2}(\mathcal{N}_2) = 0$.

Proof of Property 4: We have

$$\begin{aligned}
&\mathfrak{h}^{m_1+m_2}(\mathsf{x},\mathsf{y}) \\
&\quad\overset{(7.2)}{=} -\int_{\mathbb{R}^{M_1+M_2}} \theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y}) \log \theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y})\, \mathrm{d}\mathscr{H}^{m_1+m_2}|_{\mathcal{E}_1\times\mathcal{E}_2}(\boldsymbol{x},\boldsymbol{y}) \\
&\quad\overset{(a)}{=} -\int_{\mathbb{R}^{M_1+M_2}} \theta_{\mathsf{x}}^{m_1}(\boldsymbol{x})\theta_{\mathsf{y}}^{m_2}(\boldsymbol{y}) \log\big(\theta_{\mathsf{x}}^{m_1}(\boldsymbol{x})\theta_{\mathsf{y}}^{m_2}(\boldsymbol{y})\big)\, \mathrm{d}\mathscr{H}^{m_1+m_2}|_{\mathcal{E}_1\times\mathcal{E}_2}(\boldsymbol{x},\boldsymbol{y}) \\
&\quad\overset{(b)}{=} -\int_{\mathbb{R}^{M_1+M_2}} \theta_{\mathsf{x}}^{m_1}(\boldsymbol{x})\theta_{\mathsf{y}}^{m_2}(\boldsymbol{y}) \log\big(\theta_{\mathsf{x}}^{m_1}(\boldsymbol{x})\theta_{\mathsf{y}}^{m_2}(\boldsymbol{y})\big)\, \mathrm{d}\big(\mathscr{H}^{m_1}|_{\mathcal{E}_1} \times \mathscr{H}^{m_2}|_{\mathcal{E}_2}\big)(\boldsymbol{x},\boldsymbol{y}) \\
&\quad\overset{(c)}{=} -\int_{\mathbb{R}^{M_2}}\int_{\mathbb{R}^{M_1}} \theta_{\mathsf{x}}^{m_1}(\boldsymbol{x})\theta_{\mathsf{y}}^{m_2}(\boldsymbol{y})\big(\log\theta_{\mathsf{x}}^{m_1}(\boldsymbol{x}) + \log\theta_{\mathsf{y}}^{m_2}(\boldsymbol{y})\big)\, \mathrm{d}\mathscr{H}^{m_1}|_{\mathcal{E}_1}(\boldsymbol{x})\, \mathrm{d}\mathscr{H}^{m_2}|_{\mathcal{E}_2}(\boldsymbol{y}) \\
&\quad\overset{(d)}{=} -\int_{\mathbb{R}^{M_1}} \theta_{\mathsf{x}}^{m_1}(\boldsymbol{x}) \log\theta_{\mathsf{x}}^{m_1}(\boldsymbol{x})\, \mathrm{d}\mathscr{H}^{m_1}|_{\mathcal{E}_1}(\boldsymbol{x}) - \int_{\mathbb{R}^{M_2}} \theta_{\mathsf{y}}^{m_2}(\boldsymbol{y}) \log\theta_{\mathsf{y}}^{m_2}(\boldsymbol{y})\, \mathrm{d}\mathscr{H}^{m_2}|_{\mathcal{E}_2}(\boldsymbol{y}) \\
&\quad\overset{(6.16)}{=} \mathfrak{h}^{m_1}(\mathsf{x}) + \mathfrak{h}^{m_2}(\mathsf{y})\,.
\end{aligned}$$

Here, (a) holds because of (7.5), (b) holds because $\mathcal{E}_1$ and $\mathcal{E}_2$ are product-compatible, (c)

holds by Fubini's theorem, and (d) holds because, by (6.13), $\int_{\mathbb{R}^{M_1}} \theta_{\mathsf{x}}^{m_1}(\boldsymbol{x})\,\mathrm{d}\mathscr{H}^{m_1}|_{\mathcal{E}_1}(\boldsymbol{x}) = \int_{\mathbb{R}^{M_2}} \theta_{\mathsf{y}}^{m_2}(\boldsymbol{y})\,\mathrm{d}\mathscr{H}^{m_2}|_{\mathcal{E}_2}(\boldsymbol{y}) = 1$.

B.8 Proof of Theorem 7.6

We will use the generalized coarea formula [Federer, 1969, Th. 3.2.22] several times in our proofs. Unfortunately, the classical version only holds for sets of finite Hausdorff measure. Thus, we first present adaptation that is suited to our setting.

Theorem B.1 Let $\mathcal{E} \subseteq \mathbb{R}^{M_1+M_2}$ be an m-rectifiable set. Furthermore, let $\mathcal{E}_2 \triangleq \mathfrak{p}_{\mathsf{y}}(\mathcal{E}) \subseteq \mathbb{R}^{M_2}$ be m_2-rectifiable, $\mathscr{H}^{m_2}(\mathcal{E}_2) < \infty$, and $\mathscr{J}_{\mathfrak{p}_{\mathsf{y}}}^{\mathcal{E}} \neq 0$ $\mathscr{H}^m|_{\mathcal{E}}$-almost everywhere. Assume that $g\colon \mathcal{E} \to \mathbb{R}$ is an $\mathscr{H}^m$-measurable function satisfying either of the following properties:

(i) $g(\boldsymbol{x},\boldsymbol{y}) \geq 0$ $\mathscr{H}^m$-almost everywhere

(ii) $\int_{\mathcal{E}} |g(\boldsymbol{x},\boldsymbol{y})|\,\mathrm{d}\mathscr{H}^m(\boldsymbol{x},\boldsymbol{y}) < \infty$.

Then for all $\mathscr{H}^{m_1}$-measurable sets $\mathcal{A}_1 \subseteq \mathbb{R}^{M_1}$ and $\mathscr{H}^{m_2}$-measurable sets $\mathcal{A}_2 \subseteq \mathbb{R}^{M_2}$,

$$\int_{(\mathcal{A}_1\times\mathcal{A}_2)\cap\mathcal{E}} g(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\mathscr{H}^m(\boldsymbol{x},\boldsymbol{y}) = \int_{\mathcal{A}_2\cap\mathcal{E}_2}\left(\int_{\mathcal{A}_1\cap\mathcal{E}^{(\boldsymbol{y})}} \frac{g(\boldsymbol{x},\boldsymbol{y})}{\mathscr{J}_{\mathfrak{p}_{\mathsf{y}}}^{\mathcal{E}}(\boldsymbol{x},\boldsymbol{y})}\,\mathrm{d}\mathscr{H}^{m-m_2}(\boldsymbol{x})\right)\mathrm{d}\mathscr{H}^{m_2}(\boldsymbol{y}) \tag{B.21}$$

where $\mathcal{E}^{(\boldsymbol{y})} \triangleq \{\boldsymbol{x} \in \mathbb{R}^{M_1} : (\boldsymbol{x},\boldsymbol{y}) \in \mathcal{E}\}$. Furthermore, the set $\mathcal{A}_1 \cap \mathcal{E}^{(\boldsymbol{y})}$ is $(m-m_2)$-rectifiable for $\mathscr{H}^{m_2}$-almost every $\boldsymbol{y} \in \mathbb{R}^{M_2}$.

Proof. By Property 2 in Lemma 6.3, $\mathscr{H}^m|_{\mathcal{E}}$ is σ-finite. Thus, we can partition $\mathcal{E}$ as $\mathcal{E} = \bigcup_{i\in\mathbb{N}} \mathcal{F}_i$ with mutually disjoint sets $\mathcal{F}_i$ satisfying $\mathscr{H}^m(\mathcal{F}_i) < \infty$. For $\mathcal{A}_1 \subseteq \mathbb{R}^{M_1}$ $\mathscr{H}^{m_1}$-measurable and $\mathcal{A}_2 \subseteq \mathbb{R}^{M_2}$ $\mathscr{H}^{m_2}$-measurable, we have

$$\begin{aligned}
&\int_{(\mathcal{A}_1\times\mathcal{A}_2)\cap\mathcal{E}} g(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\mathscr{H}^m(\boldsymbol{x},\boldsymbol{y}) \\
&\qquad = \sum_{i\in\mathbb{N}} \int_{(\mathcal{A}_1\times\mathcal{A}_2)\cap\mathcal{F}_i} g(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\mathscr{H}^m(\boldsymbol{x},\boldsymbol{y}) \\
&\qquad \overset{(a)}{=} \sum_{i\in\mathbb{N}} \int_{\mathcal{A}_2\cap\mathcal{E}_2}\left(\int_{(\mathcal{A}_1\times\mathcal{A}_2)\cap\mathfrak{p}_{\mathsf{y}}^{-1}(\{\boldsymbol{y}\})\cap\mathcal{F}_i} \frac{g(\boldsymbol{x},\boldsymbol{y}')}{\mathscr{J}_{\mathfrak{p}_{\mathsf{y}}}^{\mathcal{E}}(\boldsymbol{x},\boldsymbol{y}')}\,\mathrm{d}\mathscr{H}^{m-m_2}(\boldsymbol{x},\boldsymbol{y}')\right)\mathrm{d}\mathscr{H}^{m_2}(\boldsymbol{y})
\end{aligned} \tag{B.22}$$

where (a) holds by the classical version of the general coarea formula [Federer, 1969, Th. 3.2.22] (note that $\mathcal{E}_2$ and $\mathcal{F}_i$ have finite Hausdorff measure) and because $\mathscr{J}_{\mathfrak{p}_{\mathsf{y}}}^{\mathcal{E}} \neq 0$ $\mathscr{H}^m|_{\mathcal{E}}$-almost everywhere. If (i) holds, then by Fubini's theorem, we can change the order of integration and summation in (B.22). If (ii) holds, we will apply Lebesgue's theorem of dominated convergence [Ambrosio et al., 2000, Th. 1.21] to swap integration and summation. To this end, we have to show that the functions

$$f_N(\boldsymbol{y}) \triangleq \sum_{i=1}^{N} \int_{(\mathcal{A}_1\times\mathcal{A}_2)\cap\mathfrak{p}_{\mathsf{y}}^{-1}(\{\boldsymbol{y}\})\cap\mathcal{F}_i} \frac{g(\boldsymbol{x},\boldsymbol{y}')}{\mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_{\mathsf{y}}}(\boldsymbol{x},\boldsymbol{y}')}\,\mathrm{d}\mathscr{H}^{m-m_2}(\boldsymbol{x},\boldsymbol{y}')$$

are absolutely bounded by an $\mathscr{H}^{m_2}|_{\mathcal{A}_2\cap\mathcal{E}_2}$-integrable function. By the triangle inequality, we obtain

$$\begin{aligned}
|f_N(\boldsymbol{y})| &\le \sum_{i=1}^{N} \int_{(\mathcal{A}_1\times\mathcal{A}_2)\cap\mathfrak{p}_{\mathsf{y}}^{-1}(\{\boldsymbol{y}\})\cap\mathcal{F}_i} \frac{|g(\boldsymbol{x},\boldsymbol{y}')|}{\mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_{\mathsf{y}}}(\boldsymbol{x},\boldsymbol{y}')}\,\mathrm{d}\mathscr{H}^{m-m_2}(\boldsymbol{x},\boldsymbol{y}')\\
&\le \sum_{i\in\mathbb{N}} \int_{(\mathcal{A}_1\times\mathcal{A}_2)\cap\mathfrak{p}_{\mathsf{y}}^{-1}(\{\boldsymbol{y}\})\cap\mathcal{F}_i} \frac{|g(\boldsymbol{x},\boldsymbol{y}')|}{\mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_{\mathsf{y}}}(\boldsymbol{x},\boldsymbol{y}')}\,\mathrm{d}\mathscr{H}^{m-m_2}(\boldsymbol{x},\boldsymbol{y}')\,.
\end{aligned} \tag{B.23}$$

We claim that the right-hand side of (B.23) is an integrable function. Indeed, we have

$$\begin{aligned}
&\int_{\mathcal{A}_2\cap\mathcal{E}_2}\bigg(\sum_{i\in\mathbb{N}} \int_{(\mathcal{A}_1\times\mathcal{A}_2)\cap\mathfrak{p}_{\mathsf{y}}^{-1}(\{\boldsymbol{y}\})\cap\mathcal{F}_i} \frac{|g(\boldsymbol{x},\boldsymbol{y}')|}{\mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_{\mathsf{y}}}(\boldsymbol{x},\boldsymbol{y}')}\,\mathrm{d}\mathscr{H}^{m-m_2}(\boldsymbol{x},\boldsymbol{y}')\bigg)\mathrm{d}\mathscr{H}^{m_2}(\boldsymbol{y})\\
&\quad\overset{(a)}{=} \sum_{i\in\mathbb{N}}\int_{\mathcal{A}_2\cap\mathcal{E}_2}\bigg(\int_{(\mathcal{A}_1\times\mathcal{A}_2)\cap\mathfrak{p}_{\mathsf{y}}^{-1}(\{\boldsymbol{y}\})\cap\mathcal{F}_i} \frac{|g(\boldsymbol{x},\boldsymbol{y}')|}{\mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_{\mathsf{y}}}(\boldsymbol{x},\boldsymbol{y}')}\,\mathrm{d}\mathscr{H}^{m-m_2}(\boldsymbol{x},\boldsymbol{y}')\bigg)\mathrm{d}\mathscr{H}^{m_2}(\boldsymbol{y})\\
&\quad\overset{(b)}{=} \sum_{i\in\mathbb{N}}\int_{(\mathcal{A}_1\times\mathcal{A}_2)\cap\mathcal{F}_i}|g(\boldsymbol{x},\boldsymbol{y})|\,\mathrm{d}\mathscr{H}^{m}(\boldsymbol{x},\boldsymbol{y})\\
&\quad= \int_{(\mathcal{A}_1\times\mathcal{A}_2)\cap\mathcal{E}}|g(\boldsymbol{x},\boldsymbol{y})|\,\mathrm{d}\mathscr{H}^{m}(\boldsymbol{x},\boldsymbol{y})\\
&\quad\le \int_{\mathcal{E}}|g(\boldsymbol{x},\boldsymbol{y})|\,\mathrm{d}\mathscr{H}^{m}(\boldsymbol{x},\boldsymbol{y})\\
&\quad\overset{(c)}{<} \infty
\end{aligned}$$

where (a) holds by Fubini's theorem, (b) holds by the general coarea formula [Federer, 1969, Th. 3.2.22], and (c) holds due to (ii).

