Explicit Bounds for Entropy Concentration under Linear Constraints*

Kostas N. Oikonomou†
AT&T Labs Research
Middletown, NJ 07748, U.S.A.

August 1, 2011

Abstract

Consider the construction of an object composed of m parts by distributing n units to those parts. For example, say we are assigning n balls to m boxes. Each assignment results in a certain count vector, specifying the number of balls allocated to each box. If only assignments satisfying a set of constraints that are linear in these counts are allowable, and m is fixed while n increases, most assignments that satisfy the constraints result in frequency vectors (normalized counts) whose entropy approaches that of the maximum entropy vector satisfying the constraints. This phenomenon of "entropy concentration" is known in various forms, and is one of the justifications of the maximum entropy method, one of the most powerful tools for solving problems with incomplete information. The appeal of entropy concentration comes from the simplicity of the argument: it is based purely on counting.

Existing proofs of the concentration phenomenon are based on limits or asymptotics. Here we present non-asymptotic, explicit lower bounds on n for a number of variants of the concentration result to hold to any prescribed accuracies, taking into account the fact that allocations of discrete units can satisfy constraints only approximately. The results are illustrated with examples on die tossing, vehicle or network traffic, and the probability distribution of the length of a $G/G/1$ queue.

1 Introduction

Consider a process which is repeated n times and each repetition has m possible outcomes. For concreteness we may think of assigning n balls to m labelled boxes, where each box can hold any number of balls. The first ball can go into any box, the second ball can go into any box, ..., and the nth ball can go into any box. Each assignment or allocation is thus a

*Keywords: maximum entropy, concentration, bounds, linear constraints, tolerances
†*Email:* ko@research.att.com

sequence of n box labels and results in some number ν_1 of balls in box 1, ν_2 in box 2, etc., where the ν_i are $\geqslant 0$ and sum to n. There are m^n possible assignments in all, and many of them can lead to the same count vector $\nu = (\nu_1, \ldots, \nu_m)$. We refer to these assignments as the *realizations* of the count vector.

The arrangement of n balls into m boxes can represent the construction of any object consisting of m distinguishable parts out of n identical units. So if the balls represent pixels of an image, the attributes of color and (suitably discretized) intensity are ascribed to the boxes to which the pixels are assigned. Then the count vector is thought of as a 2-dimensional matrix with rows labelled by intensity and columns by color. Other examples are people categorized by age, height, and weight, vehicles classified by weight, size, and fuel economy, packets in a communications network with attributes of origin, destination, size, and timestamp, etc. The object can even be a (discrete) probability distribution. When the process simply represents the classification of n units by m discrete or discretized attibutes, it is known as a (multi-dimensional) contingency table.

Now consider imposing constraints $\mathcal{C}$ on the allowable assignments, expressed as a set of *linear* relations on the elements of the *frequency* vector $f = (\nu_1/n, \ldots, \nu_m/n)$ corresponding to the counts ν. E.g. $5f_1 - 17.4f_2 \geqslant 0.131$, $f_{12} \leqslant f_{15}$, etc. As n grows, the frequency vectors of more and more of the assignments that satisfy the constrtaints will have *entropy* closer and closer to that of a particular m-vector φ^*, the vector of *maximum entropy* H^* subject to the constraints $\mathcal{C}$. (We denote this vector by φ^*, as opposed to f^*, to emphasize that its entries are, in general, not rational.) This result is known, more or less, in many forms: the original is E.T. Jaynes's "entropy concentration theorem" [Jay82], [Jay83], in the information theory literature it is the "conditional limit theorem" [CT91], and in computer science there is "strong entropy concentration" [Gr1], [Gr8][1]. All these results involve limits or asymptotics in one way or another, i.e. in the statement

EC: given an $\varepsilon > 0$ and an $\eta > 0$, there is a $N(\varepsilon, \eta)$ such that for all $n \geqslant N(\varepsilon, \eta)$, the fraction of assignments that satisfy $\mathcal{C}$ and have a frequency vector with entropy within η of H^* is at least $1 - \varepsilon$,

one or more of the quantities ε, η, or N is not given explicitly. For example, it is well-known that the fraction of assignments that don't satisfy $\mathcal{C}$ is $O(e^{-\eta n H^*})$. Note that *EC* is simply a problem of *counting*; there is no uncertainty, no randomness, and there are no probabilities anywhere. (However, the results can be applied to the *derivation* of probability distributions.) Our purpose in this paper is to derive *explicit* expressions for $N(\varepsilon, \eta)$ assuming that the maximum entropy vector $\varphi^* \in \mathbb{R}^m$ and its entropy H^* are known[2]. Given a concrete problem with incomplete information, these expressions allow

[1] When we say "more or less" and "in many forms" we mean that similar statements are made about similar or the same things, but it is not our purpose here to enter into a detailed comparison.

[2] "Explicit" means that there is not a single O, not even a Θ, and much less an Ω to be found in the whole paper.

us to assess the "reliability" of the MaxEnt solution to it as we illustrate in §4. We also establish a number of new results, as detailed at the end of this section.

Before proceeding, we give a very simple illustration. Consider assigning 5 balls to 3 boxes labelled A, B, C without any constraints at all (other than the fact that all balls must be assigned, i.e. the frequencies must add up to 1). There are $3^5 = 243$ possible assignments, e.g. A, A, B, A, C, meaning that the first two balls go into box A, the third into box B, the fourth into A again, and the fifth into box C. Table 1.1 lists the possible box occupancies or count vectors, and the number of *realizations* of each count vector, denoted by #, i.e. the number of assignments that result in this vector. These numbers are given by multinomial coefficients, e.g. $\#(3,0,2) = \binom{5}{3,0,2} = 10$. Finally the table gives the entropy $H(f) = -\sum_i f_i \ln f_i$ of the frequency vector $f = \nu/5$ corresponding to each count vector ν. Both Table 1.1 and its graphical counterpart, Fig. 1.1, show the beginnings of the concentration phenomenon even in this very small case.

count vector ν	$\#\nu$	$H(f)$	count vector ν	$\#\nu$	$H(f)$	count vector ν	$\#\nu$	$H(f)$
5,0,0	1	0	3,1,1	**20**	0.950	0,2,3	10	0.673
4,0,1	5	0.500	2,1,2	**30**	1.055	2,3,0	10	0.673
3,0,2	10	0.673	1,1,3	**20**	0.950	1,3,1	**20**	0.950
2,0,3	10	0.673	0,1,4	5	0.500	0,3,2	10	0.673
1,0,4	5	0.500	3,2,0	10	0.673	1,4,0	5	0.500
0,0,5	1	0	2,2,1	**30**	1.055	0,4,1	5	0.500
4,1,0	5	0.500	1,2,2	**30**	1.055	0,5,0	1	0

Table 1.1: $m = 3, n = 5$. The $3^5 = 243$ realizations/assigments exhibit rudimentary entropy concentration: 150 of them have frequency vectors with entropy within 23% of $H^* = \ln 3 = 1.099$.

The main point of this small example is to re-emphasize that statement *EC* has to do simply with counting, and nothing to do with any probability considerations. Nevertheless, the reader might still think that we are simply avoiding the introduction of a uniform p.d. on the set of all m^n possible assignments, and that without the assumption that all these assignments or outcomes are equally likely, the concentration statement *EC* really has little "practical" significance[3]. In fact, quite the opposite is true: the phenomenon of entropy concentration *justifies* the assumption of uniformity in the absence of any other knowledge, i.e. Laplace's famous "principle of indifference" or "principle of insufficient reason"! Indeed, in the absence of any constraints other than that the frequencies must sum to 1, entropy concentration shows that the *uniform* frequeny distribution is simply the one that can be realized in the greatest number of ways, or *most likely* to be realized ([Jay82], §IV), and is therefore to be preferred; we see indications of this even in our small example. This is a reversal in our usual way of thinking.

In the following development we will take the dimension m of the problem to be given

[3]Even the author has— occasionally—fallen prey to this ingrained viewpoint.

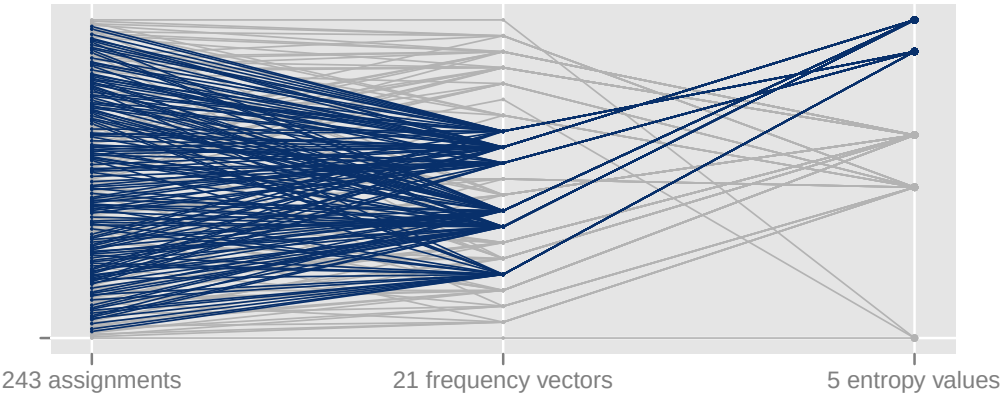


Figure 1.1: Graphical representation of the data of Table 1.1: assignments $\to$ count/frequency vectors $\to$ entropies. The 6 frequency vectors with the two highest entropies have more than half of all the realizations, and the most likely vectors to be realized are the ones closest to the uniform $(1/3, 1/3, 1/3)$.

and fixed, and concern ourselves solely with n. For a given n, we denote the set of all n-frequency vectors by F_n, i.e. $F_n = \{(f_1, \ldots, f_m) | f_i = \nu_i/n, \nu_1 + \cdots + \nu_m = n\}$. We represent the constraints $\mathcal{C}$ on the frequency vector f or count vector ν by

$$Af \leqslant b \quad \Leftrightarrow \quad A\nu \leqslant bn, \qquad f \in F_n, \nu \in \mathbb{N}, \tag{1.1}$$

where the m-column matrix A and vector b are real constants, independent of n. (At this level of generality we think of equalities as represented by pairs of inequalities; more detail is introduced in §2.) Such inequality constraints are quite expressive: as just one example we mention [Tho79], where inequalities are used to represent the limited information/uncertainty concerning oil spill scenarios. Treating the frequencies as reals instead of rationals, we assume that the constraints (1.1) are satisfiable. Then they define a (non-empty) polytope in $\mathbb{R}^m$, and maximizing the entropy

$$H(x) \;=\; -\sum_{1 \leqslant i \leqslant m} x_i \ln x_i \qquad \text{subject to} \quad A'x \leqslant b', \tag{1.2}$$

where A', b' are A, b augmented with $\sum_i x_i = 1$ and $x_i \geqslant 0$, is a strictly concave maximization problem (see e.g. [ADSZ88]) which has a unique solution $\varphi^* \in \mathbb{R}^m$ with maximum entropy H^*. The elements of φ^* are non-negative reals, with values independent of n.

When dealing with the discrete balls-and-boxes problem, some care is required in connection with *equalities* in the constraints (1.1), whether these equalities are explicit or

implied by inequalities. For example, suppose one constraint is $f_1 + f_2 = 1/139$. This constraint is not satisfiable unless n is a multiple of 139, rendering the statement EC above impossible. When n is large however, in many cases it is perfectly acceptable if the equalities are simply satisfied to a good approximation; in fact, the same goes for the inequalities. For this reason we will assume that the constraints (1.1) need to be satisfied *only approximately*, to within a tolerance $\delta > 0$. Besides being necessary for a rigorous development, this tolerance may also be regarded as reflecting some uncertainty in the exact values of A and b.

In addition to introducing a tolerance in the constraints, we will also develop a more intuitive variant of the concentration result EC, around the MaxEnt *vector* φ^* instead of the MaxEnt *value* H^*. So we will be establishing a modified and more general version of statement EC:

EC': given positive tolerances δ, ε, and η or ϑ, as described in Table 1.2, there is a $N(\delta, \varepsilon, \eta)$ or $N(\delta, \varepsilon, \vartheta)$ such that for all $n \geqslant N$, the fraction of assignments that satisfy $\mathcal{C}$ to accuracy δ and have a frequency vector with entropy within η of H^* or no farther than ϑ in norm from φ^*, is at least $1 - \varepsilon$.

A simpler way of saying this is that as the size n of the problem increases, there is a count vector ν^*

1. whose corresponding frequency vector f^* is arbitrarily close to φ^*, and satisfies the constraints to any prescribed accuracy, and

2. out of all the assignments that satisfy the constraints to this accuracy, the fraction that realize a vector as close as desired to f^*, either in entropy or in norm, is arbitrarily near 1.

This way of expressing the concentration result smacks of asymptotics, but we keep the more precise statement EC' in mind.

δ:	relative tolerance in satisfying the constraints
ε:	concentration tolerance, on number of realizations
η:	relative tolerance in deviation from the MaxEnt value H^*
ϑ:	absolute tolerance in deviation from the MaxEnt vector φ^*

Table 1.2: Tolerances for the entropy concentration results.

In addition to EC', we will prove a more forceful variant which refers solely to the realizations of the vector f^* itself, as opposed to those of a whole set of vectors close to it:

EC'': given positive tolerances δ, ε, and η or ϑ, as described in Table 1.2, there is a $N(\delta, \varepsilon, \eta)$ or $N(\delta, \varepsilon, \vartheta)$ such that for all $n \geqslant N$, the vector $f^* \in F_n$ has more than $1/\varepsilon$ times the realizations of the whole set of vectors that satisfy $\mathcal{C}$ to accuracy δ but have entropy not within η of H^* or are farther than ϑ in norm from φ^*.

Summary In §2 we go into more detail on the various tolerances, in particular δ, which relates the exact solution[4] of the continuous maximum entropy problem to the approximate solution of the discrete counting problem. The main idea is that for a given n we obtain an optimal count vector ν^* by *rounding* and *adjusting* the vector $n\varphi^*$. We then show for each of the desired properties how large n must be for ν^* and the frequency vector $f^* = \nu^*/n$ to have this property. These results are put together in §3, where we establish statement EC'. In §3.1 we prove EC' with a tolerance η on the deviation from the maximum entropy *value*, and in §3.2 we discuss how our bound compares with the well-known asymptotic result of E. T. Jaynes. We derive the result EC'' on the vector f^* itself in §3.2.2. In §3.3 we establish EC' and EC'' with the more intuitive tolerance ϑ on the norm of deviation from the maximum entropy *vector*. We use this norm version in §3.4 to point out that our concentration results also apply to the derivation of discrete probability *distributions* by the method of maximum entropy. Thus the exact, non-asymptotic concentration phenomenon is a very powerful justification for the most common use of MaxEnt. The other major justification is various axiomatic formulations, but the simple statements of the concentration results and the purely combinatorial character of their derivation have a force of their own[5]. In §3.4 we also give a further elaboration that provides a *quantitative* justification of the principle of indifference or insufficient reason.

