Strong Entropy Concentration, Game Theory and Algorithmic Randomness

Peter Grünwald[⋆1]

EURANDOM
Postbus 513
5600 MB Eindhoven
The Netherlands pdg@cwi.nl

Abstract. We give a characterization of Maximum Entropy/Minimum Relative Entropy inference by providing two 'strong entropy concentration' theorems. These theorems unify and generalize Jaynes' 'concentration phenomenon' and Van Campenhout and Cover's 'conditional limit theorem'. The theorems characterize exactly in what sense a 'prior' distribution Q conditioned on a given constraint and the distribution $\tilde{P}$ minimizing $D(P||Q)$ over all P satisfying the constraint are 'close' to each other. We show how our theorems are related to 'universal models' for exponential families, thereby establishing a link with Rissanen's MDL/stochastic complexity. We then apply our theorems to establish the relationship (A) between entropy concentration and a game-theoretic characterization of Maximum Entropy Inference due to Topsøe and others; (B) between maximum entropy distributions and sequences that are random (in the sense of Martin-Löf/Kolmogorov) with respect to the given constraint. These two applications have strong implications for the use of Maximum Entropy distributions in *sequential prediction tasks*, both for the logarithmic loss and for general loss functions. We identify circumstances under which Maximum Entropy predictions are almost optimal.

1 Introduction

Jaynes' Maximum Entropy (MaxEnt) Principle is a well-known principle for inductive inference [5, 7, 24, 15, 25, 4, 10, 19]. It has been applied to statistical and machine learning problems ranging from protein modeling so stock market prediction [17]. One of its characterizations (some would say 'justifications') is the so-called *concentration phenomenon* [13, 14]. Here is an informal version of this phenomenon, in Jaynes' words:

⋆ Also: CWI, Kruislaan 413, 1098 SJ Amsterdam. The author would like to thank Richard Gill and Phil Dawid for very useful discussions regarding resp. conditioning on measure 0-events and the concentration phenomenon. Part of this research was performed while the author was at the Computer Science Department of Stanford University, financed by an NWO (Netherlands Organization of Scientific Research) TALENT Grant.

> "If the information incorporated into the maximum-entropy analysis includes all the constraints actually operating in the random experiment, then the distribution predicted by maximum entropy is overwhelmingly the most likely to be observed experimentally." [16, Page 1124]

For the case in which a prior distribution over the domain at hand is available, Van Campenhout and Cover [25, 4] have proven the related *conditional limit theorem*. In Part I of this paper (Sections 2-4), we provide a strong generalization of both the concentration phenomenon and the conditional limit theorem. In Part II (Section 5) we apply this. We first show how our theorems can be used to construct universal models for exponential families, thereby establishing a link with Rissanen's Minimum Description Length Principle. We then extend an existing game-theoretic characterization of Maximum Entropy due to Topsøe [24]. Finally we combine the results of Part I with the theory of algorithmic (Martin-Löf/Kolmogorov) randomness. This allows us to substantiate the often-heard informal claim that 'adopting the Maximum Entropy distribution leads to good predictions if the data are random with respect to the given constraint' and to make precise informal notions like '*all* constraints *actually* operating in an environment' as used in Jaynes' statement above. We end by discussing implications of our results in Part II for (sequential) *prediction*. We identify circumstances in which Maximum Entropy distributions lead to almost optimal predictions. The proofs of Theorem 1 and Theorem 3 are in the paper; proofs of other theorems, as well as extended discussion, can be found in the technical report [12].

2 Informal Overview

Maximum Entropy Let X be a random variable taking values in some set $\mathcal{X}$, which (only for the time being!) we assume to be finite: $\mathcal{X} = \{1, \ldots, m\}$. Let P, Q be distributions for $\mathcal{X}$ with probability mass functions p and q. We define $\mathbf{H}_Q(P)$, *the Q-entropy of P*, as

$$\mathbf{H}_Q(P) = -E_P[\log \frac{p(x)}{q(x)}] = -D(P||Q) \tag{1}$$

In the usual MaxEnt setting[1], we are given a 'prior' distribution Q and a *moment constraint*:

$$E[T(X)] = \tilde{t} \tag{2}$$

where T is some function $T : \mathcal{X} \to \mathbf{R}^k$ for some $k > 0$. We define, if it exists, $\tilde{P}$ to be the unique distribution over $\mathcal{X}$ that maximizes the Q-entropy over all distributions (over $\mathcal{X}$) satisfying (2):

$$\tilde{P} = \arg \max_{\{P: E_P[T(X)] = \tilde{t}\}} \mathbf{H}_Q(P) = \arg \min_{\{P: E_P[T(X)] = \tilde{t}\}} D(P||Q) \tag{3}$$

[1] More general formulations with arbitrary convex constraints exist [5], but here we stick to constraints of form (2).

The MaxEnt Principle then tells us that, in absence of any further knowledge about the 'true' or 'posterior' distribution according to which data are distributed, our best guess for it is $\tilde{P}$. In practical problems we are usually not given a constraint of form (2). Rather we are given an *empirical constraint* of the form

$$\frac{1}{n}\sum_{i=1}^{n} T(X_i) = \tilde{t} \quad \text{which we always abbreviate to '}\overline{T^{(n)}} = \tilde{t}\text{'} \tag{4}$$

The MaxEnt Principle is then usually applied as follows: suppose we are given an empirical constraint of form (4). We then have to make predictions about new data coming from the same source. In absence of knowledge of any 'true' distribution generating this data, we should make our predictions based on the MaxEnt distribution $\tilde{P}$ for the moment constraint (2) corresponding to empirical constraint (4). $\tilde{P}$ is extended to several outcomes by taking the product distribution.

The Concentration Phenomenon and The Conditional Limit Theorem Why should this procedure make any sense? Here is one justification. If $\mathcal{X}$ is finite, and in the absence of any prior knowledge beside the constraint, one usually picks the uniform distribution for Q. In this case, Jaynes' 'concentration phenomenon' applies[2]. It says that for all $\epsilon > 0$,

$$Q^n(\sup_{j\in\mathcal{X}} |\frac{1}{n}\sum_{i=1}^{n} I_j(X_i) - \tilde{P}(X=j)| > \epsilon \mid \overline{T^{(n)}} = \tilde{t}) = O(e^{-cn}) \tag{5}$$

for some constant c depending on ϵ. Here Q^n is the n-fold product distribution of Q, and I is the indicator function: $I_j(x) = 1$ if $x = j$ and 0 otherwise. In words, for the overwhelming majority among the sequences satisfying the constraint, the empirical frequencies are close to the maximum entropy probabilities. It turns out that (5) still holds if Q is non-uniform. For an illustration we refer to Example 1. A closely related result (Theorem 1, [25]) is the Van Campenhout-Cover conditional limit theorem[3], which says that

$$\lim_{\substack{n\to\infty \\ n\tilde{t}\in\mathbf{N}}} Q^1(\cdot \mid \overline{T^{(n)}} = \tilde{t}) = \tilde{P}^1(\cdot) \tag{6}$$

where $Q^1(\cdot \mid \overline{T^{(n)}} = \tilde{t})$ and $\tilde{P}^1(\cdot)$ refer to the marginal distribution of X_1 under $Q(\cdot \mid \overline{T^{(n)}} = \tilde{t})$ and $\tilde{P}$ respectively.

Our Results Both theorems above say that for some sets $\mathcal{A}$,

$$Q^n(\mathcal{A} \mid \overline{T^{(n)}} = \tilde{t}) \approx \tilde{P}^n(\mathcal{A}) \tag{7}$$

[2] We are referring here to the version in [13]. The theorem in [14] extends this in a direction different from the one we consider here.