Therefore, in either of the cases (i) or (ii), we can swap summation and integration on the right-hand side of (B.22). We thus obtain

$$\begin{aligned}
&\int_{(\mathcal{A}_1\times\mathcal{A}_2)\cap\mathcal{E}} g(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\mathscr{H}^{m}(\boldsymbol{x},\boldsymbol{y})\\
&\quad= \int_{\mathcal{A}_2\cap\mathcal{E}_2}\bigg(\sum_{i\in\mathbb{N}} \int_{(\mathcal{A}_1\times\mathcal{A}_2)\cap\mathfrak{p}_{\mathsf{y}}^{-1}(\{\boldsymbol{y}\})\cap\mathcal{F}_i} \frac{g(\boldsymbol{x},\boldsymbol{y}')}{\mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_{\mathsf{y}}}(\boldsymbol{x},\boldsymbol{y}')}\,\mathrm{d}\mathscr{H}^{m-m_2}(\boldsymbol{x},\boldsymbol{y}')\bigg)\mathrm{d}\mathscr{H}^{m_2}(\boldsymbol{y})\\
&\quad= \int_{\mathcal{A}_2\cap\mathcal{E}_2}\bigg(\int_{(\mathcal{A}_1\times\mathcal{A}_2)\cap\mathfrak{p}_{\mathsf{y}}^{-1}(\{\boldsymbol{y}\})\cap\mathcal{E}} \frac{g(\boldsymbol{x},\boldsymbol{y}')}{\mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_{\mathsf{y}}}(\boldsymbol{x},\boldsymbol{y}')}\,\mathrm{d}\mathscr{H}^{m-m_2}(\boldsymbol{x},\boldsymbol{y}')\bigg)\mathrm{d}\mathscr{H}^{m_2}(\boldsymbol{y})\\
&\quad\overset{(a)}{=} \int_{\mathcal{A}_2\cap\mathcal{E}_2}\bigg(\int_{(\mathcal{A}_1\times\mathcal{A}_2)\cap\mathfrak{p}_{\mathsf{y}}^{-1}(\{\boldsymbol{y}\})\cap\mathcal{E}} \frac{g(\boldsymbol{x},\boldsymbol{y})}{\mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_{\mathsf{y}}}(\boldsymbol{x},\boldsymbol{y})}\,\mathrm{d}\mathscr{H}^{m-m_2}(\boldsymbol{x},\boldsymbol{y}')\bigg)\mathrm{d}\mathscr{H}^{m_2}(\boldsymbol{y})\\
&\quad\overset{(b)}{=} \int_{\mathcal{A}_2\cap\mathcal{E}_2}\bigg(\int_{\mathcal{A}_1\cap\mathcal{E}^{(\boldsymbol{y})}} \frac{g(\boldsymbol{x},\boldsymbol{y})}{\mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_{\mathsf{y}}}(\boldsymbol{x},\boldsymbol{y})}\,\mathrm{d}\mathscr{H}^{m-m_2}(\boldsymbol{x})\bigg)\mathrm{d}\mathscr{H}^{m_2}(\boldsymbol{y})
\end{aligned}$$

where (a) holds because $\boldsymbol{y}' = \boldsymbol{y}$ for all $(\boldsymbol{x}, \boldsymbol{y}') \in \mathfrak{p}_{\mathsf{y}}^{-1}(\{\boldsymbol{y}\})$, and (b) holds because the Hausdorff measure does not depend on the ambient space [Ambrosio et al., 2000, Remark 2.48], i.e., integration with respect to $\mathscr{H}^{m-m_2}$ on the affine subspace $\mathfrak{p}_{\mathsf{y}}^{-1}(\{\boldsymbol{y}\}) \subseteq \mathbb{R}^{M_1+M_2}$ and on $\mathbb{R}^{M_1}$ coincide. Thus, we have shown (B.21).

By [Federer, 1969, Th. 3.2.22], the sets $\mathfrak{p}_{\mathsf{y}}^{-1}(\{\boldsymbol{y}\}) \cap \mathcal{F}_i$ are $(m - m_2)$-rectifiable for $\mathscr{H}^{m_2}$-almost every $\boldsymbol{y} \in \mathbb{R}^{M_2}$. By Property 5 in Lemma 6.3, also the union $\bigcup_{i\in\mathbb{N}} \mathfrak{p}_{\mathsf{y}}^{-1}(\{\boldsymbol{y}\}) \cap \mathcal{F}_i = \mathfrak{p}_{\mathsf{y}}^{-1}(\{\boldsymbol{y}\}) \cap \mathcal{E}$ is $(m - m_2)$-rectifiable for $\mathscr{H}^{m_2}$-almost every $\boldsymbol{y} \in \mathbb{R}^{M_2}$. The Lipschitz mapping $\mathfrak{p}_{\mathsf{x}} \colon \mathbb{R}^{M_1+M_2} \to \mathbb{R}^{M_1}$, $\mathfrak{p}_{\mathsf{x}}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}$ satisfies

$$\begin{aligned}
\mathfrak{p}_{\mathsf{x}}\big(\mathfrak{p}_{\mathsf{y}}^{-1}(\{\boldsymbol{y}\}) \cap \mathcal{E}\big) &= \{\boldsymbol{x} \in \mathbb{R}^{M_1} : (\boldsymbol{x}, \boldsymbol{y}') \in \mathfrak{p}_{\mathsf{y}}^{-1}(\{\boldsymbol{y}\}) \cap \mathcal{E}\} \\
&= \{\boldsymbol{x} \in \mathbb{R}^{M_1} : (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{E}\} \\
&= \mathcal{E}^{(\boldsymbol{y})}.
\end{aligned}$$

Thus, $\mathcal{E}^{(\boldsymbol{y})}$ is obtained via a Lipschitz mapping from the set $\mathfrak{p}_{\mathsf{y}}^{-1}(\{\boldsymbol{y}\}) \cap \mathcal{E}$, which is $(m - m_2)$-rectifiable for $\mathscr{H}^{m_2}$-almost every $\boldsymbol{y} \in \mathbb{R}^{M_2}$. Therefore, by Property 3 in Lemma 6.3, $\mathcal{E}^{(\boldsymbol{y})}$ is again $(m - m_2)$-rectifiable for $\mathscr{H}^{m_2}$-almost every $\boldsymbol{y} \in \mathbb{R}^{M_2}$. Finally, by Property 1 in Lemma 6.3, the same is true for $\mathcal{A}_1 \cap \mathcal{E}^{(\boldsymbol{y})}$. □

We now proceed to the proof of Theorem 7.6.

Proof of Property 1: We have for any $\mathscr{H}^{m_2}$-measurable set $\mathcal{A} \subseteq \mathbb{R}^{M_2}$

$$\begin{aligned}
\mu\mathsf{y}^{-1}(\mathcal{A}) &= \Pr\{\mathsf{y} \in \mathcal{A}\} \\
&= \Pr\{(\mathsf{x}, \mathsf{y}) \in \mathbb{R}^{M_1} \times \mathcal{A}\} \\
&= \int_{(\mathbb{R}^{M_1} \times \mathcal{A}) \cap \mathcal{E}} \theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\mathscr{H}^m(\boldsymbol{x}, \boldsymbol{y}) \\
&\overset{(a)}{=} \int_{\mathcal{A} \cap \widetilde{\mathcal{E}}_2} \bigg(\int_{\mathcal{E}^{(\boldsymbol{y})}} \frac{\theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x}, \boldsymbol{y})}{\mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_{\mathsf{y}}}(\boldsymbol{x}, \boldsymbol{y})} \, \mathrm{d}\mathscr{H}^{m-m_2}(\boldsymbol{x}) \bigg) \mathrm{d}\mathscr{H}^{m_2}(\boldsymbol{y}) \\
&= \int_{\mathcal{A}} \bigg(\int_{\mathcal{E}^{(\boldsymbol{y})}} \frac{\theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x}, \boldsymbol{y})}{\mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_{\mathsf{y}}}(\boldsymbol{x}, \boldsymbol{y})} \, \mathrm{d}\mathscr{H}^{m-m_2}(\boldsymbol{x}) \bigg) \mathrm{d}\mathscr{H}^{m_2}|_{\widetilde{\mathcal{E}}_2}(\boldsymbol{y})
\end{aligned} \tag{B.24}$$

where in (a) we used (B.21) with $g(\boldsymbol{x}, \boldsymbol{y}) = \theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x}, \boldsymbol{y}) \geq 0$. For an $\mathscr{H}^{m_2}$-measurable set $\mathcal{A}$ satisfying $\mathscr{H}^{m_2}|_{\widetilde{\mathcal{E}}_2}(\mathcal{A}) = 0$, (B.24) implies $\mu\mathsf{y}^{-1}(\mathcal{A}) = 0$, i.e., $\mu\mathsf{y}^{-1} \ll \mathscr{H}^{m_2}|_{\widetilde{\mathcal{E}}_2}$. Thus, according to Definition 6.9, y is m_2-rectifiable.

Proof of Property 2: By Property 6 in Corollary 6.10 and because $\mu\mathsf{y}^{-1} \ll \mathscr{H}^{m_2}|_{\widetilde{\mathcal{E}}_2}$, there exists a support $\mathcal{E}_2 \subseteq \widetilde{\mathcal{E}}_2$ of the random variable y.

Proof of Property 3: Again by (B.24), $\int_{\mathcal{E}^{(\boldsymbol{y})}} \frac{\theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y})}{\mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_{\mathsf{y}}}(\boldsymbol{x},\boldsymbol{y})} \, \mathrm{d}\mathscr{H}^{m-m_2}(\boldsymbol{x})$ is the Radon-Nikodym derivative $\frac{\mathrm{d}\mu\mathsf{y}^{-1}}{\mathrm{d}\mathscr{H}^{m_2}|_{\widetilde{\mathcal{E}}_2}}$. By (6.10), $\frac{\mathrm{d}\mu\mathsf{y}^{-1}}{\mathrm{d}\mathscr{H}^{m_2}|_{\widetilde{\mathcal{E}}_2}}$ equals $\theta^{m_2}_{\mathsf{y}}(\boldsymbol{y})$. This implies (7.12).

Proof of Property 4: Similar to (B.24), we obtain

$$\mathfrak{h}^{m_2}(\mathsf{y}) = -\int_{\mathcal{E}_2} \theta^{m_2}_{\mathsf{y}}(\boldsymbol{y}) \log \theta^{m_2}_{\mathsf{y}}(\boldsymbol{y}) \, \mathrm{d}\mathscr{H}^{m_2}(\boldsymbol{y})$$

$$\overset{(7.12)}{=} -\int_{\mathcal{E}_2} \left(\int_{\mathcal{E}(\boldsymbol{y})} \frac{\theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y})}{\mathscr{J}^{\mathcal{E}}_{\mathsf{p}_\mathsf{y}}(\boldsymbol{x},\boldsymbol{y})} \, \mathrm{d}\mathscr{H}^{m-m_2}(\boldsymbol{x}) \right) \times \log \left(\int_{\mathcal{E}(\boldsymbol{y})} \frac{\theta^m_{(\mathsf{x},\mathsf{y})}(\tilde{\boldsymbol{x}},\boldsymbol{y})}{\mathscr{J}^{\mathcal{E}}_{\mathsf{p}_\mathsf{y}}(\tilde{\boldsymbol{x}},\boldsymbol{y})} \, \mathrm{d}\mathscr{H}^{m-m_2}(\tilde{\boldsymbol{x}}) \right) \mathrm{d}\mathscr{H}^{m_2}(\boldsymbol{y})$$

$$= -\int_{\mathcal{E}_2} \left(\int_{\mathcal{E}(\boldsymbol{y})} \frac{\theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y})}{\mathscr{J}^{\mathcal{E}}_{\mathsf{p}_\mathsf{y}}(\boldsymbol{x},\boldsymbol{y})} \times \log \left(\int_{\mathcal{E}(\boldsymbol{y})} \frac{\theta^m_{(\mathsf{x},\mathsf{y})}(\tilde{\boldsymbol{x}},\boldsymbol{y})}{\mathscr{J}^{\mathcal{E}}_{\mathsf{p}_\mathsf{y}}(\tilde{\boldsymbol{x}},\boldsymbol{y})} \, \mathrm{d}\mathscr{H}^{m-m_2}(\tilde{\boldsymbol{x}}) \right) \mathrm{d}\mathscr{H}^{m-m_2}(\boldsymbol{x}) \right) \mathrm{d}\mathscr{H}^{m_2}(\boldsymbol{y})$$

$$\overset{(a)}{=} -\int_{\mathcal{E}} \theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y}) \log \left(\int_{\mathcal{E}(\boldsymbol{y})} \frac{\theta^m_{(\mathsf{x},\mathsf{y})}(\tilde{\boldsymbol{x}},\boldsymbol{y})}{\mathscr{J}^{\mathcal{E}}_{\mathsf{p}_\mathsf{y}}(\tilde{\boldsymbol{x}},\boldsymbol{y})} \, \mathrm{d}\mathscr{H}^{m-m_2}(\tilde{\boldsymbol{x}}) \right) \mathrm{d}\mathscr{H}^{m}(\boldsymbol{x},\boldsymbol{y})$$

where in (a) we used (B.21) with $\mathcal{A}_1 = \mathbb{R}^{M_1}$, $\mathcal{A}_2 = \mathbb{R}^{M_2}$, and

$$g(\boldsymbol{x},\boldsymbol{y}) = \theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y}) \log \left(\int_{\mathcal{E}(\boldsymbol{y})} \frac{\theta^m_{(\mathsf{x},\mathsf{y})}(\tilde{\boldsymbol{x}},\boldsymbol{y})}{\mathscr{J}^{\mathcal{E}}_{\mathsf{p}_\mathsf{y}}(\tilde{\boldsymbol{x}},\boldsymbol{y})} \, \mathrm{d}\mathscr{H}^{m-m_2}(\tilde{\boldsymbol{x}}) \right).$$

(Here, $g(\boldsymbol{x},\boldsymbol{y})$ is $\mathscr{H}^m|_{\mathcal{E}}$-integrable by assumption, i.e., Condition (ii) in Theorem B.1 is satisfied.) Thus, (7.13) holds.

B.9 Proof of Theorem 7.8

Proof of Property 1: We have for any $\mathscr{H}^2$-measurable set $\mathcal{A} \subseteq \mathbb{R}^{M_2}$

$$\begin{aligned}
\mu\mathsf{y}^{-1}(\mathcal{A}) &= \Pr\{\mathsf{y} \in \mathcal{A}\} \\
&= \Pr\{(\mathsf{x},\mathsf{y}) \in \mathbb{R}^{M_1} \times \mathcal{A}\} \\
&= \int_{(\mathbb{R}^{M_1} \times \mathcal{A}) \cap \mathcal{E}} \theta^{m_1+m_2}_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y}) \, \mathrm{d}\mathscr{H}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y}) \\
&\overset{(a)}{=} \int_{(\mathbb{R}^{M_1} \times \mathcal{A}) \cap (\mathcal{E}_1 \times \mathcal{E}_2)} \theta^{m_1+m_2}_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y}) \, \mathrm{d}\mathscr{H}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y}) \\
&= \int_{\mathbb{R}^{M_1} \times \mathcal{A}} \theta^{m_1+m_2}_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y}) \, \mathrm{d}\mathscr{H}^{m_1+m_2}|_{\mathcal{E}_1 \times \mathcal{E}_2}(\boldsymbol{x},\boldsymbol{y}) \\
&\overset{(b)}{=} \int_{\mathbb{R}^{M_1} \times \mathcal{A}} \theta^{m_1+m_2}_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y}) \, \mathrm{d}(\mathscr{H}^{m_1}|_{\mathcal{E}_1} \times \mathscr{H}^{m_2}|_{\mathcal{E}_2})(\boldsymbol{x},\boldsymbol{y}) \\
&\overset{(c)}{=} \int_{\mathcal{A}} \left(\int_{\mathcal{E}_1} \theta^{m_1+m_2}_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y}) \, \mathrm{d}\mathscr{H}^{m_1}(\boldsymbol{x}) \right) \mathrm{d}\mathscr{H}^{m_2}|_{\mathcal{E}_2}(\boldsymbol{y})
\end{aligned} \qquad \text{(B.25)}$$

where (a) holds because, by the product-compatibility of x and y, $\mathcal{E} \subseteq \mathcal{E}_1 \times \mathcal{E}_2$ and because, by Property 4 in Corollary 6.10, $\theta^{m_1+m_2}_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y})$ is zero for $\mathscr{H}^{m_1+m_2}$-almost all $(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{E}^c$, (b) holds because $\mathcal{E}_1$ and $\mathcal{E}_2$ are product-compatible, and (c) holds by Fubini's theorem. By (B.25), $\int_{\mathcal{E}_1} \theta^{m_1+m_2}_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y}) \, \mathrm{d}\mathscr{H}^{m_1}(\boldsymbol{x})$ is the Radon-Nikodym derivative $\frac{\mathrm{d}\mu\mathsf{y}^{-1}}{\mathrm{d}\mathscr{H}^{m_2}|_{\mathcal{E}_2}}$, which, by (6.10), equals $\theta^{m_1}_{\mathsf{y}}(\boldsymbol{y})$. This implies (7.15).

