# Frequency Moments Inverse Problem and Maximum (Shannon vs. Rényi-Tsallis) Entropy *

*A case study in MaxProb*

M. Grendár, Jr. (umergren@savba.sk)
*Institute of Measurement Science, Slovak Academy of Sciences (SAS) Dúbravská cesta 9, Bratislava, 842 19, Slovakia and Institute of Mathematics and Informatics of Mathematical Institute SAS and of Matej Bel University, Banská Bystrica, Slovakia.*

M. Grendár

**Abstract.** An ill-posed inverse problem which lays in making a choice of probability distribution from set $\mathcal{H}$ of distributions is considered, in a non-standard setting where $\mathcal{H}$ is given by constraints – known as frequency moment constraints – which are non-linear (in probabilities). From perspective of their probabilistic justification(s) two methods for making the choice are studied: a) maximization of Shannon's entropy (MaxEnt), b) maximization of Rényi-Tsallis entropy (maxTent). Concerning MaxEnt, its large deviations asymptotic justification, known as conditioned weak law of large numbers, seems to be yet not developed for this case (of non-convex $\mathcal{H}$ and non-unique maximum). Hovewer, MaxProb justification of MaxEnt is valid also in this setting, implying that distribution chosen by maxTent is not the most probable one.

The case of Tsallis' $X$-frequency constraints is studied, as well. Two sub-classes of the feasible set are constructed, where it can be directly seen that maxTent choice is not only different than the most probable distribution, it is also asymptotically improbable. The findings, taken together, brings from probability viewpoint maxTent under suspect.

**Keywords:** nonconvex feasible set, nonlinear constraints, maximum entropy methods, probabilistic justifications, statistical physics

**Abbreviations:** MaxEnt – method of maximization of Shannon's entropy, maxTent – method of maximization of Rényi-Tsallis entropy, CWLLN – Conditioned Weak Law of Large Numbers

**AMS 2000 subject classification:** 80A23, 65R32, 60F10, 94A17

*with Mar*
*July 23, 2002*

* It is a pleasure to thank Aleš Gottvald, George Judge and Jonathan Smith for hints, discussions and notes on subject of this study. Lapses are mine.

**Table of Contents**

## 1. Preliminaries, generalities, context

*Setup.* Let $\mathbf{q}$ be a probability mass function (pmf), defined on $m$-element support. Let $\Pi_n$ be a set of all empirical pmf's which can be based on random samples of size $n$ and which supposedly come from $\mathbf{q}$. Let $\mathcal{H}_n \subseteq \Pi_n$.

*Problem.* Given this information, make a choice of an empirical pmf from $\mathcal{H}_n$.

**Example 1**: Let the feasible set of empirical pmf's (also called types, occurrence vectors) comprises of such types: 1) which are defined on 3-element support and 2) which have probability of one of support-points equal 2/3. Let $n$ be sufficiently large. Choose a type from this set. $\diamond$

If $\mathcal{H}_n$ contains more than one type, the task of making a choice is an ill-posed (underdetermined) inverse problem (IPIP). In general three attitudes towards it can be recognized: 1) (pragmatic) choice is made by some method and appropriateness of the method is judged by some criterion, 2) (MaxEnt) choice is made by maximization of Shannon's entropy (or relative entropy, if $\mathbf{q}$ is non-uniform), 3) (fundamentalist) the problem is non-sense (other name for its ill-posed-ness), hence it is meaningless to make a choice.

Many maxentists can be subsumed under the category of pragmatists. In other words, many pragmatists use method of maximization of Shannon's entropy (or relative entropy) (abbreviated as MaxEnt) for making the choice. Clearly, there are plenty of other ways of making the choice, and in particular several other objective functions to extremize. What makes the IPIP interesting from mathematical point of view is search for a justification of particular mechanism of distribution selection.

Boltzmann's combinatorial argument was turned into a formal mathematical technique: large deviation theory (see for instance [7]). The theory provides a probabilistic argument for use of Shannon's entropy (or its generalization: relative entropy) as the only reasonable objective function for solving the IPIP problem *when $n$ gets large*, under some special (though frequently met in applications) conditions. The argument shows (loosely speaking) that any empirical distribution other than the one which maximizes relative entropy on the feasible set $\mathcal{H}_\infty$ becomes improbable to come from $\mathbf{q}$ (in measure theoretic sense) as sample size grows large. To establish this result (which is known as conditioned weak law of large numbers, (cf. [22], [6]) two assumptions are usually needed: 1) uniqueness of the entropy maximizer, 2) convexity of the constraint set.

**Example 1, cont'd**: Would you make the same choice, having known that the feasible set consists of types which were obtained from random samples of size $n = 12$? $\diamond$

Another argument (dubbed MaxProb) which arises as well from Boltzmann's combinatorics, requires weaker assumptions and makes a different claim: the most probable type $\hat{\mathbf{n}} \in \mathcal{H}_n$ which can be drawn from $\mathbf{q}$ converges normalized by $n$ to a pmf $\hat{\mathbf{p}} \in \mathcal{H}_\infty$ which has in the set $\mathcal{H}_\infty$ maximal value of relative entropy $H(\mathbf{p}, \mathbf{q})$ (or Shannon's entropy if $\mathbf{q}$ is uniform). The argument holds under linear as well as non-linear constraints (convex as well as non-convex feasible set $\mathcal{H}$).