[3] This theorem too has later been extended in several directions different from the one considered here [6]; see the discussion at the end of Section 4.

In the concentration phenomenon, the set $\mathcal{A} \subset \mathcal{X}^n$ is about the frequencies of individual outcomes in the sample. In the conditional limit theorem $\mathcal{A} \subset \mathcal{X}^1$ only concerns the first outcome. One might conjecture that (7) holds asymptotically in a much wider sense, namely for *just about any set whose probability one may be interested in.* For examples of such sets see Example 1. In Theorems 1 and 2 we show that (7) indeed holds for a very large class of sets; moreover, we give an explicit indication of the error one makes if one approximates $Q(\mathcal{A} \mid \overline{T^{(n)}} = \tilde{t})$ by $\tilde{P}(\mathcal{A})$. In this way we unify and strengthen both the concentration phenomenon and the conditional limit theorem. To be more precise, let $\{\mathcal{A}_n\}$, with $\mathcal{A}_i \subset \mathcal{X}^i$ be a sequence of 'typical' sets for $\tilde{P}$ in the sense that $\tilde{P}^n(\mathcal{A}_n)$ goes to 1 sufficiently fast. Then broadly speaking Theorem 1 shows that $Q^n(\mathcal{A}_n \mid \overline{T^{(n)}} = \tilde{t})$ goes to 1 too, 'almost' as fast as $\tilde{P}^n(\mathcal{A}_n)$. Theorem 2, our main theorem, says that, if m is an arbitrary increasing sequence with $\lim_{n\to\infty} m/n = 0$, then for *every* (measurable) sequence $\{\mathcal{A}_m\}$ (i.e. not just the typical ones), with $\mathcal{A}_m \subset \mathcal{X}^m$, $\tilde{P}^n(\mathcal{A}_m) \to Q^n(\mathcal{A}_m \mid \overline{T^{(n)}} = \tilde{t})$. Then, in Section 5, we first connect this to the notion of 'universal models' as arising in the MDL (Minimum Description Length) approach to inductive inference. We next show in what sense our strong concentration phenomena really provide a 'justification', not just a characterization, of MaxEnt. We show first (Theorem 3) that our concentration phenomenon implies that the MaxEnt distribution $\tilde{P}$ *uniquely* achieves the best minimax logarithmic loss achievable for sequential prediction of samples satisfying the constraint. We also show (Theorem 4) that for sequences that are algorithmically *random* relative to the constraint, $\tilde{P}$ achieves good loss also for loss functions other than the logarithmic loss.

3 Mathematical Preliminaries

The Sample Space From now on we assume a sample space $\mathcal{X} \subseteq \mathbf{R}^l$ for some $l > 0$ and let X be the random vector with $X(x) = x$ for all $x \in \mathcal{X}$. We reserve the symbol Q to refer to a distribution for X called the *prior distribution* (formally, Q is a distribution over $(\mathcal{X}, \sigma(X))$ where $\sigma(X)$ is the Borel-σ-algebra generated by X). We will be interested in sequences of i.i.d. random variables $X_1, X_2, \ldots$, all distributed according to Q. Whenever no confusion can arise, we use Q also to refer to the joint (product) distribution of $\times_{i\in\mathbf{N}} X_i$. Otherwise, we use Q^m to denote the m-fold product distribution of Q. The sample $(X_1, \ldots, X_m)$ will also be written as $X^{(m)}$.

The Constraint Functions T Let $T = (T_{[1]}, \ldots, T_{[k]})$ be a k-dimensional random vector that is $\sigma(X)$-measurable. We refer to the event $\{x \in \mathcal{X} \mid T(x) = t\}$ both as '$T(X) = t$' and as '$T = t$'. Similarly we write $T_i = t$ as an abbreviation of $T(X_i) = t$ and $T^{(n)}$ as short for $(T(X_1), \ldots, T(X_n))$. The *average* of n observations of T will be denoted by $\overline{T^{(n)}} := n^{-1} \sum_{i=1}^{n} T(X_i)$. We assume that the support of $\mathcal{X}$ is either countable (in which case the prior distribution Q admits a probability mass function) or that it is a connected subset of $\mathbf{R}^l$ for some $l > 1$ (in which case we assume that Q has a bounded continuous density with respect to Lebesgue measure). In both cases, we denote the probability mass function/density by q. If $\mathcal{X}$ is countable, we shall further assume that T is of the *lattice form* (which it will be in most applications):

Definition 1. [9, Page 490] *A k-dimensional lattice random vector $T = (T_{[1]}, \ldots, T_{[k]})$ is a random vector for which there exists real-valued $b_1, \ldots, b_k$ and $h_1, \ldots, h_k$ such that, for $1 \leq j \leq k$, $\forall x \in \mathcal{X} : T_{[j]}(x) \in \{b_j + sh_j \mid s \in \mathbf{N}\}$. We call the largest h_i for which this holds the* span *of $T_{[i]}$.*

If X is continuous, we shall assume that T is 'regular':

Definition 2. *We say a k-dimensional random vector is of* regular continuous form *if its distribution under Q admits a bounded continuous density with respect to Lebesgue measure.*

Maximum Entropy Throughout the paper, log is used to denote logarithm to base 2. Let P, Q be distributions for $\mathcal{X}$. We define $\mathbf{H}_Q(P)$, *the Q-entropy of P*, as

$$\mathbf{H}_Q(P) = -D(P||Q) \tag{8}$$

This is defined even if P or Q have no densities, see [5]. Assume we are given a constraint of form (2), i.e. $E_P[T(X)] = \tilde{t}$. Here $T = (T_{[1]}, \ldots, T_{[k]}), \tilde{t} = (\tilde{t}_{[1]}, \ldots, \tilde{t}_{[k]})$. We define, if it exists, $\tilde{P}$ to be the unique distribution over $\mathcal{X}$ that maximizes the Q-entropy over all distributions (over $(\mathcal{X}, \sigma(X))$) satisfying (2). That is, $\tilde{P}$ is given by (3). If Condition 1 below holds, then $\tilde{P}$ exists and is given by the exponential form (9), as expressed in the proposition below. In the condition, the notation $a^T b$ refers to the dot product between a and b.

Condition 1: There exists $\tilde{\beta} \in \mathbf{R}^k$ such that $Z(\tilde{\beta}) = \int_{x \in \mathcal{X}} \exp(-\tilde{\beta}^T T(x)) dQ(x)$ is finite and the distribution $\tilde{P}$ with density (with respect to Q)

$$\tilde{p}(x) := \frac{1}{Z(\tilde{\beta})} e^{-\tilde{\beta}^T T(x)} \tag{9}$$

satisfies $E_{\tilde{P}}[T(X)] = \tilde{t}$.

Proposition 1 ([5]). *Assume Condition 1 holds for Constraint (2). Then it holds for only one $\beta \in \mathbf{R}^k$ and* inf $\{D(P||Q) \mid P : E_P[T(X)] = \tilde{t}\}$ *is attained by (and only by) the $\tilde{P}$ given by (9).*

If Condition 1 holds, then $\tilde{t}$ determines both $\tilde{\beta}$ and $\tilde{P}$. In our theorems, we shall simply assume that Condition 1 holds. A sufficient (by no means necessary!) requirement for Condition 1 is for example that Q has bounded support; see [5] for a more precise characterization. We will also assume in our theorems the following natural condition:

Condition 2: The 'T-covariance matrix' Σ with $\Sigma_{ij} = E_{\tilde{P}}[T_{[i]} T_{[j]}] - E_{\tilde{P}}[T_{[i]}] E_{\tilde{P}}[T_{[j]}]$ is invertible.

