For the Sharma-Mittal entropy framework: What is the most useful question? What is the least useful question? How does this depend on the number of possible feature (answer) values, and the number of possible categories?

**Case 1: the number of possible answer values equals the number of possible category values.**

I’ll start with the easy part. It seems clear that (up to a possible tie), if you measure questions’ value with Sharma Mittal relevance, then you can’t do better than starting out in a state of maximal uncertainty, and ending in a state of zero uncertainty. (For whatever it’s worth, my simulations do reliably find this solution, various places in the Sharma Mittal space.)

What I mean is that no matter where you are in the Sharma Mittal space, you can’t do better (for the case of two answers and a binary category) than the following question:

|  |  |  |
| --- | --- | --- |
|  | k1 | k2 |
| q1 | 0.5 | 0 |
| q2 | 0 | 0.5 |

If there are three answers and a three-way category, then you can’t do better than the following:

|  |  |  |  |
| --- | --- | --- | --- |
|  | k1 | k2 | k3 |
| q1 | 1/3 | 0 | 0 |
| q2 | 0 | 1/3 | 0 |
| q3 | 0 | 0 | 1/3 |

… and so on.

Now, the above solutions are not unique, in general; for instance, for Hartley (ord=0,deg=1), the following question (and many others) tie for most useful, for the two-category, binary-feature case:

|  |  |  |
| --- | --- | --- |
|  | k1 | k2 |
| q1 | 0.96 | 0 |
| q2 | 0 | 04 |

**Case 2: the number of answer values is greater than the number of category values**

This case also turns out to be easy, and I think it makes sense. The solutions that my code finds, generally speaking, is to have maximal uncertainty in the category values (equal prior distribution across the categories), and to make it so that the category will be known with certainty once the question’s answer is known. However, there are many ways to achieve this, as the following example questions illustrate.

|  |  |  |
| --- | --- | --- |
|  | k1 | k2 |
| q1 | 0.50 | 0 |
| q2 | 0 | 0.50 |
| q3 | 0 | 0 |

|  |  |  |
| --- | --- | --- |
|  | k1 | k2 |
| q1 | 0.50 | 0 |
| q2 | 0 | 0.01 |
| q3 | 0 | 0.49 |

From an information-theoretic standpoint, note that the uncertainty in the answer is no longer reliably tied to the information gain in the question. That this must be so should be obvious because, relative to a fixed number of category values, if we allow many answer values, we can have arbitrarily high uncertainty in the question’s answer, without increasing uncertainty in the underlying category. Imagine a binary category, with 1000 equally probable answer values, in which there is maximal uncertainty about the category, and the category will be known for sure after the answer. In the example below, there are almost 10 bits of Shannon entropy in the outcome to the question, but only 1 bit of Shannon entropy in the true category; hence, even a completely definitive question has a lot more intrinsic uncertainty than the maximum possible information gain, which is the one bit of Shannon entropy in the true category.

|  |  |  |
| --- | --- | --- |
|  | k1 | k2 |
| q1 | 0.001 | 0 |
| … | … | … |
| q500 | 0.001 | 0 |
| q501 | 0 | 0.001 |
| … | … | … |
| q1000 | 0 | 0.001 |

To summarize, it turns out that there is no need to use more answer values than there are question values, to obtain a (non-unique) most useful question, in the case with more answer values than category values. We can just use the recipe from case 1, where the number of answer values and the number of category values is the same, and put zero probability in the answer values beyond (e.g.) the first two answer values, in the case of (e.g.) binary categories.