Besides the two probabilistic arguments, several axiomatic approaches were developed, to support choice of Shannon's entropy (relative entropy) as the only objective function for solving the IPIP. The most widely accepted one (cf. [5]) requires assumption of linearity of the constraints. A non-axiomatic argument based on potential-probability density relationship and complementarity (cf. [10]) is restricted to the linear constraints as well. Also a game-theoretic view of MaxEnt (see [12]) seems to concentrate on the linear constraints.

Taken together: 1) for the linear constraints (or convex feasible set) case of the IPIP with $n \to \infty$ choice of other than the most-entropic distribution is not choice of the most-probable one (recall MaxProb), moreover it is choice of almost improbable distribution (recall the conditioned law). 2) For non-linear constraints – precisely: non-convex $\mathcal{H}$ – conditioned weak law of large numbers seems to be not yet studied[1], hence the large deviation argument seems to be missing (which is somehow in variation with [14]). Axiomatic grounding is at its best limited to linear constraints as well. For now, MaxProb argument in favor of Shannon's entropy seems to be the only one applicable also for the studied setting of non-convex feasible sets, non-linear constraints.

## 1.1. MaxProb/MaxEnt

For a reference, the MaxProb to MaxEnt convergence will be recalled. Let $\mathcal{H}_n$ comprises all types $\mathbf{n}$ which satisfy an adding-up constraint $\sum_{i=1}^m n_i = n$. Let $\mathcal{P} = \{\mathbf{p} : \sum_{i=1}^m p_i = 1\}$ be a set of all probability mass functions. Let $\hat{\mathbf{n}}$ be the most probable type from $\mathcal{H}_n$, which can come from $\mathbf{q}$. With other notation remained unchanged Thm 1 of [9] can be recalled:

*Theorem.* Let $\hat{\mathbf{n}}$ be the most probable type from $\mathcal{H}_n$, which can come from $\mathbf{q}$. Let $n \to \infty$. Then $\frac{\hat{\mathbf{n}}}{n} = \hat{\mathbf{p}}$, where $\hat{\mathbf{p}} = \arg\max_{\mathbf{p} \in \mathcal{P}} H(\mathbf{p}, \mathbf{q})$, and

[1] See also discussion at Postscript 1.

$H(\mathbf{p},\mathbf{q}) = -\sum_{i=1}^{m} p_i \log(p_i/q_i)$ is relative entropy of $\mathbf{p}$ with respect to $\mathbf{q}$.

The following Corollary expands its scope:

*Corollary.* The claim of Theorem remains valid also if some other differentiable constraint $F(\frac{\mathbf{n}}{n}) = 0$ is employed to form the working set $\mathcal{H}_n$ and corresponding constraint $F(\mathbf{p}) = 0$ is added to relative entropy maximization.

The MaxProb/MaxEnt theorem[2] gives a probabilistic justification to MaxEnt and also shows that the proper place of MaxEnt is at the asymptotic, only.

### 1.2. Aims

This research report brings results of study of a particular case of non-linear constraints (non-convex $\mathcal{H}$) which are known as 'frequency moment constraints'. Few examples, intended to illustrate the above mentioned convergence of the most probable type(s) to the most entropic pmf(s) are reported. MaxEnt pmf(s) are characterized. Characteristics of choice made by method of maximization of Rényi-Tsallis entropy (maxTent) are offered as well. Convergence of MaxProb to MaxEnt is illustrated also for $X$-frequency constraints, which are currently used at maxTent framework. Two particular cases of the constraints are presented, which allow to illustrate that maxTent-made choice of pmf is not only different than the most probable one, it is even improbable.

## 2. Frequency moment constraints

This study was triggered by an interesting paper by Romera et al. (cf. [18]) on frequency moment problem. There, links to statistical considerations of the frequency moments as well as to their applications in physics can be found.

In the simplest case of single frequency moment constraint, feasible set of distributions is defined as $\mathcal{H} = \{\mathbf{p} : \sum_{i=1}^{m} p_i^q - a = 0, \sum_{i=1}^{m} p_i - 1 = 0\}$, where $q, a \in \mathbf{R}$. If $m > 2$, the problem of choice of pmf becomes ill-posed. Note that the first constraint is non-linear in $\mathbf{p}$ and $\mathcal{H}$ is non-convex.

[2] See [9] for its proof, illustrative example under classic moments and further discussion. See also [11].

## 2.1. Characteristics of MaxEnt choice

### 2.1.1. *Analytic aspects*

The non-convexity of feasible set makes the problem of maximization of an objective function (even when it is concave, as it is the case of Shannon's entropy) analytically unsolvable (at least via standard Lagrange multiplier technique). Critical value of $p_i$ is expressed implicitly as: $p_i(\lambda) = k(\lambda)e^{-\lambda p_i}$, where $k(\lambda) = \sum e^{-\lambda p_i}$. The second derivatives does not allow for assessing whether maximum or minimum was attained at the critical point. Moreover, even if the critical point was the point of maximum, it could be of a local one, only.