The expressions for $N(\delta, \varepsilon, \eta)$ or $N(\delta, \varepsilon, \vartheta)$ use the solution φ^* to the maximum entropy problem, the value H^* of the maximum entropy, and norms of the matrix A and vector b defining the constraints. A main point of the paper is the computations made possible by the bounds. Therefore in §4 we give three examples with detailed numerical results on the lower bound N, one using the classic die-tossing experiment, one involving a vehicle or network traffic problem, and one having to do with a simple queue. We also discuss the relationship to Sanov's bound (from information theory), and the exact computation of the number of count vectors satisfying the constraints.

The proofs of all of our results are given in the Appendix, so as not to disrupt the flow of the exposition.

2 Basic results: tolerances

We define the rounding of a positive real number x to an integer $[x]$ in the usual way, so that it satisfies $|x - [x]| \leqslant 1/2$. Given an $n \in \mathbb{N}$, from the MaxEnt vector φ^* we derive a count vector ν^* and a frequency vector f^* by a process of *rounding* and *adjusting*:

[4]Sometimes this solution may be analytical, and if it is numerical we assume it is sufficiently accurate to be called "exact".

[5]The maximum entropy method itself is not the subject of this paper. There is an extensive body of work on it: for example the works [Ros83], [Jay03] of E. T. Jaynes, the books [Tri69], [KK92], the series of MaxEnt conference proceedings [ME98] and [ME99], etc.

Definition 2.1 *Given φ^* and $n \geqslant m$, let $\tilde{\nu} = [n\varphi^*]$ and set $d = \sum_i \tilde{\nu}_i - n$. If $d = 0$, let $\nu^* = \tilde{\nu}$. Otherwise, if $d < 0$, add 1 to $|d|$ elements of $\tilde{\nu}$ that were rounded down, and if $d > 0$, subtract 1 from $|d|$ elements that were rounded up. Let the resulting vector be ν^*, and define $f^* = \nu^*/n$, $f^* \in F_n$.*

Unlike φ^*, both of the vectors ν^* and f^* depend on n, but we will not indicate this explicitly to avoid burdensome notation. The adjustment of $\tilde{\nu}$ in Definition 2.1 ensures that the result ν^* indeed sums to n, so f^* is a proper frequency vector. (This adjustment is always possible because if $d \neq 0$, there must be at least $|d|$ elements of $n\varphi^*$ that were rounded to their floors if $d < 0$, or to their ceilings if $d > 0$. And $|d| \leqslant \lfloor m/2 \rfloor$ by the definition of rounding.)

The fundamental observation is that when n is large enough, f^* is arbitrarily close to φ^*:

Proposition 2.1 *Given any $\gamma > 0$, the frequency vector f^* is s.t.*

$$n \geqslant \frac{1}{\gamma} \;\Rightarrow\; \|f^* - \varphi^*\|_\infty \leqslant \gamma, \qquad n \geqslant \frac{3\mu^*}{4\gamma} \;\Rightarrow\; \|f^* - \varphi^*\|_1 \leqslant \gamma,$$

where μ^* is the number of non-zero elements of f^* (and φ^*).

Recall that the ℓ_1 norm is the sum of the absolute values, whereas the ℓ_∞ norm is the maximum of the absolute values ([HJ90], §5.5).

The MaxEnt vector φ^* satisfies the constraints (1.1) exactly. Now we show how f^* satisfies them approximately, and how ν^* satisfies the scaled constraints approximately.

2.1 Constraints on frequency vectors

All constraints on the frequency vectors can be expressed by inequalities as in (1.1). An equality, e.g. $5f_1 + 3f_2 - f_3 = 0.34$, can be formulated as two inequalities, $5f_1 + 3f_2 - f_3 \leqslant 0.34$ and $-(5f_1 + 3f_2 - f_3) \leqslant -0.34$. In practice however, we may, for example, consider equalities to be more important than inequalities, and may want to assign different tolerances to them. Further, if we want to use tolerances that are relative to the magnitudes of the elements of b, the presence of zeros requires special treatment. For these reasons we will separate the constraints (1.1) into four categories: equalities with non-zeros on the r.h.s., inequalities with non-zeros on the r.h.s., equalities with zeros, and inequalities with zeros.

We represent the first category using a matrix $A^=$ and vector $b^=$ as $A^= x = b^=$, where all elements of $b^=$ are non-zero. We want f^* to satisfy the equality constraints with a *maximum error* which is no more than a constant $\delta^= > 0$ times the smallest element of $b^=$ in absolute value. So we require

$$A^= f^* \;=\; b^= + \beta \qquad \text{with} \qquad \|\beta\|_\infty \leqslant\; \delta^= |b^=|_{\min}. \tag{2.1}$$

Similarly, formulating the inequalities with non-zeros as $A^{\leqslant}x \leqslant b^{\leqslant}$, we require

$$A^{\leqslant}f^* \;\leqslant\; b^{\leqslant} + \beta \qquad \text{with} \qquad \|\beta\|_\infty \leqslant\; \delta^{\leqslant}|b^{\leqslant}|_{\min}. \tag{2.2}$$

(The βs in (2.1) and (2.2) are different, but we don't want to complicate the notation.) Coming to the constraints with 0s on the r.h.s., e.g. $f_2 = f_3$, $f_4 \geqslant f_1 + f_5$, etc., we require

$$A^{=0}f^* \;=\; \zeta \qquad \text{with} \qquad \|\zeta\|_\infty \leqslant \delta^{=0} \tag{2.3}$$

and

$$A^{\leqslant 0}f^* \;\leqslant\; \zeta \qquad \text{with} \qquad \|\zeta\|_\infty \leqslant \delta^{\leqslant 0} \tag{2.4}$$

for some positive $\delta^{=0}$ and $\delta^{\leqslant 0}$.

φ^* satisfies all of the constraints exactly. So any $f \in F_n$ close enough to φ^* should satisfy (2.1) to (2.4) for any positive tolerances. Indeed, using the abbreviation $\delta = (\delta^{=}, \delta^{\leqslant}, \delta^{=0}, \delta^{\leqslant 0})$,

Proposition 2.2 *Given any $\delta > 0$, set*

$$\vartheta_\infty \;=\; \min\left(\frac{\delta^{=}|b^{=}|_{\min}}{|||A^{=}|||_\infty}, \frac{\delta^{\leqslant}|b^{\leqslant}|_{\min}}{|||A^{\leqslant}|||_\infty}, \frac{\delta^{=0}}{|||A^{=0}|||_\infty}, \frac{\delta^{\leqslant 0}}{|||A^{\leqslant 0}|||_\infty}\right),$$

or ∞ if there are no constraints. Then any $f \in F_n$ such that $\|f - \varphi^*\|_\infty \;\leqslant\; \vartheta_\infty$ satisfies (2.1), (2.2), (2.3), and (2.4).

Recall that the infinity norm $|||\cdot|||_\infty$ of a matrix is the maximum of the ℓ_1 norms of the rows. By Proposition 2.1, f^* satisfies Proposition 2.2 if $n \geqslant 1/\vartheta_\infty$.

2.2 Entropy

Turning now to entropy, we point out that if a frequency vector f is close enough to φ^*, its entropy can be as close to H^* as desired:

Proposition 2.3 *For any $\gamma > 0$, if f is s.t. $\|f - \varphi^*\|_\infty \leqslant \gamma$, then $H^* - H(f) \leqslant m\gamma \ln 1/\gamma$.*

It follows that

Proposition 2.4 *Given an entropy tolerance $\eta > 0$ and $\eta \leqslant m/(21H^*)$, if f is s.t.*

$$\|f - \varphi^*\|_\infty \;\leqslant\; \frac{2}{3}\frac{\eta H^*}{\ln\left(m/(\eta H^*)\right)},$$

then $(1-\eta)H^* \leqslant H(f) \leqslant H^*$.

(The condition $\eta \leqslant m/(21H^*)$ is not much of a restriction, and is explained in the proof.) In view of Proposition 2.1, f^* will satisfy Proposition 2.4 when n is large enough. Proposition 2.4 is used to establish Lemma 3.2 in §3.1.

3 Concentration

We establish the concentration result EC' stated in the Introduction: in §3.1 we prove the first version, expressed in terms of deviation from the maximum entropy *value* H^*, and in §3.3 we prove the second version, phrased in terms of deviation from the MaxEnt *vector* φ^*. We also establish the statement EC'' in its two versions. To avoid cumbersome notation in what follows, we denote the tolerances on the constraints collectively by $\delta = (\delta^{=}, \delta^{\leqslant}, \delta^{=0}, \delta^{\leqslant 0})$.

3.1 Maximum entropy value

Let $\mathcal{C}(\delta)$ be the set of m-vectors that satisfy the constaints (2.1) to (2.4) to accuracy δ:

$$\mathcal{C}(\delta) = \{x \in \mathbb{R}^m \mid x \text{ satisfies (2.1) to (2.4) with } \delta = (\delta^{=}, \delta^{\leqslant}, \delta^{=0}, \delta^{\leqslant 0})\}. \tag{3.1}$$

Now given an $\eta > 0$, consider the following two sets of frequency vectors[6]. $A_n(\delta, \eta)$ is the set of vectors in F_n that lie in $\mathcal{C}(\delta)$ and have entropy at least $(1-\eta)H^*$:

$$A_n(\delta, \eta) = \{f \in F_n \cap \mathcal{C}(\delta), H(f) \geqslant (1-\eta)H^*\}. \tag{3.2}$$

$B_n(\delta, \eta)$ is the complementary set of frequency vectors, i.e. those in $\mathcal{C}(\delta)$ but with entropy less than $(1-\eta)H^*$:

$$B_n(\delta, \eta) = \{f \in F_n \cap \mathcal{C}(\delta), H(f) < (1-\eta)H^*\}. \tag{3.3}$$

Clearly, $F_n = A_n(\delta, \eta) \cup B_n(\delta, \eta)$ irrespective of the values of δ and η.

The number of realizations $\#f$ of a frequency vector f is related to its entropy $H(f)$. A simple result is Lemmas II.1 and II.2 of [Csi99]

$$\forall f \in F_n, \qquad \frac{1}{\binom{n+m-1}{m-1}} e^{nH(f)} \;\leqslant\; \#f \;\leqslant\; e^{nH(f)}, \tag{3.4}$$

but a much more precise result is

Proposition 3.1 *Given $f \in F_n$, let $f_1, \ldots, f_\mu, \mu \geqslant 1$, be its non-zero elements ($\#f$ does not change when f is permuted). Define*

$$S(f, n) \;=\; \frac{1}{(2\pi n)^{\frac{\mu-1}{2}}} \frac{1}{\sqrt{f_1 \cdots f_\mu}}.$$

Then $\#f$ is bounded as follows:

$$e^{-\frac{1}{12n}\sum_{i=1}^{\mu} 1/f_i} e^{nH(f)} \leqslant \frac{\#f}{S(f,n)} \;\leqslant e^{\frac{1}{12n}} e^{nH(f)}.$$

[6] Our notation is similar to that of [CT91], §12.6. The set A_n is not to be confused with the matrix A of (1.1).

(The bounds hold even when $\mu = 1$ and $\#f = 1$.)

Using the bounds of Proposition 3.1, we will now show that given any $\varepsilon > 0$, there is a number $N = N(\varepsilon)$ s.t. if $n \geqslant N$, then all but a fraction ε of the realizations/assignments that satisfy the constraints have frequencies in the set $A_n(\delta, \eta)$:

$$\frac{\#A_n(\delta,\eta)}{\#A_n(\delta,\eta) + \#B_n(\delta,\eta)} \;=\; \frac{\#A_n(\delta,\eta)}{\#(\,F_n \cap \mathcal{C}(\delta)\,)} \;\geqslant\; 1 - \varepsilon. \tag{3.5}$$

The proof consists of deriving a lower bound on $\#A_n$ and an upper bound on $\#B_n$, taking their ratio, and deriving a lower bound on n so as to ensure that the ratio is at least $1+1/\varepsilon$. It is similar in spirit to the proof of the "conditional limit theorem", Theorem 12.6.2 of [CT91].

First, the upper bound on $\#B_n$. Recall from the beginning of §3 that we are using the abbreviated notation $\delta = (\delta^{=}, \delta^{\leqslant}, \delta^{=0}, \delta^{\leqslant 0})$.

Lemma 3.1 *Given any $\delta, \eta > 0$,*

$$\#B_n(\delta,\eta) \;< 4.004\,\sqrt{2\pi}\,0.6^m\,n^{\frac{m-1}{2}}\,e^{n(1-\eta)H^*}, \tag{3.6}$$

where the numerical constants assume that $n \geqslant 100$.

This bound is independent of δ.

For our lower bound on $\#A_n$ we need an auxiliary lower bound, on the number of frequency vectors that lie in an m-dimensional cube centered at φ^* and of side 2ϑ:

Proposition 3.2 *Let $\mu^* \geqslant 1$ be the number of non-zero elements of φ^*, $\varphi^*_{\max}$ be its largest element, and $\varphi^*_{\min}$ its smallest non-zero element. Let ϑ be a positive number s.t. $\vartheta \leqslant \varphi^*_{\max}$ and $\vartheta \leqslant (\mu^* - 1)\,\varphi^*_{\min}$. Then the set $\{f \in F_n \mid \|f - \varphi^*\|_\infty \leqslant \vartheta\}$ contains at least*

$$\left\lfloor n\vartheta\left(\frac{1}{m-1} + \frac{1}{\mu^*-1}\right)\right\rfloor^{\mu^*-1} \left\lfloor\frac{n\vartheta}{m-1}\right\rfloor^{m-\mu^*} \;=\; \Lambda\,(n,\vartheta,\mu^*)$$

elements. If $\mu^* = 1$, the first factor in this expression and the second condition on ϑ are absent.

The two extreme cases, when all the elements of φ^* are non-zero, and when only one is non-zero, yield, respectively, $\lfloor 2n\vartheta/(m-1)\rfloor^{m-1}$ and $\lfloor n\vartheta/(m-1)\rfloor^{m-1}$. Fig. 3.1 illustrates the difference in the case $m = 2$.

Now let $\alpha \in (0,1)$ be a parameter, and define the number

$$\vartheta_0 \;= \min\left(\vartheta_\infty, \frac{2}{3}\frac{\alpha\eta H^*}{\ln\left(m/(\alpha\eta H^*)\right)}, \varphi^*_{\min}\right) \tag{3.7}$$

where ϑ_∞ has been specified in Proposition 2.2 and $\varphi^*_{\min}$ in Proposition 3.2. ϑ_0 depends on δ, η and α, which is specified in Theorem 3.1 below. Our lower bound on $\#A_n$, the number of realizations of A_n, is based on a lower bound on the size of A_n, obtained by using the tolerance ϑ_0 in Proposition 3.2:

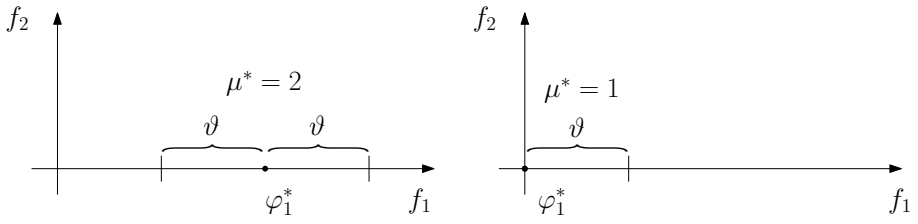


Figure 3.1: Illustration of Proposition 3.2 in two dimensions when $\mu^* = 2$ and when $\mu^* = 1$.