Σ is guaranteed to exist by Condition 1 (see any book with a treatment of exponential families, for example, [18]) and will be singular only if either $\tilde{t}_j$ lies at the boundary of the range of $T_{[j]}$ for some j or if some of the $T_{[j]}$ are affine combinations of the others. In the first case, the constraint $T_{[j]} = \tilde{t}_j$ can be replaced by restricting the sample space to $\{x \in \mathcal{X} \mid T_{[j]}(x) = \tilde{t}_j\}$ and considering the remaining constraints for the new sample space. In the second case, we can remove some of the $T_{[i]}$ from the constraint without changing the set of distributions satisfying it, making Σ once again invertible.

4 The Concentration Theorems

Theorem 1. (the concentration phenomenon for typical sets, lattice case) *Assume we are given a constraint of form (2) such that T is of the lattice form and $h = (h_1, \ldots, h_k)$ is the span of T and such that conditions 1 and 2 hold. Then there exists a sequence $\{c_i\}$ satisfying*

$$\lim_{n\to\infty} c_n = \frac{\prod_{j=1}^{k} h_j}{\sqrt{(2\pi)^k \det \Sigma}}$$

such that

1. *Let $\mathcal{A}_1, \mathcal{A}_2, \ldots$ be an arbitrary sequence of sets with $\mathcal{A}_i \subset \mathcal{X}^i$. For all n with $Q(T_n = \tilde{t}) > 0$, we have:*

$$\tilde{P}(\mathcal{A}_n) \geq n^{-k/2} c_n Q(\mathcal{A}_n \mid \overline{T^{(n)}} = \tilde{t}). \tag{10}$$

Hence if $\mathcal{B}_1, \mathcal{B}_2, \ldots$ is a sequence of sets with $\mathcal{B}_i \subset \mathcal{X}^i$ whose probability tends to 1 under $\tilde{P}$ in the sense that $1 - \tilde{P}(\mathcal{B}_n) = O(f(n)n^{-k/2})$ for some function $f : \mathbf{N} \to \mathbf{R}$; $f(n) = o(1)$, then $Q(\mathcal{B}_n | \overline{T^{(n)}} = \tilde{t})$ tends to 1 in the sense that $1 - Q(\mathcal{B}_n | \overline{T^{(n)}} = \tilde{t}) = O(f(n))$.

2. *If for all n, $\mathcal{A}_n \subseteq \{x^{(n)} \mid n^{-1} \sum_{i=1}^{n} T(x_i) = \tilde{t}\}$ then (10) holds with equality.*

Theorem 1 has applications for coding/compression, Minimum Description Length inference and prediction. These are discussed in Section 5. The relation of the Theorem to Jaynes' original concentration phenomenon is discussed at the end of this section.

Proof. We need the following theorem[4]:

[4] Feller gives the local central limit theorem only for 1-dimensional lattice random variables with $E[T] = 0$ and $\mathrm{var}[T] = 1$; extending the proof to k-dimensional random vectors with arbitrary means and covariances is, however, completely straightforward: see XV.7 (page 494) of [9].

Theorem **('local central limit theorem for lattice random variables', [9], page 490)** Let $T = (T_{[1]}, \ldots, T_{[k]})$ be a lattice random vector and $h_1, \ldots, h_k$ be the corresponding spans as in Definition 1; let $E_P[T(X)] = t$ and suppose that P satisfies Condition 2 with T-covariance matrix Σ. Let $X_1, X_2, \ldots$ be i.i.d. with common distribution P. Let V be a closed and bounded set in $\mathbf{R}^k$. Let $v_1, v_2, \ldots$ be a sequence in V such that for all n, $P(\sum_{i=1}^n (T_i - t)/\sqrt{n} = v_n) > 0$. Then as $n \to \infty$,

$$\frac{n^{k/2}}{\prod_{j=1}^k h_j} P\left(\frac{\sum_{i=1}^n (T_i - t)}{\sqrt{n}} = v_n\right) - \aleph(v_n) \to 0.$$

Here $\aleph$ is the density of a k-dimensional normal distribution with mean vector $\mu = t$ and covariance matrix Σ.

The theorem shows that there exists a sequence $d_1, d_2, \ldots$ with $\lim_{n\to\infty} d_n = 1$ such that, for all n with $P(\sum_{i=1}^n (T_i - t) = \mathbf{0}) > 0$,

$$\frac{\frac{n^{k/2}}{\prod_{j=1}^k h_j} P\left(\frac{\sum_{i=1}^n (T_i - t)}{\sqrt{n}} = \mathbf{0}\right)}{\aleph(0)} = \frac{\sqrt{(2\pi n)^k \det \Sigma}}{\prod_{j=1}^k h_j} P\left(\frac{1}{n}\sum_{i=1}^n T_i = t\right) = d_n \qquad (11)$$

The proof now becomes very simple. First note that $\tilde{P}(\mathcal{A}_n \mid \overline{T^{(n)}} = \tilde{t}) = Q(\mathcal{A}_n \mid \overline{T^{(n)}} = \tilde{t})$ (write out the definition of conditional probability and realize that $\exp(-\tilde{\beta}^T T(x)) = \exp(-\tilde{\beta}^T \tilde{t}) =$ constant for all x with $T(x) = \tilde{t}$. Use this to show that

$$\begin{aligned}\tilde{P}(\mathcal{A}_n) &\geq \tilde{P}(\mathcal{A}_n, \overline{T^{(n)}} = \tilde{t}) = \tilde{P}(\mathcal{A}_n \mid \overline{T^{(n)}} = \tilde{t})\tilde{P}(\overline{T^{(n)}} = \tilde{t}) \\ &= Q(\mathcal{A}_n \mid \overline{T^{(n)}} = \tilde{t})\tilde{P}(\overline{T^{(n)}} = \tilde{t}).\end{aligned} \qquad (12)$$

Clearly, with $\tilde{P}$ in the rôle of P, the local central limit theorem is applicable to random vector T. Then, by (11), $\tilde{P}(\overline{T^{(n)}} = \tilde{t}) = (\prod_{j=1}^k h_j)/\sqrt{(2\pi n)^k \det \Sigma} d_n$. Defining $c_n := \tilde{P}(\overline{T^{(n)}} = \tilde{t}) n^{k/2}$ finishes the proof of item 1. For item 2, notice that in this case (12) holds with equality; the rest of the proof remains unchanged.