### 2.1.2. *MaxProb/MaxEnt*

Using of MaxEnt for choosing a pmf from $\mathcal{H}$ is justified on the MaxProb ground. The convergence of MaxProb to MaxEnt (recall Sect 1.1) will be illustrated by the following Example.

**Example 2**: Let $q = 2$, $m = 3$ and $a = 0.42$ (the value was obtained for $p = [0.5\ 0.4\ 0.1]$.

For $n = 10, 30, 330, 1000, 2000$ the working sets $\mathcal{H}_n$ were constructed. For example, $\mathcal{H}_{10}$ contains $n = [5\ 4\ 1]$ and all its permutations (ie. [5 1 4], etc). This will be called group of types. $\mathcal{H}_{30}$ contains two groups: [15 12 3] and [17 8 5]. The last one (corresponding to frequencies [0.566 0.266 0.166]) has higher probability of coming from uniform prior generator. For $n = 330$ the feasible set comprises groups [0.0939 0.4333 0.4727], [0.5666 0.2666 0.1666], [0.1 0.4 0.5] and [0.1939 0.2333 0.5727], which has the highest probability of being generated.

For each $n$, among the feasible types, the most probable one which could come from a uniform prior generator was picked up. They are stated at the Table I.

Table 1. The most probable type, for growing $n$.

| $n$ | $\hat{n}_1/n$ | $\hat{n}_2/n$ | $\hat{n}_3/n$ |
|---|---|---|---|
| 10 | 0.1 | 0.4 | 0.5 |
| 30 | 0.166 | 0.266 | 0.566 |
| 330 | 0.1939 | 0.2333 | 0.5727 |
| 1000 | 0.1990 | 0.2280 | 0.5730 |
| 2000 | 0.2080 | 0.2185 | 0.5735 |
| $\hat{\mathbf{p}}$ | 0.2131 | 0.2131 | 0.5737 |

Clearly, the most probable type (hence also group of types) converges to a pmf (group of pmf's) which maximizes Shannon's entropy. $\diamond$

2.1.3. *Algebraic aspects*

Observe, that the most entropic pmf has two of probabilities equal. This can be elucidated by the following elementary considerations: let the feasible set be constrained further by additional requirement $p_1 = p_2 = p_3$. This makes $\mathbf{p}_0 = [1/3 \;\; 1/3 \;\; 1/3]$ the only pmf in working set. Clearly, the pmf is indeed in the set only if $a \equiv a_0 = 1/3$, ie. the 'centre of mass' of $\sum_{i=1}^{3} p_i^2$. If $a \neq a_0$ then $\mathbf{p}_0$ is not in $\mathcal{H}$, and the most entropic pmf should be searched among those pmf's which have two of probabilities equal; say $p_1 = p_2$ (hence $p_3 = 1 - 2p_1$).

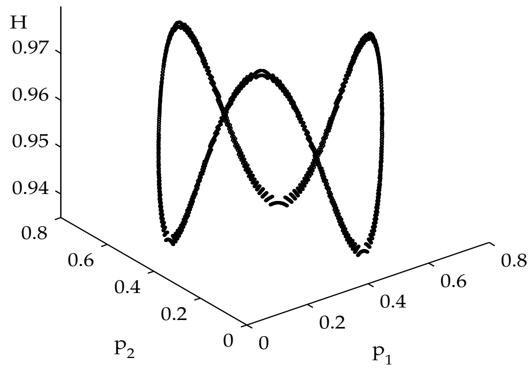


*Figure 1.* Feasible values of Shannon'e entropy for $m = 3$, $q = 2$, $a = 0.42$.

The additional requirement turns the under-determined conditions into a quadratic equation which is solved by either $p_1 = 0.2131$ or $p_1 = 0.4535$. Hence the restricted working set comprises two groups of pmf's [0.2131 0.2131 0.5737] and [0.4535 0.4535 0.0930]. The first type has Shannon's entropy $H_U = 0.9777$, the second $H_L = 0.9381$. It does not surprise that pmf's from the original set $\mathcal{H}$ (ie. those who may have all three probabilities different) have Shannon's entropy within the bounds which are set up by the $H_L$ and $H_U$.

Figure 1 depicts possible values of Shannon's entropy (and implicitly also feasible sets of pmf's) for the studied Example. The two groups of pmf's with $H_L$ and $H_U$ are easy to recognize.

This is obviously, not a property specific to the studied example with specific choice of $q = 2$ and $m = 3$. In general, the finding permits to state the following

*Lemma.* Let $\mathcal{H} = \{\mathbf{p} : \sum p_i^q - a = 0, \sum p_i - 1 = 0\}$, where $\mathbf{p} \in \mathbf{R}^m$ and $q \in \mathbf{Z}$. Let $m > q$. Then $\hat{\mathbf{p}} \in \mathcal{H}$ such that $H(\mathbf{p}) \leq H(\hat{\mathbf{p}})$ for any $\mathbf{p} \in \mathcal{H}$, is such that $\hat{p}_1 = \hat{p}_2 = \ldots = \hat{p}_{m-1}$, where $\hat{p}_1$ is one of solutions of the following algebraic equation: $(m-1)\hat{p}_1^q + (1-(m-1)\hat{p}_1)^q - a = 0$.

Notes:

1) Clearly, among the pmf's which are implied by solution of the equation, the $\hat{\mathbf{p}}$ is the one with highest value of $H$.