Proof of Property 2: Similar to (B.25), we obtain

$$
\begin{aligned}
\mathfrak{h}^{m_2}(\mathsf{y}) &= -\int_{\mathcal{E}_2} \theta_{\mathsf{y}}^{m_2}(\boldsymbol{y}) \log \theta_{\mathsf{y}}^{m_2}(\boldsymbol{y})\, \mathrm{d}\mathscr{H}^{m_2}(\boldsymbol{y}) \\
&\overset{(7.15)}{=} -\int_{\mathcal{E}_2} \bigg(\int_{\mathcal{E}_1} \theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y})\, \mathrm{d}\mathscr{H}^{m_1}(\boldsymbol{x}) \bigg) \\
&\qquad\qquad \times \log \bigg(\int_{\mathcal{E}_1} \theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\tilde{\boldsymbol{x}},\boldsymbol{y})\, \mathrm{d}\mathscr{H}^{m_1}(\tilde{\boldsymbol{x}}) \bigg)\, \mathrm{d}\mathscr{H}^{m_2}(\boldsymbol{y}) \\
&= -\int_{\mathcal{E}_2} \bigg(\int_{\mathcal{E}_1} \theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y}) \\
&\qquad\qquad \times \log \bigg(\int_{\mathcal{E}_1} \theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\tilde{\boldsymbol{x}},\boldsymbol{y})\, \mathrm{d}\mathscr{H}^{m_1}(\tilde{\boldsymbol{x}}) \bigg) \mathrm{d}\mathscr{H}^{m_1}(\boldsymbol{x}) \bigg)\, \mathrm{d}\mathscr{H}^{m_2}(\boldsymbol{y}) \\
&\overset{(a)}{=} -\int_{\mathcal{E}_1\times\mathcal{E}_2} \theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y}) \\
&\qquad\qquad \times \log \bigg(\int_{\mathcal{E}_1} \theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\tilde{\boldsymbol{x}},\boldsymbol{y})\, \mathrm{d}\mathscr{H}^{m_1}(\tilde{\boldsymbol{x}}) \bigg)\, \mathrm{d}(\mathscr{H}^{m_1}|_{\mathcal{E}_1} \times \mathscr{H}^{m_2}|_{\mathcal{E}_2})(\boldsymbol{x},\boldsymbol{y}) \\
&\overset{(b)}{=} -\int_{\mathcal{E}_1\times\mathcal{E}_2} \theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y}) \log \bigg(\int_{\mathcal{E}_1} \theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\tilde{\boldsymbol{x}},\boldsymbol{y})\, \mathrm{d}\mathscr{H}^{m_1}(\tilde{\boldsymbol{x}}) \bigg)\, \mathrm{d}\mathscr{H}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y}) \\
&\overset{(c)}{=} -\int_{\mathcal{E}} \theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y}) \log \bigg(\int_{\mathcal{E}_1} \theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\tilde{\boldsymbol{x}},\boldsymbol{y})\, \mathrm{d}\mathscr{H}^{m_1}(\tilde{\boldsymbol{x}}) \bigg)\, \mathrm{d}\mathscr{H}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y})
\end{aligned}
$$

where (a) holds by Fubini's theorem, (b) holds because $\mathcal{E}_1$ and $\mathcal{E}_2$ are product-compatible, and (c) holds because, by the product-compatibility of x and y, $\mathcal{E} \subseteq \mathcal{E}_1 \times \mathcal{E}_2$ and because, by Property 4 in Corollary 6.10, $\theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y}) = 0$ for $\mathscr{H}^{m_1+m_2}$-almost all $(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{E}^c$. Thus, (7.16) holds.

B.10 Proof of Theorem 7.9

Proof of Property 1: Let $\mathcal{A}_1 \subseteq \mathbb{R}^{M_1}$ and $\mathcal{A}_2 \subseteq \mathbb{R}^{M_2}$ be Borel sets. Then

$$
\begin{aligned}
\Pr\{(\mathsf{x},\mathsf{y}) \in \mathcal{A}_1 \times \mathcal{A}_2\} &\overset{(7.22)}{=} \int_{\mathcal{A}_2} \Pr\{\mathsf{x} \in \mathcal{A}_1 \,|\, \mathsf{y} = \boldsymbol{y}\}\, \mathrm{d}\mu\mathsf{y}^{-1}(\boldsymbol{y}) \\
&\overset{(a)}{=} \int_{\mathcal{A}_2} \Pr\{\mathsf{x} \in \mathcal{A}_1 \,|\, \mathsf{y} = \boldsymbol{y}\}\, \theta_{\mathsf{y}}^{m_2}(\boldsymbol{y})\, \mathrm{d}\mathscr{H}^{m_2}|_{\mathcal{E}_2}(\boldsymbol{y}) \\
&\overset{(b)}{=} \int_{\mathcal{A}_2} \Pr\{\mathsf{x} \in \mathcal{A}_1 \,|\, \mathsf{y} = \boldsymbol{y}\}\, \theta_{\mathsf{y}}^{m_2}(\boldsymbol{y})\, \mathrm{d}\mathscr{H}^{m_2}|_{\widetilde{\mathcal{E}}_2}(\boldsymbol{y})
\end{aligned}
\tag{B.26}
$$

where (a) holds because, by (6.10), $\theta_{\mathsf{y}}^{m_2}$ is equal to the Radon-Nikodym derivative $\frac{\mathrm{d}\mu\mathsf{y}^{-1}}{\mathrm{d}\mathscr{H}^{m_2}|_{\mathcal{E}_2}}$, and (b) holds because, by Property 4 in Corollary 6.10, $\theta_{\mathsf{y}}^{m_2}(\boldsymbol{y}) = 0$ for $\mathscr{H}^{m_2}$-almost all $\boldsymbol{y} \in \widetilde{\mathcal{E}}_2^c$. On the other hand, we have

$$
\Pr\{(\mathsf{x},\mathsf{y}) \in \mathcal{A}_1 \times \mathcal{A}_2\} = \int_{(\mathcal{A}_1\times\mathcal{A}_2)\cap\mathcal{E}} \theta_{(\mathsf{x},\mathsf{y})}^{m}(\boldsymbol{x},\boldsymbol{y})\, \mathrm{d}\mathscr{H}^{m}(\boldsymbol{x},\boldsymbol{y})
$$

$$\overset{(a)}{=} \int_{\mathcal{A}_2 \cap \widetilde{\mathcal{E}}_2} \left(\int_{\mathcal{A}_1 \cap \mathcal{E}^{(\boldsymbol{y})}} \frac{\theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y})}{\mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_\mathsf{y}}(\boldsymbol{x},\boldsymbol{y})} \,\mathrm{d}\mathscr{H}^{m-m_2}(\boldsymbol{x}) \right) \mathrm{d}\mathscr{H}^{m_2}(\boldsymbol{y})$$
$$= \int_{\mathcal{A}_2} \left(\int_{\mathcal{A}_1 \cap \mathcal{E}^{(\boldsymbol{y})}} \frac{\theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y})}{\mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_\mathsf{y}}(\boldsymbol{x},\boldsymbol{y})} \,\mathrm{d}\mathscr{H}^{m-m_2}(\boldsymbol{x}) \right) \mathrm{d}\mathscr{H}^{m_2}|_{\widetilde{\mathcal{E}}_2}(\boldsymbol{y}) \quad \text{(B.27)}$$

where in (a) we used (B.21) with $g(\boldsymbol{x},\boldsymbol{y}) = \theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y}) \geq 0$. Combining (B.26) and (B.27), we obtain that for $\mathscr{H}^{m_2}|_{\widetilde{\mathcal{E}}_2}$-almost every $\boldsymbol{y}$ and every $\mathscr{H}^{m_1}$-measurable set $\mathcal{A}_1 \subseteq \mathbb{R}^{M_1}$

$$\Pr\{\mathsf{x} \in \mathcal{A}_1 \,|\, \mathsf{y} = \boldsymbol{y}\}\, \theta^{m_2}_{\mathsf{y}}(\boldsymbol{y}) = \int_{\mathcal{A}_1 \cap \mathcal{E}^{(\boldsymbol{y})}} \frac{\theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y})}{\mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_\mathsf{y}}(\boldsymbol{x},\boldsymbol{y})} \,\mathrm{d}\mathscr{H}^{m-m_2}(\boldsymbol{x})\,. \quad \text{(B.28)}$$

Because (B.28) holds for $\mathscr{H}^{m_2}|_{\widetilde{\mathcal{E}}_2}$-almost every $\boldsymbol{y}$ and $\mathcal{E}_2 \subseteq \widetilde{\mathcal{E}}_2$, (B.28) also holds for $\mathscr{H}^{m_2}|_{\mathcal{E}_2}$-almost every $\boldsymbol{y}$. Furthermore, because $\mathcal{E}_2$ is a support of y, we have $\theta^{m_2}_{\mathsf{y}}(\boldsymbol{y}) > 0$ $\mathscr{H}^{m_2}|_{\mathcal{E}_2}$-almost everywhere. Thus, we obtain for $\mathscr{H}^{m_2}|_{\mathcal{E}_2}$-almost every $\boldsymbol{y}$ and every $\mathscr{H}^{m_1}$-measurable set $\mathcal{A}_1 \subseteq \mathbb{R}^{M_1}$

$$\Pr\{\mathsf{x} \in \mathcal{A}_1 \,|\, \mathsf{y} = \boldsymbol{y}\} = \int_{\mathcal{A}_1 \cap \mathcal{E}^{(\boldsymbol{y})}} \frac{\theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y})}{\mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_\mathsf{y}}(\boldsymbol{x},\boldsymbol{y})\, \theta^{m_2}_{\mathsf{y}}(\boldsymbol{y})} \,\mathrm{d}\mathscr{H}^{m-m_2}(\boldsymbol{x})$$
$$= \int_{\mathcal{A}_1} \frac{\theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y})}{\mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_\mathsf{y}}(\boldsymbol{x},\boldsymbol{y})\, \theta^{m_2}_{\mathsf{y}}(\boldsymbol{y})} \,\mathrm{d}\mathscr{H}^{m-m_2}|_{\mathcal{E}^{(\boldsymbol{y})}}(\boldsymbol{x})\,. \quad \text{(B.29)}$$

Therefore, $\Pr\{\mathsf{x} \in \cdot \,|\, \mathsf{y} = \boldsymbol{y}\} \ll \mathscr{H}^{m-m_2}|_{\mathcal{E}^{(\boldsymbol{y})}}$. By Theorem B.1, the set $\mathcal{E}^{(\boldsymbol{y})}$ is $(m-m_2)$-rectifiable for $\mathscr{H}^{m_2}|_{\mathcal{E}_2}$-almost every $\boldsymbol{y}$. Hence, according to Definition 6.9, $(\mathsf{x} \,|\, \mathsf{y} = \boldsymbol{y})$ is $(m-m_2)$-rectifiable for $\mathscr{H}^{m_2}|_{\mathcal{E}_2}$-almost every $\boldsymbol{y}$.

Proof of Property 2: By (B.29), $\frac{d\Pr\{\mathsf{x}\in\cdot\,|\,\mathsf{y}=\boldsymbol{y}\}}{d\mathscr{H}^{m-m_2}|_{\mathcal{E}^{(\boldsymbol{y})}}} = \frac{\theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y})}{\mathscr{J}^{\mathcal{E}}_{\mathfrak{p}_\mathsf{y}}(\boldsymbol{x},\boldsymbol{y})\,\theta^{m_2}_{\mathsf{y}}(\boldsymbol{y})}$ for $\mathscr{H}^{m_2}|_{\mathcal{E}_2}$-almost every $\boldsymbol{y}$. Furthermore, by (6.10), $\frac{d\Pr\{\mathsf{x}\in\cdot\,|\,\mathsf{y}=\boldsymbol{y}\}}{d\mathscr{H}^{m-m_2}|_{\mathcal{E}^{(\boldsymbol{y})}}} = \theta^{m-m_2}_{(\mathsf{x}\,|\,\mathsf{y}=\boldsymbol{y})}$. Hence, we see that (7.23) holds.

Proof of Property 3: By (6.17), we have

$$\mathfrak{h}^{m-m_2}(\mathsf{x} \,|\, \mathsf{y} = \boldsymbol{y}) = -\int_{\mathcal{E}^{(\boldsymbol{y})}} \theta^{m-m_2}_{(\mathsf{x}\,|\,\mathsf{y}=\boldsymbol{y})}(\boldsymbol{x}) \log \theta^{m-m_2}_{(\mathsf{x}\,|\,\mathsf{y}=\boldsymbol{y})}(\boldsymbol{x}) \,\mathrm{d}\mathscr{H}^{m-m_2}(\boldsymbol{x})\,.$$

The result (7.24) then follows by (7.23).

B.11 Proof of Theorem 7.10

Proof of Property 1: Let $\mathcal{A}_1 \subseteq \mathbb{R}^{M_1}$ and $\mathcal{A}_2 \subseteq \mathbb{R}^{M_2}$ be Borel sets. Then

$$\Pr\{(\mathsf{x},\mathsf{y}) \in \mathcal{A}_1 \times \mathcal{A}_2\} \overset{(7.22)}{=} \int_{\mathcal{A}_2} \Pr\{\mathsf{x} \in \mathcal{A}_1 \,|\, \mathsf{y} = \boldsymbol{y}\} \,\mathrm{d}\mu\mathsf{y}^{-1}(\boldsymbol{y})$$
$$\overset{(a)}{=} \int_{\mathcal{A}_2} \Pr\{\mathsf{x} \in \mathcal{A}_1 \,|\, \mathsf{y} = \boldsymbol{y}\}\, \theta^{m_2}_{\mathsf{y}}(\boldsymbol{y}) \,\mathrm{d}\mathscr{H}^{m_2}|_{\mathcal{E}_2}(\boldsymbol{y}) \quad \text{(B.30)}$$

where (a) holds because, by (6.10), $\theta_{\mathsf{y}}^{m_2}$ is equal to the Radon-Nikodym derivative $\frac{\mathrm{d}\mu\mathsf{y}^{-1}}{\mathrm{d}\mathscr{H}^{m_2}|_{\mathcal{E}_2}}$. On the other hand, we have

$$\begin{aligned}
\Pr\{(\mathsf{x},\mathsf{y}) \in \mathcal{A}_1 \times \mathcal{A}_2\} &= \int_{(\mathcal{A}_1\times\mathcal{A}_2)\cap\mathcal{E}} \theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\mathscr{H}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y}) \\
&\stackrel{(a)}{=} \int_{(\mathcal{A}_1\times\mathcal{A}_2)\cap(\mathcal{E}_1\times\mathcal{E}_2)} \theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\mathscr{H}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y}) \\
&= \int_{\mathcal{A}_1\times\mathcal{A}_2} \theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\mathscr{H}^{m_1+m_2}|_{\mathcal{E}_1\times\mathcal{E}_2}(\boldsymbol{x},\boldsymbol{y}) \\
&\stackrel{(b)}{=} \int_{\mathcal{A}_1\times\mathcal{A}_2} \theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}(\mathscr{H}^{m_1}|_{\mathcal{E}_1}\times\mathscr{H}^{m_2}|_{\mathcal{E}_2})(\boldsymbol{x},\boldsymbol{y}) \\
&\stackrel{(c)}{=} \int_{\mathcal{A}_2}\left(\int_{\mathcal{A}_1} \theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\mathscr{H}^{m_1}|_{\mathcal{E}_1}(\boldsymbol{x})\right)\mathrm{d}\mathscr{H}^{m_2}|_{\mathcal{E}_2}(\boldsymbol{y}) \qquad \text{(B.31)}
\end{aligned}$$

where (a) holds because, by the product-compatibility of x and y, $\mathcal{E} \subseteq \mathcal{E}_1 \times \mathcal{E}_2$ and because, by Property 4 in Corollary 6.10, $\theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y}) = 0$ for $\mathscr{H}^{m_1+m_2}$-almost all $(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{E}^c$, (b) holds because $\mathcal{E}_1$ and $\mathcal{E}_2$ are product-compatible, and (c) holds by Fubini's theorem. Combining (B.30) and (B.31), we obtain that for $\mathscr{H}^{m_2}|_{\mathcal{E}_2}$-almost every $\boldsymbol{y}$ and every Borel set $\mathcal{A}_1 \subseteq \mathbb{R}^{M_1}$

$$\Pr\{\mathsf{x} \in \mathcal{A}_1 \,|\, \mathsf{y} = \boldsymbol{y}\}\theta_{\mathsf{y}}^{m_2}(\boldsymbol{y}) = \int_{\mathcal{A}_1} \theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\mathscr{H}^{m_1}|_{\mathcal{E}_1}(\boldsymbol{x})\,.$$

Because $\mathcal{E}_2$ is a support of y, we have $\theta_{\mathsf{y}}^{m_2}(\boldsymbol{y}) > 0$ $\mathscr{H}^{m_2}|_{\mathcal{E}_2}$-almost everywhere. Thus, we obtain for $\mathscr{H}^{m_2}|_{\mathcal{E}_2}$-almost every $\boldsymbol{y}$ and every Borel set $\mathcal{A}_1 \subseteq \mathbb{R}^{M_1}$

$$\Pr\{\mathsf{x} \in \mathcal{A}_1 \,|\, \mathsf{y} = \boldsymbol{y}\} = \int_{\mathcal{A}_1} \frac{\theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y})}{\theta_{\mathsf{y}}^{m_2}(\boldsymbol{y})}\,\mathrm{d}\mathscr{H}^{m_1}|_{\mathcal{E}_1}(\boldsymbol{x})\,. \qquad \text{(B.32)}$$

Therefore, $\Pr\{\mathsf{x} \in \cdot \,|\, \mathsf{y} = \boldsymbol{y}\} \ll \mathscr{H}^{m_1}|_{\mathcal{E}_1}$. The set $\mathcal{E}_1$ is m_1-rectifiable and thus, according to Definition 6.9, $(\mathsf{x} \,|\, \mathsf{y} = \boldsymbol{y})$ is m_1-rectifiable for $\mathscr{H}^{m_2}|_{\mathcal{E}_2}$-almost every $\boldsymbol{y}$.

