Maximizing Entropy with an Expectation Constraint and One-Parameter Exponential Families of Distributions: A Reexamination

Other titles in Foundations and Trends® in Communications and Information Theory

Codes for Adversaries: Between Worst-Case and Average-Case Jamming
Bikash Kumar Dey, Sidharth Jaggi, Michael Langberg,
Anand D. Sarwate and Yihan Zhang
ISBN: 978-1-63828-460-4

Ultra-Reliable Low-Latency Communications: Foundations, Enablers, System Design, and Evolution Towards 6G
Nurul Huda Mahmood, Italo Atzeni, Eduard Axel Jorswieck and
Onel Luis Alcaraz López
ISBN: 978-1-63828-180-1

Probabilistic Amplitude Shaping
Georg Böcherer
ISBN: 978-1-63828-178-8

Reed-Muller Codes
Emmanuel Abbe, Ori Sberlo, Amir Shpilka and Min Ye
ISBN: 978-1-63828-144-3

Topics and Techniques in Distribution Testing: A Biased but Representative Sample
Clément L. Canonne
ISBN: 978-1-63828-100-9

Codes in the Sum-Rank Metric: Fundamentals and Applications
Umberto Martínez-Peñas, Mohannad Shehadeh and
Frank R. Kschischang
ISBN: 978-1-63828-030-9

Maximizing Entropy with an Expectation Constraint and One-Parameter Exponential Families of Distributions: A Reexamination

David L. Neuhoff
University of Michigan
neuhoff@umich.edu

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Communications and Information Theory

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

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

The preferred citation for this publication is

D. L. Neuhoff. *Maximizing Entropy with an Expectation Constraint and One-Parameter Exponential Families of Distributions: A Reexamination.* Foundations and Trends® in Communications and Information Theory, vol. 21, no. 5, pp. 589–846, 2024.

ISBN: 978-1-63828-481-9

Foundations and Trends® in Communications and Information Theory
Volume 21, Issue 5, 2024

Editorial Board

Editorial Scope

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

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

Information for Librarians

Foundations and Trends® in Communications and Information Theory, 2024, Volume 21, 4 issues. ISSN paper version 1567-2190. ISSN online version 1567-2328 . Also available as a combined paper and online subscription.

Contents

Maximizing Entropy with an Expectation Constraint and One-Parameter Exponential Families of Distributions: A Reexamination

David L. Neuhoff

Department of Electrical Engineering and Computer Science, University of Michigan, USA; neuhoff@umich.edu

ABSTRACT

The usual answer to the question "What probability distribution maximizes entropy or differential entropy of a random variable X subject to the constraint that the expected value of a real-valued function g applied to X has a specified value μ?" is an exponential distribution (probability mass or probability density function), with $g(x)$ in the exponent multiplied by a parameter λ, and with the parameter chosen so the exponential distribution causes the expected value of $g(X)$ to equal μ. The latter is called *moment matching*. While it is well-known that, when there are multiple expected value constraints, there are functions and expected value specifications for which moment matching is not possible, it is not well-known that this can happen when there is a single expected value constraint and a single parameter.

This motivates the present monograph, whose goal is to reexammine the question posed above, and to derive its

David L. Neuhoff (2024), "Maximizing Entropy with an Expectation Constraint and One-Parameter Exponential Families of Distributions: A Reexamination", Foundations and Trends® in Communications and Information Theory: Vol. 21, No. 5, pp 589–846. DOI: 10.1561/0100000132.

answer in an accessible, self-contained and complete fashion. It also derives the maximum entropy/differential entropy when there is a constraint on the support of the probability distributions, when there is only a bound on expected value and when there is a variance constraint. Properties of the resulting maximum entropy/differential entropy as a function of μ are derived, such as its convexity and its monotonicities. Example functions are presented, including many for which moment matching is possible for all relevant values of μ, and some for which it is not. Indeed, there can be only subtle differences between the two kinds of functions.

As one-parameter exponential probability distributions play a dominant role, one section of this monograph provides a self-contained discussion and derivation of their properties, such as the finiteness and continuity of the exponential normalizing constant (sometimes called the partition function) as λ varies, the finiteness, continuity, monotonicity and limits of the expected value of $g(X)$ under the exponential distribution as λ varies, and similar issues for entropy and differential entropy. Most of these are needed in deriving the maximum entropy/differential entropy or the properties of the resulting function of μ.

Aside from addressing the question posed initially, this monograph can be viewed as a warmup for discussions of maximizing entropy/differential entropy with multiple expected value constraints and of multiparameter exponential families. It also provides a small taste of information geometry.

1

Introduction

1.1 Goals

The first goal of this monograph is to derive, in an accessible, self-contained and complete fashion, the largest possible *entropy*,[1] $H(X)$, of a discrete random variable X for which an alphabet A_X is specified (for example, $\{1, 2, \ldots, n\}$ or $\{1, 2, 3, \ldots\}$), and the expected value, $E[g(X)]$, of some real-valued function g applied to X is constrained to have a specified value μ. The second goal is to derive, in the same manner, the largest possible *differential entropy*, $H_d(X)$, of a real-valued continuous random variable X with, again, a constraint that the expected value of some real-valued function g has a specified value μ, and also a constraint that the support of X be a subset of some specified set S (for example, $S = (-\infty, \infty), [0, \infty)$ or $[0, 1]$). As one-parameter exponential families of probability distributions play a principal role in the solution to the maximum entropy and differential entropy questions, the last goal is to carefully derive their principal properties.

Entropy, $H(X)$, of a discrete random variable X is an information-theoretic measure of the randomness of, or uncertainty in, the outcome

[1] Italics indicates a term with a meaning that is either defined here or will be defined shortly.

of X. It is defined by the formula

$$H(X) \triangleq -\sum_{x \in A_X} p(x) \log p(x), \tag{1.1}$$

where A_X is the alphabet of X, i.e., the set of its possible outcomes, p is its probability mass function (pmf) ($p(x) \triangleq \Pr(X = x)$), and the logarithm is base-2. Indeed, all logarithms in this monograph are base-2 unless specified otherwise. Entropy can also be viewed as a property or function of the pmf p, and consequently p will often be added as a subscript, as in $H_p(X)$.

Similarly, the differential entropy, $H_d(X)$, of a continuous, real-valued random variable X is an information-theoretic relative measure of the randomness of, or uncertainty in, the outcome of X. It is defined by the formula

$$H_d(X) = -\int_{-\infty}^{\infty} p(x) \log p(x) dx, \tag{1.2}$$

where p is the probability density function (pdf) of X (for example, $\Pr(a \leq X \leq b) = \int_a^b p(x)\, dx$), and the logarithm is again base-2. Differential entropy can also be viewed as a property or function of the pdf p, and consequently, p will often be added as a subscript, as in $H_{d,p}(X)$. As for entropy, the appendix provides justification for viewing differential entropy as a relative measure of randomness or uncertainty, and discusses some of its operational significance, for example, in lossy source coding.

Following next, Sections 1.2–1.4 are intended to give a general idea of what is to be found in this monograph, and how its contents relate to the literature on maximizing entropy and differential entropy subject to expected value constraints. Since these subsections are not self-contained discussions, they might not be well understood by a newcomer to the subject. Accessible, self-contained explanations come in subsequent sections. Section 1.5 provides examples of applications where maximizing entropy subject to expected value constraints are useful. Section 1.6 describes the closely related task of minimizing divergence subject to an expected value constraint. Section 1.7 provides the rationale for why this monograph focuses exclusively on maximizing entropy and differential entropy of just one random variable with just one expected value constraint. Finally, Section 1.8 discusses the manner

in which this monograph aims to be accessible, self-contained and complete.

1.2 Maximizing Entropy

This subsection focuses on the first stated goal of this monograph, namely, maximizing entropy of a discrete random variable subject to an expected value constraint. Maximizing differential entropy is deferred to the next subsection.

For most discrete[2] alphabets A_X, real-valued functions g, and specified values μ, Section 2 of this monograph will carefully derive the well-known result[3] that the largest possible entropy of a discrete random variable X with alphabet A_X and expected value of $g(X)$ equal to μ occurs when, and only when, X has a probability mass function with the exponential form

$$q_\lambda(x) = \frac{2^{-\lambda g(x)}}{\Sigma_\lambda}, \quad x \in A_X, \tag{1.3}$$

and the resulting largest entropy is

$$H_{max}(\mu) = \lambda\mu - \log \Sigma_\lambda, \tag{1.4}$$

where $\Sigma_\lambda = \sum_{x \in A_X} 2^{-\lambda g(x)}$ is a normalizer chosen to make q_λ sum to one, and λ is a real value chosen to make the expected value of $g(X)$ equal μ when q_λ governs the distribution of X. The resulting distribution is commonly said to be a *moment-matching exponential-form* distribution.

Let us call attention to the fact that, in the previous paragraph, we qualified the well-known result with for "most" alphabets and functions. We did this because the above result applies when and only when

(i) There exists at least one pmf p such that $E_p[g(X)] = \mu$,[4] and

[2] In this monograph, a *discrete* set is a finite or countably infinite set.