2) Similar claim can be made for $q$ such that $1/q \in \mathbf{Z}$.

3) The vector-point of maximum is not unique, any permutation od $\hat{\mathbf{p}}$ is as well point of maximum of Shannon's entropy on the set $\mathcal{H}$.

**Example 3**: Let $m = 3$, and $q = 3$. Let $a = 0.2440$. As the above Lemma suggests, $\hat{\mathbf{p}}$ should be one of solutions of the algebraic equation implied by $p_1 \equiv p_2$. Thus, it should solve $(1-2p_1)^3 + 2p_1^3 - a = 0$. Recalling Note 1, the most entropic pmf should be $\hat{\mathbf{p}} = [0.1939\ 0.1939\ 0.6122]$.

Table 2. The most probable type, for growing $n$.

| $n$ | $\hat{n}_1/n$ | $\hat{n}_2/n$ | $\hat{n}_3/n$ |
|---|---|---|---|
| 300 | 0.1 | 0.3 | 0.6 |
| 1000 | 0.1820 | 0.2060 | 0.6120 |
| $\hat{\mathbf{p}}$ | 0.1939 | 0.1939 | 0.6122 |

Analytical check of the claim (by constrained maximization techniques) is impossible (recall Sect. 2.1.1). Even numerically it appears to be unviable task (unlike the one of previous Example). MaxProb thus appears here also as a way of checking the Lemma. As the Table II shows, MaxProb type indeed converges to the pmf suggested by the Lemma. $\diamond$

### 2.1.4. *Summary of findings*

To sum up findings (analytical and those of the $q = 2$, $m = 3$ example) concerning general properties of pmf which MaxEnt chooses from frequency moment constraint based set recall that: 1) the feasible set is not convex, hence a local maximum (of whatever objective function, and if any at all) can be nothing else but a local maximum, 2) the MaxEnt distribution is not unique (ie. the entire group of all permutations of a distribution is 'point of maximum'.), 3) the MaxProb type is not unique: any permutation of pmf is also point of maximum, 4) $m - 1$ of MaxEnt pmf probabilities are equal; they are one of solutions of the implied algebraic equation, 5) hence obviously: the MaxEnt distribution is not the classic exponential and last but the most important: 6) it was illustrated that MaxProb group of types converges to MaxEnt group of pmf's.

### 2.2. Maximum Rényi-Tsallis entropy

Frequency moment constraints can be viewed as a special case of non-linear constraints which were introduced into statistical mechanics by Tsallis (cf. [19]). They will be discussed at Section 3, in detail. Tsallis' constraints (referred to as $X$-frequency constraints) are accompanied by a particular form of objective function which should be used for choice-making; usually it is called Tsallis' entropy:

$$H_T(\mathbf{p}) = \frac{1 - \sum_{i=1}^m p_i^q}{q-1}$$

where $q \in \mathbf{R}$, $q \neq 1$.

Maximization of Tsallis' entropy under Tsallis' $X$-frequency constraints satisfies the elementary requirement of backward compatibility: when the Tsallis' $X$-frequency constraints reduce to the classic linear moment constraints, the Tsallis' entropy reduces to Shannon's one (it happens for $q \to 1$).

Non-shannon-ian forms of entropies were around for long time. Some of them fall into category of convex statistical distances, and their mathematical properties are well-studied (cf. [15]). However, little seems to be known about their performance and probabilistic backing for use as criteria to solve IPIP. Rényi's enropy (see [17]):

$$H_R(\mathbf{p}) = \frac{1}{1-q} \log\left(\sum_{i=1}^m p_i^q\right)$$

is perhaps the most known among them. The Tsallis twist resurrected Rényi's entropy, since (as it is easy to see) from the perspective of maximization, any of them is a monotone transformation of the other one. Hence, Rényi's entropy attains maximum at the same point as Tsallis' entropy. (Another issue is that whilst the former is an additive entropy, the latter is non-additive, cf. [1], [13]).

Since the frequency constraints can be viewed as a special case of Tsallis' $X$-frequency constraints, it would be of interest to find, which of distributions from $\mathcal{H}$ is chosen by the method of maximization of Rényi-Tsallis entropy. This is not difficult to find: notice that the objective function is effectively from maximization point (ie. $\sum_{i=1}^m p_i^q$) identical with the frequency constraint: $\sum_{i=1}^m p_i^q - a = 0$. Hence, value of Tsallis entropy (as well as of Rényi, recall the monotonicity) is *the same* for any $\mathbf{p} \in \mathcal{H}$.

Hence, the 7-th point which has to be added to the Summary of findings on frequency constraint is, loosely speaking, that maxTent refuses to make a choice. Stated from other side, maxTent does not choose the most probable distribution.

### 2.3. A note on the most probable $a$

For a moment, consider the simplest of linear moment constraints: $\mathcal{H} = \{\mathbf{p} : \sum_{i=1}^{m} p_i - 1\}$, ie. the adding-up constraint, only. Obviously, for any $n = km$, $k \in \mathbf{Z}$, the most probable type which can come from uniform generator is $[n/m \;\; n/m \;\; \dots \;\; n/m]$. The same type is also the most probable in the case of frequency moment constraint, if $a = a_0 \equiv \sum_{i=1}^{m}(1/m)^q$.