Proof of Property 2: By (B.32), $\frac{d\Pr\{\mathsf{x}\in\cdot\,|\,\mathsf{y}=\boldsymbol{y}\}}{d\mathscr{H}^{m_1}|_{\mathcal{E}_1}} = \frac{\theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y})}{\theta_{\mathsf{y}}^{m_2}(\boldsymbol{y})}$ for $\mathscr{H}^{m_2}|_{\mathcal{E}_2}$-almost every $\boldsymbol{y}$. Furthermore, by (6.10), $\frac{d\Pr\{\mathsf{x}\in\cdot\,|\,\mathsf{y}=\boldsymbol{y}\}}{d\mathscr{H}^{m_1}|_{\mathcal{E}_1}} = \theta_{(\mathsf{x}\,|\,\mathsf{y}=\boldsymbol{y})}^{m_1}$. Hence, we see that (7.25) holds.

Proof of Property 3: By (6.17), we have

$$\mathfrak{h}^{m_1}(\mathsf{x} \,|\, \mathsf{y} = \boldsymbol{y}) = -\int_{\mathcal{E}_1} \theta_{(\mathsf{x}\,|\,\mathsf{y}=\boldsymbol{y})}^{m_1}(\boldsymbol{x}) \log \theta_{(\mathsf{x}\,|\,\mathsf{y}=\boldsymbol{y})}^{m_1}(\boldsymbol{x})\,\mathrm{d}\mathscr{H}^{m_1}(\boldsymbol{x})\,.$$

The result (7.26) then follows by (7.25).

B.12 Proof of Theorem 7.12

Starting from (7.27), we have

$$\mathfrak{h}^{m-m_2}(\mathsf{x} \,|\, \mathsf{y}) = \int_{\mathcal{E}_2} \theta_{\mathsf{y}}^{m_2}(\boldsymbol{y})\,\mathfrak{h}^{m-m_2}(\mathsf{x} \,|\, \mathsf{y} = \boldsymbol{y})\,\mathrm{d}\mathscr{H}^{m_2}(\boldsymbol{y})$$

$$
\begin{aligned}
&\overset{(7.24)}{=} - \int_{\mathcal{E}_2} \theta_{\mathsf{y}}^{m_2}(\boldsymbol{y}) \bigg(\int_{\mathcal{E}(\boldsymbol{y})} \frac{\theta_{(\mathsf{x},\mathsf{y})}^{m}(\boldsymbol{x},\boldsymbol{y})}{\mathscr{J}_{\mathfrak{p}_{\mathsf{y}}}^{\mathcal{E}}(\boldsymbol{x},\boldsymbol{y})\, \theta_{\mathsf{y}}^{m_2}(\boldsymbol{y})} \\
&\qquad\qquad \times \log\bigg(\frac{\theta_{(\mathsf{x},\mathsf{y})}^{m}(\boldsymbol{x},\boldsymbol{y})}{\mathscr{J}_{\mathfrak{p}_{\mathsf{y}}}^{\mathcal{E}}(\boldsymbol{x},\boldsymbol{y})\, \theta_{\mathsf{y}}^{m_2}(\boldsymbol{y})} \bigg) \mathrm{d}\mathscr{H}^{m-m_2}(\boldsymbol{x}) \bigg) \mathrm{d}\mathscr{H}^{m_2}(\boldsymbol{y}) \\
&= - \int_{\mathcal{E}_2} \bigg(\int_{\mathcal{E}(\boldsymbol{y})} \frac{\theta_{(\mathsf{x},\mathsf{y})}^{m}(\boldsymbol{x},\boldsymbol{y})}{\mathscr{J}_{\mathfrak{p}_{\mathsf{y}}}^{\mathcal{E}}(\boldsymbol{x},\boldsymbol{y})} \\
&\qquad\qquad \times \log\bigg(\frac{\theta_{(\mathsf{x},\mathsf{y})}^{m}(\boldsymbol{x},\boldsymbol{y})}{\mathscr{J}_{\mathfrak{p}_{\mathsf{y}}}^{\mathcal{E}}(\boldsymbol{x},\boldsymbol{y})\, \theta_{\mathsf{y}}^{m_2}(\boldsymbol{y})} \bigg) \mathrm{d}\mathscr{H}^{m-m_2}(\boldsymbol{x}) \bigg) \mathrm{d}\mathscr{H}^{m_2}(\boldsymbol{y}) \\
&\overset{(a)}{=} - \int_{\mathcal{E}} \theta_{(\mathsf{x},\mathsf{y})}^{m}(\boldsymbol{x},\boldsymbol{y}) \log\bigg(\frac{\theta_{(\mathsf{x},\mathsf{y})}^{m}(\boldsymbol{x},\boldsymbol{y})}{\mathscr{J}_{\mathfrak{p}_{\mathsf{y}}}^{\mathcal{E}}(\boldsymbol{x},\boldsymbol{y})\, \theta_{\mathsf{y}}^{m_2}(\boldsymbol{y})} \bigg) \mathrm{d}\mathscr{H}^{m}(\boldsymbol{x},\boldsymbol{y}) \\
&= - \int_{\mathcal{E}} \theta_{(\mathsf{x},\mathsf{y})}^{m}(\boldsymbol{x},\boldsymbol{y}) \bigg(\log\bigg(\frac{\theta_{(\mathsf{x},\mathsf{y})}^{m}(\boldsymbol{x},\boldsymbol{y})}{\theta_{\mathsf{y}}^{m_2}(\boldsymbol{y})} \bigg) - \log \mathscr{J}_{\mathfrak{p}_{\mathsf{y}}}^{\mathcal{E}}(\boldsymbol{x},\boldsymbol{y}) \bigg) \mathrm{d}\mathscr{H}^{m}(\boldsymbol{x},\boldsymbol{y}) \\
&\overset{(6.12)}{=} -\mathbb{E}_{(\mathsf{x},\mathsf{y})}\bigg[\log\bigg(\frac{\theta_{(\mathsf{x},\mathsf{y})}^{m}(\mathsf{x},\mathsf{y})}{\theta_{\mathsf{y}}^{m_2}(\mathsf{y})} \bigg) \bigg] + \mathbb{E}_{(\mathsf{x},\mathsf{y})}\big[\log \mathscr{J}_{\mathfrak{p}_{\mathsf{y}}}^{\mathcal{E}}(\mathsf{x},\mathsf{y}) \big]
\end{aligned}
$$

where in (a) we used (B.21) with $\mathcal{A}_1 = \mathbb{R}^{M_1}$, $\mathcal{A}_2 = \mathbb{R}^{M_2}$, and

$$
g(\boldsymbol{x},\boldsymbol{y}) = \theta_{(\mathsf{x},\mathsf{y})}^{m}(\boldsymbol{x},\boldsymbol{y}) \log\bigg(\frac{\theta_{(\mathsf{x},\mathsf{y})}^{m}(\boldsymbol{x},\boldsymbol{y})}{\mathscr{J}_{\mathfrak{p}_{\mathsf{y}}}^{\mathcal{E}}(\boldsymbol{x},\boldsymbol{y})\, \theta_{\mathsf{y}}^{m_2}(\boldsymbol{y})} \bigg). \tag{B.33}
$$

(Here, $g(\boldsymbol{x},\boldsymbol{y})$ is $\mathscr{H}^m|_{\mathcal{E}}$-integrable by assumption, i.e., Condition (ii) in Theorem B.1 is satisfied.) Thus, (7.28) holds.

B.13 Proof of Theorem 7.13

By the product-compatibility of x and y, we have a support $\mathcal{E}_1$ of x, a support $\mathcal{E}_2$ of y, and a support $\mathcal{E} \subseteq \mathcal{E}_1 \times \mathcal{E}_2$ of (x,y) . Starting from (7.27) with $m = m_1 + m_2$, we obtain

$$
\begin{aligned}
\mathfrak{h}^{m_1}(\mathsf{x}\,|\,\mathsf{y}) &= \int_{\mathcal{E}_2} \theta_{\mathsf{y}}^{m_2}(\boldsymbol{y})\, \mathfrak{h}^{m_1}(\mathsf{x}\,|\,\mathsf{y} = \boldsymbol{y})\, \mathrm{d}\mathscr{H}^{m_2}(\boldsymbol{y}) \\
&\overset{(7.26)}{=} - \int_{\mathcal{E}_2} \theta_{\mathsf{y}}^{m_2}(\boldsymbol{y}) \bigg(\int_{\mathcal{E}_1} \frac{\theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y})}{\theta_{\mathsf{y}}^{m_2}(\boldsymbol{y})} \log\bigg(\frac{\theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y})}{\theta_{\mathsf{y}}^{m_2}(\boldsymbol{y})} \bigg) \mathrm{d}\mathscr{H}^{m_1}(\boldsymbol{x}) \bigg) \mathrm{d}\mathscr{H}^{m_2}(\boldsymbol{y}) \\
&= - \int_{\mathcal{E}_2} \bigg(\int_{\mathcal{E}_1} \theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y}) \log\bigg(\frac{\theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y})}{\theta_{\mathsf{y}}^{m_2}(\boldsymbol{y})} \bigg) \mathrm{d}\mathscr{H}^{m_1}(\boldsymbol{x}) \bigg) \mathrm{d}\mathscr{H}^{m_2}(\boldsymbol{y}) \\
&\overset{(a)}{=} - \int_{\mathbb{R}^{M_1+M_2}} \theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y}) \log\bigg(\frac{\theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y})}{\theta_{\mathsf{y}}^{m_2}(\boldsymbol{y})} \bigg) \mathrm{d}(\mathscr{H}^{m_1}|_{\mathcal{E}_1} \times \mathscr{H}^{m_2}|_{\mathcal{E}_2})(\boldsymbol{x},\boldsymbol{y}) \\
&\overset{(b)}{=} - \int_{\mathbb{R}^{M_1+M_2}} \theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y}) \log\bigg(\frac{\theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y})}{\theta_{\mathsf{y}}^{m_2}(\boldsymbol{y})} \bigg) \mathrm{d}\mathscr{H}^{m_1+m_2}|_{\mathcal{E}_1\times\mathcal{E}_2}(\boldsymbol{x},\boldsymbol{y}) \\
&\overset{(c)}{=} - \int_{\mathcal{E}} \theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y}) \log\bigg(\frac{\theta_{(\mathsf{x},\mathsf{y})}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y})}{\theta_{\mathsf{y}}^{m_2}(\boldsymbol{y})} \bigg) \mathrm{d}\mathscr{H}^{m_1+m_2}(\boldsymbol{x},\boldsymbol{y})
\end{aligned}
$$

$$\overset{(6.12)}{=} -\mathbb{E}_{(\mathsf{x},\mathsf{y})}\left[\log\left(\frac{\theta^{m_1+m_2}_{(\mathsf{x},\mathsf{y})}(\mathsf{x},\mathsf{y})}{\theta^{m_2}_{\mathsf{y}}(\mathsf{y})}\right)\right]$$

where (a) holds by Fubini's theorem, (b) holds because $\mathcal{E}_1$ and $\mathcal{E}_2$ are product-compatible, and (c) holds because $\mathcal{E} \subseteq \mathcal{E}_1 \times \mathcal{E}_2$ and because, by Property 4 in Corollary 6.10, $\theta^{m_1+m_2}_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y}) = 0$ for $\mathscr{H}^{m_1+m_2}$-almost all $(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{E}^c$. Thus, (7.29) holds.