[3] It can be found in information theory textbooks, see for example, [34, p. 296], [76, Prob. 1.8], [12, Chap. 11], [78, p. 308], [13, Chap. 12], [101, Sec. 2.9]), [82, pp. 99,100], and in books, articles and websites devoted to the maximum entropy principle and to the many application areas for which the maximum entropy principle is useful, see for example, [3], [8], [26], [29], [35]–[37], [46]–[49], [64], [77], [81], [84]–[87], [89], [92], [96], [97], [100].

[4] As with H for entropy, we will usually subscript an E denoting expected value with the distribution that governs the random variable.

(ii) There exists a λ that makes the expected value of $g(X)$ equal μ when q_λ governs the distribution of X.

As reviewed in Section 5, it is also well known that the above result generalizes to scenarios in which the expected values of N real-valued functions $g_1, \ldots, g_N$ on discrete alphabet A_X are specified to have values $\mu_1, \ldots, \mu_N$, sees for example, [12, Chap. 11], [13, Chap. 12], and also to scenarios in which X is a discrete-time, discrete-valued stationary random process $X_1, X_2, \ldots$ for which it is desired to maximize *entropy-rate*[5] subject to a constraint that the expected value of some real-valued function g applied to any one of the random variables equals a specified value μ [77].

Let us focus on the former scenario, for which it is known that, in most instances, the largest entropy of X is well known to be

$$H_{max}(\mu_1, \ldots, \mu_N) = \sum_{n=1}^{N} \lambda_n \mu_n - \log\left(\sum_{x \in A_X} 2^{-\lambda_n g_n x)}\right), \tag{1.5}$$

where $\lambda_1, \ldots, \lambda_N$ are chosen so the following exponential-form probability mass function $q_{\lambda_1,\ldots,\lambda_N}$ causes $E_{q_{\lambda_1,\ldots,\lambda_N}}[g_n(X)] = \mu_n$, for $n = 1, \ldots, N$:

$$q_{\lambda_1,\ldots,\lambda_N}(x) = \frac{2^{-\sum_{n=1}^{N} \lambda_n g_n(x)}}{\Sigma_{\lambda_1,\ldots,\lambda_N}}, \tag{1.6}$$

where

$$\Sigma_{\lambda_1,\ldots,\lambda_N} = \sum_{x \in A_X} 2^{-\sum_{n=1}^{N} \lambda_n g_n(x)}. \tag{1.7}$$

This, again, is called moment-matching. As with the previous scenario having just one expected value constraint, it is important to note the result expressed in (1.5) and (1.6) applies only when

(i) There exists a pmf p on A_X such that $E_p[g_n(X)] = \mu_n$ for $n = 1, \ldots, N$, and consequently, $H_{max}(\mu_1, \ldots, \mu_N)$ is well-defined, and

[5] The entropy-rate of a discrete-time, discrete-valued stationary random process is $H_\infty(X) = \lim_{N \to \infty} \frac{1}{N} H(X_1, \ldots, X_N)$.

(ii) There exists $\lambda_1, \ldots, \lambda_N$ that cause the desired moment-matching, which includes the requirement that $\Sigma_{\lambda_1,\ldots,\lambda_N}$ be finite.

Let us now consider the feasibility of satisfying conditions (i) and (ii), both for one expected value constraint ($N = 1$), and for multiple expected value constraints ($N \geq 2$).

In regard to Condition (i), it is straightforwardly shown in the Appendix, that when $N = 1$ there is a pmf p such that $E_p[g(X)] = \mu$ whenever μ lies between the infimum and supremum values of g, and also sometimes when μ is the infimum or supremum of g. On the other hand, when $N \geq 2$, it is easy to construct a set of real-valued functions $g_1, \ldots, g_N$ on A_X, a set of values $\mu_1, \ldots, \mu_N$, and a set of pmfs $p_1, \ldots, p_N$ on A_X such that for each $n \in \{1, \ldots, N\}$, μ_n lies between the infimum and supremum values of g_n, and $E_{p_n}[g_n(X)] = \mu_n$, but there is no single pmf p on A_X such that $E_p[g_n(X)] = \mu_n$ for each n. (A simple example is $A_X = \{a, b\}$, $g_1(a) = 1$, $g_1(b) = 0$, $g_2(a) = 0$, $g_2(b) = 1$, and $\mu_1 = \mu_2 = 1$.) We conclude that, when $N \geq 2$, $H_{max}(\mu_1, \ldots, \mu_N)$ is well-defined only for a subset of the possible $\mu_1, \ldots, \mu_N$ values. We call the set of such $\mu_1, \ldots, \mu_N$ the *range of* $(E[g_1(X)], \ldots, E[g_N(X)])$, and also, the *domain of* H_{max}. In summary, the domain of H_{max} is easily found for $N = 1$ and can be nontrivial for $N \geq 2$.

In regard to Condition (ii), it is has long been known that there exists $N \geq 2$, a set of functions $g_1, \ldots, g_N$ and a set of values $\mu_1, \ldots, \mu_N$ in the domain of H_{max} for which there does not exist $\lambda_1, \ldots, \lambda_N$ that causes the desired moment matching. See, for example, [18], [7, p. 86], [16]. Thus, the solution expressed in (1.5) and (1.6) is not a panacea for finding $H_{max}(\mu_1, \ldots, \mu_N)$ in all cases, but rather a solution that works in many cases. In particular, the solution in (1.5) and (1.6) applies to every $(\mu_1, \ldots, \mu_N)$ in the set of possible values of $(E_{q_{\underline{\lambda}}}[g_1(X)], \ldots, E_{q_{\underline{\lambda}}}[g_N(X)])$ as $\underline{\lambda} = (\lambda_1, \ldots, \lambda_N)$ varies. Let us call this the *exponential mean space*.

Let us now return to the case of one expected value constraint, i.e., $N = 1$, which is the main focus of this monograph. Here, it is natural to wonder if the exponential mean space is something simple, just as the domain of H_{max} is simple when $N = 1$.

On the one hand, if the alphabet A_X is finite, it is straightforward to show (see, for example, Section 2.1), that for any μ in the domain

of H_{max}, there is a value λ such that $E_{q_\lambda}[g(X)] = \mu$, and consequently, $H_{max}(\mu)$ is given by (1.4) and the maximizing pmf is given by (1.3). That is, the exponential moment space simply equals the domain of H_{max}.

On the other hand, when the alphabet A_X is countably infinite, the existence of a moment-matching exponential-form pmf is not at all obvious. For one thing, Σ_λ can be infinite for all λ, in which case q_λ and $E_{q_\lambda}[g(X)]$ are undefined for all λ. For example, this happens when g is bounded.

Accordingly, two questions naturally arise for the case of one expected value constraint ($N = 1$):

1. Are there non-idiosyncratic examples for which moment matching is not possible?

2. If the answer to the previous questions is "yes", then under what conditions is moment matching possible?

Surprisingly, the basic treatments in the information theory literature of maximizing entropy with expected value constraints [12], [13], [34], [77], [78], [101] do not address these questions.

The present monograph originated when the author unsuccessfully searched the literature for one or the other of the following:

(a) A result showing that, if g is unbounded on a countably infinite alphabet A_X and also on every infinite subset thereof,[6] and if there exists a probability mass function p on a countably infinite alphabet A_X such that $E_p[g(X)] = \mu$ (so $H_{max}(\mu)$ is well-defined), then moment matching is possible, i.e., there exists a value λ such that $E_{q_\lambda}[g(X)] = \mu$, in which case $H_{max}(\mu)$ is given by (1.4) with the moment-matching value of λ, or

(b) An example having

   i. A function g that is unbounded on a countably infinite alphabet A_X, as well as on all infinite subsets of A_X,

   ii. A finite value μ,

[6] These assumptions avoid a case in which moment-matching is not possible due to $\Sigma_\lambda = \infty$ for all λ.

iii. A probability mass function p on A_X such that $E_p[g(X)] = \mu$ (so $H_{max}(\mu)$ is well defined), and

iv. No choice of λ such that $E_{q_\lambda}[g(X)] = \mu$ (so $H_{max}(\mu)$ is not determined by moment-matching).

Having no luck searching the literature, the author began his own investigation. After many alternations between attempting to prove a result like (a) and seeking a counterexample like (b), a counterexample was found, namely, a countably infinite alphabet A_X, a real-valued function g that is unbounded on A_X, as well as on every infinite subset of A_X, and a value μ for which there exists a pmf p on A_X such that $E_p[g(X)] = \mu$, but $E_{q_\lambda}[g(X)] < \mu$ for every λ such that $\Sigma_\lambda < \infty$. (Indeed, for this example, $E_p[g(X)]$ can be arbitrarily large, but $E_{q_\lambda}[g(X)]$ is bounded over all choices of λ such $\Sigma_\lambda < \infty$, i.e., such that q_λ is well defined.)

It next became important to find conditions on A_X, g and μ under which moment-matching yields $H_{max}(\mu)$. A couple of such sufficient conditions were found, which are presented in Section 2.

A subsequent re-examination of the literature for multiple expected value constraints found counterexamples (as mentioned earlier) for the case of multiple expected value constraints, did not find counterexamples for the case of one expected value constraint, and did find that, as discussed in Section 5, the sufficient conditions discovered by the author are specializations to one expected value constraint of sufficient conditions in the literature that apply to multiple expected value constraints. However, for someone interested in just the single expected value constraint, the theory in the literature leading to these conditions is, to say the least, not easily accessible.