A noteworthy property of the frequency moment constraint is that $a_0$ is not the most probable value which sum of squared frequencies can take up.

The following example illustrates the point.

**Example 4**: Let pmfs/types are defined on three-element support. Let $n = 30$. Few of the possible values of $\sum_{i=1}^{3}(n_i/n)^2$ are at the Table III, together with their probabilities $P$ of coming from uniform generator (ie. $P$ is sum of probabilities of all types which have the particular value of frequency moment).

Table 3. The $r$-th lowest value of $a$ and its probability.

| $r = ord(a)$ | $a$ | $P(a)$ |
|---|---|---|
| 1 | 1/3 | 0.0269 |
| 2 | 0.3356 | 0.1470 |
| 3 | 0.34 | 0.1214 |
| 4 | 0.3422 | 0.1103 |
| 5 | 0.3489 | 0.1651 |
| 6 | 0.3533 | 0.0679 |
| 7 | 0.36 | 0.0507 |

Figure 2 depicts the data for broader range.

The most probable value of frequency moment on random sample of size $n = 30$ which could come from uniform generator is $a = 0.3489$, and it is different than $a_0 = 1/3$. $\diamond$

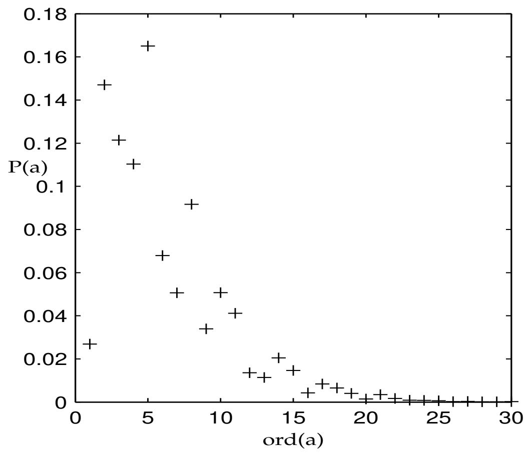


*Figure 2.* Probability of obtaining $r$-th smallest value of $a$

## 3. $X$-frequency moment constraints

*Setup.* Let, in addition, $\mathbf{q}$ be pmf of a random variable $X$ which can take up $m$ different values $\mathbf{x} = [x_1, x_2, \dots, x_m]$.

The first issue which has to be considered, is how to extend the frequency constraints. There were (at least) two ways suggested in the literature:

$$\begin{aligned} \mathcal{H}^A &= \left\{ \mathbf{p} : \sum_{i=1}^m p_i^q x_i - a = 0, \sum_{i=1}^m p_i - 1 = 0 \right\} \\ \mathcal{H}^B &= \left\{ \mathbf{p} : \sum_{i=1}^m p_i^q (x_i - b) = 0, \sum_{i=1}^m p_i - 1 = 0 \right\} \end{aligned}$$

The first form seems to be natural, from mathematical point of view, since it reduces to the frequency constraint when $\mathbf{x}$ approaches to $\mathbf{1}$. At the Tsallis' entropy framework $\mathcal{H}^A$ was superseded by $\mathcal{H}^B$ for physics reasons (see [21]). Here, both forms of $X$-frequency constraints will be considered.

### 3.1. Analytic side

A critical pmf (obtained by Lagrange multipliers technique) has the following form

$$p_i(\lambda) \propto \begin{cases} e^{-\lambda x_i p_i^{q-1}} & \text{for MaxEnt on } \mathcal{H}^A \\ e^{-\lambda (x_i - b) p_i^{q-1}} & \text{for MaxEnt on } \mathcal{H}^B \\ (1 + \lambda x_i (q-1))^{1/(1-q)} & \text{for maxTent on } \mathcal{H}^A \\ (1 + \lambda (x_i - b)(q-1))^{1/(1-q)} & \text{for maxTent on } \mathcal{H}^B \end{cases}$$

Note that for MaxEnt the form is an implicit one. The critical pmf can be either minimum or maximum, and local rather than the global one.

### 3.2. MaxEnt vs. maxTent

#### 3.2.1. *General setting*

MaxProb argument applies for the $X$-frequency constraints. The next two Examples are to illustrate the convergence.

**Example 5**: Let $m = 3$, $q = 2$, and let feasible set be the one of current use in maxTent: $\mathcal{H}^B$. Let $b = 1.447368$ (the value was obtained for $\mathbf{p} = [0.5\ 0.3\ 0.2]$).

The most entropic pmf (obtained numerically only) is $\hat{\mathbf{p}} = [0.50569\ 0.28065\ 0.21366]$. Value of Shannon's entropy at the maximum is $\hat{H} = 1.03116$.

The most Tsallis-entropic pmf (numerically obtained) is $\hat{\mathbf{p}}_T = [0.50235\ 0.29163\ 0.20601]$ with maxTent value 0.62015.