Lemma 3.2 *Given any $\delta, \eta > 0$, and some $\alpha \in (0,1)$, we have*

$$|A_n(\delta,\eta)| \;\geqslant\; \Lambda\left(n, \vartheta_0, \mu^*\right)$$

and

$$\#A_n(\delta,\eta) \;\geqslant \Lambda\left(n, \vartheta_0, \mu^*\right) \sqrt{2\pi} \left(\frac{\mu^*}{2\pi}\right)^{\mu^*/2} e^{-\frac{\mu^*}{12}} n^{-\frac{\mu^*-1}{2}} e^{n(1-\alpha\eta)H^*}.$$

Typically $\mu^* = m$, and then $n \geqslant (m-1)/(2\vartheta_0)$ is necessary for $A_n(\delta,\eta)$ to not be empty and for $\#A_n(\delta,\eta)$ to be at least 1. Otherwise we need $n \geqslant (m-1)/\vartheta_0$.

The main result following from Lemmas 3.1 and 3.2 depends on many parameters and we have taken some care to reduce the slack in the bounds, so it reads more like the specification of an algorithm rather than a theorem:

Theorem 3.1 *Given any $\delta, \varepsilon, \eta > 0$, let $\alpha \in (0,1)$ be a parameter whose value is specified below. With μ^* the number of non-zero elements of φ^*, define the constants*

$$C_1 = \frac{0.5(m+\mu^*) - 1}{(1-\alpha)\eta H^*}, \qquad C_2 = \frac{m \ln 0.6 + (0.5 \ln 2\pi + 1/12 - 0.5 \ln \mu^*)\mu^* + \ln\left(\frac{1/\varepsilon+1}{0.249}\right)}{(1-\alpha)\eta H^*},$$

and set

$$N(\alpha) \;=\; \begin{cases} 1.5\, C_1 \ln(C_1 + C_2) + C_2, & \text{if } C_2 > 0 \text{ and } C_1 + C_2 \geqslant 21, \\ 1.5\, C_1 \ln C_1 + C_2, & \text{if } C_2 \leqslant 0. \end{cases}$$

Let $\hat{\alpha} \in (0,1)$ be the solution of the equation

$$N(\alpha) = \frac{m-1}{[\![2, \mu^* = m]\!]\, \vartheta_0(\alpha)},$$

where the notation $[\![x, B]\!]$ yields x if boolean condition B holds and 1 otherwise, and ϑ_0 is given by (3.7). Finally set

$$N \;=\; N(\hat{\alpha}) \;=\; \frac{m-1}{[\![2, \mu^* = m]\!]} \max\left(\frac{3 \ln\left(m/(\hat{\alpha}\eta H^*)\right)}{2\hat{\alpha}\eta H^*}, \frac{1}{\vartheta_\infty}, \frac{1}{\varphi^*_{\min}}\right),$$

where ϑ_∞ is specified in Proposition 2.2 and $\varphi^*_{\min}$ in Proposition 3.2. Then for all $n \geqslant N$ we have

$$\frac{\#\{f \in F_n \cap \mathcal{C}(\delta), H(f) \geqslant (1-\eta)H^*\}}{\#\{f \in F_n \cap \mathcal{C}(\delta)\}} \;\geqslant\; 1-\varepsilon.$$

Given any tolerances $\delta, \varepsilon, \eta$, the theorem shows how to calculate a number $N(\delta, \varepsilon, \eta)$ s.t. if $n \geqslant N(\delta, \varepsilon, \eta)$, then all but the fraction ε of the assignments of the n objects to the m boxes that satisfy the constraints to accuracy δ have entropy within $1-\eta$ of the maximum. An analogue of this result, but phrased in terms of a deviation ϑ from the maximum entropy vector φ^*, is given in §3.3.

3.2 Discussion

There are a few things to note in connection with Theorem 3.1:

1. The first term in the expression for N depends on the adjusted deviation $\Delta H = \hat{\alpha}\eta H^*$ from the value of the maximum entropy, the second depends on the tolerances δ for satisfying the constraints (§2.1), and the third depends on the smallest non-zero element of the solution φ^* to the maximum entropy problem. ε is hiding in the value of $\hat{\alpha}$, see 5 below.

2. If the constraints (1.1) do not force any element of φ^* to be 0, we will simply have $\mu^* = m$.

3. Roughly speaking, N is at least m/ϑ_∞, at least $m/\varphi^*_{\min}$, and at least $(m/\Delta H)\ln(m/\Delta H)$ as well.

4. By examining the expressions for C_1 and C_2 it is clear that the value of $N(\alpha)$ is sensitive to η, but not very sensitive to ε. This carries over to N, and is illustrated numerically in §4.

5. The l.h.s. of the equation determining the parameter α depends on ε and η, whereas the r.h.s. depends only on δ. The optimal value $\hat{\alpha}$ depends weakly on ε, and is essentially a function of η. This is illustrated in §4.2.

6. Finally, the assumptions $n \geqslant 100$ in Lemma 3.1 and $C_1 + C_2 \geqslant 21$ in Theorem 3.1 are very easily satisfied in applications.

3.2.1 Comparison with the results of Jaynes

We compare Theorem 3.1 with the original concentration theorem of E.T. Jaynes:

Theorem 3.2 *([Jay82], or [Jay83], Ch. 11) Suppose the constraints consist of ℓ linearly-independent equalities, and set $\Delta H = H^* - H$. Then, as $n \to \infty$, $2N\Delta H = \chi^2_{m-\ell-1}(\varepsilon)$, where χ^2_k is the chi-squared distribution with k degrees of freedom.*

This says that

$$\varepsilon \;\sim\; \frac{1}{\Gamma(s+1)} \int_{n\Delta H}^{\infty} e^{-x} x^s ds, \tag{3.8}$$

where $s = (m-\ell-1)/2 - 1$ and the r.h.s. represents the tail of the chi-squared density. This tail is the normalized incomplete gamma function, with the asymptotic expansion $\frac{1}{\Gamma(s+1)} e^{-n\Delta H} (n\Delta H)^s \left(1 + \frac{s}{n\Delta H} + \cdots\right)$ (see, e.g. [AS72], eq. 6.5.32) . Thus ignoring ℓ (but see §3.2.3) , and retaining only the first term of the above series, it can be seen that (3.8) requires $n\Delta H \geqslant s \ln(n\Delta H) - \ln(\varepsilon\Gamma(s+1))$. This translates to $n \geqslant C_1 \ln n + C_2$, where the constants are

$$C_1 = \frac{s}{\Delta H}, \qquad C_2 = \frac{1}{\Delta H} \left(s \ln \Delta H - \ln(\varepsilon\Gamma(s+1))\right), \qquad s = \frac{m-3}{2}.$$

Comparing this with the N_1 of Theorem 3.1, we see that there is qualitative agreement between our exact bound and Jaynes's asymptotic result, and the C_1's are similar. (In fact, it follows from Lemma 3.2 and Proposition 3.2, that *asymptotically* our C_1 is better.)

3.2.2 The MaxEnt vector itself

It may seem in some sense unsatisfactory that Theorem 3.1 says that an *entire set* of vectors around the MaxEnt vector φ^* is dominant. Indeed, using elements in the proof of Theorem 3.1 it is possible to say something about just f^* itself. The result can be stated more simply than Theorem 3.1, holds for smaller n, and even shows that f^* is closer than η to φ^* in entropy:

Lemma 3.3 *Given any $\delta, \varepsilon, \eta > 0$, let $\hat{\alpha} \in (0,1)$ be the solution of $N(\alpha) = 1/\vartheta_0(\alpha)$, with $N(\alpha), \vartheta_0(\alpha)$ as in Theorem 3.1. Then if*

$$n \;\geqslant\; \max\left(\frac{3\ln\left(m/(\hat{\alpha}\eta H^*)\right)}{2\hat{\alpha}\eta H^*}, \frac{1}{\vartheta_\infty}, \frac{1}{\varphi^*_{\min}}\right),$$

the frequency vector f^* is such that

$$f^* \in A_n(\delta, \hat{\alpha}\eta) \qquad \text{and} \qquad \frac{\#f^*}{\#\{f \in F_n \cap \mathcal{C}(\delta), H(f) < (1-\eta)H^*\}} \;>\; \frac{1}{\varepsilon}.$$

This is the statement EC'' in the Introduction. In simple terms, it says that for n about m times smaller than what Theorem 3.1 requires, the MaxEnt frequency vector f^*, whose entropy differs from H^* by less than $\hat{\alpha}\eta$, has all by itself $1/\varepsilon$ times as many realizations as the entire set of vectors that satisfy the constraints but have entropies not within η of H^*. See Fig. 3.2.

The closeness of the excluded vectors to f^* is controlled by η and can be made as tight as desired. Nevertheless, we cannot exclude *everything* around f^*: the simplest counter-example is n even, $m = 2$, and no constraints; then $\binom{n}{n/2} / \binom{n}{n/2\pm1} \to 1$ as n increases.

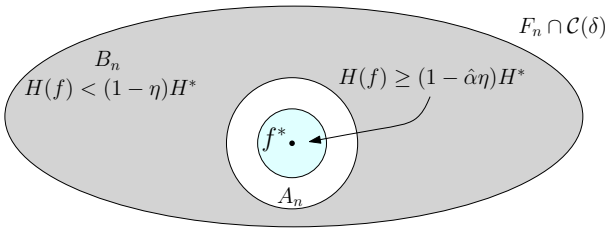


Figure 3.2: The sets $A_n(\delta,\eta), B_n(\delta,\eta)$, and $A_n(\delta,\hat{\alpha}\eta)$ for Lemma 3.3.

3.2.3 Improvements

In perusing the various results and their proofs with an eye toward improvements, we notice that if the constraints (1.1) include (linearly-independent) equalities, say ℓ of them, this would reduce the dimension m of of F_n by ℓ. Our results could then be re-worked using a notation such as $F_{n,m-\ell}$, which makes the dimension explicit. We will not pursue this improvement further here. Another possibility is to improve the bound of Proposition 2.4 as noted in its proof. A shortcoming in the results on which Theorem 3.1 is based is that the bound on $\#A_n$ is sensitive to δ but the bound on $\#B_n$ is not.

3.3 Maximum entropy vector

The development in §3.1 used a tolerance η in deviation from the maximum entropy *value* H^*. Here we re-cast this development in terms of a more intuitive tolerance ϑ in deviation from the maximum entropy *vector* φ^*. This formulation will be very useful when we deal with probability distributions in §3.4. So, given a $\vartheta > 0$, we re-define the sets A_n and B_n of (3.2) and (3.3) as

$$A'_n(\delta,\vartheta) = \{f \in F_n \cap \mathcal{C}(\delta), \|f-\varphi^*\|_1 \leqslant \vartheta\},\ B'_n(\delta,\vartheta) = \{f \in F_n \cap \mathcal{C}(\delta), \|f-\varphi^*\|_1 > \vartheta\}. \quad (3.9)$$

These sets form a partition of $F_n \cap \mathcal{C}(\delta)$ for any δ and any suitably small ϑ, as was the case in §3.1.

To count the realizations of these sets we need a connection between differences in norm and differences in entropy. If f is close to φ^* in norm, its entropy is close to H^*, and if it is far from φ^*, its entropy cannot be too close to H^*:

Proposition 3.3 *Given $0 < \vartheta \leqslant 1/2$, then*

$$\begin{aligned} \|f-\varphi^*\|_1 \leqslant \vartheta \quad &\Rightarrow \quad H(f) \geqslant H^* - \vartheta \ln(m/\vartheta), \\ \|f-\varphi^*\|_1 > \vartheta \quad &\Rightarrow \quad H(f) < H^* - \vartheta^2/2. \end{aligned}$$

Now with the definitions (3.9) we have analogues of the results of §3.1, beginning with an analogue of Lemma 3.1:

Lemma 3.4 *Given any $\delta, \vartheta > 0$,*

$$\#B'_n(\delta,\vartheta) \ < 4.004\,\sqrt{2\pi}\,0.6^m\,n^{\frac{m-1}{2}}\,e^{n(H^*-\vartheta^2/2)}, \tag{3.10}$$

where the numerical constants assume that $n \geqslant 100$.

Next is an analogue of (3.7): we define

$$\vartheta'_0 \ = \ \min\left(\vartheta_\infty, \alpha\vartheta, \varphi^*_{\min}\right), \tag{3.11}$$

where $\alpha \in (0,1)$ is a parameter on which we elaborate in Proposition 3.4 and Theorem 3.3 below. Finally, an analogue of Lemma 3.2, for the set A'_n:

Lemma 3.5 *Given any $\delta, \vartheta > 0$, and some $\alpha \in (0,1)$, we have*

$$|A'_n(\delta,\vartheta)| \geqslant \Lambda\left(n, \vartheta'_0, \mu^*\right)$$

with $\Lambda(\cdot)$ defined in Proposition 3.2, and

$$\#A'_n(\delta,\vartheta) \ \geqslant \Lambda\left(n,\vartheta'_0,\mu^*\right)\sqrt{2\pi}\left(\frac{\mu^*}{2\pi}\right)^{\mu^*/2} e^{-\frac{\mu^*}{12}} n^{-\frac{\mu^*-1}{2}} e^{n(H^*-h(\alpha\vartheta))},$$

where $h(\vartheta) = \vartheta\ln(m/\vartheta)$.

From these two lemmas, the ratio $\#A'_n/\#B'_n$ is bounded from below by an exponential in n of the form $e^{n\psi(\alpha,\vartheta)}$, divided by a polynomial in n. The coefficient of n in the exponential, the analogue of the ΔH of §3.2, is critical:

Proposition 3.4 *Consider the function*

$$\psi(\alpha,\vartheta) = \frac{1}{2}\vartheta^2 - \alpha\vartheta\ln\frac{m}{\alpha\vartheta}, \qquad \alpha,\vartheta \in (0,1), \quad m < \frac{1}{2}\vartheta^3 e^{1/\vartheta}.$$

For fixed ϑ, $\psi(\alpha,\vartheta)$ is positive for $\alpha \leqslant \vartheta^2/2$ and increases as α decreases. The equation $\psi(\alpha,\vartheta) = 0$ has a root $\alpha_0 \in (\vartheta^2/2, 1)$.

(The condition on m does not impose a significant restriction in practice; even for ϑ as large as 0.04, it requires $m \leqslant 2.3 \cdot 10^6$.)