Accordingly, as mentioned earlier, the first goal of the present monograph is to present the derivation of the maximum entropy with a single expected value constraint, including both the counterexample and sufficient conditions, in an accessible and self-contained fashion. The monograph also aspires to be complete. For example, it presents some properties of a function g on a countably infinite alphabet A_X that cause the maximum entropy function, $H_{max}(\mu)$, to be infinite for all μ.

Of course the degree to which this monograph is accessible and self-contained, and complete is somewhat subjective. Section 1.8 mentions the ways in which it attempts to have these characteristics.

1.3 Maximizing Differential Entropy

Analysis similar to that used to address entropy maximization can be used to address the second stated goal of this monograph, namely, to find the maximum differential entropy, $H_d(X)$, of a continuous, real-valued random variable X with the constraints that the expected value of some real-valued function g applied to X equals a target value μ, and the support of X is contained in a specified constraint set S.

Section 3 will carefully derive the well-known result[7] that, in many cases, differential entropy, $H_d(X)$, is maximized when, and only when, X has a probability density function (pdf) with the exponential form

$$q_\lambda(x) = \begin{cases} \dfrac{2^{-\lambda g(x)}}{\Sigma_\lambda}, & x \in S \\ 0, & x \notin S \end{cases}, \tag{1.8}$$

and the resulting largest differential entropy is

$$H_{dmax}(\mu) = \lambda\mu - \log \Sigma_\lambda, \tag{1.9}$$

where $\Sigma_\lambda = \int_S 2^{-\lambda g(x)}\, dx$ is a normalizer chosen to make q_λ integrate to one, and λ is a real value chosen to make the expected value of $g(X)$ equal μ when q_λ governs the distribution of X. That is, moment-matching applies once again.

As in Section 1.2, this type of analysis is well known to extend to maximizing differential entropy with multiple expected value constraints, $E[g_i(X)] = \mu_i$, $i = 1, \ldots, N$, resulting in expressions for $H_{dmax}(\mu_1, \ldots, \mu_N)$ and the maximizing exponential-form pdf analogous to those given by (1.5)–(1.7). It also extends to scenarios in which X is a

[7] It can be found in information theory textbooks, (see, for example, [34, p. 296 ff.], [12, Chap. 11], [13, Chap. 12], [101, Sec. 10.6]), and in books, articles and websites devoted to the maximum entropy principle and to the many application areas for which the maximum entropy principle is useful, see for example, [8], [26], [29], [35], [47]–[49], [84], [97], [100].

discrete-time, continuous-valued stationary random process $X_1, X_2, \ldots$ for which it is desired to maximize *differential entropy-rate*[8] subject to a constraint that the expected value of some real-valued function g applied to any one of the random variables equals a specified value μ.

Let us focus on the first scenario. Then, as in the maximizing entropy case, it is important to note that this analysis applies only when

(i) There exists a pdf p with support contained in S such that $E_p[g_n(X)] = \mu_n$ for $n = 1, \ldots, N$, and consequently, $H_{dmax}(\mu_1, \ldots, \mu_N)$ is well-defined, and

(ii) There exist $\lambda_1, \ldots, \lambda_N$ that cause the desired moment-matching, which includes the requirement that $\Sigma_{\lambda_1,\ldots,\lambda_N}$ be finite.

In regard to Condition (i), it is straightforwardly shown in the Appendix, that when $N = 1$ (one expected value constraint), there is a pdf p such that $E_p[g(X)] = \mu$ whenever μ lies between the essential infimum and essential supremum of g, and also sometimes when μ is the essential infimum or essential supremum. On the other hand, if $N \geq 2$, then as with maximizing entropy, there are examples for which $H_{dmax}(\mu_1, \ldots, \mu_N)$ is undefined due to there not being a single pdf p such that $E_p[g_i(X)] = \mu_n$, for $i = n, \ldots, N$. In regard to Condition (ii), let us focus on the case of just one expected value constraint, which is the main maximizing-differential-entropy concern of this monograph. Let us also assume for the discussion in this introduction that g is continuous or piecewise continuous, so as to avoid having to deal with certain technicalities.[9]

On the one hand, when S has finite measure[10] and g is bounded, the situation is similar to maximizing entropy when A_X is finite. Specifically, it is straightforward to show (see, for example, Section 3.2) that

[8] The differential entropy-rate of a discrete-time, discrete-valued stationary random process is $H_{d,\infty}(X) = \lim_{n\to\infty} \frac{1}{n} H_d(X_1, \ldots, X_n)$.

[9] No such assumption is made in Section 3.

[10] In this monograph, the measure of an interval is its length, the measure of a countable union of disjoint intervals is the sum of their lengths, and more generally, the measure of an arbitrary subset of the reals is its Lebesgue measure, unless it is an idiosyncratic set for which Lebsgue measure is undefined. Such sets will not arise in this monograph.

for any μ between the essential minimum and maximum values of g, there is a probability mass function p on A_X such that $E_p[g(X)] = \mu$, and so $H_{dmax}(\mu)$ is well defined. Furthermore, for any such μ, it is straightforward to show (see, for example, Section 3.2) there is a value λ such that $E_{q_\lambda}[g(X)] = \mu$, and consequently, $H_{dmax}(\mu)$ is given by the moment-matching expressions (1.8) and (1.9).

On the other hand, when g is unbounded and S has finite or infinite measure, the situation is somewhat similar to maximizing entropy when A_X is countably infinite. For one thing, in these cases Σ_λ can be infinite for some or all values of λ (depending on g). For example, if S has infinite measure, then $\Sigma_\lambda = \infty$ for all λ when g is bounded on S, or on just an infinite-measure subset of S.

Most importantly, no proof or counterexample could be found in the literature to the hypothesis (similar to (a) for H_{max} in Section 1.2) that, if g is unbounded on S, and also on every subset of S with infinite measure,[11] and if there exists a probability density function p with support contained in S such that $E_p[g(X)] = \mu$ (so $H_{dmax}(\mu)$ is well-defined), then there necessarily exists a value λ such that $E_{q_\lambda}[g(X)] = \mu$, in which case $H_{dmax}(\mu)$ is well defined and given by moment matching. However, the author discovered that a pdf version of the pmf creating a counterexample in the maximum entropy scenario, provides a counterexample for the maximum differential entropy scenario, and also that there are sufficient conditions for moment matching to succeed that are similar to those for a countably infinite alphabet. As in the maximizing-entropy case, these conditions turned out to be specializations to one expected value constraint of the sufficient conditions in the literature that apply to multiple expected value constraints.

Accordingly, as mentioned earlier, the second goal of the present monograph is to present the derivation of the maximum differential entropy with a single expected value constraint, including both the counterexample and sufficient conditions, in an accessible, self-contained and complete as possible fashion.

[11] These assumptions avoid a case in which moment-matching is not possible due to $\Sigma_\lambda = \infty$ for all λ.

1.4 Properties of One-Parameter Exponential Families of Probability Distributions

As previously mentioned the distributions that maximize entropy and differential entropy with an expected value constraint almost always have an exponential form with the function g in the exponent multiplied by a parameter λ. Varying λ creates a family of similar probability distributions, called a *one-parameter exponential family*. Section 6 introduces one-parameter exponential families in a self-contained manner and develops a number of their properties, most of which are needed (Sections 2 and 3). Aside from being useful in these derivations, exponential families have wide application in many fields of endeavor. This section can serve as useful introduction one-parameter families and a self-contained derivation of many of their properties. While most properties in this section are well known in the literature, the derivations given here are independent of those in the literature, and may or may not match them.

1.5 Applications of Maximizing Entropy with an Expected Value Constraint

This subsection describes a number of application areas in which maximizing entropy with an expected value constraint arises. Many of them are for a situation more general than that considered in this monograph. For example, some involve maximizing entropy with multiple expected value constraints, and some involve maximizing entropy-rate of a stationary random process with an expected value constraint.

The first application areas come from information theory and communications. The next described application area, is statistical physics, aka statistical mechanics. Indeed, it was the first area to use maximizing entropy with an expected value constraint. In turn, this usage motivated its usage in many fields of science and engineering as a method for choosing an appropriate probability distribution, as outlined in the final group of application areas. We make no claim that the areas described below include all, or even most, fields in which maximizing entropy with an expected value constraint has been applied.

1.5.1 Capacity of Discrete Noiseless Channels with Average Cost Constraints

The capacity of a communication channel is the maximum rate in bits per second (for a continuous-time channel), or bits per channel use (for a discrete-time channel), at which information bits can be reliably transmitted across the channel with encoding before transmission and decoding after transmission. Shannon theory, which originated with Shannon's seminal 1948 work [93], shows that, under ordinary assumptions about the channel (e.g., stationary and memoryless), the capacity equals the largest possible mutual information between the channel input and output when the input is random. Some channels, namely *discrete noiseless channels*, can be modeled as having their output equal their input, provided the input sequence satisfies a certain constraint. For such a channel, the capacity reduces to the largest possible entropy per unit time of any random channel input satisfying the constraint or, sometimes, an expected value version of the constraint.