Table 4. The most probable type, for growing $n$.

| $n$ | $\hat{n}_1/n$ | $\hat{n}_2/n$ | $\hat{n}_3/n$ |
|---|---|---|---|
| 100 | 0.5000 | 0.3000 | 0.2000 |
| 1000 | 0.5050 | 0.2830 | 0.2120 |
| 2000 | 0.5055 | 0.2810 | 0.2135 |
| $\hat{\mathbf{p}}$ | 0.50529 | 0.28084 | 0.21387 |
| $\hat{\mathbf{p}}_T$ | 0.50194 | 0.29184 | 0.20622 |

Let $n = 100, 1000, 2000$. For each $n$, among the feasible types, the most probable one which could come from a uniform prior generator was picked up. The resulting most probable types are listed at the Table IV. It clearly demonstrates that as $n \to \infty$, the most probable type converges to a pmf which maximizes Shannon's entropy on the set $\mathcal{H}^B$. $\diamond$

**Example 6**: Let now $\mathcal{H}^A$ be investigated. Let $a = 0.55$ (corresponding to the same $\mathbf{p}$ as in the above Example).

The most Shannon-entropic pmf (obtained numerically) is [0.50216 0.28952 0.20832] with value of the entropy $H = 1.03156$.

maxTent-made choice is $\hat{\mathbf{p}}_T = [0.50122\ 0.29235\ 0.20644]$ with value of the entropy $H_T = 0.62070$. (Note that contrary to [16], the maximum is different than the one obtained under the previous constraints.)

Table 5. The most probable type, for growing $n$.

| $n$ | $\hat{n}_1/n$ | $\hat{n}_2/n$ | $\hat{n}_3/n$ |
|---|---|---|---|
| 100 | 0.5000 | 0.3000 | 0.2000 |
| 1000 | 0.5020 | 0.2890 | 0.2090 |
| $\hat{\mathbf{p}}$ | 0.50216 | 0.28952 | 0.20832 |
| $\hat{\mathbf{p}}_T$ | 0.50122 | 0.29235 | 0.20644 |

For $n = 100, 1000$ the most probable type (which could come from uniform $\mathbf{q}$) of $\mathcal{H}_n$ was chosen. Table V shows the convergence to the pmf which is chosen by MaxEnt from the corresponding $\mathcal{H}_\infty$ set. $\diamond$

### 3.2.2. *Two special cases*

Whilst the above Examples were intended to illustrate numerically the MaxProb/MaxEnt convergence for both kinds of $X$-frequency constraints, the next two examples are constructed in such a way, that they allow to show, that the most probable type becomes also asymptotically 'the only probable pmf'. Both the examples cover $\mathcal{H}^A$ and $\mathcal{H}^B$ form of constraints at once.

In the first Example $\mathcal{H}^A$ reduces by a choice of $\mathbf{x}$ into convex set, and there the conditioned weak law of large numbers applies, proving that types concentrates around the MaxEnt pmf as sample size gets big.

**Example 7**: With no loss for the point to be made, let us stick to $\mathcal{H}^A$. Let $m = 3$, $q = 2$ and $\mathbf{x} = [-1\ 0\ 1]$, and let $a = 0$ (obtained for instance from any pmf for which $p_1 = p_3$). Then the working set $\mathcal{H}^A = \{\mathbf{p} : \sum_{i=1}^3 x_i p_i^2 - a = 0, \sum_{i=1}^3 p_i - 1 = 0\}$ becomes $\mathcal{H}^A = \{\mathbf{p} : p_1^2 = p_3^2, \sum_{i=1}^3 p_i - 1 = 0\}$, which effectively reduces to $\mathcal{H}^A = \{\mathbf{p} : p_2 = 1 - 2p_1\}$ (since $\mathbf{p}$ is nonnegative). The set is convex. It allows to invoke the conditioned law of large numbers, which complements MaxProb argument, by claiming that types other than those which are close to the MaxEnt pmf becomes impossible as $n$ gets large. Stated from the point of maximum Rényi-Tsallis entropy method, the maxTent misses the most probable type, which here also becomes the only possible one as $n \to \infty$. $\diamond$

The second example illustrates the entropy concentration as well. There a non-convex set $\mathcal{H}^A$ (or $\mathcal{H}^B$, it does not make difference, here) is constructed in such a way that it consists of two types, only.

**Example 8**: To the previous Example setup add a requirement $\sum_{i=1}^{3} p_i^2 - c = 0$, which narrows the working set down to $\mathcal{H}^A = \{\mathbf{p} : p_1^2 = p_3^2, \sum_{i=1}^{m} p_i^2 - c = 0, \sum_{i=1}^{3} p_i - 1 = 0\}$. Though it is a system of 3 equations with 3 unknowns, the unknowns are not determined uniquely (a case which clearly cannot happen with traditional linear constraints) since $p_1$ (let's hang with it), is a solution of quadratic equation. For $c$ intentionally chosen to be 0.5 the two possible values of $p_1$ are 1/2 and 1/6. Hence $p_a = [0.5\ 0\ 0.5]$ and $p_b = [1/6\ \ 2/3\ \ 1/6]$ are the two possible pmf's at $\mathcal{H}^A$. Corresponding working set of types is obviously $\mathcal{H}_n = \{\mathbf{n} : 2n_1^2 + n_2^2 = 0.5n^2, 2n_1 + n_2 = n\}$, and it comprises the only two possible types: $[n_1/2\ \ 0\ \ n_1/2]$ and $[n_1/6\ \ 2n_1/3\ \ n_1/6]$, for any $n = 6k$. For $n = 6k$, with some $k \in \mathbf{Z}$, probabilities of generating the two possible types by a uniform generator are stated at the Table VI.