With the above, we finally have the analogue of Theorem 3.1. Again, the statement is much more like that of an algorithm for computing N and the main feature is that H^* does not appear anywhere:

Theorem 3.3 *Given any $\delta, \varepsilon > 0$ and $0 < \vartheta < 1/2$, assume that $m < 1/2\vartheta^3 e^{1/\vartheta}$. Define the constants*

$$C_1 = \frac{0.5(m+\mu^*)-1}{\psi(\alpha,\vartheta)}, \qquad C_2 = \frac{m\ln 0.6 + (0.5\ln 2\pi + 1/12 - 0.5\ln\mu^*)\mu^* + \ln\left(\frac{1/\varepsilon+1}{0.249}\right)}{\psi(\alpha,\vartheta)}.$$

The numerators are the same as in Theorem 3.1, and the function $\psi(\alpha,\vartheta)$ is defined in Proposition 3.4. Set

$$N(\alpha) \;=\; \begin{cases} 1.5\,C_1\ln(C_1+C_2)+C_2, & \text{if } C_2 > 0 \text{ and } C_1+C_2 \geqslant 21, \\ 1.5\,C_1\ln C_1 + C_2, & \text{if } C_2 \leqslant 0, \end{cases}$$

as in Theorem 3.1. Let $\alpha_0 \in (\vartheta^2/2, 1)$ be the root of $\psi(\alpha,\vartheta) = 0$, and $\hat{\alpha} \in (0, \alpha_0)$ be the solution of the equation

$$N(\alpha) = \frac{m-1}{[\![2, \mu^* = m]\!]\,\vartheta_0'(\alpha)},$$

where the notation $[\![.,.]\!]$ was defined in Theorem 3.1 and ϑ_0' is given by (3.11). Finally set

$$N \;=\; N(\hat{\alpha}) \;=\; \frac{m-1}{[\![2, \mu^* = m]\!]\min\left(\hat{\alpha}\vartheta, \vartheta_\infty, \varphi^*_{\min}\right)},$$

where ϑ_∞ has been specified in Proposition 2.2 and $\varphi^*_{\min}$ in Proposition 3.2. Then for all $n \geqslant N$ we have

$$\frac{\#\{f \in F_n \cap \mathcal{C}(\delta), \|f-\varphi^*\|_1 \leqslant \vartheta\}}{\#\{f \in F_n \cap \mathcal{C}(\delta)\}} \;\geqslant\; 1-\varepsilon.$$

This is the desired result, using deviation in norm from the MaxEnt vector φ^*, instead of difference in entropy from H^*. It says that the set of frequency vectors that are within ϑ of φ^* in ℓ_1 norm has all but the fraction ε of the realizations that satisfy the constraints to the prescribed accuracy δ.

Comments similar to those made on Theorem 3.1 apply here also, and in addition we have a mild restriction on m. Further, recalling the traditional view of entropy concentration in Fig. 1.1, because the norm is a more intuitive measure of closeness, Theorem 3.3 in effect says that concentration occurs *earlier*, at the "vector", instead of the "entropy" stage.

We also have the analogue of Lemma 3.3 on the MaxEnt vector f^* itself. Again, its statement is simpler than that of Theorem 3.3, it holds for somewhat smaller n when $\mu^* < m$, and in fact it establishes that f^* is closer than ϑ to φ^* in norm:

Lemma 3.6 *Given any $\delta, \varepsilon > 0$ and $0 < \vartheta < 1/2$, let $m < 1/2\vartheta^3 e^{1/\vartheta}$. Let $\hat{\alpha} \in (0,1)$ be the solution of $N(\alpha) = 3\mu^*/(4\vartheta_0'(\alpha))$, with $N(\alpha)$ and $\vartheta_0'(\alpha)$ as in Theorem 3.3. Then if*

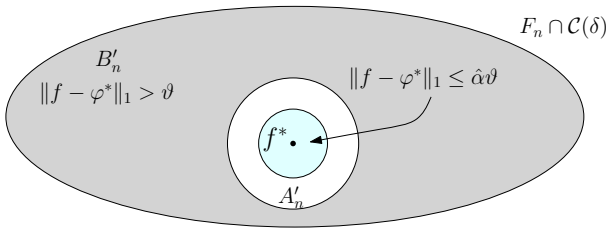
$$n \;\geqslant\; \frac{3}{4}\mu^* \max\left(\frac{1}{\hat{\alpha}\vartheta}, \frac{1}{\vartheta_\infty}, \frac{1}{\varphi^*_{\min}}\right),$$

the frequency vector f^* is s.t.

$$f^* \in A'_n(\delta, \hat{\alpha}\vartheta) \qquad \text{and} \qquad \frac{\#f^*}{\#\{f \in F_n \cap \mathcal{C}(\delta), \|f - \varphi^*\|_1 > \vartheta\}} \; > \; \frac{1}{\varepsilon}.$$

We can paraphrase this as (recall EC'' in the Introduction)

> The MaxEnt frequency vector f^*, which is no farther than $\hat{\alpha}\vartheta$ in ℓ_1 norm from φ^*, has $1/\varepsilon$ times as many realizations as the entire set of vectors that satisfy the constraints to the prescribed accuracies but differ from φ^* by more than ϑ in ℓ_1 norm:



As far as the *number* of excluded vectors goes, i.e. the size of the set $A'_n(\delta, \vartheta)$, we have Lemma 3.5. This points out that even though we can make the tolerance ϑ as small as desired, the number of excluded vectors around f^* in Lemma 3.6 does not necessarily become small.

Finally, we note that the result of Theorem 3.3 might be improved by tightening the bounds of Proposition 3.3 in the ways indicated in its proof.

3.4 Maximum entropy probability distributions

The maximum entropy method, MaxEnt, is most commonly presented as the solution to the problem of inferring a unique probability distribution from limited information (constraints). A very appealing construction of such distributions in the discrete case is the "Wallis derivation", given by Jaynes in [Jay03], §11.4. The main idea is that the n units to be allocated to the m boxes are thought of as *probability quanta*, each of size $1/n$. These quanta are used to construct rational approximations to an m-vector with real entries. When n is large, the result is the most likely (greatest number of realizations) *probability distribution* that satisfies the constraints, which we have denoted φ^*. The norm formulation of entropy concentration in §3.3 lends itself perfectly to obtaining non-asymptotic bounds in this situation.

To emphasize that here we are viewing vectors in F_n as discrete probability distributions, we use p in place of f and P_n in place of F_n. Thus Lemma 3.6 becomes

Corollary 3.1 *Given any $\delta, \varepsilon > 0$ and $0 < \vartheta < 1/2$, and the* MaxEnt *vector φ^*, if $m < 1/2\vartheta^3 e^{1/\vartheta}$ and*

$$n \geqslant \frac{3}{4}\mu^* \max\left(\frac{1}{\hat{\alpha}\vartheta}, \frac{1}{\vartheta_\infty}, \frac{1}{\varphi^*_{\min}}\right)$$

where $\hat{\alpha}$ is as in Lemma 3.6, the discrete p.d. p^* with rational elements obtained from ν^* as specified in Definition 2.1 is s.t.

$$p^* \in A'_n(\delta, \hat{\alpha}\vartheta) \qquad \text{and} \qquad \frac{\#p^*}{\#\{p \in P_n \cap \mathcal{C}(\delta), \|p - \varphi^*\|_1 > \vartheta\}} > \frac{1}{\varepsilon}.$$

Corollary 3.1 increases the applicability of our concentration results significantly, as the principal use of the MaxEnt method is to infer probability distributions[7]. The lower bound on n is simple: the first term depends on the desired tolerance ϑ and concentration factor ε, the second just on the desired accuracies δ, and the third on the maximum entropy solution φ^*. We illustrate this result in §4.3 using the probability distribution of the length of a queue.

We mentioned Laplace's principle of indifference in the Introduction. Corollary 3.1 allows us to derive a *quantified* justification of this principle from entropy concentration:

Corollary 3.2 *Given any $\varepsilon > 0$ and $0 < \vartheta \leqslant \min(0.09, 1/m)$, let $\hat{\alpha} \in (0, 1)$ be the solution of the equation*

$$N(\alpha) = \frac{3m}{4\alpha\vartheta},$$

where $N(\alpha)$ is defined in Theorem 3.3. Let $u^* \in P_n$ be the rational p.d. obtained from the uniform MaxEnt *vector $(1/m, \ldots, 1/m)$ according to Definition 2.1. Then for any $n \geqslant N(\hat{\alpha})$, u^* is s.t.*

$$\frac{\#u^*}{\#\{p \in P_n, \|p - (1/m, \ldots, 1/m)\|_1 > \vartheta\}} > \frac{1}{\varepsilon} \quad \text{and} \quad \|u^* - (1/m, \ldots, 1/m)\|_1 \leqslant \hat{\alpha}\vartheta.$$

Consider constructing a discrete m-element p.d. from quanta of size $1/n$ in the absence of any constraints at all on the frequencies, that is in the situation of complete ignorance, apart from the fact that there are m mutually-exclusive possibilities. Corollary 3.2 says that if $n \geqslant N(\hat{\alpha})$, there is a dominant p.d. $u^* \in P_n$ which has $1/\varepsilon$ times more realizations than the entrire set of p.d.'s in P_n which differ from $(1/m, \ldots, 1/m)$ by more than ϑ in ℓ_1 norm. Further, this dominant p.d. is *uniform* to within $\hat{\alpha}\vartheta$ in ℓ_1 norm.

For example, with $n \geqslant 1232818$, the p.d. u^* has at least 10^8 times as as many realizations as the entire set of p.d.'s that differ from $(0.5, 0.5)$ by more than 0.01 in ℓ_1 norm; further, u^* is no farther than $1.22 \cdot 10^{-6}$ in ℓ_1 norm from $(0.5, 0.5)$.

[7]The application to probability distributions invites comparison with the concentration of measure results in [DP09].

4 Examples

We begin with a simple example of die tosses, and then give an example involving a traffic problem and an example having to do with the probability distribution of the size of a queue. In the first two examples the units (balls) out of which we construct the composite object are clearly distinguishable, but we do not make any use of their distinguishing characteristics. In the last example, one would be hard pressed to say that the units can be distinguished in any way.

The examples follow this recipe:

1. Formulate the problem with constraints on frequencies, treating them as continuous (real numbers).
2. Solve to find the MaxEnt vector $\varphi^* \in \mathbb{R}^m$ and its entropy H^*.
3. Define tolerances δ, linking the continuous problem to the discrete allocation problem.
4. Choose ε and η or ϑ to calculate N.
5. For any $n \geqslant N$, construct the integral count vector ν^* by rounding and adjusting $n\varphi^*$, and the rational frequency vector f^* as ν^*/n. If we are talking about discrete p.d.'s p, as in §3.4, interpret f^* as p^*.

4.1 Die tosses

We use E.T. Jaynes's classic example of tossing a die (see [Jay82] or [Jay83], Ch. 11) to illustrate the parameters δ, ε, and η appearing in Theorem 3.1, to compare with the results of [Jay82] which use the asymptotic chi-squared approximation (Theorem 3.2), and to relate entropy concentration to Sanov's theorem in information theory.

Jaynes considers 1000 tosses of a die in two situations: first, no other information at all is known (including fairness or biasedness of the die), and second, it is also known that the average of the 1000 tosses is 4.5. What can be said in each case about the number of times that each face occurred?

4.1.1 Entropy concentration

The die tosses can be thought of as assignments of $n = 1000$ balls to $m = 6$ boxes. In the first case the MaxEnt solution is $\varphi^* = (1/6, \ldots, 1/6)$ with $H^* = \ln 6 = 1.7918$. No element of φ^* is 0, so we have $\mu^* = m = 6$. For this example Jaynes uses two values of ε, 0.05 and 0.005. From Theorem 3.2, these imply entropy deviations $2000\Delta H = \chi^2_5(0.05) = 11.07$ and $2000\Delta H = \chi^2_5(0.005) = 16.75$, which translate to $\eta = 0.00309$ and $\eta = 0.00467$ in our formulation. Part (a) of Table 4.1 lists the N and $\hat{\alpha}$ of Theorem 3.1 for $\eta = 0.00309$ and various ε, starting from 0.05. Part (b) does the same for $\eta = 0.00467$.

η	ε	$\hat{\alpha}$	N
0.00309	0.05	0.340	**16071**
	0.005	0.326	16858
	$5 \cdot 10^{-6}$	0.295	18866
	$5 \cdot 10^{-12}$	0.255	22223
	$5 \cdot 10^{-18}$	0.227	25261
	$5 \cdot 10^{-36}$	0.175	33739

(a)

η	ε	$\hat{\alpha}$	N
0.00467	0.05	0.340	10065
	0.005	0.326	**10585**
	$5 \cdot 10^{-6}$	0.293	11913
	$5 \cdot 10^{-12}$	0.252	14132
	$5 \cdot 10^{-18}$	0.224	16141
	$5 \cdot 10^{-36}$	0.172	21747

(b)

η	ε	$\hat{\alpha}$	N
0.0067	0.01	0.330	6945
	0.0001	0.304	**7597**
	10^{-8}	0.270	8704
	10^{-16}	0.226	10633
	10^{-32}	0.176	14144
	10^{-64}	0.125	20771

(c)

Table 4.1: The N of Theorem 3.1 for Jaynes' die tosses. (a): no other information and $\eta = 0.00309$, (b): no other information and $\eta = 0.00467$, (c): mean of 4.5 and $\eta = 0.0067$, $\delta^{=} = 0.00467$.

In the second case, when the mean of the 1000 tosses is also known, the information is expressed by $A^{=} = (1, 2, 3, 4, 5, 6)$, $b^{=} = (4.5)$, and we have $\varphi^* = (0.0543, 0.0788, 0.1142, 0.1654, 0.2398, 0.3475)$ with $H^* = 1.61358$. Now $|||A^{=}|||_\infty = 21, |b^{=}|_{\min} = 4.5$, and by Proposition 2.1 the achievable tolerance $\delta^{=}$ is 0.00467. Here Jaynes takes $\varepsilon = 0.0001$, leading to $2000\Delta H = 0.012$ and $\eta = 0.0067$. Part (c) of Table 4.1 lists N under these conditions.

To interpret the third row of Table 4.1(c), for example, keep in mind that there are $6^{8704} \approx 10^{6773}$ possible sequences of 8704 tosses, and $\binom{8704+5}{5} \approx 4.2 \cdot 10^{17}$ possible frequency/count vectors, of which about 1 in 44000 has average equal to 4.5 (see §4.1.3). Then this row of the table says that

> Out of all the possible sequences of 8704 or more tosses whose frequency vectors satisfy the equality constraint (mean) to relative accuracy 0.00467, at most one in 10^8 has frequencies/counts with entropy less than 99.33% of the maximum.

We see from Table 4.1 that

- The asymptotic χ^2 result and our exact bound are quite far apart: for all of the bold entries in the table, the χ^2 result is 1000. Only a small part of this difference can be attributed to our ignoring equality constraints in the dimension of F_n, recall §3.2.3.
- N is very insensitive to ε: in the whole table, ε decreases by 30 orders of magnitude before N so much as doubles.

- The effect of optimizing the parameter α appearing in Theorem 3.1 can be significant, as illustrated in Fig. 4.1.