B.14 Proof of Theorem 7.18

We first note that the product measure $\mu\mathsf{x}^{-1} \times \mu\mathsf{y}^{-1}$ can be interpreted as the joint measure of independent random variables $\tilde{\mathsf{x}}$ and $\tilde{\mathsf{y}}$, where $\tilde{\mathsf{x}}$ has the same distribution as x and $\tilde{\mathsf{y}}$ has the same distribution as y. Because x is m_1-rectifiable and y is m_2-rectifiable, the same holds for $\tilde{\mathsf{x}}$ and $\tilde{\mathsf{y}}$, respectively. Furthermore, the Hausdorff densities satisfy $\theta^{m_1}_{\tilde{\mathsf{x}}}(\boldsymbol{x}) = \theta^{m_1}_{\mathsf{x}}(\boldsymbol{x})$ and $\theta^{m_2}_{\tilde{\mathsf{y}}}(\boldsymbol{y}) = \theta^{m_2}_{\mathsf{y}}(\boldsymbol{y})$. By Properties 1–3 in Theorem 7.4, the joint random variable $(\tilde{\mathsf{x}},\tilde{\mathsf{y}})$ is (m_1+m_2)-rectifiable with (m_1+m_2)-dimensional Hausdorff density

$$\theta^{m_1+m_2}_{(\tilde{\mathsf{x}},\tilde{\mathsf{y}})}(\boldsymbol{x},\boldsymbol{y}) = \theta^{m_1}_{\tilde{\mathsf{x}}}(\boldsymbol{x})\theta^{m_2}_{\tilde{\mathsf{y}}}(\boldsymbol{y}) = \theta^{m_1}_{\mathsf{x}}(\boldsymbol{x})\theta^{m_2}_{\mathsf{y}}(\boldsymbol{y}) \tag{B.34}$$

and support $\mathcal{E}_1 \times \mathcal{E}_2$. The rectifiability of $(\tilde{\mathsf{x}},\tilde{\mathsf{y}})$ with support $\mathcal{E}_1 \times \mathcal{E}_2$ implies that the measure $\mu\mathsf{x}^{-1} \times \mu\mathsf{y}^{-1}$ is (m_1+m_2)-rectifiable and

$$\mu\mathsf{x}^{-1} \times \mu\mathsf{y}^{-1} \ll \mathscr{H}^{m_1+m_2}|_{\mathcal{E}_1\times\mathcal{E}_2}\,. \tag{B.35}$$

Proof of Statement 1 (case $m = m_1 + m_2$): For any measurable set $\mathcal{A} \subseteq \mathbb{R}^{M_1+M_2}$, we have

$$\begin{aligned}
\mu(\mathsf{x},\mathsf{y})^{-1}(\mathcal{A}) &\overset{(6.11)}{=} \int_{\mathcal{A}} \theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\mathscr{H}^m|_{\mathcal{E}}(\boldsymbol{x},\boldsymbol{y})\\
&\overset{(a)}{=} \int_{\mathcal{A}} \theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\mathscr{H}^m|_{\mathcal{E}_1\times\mathcal{E}_2}(\boldsymbol{x},\boldsymbol{y})\\
&\overset{(b)}{=} \int_{\mathcal{A}} \frac{\theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y})}{\theta^{m_1}_{\mathsf{x}}(\boldsymbol{x})\theta^{m_2}_{\mathsf{y}}(\boldsymbol{y})}\theta^{m_1}_{\mathsf{x}}(\boldsymbol{x})\theta^{m_2}_{\mathsf{y}}(\boldsymbol{y})\,\mathrm{d}\mathscr{H}^m|_{\mathcal{E}_1\times\mathcal{E}_2}(\boldsymbol{x},\boldsymbol{y})\\
&\overset{(B.34)}{=} \int_{\mathcal{A}} \frac{\theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y})}{\theta^{m_1}_{\mathsf{x}}(\boldsymbol{x})\theta^{m_2}_{\mathsf{y}}(\boldsymbol{y})}\theta^m_{(\tilde{\mathsf{x}},\tilde{\mathsf{y}})}(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\mathscr{H}^m|_{\mathcal{E}_1\times\mathcal{E}_2}(\boldsymbol{x},\boldsymbol{y})\\
&\overset{(c)}{=} \int_{\mathcal{A}} \frac{\theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y})}{\theta^{m_1}_{\mathsf{x}}(\boldsymbol{x})\theta^{m_2}_{\mathsf{y}}(\boldsymbol{y})}\,\mathrm{d}\big(\mu\mathsf{x}^{-1}\times\mu\mathsf{y}^{-1}\big)(\boldsymbol{x},\boldsymbol{y})\,.
\end{aligned} \tag{B.36}$$

Here, (a) holds because $\mathcal{E} \subseteq \mathcal{E}_1 \times \mathcal{E}_2$ and because, by Property 4 in Corollary 6.10, $\theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y}) = 0$ on $\mathcal{E}^c$, (b) holds because $\theta^{m_1}_{\mathsf{x}}(\boldsymbol{x})\theta^{m_2}_{\mathsf{y}}(\boldsymbol{y}) > 0$ on $\mathcal{E}_1 \times \mathcal{E}_2$, and (c) holds because, by (6.10), $\theta^m_{(\tilde{\mathsf{x}},\tilde{\mathsf{y}})} = \frac{d(\mu\mathsf{x}^{-1}\times\mu\mathsf{y}^{-1})}{\mathrm{d}\mathscr{H}^m|_{\mathcal{E}_1\times\mathcal{E}_2}}$ $\mathscr{H}^m|_{\mathcal{E}_1\times\mathcal{E}_2}$-almost everywhere. By (B.36), we obtain that $\mu(\mathsf{x},\mathsf{y})^{-1} \ll \mu\mathsf{x}^{-1} \times \mu\mathsf{y}^{-1}$ with Radon-Nikodym derivative

$$\frac{\mathrm{d}\mu(\mathsf{x},\mathsf{y})^{-1}}{\mathrm{d}\big(\mu\mathsf{x}^{-1}\times\mu\mathsf{y}^{-1}\big)}(\boldsymbol{x},\boldsymbol{y}) = \frac{\theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y})}{\theta^{m_1}_{\mathsf{x}}(\boldsymbol{x})\theta^{m_2}_{\mathsf{y}}(\boldsymbol{y})}\,. \tag{B.37}$$

Inserting (B.37) into (7.43) yields

$$\begin{aligned} I(\mathsf{x};\mathsf{y}) &= \int_{\mathbb{R}^{M_1+M_2}} \log\left(\frac{\theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y})}{\theta^{m_1}_{\mathsf{x}}(\boldsymbol{x})\theta^{m_2}_{\mathsf{y}}(\boldsymbol{y})}\right) \mathrm{d}\mu(\mathsf{x},\mathsf{y})^{-1}(\boldsymbol{x},\boldsymbol{y}) \\ &\overset{(6.10)}{=} \int_{\mathcal{E}} \theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y}) \log\left(\frac{\theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y})}{\theta^{m_1}_{\mathsf{x}}(\boldsymbol{x})\theta^{m_2}_{\mathsf{y}}(\boldsymbol{y})}\right) \mathrm{d}\mathscr{H}^m(\boldsymbol{x},\boldsymbol{y}) \end{aligned} \tag{B.38}$$

which is (7.44). Furthermore, we can rewrite (B.38) as

$$\begin{aligned} I(\mathsf{x};\mathsf{y}) &\overset{(6.12)}{=} \mathbb{E}_{(\mathsf{x},\mathsf{y})}\left[\log\left(\frac{\theta^m_{(\mathsf{x},\mathsf{y})}(\boldsymbol{x},\boldsymbol{y})}{\theta^{m_1}_{\mathsf{x}}(\boldsymbol{x})\theta^{m_2}_{\mathsf{y}}(\boldsymbol{y})}\right)\right] \\ &= \mathbb{E}_{(\mathsf{x},\mathsf{y})}[\log\theta^m_{(\mathsf{x},\mathsf{y})}(\mathsf{x},\mathsf{y})] - \mathbb{E}_{(\mathsf{x},\mathsf{y})}[\log\theta^{m_1}_{\mathsf{x}}(\mathsf{x})] - \mathbb{E}_{(\mathsf{x},\mathsf{y})}[\log\theta^{m_2}_{\mathsf{y}}(\mathsf{y})] \\ &= -\mathfrak{h}^m(\mathsf{x},\mathsf{y}) - \mathbb{E}_{\mathsf{x}}[\log\theta^{m_1}_{\mathsf{x}}(\mathsf{x})] - \mathbb{E}_{\mathsf{y}}[\log\theta^{m_2}_{\mathsf{y}}(\mathsf{y})] \\ &= -\mathfrak{h}^m(\mathsf{x},\mathsf{y}) + \mathfrak{h}^{m_1}(\mathsf{x}) + \mathfrak{h}^{m_2}(\mathsf{y}) \end{aligned} \tag{B.39}$$

which is (7.45).

Finally, we obtain (7.46) by inserting (7.36) into (B.39). The second expression in (7.46) is obtained by symmetry.

Proof of Statement 2 (case $m < m_1 + m_2$): We first show that $\mu(\mathsf{x},\mathsf{y})^{-1} \not\ll \mu\mathsf{x}^{-1} \times \mu\mathsf{y}^{-1}$. To this end, we assume that $\mu(\mathsf{x},\mathsf{y})^{-1} \ll \mu\mathsf{x}^{-1} \times \mu\mathsf{y}^{-1}$ and will obtain a contradiction. Using (B.35), we have $\mu(\mathsf{x},\mathsf{y})^{-1} \ll \mu\mathsf{x}^{-1} \times \mu\mathsf{y}^{-1} \ll \mathscr{H}^{m_1+m_2}|_{\mathcal{E}_1\times\mathcal{E}_2}$. By Property 4 in Lemma 6.3 and because $m_1 + m_2 > m$, we obtain $\mathscr{H}^{m_1+m_2}(\mathcal{E}) = 0$, and thus, $\mathscr{H}^{m_1+m_2}|_{\mathcal{E}_1\times\mathcal{E}_2}(\mathcal{E}) \le \mathscr{H}^{m_1+m_2}(\mathcal{E}) = 0$. On the other hand, $\mu(\mathsf{x},\mathsf{y})^{-1}(\mathcal{E}) = 1$. Thus, we have a contradiction to $\mu(\mathsf{x},\mathsf{y})^{-1} \ll \mathscr{H}^{m_1+m_2}|_{\mathcal{E}_1\times\mathcal{E}_2}$. Hence, $\mu(\mathsf{x},\mathsf{y})^{-1} \not\ll \mu\mathsf{x}^{-1} \times \mu\mathsf{y}^{-1}$ and, by (7.43), $I(\mathsf{x};\mathsf{y}) = \infty$.

B.15 Proof of Theorem 7.22

Property 1 follows immediately from Lemma 7.20. Indeed, because of (7.50), equation (7.49) can be rewritten as

$$\lim_{n\to\infty} \Pr\{\mathsf{x}_{1:n} \notin \mathcal{A}^{(n)}_\varepsilon\} = 0\,.$$

This implies that there exists $n_0 \in \mathbb{N}$ such that $\Pr\{\mathsf{x}_{1:n} \in \mathcal{A}^{(n)}_\varepsilon\} > 1-\delta$, for all $n > n_0$.

We will next prove Property 2. For $\boldsymbol{x}_{1:n} \in \mathcal{A}^{(n)}_\varepsilon$, it follows from (7.50) that

$$-\frac{1}{n}\sum_{i=1}^n \log\theta^m_{\mathsf{x}}(\boldsymbol{x}_i) - \mathfrak{h}^m(\mathsf{x}) \le \varepsilon$$

which implies

$$\exp\left(\sum_{i=1}^n \log\theta^m_{\mathsf{x}}(\boldsymbol{x}_i)\right) \ge e^{-n(\mathfrak{h}^m(\mathsf{x})+\varepsilon)}$$

or equivalently

$$\prod_{i=1}^{n} \theta_{\mathsf{x}}^{m}(\boldsymbol{x}_i) \geq e^{-n(\mathfrak{h}^m(\mathsf{x})+\varepsilon)} . \tag{B.40}$$

We have

$$\begin{aligned}
1 \geq\ & \Pr\{\mathsf{x}_{1:n} \in \mathcal{A}_\varepsilon^{(n)}\} \\
=\ & \int_{\mathcal{A}_\varepsilon^{(n)}} \theta_{(\mathsf{x}_1,\ldots,\mathsf{x}_n)}^{nm}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)\,\mathrm{d}\mathscr{H}^{nm}|_{\mathcal{E}^n}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) \\
\overset{(7.6)}{=} & \int_{\mathcal{A}_\varepsilon^{(n)}} \left(\prod_{i=1}^{n} \theta_{\mathsf{x}}^{m}(\boldsymbol{x}_i)\right) \mathrm{d}\mathscr{H}^{nm}|_{\mathcal{E}^n}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) \\
\overset{(B.40)}{\geq} & \int_{\mathcal{A}_\varepsilon^{(n)}} e^{-n(\mathfrak{h}^m(\mathsf{x})+\varepsilon)}\,\mathrm{d}\mathscr{H}^{nm}|_{\mathcal{E}^n}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) \\
=\ & e^{-n(\mathfrak{h}^m(\mathsf{x})+\varepsilon)} \mathscr{H}^{nm}|_{\mathcal{E}^n}(\mathcal{A}_\varepsilon^{(n)}) \\
\overset{(a)}{=}\ & e^{-n(\mathfrak{h}^m(\mathsf{x})+\varepsilon)} \mathscr{H}^{nm}(\mathcal{A}_\varepsilon^{(n)})
\end{aligned} \tag{B.41}$$

where (a) holds because $\mathcal{A}_\varepsilon^{(n)} \subseteq \mathcal{E}^n$. The inequality (B.41) is equivalent to (7.51).

It remains to prove Property 3. For $\boldsymbol{x}_{1:n} \in \mathcal{A}_\varepsilon^{(n)}$, it follows from (7.50) that

$$-\frac{1}{n}\sum_{i=1}^{n} \log \theta_{\mathsf{x}}^{m}(\boldsymbol{x}_i) - \mathfrak{h}^m(\mathsf{x}) \geq -\varepsilon$$

which implies

$$\prod_{i=1}^{n} \theta_{\mathsf{x}}^{m}(\boldsymbol{x}_i) \leq e^{-n(\mathfrak{h}^m(\mathsf{x})-\varepsilon)} . \tag{B.42}$$

By Property 1, there exists $n_0 \in \mathbb{N}$ such that $\Pr\{\mathsf{x}_{1:n} \in \mathcal{A}_\varepsilon^{(n)}\} > 1-\delta$ for all $n > n_0$. Thus, we have for $n > n_0$

$$\begin{aligned}
1-\delta <\ & \Pr\{\mathsf{x}_{1:n} \in \mathcal{A}_\varepsilon^{(n)}\} \\
=\ & \int_{\mathcal{A}_\varepsilon^{(n)}} \theta_{(\mathsf{x}_1,\ldots,\mathsf{x}_n)}^{nm}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)\,\mathrm{d}\mathscr{H}^{nm}|_{\mathcal{E}^n}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) \\
\overset{(7.6)}{=} & \int_{\mathcal{A}_\varepsilon^{(n)}} \left(\prod_{i=1}^{n} \theta_{\mathsf{x}}^{m}(\boldsymbol{x}_i)\right) \mathrm{d}\mathscr{H}^{nm}|_{\mathcal{E}^n}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) \\
\overset{(B.42)}{\leq} & \int_{\mathcal{A}_\varepsilon^{(n)}} e^{-n(\mathfrak{h}^m(\mathsf{x})-\varepsilon)}\,\mathrm{d}\mathscr{H}^{nm}|_{\mathcal{E}^n}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) \\
=\ & e^{-n(\mathfrak{h}^m(\mathsf{x})-\varepsilon)} \mathscr{H}^{nm}|_{\mathcal{E}^n}(\mathcal{A}_\varepsilon^{(n)}) \\
\overset{(a)}{=}\ & e^{-n(\mathfrak{h}^m(\mathsf{x})-\varepsilon)} \mathscr{H}^{nm}(\mathcal{A}_\varepsilon^{(n)})
\end{aligned} \tag{B.43}$$

where (a) holds because $\mathcal{A}_\varepsilon^{(n)} \subseteq \mathcal{E}^n$. The inequality (B.43) is equivalent to (7.52).

B.16 Proof of Lemma 8.2

We will use two alternative characterizations of the Kullback-Leibler divergence $D_{\mathrm{KL}}(\mu\|\nu)$ between probability measures μ and ν on $\mathbb{R}^M$. Usually, the Kullback-Leibler divergence is defined by [Kullback and Leibler, 1951]

$$D_{\mathrm{KL}}(\mu\|\nu) \triangleq \begin{cases} \int_{\mathbb{R}^M} \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\right) \mathrm{d}\mu(x) & \text{if } \mu \ll \nu \\ \infty & \text{else.} \end{cases} \tag{B.44}$$

By the Gelfand-Yaglom-Perez theorem [Gray, 1990, Lem. 5.2.3], this definition coincides with the following: Let $\mathfrak{P}$ denote the set of all finite, measurable partitions of $\mathbb{R}^M$, i.e., for $\mathfrak{Q} = \{\mathcal{A}_1, \dots, \mathcal{A}_N\} \in \mathfrak{P}$ the sets $\mathcal{A}_i$ are mutually disjoint and measurable and satisfy $\bigcup_{i=1}^N \mathcal{A}_i = \mathbb{R}^M$. Then

$$D_{\mathrm{KL}}(\mu\|\nu) = \sup_{\mathfrak{Q}\in\mathfrak{P}} \sum_{\mathcal{A}\in\mathfrak{Q}} \mu(\mathcal{A}) \log\left(\frac{\mu(\mathcal{A})}{\nu(\mathcal{A})}\right). \tag{B.45}$$

For our setting, we have to generalize the equivalence of the definitions (B.44) and (B.45) to measures ν that are not necessarily probability measures. Although the Kullback-Leibler divergence is usually defined only for probability measures, we will use (B.44) to define the Kullback-Leibler divergence $D_{\mathrm{KL}}(\mu\|\nu)$ for an arbitrary finite measure ν. Based on this definition, we then obtain again the expression (B.45):

Lemma B.2 Let μ be a probability measure and ν be a finite measure on $\mathbb{R}^M$, i.e., $\nu(\mathbb{R}^M) < \infty$. Then

$$D_{\mathrm{KL}}(\mu\|\nu) = \sup_{\mathfrak{Q}\in\mathfrak{P}} \sum_{\mathcal{A}\in\mathfrak{Q}} \mu(\mathcal{A}) \log\left(\frac{\mu(\mathcal{A})}{\nu(\mathcal{A})}\right). \tag{B.46}$$

Proof. We consider the cases $\mu \ll \nu$ and $\mu \not\ll \nu$ separately.