Example 1. The 'Brandeis dice example' is a toy example frequently used by Jaynes and others in discussions of the MaxEnt formalism [13]. Let $\mathcal{X} = \{1, \ldots, 6\}$ and X be the outcome in one throw of some given die. We initially believe (e.g. for reasons of symmetry) that the distribution of X is uniform. Then $Q(X = j) = 1/6$ for all j and $E_Q[X] = 3.5$. We are then told that the average number of spots is $E[X] = 4.5$ rather than 3.5. As calculated by Jaynes, the MaxEnt distribution $\tilde{P}$ given this constraint is given by

$$(\tilde{p}(1), \ldots, \tilde{p}(6)) = (0.05435, 0.07877, 0.11416, 0.16545, 0.23977, 0.34749). \qquad (13)$$

By the Chernoff bound, for every $j \in \mathcal{X}$, every $\epsilon > 0$, $\tilde{P}(|n^{-1}\sum_{i=1}^n I_j(X_i) - \tilde{p}(j)| > \epsilon) < 2\exp(-nc)$ for some constant $c > 0$ depending on ϵ; here

$I_j(X)$ is the indicator function for $X = j$. Theorem 1 then implies that $Q(|n^{-1}\sum_{i=1}^n I_j(X_i) - \tilde{p}(j)| > \epsilon | \overline{T^{(n)}} = \tilde{t}) = O(\sqrt{n}e^{-nc}) = O(e^{-nc'})$ for some $c' > 0$. In this way we recover Jaynes' original concentration phenomenon (5): the fraction of sequences satisfying the constraint with frequencies close to MaxEnt probabilities $\tilde{p}$ is overwhelmingly large. Suppose now we receive new information about an additional constraint: $P(X = 4) = P(X = 5) = 1/2$. This can be expressed as a moment constraint by $E[(I_4(X), I_5(X))^T] = (0.5, 0.5)^T$. We can now either use $\tilde{P}$ defined as in (13) in the rôle of prior Q and impose the new constraint $E[(I_4(X), I_5(X))^T] = (0.5, 0.5)^T$, or use uniform Q and impose the combined constraint $E[T] = E[(T_{[1]}, T_{[2]}, T_{[3]})^T] = (4.5, 0.5, 0.5)^T$, with $T_{[1]} = X, T_{[2]} = I_4(X), T_{[3]} = I_5(X)$. In both cases we end up with a new MaxEnt distribution $\tilde{\tilde{p}}(4) = \tilde{\tilde{p}}(5) = 1/2$. This distribution, while still consistent with the original constraint $E[X] = 4.5$, rules out the vast majority of sequences satisfying it. However, we can apply our concentration phenomenon again to the new MaxEnt distribution $\tilde{\tilde{P}}$. Let $\mathcal{I}_{j,j',\epsilon}$ denote the event that

$$|\frac{1}{n}\sum_{i=1}^{n} I_j(X_i) - \frac{\sum_{i=1}^{n-1} I_{j'}(X_i)I_j(X_{i+1})}{\sum_{i=1}^{n-1} I_{j'}(X_i)}| > \epsilon.$$

According to $\tilde{\tilde{P}}$, we still have that $X_1, X_2, \ldots$ are i.i.d. Then by the Chernoff bound, for each $\epsilon > 0$, for $j, j' \in \{4, 5\}$, $\tilde{\tilde{P}}(\mathcal{I}_{j,j',\epsilon})$ is exponentially small. Theorem 1 then implies that $Q^n(\mathcal{I}_{j,j'\epsilon} \mid \overline{T^{(n)}} = (4.5, 0.5, 0.5)^T)$ is exponentially small too: for the overwhelming majority of samples satisfying the combined constraint, the sample will look just as if it had been generated by an i.i.d. process, even though $X_1, \ldots, X_n$ are obviously not *completely* independent under $Q^n(\cdot | \overline{T^{(n)}} = (4.5, 0.5, 0.5)^T)$.

There also exists a version of Theorem 1 for continuous-valued random vectors. This is given, along with the proof, in technical report [12].

There are a few limitations to Theorem 1: (1) we must require that $\tilde{P}(\mathcal{A}_n)$ goes to 0 or 1 as $n \to \infty$; (2) the continuous case needed a separate statement, which is caused by the more fundamental (3) it turns out that the proof technique used cannot be adapted to point-wise conditioning on $\overline{T^{(n)}} = \tilde{t}$ in the continuous case, see [12]. Theorem 2 overcomes all these problems. The price we pay is that, when conditioning on $\overline{T^{(n)}} = \tilde{t}$, the sets $\mathcal{A}_m$ must only refer to $X_1, \ldots, X_m$ where m is such that $m/n \to 0$; for example, $m = \lceil n/\log n \rceil$ will work. Whenever in the case of continuous-valued T we write $Q(\cdot \mid \overline{T^{(n)}} = t)$ or $\tilde{P}(\cdot \mid \overline{T^{(n)}} = t)$ we refer to the continuous version of these quantities. These are easily shown to exist, see [12]. Recall that (for $m < n$) $Q^m(\cdot \mid \overline{T^{(n)}} = \tilde{t})$ refers to the marginal distribution of $X_1, \ldots, X_m$ conditioned on $\overline{T^{(n)}} = \tilde{t}$. It is implicitly understood in the theorem that in the lattice case, n ranges only over those values for which $Q(\overline{T^{(n)}} = \tilde{t}) > 0$.

Theorem 2. (Main Theorem: the Strong Concentration Phenomenon/ Strong Conditional Limit Theorem) *Let $\{m_i\}$ be an increasing sequence* *with* $m_i \in \mathbf{N}$*, such that* $\lim_{n\to\infty} m_n/n = 0$*. Assume we are given a constraint of form (2) such that* T *is of the regular continuous form or of the lattice form and suppose that Conditions 1 and 2 are satisfied. Then as* $n \to \infty$*,* $Q^{m_n}(\cdot \mid \overline{T^{(n)}} = \tilde{t})$ *converges weakly*[5] *to* $\tilde{P}^{m_n}(\cdot)$*.*

The proof (using the same key idea, but involving much more work than the proof of Theorem 1) is in technical report [12].

Related Results Theorem 1 is related to Jaynes' original concentration phenomenon, the proof of which is based on Stirling's approximation of the factorial. Another closely related result (also based on Stirling's approximation) is in Example 5.5.8 of [20]. Both results can be easily extended to prove the following weaker version of Theorem 1, item 1: $\tilde{P}(\mathcal{A}_n) \geq n^{-|\mathcal{X}|} c_n Q(\mathcal{A}_n | \overline{T^{(n)}} = \tilde{t})$ where c_n tends to some constant. Note that in this form, the theorem is void for infinite sample spaces. In [14] the original concentration phenomenon is extended in a direction somewhat different from Theorem 1; it would be interesting to study the relations.

Theorem 2 is similar to the original 'conditional limit theorems' (Theorems 1 and 2) of Van Campenhout and Cover [25]. We note that the preconditions for our theorem to hold are weaker and the conclusion is stronger than for the original conditional limit theorems, the main novelty being that Theorem 2 supplies us with an explicit bound on how fast m can grow as n tends to infinity. The conditional limit theorem was later extended by Csiszár [6]. His setting is considerably more general than ours (e.g. allowing for general convex constraints rather than just moment constraints), but his results also lack an explicit estimate of the rate at which m can increase with n. Csiszár [6] and Cover and Thomas [4] (where a simplified version of the conditional limit theorem is proved) both make the connection to large deviation results, in particular Sanov's theorem. As shown in the latter reference, weak versions of the conditional limit theorem can be interpreted as immediate consequences of Sanov's theorem.

5 Applications

For simplicity we restrict ourselves in this section to countable sample spaces $\mathcal{X}$ and we identify probability mass functions with probability distributions. Subsections 5.1 and 5.2 make frequent use of coding-theoretic concepts which we now briefly review (Sections 5.3 and 6 can be read without knowledge of coding/information theory).