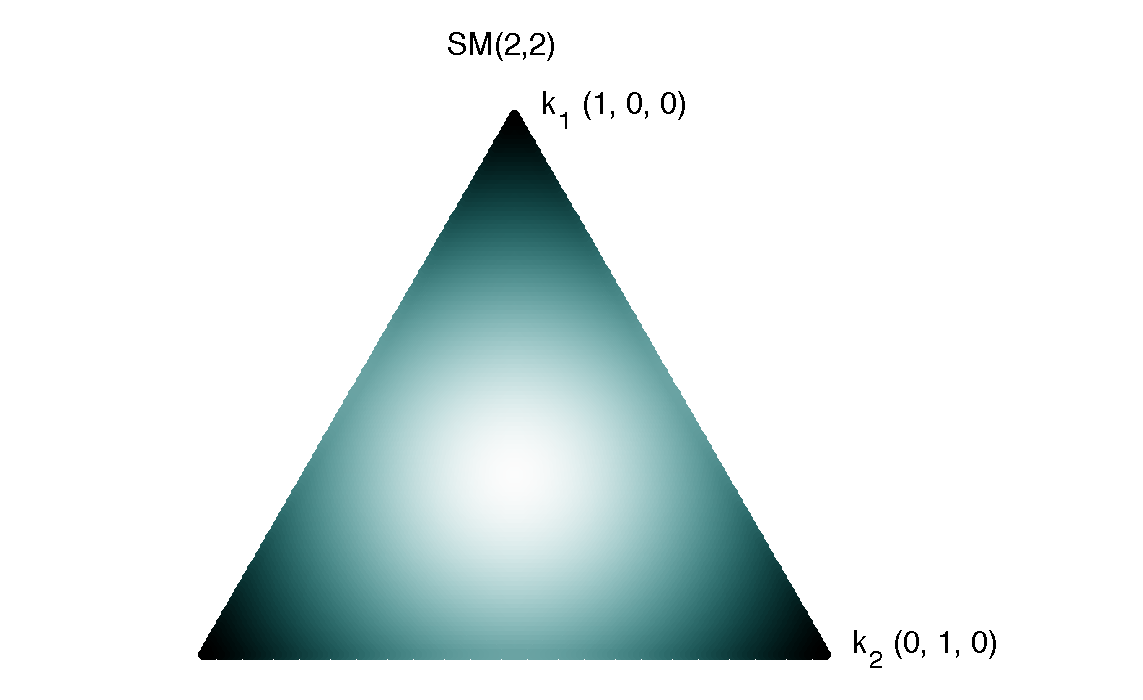
**Case 3: there are more category values than question values**

Suppose that the question Q={q1,q2} is binary but there is a three-valued category K={k1,k2,k3}. How do we populate the joint distribution of answer values and category values so as to have the most useful question?

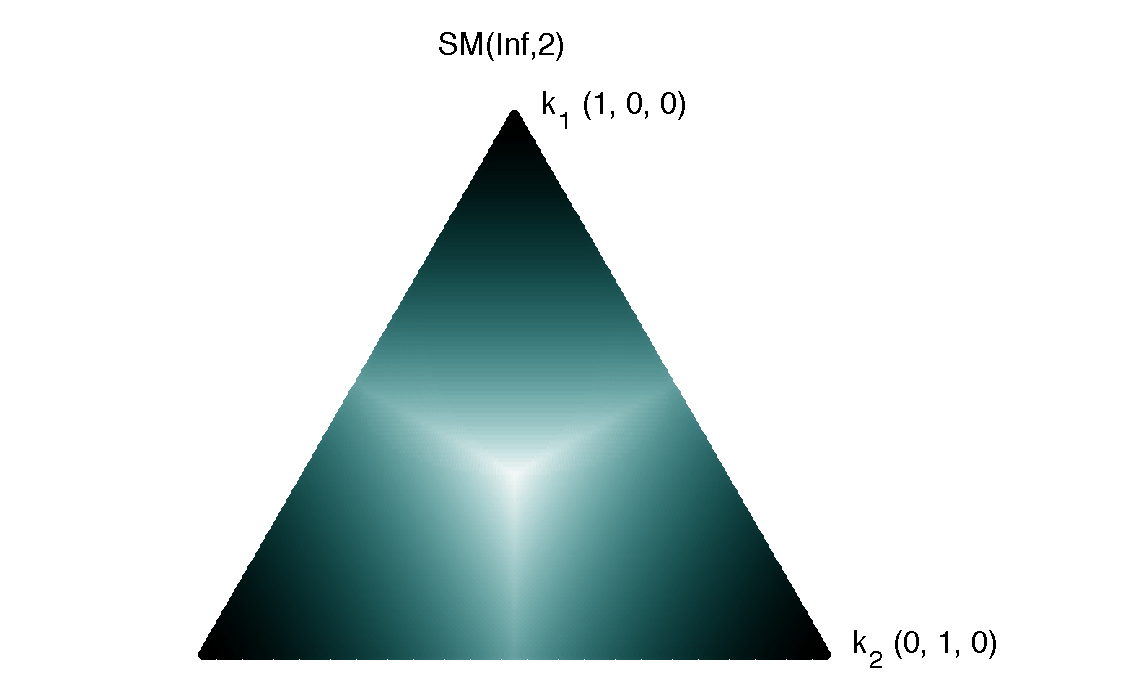
|  |  |  |  |
| --- | --- | --- | --- |
|  | k1 | k2 | k3 |
| q1 |  |  |  |
| q2 |  |  |  |