Case $\mu \not\ll \nu$: In this case, there exists a set $\mathcal{A}_0$ such that $\mu(\mathcal{A}_0) > 0$ and $\nu(\mathcal{A}_0) = 0$. We then choose the partition $\mathfrak{Q}_0 = \{\mathcal{A}_0, \mathcal{A}_0^c\}$ and obtain

$$\begin{aligned} \sup_{\mathfrak{Q}\in\mathfrak{P}} \sum_{\mathcal{A}\in\mathfrak{Q}} \mu(\mathcal{A}) \log\left(\frac{\mu(\mathcal{A})}{\nu(\mathcal{A})}\right) &\geq \sum_{\mathcal{A}\in\mathfrak{Q}_0} \mu(\mathcal{A}) \log\left(\frac{\mu(\mathcal{A})}{\nu(\mathcal{A})}\right) \\ &= \mu(\mathcal{A}_0) \log\left(\frac{\mu(\mathcal{A}_0)}{\nu(\mathcal{A}_0)}\right) + \mu(\mathcal{A}_0^c) \log\left(\frac{\mu(\mathcal{A}_0^c)}{\nu(\mathcal{A}_0^c)}\right) \\ &= \infty \end{aligned}$$

which coincides with $D_{\mathrm{KL}}(\mu\|\nu)$ according to (B.44).

Case $\mu \ll \nu$: We first note that $\nu(\mathbb{R}^M) \neq 0$. Indeed, assuming $\nu(\mathbb{R}^M) = 0$ would imply $\mu(\mathbb{R}^M) = 0$ due to the absolute continuity $\mu \ll \nu$. This is a contradiction since we assumed that μ is a probability measure, i.e., $\mu(\mathbb{R}^M) = 1$.

The measure $\hat{\nu}$ defined by $\hat{\nu}(\mathcal{A}) \triangleq \nu(\mathcal{A})/\nu(\mathbb{R}^M)$ is a probability measure and, by (B.45),

$$\begin{aligned} D_{\mathrm{KL}}(\mu\|\hat{\nu}) &= \sup_{\mathfrak{Q}\in\mathfrak{P}} \sum_{\mathcal{A}\in\mathfrak{Q}} \mu(\mathcal{A}) \log\left(\frac{\mu(\mathcal{A})}{\hat{\nu}(\mathcal{A})}\right) \\ &= \sup_{\mathfrak{Q}\in\mathfrak{P}} \sum_{\mathcal{A}\in\mathfrak{Q}} \mu(\mathcal{A}) \log\left(\frac{\mu(\mathcal{A})\nu(\mathbb{R}^M)}{\nu(\mathcal{A})}\right) \\ &= \sup_{\mathfrak{Q}\in\mathfrak{P}} \left(\sum_{\mathcal{A}\in\mathfrak{Q}} \mu(\mathcal{A}) \log \nu(\mathbb{R}^M) + \sum_{\mathcal{A}\in\mathfrak{Q}} \mu(\mathcal{A}) \log\left(\frac{\mu(\mathcal{A})}{\nu(\mathcal{A})}\right)\right) \\ &\overset{(a)}{=} \log \nu(\mathbb{R}^M) + \sup_{\mathfrak{Q}\in\mathfrak{P}} \sum_{\mathcal{A}\in\mathfrak{Q}} \mu(\mathcal{A}) \log\left(\frac{\mu(\mathcal{A})}{\nu(\mathcal{A})}\right) \end{aligned} \tag{B.47}$$

where (a) holds because $\sum_{\mathcal{A}\in\mathfrak{Q}} \mu(\mathcal{A}) = \mu(\mathbb{R}^M) = 1$ for all $\mathfrak{Q} \in \mathfrak{P}$. On the other hand, by (B.44), we have

$$\begin{aligned} D_{\mathrm{KL}}(\mu\|\hat{\nu}) &= \int_{\mathbb{R}^M} \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\hat{\nu}}(x)\right) \mathrm{d}\mu(x) \\ &\overset{(a)}{=} \int_{\mathbb{R}^M} \log\left(\nu(\mathbb{R}^M)\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\right) \mathrm{d}\mu(x) \\ &= \int_{\mathbb{R}^M} \log \nu(\mathbb{R}^M)\, \mathrm{d}\mu(x) + \int_{\mathbb{R}^M} \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\right) \mathrm{d}\mu(x) \\ &\overset{(b)}{=} \log \nu(\mathbb{R}^M) + \int_{\mathbb{R}^M} \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\right) \mathrm{d}\mu(x) \\ &= \log \nu(\mathbb{R}^M) + D_{\mathrm{KL}}(\mu\|\nu)\,. \end{aligned} \tag{B.48}$$

Here, (a) holds because $\frac{\mathrm{d}\mu}{\mathrm{d}\hat{\nu}} = \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\frac{\mathrm{d}\nu}{\mathrm{d}\hat{\nu}}$ and $\frac{\mathrm{d}\nu}{\mathrm{d}\hat{\nu}} = \nu(\mathbb{R}^M)$, and (b) holds because $\int_{\mathbb{R}^M} \mathrm{d}\mu(x) = \mu(\mathbb{R}^M) = 1$. Combining (B.47) and (B.48), we obtain

$$\log \nu(\mathbb{R}^M) + \sup_{\mathfrak{Q}\in\mathfrak{P}} \sum_{\mathcal{A}\in\mathfrak{Q}} \mu(\mathcal{A}) \log\left(\frac{\mu(\mathcal{A})}{\nu(\mathcal{A})}\right) = \log \nu(\mathbb{R}^M) + D_{\mathrm{KL}}(\mu\|\nu)\,. \tag{B.49}$$

Here, $\log \nu(\mathbb{R}^M)$ is finite because $\nu(\mathbb{R}^M)$ is nonzero and finite. Therefore, (B.49) implies (B.46). □

We now proceed to the proof of Lemma 8.2. For an m-rectifiable random variable x with support $\mathcal{E}$ satisfying $\mathscr{H}^m(\mathcal{E}) < \infty$, we can rewrite the m-dimensional entropy $\mathfrak{h}^m(\mathsf{x})$ using (B.46). To this end, we first note that $\mathfrak{h}^m(\mathsf{x})$ can be interpreted as a negative Kullback-Leibler divergence with respect to the Hausdorff measure $\mathscr{H}^m|_{\mathcal{E}}$. Starting from (6.16), we obtain

$$\begin{aligned} \mathfrak{h}^m(\mathsf{x}) &= -\int_{\mathbb{R}^M} \log \theta^m_{\mathsf{x}}(\boldsymbol{x})\, \mathrm{d}\mu\mathsf{x}^{-1}(\boldsymbol{x}) \\ &\overset{(6.10)}{=} -\int_{\mathbb{R}^M} \log\left(\frac{\mathrm{d}\mu\mathsf{x}^{-1}}{\mathrm{d}\mathscr{H}^m|_{\mathcal{E}}}(\boldsymbol{x})\right) \mathrm{d}\mu\mathsf{x}^{-1}(\boldsymbol{x}) \\ &\overset{(B.44)}{=} -D_{\mathrm{KL}}\big(\mu\mathsf{x}^{-1}\|\mathscr{H}^m|_{\mathcal{E}}\big)\,. \end{aligned}$$

Here, for the last step, we used that $\mathcal{E}$ is a support of $\mathbf{x}$, which implies $\mu\mathbf{x}^{-1} \ll \mathscr{H}^m|_{\mathcal{E}}$ (see Definition 6.9). By (B.46), we thus obtain

$$\mathfrak{h}^m(\mathbf{x}) = -\sup_{\mathfrak{Q}\in\mathfrak{P}} \sum_{\mathcal{A}\in\mathfrak{Q}} \mu\mathbf{x}^{-1}(\mathcal{A}) \log\left(\frac{\mu\mathbf{x}^{-1}(\mathcal{A})}{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A})}\right). \tag{B.50}$$

Because $\mu\mathbf{x}^{-1}(\mathcal{E}^c) = 0$ and $\mathscr{H}^m|_{\mathcal{E}}(\mathcal{E}^c) = 0$, we have for all $\mathfrak{Q} \in \mathfrak{P}$

$$\begin{aligned}\sum_{\mathcal{A}\in\mathfrak{Q}} \mu\mathbf{x}^{-1}(\mathcal{A}) \log\left(\frac{\mu\mathbf{x}^{-1}(\mathcal{A})}{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A})}\right) &= \sum_{\mathcal{A}\in\mathfrak{Q}} \mu\mathbf{x}^{-1}(\mathcal{A}\cap\mathcal{E}) \log\left(\frac{\mu\mathbf{x}^{-1}(\mathcal{A}\cap\mathcal{E})}{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}\cap\mathcal{E})}\right)\\ &= \sum_{\mathcal{A}'\in\widetilde{\mathfrak{Q}}} \mu\mathbf{x}^{-1}(\mathcal{A}') \log\left(\frac{\mu\mathbf{x}^{-1}(\mathcal{A}')}{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}')}\right)\end{aligned}$$

where $\widetilde{\mathfrak{Q}} \triangleq \{\mathcal{A}\cap\mathcal{E} : \mathcal{A}\in\mathfrak{Q}\} \in \mathfrak{P}^{(\mathcal{E})}_{m,\infty}$. Hence, the supremum in (B.50) does not change if we replace $\mathfrak{P}$ by $\mathfrak{P}^{(\mathcal{E})}_{m,\infty}$, and thus,

$$\mathfrak{h}^m(\mathbf{x}) = -\sup_{\mathfrak{Q}\in\mathfrak{P}^{(\mathcal{E})}_{m,\infty}} \sum_{\mathcal{A}\in\mathfrak{Q}} \mu\mathbf{x}^{-1}(\mathcal{A}) \log\left(\frac{\mu\mathbf{x}^{-1}(\mathcal{A})}{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A})}\right). \tag{B.51}$$

Swapping minus sign and supremum in (B.51), we obtain

$$\mathfrak{h}^m(\mathbf{x}) = \inf_{\mathfrak{Q}\in\mathfrak{P}^{(\mathcal{E})}_{m,\infty}} \left(-\sum_{\mathcal{A}\in\mathfrak{Q}} \mu\mathbf{x}^{-1}(\mathcal{A}) \log\left(\frac{\mu\mathbf{x}^{-1}(\mathcal{A})}{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A})}\right)\right) \tag{B.52}$$

$$= \inf_{\mathfrak{Q}\in\mathfrak{P}^{(\mathcal{E})}_{m,\infty}} \left(-\sum_{\mathcal{A}\in\mathfrak{Q}} \mu\mathbf{x}^{-1}(\mathcal{A}) \log\mu\mathbf{x}^{-1}(\mathcal{A}) + \sum_{\mathcal{A}\in\mathfrak{Q}} \mu\mathbf{x}^{-1}(\mathcal{A}) \log\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A})\right)$$

$$= \inf_{\mathfrak{Q}\in\mathfrak{P}^{(\mathcal{E})}_{m,\infty}} \left(H([\mathbf{x}]_{\mathfrak{Q}}) + \sum_{\mathcal{A}\in\mathfrak{Q}} \mu\mathbf{x}^{-1}(\mathcal{A}) \log\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A})\right). \tag{B.53}$$

Here, (B.52) is (8.2) and (B.53) is (8.3).

B.17 Proof of Theorem 8.3

B.17.1 Proof of Lower Bound (8.4)

Let $\mathfrak{Q} \in \mathfrak{P}^{(\mathcal{E})}_{m,\delta}$ be an (m,δ)-partition of $\mathcal{E}$ according to Definition 8.1, i.e., $\mathfrak{Q} = \{\mathcal{A}_1,\ldots,\mathcal{A}_N\}$ where $\mathscr{H}^m(\mathcal{A}_i) \le \delta$ for all $i \in \{1,\ldots,N\}$ and $\bigcup_{i=1}^N \mathcal{A}_i = \mathcal{E}$. Note that $\mathfrak{Q}$ also belongs to $\mathfrak{P}^{(\mathcal{E})}_{m,\infty}$. Then, starting from (8.3), we obtain

$$\begin{aligned}\mathfrak{h}^m(\mathbf{x}) &= \inf_{\mathfrak{Q}'\in\mathfrak{P}^{(\mathcal{E})}_{m,\infty}} \left(H([\mathbf{x}]_{\mathfrak{Q}'}) + \sum_{\mathcal{A}\in\mathfrak{Q}'} \mu\mathbf{x}^{-1}(\mathcal{A}) \log\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A})\right)\\ &\le H([\mathbf{x}]_{\mathfrak{Q}}) + \sum_{i=1}^N \mu\mathbf{x}^{-1}(\mathcal{A}_i) \log\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_i)\end{aligned}$$

$$\overset{(a)}{\leq} H([\mathsf{x}]_{\mathfrak{Q}}) + \sum_{i=1}^{N} \mu\mathsf{x}^{-1}(\mathcal{A}_i) \log \delta$$
$$\overset{(b)}{=} H([\mathsf{x}]_{\mathfrak{Q}}) + \log \delta$$

where (a) holds because $\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_i) \leq \delta$ and (b) holds because $\sum_{i=1}^{N} \mu\mathsf{x}^{-1}(\mathcal{A}_i) = \mu\mathsf{x}^{-1}(\mathcal{E}) = 1$. Multiplying by $\operatorname{ld} e$, we have equivalently

$$(\mathfrak{h}^m(\mathsf{x}) - \log \delta) \operatorname{ld} e \leq H([\mathsf{x}]_{\mathfrak{Q}}) \operatorname{ld} e \,. \tag{B.54}$$

By (8.1), we have

$$H([\mathsf{x}]_{\mathfrak{Q}}) \operatorname{ld} e \leq L^*([\mathsf{x}]_{\mathfrak{Q}}) \,. \tag{B.55}$$

Combining (B.54) and (B.55), we obtain

$$(\mathfrak{h}^m(\mathsf{x}) - \log \delta) \operatorname{ld} e \leq L^*([\mathsf{x}]_{\mathfrak{Q}})$$

which implies (8.4).

B.17.2 Proof of Upper Bound (8.5)

We first state two preliminary results.