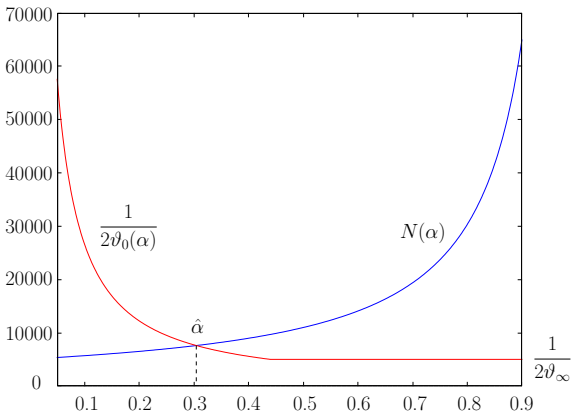


Figure 4.1: The quantities $1/(2\vartheta_0(\alpha))$ and $N(\alpha)$ of Theorem 3.1 vs. α in the case $N = 7597$ of Table 4.1(c).

4.1.2 Sanov bound

Sanov's theorem ([CT91], Theorem 12.4.1) bounds the probability of a set of n-sequences in terms of the distribution with minimum cross-entropy (relative entropy) in this set, assuming that the p.d. generating the sequences is known[8]. The theorem involves sets of sequences and maximum entropy, so it is useful to understand how it relates to the entropy concentration results. The Sanov theorem is usually expressed in the terminology of the "theory of types", see Table 4.2. Stated in our terminology,

Theorem. (Sanov, equiprobable version) *If $\mathcal{C}$ is a subset of $\mathbb{R}^m$ and all n-sequences of m symbols are equiprobable, then*

$$\Pr(\mathcal{C} \cap F_n) \;\leqslant\; \frac{1}{m^n}\binom{n+m-1}{n} e^{nH^*}$$

where H^* is the entropy of the maximum entropy distribution in $\mathcal{C}$.

[The proof of this version of the theorem is simple: using Table 4.2 to translate probability to #, what is to be proved reduces to $\#(\mathcal{C} \cap F_n) \leqslant \binom{n+m-1}{n} e^{nH^*}$. But $\binom{n+m-1}{n} = |F_n|$ is an upper bound on $|\mathcal{C} \cap F_n|$, and then we use the bound of (3.4).]

First we give a numerical example. Take the set $\mathcal{C}$ defined by $f_i \geqslant 0, \sum_i f_i = 1, \sum_{i=1}^6 i f_i = 4.5$ which we considered in §4.1.1. Then $H^* = 1.61358$, so with $n = 9542$ the theorem gives

[8]The perhaps more familiar Chernoff bound follows from Sanov's theorem. See e.g. [DP09] where the Chernoff bound is expressed in terms of relative entropy.

type	frequency vector
type class	set of realizations of a type
size of type class C	$\#C$
probability of class C under uniform p.d., $1/m^n$	$\#C/m^n$

Table 4.2: Information theory (theory of types) terms in [CT91] and [Csi99] on the left, and our terms on the right.

$7.88 \cdot 10^{-714}$ as an upper bound for the probability of $\mathcal{C} \cap F_{9542}$. Therefore, by the last row of Table 4.2, $\#(\mathcal{C} \cap F_{9542}) \leqslant 1.04 \cdot 10^{6712}$. This is a big number, but $6^{9542} \approx 10^{7425}$ is much bigger still, leading to the small probability.

Recalling §3.1, we see that the Sanov result is an *upper bound* on the number of realizations of the sequences in the set A_n defined in (3.2). Lemma 3.2 is a lower bound on this number, and Theorem 3.1 is a lower bound on the *ratio* of this number to the *complementary* number $\#B_n$. (See (A.10) in the proof of the theorem in the Appendix.) In the Sanov bound, the set A_n is interpreted as a set of "bad" or undesirable sequences whose probability we want to limit. On the contrary, in the entropy concentration results, A_n is viewed as the "good" set of interest, whose dominance we want to demonstrate, whereas B_n is the undesirable set. The concentration results then show not only that the good set A_n has a lot of realizations (Lemma 3.2), but that in fact its realizations dominate those of the bad set B_n. In other words, the concentration result answers the question

> If we adopt the set A_n, or even the vector f^* itself, as our prediction or estimate in the face of the limited information, *how reliable* is this prediction? What about these other possible frequency vectors that also accord with the given information?

4.1.3 The exact number of frequency vectors satisfying the constraints

The n tosses of the die with average known to equal 4.5 are characterized by a count vector $\nu = (\nu_1, \ldots, \nu_6)$ s.t. $\nu_i \geqslant 0$, $\nu_1 + \cdots + \nu_6 = n$, and $2(1\nu_1 + 2\nu_2 + \cdots + 6\nu_6) = 9n$. These constraints define a polytope $\mathcal{C}$ in $\mathbb{N}^6$ which depends on n. Using the theory and algorithms in [VWBC05] for counting lattice points in parametric polytopes, it is possible to compute the exact number of lattice points in this polytope, i.e. the number of vectors $\nu \in \mathbb{N}^6$ satisfying the constraints (exactly), as a function of n. The result is a long expression, polynomial in floors of sub-expressions linear in n; a much simpler slight approximation (see [BV08] for the details) is the ordinary polynomial

$$|\mathcal{C} \cap F_n| \;= \frac{19n^4}{11520} + \frac{n^3}{32} + \frac{113n^2}{576} + \frac{2101723n}{4196000} + \frac{225740219}{755280000}.$$

There is some distance between the easy upper bound $|\mathcal{C} \cap F_n| \leqslant |F_n| = \binom{n+5}{n}$ and this exact result: for $n = 1000$ the bound is $8.46 \cdot 10^{12}$, whereas the exact result is $1.680752 \cdot 10^9$, quite

a bit smaller. (But if we reduce m to 4, reflecting the correct dimension of F_n, recall §3.2.3, the bound improves to $4.2 \cdot 10^{10}$.) For $n = 8704$, the above exact result is $9.48684 \cdot 10^{12}$ and the bound is $4.2 \cdot 10^{17}$.

4.2 Vehicle or network traffic

Five cities are connected by two highways as shown in Fig. 4.2. The total number of cars in the 5 cities is known. From measurements made on one day, the number of cars that left

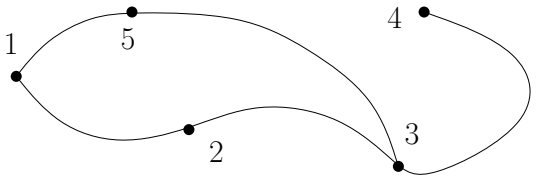


Figure 4.2: Five cities connected by two highways.

cities 1, 2, and 4 is known. The number of cars that travelled the highway segment from 2 to 3 is also known, and finally it is known that at least a certain number travelled the 5 to 3 segment (this segment was observed for only part of the day). Given this information, what is the most likely number (fraction) of cars that travelled between each pair of cities on that day? This is the 5×5 matrix $C = [c_{ij}]$; c_{ii} represents the fraction of cars that left city i and returned to it. (Clearly, instead of cars, we could be considering packets or other units of traffic in a communications network.) Our information is

$$\sum_{1 \leqslant j \leqslant 5} c_{ij} = r_i, \quad i = 1, 2, 4, \qquad c_{13} + c_{14} + c_{23} + c_{24} = s_{23}, \quad c_{13} + c_{53} + c_{14} \geqslant s_{53},$$

where the r_i and s_{ij} are also fractions of the total. Suppose that $(r_1, r_2, r_4) = (0.13, 0.25, 0.1)$, and $(s_{23}, s_{53}) = (0.11, 0.07)$. Then the MaxEnt vector $\varphi^* = (c^*_{11}, \ldots, c^*_{15}, c^*_{21}, \ldots, c^*_{25}, \ldots)$ arranged in matrix form is

$$\begin{pmatrix} 0.030790 & 0.030790 & 0.018816 & 0.018816 & 0.030790 \\ 0.059210 & 0.059210 & 0.036184 & 0.036184 & 0.059210 \\ 0.052 & 0.052 & 0.052 & 0.052 & 0.052 \\ 0.02 & 0.02 & 0.02 & 0.02 & 0.02 \\ 0.052 & 0.052 & 0.052 & 0.052 & 0.052 \end{pmatrix},$$

with $H^* = 3.1419$. We also have

$$|||A^{=}|||_\infty = 5, \quad |||A^{\leqslant}|||_\infty = 3, \qquad |b^{=}|_{\min} = 0.1, \quad |b^{\leqslant}|_{\min} = 0.07.$$

None of the elements of φ^* is 0, so $\mu^* = m = 25$. Table 4.3 lists some values of N obtained from Theorems 3.1, 3.3, and Lemma 3.6 assuming the tolerances $\delta = (0.005, 0.01, \infty, \infty)$ for satisfying the constraints.

η	ε	$\hat{\alpha}$	$N_{\mathrm{Th}\,3.1}$
0.01	10^{-6}	0.9163	120000
	10^{-12}	0.9126	
	10^{-24}	0.8995	
0.005		0.8326	
		0.8257	
		0.7991	
0.001		0.3567	160807
		0.3471	165692
		0.3171	183001

ϑ	ε	$\hat{\alpha}$	$N_{\mathrm{Th}\,3.3}$	$N_{\mathrm{Le}\,3.6}$
0.05	10^{-6}	0.9833	416022	489889
	10^{-12}	0.9825	428261	502121
	10^{-24}	0.9799	471616	545477
0.01		0.9163	$1.35 \cdot 10^7$	$1.58 \cdot 10^7$
		0.9126	$1.38 \cdot 10^7$	$1.62 \cdot 10^7$
		0.8995	$1.49 \cdot 10^7$	$1.72 \cdot 10^7$
0.005		0.8326	$5.96 \cdot 10^7$	$6.96 \cdot 10^7$
		0.8253	$6.08 \cdot 10^7$	$7.08 \cdot 10^7$
		0.7991	$6.51 \cdot 10^7$	$7.52 \cdot 10^7$

Table 4.3: Traffic example with $\delta = (0.005, 0.01, \infty, \infty)$. Empty entries indicate repetition of previous values. Left: the $N(\delta, \varepsilon, \eta)$ of Theorem 3.1. 120000 is the value of $(m-1)/(2\vartheta_\infty)$. Right: the $N(\delta, \varepsilon, \vartheta)$ of Theorem 3.3 and Lemma 3.6.

In this problem it makes much more sense to think of the frequency or count vectors as (traffic) *matrices*. Consider the 3d row of Table 4.3 on the right. With $n = 545500$, by Definition 2.1 we get the count matrix

$$\nu^* = \begin{pmatrix} 16796 & 16796 & 10265 & 10264 & 16796 \\ 32299 & 32299 & 19738 & 19738 & 32299 \\ 28366 & 28366 & 28366 & 28366 & 28366 \\ 10910 & 10910 & 10910 & 10910 & 10910 \\ 28366 & 28366 & 28366 & 28366 & 28366 \end{pmatrix}.$$

How are we to interpret this? First, the number of all possible 5×5 count matrices with total sum 545500 is $|F_{545500}| = \binom{545500+24}{24} = 7.77 \cdot 10^{113}$. Second, $1.171 \cdot 10^{104}$ of these matrices satisfy the constraints. (To find this number we express (r_1, r_2, r_4), (s_{23}, s_{53}), and δ as rationals, resulting in inequalities such as

$$\begin{array}{rcccl} (13/100 - 1/2000)n & \leqslant & \nu_{11} + \nu_{12} + \nu_{13} + \nu_{14} + \nu_{15} & \leqslant & (13/100 + 1/2000)n, \\ (25/100 - 1/2000)n & \leqslant & \nu_{21} + \nu_{22} + \nu_{23} + \nu_{24} + \nu_{25} & \leqslant & (25/100 + 1/2000)n, \\ (7/100 - 7/10000)n & \leqslant & \nu_{13} + \nu_{53} + \nu_{14}, & & \end{array}$$

etc, and proceed as in §4.1.3 to find the number of lattice points in the polytope.) So about 1 out of every 10^{10} of the possible traffic matrices satisfies the constraints. Third, each of these matrices can be realized in many ways (multinomial coefficient). Now the entry for $N_{\mathrm{Th}\,3.3}$ in the third row of Table 4.3 on the right says:

> Consider all the assignments of the 545500 cars to the 25 matrix elements that result in one of the $1.171 \cdot 10^{104}$ matrices that satisfy the constraints to accuracy δ. Only *one in* 10^{24} of these assignments results in a matrix that deviates from ν^* by more than 27300 in ℓ_1 norm[9].

[9]This follows from the fact that $\|f - \varphi^*\|_1 \leqslant \vartheta \;\Rightarrow\; \|\nu - \nu^*\|_1 \leqslant n\vartheta + m$.

And the entry for $N_{\mathrm{Le}\,3.6}$ says something more impressive:

> The MaxEnt matrix ν^* can be realized in 10^{24} as many ways as the *entire set* of $1.171 \cdot 10^{104}$ matrices that satisfy the constraints but differ from ν^* by more than 27300 in ℓ_1 norm. ν^*/n differs from φ^* by no more than 43.75 in ℓ_1 norm[10].

4.3 Queue length distribution

Suppose we have a single-server $G/G/1$ queue of finite capacity $c \in \mathbb{N}$, in which customers arrive at rate λ and experience a mean waiting time $\bar{W}$. The known or measured λ and $\bar{W}$ imply (Little's law) a mean queue length $\bar{L} = \lambda \bar{W} \in \mathbb{R}$. So we consider the system under the following two states of knowledge: (a) besides c, only the mean queue length $\bar{L}$ is known, (b) in addition, we know that the probability that the queue is empty is in the interval $[a_1, b_1]$, and the probability that it is full is in $[a_2, b_2]$. What can be said in these two scenarios about the distribution $p_0, p_1, \ldots, p_c$ of the queue's length L? (This is a simple example; much more complex MaxEnt queueing problems are addressed in [Kou94], [BGdMT06].)