Recall that by the Kraft Inequality [4], for every prefix code with lengths L over symbols from a countable alphabet $\mathcal{X}^n$, there exists a (possibly sub-additive) probability mass function p over $\mathcal{X}^n$ such that for all $x^{(n)} \in \mathcal{X}^n$,

[5] That is, for all sequences $\{\mathcal{A}_m\}$ where each $\mathcal{A}_m$ is a measurable continuity set $\mathcal{A}_m \subseteq \mathcal{X}^m$, $Q^{m_n}(\mathcal{A}_{m_n} \mid \overline{T^{(n)}} = \tilde{t}) \to \tilde{P}^{m_n}(\mathcal{A}_{m_n})$. A 'continuity set' $\mathcal{A}_m$ is a set such that the $\tilde{P}^m$-probability of the *boundary* of the set $\mathcal{A}_m$ is 0; in our case, *all* measurable sets $\mathcal{A}_m$ are continuity sets. See Theorem 2.1 of [3].

$L(x^{(n)}) = -\log p(x^{(n)})$. We will call this p the 'probability (mass) function corresponding to L'. Similarly, for every probability mass function p over $\mathcal{X}^n$ there exists a (prefix) code with lengths $L(x^{(n)}) = \lceil -\log p(x^{(n)}) \rceil$. Neglecting the round-off error, we will simply say that for every p, there exists a code with lengths $L(x^{(n)}) = -\log p(x^{(n)})$. We call the code with these lengths 'the code corresponding to p'. By the information inequality [4], this is also the most efficient code to use if data $X^{(n)}$ were actually distributed according to p.

We can now see that Theorem 1, item 2, has important implications for coding. Consider the following special case of Theorem 1, which obtains by taking $\mathcal{A}_n = \{x^{(n)}\}$ and logarithms:

Corollary 1. (the concentration phenomenon, coding-theoretic formulation) *Assume we are given a constraint of form (2) such that T is of the lattice form and $h = (h_1, \ldots, h_k)$ is the span of T and such that conditions 1 and 2 hold. For all n, all $x^{(n)}$ with $n^{-1} \sum_{i=1}^{n} T(x_i) = \tilde{t}$, we have*

$$
\begin{aligned}
&-\log \tilde{p}(x^{(n)}) = -\log q(x^{(n)} \mid \tfrac{1}{n} \textstyle\sum_{i=1}^{n} T(X_i) = \tilde{t}) + \\
&+ \tfrac{k}{2} \log 2\pi n + \log \sqrt{\det \Sigma} - \textstyle\sum_{j=1}^{k} \log h_j + o(1) = \\
&-\log q(x^{(n)} \mid \tfrac{1}{n} \textstyle\sum_{i=1}^{n} T(X_i) = \tilde{t}) + \tfrac{k}{2} \log n + O(1).
\end{aligned} \tag{14}
$$

In words, this means the following: let $x^{(n)}$ be a sample distributed according to Q, Suppose we are given the information that $n^{-1} \sum_{i=1}^{n} T(x_i) = \tilde{t}$. Then, by the information inequality, the most efficient code to encode $x^{(n)}$ is the one based on $q(\cdot | \overline{T^{(n)}} = \tilde{t})$ with lengths $-\log q(x^{(n)} \mid \overline{T^{(n)}} = \tilde{t})$. Yet if we encode $x^{(n)}$ using the code with lengths $-\log \tilde{p}(\cdot)$ (which would be the most efficient had $x^{(n)}$ been generated by $\tilde{p}$) then the number of extra bits we need is only of the order $(k/2) \log n$. That means, for example, that the number of additional bits we need *per outcome* goes to 0 as n increases. These and other consequences of Corollary 1 will be exploited in the next three subsections.

5.1 Connection to MDL, Stochastic Complexity, Two-Part Codes

Universal Models play a fundamental rôle in modern versions of the MDL (Minimum Description Length) approach to inductive inference and model selection [2, 22]. For details about universal models and codes as well as all coding-theoretic concepts appearing in this section, we refer to [2]. The material in the present section is not needed to understand later sections.

Let $\mathcal{M}_k = \{P_\theta(\cdot) \mid \theta \in \Gamma_k\}$, where $\Gamma_k \subseteq \mathbf{R}^k$ is a k-dimensional parametric class of i.i.d. distributions for sample space $\mathcal{X}$. Let C be a code for alphabet $\mathcal{X}^n$, with lengths L_C and define the *regret* $R_C(\cdot)$ such that for all $x^{(n)}$,

$$L_C(x^{(n)}) = -\log p_{\hat{\theta}(x^{(n)})}(x^{(n)}) + R_C(x^{(n)}),$$

where $\hat{\theta}(x^{(n)})$ is the (ML) Maximum Likelihood estimator in $\mathcal{M}_k$ for data $x^{(n)}$, assumed to exist. Roughly speaking, a *universal code* for sequences of length n is a code C such that the regret $r_C(x^{(n)})$ is small uniformly for all or (in some sense) 'most' $x^{(n)}$. A *universal model* is the probability distribution corresponding to a universal code.

It is well-known [2, 22] that, under mild regularity conditions, there exist universal codes C for $\mathcal{M}_k$ with lengths $L_C(x^{(n)}) = -\log p_{\hat{\theta}(x^{(n)})}(x^{(n)}) + \frac{k}{2}\log n + O(1)$, leading to regret

$$R_C(x^{(n)}) = \frac{k}{2}\log n + O(1) \tag{15}$$

Usually (15) holds uniformly for all sequences $x_1, x_2, \ldots$. (we sometimes need to restrict ourselves to a compact subset of Γ_k in order to make (15) uniformly true). It is also known that (15) is in some sense (up to $O(1)$) the best regret that can be achieved [21, 2]. Therefore, *every* code that achieves (15) is usually called a 'universal code', and its corresponding distribution 'universal model'. Until very recently there were four known ways to construct a universal model for a given class $\mathcal{M}_k$: the two-part code, the Bayesian mixture-code , the Shtarkov-normalized-maximum-likelihood (NML) code and the predictive or 'prequential' code, see [2]. These four methods, while superficially very different, all share the same asymptotic lengths (15). Under further regularity conditions on $\mathcal{M}_k$ and if the code C that is used is allowed to depend on sample size n, (15) the Shtarkov-NML and two-part codes can be refined to give [2]:

$$R_C(x^{(n)}) = \frac{k}{2}\log\frac{n}{2\pi} + \log\int_{\Gamma_k}\sqrt{\det I(\theta)}d\theta + o(1), \tag{16}$$

where $I(\theta)$ is the (expected) Fisher information matrix of θ. Quite recently, Rissanen [23] showed that the regret (16) is the best that can be achieved under at least three different definitions of optimality. $L_C(x^{(n)}) = -\log p_{\hat{\theta}(x^{(n)})}(x^{(n)}) + R_C(x^{(n)})$, with $R_C(x^{(n)})$ given by (16), is called the 'stochastic complexity of $x^{(n)}$ relative to $\mathcal{M}_k$'.

In the same recent reference [23], Rissanen implicitly introduced a new type of universal code that achieves regret (16). We illustrate this kind of code for the simple case where $\mathcal{M}_k$ is a k-dimensional exponential family with finite sample space $\mathcal{X}$. Let then $\mathcal{M}_k = \{P_\theta(\cdot) \mid \theta \in \Gamma_k\}$ be a k-parameter exponential family for $\mathcal{X}$ with Γ_k the mean-value parameter space, q the background measure and sufficient statistic $T = (T_{[1]}, \ldots, T_{[k]})$. Then $p_\theta = \tilde{p}$ with $\tilde{p}$ given by (9), and $E_{p_\theta}[T] = \tilde{t} = \theta$.