We might wish to have a question with maximal uncertainty in its outcome, or to have maximal uncertainty in the category values, or to have maximal chance of ending up with certainty given the answer to the question, etc. These **various objectives, importantly, cannot in general be achieved simultaneously**, at arbitrary places in the Sharma-Mittal space, and definitely not for the entropy metrics that we think have a shot at being plausible psychologically. Our standard entropy metrics have very different opinions about what the most useful question is in this scenario. I find it insightful to consider the solutions that are found by our standard entropy metrics (and others), and I illustrate them below. We might even consider putting this in the paper, where we introduce the standard entropy models.

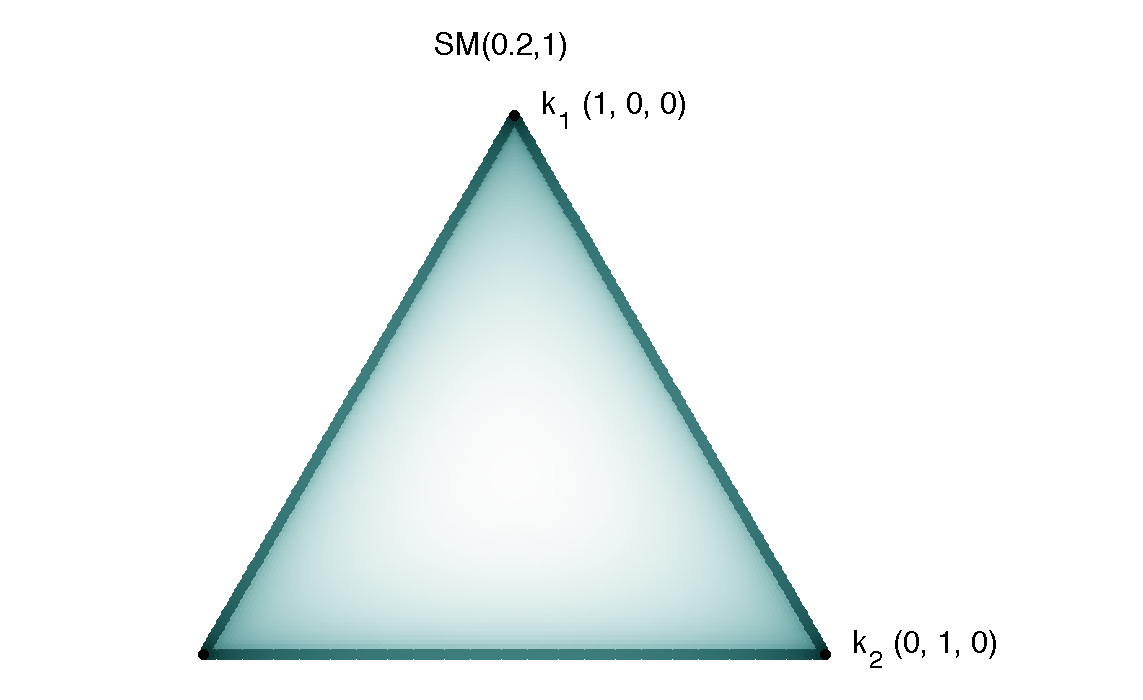
I will plot the questions on a **ternary plot, which I illustrate below**, using quadratic entropy. In this plot, white represents maximum entropy, and black represents zero entropy. [In every plot, I scale so that the largest entropy number corresponds to white, and the smallest entropy number corresponds to black.] Intermediate entropy values are varying shades of blue-green. Darker is always better. What we see is that if P(k1)=1, then there is zero entropy (top of triangle). Similarly if P(k2)=1 (bottom right of triangle) or if P(k3)=1 (bottom left of triangle), entropy is zero. Entropy is highest (as it is, up to a tie, for all SM entropies) if P(k1)=P(k2)=P(k3)=1/3, which is the middle of the triangle. Note that the middle of the triangle is not halfway from bottom to top. Rather, it is 1/3 of the way up from the bottom, in the left-to-right middle. This is the point that is equidistant from the three vertices.