Using what was said in §3.4, here we have the problem of inferring a unique discrete p.d. $(p_0, p_1, \ldots, p_c) \in \mathbb{R}^{c+1}$ from the information

$$1p_1 + 2p_2 + \cdots + cp_c = \bar{L}, \tag{4.1}$$

in the first case, and from

$$1p_1 + 2p_2 + \cdots + cp_c = \bar{L}, \quad p_0 \in [a_1, b_1]\,,\ p_c \in [a_2, b_2] \tag{4.2}$$

in the second case. We interpret information (4.1) as assigning n probability quanta of size $1/n$ to $m = c+1$ boxes under the constraint that the "mean box index" must equal $\bar{L}$, and information (4.2) as imposing the additional constraints that the fraction assigned to box 0 must be between a_1 and b_1, while that assigned to box c must be between a_2 and b_2. If we take $c = 12$, $\bar{L} = 5.63$, the MaxEnt p.d. φ^* in the first case is, as expected, geometric:

$$\varphi^*_k = 0.0897 \cdot 0.9739^k, \qquad \text{with} \quad H^* = 2.5600.$$

Now we add information to the effect that the probability of being empty is larger than expected while that of being full is smaller than expected, expressed by $p_0 \in [0.12, 0.14]$, $p_{12} \in [0.01, 0.04]$. We get a distribution of lower entropy, geometric between 1 and 11:

$$\varphi^*_0 = 0.12, \quad \varphi^*_k = 0.0768 \cdot 0.987^k, \quad \varphi^*_{12} = 0.04 \qquad \text{with} \quad H^* = 2.5432. \tag{4.3}$$

To investigate the concentration around $p^* = \nu^*/n$, in case (4.1) we have $|||A^{=}|||_\infty = 78, |b^{=}|_{\min} = 5.63$, and in case (4.2) we add $|||A^{\leqslant}|||_\infty = 1, |b^{\leqslant}|_{\min} = 0.01$. If we choose the

[10] In this case we have $\hat{\alpha} = 6.87 \cdot 10^{-4}$.

tolerances $\delta^{=} = 10^{-5}, \delta^{\leqslant} = 10^{-3}$, Table 4.4 lists some values of N obtained from Theorem 3.3 and from Corollary 3.1 for $\varepsilon = 10^{-20}$. The results are the same for both scenarios even though the φ^* with bounded p_0 and p_{12} has lower entropy, for two reasons: first, because H^* does not appear in Theorem 3.3 or Corollary 3.1; second, because ϑ_∞ (Proposition 2.2) is the same in both cases. To interpret the first line of Table 4.4, suppose we choose

$(\delta^{=}, \delta^{\leqslant})$	ϑ	ε	ϑ_∞	$N_{\mathrm{Th}\,3.3}$	$\hat{\alpha}$	$N_{\mathrm{Co}\,3.1}$
$10^{-5}, 10^{-3}$	0.01	10^{-20}	$7.22 \cdot 10^{-7}$	$8.31 \cdot 10^{6}$	$1.76 \cdot 10^{-4}$	$1.35 \cdot 10^{7}$
	0.001			$1.00 \cdot 10^{9}$	$8.37 \cdot 10^{-6}$	$1.17 \cdot 10^{9}$
	10^{-4}			$1.23 \cdot 10^{11}$	$6.84 \cdot 10^{-7}$	$1.42 \cdot 10^{11}$
	10^{-5}			$1.45 \cdot 10^{13}$	$5.70 \cdot 10^{-8}$	$1.68 \cdot 10^{13}$
	10^{-6}			$1.67 \cdot 10^{15}$	$5.02 \cdot 10^{-9}$	$1.94 \cdot 10^{15}$

Table 4.4: Denominator N of the rational approximation $p^* \in P_N$ to the MaxEnt p.d. $\varphi^* \in \mathbb{R}^{13}$, from Theorem 3.3 and from Corollary 3.1. Results are the same whether only the mean is known, or bounds on p_0 and p_{12} are also known. Recall that ϑ_∞ depends only on δ. Empty entries signify no change from above.

$n = 13500000$. We then find

$$p^* = \frac{1}{13500000}\,(1620000, 964722, 977442, 990323, 1003387, 1016618, 1030037, \\ 1043613, 1057372, 1071308, 1085429, 1099749, 540000). \tag{4.4}$$

By Corollary 3.1, this rational approximation to the MaxEnt p.d. (4.3) has at least 10^{20} times more realizations than the entire set of p.d.'s which satisfy the constraints to accuracy δ but differ from the p.d. (4.3) by more than 0.01 in ℓ_1 norm.

5 Conclusion

The phenomenon of entropy concentration appears when a large number of units is allocated to containers subject to constraints that are linear functions of the numbers of units in each container: most allocations will result in frequency (normalized count) vectors with entropy close to that of the vector of maximum entropy that satisfies the constraints. Asymptotic proofs of this phenomenon are known, beginning with the work of E. T. Jaynes, but here we provided explicit bounds on how large the number of units must be for concentration to any desired degree to occur, and at the same time dealt with the fact that constraints cannot be satisfied exactly by rational frequencies, but only to some prescribed tolerances. We also established a perhaps more useful version of the concentration result, in terms of deviation from the maximum entropy vector, instead of the usual maximum entropy value, as well as results that pertain to the maximum entropy vector itself and not to a whole set of vectors around it. Because of its conceptual simplicity and minimality of assumptions, entropy concentration is a powerful justification of the widely-used discrete

MaxEnt method (the other being axiomatic formulations), and we believe that the explicit, non-asymptotic bounds strengthen it considerably. All of our results were illustrated with detailed numerical examples.

Acknowledgements

Thanks to David Applegate for discussing with me what can be done, Howard Karloff for telling me what can't be done, Steve Korotky for our many discussions on entropy and networks, and Neil Sloane for many informative discussions, answering many questions, in particular about latttice points, and for his careful reading of the paper.

A Proofs

Proof of Proposition 2.1

Rounding ensures that $\|\tilde{\nu} - n\varphi^*\|_\infty \leqslant 1/2$. From the explanation after Definition 2.1, the adjustment of $\tilde{\nu}$to ν^* ensures $\|\nu^* - n\varphi^*\|_\infty \leqslant 1$, which establishes the ℓ_∞ claim. In more detail, the 0 elements of ν^* coincide with those of φ^*, and this adjustment causes at most $\lfloor \mu^*/2 \rfloor$ of the non-zero elements of ν^* to differ from the corresponding elements of $n\varphi^*$ by $\leqslant 1$, so $\|\nu^* - n\varphi^*\|_1 \leqslant 1 \cdot \lfloor \mu^*/2 \rfloor + (1/2) \cdot (\mu^* - \lfloor \mu^*/2 \rfloor) \leqslant 3\mu^*/4$. Hence $\|\nu^*/n - \varphi^*\|_1 \leqslant (3/4)\mu^*/n$, which establishes the claim for the ℓ_1 norm.

Proof of Proposition 2.2

Beginning with the equality constraints (2.1), note that $A^{=}\varphi^* = b^{=} \;\Leftrightarrow\; A^{=}(\varphi^* - f) + A^{=}f = b^{=}$. Set $A^{=}(f - \varphi^*) = \beta$. Then we have $A^{=}f = b^{=} + \beta$, with $\|\beta\|_\infty = \|A^{=}(f - \varphi^*)\|_\infty$. Now $\|A^{=}(f - \varphi^*)\|_\infty \leqslant |||A^{=}|||_\infty \|f - \varphi^*\|_\infty$, where $|||\cdot|||_\infty$ denotes the matrix infinity norm (also known as the "maximum row sum" norm). The inequality holds because the vector norm $\|\cdot\|_\infty$ is *compatible* with the (rectangular) matrix norm $|||\cdot|||_\infty$ (see [HJ90], §5.7). So if we make $\|f - \varphi^*\|_\infty \leqslant \delta^{=}|b^{=}|_{\min} / |||A^{=}|||_\infty$, we will have $\|\beta\|_\infty \leqslant \delta^{=}|b^{=}|_{\min}$, as required by (2.1). The inequality constraints (2.2) are handled in exacty the same way.

Finally, for the equalities with zeros (2.3) we have $A^{=0}f = \zeta$, where $\|\zeta\|_\infty = \|A^{=0}(f - \varphi^*)\|_\infty$.

Proof of Proposition 2.3

Theorem 16.3.2 of [CT91] is a similar result, but in terms of the ℓ_1 norm. The function $f(x) = -x \ln x$, $x \in [0, 1]$, is concave and has a maximum at $x = 1/e$. Let $a > 0$ be $\leqslant \gamma$, and consider the difference of the values of $f(x)$ at two points that are a apart: $g(x) = f(x + a) - f(x)$. Since $g'(x) \leqslant 0$ always, the maximum of $g(x)$ occurs at $x = 0$

and equals $-a \ln a$. So if $\gamma \leqslant 1/e$,

$$\forall x, \quad \max_{0 \leqslant a \leqslant \gamma} f(x+a) - f(x) \;=\; \gamma \ln \frac{1}{\gamma}, \qquad \gamma \leqslant \frac{1}{e}.$$

(This is tighter than what we would get by simply applying the defining inequality of concavity to f.) We have now shown that if $|f_i - \varphi_i^*| \leqslant \gamma$, then $|f_i \ln f_i - \varphi_i^* \ln \varphi_i^*| \leqslant \gamma \ln 1/\gamma$. The result of the proposition follows.

Proof of Proposition 2.4

Using Proposition 2.3, we want to find a $\hat{\gamma}$ s.t. $m\gamma \ln 1/\gamma \leqslant \eta H^*$ for all $\gamma \leqslant \hat{\gamma}$. Setting $y = 1/\gamma$ and $\zeta = m/(\eta H^*)$, we want to find a $\hat{y}$ s.t. for all $y \geqslant \hat{y}$, $y \geqslant \zeta \ln y$. We claim that this inequality, where $\zeta \gg 1$ and y is expected to be $\gg 1$, is satisfied by $\hat{y} = (1+c)\,\zeta \ln \zeta$, for any $c > 0$. Indeed,

$$\hat{y} \geqslant \zeta \ln \hat{y} \;\Leftrightarrow\; \zeta^{1+c} \geqslant (1+c)\,\zeta \ln \zeta \;\Leftrightarrow\; \zeta^c / \ln \zeta \geqslant 1 + c, \tag{A.1}$$

which is possible for any $c > 0$ if ζ is large enough. With $c = 0.5$, this condition is $\sqrt{\zeta}/\ln \zeta \geqslant 3/2$. But this holds for $\zeta \geqslant 21$, a very mild requirement. Finally, the function $y/\ln y$ is increasing for $y \geqslant 1$, so the l.h.s. of (A.1) will hold for all $y \geqslant \hat{y}$ as desired. We have now shown that $H^* - H(f) \leqslant \eta H^*$ will hold if f is s.t.

$$\|f - \varphi^*\|_\infty \;\leqslant\; \frac{2}{3} \frac{\eta H^*}{\ln(m/\eta H^*)},$$

where the "2/3" can be tightened to $1/(1+c)$.

Proof of Proposition 3.1

Begin with $\#f = \binom{n}{nf_1, \ldots, nf_m} = \binom{n}{nf_1, \ldots, nf_\mu}$ and use the fact that

$$\ln x! \;=\; x \ln x - x + \frac{1}{2} \ln x + \ln \sqrt{2\pi} + \frac{\vartheta}{12x}, \quad \vartheta \in (0,1), \tag{A.2}$$

which is defined for all $x > 0$ by $x! = \Gamma(x+1)$. Then we find

$$\ln \#f \;=\; nH(f) - (\mu - 1) \ln \sqrt{2\pi n} - \sum_{f_i > 0} \ln \sqrt{f_i} + \frac{\vartheta_0}{12n} - \sum_{f_i > 0} \frac{\vartheta_i}{12 n f_i}.$$

Finally, for the upper bound in Prop. 3.1, the sum of the last two terms is maximized when $\vartheta_0 = 1, \vartheta_i = 0$. For the lower bound, it is minimized when $\vartheta_0 = 0, \vartheta_i = 1$.

Proof of Lemma 3.1

We begin by observing that the sum over $F_n \cap \mathcal{C}(\delta)$ is bounded above by the sum over all of F_n and then use the bound of Proposition 3.1 on $\#f$ to find

$$\#B_n(\delta,\eta) \leqslant \sum_{\substack{f \in F_n \\ H(f) < (1-\eta)H^*}} \#f \leqslant e^{n(1-\eta)H^*} e^{\frac{1}{12n}} \sum_{f\in F_n} S(f,n). \tag{A.3}$$

To evaluate the last sum, let $F_n^{(\mu)}$ be the subset of F_n consisting of vectors with μ non-zero elements. Since the $F_n^{(\mu)}$ form a partition of F_n,

$$\sum_{f\in F_n} S(f,n) = \sum_{\mu=1}^{m} \sum_{f\in F_n^{(\mu)}} S(f,n) = \sum_{\mu=1}^{m} \binom{m}{\mu} \sum_{\substack{f_1=\nu_1/n,\ldots,f_\mu=\nu_\mu/n \\ \nu_1+\cdots+\nu_\mu=n,\ \nu_i\geqslant 1}} S(f,n),$$

where the $\binom{m}{\mu}$ comes from the fact that as pointed out in Proposition 3.1, $\#f$ depends only on the non-zero elements and not on their positions. Thus

$$\sum_{f\in F_n} S(f,n) = \sum_{\mu=1}^{m} \binom{m}{\mu} \frac{1}{(2\pi n)^{\frac{\mu-1}{2}}} \sum_{\substack{\nu_1+\cdots+\nu_\mu=n \\ \nu_1,\ldots,\nu_\mu\geqslant 1}} \frac{(\sqrt{n})^\mu}{\sqrt{\nu_1\cdots\nu_\mu}}. \tag{A.4}$$

We now need an auxiliary result on the inner sum in (A.4):

Proposition A.1 *For any $\mu \geqslant 2$,*

$$\sum_{\substack{\nu_1+\cdots+\nu_\mu=n \\ \nu_1,\ldots,\nu_\mu\geqslant 1}} \frac{1}{\sqrt{\nu_1\cdots\nu_\mu}} < \frac{\pi^{\mu/2}}{\Gamma(\mu/2)} n^{\mu/2-1}.$$

Proposition A.1 is proved separately later. Using this result in (A.4),

$$\sum_{f\in F_n} S(f,n) = \sqrt{2\pi/n} \sum_{\mu=1}^{m} \binom{m}{\mu} \left(\frac{n}{2}\right)^{\mu/2} \frac{1}{\Gamma(\mu/2)} < 4\sqrt{2\pi/n} \left(1+\sqrt{n/4}\right)^m < 4\sqrt{2\pi} 0.6^m n^{\frac{m-1}{2}}.$$

For the first inequality we used $\Gamma(\mu/2) \geqslant 2^{\mu/2-2}$ and for the second we assumed that $n \geqslant 100$ and $m \geqslant 2$. Combining the above with (A.3), and again assuming $n \geqslant 100$ we obtain the result of the lemma.

Proof of Proposition 3.2

Ignoring the rational requirement for the moment and denoting f by x, only $x_1, \ldots, x_{m-1}$ are independent, so our set is the subset of $\mathbb{R}^{m-1}$ belonging to

$$\begin{array}{ll} |x_i - \varphi_i^*| \leqslant \vartheta, \quad i = 1, \ldots, m-1, & (m-1)\text{-dimensional cube,} \\ |(x_1 + \cdots + x_{m-1}) - (\varphi_1^* + \cdots + \varphi_{m-1}^*)| \leqslant \vartheta, & \text{region between two hyperplanes,} \\ x_1 + \cdots + x_{m-1} \leqslant 1, \quad x_i \geqslant 0, & \text{unit } (m-1)\text{-dimensional simplex.} \end{array} \tag{A.5}$$

We will construct inside this set an $(m-1)$-dimensional rectangular parallelepiped P whose intersection with F_n is easy to count. To construct P we will determine its two extreme points, the one with the largest coordinates, $\varphi^* + y$, and the one with the smallest, $\varphi^* - z$, where $y_i, z_i \geqslant 0$.