Table 6. The most probable type for growing $n$ becomes the only-possible.

| $n$ | $\hat{n}_1/n$ | $\hat{n}_2/n$ | $\hat{n}_3/n$ | $P(\mathbf{n})$ | $P(\mathbf{n}\vert\mathcal{H}_n)$ |
|---|---|---|---|---|---|
| 12 | 6 | 0 | 6 | 0.00173867 | 0.2373 |
| | 2 | 8 | 2 | 0.00558858 | 0.7627 |
| 120 | 60 | 0 | 60 | 5.376e-23 | 6.117e-9 |
| | 20 | 80 | 20 | 8.788e-15 | $\approx 1$ |
| 1200 | 600 | 0 | 600 | 1.129e-213 | 0 |
| | 200 | 800 | 200 | 3.771e-124 | 1 |
| $\hat{\mathbf{p}}$ | 1/6 | 2/3 | 1/6 | | |
| $\hat{\mathbf{p}}_T$ | both | | | | |

Clearly, the most probable type converges to the most entropic distribution. Moreover, as it is seen, the most probable type becomes the only possible one as $n$ gets large.

Hence, at this artificially constructed example (which nevertheless covers a meaningful setups) the most probable type converges to the most entropic pmf, which becomes the only possible pmf at the feasible set as $n$ gets large.

Value of Tsallis' entropy is the same for both feasible pmf's. Hence, maxTent fails to make a choice of the feasible set. $\diamond$

### 3.3. Summary

MaxProb justification of MaxEnt was illustrated for both forms of $X$-frequency constraints. Moreover, for two special cases of the constraints, it was shown that the MaxEnt-made choice of distribution for sufficiently large $n$ is identical with the asymptotically only-possible distribution. Since maxTent choice is different than the MaxEnt one it obviously implies that maxTent fails to choose the only-possible distribution.

Rather than summing up the summaries, it will be better to end the study by

## 4. Implications

which it *suggests* concerning application of the two methods for *modelling data*, in particular at Statistical Mechanics where $n$ is indeed large: there, MaxEnt models by the most probable, which is at the same time the only possible. maxTent models by impossible (which is, of course, possible :).

The findings about maxTent are only a by-product of the study. Its main aim was to illustrate generality of MaxProb method and to look at properties of MaxEnt at the non-linear/non-convex setting. As it was shown, there MaxEnt loses many of its usual properties[3], however it keeps being the asymptotic form of MaxProb.

## 5. Postscript 1: A conditioned law of large numbers

The standard form of conditioned weak law of large numbers (CWLLN) requires assumption of convexity of $\mathcal{H}$ and uniqueness of $\hat{\mathbf{p}}$ to state that types from $\mathcal{H}_n$ normalized by $n$ (ie. relative frequencies) converge in probability to the pmf $\hat{\mathbf{p}}$ which has maximal value of relative entropy in $\mathcal{H}$.

Though the same claim cannot be directly made for the case of non-convex $\mathcal{H}$ and non-unique $\hat{\mathbf{p}}$, an usual proof of CWLLN (see [4]) makes

[3] Point of maximum might be non-unique, recovered pmf is not exponential. It obviously does not mean that it cannot be brought into the canonical exponential form. Any vector of non-negative numbers which add up to one is MaxEnt canonical distribution, recall [3], Thm. 4.1.

it instantly possible to state a convergence in probability for the general setup in the following form:

*Theorem.* Let $\mathcal{H}_n$ be a non-empty set of types $\mathbf{n}$, which can be based of random samples of size $n$ from population $\mathbf{q}$. Let non-empty $\mathcal{H}_\infty$ be the corresponding set of pmf's and let $\hat{H}$ be such that $\hat{H} \geq H(\mathbf{p}, \mathbf{q})$ for any $\mathbf{p} \in \mathcal{H}$. Then for every $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr\left(\left|H\left(\frac{\mathbf{n}}{n}, \mathbf{q}\right) - \hat{H}\right| < \epsilon \mid \mathbf{n} \in \mathcal{H}_n, \mathbf{q}\right) = 1$$

To state it in words (it means: loosely): as $n$ gets large, it becomes virtually impossible that $\mathbf{q}$ will generate types, which have value of relative entropy different than $\hat{H}$. Perhaps, at this general setup (ie. with no non-trivial assumption about $\mathcal{H}$) no further claim than that of the above Theorem can be made. Being it so, or not, the Theorem already indicates that the choice which maxTent makes (under the frequency or $X$-frequency constraints) is choice of asymptotically impossible type.

## 6. Postscript 2: Bake yours own maxent

Ingredients which will be needed: yours 1) favorite distribution (or class of distributions) (ie. exponential family), and 2) preferred constraints (ie. mean consistency constraints).

Receipt, suggested at [2], is simple: make a backward engineering to find out what is the form which sought objective function should take, if its maximization under the preferred constraints has to lead to the favorite distribution. Then, check whether the function satisfies (at least some of) axioms of entropy. If yes, you are free to promote it everywhere.