Shannon's Discrete Noiseless Channels with an Average Cost Constraint

In his original work, Shannon [93] gave the first examples of discrete noiseless channels, including a model for telegraphy as a continuous-time, discrete-alphabet channel in which (a) the possible inputs are *dot*, *dash*, *letter space* and *word space*, (b) each input has a specified duration in seconds, and (c) the constraint on channel inputs is that two consecutive spaces are not permitted. He modeled this constraint with a graph having edges labeled with input symbols and their corresponding durations. This was done in such a way that the sequences obtainable by reading the symbol labels on walks through the graph are precisely the allowable input sequences.

For continuous-time or discrete-time channels that can be modeled by such graph labelings, Shannon found that capacity, i.e., the largest possible ratio of input symbol entropy to average symbol duration is obtained by an assignment of exponential conditional probabilities to the edges of the graph, with the exponent for the conditional probability assigned to an edge being proportional to the duration of the symbol

produced by that edge. The resulting *sequence-of-edges random process* is stationary and Markov, and entropy is, actually, *entropy-rate*, $H_\infty(X) = \lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n)$. The capacity-attaining channel input process is that which results from reading the symbols on the edges produced by the Markov process. This result also applies to scenarios where symbol costs other than symbol durations label the edges.

In 1983, McEliece and Rodemich [77] found the solution to a closely related problem, namely, that of finding the largest possible entropy-rate, $H_{max}(C)$, of any stationary random process $\{X_n\}$ that (i) is *consistent* with the symbols labeling edges of a given finite graph, and (ii) causes the expected value, $E[c(X_n)]$, of a given edge cost function c to have value C, or less. Here, *consistent* means that (a) $\{X_n\}$ results from reading the symbol labels of the edges produced by a stationary, sequence-of-edges random process $\{Z_n\}$, and (b) $\{Z_n\}$ *respects the graph* in the sense that the only possible values for Z_{n+1} are edges beginning at the graph vertex at which Z_n terminates.

Clearly, the McEliece-Rodemich problem is a maximum-entropy-with-expected-value-constraint problem, albeit in a setting in which it is the entropy-rate of a stationary random process that is being maximized, rather than the entropy of a single random variable, as in the focus of the present monograph. Their paper shows that the entropy-maximizing sequence-of-edges random process is Markov, which Shannon stated, but did not explicitly show, and that its conditional probabilities for edges stemming from each vertex of the graph are exponential, with the exponent being proportional to the cost of the symbol produced by that edge. From this, one sees that the exponential nature of the entropy maximizing distribution carries over from the one-random-variable case considered in the present monograph to this more general setting.

We also note that the result of [77] was extended somewhat in [53], and that maximizing entropy-rate of a finite-valued, stationary Markov process with average cost constraints was also the subject of [46].

Finally, note that the solution to Shannon's telegraphy scenario can be derived from that for the McEliece-Rodemich scenario in that the largest possible ratio of entropy to average symbol duration/cost is $\frac{H_{max}(C^*)}{C^*}$, where C^* is chosen so that in a plot of $H_{max}(C)$ vs. C, a

straight line from the origin to the point $(C^*, H_{max}(C^*)$ is tangent to the plot at C^*, i.e., $\frac{H_{max}(C^*)}{C^*} = \frac{d}{dc}H_{max}(c)\big|_{c=C^*}$.

Storage Channels with an Average Cost Constraint

The process of writing and reading a magnetic storage disk is another scenario[12] that can be modeled as a discrete noiseless channel with an input constraint, see for example, [92]. It is a binary channel with 0's and 1's as the inputs and outputs. For example, the "(2,7) constraint", requires at least two 0's after every stored 1, and no more than 7. Because some binary sequences are not allowed for storage, an encoded binary sequence satisfying the constraints will be longer than the binary data sequence it encodes. The obvious goal is to encode so as to maximize the rate, i.e., the average number of data bits per stored bit. The maximum rate of any such constrained code is the capacity of the channel model, and as before, capacity reduces to the maximum entropy-rate of any binary, stationary random input process whose successive random variables satisfy the input constraints with probability one.

Ordinarily, the constraints associated with a channel model are represented with a finite graph, each edge of which is labeled with a 0 or 1 in such a way that walks through the graph produce the allowable stored sequences as the sequence of edge labels, and no others. For such a model, capacity equals the maximum entropy-rate of any binary, stationary sequence-of-edges random process consistent with the graph.

Encoding with an *average cost constraint* is an alternative to the encoding with a *hard constraint* described above [38], [54], [55], [58], [66], [69]. In this case, the channel is modeled with a graph on which each edge is labeled with both a channel input symbol (0 or 1) and a cost. Rather than strictly restrict the channel input sequences, a constraint is placed on the average cost per edge of the input sequence. For example, the graph describing some particular hard constraint, such as the (2,7) constraint, could be augmented by adding additional edges, assigning 0 as the cost to each edge allowed by the hard constraint, and assigning nonzero costs to the augmented edges. Capacity becomes a function of a target average cost C. As before, information theory shows that

[12] Writing and reading compact disks presents a similar scenario.

capacity equals the maximum entropy-rate of any binary, stationary random process with average cost at most C that is consistent with the symbols labeling the edges of the graph. Thus, finding capacity is again a maximum-entropy-with-expected-value-constraint problerm. Moreover, as shown by the result in [77], the maximizing edge random process is Markov, in the sense that the edge at time $n+1$ will be any of the potential successors to the edge at time n, with a probability that is independent of what edges occurred prior to n, and the edge transition probabilities are exponential, with exponents proportional to edge costs.

NAND flash memory is another storage medium for which maximizing entropy subject to an average cost constraint is relevant, see for example, [39], [67], [68], [70]. It is a multilevel medium. For example, each memory cell might store one of four possible values. The issue is that memory cells wear out after a certain number of writings and erasings (on the order of 10^5 or 10^6 such events), and that writing and erasing certain cell values cause more wear than others. Accordingly, a cost can be assigned to each potential cell value, and it makes sense to encode data to be stored into a sequence of cell values with a constraint on the average cost, and thereby using lower costs cell values more frequently than higher cost cell values. With such a constraint, the maximum encoding rate (in data bits per cell) is the maximum possible entropy of any random variable whose alphabet is the set of allowed cell values subject to the constraint that expected cost of cell values is limited to some preset value.

1.5.2 Shaping Codes

The coding for NAND flash memory just described is an example of what is sometimes called a *shaping code*, in that it shapes a distribution of values. So, too, can the coding for discrete-noiseless-channels-with-average-cost-constraints be considered to be a shaping code. In the communications and information theory literature, the term "shaping" can mean several different things.

Shaping Codes for an Additive Gaussian Channel

The term "shaping" appears first to have been used in the context of designing signal constellations[13] to be used when reliably transmitting data[14] over a communication channel that adds Gaussian noise to the transmitted signal, i.e., over an *additive Gaussian channel* (AGC), [24]. Such constellations are often created from some large or infinite set of N-dimensional signals ($N = 2$ is common) having a regular structure, such as a lattice, by including all members of the set that lie within some specified N-dimensional bounding region. For example, an N-dimensional cube centered at the origin is the plain-vanilla choice. With high data rate, low error probability and low consumed power as the goals of such a system, it was observed in [24] that choosing the bounding region to be an N-dimensional sphere, rather than a cube, induces a reduction in the power required to attain a specified data rate and error probability. This was referred to as a "shaping gain".

From a higher level viewpoint, it was also recognized in [24] that one should consider the sequence of signal constellation component values produced by a transmitter as having a probability distribution that depends on the shape of the bounding region. Information theory indicates that the ideal form of this distribution is Gaussian, and it was recognized in [24] that a circular bounding region creates a distribution more similar to Gaussian than a cubic bounding region, and that this can be considered a source of the aforementioned shaping gain. Relaxing the requirement that the signal constellation consists of all vectors in a given bounding region, but encoding data so as to attain a Gaussian-like distribution of component values is an additional form of shaping. It, too, induces a gain.

[13] A *signal constellation* is a finite set of, say, M signals, each a vector, or equivalently, sequence, in an N-dimensional Euclidean space (often $N = 2$). The constellation is used to transmit binary data across a channel that adds noise, typically Gaussian noise. One possibility is that each signal in the constellation conveys one particular block of $\log_2 M$ data bits. Another, which involves *coding*, uses a codebook of m codewords, each a sequence of n signals in the constellation. In this case, each block of $\log_2 m$ data bits is encoded for transmission into one of the codewords.

[14] Data is assumed to be binary, with successive bits being independent and equally likely to be 0 or 1.

Another higher level viewpoint in [24] was that encoding data so as to use lower energy signals more frequently than higher energy signals is another way to attain a shaping gain.[15] It is a shaping of the probability distribution on the signals in the constellation. (It also shapes the signal component distribution.)

Since [24], much work has focused on developing, optimizing and analyzing codes that do shaping for the additive Gaussian channel in one or more of the senses described previously, see for example, [9], [10], [19]–[22], [24], [25], [51], [52], [59], [60], [63], [71], [79], [101]. Some of these works have used maximizing entropy subject to an expected value (namely, power) constraint as a guiding criteria, or as a vehicle for obtaining performance bounds, see for example, [21], [25], [59].

Shaping Codes for DNA Synthesis

Storing data in man-made DNA sequences is a promising storage methodology of the future, see for example, [11], [30]. In the approach considered in [65], data is encoded into DNA sequences in a sequence of steps. With the goal of maximizing the amount of data that can be stored with a given number of steps, i.e., with a given synthesis time, [65] uses a maximum-entropy-with-expected-value-constraint scenario to analyze the storage capacity of a DNA scheme with a constraint on the average number of synthesis steps.