Lemma B.3 Let x be an m-rectifiable random variable with $m \geq 1$ and support $\mathcal{E}$ satisfying $\mathscr{H}^m(\mathcal{E}) < \infty$. Furthermore, let $\mathfrak{Q} = \{\mathcal{A}_1, \ldots, \mathcal{A}_N\} \in \mathfrak{P}_{m,\infty}^{(\mathcal{E})}$ satisfy $\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_i) \neq 0$ for $i \in \{1, \ldots, N\}$. Let $\mathcal{A}_i$ be the disjoint union of $\mathcal{A}_{i,1}, \ldots, \mathcal{A}_{i,k_i}$, i.e., $\mathcal{A}_i = \bigcup_{j=1}^{k_i} \mathcal{A}_{i,j}$ and $\mathcal{A}_{i,j_1} \cap \mathcal{A}_{i,j_2} = \emptyset$ for $j_1 \neq j_2$. Finally, let $\widetilde{\mathfrak{Q}} \triangleq \{\mathcal{A}_{1,1}, \ldots, \mathcal{A}_{1,k_1}, \ldots, \mathcal{A}_{N,1}, \ldots, \mathcal{A}_{N,k_N}\}$. Then

$$-\sum_{\mathcal{A} \in \mathfrak{Q}} \mu\mathsf{x}^{-1}(\mathcal{A}) \log\left(\frac{\mu\mathsf{x}^{-1}(\mathcal{A})}{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A})}\right) \geq -\sum_{\mathcal{A} \in \widetilde{\mathfrak{Q}}} \mu\mathsf{x}^{-1}(\mathcal{A}) \log\left(\frac{\mu\mathsf{x}^{-1}(\mathcal{A})}{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A})}\right) \tag{B.56}$$

Proof. The inequality (B.56) can be written as

$$-\sum_{i=1}^{N} \mu\mathsf{x}^{-1}(\mathcal{A}_i) \log\left(\frac{\mu\mathsf{x}^{-1}(\mathcal{A}_i)}{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_i)}\right) \geq -\sum_{i=1}^{N} \sum_{j=1}^{k_i} \mu\mathsf{x}^{-1}(\mathcal{A}_{i,j}) \log\left(\frac{\mu\mathsf{x}^{-1}(\mathcal{A}_{i,j})}{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_{i,j})}\right).$$

Therefore, it suffices to show that

$$\mu\mathsf{x}^{-1}(\mathcal{A}_i) \log\left(\frac{\mu\mathsf{x}^{-1}(\mathcal{A}_i)}{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_i)}\right) \leq \sum_{j=1}^{k_i} \mu\mathsf{x}^{-1}(\mathcal{A}_{i,j}) \log\left(\frac{\mu\mathsf{x}^{-1}(\mathcal{A}_{i,j})}{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_{i,j})}\right)$$

for $i \in \{1, \ldots, N\}$. This latter inequality follows from the log sum inequality [Cover and Thomas, 2006, Th. 2.7.1]. □

Lemma B.4 Let x be an m-rectifiable random variable with $m \geq 1$ and support $\mathcal{E}$ satisfying $\mathscr{H}^m(\mathcal{E}) < \infty$. For $\varepsilon > 0$ there exists a partition $\mathfrak{Q} \in \mathfrak{P}_{m,\infty}^{(\mathcal{E})}$ such that $\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}) \neq 0$ for $\mathcal{A} \in \mathfrak{Q}$ and

$$\mathfrak{h}^m(\mathsf{x}) > -\sum_{\mathcal{A}\in\mathfrak{Q}} \mu\mathsf{x}^{-1}(\mathcal{A}) \log\left(\frac{\mu\mathsf{x}^{-1}(\mathcal{A})}{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A})}\right) - \frac{\varepsilon}{2\,\mathrm{ld}\,e}\,. \tag{B.57}$$

Proof. By (8.2), there exists a partition $\widetilde{\mathfrak{Q}} = \{\mathcal{A}_1, \dots, \mathcal{A}_N\} \in \mathfrak{P}_{m,\infty}^{(\mathcal{E})}$ such that

$$\mathfrak{h}^m(\mathsf{x}) > -\sum_{\mathcal{A}\in\widetilde{\mathfrak{Q}}} \mu\mathsf{x}^{-1}(\mathcal{A}) \log\left(\frac{\mu\mathsf{x}^{-1}(\mathcal{A})}{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A})}\right) - \frac{\varepsilon}{2\,\mathrm{ld}\,e}\,. \tag{B.58}$$

If $\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_i) \neq 0$ for $i \in \{1, \dots, N\}$ we are done. Otherwise, assume without loss of generality that $\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_i) = 0$ for $i \in \{k+1, \dots, N\}$, and $\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_i) \neq 0$ for $i \in \{1, \dots, k\}$. Because $\mu\mathsf{x}^{-1} \ll \mathscr{H}^m|_{\mathcal{E}}$, we also have $\mu\mathsf{x}^{-1}(\mathcal{A}_i) = 0$ for $i \in \{k+1, \dots, N\}$. We define $\mathfrak{Q} \triangleq \big\{\mathcal{A}_1 \cup \bigcup_{i=k+1}^N \mathcal{A}_i, \mathcal{A}_2, \dots, \mathcal{A}_k\big\}$. Note that $\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}) > 0$ for $\mathcal{A} \in \mathfrak{Q}$. Furthermore,

$$\begin{aligned}
&-\sum_{\mathcal{A}\in\mathfrak{Q}} \mu\mathsf{x}^{-1}(\mathcal{A}) \log\left(\frac{\mu\mathsf{x}^{-1}(\mathcal{A})}{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A})}\right)\\
&\qquad= -\mu\mathsf{x}^{-1}\bigg(\mathcal{A}_1 \cup \bigcup_{i=k+1}^N \mathcal{A}_i\bigg) \log\left(\frac{\mu\mathsf{x}^{-1}\big(\mathcal{A}_1 \cup \bigcup_{i=k+1}^N \mathcal{A}_i\big)}{\mathscr{H}^m|_{\mathcal{E}}\big(\mathcal{A}_1 \cup \bigcup_{i=k+1}^N \mathcal{A}_i\big)}\right)\\
&\qquad\qquad - \sum_{i=2}^k \mu\mathsf{x}^{-1}(\mathcal{A}_i) \log\left(\frac{\mu\mathsf{x}^{-1}(\mathcal{A}_i)}{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_i)}\right)\\
&\qquad\overset{(a)}{=} -\mu\mathsf{x}^{-1}(\mathcal{A}_1) \log\left(\frac{\mu\mathsf{x}^{-1}(\mathcal{A}_1)}{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_1)}\right) - \sum_{i=2}^k \mu\mathsf{x}^{-1}(\mathcal{A}_i) \log\left(\frac{\mu\mathsf{x}^{-1}(\mathcal{A}_i)}{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_i)}\right)\\
&\qquad\overset{(b)}{=} -\sum_{i=1}^N \mu\mathsf{x}^{-1}(\mathcal{A}_i) \log\left(\frac{\mu\mathsf{x}^{-1}(\mathcal{A}_i)}{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_i)}\right)\\
&\qquad= -\sum_{\mathcal{A}\in\widetilde{\mathfrak{Q}}} \mu\mathsf{x}^{-1}(\mathcal{A}) \log\left(\frac{\mu\mathsf{x}^{-1}(\mathcal{A})}{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A})}\right)
\end{aligned} \tag{B.59}$$

where (a) holds because

$$\mu\mathsf{x}^{-1}\bigg(\mathcal{A}_1 \cup \bigcup_{i=k+1}^N \mathcal{A}_i\bigg) = \mu\mathsf{x}^{-1}(\mathcal{A}_1) + \underbrace{\mu\mathsf{x}^{-1}\bigg(\bigcup_{i=k+1}^N \mathcal{A}_i \setminus \mathcal{A}_1\bigg)}_{=0} = \mu\mathsf{x}^{-1}(\mathcal{A}_1)$$

and

$$\mathscr{H}^m|_{\mathcal{E}}\bigg(\mathcal{A}_1 \cup \bigcup_{i=k+1}^N \mathcal{A}_i\bigg) = \mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_1) + \underbrace{\mathscr{H}^m|_{\mathcal{E}}\bigg(\bigcup_{i=k+1}^N \mathcal{A}_i \setminus \mathcal{A}_1\bigg)}_{=0} = \mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_1)$$

and (b) holds because $\mu\mathbf{x}^{-1}(\mathcal{A}_i) = 0$ for $i \in \{k+1, \ldots, N\}$. Inserting (B.59) into (B.58) results in (B.57). $\square$

We now proceed to the proof of (8.5). Let $\mathfrak{Q} = \{\mathcal{A}_1, \ldots, \mathcal{A}_N\} \in \mathfrak{P}^{(\mathcal{E})}_{m,\infty}$ satisfy (B.57) and $\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_i) \neq 0$ (which exists due to Lemma B.4). Furthermore, let $\varepsilon' \triangleq \frac{\varepsilon}{2\operatorname{ld} e}$. We define

$$\delta_\varepsilon \triangleq \left(1 - e^{-\varepsilon'}\right) \min_{i \in \{1,\ldots,N\}} \mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_i) \tag{B.60}$$

which is nonzero because $\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_i) \neq 0$. For $\delta \in (0, \delta_\varepsilon)$, we partition each set $\mathcal{A}_i$ into $\left\lceil \frac{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_i)}{\delta} \right\rceil$ disjoint subsets $\mathcal{A}_{i,j}$ of equal Hausdorff measure, i.e.,

$$\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_{i,j}) = \frac{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_i)}{\left\lceil \frac{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_i)}{\delta} \right\rceil}\,.$$

Using the shorthand notation $J_{i,\delta} \triangleq \frac{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_i)}{\delta}$, we obtain

$$\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_{i,j}) = \frac{J_{i,\delta}}{\lceil J_{i,\delta} \rceil}\,\delta \leq \delta\,. \tag{B.61}$$

Let us denote by $\mathfrak{Q}_\delta$ the partition of $\mathcal{E}$ containing all sets $\mathcal{A}_{i,j}$. Then (B.61) implies $\mathfrak{Q}_\delta \in \mathfrak{P}^{(\mathcal{E})}_{m,\delta}$. Furthermore,

$$\begin{aligned}
\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_{i,j}) &= \frac{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_i)}{\lceil J_{i,\delta} \rceil} \\
&= \frac{J_{i,\delta}}{\lceil J_{i,\delta} \rceil}\,\delta \\
&= \frac{\lceil J_{i,\delta} \rceil - \left(\lceil J_{i,\delta} \rceil - J_{i,\delta}\right)}{\lceil J_{i,\delta} \rceil}\,\delta \\
&= \left(1 - \frac{\lceil J_{i,\delta} \rceil - J_{i,\delta}}{\lceil J_{i,\delta} \rceil}\right)\delta \\
&\overset{(a)}{>} \left(1 - \frac{1}{\lceil J_{i,\delta} \rceil}\right)\delta
\end{aligned} \tag{B.62}$$

where (a) holds because $\lceil J_{i,\delta} \rceil - J_{i,\delta} < 1$. We can bound $\lceil J_{i,\delta} \rceil$ as

$$\lceil J_{i,\delta} \rceil \geq \frac{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_i)}{\delta} > \frac{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_i)}{\delta_\varepsilon} \overset{\text{(B.60)}}{=} \frac{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_i)}{(1 - e^{-\varepsilon'}) \min_{i' \in \{1,\ldots,N\}} \mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_{i'})} \geq \frac{1}{1 - e^{-\varepsilon'}}\,. \tag{B.63}$$

Inserting (B.63) into (B.62), we obtain for all sets $\mathcal{A}_{i,j} \in \mathfrak{Q}_\delta$

$$\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A}_{i,j}) > \left(1 - \frac{1}{\frac{1}{1 - e^{-\varepsilon'}}}\right)\delta = e^{-\varepsilon'}\,\delta\,. \tag{B.64}$$

Combining our results yields

$$\begin{aligned}
\mathfrak{h}^m(\mathbf{x}) &\overset{\text{(B.57)}}{>} - \sum_{\mathcal{A}\in\mathfrak{Q}} \mu\mathbf{x}^{-1}(\mathcal{A}) \log\left(\frac{\mu\mathbf{x}^{-1}(\mathcal{A})}{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A})}\right) - \varepsilon' \\
&\overset{\text{(B.56)}}{\geq} - \sum_{\mathcal{A}\in\mathfrak{Q}_\delta} \mu\mathbf{x}^{-1}(\mathcal{A}) \log\left(\frac{\mu\mathbf{x}^{-1}(\mathcal{A})}{\mathscr{H}^m|_{\mathcal{E}}(\mathcal{A})}\right) - \varepsilon' \\
&\overset{\text{(B.64)}}{>} - \sum_{\mathcal{A}\in\mathfrak{Q}_\delta} \mu\mathbf{x}^{-1}(\mathcal{A}) \log\left(\frac{\mu\mathbf{x}^{-1}(\mathcal{A})}{e^{-\varepsilon'}\delta}\right) - \varepsilon' \\
&= - \sum_{\mathcal{A}\in\mathfrak{Q}_\delta} \mu\mathbf{x}^{-1}(\mathcal{A}) \log\left(\mu\mathbf{x}^{-1}(\mathcal{A})\right) + (\log\delta - \varepsilon') \sum_{\mathcal{A}\in\mathfrak{Q}_\delta} \mu\mathbf{x}^{-1}(\mathcal{A}) - \varepsilon' \\
&\overset{(a)}{=} - \sum_{\mathcal{A}\in\mathfrak{Q}_\delta} \mu\mathbf{x}^{-1}(\mathcal{A}) \log\left(\mu\mathbf{x}^{-1}(\mathcal{A})\right) + \log\delta - 2\varepsilon' \\
&= H([\mathbf{x}]_{\mathfrak{Q}_\delta}) + \log\delta - 2\varepsilon' \\
&\overset{(b)}{>} \frac{L^*([\mathbf{x}]_{\mathfrak{Q}_\delta}) - 1}{\operatorname{ld} e} + \log\delta - 2\varepsilon'
\end{aligned} \tag{B.65}$$

where (a) holds because $\sum_{\mathcal{A}\in\mathfrak{Q}_\delta} \mu\mathbf{x}^{-1}(\mathcal{A}) = \mu\mathbf{x}^{-1}(\mathcal{E}) = 1$ and (b) holds by the second inequality in (8.1). Finally, rewriting (B.65) (recall $\varepsilon' = \frac{\varepsilon}{2\operatorname{ld} e}$) gives

$$L^*([\mathbf{x}]_{\mathfrak{Q}_\delta}) < \mathfrak{h}^m(\mathbf{x}) \operatorname{ld} e - \log\delta \operatorname{ld} e + 1 + \varepsilon$$

which is (8.5).

B.18 Proof of Lemma 8.6

Because $\mathfrak{h}^m(\mathbf{x})$ is finite and does not depend on s, it is sufficient to show $\lim_{s\to\infty} sD + \log\gamma(s) = \infty$. We have

$$\begin{aligned}
sD + \log\gamma(s) &\overset{\text{(8.12)}}{=} sD + \log\left(\sup_{\boldsymbol{y}\in\mathbb{R}^M} \int_{\mathcal{E}} e^{-sd(\boldsymbol{x},\boldsymbol{y})} \,\mathrm{d}\mathscr{H}^m(\boldsymbol{x})\right) \\
&\overset{(a)}{=} \sup_{\boldsymbol{y}\in\mathbb{R}^M} sD + \log\left(\int_{\mathcal{E}} e^{-sd(\boldsymbol{x},\boldsymbol{y})} \,\mathrm{d}\mathscr{H}^m(\boldsymbol{x})\right) \\
&= \sup_{\boldsymbol{y}\in\mathbb{R}^M} \log\left(e^{sD} \int_{\mathcal{E}} e^{-sd(\boldsymbol{x},\boldsymbol{y})} \,\mathrm{d}\mathscr{H}^m(\boldsymbol{x})\right) \\
&= \sup_{\boldsymbol{y}\in\mathbb{R}^M} \log\left(\int_{\mathcal{E}} e^{s(D-d(\boldsymbol{x},\boldsymbol{y}))} \,\mathrm{d}\mathscr{H}^m(\boldsymbol{x})\right) \\
&\overset{(b)}{\geq} \sup_{\boldsymbol{y}\in\mathbb{R}^M} \log\left(\int_{\mathcal{E}\cap\mathcal{B}_{D/2}(\boldsymbol{y})} e^{s(D-d(\boldsymbol{x},\boldsymbol{y}))} \,\mathrm{d}\mathscr{H}^m(\boldsymbol{x})\right) \\
&\overset{(c)}{\geq} \sup_{\boldsymbol{y}\in\mathbb{R}^M} \log\left(\int_{\mathcal{E}\cap\mathcal{B}_{D/2}(\boldsymbol{y})} e^{s\frac{D}{2}} \,\mathrm{d}\mathscr{H}^m(\boldsymbol{x})\right)
\end{aligned}$$

$$= s\,\frac{D}{2} + \sup_{\boldsymbol{y}\in\mathbb{R}^M} \log \mathscr{H}^m\big(\mathcal{E}\cap\mathcal{B}_{\frac{D}{2}}(\boldsymbol{y})\big) \tag{B.66}$$

where (a) holds because $\log$ is a monotonically increasing function, (b) holds because $\exp$ is a positive function, and (c) holds because $d(\boldsymbol{x},\boldsymbol{y})) < D/2$ for all $\boldsymbol{x} \in \mathcal{B}_{\frac{D}{2}}(\boldsymbol{y})$. Because $\mu\mathbf{x}^{-1}(\mathcal{E}) = 1$ (see (6.13)) the absolute continuity $\mu\mathbf{x}^{-1} \ll \mathscr{H}^m|_{\mathcal{E}}$ implies $\mathscr{H}^m(\mathcal{E}) > 0$. Thus, there exists a $\bar{\boldsymbol{y}} \in \mathbb{R}^M$ such that $\delta \triangleq \mathscr{H}^m\big(\mathcal{E}\cap\mathcal{B}_{\frac{D}{2}}(\bar{\boldsymbol{y}})\big) > 0$. Hence, by (B.66),

$$sD + \log\gamma(s) \geq s\,\frac{D}{2} + \log\delta\,.$$

This yields for any $K > 0$

$$sD + \log\gamma(s) \geq K\,.$$

for $s > \frac{2K}{D} - \log\delta$, i.e., $\lim_{s\to\infty} sD + \log\gamma(s) = \infty$.