We will encode $x^{(n)}$ in a way similar to (but, as we shall see, still essentially different from) the two-part coding technique [2]: we first code (describe) a distribution for $\mathcal{X}^n$ and then code the data 'with the help of' this distribution. In our case, for data $x^{(n)}$, we first encode the ML estimator $\hat{\theta}(x^{(n)})$ using some code C_1 with lengths L_1. We then encode $x^{(n)}$ itself using some code C_2, making use of the fact that its ML estimator is $\hat{\theta}(x^{(n)})$. By the Kraft inequality this can be done using $L_2(x^{(n)} \mid \hat{\theta}(x^{(n)})) = -\log q(x^{(n)} \mid \overline{T^{(n)}} = \hat{\theta}(x^{(n)})) = -\log q(x^{(n)} \mid \hat{\theta}(x^{(n)}))$ bits. This leads to a code C^* that allows us to encode all $x^{(n)} \in \mathcal{X}^n$ by concatenating the code words of $\hat{\theta}(x^{(n)})$ (under C_1) and $x^{(n)}|\hat{\theta}(x^{(n)})$ (under C_2).

Since $\mathcal{X}$ is finite, $n^{-1}\sum_{i=1}^{n} T_{[j]}(X_i)$ can only take on $n \cdot |\mathcal{X}|$ distinct values. Therefore, we can choose C_1 such that $L_1(\hat{\theta}(x^{(n)})) = k \log n + k \log |\mathcal{X}|$. By Corollary 1 the code $L_2(\cdot|\cdot)$ has lengths

$$L_2(x^{(n)} \mid \hat{\theta}(x^{(n)})) = -\log p_{\hat{\theta}(x^{(n)})}(x^{(n)}) - \frac{k}{2}\log n - O(1). \tag{17}$$

Summing L_1 and L_2, we see that the total code length $L^*(x^{(n)})$ for *arbitrary* $x^{(n)}$ is bounded by $\frac{k}{2}\log n - \log p_{\hat{\theta}(x^{(n)})}(x^{(n)}) + O(1)$. Therefore, the regret satisfies (15) which suggests that C^* is a universal code for $\mathcal{M}_k$. Indeed, in the technical report [12] we sketch how, by changing C_1, we can even refine C^* such that the regret (16) is achieved.

Other relations between MDL and Maximum Entropy have been investigated by Feder [8] and Li and Vitányi [20]. In the next section we will see how Theorem 1 leads to yet another relation between minimum code length and Maximum Entropy.

5.2 Empirical Constraints and Game Theory

Recall we assume countable $\mathcal{X}$. The σ-algebra of such $\mathcal{X}$ is always tacitly taken to be the power set of $\mathcal{X}$. The σ-algebra thus being implicitly understood, we can define $\mathcal{P}(\mathcal{X})$ to be the set of all probability distributions over $\mathcal{X}$. For a product $\mathcal{X}^\infty = \times_{i\in\mathbf{N}}\mathcal{X}$ of a countable sample space $\mathcal{X}$, we define $\mathcal{P}(\mathcal{X}^\infty)$ to be the set of all distributions over the product space with the associated product σ-algebra.

In [24, 10], a characterization of Maximum Entropy distributions quite different from the present one was given. It was shown that, under regularity conditions,

$$= \sup_{p^*:E_{p^*}[T]=\tilde{t}} \ \inf_{p} \ E_{p^*}[-\log\frac{p(X)}{q(X)}] = \inf_{p} \ \sup_{p^*:E_{p^*}[T]=\tilde{t}} \ E_{p^*}[-\log\frac{p(X)}{q(X)}] \tag{18}$$

where both p and p^* are understood to be members of $\mathcal{P}(\mathcal{X})$ and $\mathbf{H}_q(\tilde{p})$ is defined as in (1). By this result, the MaxEnt setting can be thought of as a game between Nature, who can choose any p^* satisfying the constraint, and Statistician, who only knows that Nature will choose a p^* satisfying the constraint. Statistician wants to minimize his worst-case expected logarithmic loss (relative to q), where the worst-case is over all choices for Nature. It turns out that the minimax strategy for Statistician in (18) is given by $\tilde{p}$. That is,

$$\tilde{p} = \arg\inf_{p} \sup_{p^*:E_{p^*}[T]=\tilde{t}} E_{p^*}[-\log\frac{p(x)}{q(x)}]. \tag{19}$$

This gives a decision-theoretic justification of using MaxEnt probabilities which seems quite different from our concentration phenomenon. Or is it? Realizing that in practical situations we deal with empirical constraints of form (4) rather than (2) we may wonder what distribution $\hat{p}$ is minimax in the empirical version of problem (19). In this version Nature gets to choose an individual sequence rather than a distribution[6]. To make this precise, let

$$\mathcal{C}_n = \{x^{(n)} \in \mathcal{X}^n \mid n^{-1} \sum_{i=1}^{n} T(x_i) = \tilde{t}\}. \tag{20}$$

Then, for n with $\mathcal{C}_n \neq \emptyset$, $\hat{p}_n$ (if it exists) is defined by

$$\hat{p}_n := \arg \inf_{p \in \mathcal{P}(\mathcal{X}^n)} \sup_{x^{(n)} \in \mathcal{C}_n} -\log \frac{p(x_1, \ldots, x_n)}{q(x_1, \ldots, x_n)} = \arg \sup_{p \in \mathcal{P}(\mathcal{X}^n)} \arg \inf_{x^{(n)} \in \mathcal{C}_n} \frac{p(x^{(n)})}{q(x^{(n)})} \tag{21}$$

$\hat{p}_n$ can be interpreted in two ways: (1) it is the distribution that assigns 'maximum probability' (relative to q) to all sequences satisfying the constraint; (2) as $-\log(\hat{p}(x^{(n)})/q(x^{(n)})) = \sum_{i=1}^{n}(-\log \hat{p}(x_i|x_1, \ldots, x_{i-1}) + \log q(x_i|x_1, \ldots, x_{i-1}))$, it is also the p that minimizes cumulative worst-case logarithmic loss relative to q when used for sequentially predicting $x_1, \ldots, x_n$.

One immediately verifies that $\hat{p}_n = q^n(\cdot \mid \overline{T^{(n)}} = \tilde{t})$: the solution to the empirical minimax problem is just the conditioned prior, which we know by Theorems 1 and 2 is in some sense very close to $\tilde{p}$. However, for no single n, $\tilde{p}$ is exactly equal to $q^n(\cdot \mid \overline{T^{(n)}} = \tilde{t})$. Indeed, $q^n(\cdot \mid \overline{T^{(n)}} = \tilde{t})$ assigns zero probability to any sequence of length n not satisfying the constraint. This means that using q in prediction tasks against the logarithmic loss will be problematic if the constraint only holds approximately (as we will discuss in more detail in the journal version of this paper) and/or if n is unknown in advance. In the latter case, it is impossible to use $q(\cdot \mid \overline{T^{(n)}} = \tilde{t})$ for prediction without modification. The reason is that there exist sequences $x^{(n_2)}$ of length $n_2 > n_1$ satisfying the constraint such that $q(x^{(n_2)}|x^{(n_1)} \in \mathcal{C}_{n_1}) = 0$. We may guess that in this case (n not known in advance), the MaxEnt distribution $\tilde{p}$, rather than $q(\cdot|\overline{T^{(n)}} = \tilde{t})$ is actually the optimal distribution to use for prediction. The following theorem shows that this is indeed so:

Theorem 3. *Let $\mathcal{X}$ be a countable sample space. Assume we are given a constraint of form (2) such that T is of the lattice form, and such that Conditions 1 and 2 are satisfied. Let $\mathcal{C}_n$ be as in (20). Then the infimum in*

$$\inf_{p \in \mathcal{P}(\mathcal{X}^\infty)} \sup_{\{n \,:\, \mathcal{C}_n \neq \emptyset\}} \sup_{x^{(n)} \in \mathcal{C}_n} -\frac{1}{n} \log \frac{p(x_1, \ldots, x_n)}{q(x_1, \ldots, x_n)} \tag{22}$$

is achieved by the Maximum Entropy distribution $\tilde{p}$, and is equal to $\mathbf{H}_q(\tilde{p})$.