If $\varphi^* + y$ satisfies (A.5), then

$$y_i \leqslant \vartheta, \qquad y_1 + \cdots + y_{m-1} \leqslant \vartheta, \qquad y_1 + \cdots + y_{m-1} \leqslant \varphi_m^*, \quad y_i \geqslant -\varphi_i^*.$$

The 4th inequality is true, and the 2nd implies the first. The 2nd and 3d inequalities are satisfied if $y_1 + \cdots + y_{m-1} = \min(\vartheta, \varphi_m^*)$, and $y_1 \cdots y_{m-1}$ will be maximized if

$$y_1 = \cdots = y_{m-1} = \frac{1}{m-1} \min(\vartheta, \varphi_m^*). \tag{A.6}$$

Since φ^* has $\mu^* \geqslant 1$ non-zero elements, we can assume w.l.o.g. that $\varphi_m^* > 0$.

Similarly, for the other extreme point $\varphi^* - z$ we must have $z_1 + \cdots + z_{m-1} \leqslant \vartheta$ and $z_i \leqslant \varphi_i^*$. If some φ_i^* are 0 the corresponding z_i are 0, and w.l.o.g. we can take the non-zero z_i to be $z_1, \ldots, z_{\mu^*-1}$. Then the z_i that satisfy the inequalities and maximize the product $z_1 \cdots z_{\mu^*-1}$ are

$$z_1 = \cdots = z_{\mu^*-1} = \vartheta / (\mu^* - 1). \tag{A.7}$$

But this needs $\vartheta / (\mu^* - 1) \leqslant \varphi_i^*$ for all the non-zero φ_i^*, which we have assumed.

Thus from (A.6) and (A.7) $\mu^* - 1$ sides of P have length $y_i + z_i = \frac{1}{m-1} \min(\vartheta, \varphi_m^*) + \frac{1}{\mu^*-1}\vartheta$, and the other $m - \mu^*$ sides have length $y_i + z_i = \frac{1}{m-1} \min(\vartheta, \varphi_m^*)$. Again w.l.o.g. we can take φ_m^* to be $\varphi_{\max}^*$, the largest element of φ^*. So if we assume that $\vartheta \leqslant \varphi_{\max}^*$, P has $\mu^* - 1$ sides of length $\vartheta \left(\frac{1}{m-1} + \frac{1}{\mu^*-1} \right)$ and $m - \mu^*$ sides of length $\frac{\vartheta}{m-1}$.

Now a k-dimensional parallelepiped with sides of lengths $L_1, \ldots, L_k$, irrespective of its location in $\mathbb{R}^k$, contains at least $\lfloor L_1 \rfloor \cdots \lfloor L_k \rfloor$ lattice points, i.e. points in $\mathbb{Z}^k$. (This can be established by induction on k. For $k = 1$ it says that a segment of length L on the real axis contains at least $\lfloor L \rfloor$ integers.) Applying this to P with all its $m-1$ dimensions scaled up by n, the scaled P must contain at least

$$\left\lfloor n\vartheta \left(\frac{1}{m-1} + \frac{1}{\mu^*-1} \right) \right\rfloor^{\mu^*-1} \left\lfloor \frac{n\vartheta}{m-1} \right\rfloor^{m-\mu^*} \tag{A.8}$$

points whose coordinates are rational numbers with denominator n, i.e. vectors in F_n. The first factor and its attendant condition $\vartheta \leqslant (\mu^* - 1)\varphi_{\min}^*$ are absent if $\mu^* = 1$.

Proof of Lemma 3.2

Let $0 < \alpha < 1$ be some constant whose purpose is expained later, in the proof of Theorem 3.1. We begin by deriving a lower bound on the size of $A_n(\delta, \alpha\eta)$, a subset of $A_n(\delta, \eta)$. Consider the set $\mathcal{A} = \{f \in F_n \mid \|f - \varphi^*\|_\infty \leqslant \vartheta_0\}$. Since $\vartheta_0 \leqslant \vartheta_\infty$, Proposition 2.2 implies that any$f$ in this set also belongs to $\mathcal{C}(\delta)$. Further, by the middle expression in the definition of ϑ_0, Proposition 2.4 implies that any such f also has entropy at least $(1 - \alpha\eta)H^*$. Thus all f in the set $\mathcal{A}$ belong to $A_n(\delta, \alpha\eta)$. Finally ϑ_0 satisfies the conditions of Proposition 3.2, hence the size of $\mathcal{A}$ is bounded from below by $\Lambda(n, \vartheta_0, \mu^*)$. But $\mathcal{A} \subseteq A_n(\delta, \alpha\eta) \subseteq A_n(\delta, \eta)$, so we established the first claim of the lemma.

Now suppose that all f in $A_n(\delta, \alpha\eta)$ have at least $\mu \geqslant 1$ non-zero elements; for the purposes of this proof we may take these to be the first μ elements. Then by Proposition 3.1, if g is an arbitrary element of $A_n(\delta, \alpha\eta)$,

$$\#A_n(\delta, \alpha\eta) \;\geqslant\; |A_n(\delta, \alpha\eta)|\, e^{n(1-\alpha\eta)H^*} e^{-\frac{1}{12n}\left(\frac{1}{g_1}+\cdots+\frac{1}{g_\mu}\right)} \frac{1}{(2\pi n)^{\frac{\mu-1}{2}}} \frac{1}{\sqrt{g_1 \cdots g_\mu}}. \tag{A.9}$$

Let $\xi = (ng_1, \ldots, ng_\mu)$; this vector has integral entries, all positive, and summing to n. The maximum of $1/\xi_1 + \cdots + 1/\xi_\mu$ equals $\mu - 1 + 1/(n - \mu + 1)$, occurring when $\xi_1 = \cdots = \xi_{\mu-1} = 1$ and $\xi_\mu = n - \mu + 1$. Thus the exponential in (A.9) is at least $e^{-\mu/12}$. Further, the maximum of $\sqrt{g_1 \cdots g_\mu}$ subject to $g_1 + \cdots + g_\mu = 1$ occurs at $g_1 = \cdots = g_\mu = 1/\mu$, so the last factor in (A.9) is at least $\mu^{\mu/2}$. Finally $A_n(\delta, \eta) \supseteq A_n(\delta, \alpha\eta)$, and so (A.9) implies the second result of the lemma, but with the number μ still undetermined. By requiring $\|f - \varphi^*\|_\infty$ to be less than the smallest non-zero element of φ^*, we can ensure that there is no element of f which is 0 while the corresponding element of φ^* is positive; this is accomplished by the last term on the r.h.s. of (3.7). Thus we can take μ equal to μ^*, the number of non-zero elements of φ^*.

Proof of Theorem 3.1

The upper bound on $\#B_n$ and the lower bound on $\#A_n$ are given by Lemmas 3.1 and 3.2. Both these bounds *increase* when the entropy tolerance η decreases towards 0, as makes sense. To simplify the proof we assume that $\Lambda(n, \vartheta_0, \mu^*) \geqslant 1$. Then combining the two bounds and unifying some numerical constants

$$\frac{\#A_n(\delta, \eta)}{\#B_n(\delta, \eta)} \;>\; \frac{0.249}{0.6^m} e^{-\mu^*/12} \left(\frac{\mu^*}{2\pi}\right)^{\mu^*/2} n^{-\frac{m+\mu^*-2}{2}} e^{n(1-\alpha)\eta H^*}, \qquad n \geqslant \frac{m-1}{[\![2, \mu^* = m]\!]\vartheta_0}. \tag{A.10}$$

When everything else is fixed, this lower bound on $\#A_n/\#B_n$ (eventually) increases as $n \to \infty$, as we want it to. In general, this behavior would have been impossible if α were 1. This is why we introduced α and required it to be < 1: it serves to strictly separate our bounds on $\#A_n$ and $\#B_n$. There is freedom in choosing the value of α, which we exploit below.

To establish (3.5) we need the l.h.s. of (A.10) to be $\geqslant 1/\varepsilon+1$. This reduces to requiring

$$n \;\geqslant\; C_1 \ln n + C_2, \quad n \geqslant (m-1)/\vartheta_0, \tag{A.11}$$

where the constants C_1, C_2 are

$$C_1 = \frac{0.5(m+\mu^*)-1}{(1-\alpha)\eta H^*}, \quad C_2 = \frac{m\ln 0.6 + (0.5\ln 2\pi + 1/12 - 0.5\ln\mu^*)\mu^* + \ln\left(\frac{1/\varepsilon+1}{0.249}\right)}{(1-\alpha)\eta H^*}. \tag{A.12}$$

We will now show that (A.11) is satisfied by

$$N(\alpha) = \begin{cases} 1.5\, C_1 \ln(C_1+C_2) + C_2, & \text{if } C_2 > 0 \text{ and } C_1 + C_2 \geqslant 21, \\ 1.5\, C_1 \ln C_1 + C_2, & \text{if } C_2 \leqslant 0. \end{cases} \tag{A.13}$$

First, assume $C_2 > 0$. Setting $n = 1.5\, C_1 \ln(C_1+C_2) + C_2$ in (A.11) with $k=1$ we reduce to

$$\sqrt{C_1+C_2} \geqslant 1.5 \frac{C_1}{C_1+C_2}\ln(C_1+C_2) + \frac{C_2}{C_1+C_2}.$$

The r.h.s. of this condition is a convex combination of $1.5\ln(C_1+C_2)$ and 1, so the first condition will hold if $\sqrt{C_1+C_2} \geqslant 1.5\ln(C_1+C_2)$, which is true when $C_1+C_2 \geqslant 21$. Now let $C_2 \leqslant 0$. Putting $n = 1.5 C_1 \ln C_1 + C_2$ in (A.11), we reduce to establishing $C_1^{1.5} \geqslant 1.5\ln C_1 + C_2$, which will hold if $C_1^{1.5} \geqslant \ln C_1^{1.5}$, always true.

The r.h.s. of (A.13) depends on $\alpha \in (0,1)$, which has up to this point been left unspecified. We finally need $n \geqslant N(\alpha)$ and $n \geqslant (m-1)/\big([\![2, \mu^* = m]\!]\vartheta_0\big)$, and we observe that as $\alpha \nearrow$, the first of these bounds increases while the second decreases. Further, the first bound is finite at 0 and infinite at 1, whereas the second is infinite at 0 and finite at 1. Thus there is an optimal α which makes the two bounds equal, the $\hat{\alpha}$ which solves $N(\alpha) = (m-1)/\big([\![2, \mu^* = m]\!]\vartheta_0(\alpha)\big)$.

Proof of Proposition A.1

It seems that the inequality

$$\sum_{\substack{\nu_1+\cdots+\nu_\mu = n \\ \nu_1,\ldots,\nu_\mu \geqslant 1}} \frac{1}{\sqrt{\nu_1\cdots\nu_\mu}} \;\leqslant\; \int_{\substack{x_1+\cdots+x_\mu = n \\ x_1,\ldots,x_\mu \geqslant 0}} \frac{dx_1\cdots dx_\mu}{\sqrt{x_1\cdots x_\mu}} \;=\; n^{\mu/2-1}\frac{\pi^{\mu/2}}{\Gamma(\mu/2)},$$

should hold (see [GR80], 4.635, #4). Being unable to show this directly, we go through a more circuitous and lengthier proof.

Consider the simplest case $\mu = 2$ first. We can bound the sum as follows:

$$\sum_{\substack{\nu_1+\nu_2 = n \\ \nu_1,\nu_2 \geqslant 1}} \frac{1}{\sqrt{\nu_1\nu_2}} \;=\; \sum_{\nu=1}^{n-1} \frac{1}{\sqrt{\nu(n-\nu)}} \;<\; \int_0^n \frac{dx}{\sqrt{x(n-x)}} \;=\; \pi. \tag{A.14}$$

To see this, note that $\sum_{\nu=1}^{n/2} 1/\sqrt{\nu(n-\nu)} < \int_0^{n/2} dx/\sqrt{x(n-x)} = \pi/2$ because the sum is a lower Riemann sum for the integral. Since the summand is symmetric about $n/2$, doubling this produces the desired result.

Now consider the case of even μ, i.e. $\mu = 2\lambda$. Divide the ν_i into λ pairs, each of which sums to some number $\geqslant 2$ and these numbers in turn sum to n:

$$\sum_{\substack{\nu_1+\cdots+\nu_{2\lambda}=n \\ \nu_1,\ldots,\nu_{2\lambda}\geqslant 1}} \frac{1}{\sqrt{\nu_1\cdots\nu_{2\lambda}}} = \sum_{\substack{k_1+\cdots+k_\lambda=n \\ k_1,\ldots,k_\lambda\geqslant 2}} \left(\sum_{\substack{\nu_1+\nu_2=k_1 \\ \nu_1,\nu_2\geqslant 1}} \frac{1}{\sqrt{\nu_1\nu_2}} \cdots \sum_{\substack{\nu_{2\lambda-1}+\nu_{2\lambda}=k_\lambda \\ \nu_{2\lambda-1},\nu_{2\lambda}\geqslant 1}} \frac{1}{\sqrt{\nu_{2\lambda-1}\nu_{2\lambda}}} \right) < \pi^\lambda \sum_{\substack{k_1+\cdots+k_\lambda=n \\ k_1,\ldots,k_\lambda\geqslant 2}} 1. \quad \text{(A.15)}$$

Here the inequality follows by applying (A.14), which does not depend on n, to each of the inner sums. Further,

$$\sum_{\substack{k_1+\cdots+k_\lambda=n \\ k_1,\ldots,k_\lambda\geqslant 2}} 1 = \sum_{\substack{k_1+\cdots+k_\lambda=n-2\lambda \\ k_1,\ldots,k_\lambda\geqslant 0}} 1 = \binom{n-\lambda-1}{\lambda-1},$$

where in the first equality we assume w.l.o.g. that $2\lambda < n$, and the 2nd equality follows from the fact that the number of compositions of N into M parts (i.e. the solutions of $k_1+\cdots+k_M = N$, $k_i \geqslant 0$), is $\binom{N+M-1}{M-1}$. Finally we bound the binomial coefficient by $\binom{n-\lambda-1}{\lambda-1} < \frac{n^{\lambda-1}}{(\lambda-1)!}$, to arrive at

$$\sum_{\substack{\nu_1+\cdots+\nu_{2\lambda}=n \\ \nu_1,\ldots,\nu_{2\lambda}\geqslant 1}} \frac{1}{\sqrt{\nu_1\cdots\nu_{2\lambda}}} < \frac{\pi^\lambda}{\Gamma(\lambda)} n^{\lambda-1}. \quad \text{(A.16)}$$

Now we turn to the case of odd μ, i.e. $\mu = 2\lambda+1$. Similarly to what we did above,

$$\sum_{\substack{\nu_1+\cdots+\nu_{2\lambda}+\nu_{2\lambda+1}=n \\ \nu_1,\ldots,\nu_{2\lambda},\nu_{2\lambda+1}\geqslant 1}} \frac{1}{\sqrt{\nu_1\cdots\nu_{2\lambda}\nu_{2\lambda+1}}} = \sum_{\substack{k_1+k_2=n \\ k_1\geqslant 1, k_2\geqslant 2\lambda}} \left(\sum_{\nu_{2\lambda+1}=k_1} \frac{1}{\sqrt{\nu_{2\lambda+1}}} \sum_{\substack{\nu_1+\cdots+\nu_{2\lambda}=k_2 \\ \nu_1,\ldots,\nu_{2\lambda}\geqslant 1}} \frac{1}{\sqrt{\nu_1\cdots\nu_{2\lambda}}} \right). \quad \text{(A.17)}$$

By (A.16), the r.h.s. does not exceed

$$\frac{\pi^\lambda}{\Gamma(\lambda)} \sum_{\substack{k_1+k_2=n \\ k_1 \geqslant 1, k_2 \geqslant 2\lambda}} \frac{k_2^{\lambda-1}}{\sqrt{k_1}} \;<\; \frac{\pi^\lambda}{\Gamma(\lambda)} \sum_{k=1}^{n-1} \frac{k^{\lambda-1}}{\sqrt{n-k}},$$

and this last sum can be bounded by the integral

$$\int_0^n \frac{k^{\lambda-1}}{\sqrt{n-k}} dk \;=\; n^{\lambda-1/2} \int_0^1 \frac{x^{\lambda-1}}{\sqrt{1-x}} dx \;=\; n^{\lambda-1/2} \frac{\Gamma(\lambda)\Gamma(1/2)}{\Gamma(\lambda+1/2)}.$$

We have thus shown that for $\mu = 2\lambda + 1$,

$$\sum_{\substack{\nu_1+\cdots+\nu_{2\lambda+1}=n \\ \nu_1,\ldots,\nu_{2\lambda+1} \geqslant 1}} \frac{1}{\sqrt{\nu_1 \cdots \nu_{2\lambda+1}}} \;<\; \frac{\pi^{\lambda+1/2}}{\Gamma(\lambda+1/2)} n^{\lambda-1/2}. \tag{A.18}$$

Eqs. (A.16) and (A.18) establish the proposition for all $\mu \geqslant 2$.