B.19 Proof of Theorem 8.9

Consider the source $\mathbf{x}$ as specified in Theorem 8.9. By the source coding theorem [Gray, 1990, Th. 11.4.1], every code for $\mathbf{x}$ with expected distortion D must have a rate[1] greater than or equal to $R(D)$. Thus, if we can find an encoding function $f\colon \mathbb{R}^2 \to \{1,\dots,n\}$ and a decoding function $g\colon \{1,\dots,n\} \to \mathbb{R}^2$ such that $\mathbb{E}_{\mathbf{x}}\big[\|\mathbf{x} - g(f(\mathbf{x})\|^2\big] \leq D$, then the rate, $\log n$, must satisfy $\log n \geq R(D)$.

We directly design the composed function $q_n \triangleq g\circ f$. Because $\mathbf{x}$ has probability zero outside $\mathcal{S}_1$, we only have to define q_n on the unit circle. Furthermore, because f maps $\boldsymbol{x}$ to a set of at most n distinct values, $q_n = g\circ f$ can also attain at most n distinct values. We define q_n to map each circle segment defined by an angle interval $\big[i\frac{2\pi}{n}, (i+1)\frac{2\pi}{n}\big)$, $i \in \{0,\dots,n-1\}$, onto one associated "center" point, which is not constrained to lie on the unit circle. To this end, we only have to consider the circle segment defined by $\{\boldsymbol{x} = (\cos\phi\ \sin\phi)^{\mathrm{T}} : \phi \in [-\pi/n, \pi/n)\}$ since the problem is invariant under rotations. Because of symmetry, we choose the "center" associated with this segment to be some point $(x_1\ 0)^{\mathrm{T}}$, i.e., $q_n(\boldsymbol{x}) = (x_1\ 0)^{\mathrm{T}}$ for all $\boldsymbol{x} = (\cos\phi\ \sin\phi)^{\mathrm{T}}$ with $\phi \in [-\pi/n, \pi/n)$. The expected distortion is then obtained as

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}}\big[\|\mathbf{x} - q_n(\mathbf{x})\|^2\big] &= \int_0^{2\pi} \frac{1}{2\pi} \left\| \begin{pmatrix}\cos\phi\\ \sin\phi\end{pmatrix} - q_n\!\left(\begin{pmatrix}\cos\phi\\ \sin\phi\end{pmatrix}\right)\right\|^2 \mathrm{d}\phi \\
&= \frac{n}{2\pi}\int_{-\pi/n}^{\pi/n} \left\| \begin{pmatrix}\cos\phi\\ \sin\phi\end{pmatrix} - \begin{pmatrix}x_1\\ 0\end{pmatrix}\right\|^2 \mathrm{d}\phi \\
&= \frac{n}{2\pi}\int_{-\pi/n}^{\pi/n} \big((\cos\phi - x_1)^2 + \sin^2\phi\big)\,\mathrm{d}\phi
\end{aligned}$$

[1]Because we use the natural logarithm, the rate is measured in nats, i.e., the rate is defined using the natural logarithm. Furthermore, we do not code several symbols at once (i.e., we do not use a block code), but code only one symbol at a time. Thus the rate, which is number of nats used per source symbol is simply the logarithm of the number of codewords.

$$= \frac{n}{2\pi} \int_{-\pi/n}^{\pi/n} \left(1 + x_1^2 - 2x_1 \cos\phi\right) \mathrm{d}\phi$$

$$= 1 + x_1^2 - \frac{2nx_1}{\pi} \sin\frac{\pi}{n} \,. \tag{B.67}$$

Minimizing the expected distortion with respect to x_1 gives the optimum value of x_1 as

$$x_1^* = \frac{n}{\pi} \sin\frac{\pi}{n} \,. \tag{B.68}$$

The corresponding quantization function will be denoted by q_n^*. Inserting (B.68) into (B.67) yields $\bar{D}(n)$ in (8.26):

$$\begin{aligned}\mathbb{E}_{\mathbf{x}}\left[\|\mathbf{x} - q_n^*(\mathbf{x})\|^2\right] &= 1 + \left(\frac{n}{\pi}\sin\frac{\pi}{n}\right)^2 - 2\left(\frac{n}{\pi}\sin\frac{\pi}{n}\right)^2 \\ &= 1 - \left(\frac{n}{\pi}\sin\frac{\pi}{n}\right)^2 \\ &= \bar{D}(n) \,.\end{aligned}$$

Thus, we found a code with expected distortion $\bar{D}(n)$. The rate of this code is $\log n$. Hence, the inequality (8.25) is satisfied.

List of Abbreviations

CSI channel state information

MIMO multiple-input multiple-output

SNR signal-to-noise ratio

SIMO single-input multiple-output

SISO single-input single-output

RD rate-distortion

Bibliography

[Adhikary et al., 2013] Adhikary, A., Nam, J., Ahn, J.-Y., and Caire, G. (2013). Joint spatial division and multiplexing–the large-scale array regime. *IEEE Trans. Inf. Theory*, 59(10):6441–6463.

[Alexeev et al., 2012] Alexeev, B., Cahill, J., and Mixon, D. G. (2012). Full spark frames. *J Fourier Anal Appl*, 18(6):1167–1194.

[Ambrosio et al., 2000] Ambrosio, L., Fusco, N., and Pallara, D. (2000). *Functions of Bounded Variation and Free Discontinuity Problems*. Oxford Univ. Press, New York, NY.

[Azarin, 2009] Azarin, V. (2009). *Growth Theory of Subharmonic Functions*. Birkhäuser, Basel, Switzerland.

[Cover and Thomas, 2006] Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley, New York, NY, 2nd edition.

[Csiszár, 1974] Csiszár, I. (1974). On an extremum problem of information theory. *Studia Scientiarum Mathematicarum Hungarica*, 9(1):57–71.

[Csiszár, 2008] Csiszár, I. (2008). Axiomatic characterizations of information measures. *Entropy*, 10:261–273.

[De Lellis, 2008] De Lellis, C. (2008). *Rectifiable sets, densities and tangent measures*. Zurich Lectures in Advanced Mathematics. EMS Publishing House.

[Dörpinghaus et al., 2014] Dörpinghaus, M., Koliander, G., Durisi, G., Riegler, E., and Meyr, H. (2014). Oversampling increases the pre-log of noncoherent rayleigh fading channels. *IEEE Trans. Inf. Theory*, 60(9):5673–5681.

[Durisi and Bölcskei, 2011] Durisi, G. and Bölcskei, H. (2011). High-SNR capacity of wireless communication channels in the noncoherent setting: A primer. *Int. J. Electron. Commun. (AEÜ)*, 65(8):707–712.

[Durisi et al., 2012] Durisi, G., Morgenshtern, V. I., and Bölcskei, H. (2012). On the sensitivity of continuous-time noncoherent fading channel capacity. *IEEE Trans. Inf. Theory*, 58(10):6372–6391.

[Federer, 1969] Federer, H. (1969). *Geometric Measure Theory*. Springer, New York, NY.

[Gallager, 1968] Gallager, R. G. (1968). *Information Theory and Reliable Communication*. Wiley, New York, NY.

[Ghozlan and Kramer, 2013] Ghozlan, H. and Kramer, G. (2013). On Wiener phase noise channels at high signal-to-noise ratio. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pages 2279–2283, Istanbul, Turkey.

[Gray, 1990] Gray, R. M. (1990). *Entropy and Information Theory*. Springer, New York, NY, first edition.

[Gray, 1990] Gray, R. M. (1990). *Source Coding Theory*. Kluwer, Boston.

[Gray, 2010] Gray, R. M. (2010). *Probability, Random Processes, and Ergodic Properties*. Springer, New York.

[Gunning and Rossi, 1965] Gunning, R. C. and Rossi, H. (1965). *Analytic Functions of Several Complex Variables*. Prentice-Hall, Englewood Cliffs, NJ.

[Hardy and Wright, 1975] Hardy, G. and Wright, E. (1975). *An Introduction to the Theory of Numbers*. Oxford Univ. Press, Oxford, UK, fourth edition.

[Horn and Johnson, 1985] Horn, R. A. and Johnson, C. R. (1985). *Matrix Analysis*. Cambridge Univ. Press, Cambridge, UK.

[Jafar, 2011] Jafar, S. A. (2011). *Interference Alignment: A New Look at Signal Dimensions in a Communication Network*, volume 7 of *Foundations and Trends in Communications and Information Theory*. now publisher.

[Kawabata and Dembo, 1994] Kawabata, T. and Dembo, A. (1994). The rate-distortion dimension of sets and measures. *IEEE Trans. Inf. Theory*, 40(5):1564–1572.

[Koch, 2009] Koch, T. (2009). *On Heating Up and Fading in Communication Channels*, volume 5 of *Information Theory and its Applications*. Hartung-Gorre Verlag, Konstanz, Germany.

[Koliander et al., 2015] Koliander, G., Pichler, G., Riegler, E., and Hlawatsch, F. (2015). Entropy and source coding for integer-dimensional singular random variables. *to be submitted to IEEE Trans. Inf. Theory*.

[Koliander et al., 2014] Koliander, G., Riegler, E., Durisi, G., and Hlawatsch, F. (2014). Degrees of freedom of generic block-fading MIMO channels without a priori channel state information. *IEEE Trans. Inf. Theory*, 60(12):7760–7781.

[Kolmogorov, 1956] Kolmogorov, A. (1956). On the Shannon theory of information transmission in the case of continuous signals. *IRE Trans. Inf. Theory*, 2(4):102–108.

[Kullback and Leibler, 1951] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86.

[Lapidoth, 2005] Lapidoth, A. (2005). On the asymptotic capacity of stationary Gaussian fading channels. *IEEE Trans. Inf. Theory*, 51(2):437–446.

[Lapidoth and Moser, 2003] Lapidoth, A. and Moser, S. M. (2003). Capacity bounds via duality with applications to multiple-antenna systems on flat-fading channels. *IEEE Trans. Inf. Theory*, 49(10):2426–2467.

[Lapidoth and Moser, 2006] Lapidoth, A. and Moser, S. M. (2006). The fading number of single-input multiple-output fading channels with memory. *IEEE Trans. Inf. Theory*, 52(2):437–453.

[Liang and Veeravalli, 2004] Liang, Y. and Veeravalli, V. V. (2004). Capacity of noncoherent time-selective Rayleigh-fading channels. *IEEE Trans. Inf. Theory*, 50(12):3095–3110.

[Marzetta and Hochwald, 1999] Marzetta, T. L. and Hochwald, B. M. (1999). Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading. *IEEE Trans. Inf. Theory*, 45(1):139–157.

[Morgenshtern et al., 2010] Morgenshtern, V. I., Durisi, G., and Bölcskei, H. (2010). The SIMO pre-log can be larger than the SISO pre-log. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT 2010)*, pages 320–324, Austin, TX.

[Morgenshtern et al., 2013] Morgenshtern, V. I., Riegler, E., Yang, W., Durisi, G., Lin, S., Sturmfels, B., and Bölcskei, H. (2013). Capacity pre-log of noncoherent SIMO channels via Hironaka's theorem. *IEEE Trans. Inf. Theory*, 59(7):4213–4229.

[Moser, 2009] Moser, S. M. (2009). The fading number of multiple-input multiple-output fading channels with memory. *IEEE Trans. Inf. Theory*, 55(6):2716–2755.

[Neeser and Massey, 1993] Neeser, F. and Massey, J. (1993). Proper complex random processes with applications to information theory. *IEEE Trans. Inf. Theory*, 39(4):1293–1302.

[Posner and Rodemich, 1971] Posner, C. E. and Rodemich, E. R. (1971). Epsilon entropy and data compression. *Ann. Math. Statist.*, 42(6):2079–2125.

[Preiss, 1987] Preiss, D. (1987). Geometry of measures in R^n: Distribution, rectifiability, and densities. *Ann. Math.*, 125(3):537–643.

[Rényi, 1959] Rényi, A. (1959). On the dimension and entropy of probability distributions. *Acta Math. Hung.*, 10(1):193–215.

[Riegler et al., 2011] Riegler, E., Morgenshtern, V. I., Durisi, G., Lin, S., Sturmfels, B., and Bölcskei, H. (2011). Noncoherent SIMO pre-log via resolution of singularities. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT 2011)*, pages 2020–2024, St. Petersburg, Russia.

[Rudin, 1976] Rudin, W. (1976). *Principles of Mathematical Analysis*. McGraw-Hill, New York, NY, 3rd edition.

[Rudin, 1987] Rudin, W. (1987). *Real and Complex Analysis*. McGraw-Hill, New York, NY, 3rd edition.

[Rusek et al., 2013] Rusek, F., Persson, D., Lau, B. K., Larsson, E. G., Marzetta, T. L., Edfors, O., and Tufvesson, F. (2013). Scaling up MIMO: Opportunities and challenges with very large arrays. *IEEE Signal Processing Mag.*, 30(1):40–60.

[Schuster et al., 2009] Schuster, U. G., Durisi, G., Bölcskei, H., and Poor, H. V. (2009). Capacity bounds for peak-constrained multiantenna wideband channels. *IEEE Trans. Commun.*, 57(9):2686–2696.

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423, 623–656.

[Stotz and Bölcskei, 2012] Stotz, D. and Bölcskei, H. (2012). Degrees of freedom in vector interference channels. In *Proc. Allerton Conf. Comm. Control Comput.*, pages 1755–1760.

[Telatar, 1999] Telatar, I. E. (1999). Capacity of multi-antenna Gaussian channels. *Eur. Trans. Telecomm.*, 10(6):585–595.

[van den Essen, 2000] van den Essen, A. R. P. (2000). *Polynomial Automorphisms and the Jacobian Conjecture*. Birkhäuser, Basel, Switzerland.

[Wu and Verdú, 2010] Wu, Y. and Verdú, S. (2010). Rényi information dimension: Fundamental limits of almost lossless analog compression. *IEEE Trans. Inf. Theory*, 56(8):3721–3748.

[Yang et al., 2013] Yang, W., Durisi, G., and Riegler, E. (2013). On the capacity of large-MIMO block-fading channels. *IEEE J. Sel. Areas Commun.*, 31(2):117–132.

[Zheng and Tse, 2002] Zheng, L. and Tse, D. (2002). Communication on the Grassmann manifold: A geometric approach to the noncoherent multiple-antenna channel. *IEEE Trans. Inf. Theory*, 48(2):359–383.