Proof. Let $\mathcal{C} = \cup_{i=1}^{\infty} \mathcal{C}_i$. We need to show that for all n, for all $x^{(n)} \in \mathcal{C}$,

$$\mathbf{H}_q(\tilde{p}) = -\frac{1}{n} \log \frac{\tilde{p}(x^{(n)})}{q(x^{(n)})} = \inf_{p \in \mathcal{P}(\mathcal{X}^\infty)} \sup_{\{n \,:\, \mathcal{C}_n \neq \emptyset\}} \sup_{x^{(n)} \in \mathcal{C}_n} -\frac{1}{n} \log \frac{p(x^{(n)})}{q(x^{(n)})} \tag{23}$$

[6] To our knowledge, we are the first to analyze this 'empirical' game.

Equation (23) implies that $\tilde{p}$ reaches the inf in (22) and that the inf is equal to $\mathbf{H}_q(\tilde{p})$. The leftmost equality in (23) is a standard result about exponential families of form (9); see for example, [11, Proposition 4.1] or [22]. To prove the rightmost equality in (23), let $x^{(n)} \in \mathcal{C}_n$. Consider the conditional distribution $q(\cdot \mid x^{(n)} \in \mathcal{C}_n)$. Note that, for every distribution p_0 over $\mathcal{X}^n$, $p_0(x^{(n)}) \leq q(x^{(n)}|x^{(n)} \in \mathcal{C}_n)$ for at least one $x^{(n)} \in \mathcal{C}_n$. By Theorem 1 (or rather Corollary 1), for this $x^{(n)}$ we have

$$-\frac{1}{n}\log\frac{p_0(x^{(n)})}{q(x^{(n)})} \geq -\frac{1}{n}\log\frac{\tilde{p}(x^{(n)})}{q(x^{(n)})} - \frac{k}{2n}\log n - O(\frac{1}{n}),$$

and we see that for every distribution p_0 over $\mathcal{X}^\infty$,

$$\sup_{\{n\,:\,\mathcal{C}_n \neq \emptyset\}} \sup_{x^{(n)} \in \mathcal{C}_n} -\frac{1}{n}\log\frac{p_0(x^{(n)})}{q(x^{(n)})} \geq \sup_{\{n\,:\,\mathcal{C}_n \neq \emptyset\}} \sup_{x^{(n)} \in \mathcal{C}_n} -\frac{1}{n}\log\frac{\tilde{p}(x^{(n)})}{q(x^{(n)})},$$

which shows the rightmost equality in (23).

5.3 Maximum Entropy and Algorithmic Randomness

In the algorithmic theory of randomness, [20], one (broadly speaking) identifies *randomness of individual sequences* with incompressibility of such sequences. In this section we show that a sequence that is 'random relative to a given constraint' is 'almost' random with respect to the MaxEnt distribution $\tilde{P}$ for the constraint. The reader who is not familiar with Martin-Löf randomness is urged to move on to Theorem 4 which demonstrates the consequences of this fact for prediction based on MaxEnt distributions.

Throughout this section we assume $\mathcal{X}$ to be finite and Q to be uniform, so maximizing the entropy reduces to the 'original' Maximum (Shannon) Entropy formalism. Let $\mathcal{U} := \cup_{i=1}^{\infty}\mathcal{X}^i$. For $\mathbf{x}, \mathbf{y} \in \mathcal{U}$, $K(\mathbf{x}|\mathbf{y})$ will stand for the prefix Kolmogorov complexity of sequence $\mathbf{x}$ conditional on $\mathbf{y}$; $K(\mathbf{x})$ stands for $K(\mathbf{x}|\lambda)$ where λ is the empty sequence. For a finite set $\mathcal{C} \subset \mathcal{U}$, $K(\mathbf{x}|\mathcal{C})$ is the prefix complexity of $\mathbf{x}$ conditional on $\mathbf{x} \in \mathcal{C}$. Kolmogorov complexity is defined here with respect to some fixed universal reference prefix Turing Machine. For precise definitions of all these concepts, see Section 3.1 and Exercise 2.2.12. of [20].

Theorem (**Theorem 3.6.1 and Corollary 4.5.2 of [20]**) *An infinite sequence* $(x_1, x_2, \ldots) \in \mathcal{X}^\infty$ *is Martin-Löf random with respect to the uniform distribution iff there exists a constant* c *such that for all* n, $K(x_1, \ldots, x_n) \geq n - c$.

Here, we take this characterization of Martin-Löf randomness as basic. We will extend the notion of randomness to *sequences conditional on constraints* in an obvious manner. Let $\{\mathcal{C}_n\}$ be a sequence of constraints, where $\mathcal{C}_n \subseteq \mathcal{X}^n$ (we identify constraints with the set of sequences satisfying them). The theorem above suggests the following definition:

Definition 3. *An infinite sequence* $(x_1, x_2, \ldots) \in \mathcal{X}^\infty$ *is called random with respect to the sequence of constraints* $\{\mathcal{C}_n\}$ *(relative to the uniform distribution) iff there exists a constant* c *such that for all* n *with* $\mathcal{C}_n \neq \emptyset$, *we have* $K(x^{(n)}|\mathcal{C}_n) \geq \log|\mathcal{C}_n| - c$.

In our situation, the constraint is of form (20). Because of this simple form and since $\mathcal{X}$ is finite, there exists a fixed-length program that, for each n, when input $\langle n, \mathbf{x} \rangle$ with $\mathbf{x} \in \mathcal{X}^n$, outputs 1 iff $\mathbf{x} \in \mathcal{C}_n$ and 0 otherwise. Therefore the definition reduces to $(x_1, x_2, \ldots)$ *is random iff* $\exists c \forall n : \mathcal{C}_n \neq \emptyset \Rightarrow K(x^{(n)}|n) \geq \log |\mathcal{C}_n| - c$.

By Theorem 1, if $(x_1, x_2, \ldots)$ is random with respect to the constraints $\{\mathcal{C}_n\}$, then for all $x^{(n)} \in \mathcal{C}_n$,

$$K(x^{(n)}|n) \geq \log |\mathcal{C}_n| - O(1) = -\log \tilde{p}(x^{(n)}) - \frac{k}{2} \log n - O(1). \qquad (24)$$

In words: (see Corollary 4.5.2 of [20]) *If* $(x_1, x_2, \ldots)$ *is random with respect to the constraints* $\{\mathcal{C}_n\}$ *(relative to the uniform distribution) then* $(x_1, x_2, \ldots)$ *is 'almost' Martin-Löf random with respect to the maximum entropy distribution* $\tilde{p}$.