1.5.3 Gilbert-Varshamov Bounds for Constrained Noisy Channels

The Gilbert-Varshamov bound [28], [98] is a famous lower bound to the largest possible rate of a channel code for a discrete-alphabet, noisy channel, as a function of the minimum Hamming distance between its codewords. It has been extended to discrete noisy channels with input constraints characterized by a graph, such as those mentioned earlier [33], [50], [57], [73]. As shown in [73], [74, pp. 243 ff.] the extended Gilbert-Varshamov bound is a function of an $H_{max}(\mu)$ function, with entropy being the entropy-rate of a stationary random process on the

[15] This is also called *nonuniform signaling*, [59].

graph, and the expected value being that of the Hamming distance between graph edge labels.

1.5.4 Symbolic Dynamics

Among the many flavors of dynamical systems, maximizing entropy with an expected value constraint has been useful in the ergodic theory version of dynamical systems, sometimes called symbolic dynamics, which has close ties to information theory. For example, [42] addresses the problem of finding maximum long-term time average of a function g among sample sequences produced by a shift-invariant ergodic process. Specifically, it shows that, in the context of the *low temperature limit*, the solution is characterized as a maximum entropy-rate with an expected value constraint.

1.5.5 Statistical Physics, Statistical Mechanics

As nicely described in the review article by Pressé *et al.* [84], maximizing entropy with expected value constraints has a long history in statistical physics, aka statistical mechanics. It was first used by Gibbs [27] to justify the exponential distribution of gas molecule energies, originally posited by Maxwell [75] and Boltzmann [6].

Specifically, as described in [84], Gibbs made the following argument. Suppose (i) there are N gas particles (atoms or molecules) with varying energies, (ii) the range of possible energies is partitioned into s small cells. (iii) n_i denotes the number particles in cell i, each of which has energy ε_i, (iv) the average energy of all particles is known to be $\overline{\varepsilon}$, and (v) we wish to find the form of the occupation probabilities $p_i = \frac{n_i}{N}$, $i = 1, \ldots, s$. Gibbs asserted that, at equilibrium, the probability distribution $p_1, \ldots, p_s$ is that which maximizes entropy[16] $S \triangleq k_B \sum_i p_i \log p_i$ subject to the constraint that $\sum_i p_i \varepsilon_i = \overline{\varepsilon}$, where k_B is Boltzmann's constant. He then showed the resulting distribution has the exponential form $p_i = \frac{e^{-\beta \varepsilon_i}}{\sum_j e^{-\beta \varepsilon_j}}$, where β is chosen to make the average energy of this distribution equal $\overline{\varepsilon}$.

[16] This is the entropy of physics, not information theory.

1.5.6 Jaynes' Maximum Entropy Method in Science and Engineering

As also described in [84], inspired by Gibbs use of maximum entropy, Jaynes [40] proposed maximizing entropy with constraints, including expected value constraints, as a widely applicable technique for estimating a probability distribution in the presence of limited data. This has been widely used in many areas of science and engineering.

Specifically, in many science and engineering fields, there is a need to estimate the probability distribution of some random variable based on rather limited amounts of information, such as its mean value, its variance, the expected values of one or more functions of the variable, the alphabet/support of the random variable, or some combination of these. For such problems, Jaynes argued that best choice of distribution is that which has largest entropy (in the discrete case), or largest differential entropy (in the continuous case), and also matches the given information. The idea is that an entropy maximizing distribution is considered to be commensurate with the fewest assumptions about the variable, and is, arguably, the fairest choice of distribution. In statistics, this is called the *principle of maximum entropy* or the *maximum entropy method or approach*.

Many books have been written on the maximum entropy method and its applications, see for example, [4], [8], [15], [26], [29], [35], [36], [47]–[49], [96], [97]. For example, [47] describes its use in regional and urban planning, marketing, elections, economics, finance, insurance, accounting, spectral analysis, image reconstruction, pattern recognition, operations research, biology, medicine and agriculture, [8] describes its use in MRI, spectroscopy, plasma physics and X-ray crystallography, [36] describes its use in ecology, and [15] describes its use in biology.

As also discussed in [84], the maximum entropy method has been criticized that its reliance on maximizing entropy or differential entropy is not adequately justified. For example, why not maximize some other function? Such criticism has been ameliorated in the work of Shore and Johnson [94], [95], who show that "maximizing any function but entropy will lead to inconsistencies unless that function and entropy have identical maxima".

A few specific engineering applications of the maximum entropy method are mentioned below.

1.5.7 Natural Language Processing

The maximum entropy method has been widely used in the design of systems that perform natural language processing tasks, see for example, [3], [17], [45], [64], [81], [85]–[87], [89]. Such tasks include part-of-speech tagging (classifying words as nouns, adjectives, verbs, etc.), parsing sentences into phrases, sentiment analysis (e.g., identifying a segment of text as expressing a positive or negative sentiment), text classification (e.g., determining if a segment of text is spam), sentence boundary classification (e.g., determining if a period indicates an abbreviation or the end of a sentence), speech recognition, and machine translation of text from one language to another.

In most of these applications, it is desired to find a model for the conditional probability distribution $p(y|x)$ for an entity Y taking values in a finite set A_Y, given values of another entity X, taking values in a finite set A_X. The conditional probability model is required to match empirical training data by making the expected values of some appropriate set of *feature functions*, $f_1, \ldots, f_N$, match the empirical means of these functions on the training data. Specifically, there is a training set $(x_1, y_1), \ldots, (x_M, y_M)$, and it is required that the model conditional distribution, $p(y|x)$, satisfies the following;

$$\sum_{x \in A_X} \tilde{p}(x) \sum_{y \in A_Y} p(y|x) f_n(x,y) = \frac{1}{M} \sum_{m=1}^{M} f_n(x_m, y_m), \quad n = 1, \ldots, N, \tag{1.10}$$

where $\tilde{p}(x)$ is the frequency with which the value x occurs in $x_1, \ldots, x_M$. Among the possible conditional probability distributions that satisfy this constraint, it is argued that one should choose that which maximizes the following conditional entropy:

$$- \sum_{x \in A_X} \tilde{p}(x) \sum_{y \in A_Y} p(y|x) \log p(y|x). \tag{1.11}$$

Thus, the problem becomes a maximum entropy with expected value constraint, although it is a conditional distribution that is sought, and a conditional entropy that is to be maximized.

Straightforward analysis shows that the resulting distribution has, once again, an exponential form:

$$p(y|x) = \frac{1}{\Sigma_{\underline{\lambda}}} \exp\left\{ -\sum_{n=1}^{N} \lambda_n f_n(x,y) \right\}, \tag{1.12}$$

where $\underline{\lambda} = (\lambda_1, \ldots, \lambda_N)$, $\Sigma_{\underline{\lambda}} = \sum_{y \in A_Y} \exp\{ -\sum_{n=1}^{N} \lambda_n f_n(x,y)\}$, and $\lambda_1, \ldots, \lambda_N$ are chosen so the constraint (1.10) holds.

Often the feature functions are chosen to be binary, i.e., indicator functions of various events involving X and Y. And often the chosen set $f_1, \ldots, f_M$ is chosen by successively adding the feature function that most improves performance from a very large set of potential feature functions. Addition continues until some stopping criterion is met, such as a plateauing of performance.

For example, [3] provides a nice introduction to the maximum entropy method in the context of natural language processing, and in particular to its use in designing several components of a system for translating French text to English text.

1.5.8 Computer Vision

The maximum entropy method has also been used in computer vision to find conditional probability distributions, for example, for use in making decisions about a scene. The methods are much the same as described in Section 1.5.7 for natural language processing. For example, [31], [32] describe a maximum-entropy-based system whose input is a segment of the audio and video of a baseball game, and whose output is the set of probabilities for the following mutually exclusive highlight events to have occurred in this segment: home run, outfield hit, outfield out, infield hit, infield out, strikeout, walk, none-of-the above. The probabilities are conditioned on features computed from the segment.

1.5.9 Control Theory

Maximum entropy with an expected value constraint has been used in control theory as an optimization strategy, as for example in [5], [56]. Note the latter actually focuses on minimizing divergence, which as discussed in the next subsection produces the same sort of exponential-form distributions.

1.6 Minimizing Divergence Subject to an Expected Value Constraint: A Closely Related Task

Closely related to the task of maximizing entropy or differential entropy with an expected value constraint is the task of minimizing the *divergence* $D(p\|q)$ of a probability distribution p with respect to a reference distribution q.

Divergence is an information-theoretic measure of the similarity of two probability distribution introduced by Kullback and Leibler [62]. It also goes by a number of other names, such as relative entropy, cross-entropy,[17] information divergence, directed divergence and Kullback-Leibler distance.

In the discrete case, the divergence of a pmf p with respect to pmf q is $D(p\|q) \triangleq \sum_x p(x) \log \frac{p(x)}{q(x)}$. In the continuous case, the divergence $D(p\|q)$ of a pdf p with respect to pdf q is defined by the same formula but with the sum replaced by an integral.