**Example 9**: Let mean constraint $\sum_{i=1}^m p_i x_i = a$ be the preferred one, and let 'power law' $p_i \propto x_i^\alpha$ be the favorite distribution. Backward engineering leads to 'entropy' $H_\alpha = \sum_{i=1}^m p_i^{\frac{\alpha-1}{\alpha}}$.

## 7. Bibliographic note

Literature on Tsallis' maximum entropy method is vast (cf. [20]). Dissenting voices are rare (cf. [14], [8], [23]). LANL arXiv contains a series of preprints which document evolution of the method. Also, see March 2002 issue of Chaos, Solitons and Fractals.

## References


1. Aczél, J., "Measuring information beyond communication theory – Why some generalized information measures may be useful, others not," *Aequa Math*, **27**, 1-19, 1984.
2. Anteneodo, C. and Plastino, A. R., "Maximum entropy approach to stretched exponential probability distributions," *J. Phys. A* **32**, 1089, 1999.
3. Athreya, K. B. and Smith, J. D. H., "Canonical distributions and phase transitions," *Disc Math Prob Stat*, **20**, 167-176, 2000.
4. Cover, T. and Thomas, J., *Elements of Information Theory*, Wiley, 1991.
5. Csiszár, I., "Why least squares and Maximum Entropy. An axiomatic approach to linear inverse problems, " *Ann Stat*, **19**, 2032 -2066, 1991.
6. Csiszár, I., "Sanov property, generalized $I$-projection and a conditional limit theorem," *Ann Prob*, **12**, No. 3, 768-793, 1984.
7. Dembo, A and Zeitouni, O., *Large Deviations Techniques and Applications*, 2-nd ed., Springer, Application of Mathematics, vol. 38, 1998.
8. Gottvald, A., "Beyond the MaxEnt Principle: Bayes, Fourier, Zipf and Spirals in Metamorphoses," *Energy and Information in Non-linear Systems*, (A. Gottvald, ed.), 48-51, 2000.
9. Grendár, M., Jr. and Grendár, M., "What is the question that MaxEnt answers? A probabilistic interpretation," in *Bayesian inference and Maximum Entropy methods*, A. Mohammad-Djafari (ed.), 83-94, AIP, 2001.
10. Grendár, M., and Grendár, M., Jr., "Why Maximumu Entropy? A non-axiomatich approach," in *Bayesian inference and Maximum Entropy methods*, R. L. Fry (ed.), 375-379, AIP, 2002.
11. Grendár, M., Jr. and Grendár, M., "Maximum Entropy: clearing up mysteries," *Entropy*, **3**, 58-63, 2001.
12. Harremoës, P. and Topsøe, F., "Maximum Entropy fundamentals," *Entropy*, **3**, 191-226, 2001.
13. Jizba, P. and Arimitsu, T., "The world according to Rényi: thermodynamics of fractal systems," `arXiv:condmat/0108184`, Aug 2001.
14. La Cour, B. and Schieve, W. C., "A comment on the Tsallis maximum entropy principle," `arXiv:condmat/0009216`, Sep 2000.
15. Liese, F. and Vajda, I., *Convex Statistical Distances*, BSB B.G. Teubner Verlagsgesellschaft, Teubner - Texte zur Mathematik, Band 95, Leipzig 1987.
16. Martínez, S., Nicolás, F., Pennini, F. and Plastino, A., "Tsallis' entropy maximization procedure revisited," `arXiv:physics/0003098`, Dec 2000.
17. Rényi, A., "On measures of entropy and information," *Proc. 4th Berk. Symp. Math. Statist. and Probl.*, University of California Press, Vol. 1, 547-461, 1961.
18. Romera, E., Angulo, J. C. and Dehesa, J. S., "Reconstruction of a density from its entropic moments," *Bayesian Inference and Maximum Entropy Methods*, R. L. Fry (ed.), 449-457, AIP, 2002.
19. Tsallis, C., "Possible generalization of Boltzmann-Gibss statistics," *J Stat Phys*, **52**, 479, 1988.
20. Visit http://tsallis.cat.cbpf.br/biblio.htm for a bibliography on maxTent.
21. Tsallis, C., Mendes, R. S. and Plastino, A. R., "The role of cosntraints within generalized nonextensive statistics," Physica A, **261**, 534-554, 1998.
22. Vasicek, O. A., "A conditional law of large numbers," *Ann Prob*, **8**, 142-147, 1980.
23. Velázquez, L. and Guzmán, F., "Where the Tsallis statistic is valid?," `arXiw:condmat/0105378`, May 2001.

**Acknowledgements**

Hospitality of Banach Centre (BC) of the Institute of Mathematics of Polish Academy of Sciences, where a part of this study was performed as a part of the European Community Center of Excellence programme (package 'Information Theory and its Applications to Physics, Finance and Biology') is gratefully acknowledged. The work was also supported by the grant VEGA 1/7295/20 from the Scientific Grant Agency of the Slovak Republic. I wish to express my gratitude to Professors Feliks Przytycki, Ryszard Zieliński, Marek Męczarski, Tomasz Rychlik, Flemming Topsøe, Andrej Pázman, George Judge and to Madams Malgorzata Vielrose, Grażyna Pieścik-Bojarska and Elżbieta Szatan in regards to my stay at BC.