Proof of Lemma 3.3

First we show that if $n \geqslant 1/\vartheta_0(\alpha)$ then $f^* \in A_n(\delta, \alpha\eta)$. By Proposition 2.1, $n \geqslant 1/\vartheta_0(\alpha) \Rightarrow \|f^* - \varphi^*\|_\infty \leqslant \vartheta_\infty$, so by Proposition 2.2 $f^* \in \mathcal{C}(\delta)$. Further, $\|f^* - \varphi^*\|_\infty \leqslant \frac{2}{3} \frac{\alpha\eta H^*}{\ln(m/(\alpha\eta H^*))}$ means that $H(f^*) \geqslant (1-\alpha\eta) H^*$ by Proposition 2.4. Therefore $n \geqslant 1/\vartheta_0(\alpha)$ implies that f^* belongs to the set $A_n(\delta, \alpha\eta)$ as claimed.

Next we put a lower bound on $\# f^*$. Applying Proposition 3.1 we see that $\# f$ is $\geqslant$ the r.h.s. of (A.9) in the proof of Lemma 3.2 with $|A_n(\delta, \alpha\eta)| = 1$, so $\# f$ is $\geqslant$ the bound of Lemma 3.2 on $\# A_n(\delta, \eta)$ with $\Lambda = 1$. Then from the proof of Theorem 3.1, we see that $\# f^* / \# B_n(\delta, \eta)$ is $\geqslant$ the r.h.s. of (A.10), but with the condition on n being $n \geqslant 1/\vartheta_0(\alpha)$. The rest of the proof of Theorem 3.1 then applies, to the point where n has to satisfy $n \geqslant N(\alpha)$ and $n \geqslant 1/\vartheta_0(\alpha)$. $\hat{\alpha}$ equalizes these bounds, and the completion of the proof of Theorem 3.1 then establishes that if $n \geqslant 1/\vartheta_0(\hat{\alpha})$, $\# f^* / \# B_n(\delta, \eta) \geqslant 1/\varepsilon + 1$.

Proof of Proposition 3.3

The function $y \ln(m/y)$, $m \geqslant 2$, is increasing for $y \in (0, 1/2]$. The first implication in the proposition then follows immediately from Theorem 16.3.2 of [CT91], the ℓ_1 norm bound on entropy, which states that if two m-vectors p, q are s.t. $\|p - q\|_1 \leqslant 1/2$, then $|H(p) - H(q)| \leqslant \|p - q\|_1 \ln(m/\|p - q\|_1)$.

To prove the second implication we use Pinsker's inequality and the "triangle inequality" for cross- or relative entropy, or divergence $D(\cdot\|\cdot)$. Applied to f and φ^*, Pinsker's inequality states that $D(f\|\varphi^*) \geqslant \frac{1}{2}\|f - \varphi^*\|_1^2$ (see [CT91], Lemma 12.6.1). Then the triangle inequality, using the uniform distribution as the prior or reference distribution ([CT91],

Theorem 12.6.1), implies that $H(\varphi^*) - H(f) \geqslant D(f\|\varphi^*)$. What we want to prove follows from the above two inequalities.

Pinsker's inequality can be tightened in two ways: [OW05] show that the 1/2 can be replaced by a factor $c(\varphi^*) \geqslant 1/2$, and [FHT03] give right-hand sides that are polynomials involving powers of the norm beyond the square.

Proof of Lemma 3.4

Entirely analogous to that of Lemma 3.1, except that the set B'_n is defined by (3.9) instead of (3.3), and the factor $e^{n(1-\eta)H^*}$ in (A.3), coming from the upper bound on $\#f$ of Proposition 3.1, is replaced by the factor $e^{n(H^*-\vartheta^2/2)}$ of Proposition 3.3.

Proof of Lemma 3.5

The proof follows that of Lemma 3.2: first we lower-bound the size of $A'_n(\delta, \alpha\vartheta)$ and then the entropy of the f in it. The basic difference is that here we have ℓ_1 norms. If $\|f - \varphi^*\|_1 \leqslant \vartheta'_0$, so is $\|f - \varphi^*\|_\infty$, and then Proposition 3.2 says that the size of $A'_n(\delta, \vartheta'_0)$ is at least $\Lambda(n, \vartheta'_0, \mu^*)$. Second, concerning the entropy of $f \in A'_n(\delta, \vartheta'_0)$, by Proposition 3.3 $\|f - \varphi^*\|_1 \leqslant \vartheta'_0$ implies that $H(f)$ is at least $H^* - h(\alpha\vartheta)$. The proof then follows that of Lemma 3.2, except that the term $e^{n(1-\alpha\eta)H^*}$ in (A.9) is replaced by $e^{n(H^*-h(\alpha\vartheta))}$.

Proof of Proposition 3.4

$\partial\psi/\partial\alpha$ is always negative, and $\psi(\vartheta^2/2, \vartheta) > 0$ if $m < 1/2\vartheta^3 e^{1/\vartheta}$. This establishes the first part. For the second part, we note, in addition, that $\psi(1, \vartheta) < 0$ even for $m = 2$.

Proof of Theorem 3.3

The proof uses Lemmas 3.4 and 3.5 and is completely analogous to that of Theorem 3.1. The main feature is that H^* falls out of the new (A.10), the exponential is $e^{n\psi(\alpha,\vartheta)}$ with $\psi(\alpha, \vartheta) = \vartheta^2/2 - h(\alpha\vartheta)$, and the condition on n is now $n \geqslant (m-1)/\big([\![2, \mu^* = m]\!]\vartheta'_0\big)$. C_1, C_2 are the same as in Theorem 3.1, except for the denominators. Finally, $N(\alpha)$ is finite at $\alpha = 0$ and increases to ∞ at $\alpha = \alpha_0$, whereas $1/\vartheta'_0(\alpha)$ is infinite at $\alpha = 0$ and decreases to a finite value at $\alpha = \alpha_0$. Thus the equation $N(\alpha) = (m-1)/\big([\![2, \mu^* = m]\!]\vartheta_0(\alpha)\big)$ has a root $\hat{\alpha}$ between 0 and α_0, which equates the two sides and is therefore the optimal α.

Proof of Lemma 3.6

The proof is analogous to that of Lemma 3.3. First, by Proposition 2.1, $n \geqslant 1/\vartheta'_0(\alpha)$ implies that $f^* \in \mathcal{C}(\delta)$. Second, if $n \geqslant 3\mu^*/(4\vartheta'_0(\alpha))$ then $\|f^* - \varphi^*\|_1 \leqslant \alpha\vartheta$ by the 2nd claim of Proposition 2.1. Hence if $n \geqslant 3\mu^*/(4\vartheta'_0(\alpha))$, f^* belongs to the set $A'_n(\delta, \alpha\vartheta)$. Next, by the argument in the proof of Lemma 3.3, $\#f^*$ can be lower-bounded by the bound of

Lemma 3.5 with $\Lambda = 1$. So $\#f/\#B'_n(\delta, \vartheta)$ is lower-bounded by the new (A.10) as in the proof of Theorem 3.3 but the condition on n is now $n \geqslant 3\mu^*/(4\vartheta'_0(\alpha))$. The rest follows as in the proof of Theorem 3.3.

Proof of Corollary 3.2

In this case there are no constraints, so by Proposition 2.2 $\vartheta_\infty = \infty$. Also, $\mu^* = m$ and $\varphi^*_{\min} = 1/m$. Further, if $\vartheta < 1/m$, then $1/(\hat{\alpha}\vartheta) > m$, so the condition of Corollary 3.1 on n is $n \geqslant 3m/(4\hat{\alpha}\vartheta)$. The conditions $\vartheta < 1/m$ and $m < 1/2\vartheta^3 e^{1/\vartheta}$ are satisfied if $\vartheta \leqslant \min(0.09, 1/m)$. Finally, $\vartheta'_0(\alpha) = \alpha\vartheta$.

References

[ADSZ88] M. Avriel, W.E. Diewert, S. Schaible, and I. Zang. *Generalized Concavity*. Plenum Press, 1988.

[AS72] M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Functions*. 1972.

[BGdMT06] G. Bolch, S. Greiner, H. de Meer, and K.S. Trivedi. *Queueing Networks and Markov Chains*. Wiley, 2006.

[BV08] M. Benoit and S. Verdoolage. Polynomial approximations in the polytope model: Bringing the power of quasi-polynomials to the masses. In *Proceedings of 6th Workshop on Optimizations for DSP and Embedded Systems (ODES-6)*, pages 45–54. Boston, Massachusetts, April 2008.

[Csi99] I. Csiszár. The Method of Types. In S. Verdú and S. McLaughlin, editors, *Information Theory: 50 Years of Discovery*. IEEE Press, 1999.

[CT91] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. J. Wiley, 1991.

[DP09] D.D. Dubhashi and A. Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge, 2009.

[FHT03] A. Fedotov, P. Harremoes, and F. Topsøe. Refinements of pinsker's inequality. *IEEE Transactions on Information Theory*, 49, 2003.

[GR80] I.S. Gradshteyn and I.M. Rhyzik. *Table of Integrals, Series, and Products*. Academic Press, 1980.

[Gr1] P. Grünwald. Strong entropy concentration, game theory, and algorithmic randomness. *Lecture Notes in Artificial Intelligence*, LNAI 2111, 2001.

[Gr8] P. Grünwald. Entropy Concentration and the Empirical Coding Game. *Statistica Neerlandica*, 62(3):374–392, 2008. Also `http://arxiv.org/abs/0809.1017`.

[HJ90] R.A. Horn and C.R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[Jay82] E.T. Jaynes. On the Rationale of Maximum-Entropy Methods. *Proceedings of the IEEE*, 70(9):939–952, September 1982.

[Jay83] E.T. Jaynes. Concentration of Distributions at Entropy Maxima. In R.D. Rosenkrantz, editor, *E.T. Jaynes: Papers on Probability, Statistics, and Statistical Physics*. D. Reidel, 1983.

[Jay03] E.T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[KK92] J.N. Kapur and H.K. Kesavan. *Entropy Optimization Principles with Applications*. Academic Press, 1992.

[Kou94] D.D. Kouvatsos. Entropy maximisation and queueing network models. *Annals of Operations Research*, 48, 1994.

[ME98] *Maximum Entropy and Bayesian Methods in Science and Engineering*. Kluwer Academic Publishers, 1985-1998.

[ME99] *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*. American Institute of Physics (AIP), from 1999.

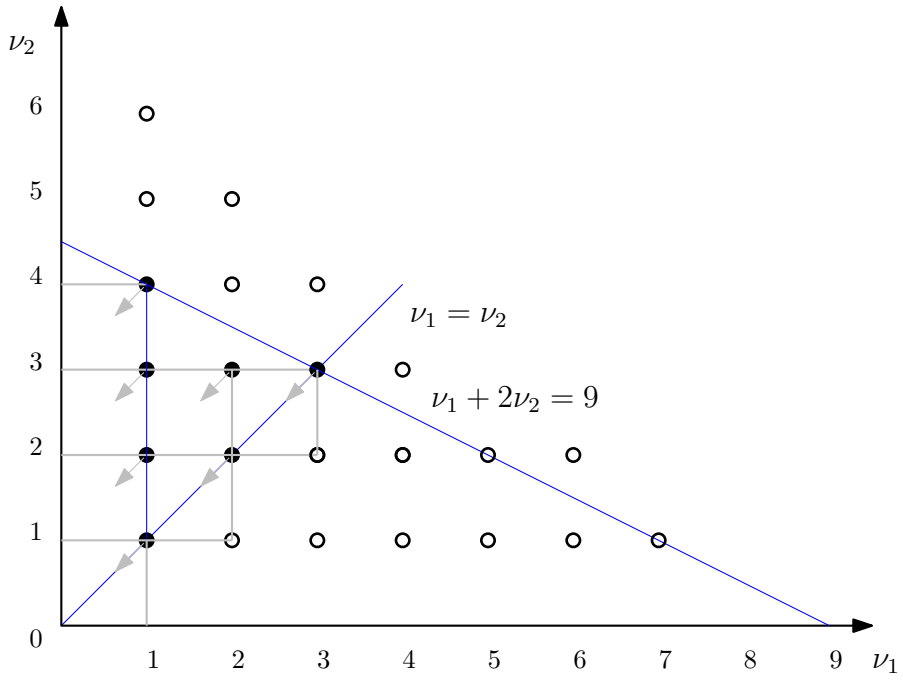
[OW05] E. Ordentlich and M.J. Weinberger. A distribution-dependent refinement of pinsker's inequality. *IEEE Transactions on Information Theory*, 51(5):1836–1840, May 2005.

[Ros83] R.D. Rosenkrantz, editor. *E. T. Jaynes: Papers on Probability, Statistics, and Statistical Physics*. D. Reidel [Kluwer], Dordrecht, The Netherlands, 1983.

[Tho79] M.U. Thomas. A generalized maximum entropy principle. *Operations Research*, 27(6), 1979.

[Tri69] M. Tribus. *Rational Descriptions, Decisions and Designs*. Pergamon Press, 1969.

[VWBC05] S. Verdoolaege, K. Woods, M. Bruynooghe, and R. Cools. Computation and Manipulation of Enumerators of Integer Projections of Parametric Polytopes. Technical report, Katholieke Universiteit Leuven, Department of Computer Science, Celestijnenlaan 200A, B-3001 Heverlee (Belgium), March 2005.


ν_2
6
5
4
3
2
1
0
1
2
3
4
5
6
7
8
9
ν_1
$\nu_1 = \nu_2$
$\nu_1 + 2\nu_2 = 9$