In a number of application areas, it is desired to find a distribution p such that the expected value $E_p[g(X)]$ of some function g matches a known measurement value μ, and in addition, p is close to some reference measure q. For example, q might be an initial estimate of p for which the expected value is not μ, or it might be an empirical distribution for which a closed form model is sought. In such applications, it is often argued that the appropriate choice of p is that which has $E_p[g(X)] = \mu$ and also minimizes $D(p\|q)$.

This minimum divergence approach was introduced by Kullback [61], not long after Jaynes introduced the maximum entropy method,

[17] In this monograph, *cross-entropy* refers to a different quantity, which is introduced in Section 2.

independently of Jaynes approach. Like Jayne's approach, the minimum divergence approach has received much attention and has been used in many applications. For example, it has appeared in the following information theory books: [12, Prob. 11.2], [13, Prob. 12.2], and [101, Prob. 10.8]), and in books, articles and websites devoted to the maximum entropy principle and to the many application areas for which the maximum entropy principle is useful, see for example, [47, Chap. 7], [48, Sections IV, V], [29, Sec. 3.3], [100, Sec. 2.3.2], [49, Sections 1,5] [36, p. 126], and [84].

Note that if the reference distribution q is uniform, i.e., constant, on a support set S with finite size (discrete case) or finite measure (continuous case), then for any p, $D(p\|q) = -\widetilde{H}_p(X) + \log|S|$, where $\widetilde{H}_p(X)$ is a proxy for the entropy of p (discrete case) or the differential entropy of p (continuous case), and where $|S|$ denotes the number of elements of S (discrete case), or the length/measure of S (continuous case). As a result, when q is uniform minimizing divergence $D(p\|q)$ is equivalent to maximizing entropy $H_p(X)$ (discrete case) or differential entropy $H_{d,p}(X)$ (continuous case).

Accordingly, for the case of finite supports, the minimum divergence approach includes the maximum entropy approach.

1.7 Rationale for Focusing on Maximizing Entropy of One Random Variable with One Expected Value Constraint

Given that there is a theory of maximizing entropy and differential entropy when there is a random process rather than a random variable, and also when there are multiple expected value constraints, one may wonder why it is worthwhile to have a monograph focused on maximizing entropy for just one random variable with just one expected value constraint. The following suggests some rationales.

1. Maximizing entropy with one expected-value constraint is sufficiently important that it deserves to be fleshed out on its own.

2. No previous discussion of maximizing entropy with one expected-value constraint considers the conditions under which moment matching yields the solution.

3. The theory for one expected value constraint can be exposited with considerably less sophisticated methods than the theory for multiple expected value constraints.

4. It is possible to say things about the one-expected-value constraint that are not easily seen from the theory for the more general case of multiple expected value constraints. For example, in most instances of this case, the set of λ's for which Σ_λ is finite has a simple form, namely, it is an infinite interval and often a one-sided infinite interval. As another example, the monotonicities of $E_{q_\lambda}[g(X)]$ and $H_{q_\lambda}(X)$ with changing λ have simple forms, as do the monotonicities of $H_{max}(\mu)$ and $H_{dmax}(\mu)$ with changing μ. In contrast, when there are multiple expected value constraints, it can be difficult to characterize the set of λ's for which Σ_λ is finite and also the monotonicities of $E_{q_\lambda}[g(X)]$ with changing λ.[18]

5. It is possible to spell out function characteristics that cause $H_{max}(\mu)$ and $H_{dmax}(\mu)$ to be infinite for all μ in their respective ranges, such as when the support of X has infinite size and the function g is bounded.

6. As mentioned earlier, it is not well known that there are instances that a conventional moment-matching, exponential-form probability distribution will not maximize entropy or differential entropy of a single random variable, subject to a single expected value constraint. That is, there are functions g and values μ for which there is a probability distribution p with $E[g(X)] = \mu$, but for no value of λ does the exponential-form distribution q_λ in (1.3) have $E_{q_\lambda}[g(X)] = \mu$ (or even $E_{q_\lambda}[g(X)] \approx \mu$). Hence, moment-matching is not possible. Furthermore, it is possible to give simple examples of functions for which moment matching is not possible that differ only slightly from functions for which moment matching is possible.

[18] As discussed in Section 5.2, λ will be a vector in the multiple-expected-value-constraint case.

7. It can be beneficial to see the maximum entropy/differential entropy derived for the basic case of one random variable and one expected value constraint, as well to become familiar with simple, one-parameter exponential families, before moving on to the more general theory, as for example in [7], [99]. That is, it can provide a warm-up, not only to the general theory of maximizing entropy/differential entropy subject to expected value constraints, but also to exponential families of probability distributions and to information geometry.

1.8 Accessibility, Self-Containment and Completeness

As mentioned earlier, the goals of this monograph include presenting the subject in an accessible, self-contained and complete manner. One of the ways in which it attempts to do this is that it defines terms with which some readers might not be familiar, or for which there might be multiple interpretations. A second is that derivations are fairly complete and leave relatively few steps to the reader. A third is that this monograph is written with the idea that some readers might be interested primarily in maximizing entropy and some primarily in maximizing differential entropy. For this reason, Section 3, which deals with maximizing differential entropy, can be read independently of Section 2, which deals with maximizing entropy. However, since many readers will have read Section 2 before Section 3, the latter frequently comments on how items in the latter relate to items in the former.

As one has probably gathered from the discussion in Sections 1.2 and 1.3, the theory of maximizing entropy or differential entropy becomes more involved when it comes to considering the feasibility of moment-matching in the cases of discrete random variables with countably infinite alphabets and continuous random variables. For example, as mentioned earlier, there is the issue that in such cases Σ_λ may be infinite for some or all values of λ, in which event, for such λ's, q_λ, $E_{q_\lambda}[g(X)]$, $H_{q_\lambda}(X)$ and $H_{d,q_\lambda}(X)$ are all undefined. There is also the issue that moment-matching might not be possible.

Accordingly, as the fourth approach to furthering accessibility, the discussions in Section 2.5 for discrete random variables with a countably

infinite alphabet begins by ignoring the possibility that Σ_λ might be infinite, and derive the basic moment-matching result, presuming that moment-matching is possible. Only subsequently, in Section 2.6, do these discussions consider the finiteness of Σ_λ and the feasibility of moment-matching. The intention is that a reader with a specific application in mind (alphabet A_X, function g and target value μ) will in many cases be able to find a value λ with $\Sigma_\lambda < \infty$ that induces moment-matching. In such cases, s/he can skip Section 2.6 and proceed to the remainder of Section 2.

As will be seen, going farther in Section 2.6 requires dealing with exponential-form probability mass functions and the infinite sums defining quantities such as the normalizer Σ_λ, the expected value $E_{q_\lambda}[g(X)]$ and the entropy $H_{q_\lambda}(X)$. For example, there are issues regarding the existence and finiteness of $E_{q_\lambda}[g(X)]$ and $H_{q_\lambda}(X)$, and their limits when λ approaches certain values. While the needed results are quite intuitive and believable, a number of their proofs involve considerable technical detail, such as references to series convergence theorems. Accordingly, as the fifth effort to enhance accessibility, many such details are postponed to Section 6. All of these are properties of exponential-form probability distributions. For example, it is shown there that as λ increases, q_λ increasingly concentrates on values of x for which $g(x)$ is smaller[19] with the result that $E_{q_\lambda}[g(X)]$ decreases. Moreover, the fact that q_λ becomes more concentrated usually causes entropy $H_{q_\lambda}(X)$ to decrease as λ increases.

Similarly, when deriving the maximum possible differential entropy with an expected value constraint, the discussion begins in Section 3.1 by ignoring the possibility that Σ_λ might be infinite and by presuming moment matching is possible. Then subsequently, in Sections 3.2 and 3.3, the discussion deals with these issues. Again, the intention is that a reader with a specific application in mind (CRVSC set S, function g, target value μ) will in many cases be able to find, for this specific application, a value λ with $\Sigma_\lambda < \infty$ that induces moment matching. In such cases, s/he can skip Sections 3.2 and 3.3.

[19] While this increasing concentration is well known, Lemma 6.1 of Section 6 provides an explicit formula quantifying this concentration.

Also, like the discrete case, there are again issues regarding the existence and finiteness of $E_{q_\lambda}[g(X)]$ and $H_{d,q_\lambda}(X)$, and their limits when λ approaches a limit. While again the needed results are almost all quite intuitive and believable, their proofs involve technical detail of the same order as for discrete random variables with countably infinite alphabets. Accordingly, these are again postponed to Section 6.

We note that a number of the derivations in Section 6 are so similar for the discrete and continuous cases that they can proceed simultaneously. To facilitate this, some specialized, nonstandard terminology, notation and conventions are introduced in Section 4. For example, a *generalized sum* is introduced that is interpreted as an ordinary sum in the discrete case and an integral in the continuous case. This section appears just before Section 5, which discusses the generalization of maximizing entropy and differential entropy subject to multiple expected value constraints, precisely so the discussion in Section 5 can proceed simultaneously for the discrete and continuous cases.

Though much of the discussion in Sections 5 and 6 applies simultaneously to both cases, the notation and conventions adopted have the characteristic that if one is interested only in the discrete case, one should not be distracted by the fact that the discussion also applies to the continuous case, and vice versa. For example, someone interested in the discrete case will simply view a generalized sum as an ordinary sum, and someone interested in the continuous case will view it as an integral.