Equation 24 suggests that for the overwhelming majority of sequences satisfying the constraint (namely, those that are random with respect to the constraint), sequentially predicting outcomes in the sequence on the basis of the MaxEnt distribution leads to almost optimal results, no matter what loss function we use. The following theorem shows that this is indeed so. It holds for general prior distributions Q and is proved in technical report [12]. Consider a loss function $\text{LOSS} : \mathcal{X} \times \Delta \to [0, \infty]$ where Δ is some space of *predictions* or *decisions*. A *prediction (decision) strategy* δ^* is a function $\delta^* : \cup_{i=0}^{\infty} \mathcal{X}^i \to \Delta$. $\delta^*(x_1, \ldots, x_n)$ is to be read as 'the prediction/decision for X_{n+1} based on initial data $(x_1, \ldots, x_n)$'. We assume

Condition 3 $\mathcal{X}$ is finite. $\text{LOSS}(x; \cdot)$ is continuous in its second argument for all $x \in \mathcal{X}$. Δ is a compact convex subspace of $\mathbf{R}^l$ for some $l > 0$.

Under this condition, there exists at least one δ attaining $\inf E_{\tilde{P}}[\text{LOSS}(X; \delta)]$. Fix any such optimal (under $\tilde{P}$) decision and denote it $\tilde{\delta}$.

Theorem 4. *Suppose that T is of lattice form and suppose Conditions 1, 2 and 3 hold. Then (letting n reach over all numbers such that $Q(\overline{T^{(n)}} = \tilde{t}) > 0$), for all decision strategies δ^*, for all $\epsilon > 0$, there exists a $c > 0$ such that*

$$Q\Big(\frac{1}{n}\Big(\sum_{i=1}^{n} \text{LOSS}(x_i; \tilde{\delta}) - \sum_{i=1}^{n} \text{LOSS}(x_i; \delta^*(x_1, \ldots, x_{i-1}))\Big) > \epsilon \;\Big|\; \overline{T^{(n)}} = \tilde{t}\Big) = O(e^{-cn}). \qquad (25)$$

6 Consequences for Prediction

We summarize the implications of our results for prediction of individual sequences based on Maximum Entropy distributions. In this section $\mathcal{X}$ is finite and Q stands for the uniform distribution. Suppose then you have to make predictions about a sequence $(x_1, \ldots, x_n)$. You know the sequence satisfies the given constraint (i.e. for some n, $x^{(n)} \in \mathcal{C}_n$, with $\mathcal{C}_n$ as in (20)), but you do not know the length n of the sequence in advance. We distinguish between the special case of the log loss function $\text{LOSS}(x; p) = -\log p(x)$ and the general case of arbitrary (computable) loss functions.

1. **(log loss)** The MaxEnt distribution $\tilde{p}$ is worst-case optimal with respect to log loss, where the worst-case is over all sequences of all lengths satisfying the constraint. This is a consequence of Theorem 3.
2. **(log loss)** Whatever sample $x^{(n)} \in \mathcal{C}_n$ arrives, the average log loss you make per outcome when you predict outcomes using $\tilde{p}$ is determined in advance and will be exactly equal to $\mathbf{H}_q(\tilde{p}) = E_{\tilde{p}}[-\log \tilde{p}(X)]$. This is also a consequence of Theorem 3.
3. **(log loss)** For the overwhelming majority of sequences satisfying the constraint, $\tilde{p}$ will be asymptotically almost optimal with respect to log loss in the following sense: the excess loss of $\tilde{p}$ over *every* other prediction strategy (including strategies depending on past data) is at most a sub-linear function of n. This is a consequence of Theorem 4. In Example 1, an example of an exceptional sequence for which $\tilde{p}$ is not optimal would be any sequence consisting of 50% fours and 50% fives.
4. **(general loss)** For *every* regular loss function LOSS (satisfying Condition 3), predicting using $\tilde{\delta}$, (that is, *acting as if the sample had been generated by* $\tilde{p}$) leads to almost optimal predictions for the overwhelming majority of sequences satisfying the constraint, in the following sense: the excess loss of $\tilde{\delta}$ over *every* other prediction strategy is at most a sub-linear function of n. This is a consequence of Theorem 4.

We stress that the fact that items (3) and (4) hold for the overwhelming majority of sequences certainly does *not* imply that they will hold on actual, real-world sequences! Often these will exhibit more regularity than the observed constraint, and then $\tilde{\delta}$ is not necessarily optimal any more.

There are two important points we have neglected so far: (1) in practice, the given constraints will often only hold approximately. (2) the results of this paper have important implications for Maximum Likelihood and Bayesian prediction of sequences based on model classes that are exponential families [1]. The reason is that the Maximum Entropy model for constraint $E[T] = \tilde{t}$ is the Maximum *Likelihood* model for the (exponential family) model class given by (9) for *every* sequence of data $x^{(n)}$ with $n^{-1}\sum_{i=1}^{n} T(x_i) = \tilde{t}$ (see e.g. [4] or [11]) . The connection will be further discussed in the journal version of this paper.

References

1. K. Azoury and M. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI '99)*, pages 31–40. Morgan Kaufmann, 1999.
2. A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760, 1998.
3. P. Billingsley. *Convergence of Probability Measures.* Wiley, 1968.
4. T.M. Cover and J.A. Thomas. *Elements of Information Theory.* Wiley Interscience, New York, 1991.

5. I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975.
6. I. Csiszár. Sanov property, generalized i-projection and a conditional limit theorem. *The Annals of Probability*, 12(3):768–793, 1984.
7. I. Csiszár. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *The Annals of Statistics*, 19(4):2032–2066, 1991.
8. M. Feder. Maximum entropy as a special case of the minimum description length criterion. *IEEE Transactions on Information Theory*, 32(6):847–849, 1986.
9. W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 2. Wiley, 1968. Third edition.
10. P. D. Grünwald. Maximum entropy and the glasses you are looking through. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI 2000)*. Morgan Kaufmann Publishers, 2000.
11. P.D. Grünwald. *The Minimum Description Length Principle and Reasoning under Uncertainty*. PhD thesis, University of Amsterdam, The Netherlands, October 1998. Available as ILLC Dissertation Series 1998-03; see www.cwi.nl/~pdg.
12. P.D. Grünwald. Strong entropy concentration, coding, game theory and randomness. Technical Report 010, EURANDOM, 2001.
13. E.T. Jaynes. Where do we stand on maximum entropy? In R.D. Levine and M. Tribus, editors, *The Maximum Entropy Formalism*, pages 15–118. MIT Press, Cambridge, MA, 1978.
14. E.T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(939-951), 1982.
15. E.T. Jaynes. *Papers on Probability, Statistics and Statistical Physics*. Kluwer Academic Publishers, second edition, 1989.
16. E.T. Jaynes. Probability theory: the logic of science. Available at ftp://bayes.wustl.edu/Jaynes.book/, 1996.
17. J. N. Kapur and H. K Kesavan. *Entropy Optimization Principles with Applications*. Academic Press, Inc., 1992.
18. R.E. Kass and P.W. Voss. *Geometrical Foundations of Asymptotic Inference*. Wiley Interscience, 1997.
19. J. Lafferty. Additive models, boosting and inference for generalized divergences. In *Proceedings of the Twelfth Annual Workshop on Computational Learning Theory (COLT '99)*, 1999.
20. M. Li and P.M.B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer-Verlag, New York, revised and expanded second edition, 1997.
21. N. Merhav and M. Feder. A strong version of the redundancy-capacity theorem of universal coding. *IEEE Transactions on Information Theory*, 41(3):714–722, 1995.
22. J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific Publishing Company, 1989.
23. J. Rissanen. Strong optimality of the normalized ML models as universal codes, 2001. To appear in *IEEE Transactions on Information Theory*.
24. F. Topsøe. Information theoretical optimization techniques. *Kybernetika*, 15(1), 1979.
25. J. van Campenhout and T. Cover. Maximum entropy and conditional probability. *IEEE Transactions on Information Theory*, IT-27(4):483–489, 1981.