Because of the nonstandard nature of these notations, conventions and terms, even fully versed probabilists and information theorists should read Section 4.

While one viewpoint is that Section 6 contains technical details that support the developments of Sections 2 and 3, another is that it can be viewed as a self-contained tutorial on one-parameter families of exponential-form pmfs induced by a function g and a range of λ's.

Sections 2 and 3 also present properties of the maximum entropy/differential entropy, $H_{max}(\mu)$ and $H_{dmax}(\mu)$ as functions of μ, such as their convexity and the fact that, for some functions, they can be infinite

for all μ.[20] Again, to facilitate accessibility some of the more arcane details are postponed to Section 7.

As a final introductory comment, we note that the Appendix derives the range of possible values of the expected value, $E_p[g(X)]$, of a function g applied to a random variable X, as its probability distribution p ranges over all possibilities. As discussed in Sections 2 and 3, this range determines the domain of the maximum entropy and differential entropy functions $H_{max}(\mu)$ and $H_{dmax}(\mu)$. It is placed at the end of the monograph rather than in Sections 2 and 3 because it proceeds simultaneously for the discrete and continuous cases.

[20] The inclusion of results delimiting when these functions are infinite for all μ is one of the ways in which this monograph attempts to be complete.

References

[1] S.-I. Amari and H. Nagaoka, "Methods of information geometry," *Oxford: Am. Math. Soc.*, vol. 191, 2000.

[2] O. Barndorff-Nielsen, *Information and Exponential Families*. New York: John Wiley & Sons, 1978.

[3] A. Berger, S. A. Della Pietra, and V. J. Della Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[4] R. M. Bevensee, *Maximum Entropy Solutions to Scientific Problems*. Englewood Cliffs: Prentice Hall, 1993.

[5] J. Bierkens and H. J. Kappen, "Explicit solution of relative entropy weighted control," *Syst. & Control Lett.*, vol. 72, pp. 36–43, 2014.

[6] L. Boltzmann, "Über das Wärmegleichgewicht zwischen mehratomigen Gasmolekülen (On the thermal equilibrium between polyatomic gas molecules)," *Wiener Berichte*, vol. 63, pp. 397–418, 1871.

[7] L. D. Brown, *Fundamentals of Statistical Exponential Families*. Hayward, CA: Inst. of Math. Stat., 1986.

[8] B. Buck and V. A. Macaulay, *Maximum Entropy in Action: A Collection of Expository Essays*. New York, NY: Clarendon Press, 1991.

[9] A. R. Calderbank and M. Klimesh, "Balanced codes and non-equiprobable signaling," *IEEE Tr. Inform. Thy.*, vol. 38, no. 3, pp. 1119–1122, 1992.

[10] A. R. Calderbank and L. H. Ozarow, "Nonequiprobable signaling on the gaussian channel," *IEEE Tr. Inform. Thy.*, vol. 36, no. 4l, pp. 726–740, 1990.

[11] G. M. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in DNA," *Science*, vol. 337, no. 6102, p. 1628, 2012.

[12] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ: John Wiley, 1991.

[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: John Wiley, 2006.

[14] I. Csiszár and P. C. Shields, "Information theory and statistics: A tutorial," *Foundations and Trends in Comm. and Inform. Thy.*, vol. 1, no. 4, pp. 417–528, 2004.

[15] A. De Martino and D. De Martino, "An introduction to the maximum entropy approach and its application to inference problems in biology," *Heliyon*, vol. 4, 2018.

[16] J. Del Castillo, "The singly truncated normal distribution: A none-steep exponential family," *Ann. Inst. Statist. Math.*, vol. 46, no. 1, pp. 58–66, 1994.

[17] S. A. Della Pietra, V. J. Della Pietra, and J. D. Lafferty, "Inducing features of random fields," *IEEE Tr. Patt. Anal. and Mach. Intel.*, vol. 19, no. 4, pp. 380–393, 1997.

[18] B. Efron, "The geometry of exponential families," *Ann. Statist.*, vol. 6, no. 2, pp. 362–376, 1978.

[19] M. V. Eyuboglu and G. D. Forney, "Trellis precoding: Combined coding, precoding and shaping for intersymbol interference channels," *IEEE Tr. Inform. Thy.*, vol. 38, no. 2, pp. 301–314, 1992.

[20] F. H. Fischer, *Precoding and Signal Shaping for Digital Transmission*. New York, NY: John Wiley and Sons, 2002.

[21] G. D. Forney Jr., "Multidimensional constellations—part II: Voronoi constellations," *IEEE J. Select. Areas Commun*, vol. 7, pp. 941–958, 1989.

[22] G. D. Forney Jr., "Trellis shaping," *IEEE Tr. Inform. Theory*, vol. 38, no. 1, pp. 281–300, 1992.

[23] G. D. Forney, *Information Theory. Unpublished course notes*. Stanford University, 1971.

[24] G. D. Forney, R. Gallager, G. Lang, F. Longstaff, and S. Qureshi, "Efficient modulation for band-limited channels," *IEEE J. Selected Areas in Comm.*, vol. 2, no. 5, pp. 632–647, 1984.

[25] G. D. Forney and L.-F. Wei, "Multidimensional constellations—part I: Introduction, figures of merit, and generalized cross constellations," *IEEE J. Selected Areas Commun.*, vol. 7, no. 6, pp. 877–878, 1989.

[26] H. Fort, *Forecasting with Maximum Entropy: The Interface Between Physics, Biology, Economics and Information Theory*. Bristol, UK: IOP Publishing, 2022.

[27] J. W. Gibbs, "On the equilibrium of heterogeneous substances," *Trans. Connecticut Acad. of Arts and Sci.*, vol. 3, pp. 108–248 (Part I), pp. 343–524 (Part II), 1875–1878.

[28] E. N. Gilbert, "A comparison of signalling alphabets," *Bell Sys. Tech. J.*, vol. 3, no. 3, pp. 504–522, 1952.

[29] A. Golan, G. Judge, and D. Miller, *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. Chichester UK: John Wiley, 1996.

[30] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. LeProust, B. Sipos, and E. Birney, "Towards practical, high-capacity, low maintenance information storage in synthesized DNA," *Nature*, vol. 494, no. 7435, pp. 77–80, 2013.

[31] Y. Gong, M. Han, W. Hua, and W. Xu, "Maximum entropy model-based baseball highlight detection and classification," *Int. J. Computer Vision and Image Understanding*, vol. 96, pp. 181–199, 2004.

[32] Y. Gong and W. Xu, *Machine Learning in Multimedia Content Analysis*. New York, NY: Springer, 2007.

[33] J. Gu and T. Fuja, "A generalized Gilbert-Varshamov bound derived via analysis of a code-search algorithm," *IEEE Tr. Inform. Thy.*, vol. 39, no. 3, pp. 1089–1093, 1993.

[34] S. Guiasu, *Information Theory with Applications*. New York: McGraw-Hill, 1977.

[35] H. Gzyl, S. Mayoral, and E. Gomes-Goncalves, *Loss Data Analysis: The Maximum Entropy Approach*. Boston, MA, USA: De Gruyter, 2018.

[36] J. Harte, *Maximum Entropy and Ecology: A Theory of Abundance, Distribution, and Energetics*. Oxford: Oxford University Press, 2011.

[37] J. Harte and E. A. Newman, "Maximum information entropy: A foundation for ecological theory," *Trends in Ecology and Evolution*, vol. 29, no. 7, pp. 384–389, 2014.

[38] C. D. Heegard, B. H. Marcus, and P. H. Siegel, "Variable-length state splitting with applications to average runlength-constrained (arc) codes," *IEEE Tr. Inform. Thy.*, vol. 37, no. 3, pp. 759–777, 1991.

[39] A. Jagmohan, M. Franceschini, L. A. Lastras-Montaão, and J. Karidis, "Adaptive endurance coding for NAND flash," in *IEEE Globecom Workshop on Application of Comm. Thy. to Emerging Memory Techniques*, Miami, Fla., USA, 1841–1845, 2010.

[40] E. T. Jaynes, "Information theory and statistical mechanics I," *Phys. Rev.*, vol. 106, pp. 620–630, 1957.

[41] F. Jelinek, *Probabilistic Information Theory*. New York: McGraw-Hill, 1968.

[42] O. Jenkinson, "Ergodic optimization in dynamical systems," *Ergodic Theory and Dynamical Systems*, vol. 39, pp 2593–2618, 2019.

[43] S. Johansen, *Theory of Regular Exponential Families*. Copenhagen: Inst. Math. Stat., 1979.

[44] R. A. Johnson, *Miller and Freund's Probability and Statistics for Engineers*, 6th ed. Upper Saddle River, N.J.: Prentice Hall, 2000.

[45] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition,* Chap. 6. Upper Saddle River, N.J.: Pearson Prentice Hall, 2009.

[46] J. Justesen and T. Hoholdt, "Maxentropic Markov chains," *IEEE Tr. Inform. Thy.*, vol. 30, no. 4, pp. 665–667, 1984.

[47] J. N. Kapur, *Maximum-Entropy Models in Science and Engineering*. New York: John Wiley and Sons, 1989.

[48] J. N. Kapur and H. K. Kesavan, *Entropy Optimization Principles with Applications*. San Diego, CA: Academic Press, 1992.

[49] Karmeshu, *Entropy Measures, Maximum Entropy Principle and Emerging Applications*. Berlin: Springer, 2003.

[50] G. Keshav and H. M. Kiah, "Evaluating the Gilbert-Varshamov bound for constrained systems," in *IEEE Int. Symp. Inform. Thy.* Espoo, Finland, pp. 1348–1353, 2022.

[51] A. K. Khandani and P. Kabal, "Shaping multidimensional signal spaces—part I: Optimum shaping, shell mapping," *IEEE Tr. Inform. Thy.*, vol. 39, no. 6, pp. 1799–1808, 1993.

[52] A. K. Khandani and P. Kabal, "Shaping multidimensional signal space—part II: Shell-addressed constellations," *IEEE Tr. Inform. Thy.*, vol. 39, no. 6, pp. 1809–1819, 1993.

[53] A. Khandekar, R. McEliece, and E. Rodemich, "The discrete noiseless channel revisited," in *Coding Communications and Broadcasting*, P. Farrell, M. D. Darnell, and B. Honary, Eds., Baldock, Hertfordshire, England: Research Studies Press Ltd, 2000.

[54] A. Khayrallah, R. Karabed, and D. L. Neuhoff, The capacity of costly noiseless channels, *IBM Res.*, Rep. FU 6040, 1988.

[55] A. Khayrallah and D. L. Neuhoff, "Coding for channels with cost constraints," *IEEE Tr. Inform. Thy.*, vol. 42, no. 3, pp. 854–867, 1996.

[56] J. Kim and I. Yang, "Maximum entropy optimal control of continuous-time dynamical systems," *IEEE Tr. Auto. Control*, vol. 68, no. 4, pp. 2018–2033, 2023.

[57] V. D. Kolesnik and V. Y. Krachkovsky, "Generating functions and lower bounds on rates for limiting error-correcting codes," *IEEE Tr. Inform Thy.*, vol. 37, no. 11, pp. 778–788, 1991.

[58] V. Y. Krachkovsky, R. Karabed, S. Yang, and B. A. Wilson, "On modulation coding for channels with cost constraints," in *Proc. IEEE Int. Symp. Inform. Thy.* Honolulu, HI, 2014, pp. 421–425.

[59] F. R. Kschischang and S. Pasupathy, "Optimal nonuniform signaling for Gaussian channels," *IEEE Tr. Inform. Thy.*, vol. 39, no. 3, pp. 913–929, 1993.

[60] F. R. Kschischang and S. Pasupathy, "Optimal shaping properties of the truncated polydisc," *IEEE Tr. Inform. Thy.*, vol. 40, no. 3, pp. 892–903, 1994.

[61] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.

[62] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, no. 1, pp. 79–86, 1951.

[63] R. Laroia, N. Farvardin, and S. A. Tretter, "On optimal shaping of multidimen-sional constellations," *IEEE Tr. Inform. Thy.*, vol. 40, no. 4, pp. 1044–1056, 1994.

[64] R. Lau, R. Rosenfeld, and S. Roukos, "Adaptive language modeling using the maximum entropy principle," in *Proc. Human Lang. Tech. Workshop*, Plainsboro, NJ, pp. 1080–1113, 1993.

[65] A. Lenz, Y. Liu, C. Rashtchian, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "Coding for efficient DNA synthesis," in *IEEE Int. Symp. Inform. Thy.* Los Angeles, CA, USA, 2020, pp. 2885–2890.

[66] A. Lenz, S. Melczer, C. Rashtchian, and P. H. Siegel, "Exact asymptotics for discrete noiseless channels," *IEEE Int'l Symp. Inform. Thy.*, pp. 2494–2498, 2023.

[67] Y. Liu, P. Huang, A. W. Bergman, and P. H. Siegel, "Rate-constrained shaping codes for structured sources," *IEEE Tr. Inform. Thy.*, vol. 66, no. 8, pp. 5261–5281, 2020.

[68] Y. Liu, P. Huang, and P. H. Siegel, "Performance of optimal data shaping codes," in *IEEE Int. Symp. Inform. Thy.* Aachen, Germany, 2017, pp. 1003–1007.

[69] Y. Liu, Y. Li, P. Huang, and P. H. Siegel, "Rate-constrained shaping codes for finite-state channels with cost," *IEEE Int. Symp. Inform. Thy.*, pp. 1354–1359, 2022.

[70] Y. L̇iu and P. H. Siegel, "Shaping codes for structured data," in *IEEE Global Commun. Conf. (GLOBECOM)*, Washington, D.C., USA, 2016, pp. 1–6.

[71] J. N. Livingston, "Shaping using variable-size regions," *IEEE Tr. Inform. Thy.*, vol. 38, no. 4, pp. 1347–1353, 1992.

[72] P. Loya, *Amazing and Aesthetic Aspects of Analysis*. New York, NY: Springer, 2017.

[73] B. H. Marcus and R. M. Roth, "Improved Gilbert-Varshamov bound for constrained systems," *IEEE Tr. Inform. Thy.*, vol. 38, no. 4, pp. 1213–1221, 1992.

[74] B. H. Marcus, R. M. Roth, and P. H. Siegel, "An introduction to coding for constrained systems," *Lecture Notes*, 2001.

[75] J. C. Maxwell, "Illustrations of the dynamical theory of gases," *Philosophical Magazine, Series 4*, vol. 19, p. 19, 1860.

[76] R. J. McEliece, *The Theory of Information and Coding*, 2nd ed. Reading, Mass: Addison-Wesley, 1977.

[77] R. J. McEliece and E. R. Rodemich, "A maximum entropy Markov chain," in *Proc. 17th Annual Conf. on Inform. Sci. and Sys.*, pp. 245–248, 1983.

[78] D. J. C. McKay, *Information Theory, Inference and Learning Algorithms*. Cambridge: Cambridge University Press, 2003.

[79] S. W. McLaughlin and A. S. Khayrallah, "Shaping codes constructed from cost constrained graphs," *IEEE Tr. Inf. Thy.*, vol. 43, no. 2, pp. 692–699, 1997.

[80] F. Nielsen, "An elementary introduction to information geometry," *Entropy*, vol. 22, no. 10, p. 1100, 2020. DOI: 10.3390/e22101100.

[81] K. Nigam, J. Lafferty, and A. McCallum, "Using maximum entropy for text classification," *IJCAI-99 Workshop on Mach. Learning for Inform. Filtering*, vol. 1, no. 1, pp. 61–67, 1999.

[82] Y. Polyanskiy and Y. Wu, *Information Theory: From Coding to Learning*. Cambridge: Cambridge University Press, 2024.

[83] E. C. Posner, E. R. Rodemich, and H. Rumsey Jr., "Epsilon entropy of stochastic processes," *Annals of Math. Stat.*, vol. 38, no. 4, pp. 1000–1020, 1967.

[84] S. Pressé, G. Kingshuk, J. Lee, and K. A. Dill, "Principles of maximum entropy and maximum caliber in statistical physics," *Reviews of Modern Physics*, vol. 85, no. 3, pp. 1115–1141, 2013.

[85] A. Ratnaparkhi, "A maximum entropy part of speech tagger," in *Conf. on Empirical Methods in Natural Lang. Proc.* Philadelphia, PA, pp. 133–142, 1996.

[86] A. Ratnaparkhi, "A simple introduction to maximum entropy models for natural language processing," *IRCS Technical Reports Series*, vol. 81, 1997.

[87] A. Ratnaparkhi, "Learning to parse natural language with maximum entropy models," *Machine Learning*, vol. 34, pp. 151–175, 1999.

[88] R. T. Rockafellar, *Convex Analysis*. Princeton: Princeton Univ. Press, 1970.

[89] R. Rosenfeld, "*Adaptive Statistical Languarge Modeling: A Maximum Entropy Approach*," Carnegie Mellon University, 1994.

[90] S. Ross, *A First Course in Probability*, 6th ed. Upper Saddle River, N.J.: Prentice Hall, 2002.

[91] H. L. Royden, *Real Analysis*, 3rd ed. New York, NY: Macmillan, 1988.

[92] K. A. Schouhamer Immink, "Innovation in constrained codes," *IEEE Comm. Mag.*, vol. 60, no. 10, pp. 20–24, 2022.

[93] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.

[94] J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. Inform. Thy.*, vol. 26, no. 1, pp. 26–37, 1980.

[95] J. Shore and R. Johnson, "Properties of cross-entropy minimization," *IEEE Trans. Inform. Thy.*, vol. 27, no. 4, pp. 472–482, 1981.

[96] T. Squartini and D. Garlaschelli, *Maximum-Entropy Networks: Pattern Detection, Network Reconstruction and Graph Combinatorics*. Springer, 2017.

[97] H. Theil and D. G. Fiebig, *Exploiting Continuity: Maximum Entropy Estimation of Continuous Distributions*. Cambridge, MA: Ballinger, 1984.

[98] R. R. Varshamov, "Estimate of the number of signals in error correcting codes," *Dokl. Akad. Nauk SSSR*, vol. 117, pp. 739–741, 1957.

[99] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, vol. 1, no. 1–2, pp. 1–305, 2008.

[100] N. Wu, *The Maximum Entropy Method*. New York, NY: Springer-Verlag, 1997.

[101] R. W. Yeung, *Information Theory and Network Coding*. New York, NY: Springer, 2008.