We can contrast quadratic entropy with error entropy (below), and already get some insight about them. Visually, whereas quadratic entropy looks like it is made out of concentric circles (we should check if this is indeed true), error entropy looks like a triangle from some sci fi thriller, light emanating from the maximum entropy point, along the points where any two categories are equally probable, decreasing slightly as it goes to the edges. The upshot is that error entropy cares very much about whether you know what is most probable, and doesn’t care very much whether you know anything for sure.

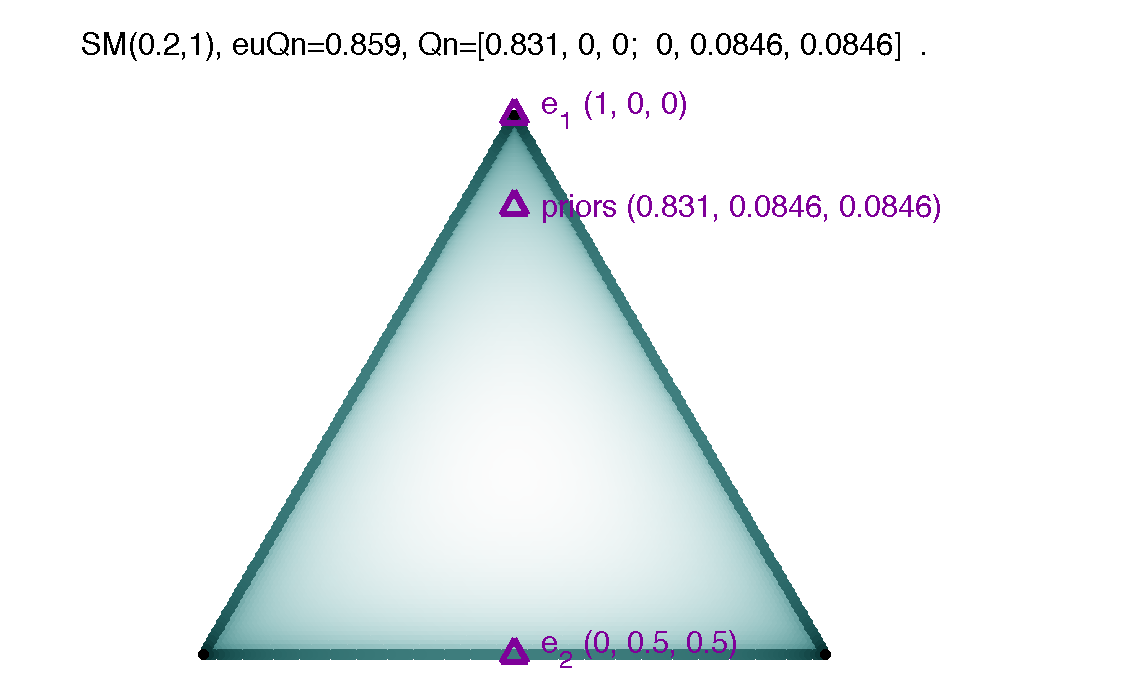


Finally, let’s consider SM(ord=0.2,deg=1), below. This is similar to Hartley entropy, but has more interesting continuous variation as you move among possible three-item probability distributions. This entropy cares only a small amount about which category is most probable. What is cares the most about is the number of categories that are possible: entropy is high if three categories are possible; entropy is lower (edges of triangle) if two categories are possible; entropy is zero (vertices—you have to look closely, but you can see they are black) if only one category is possible.



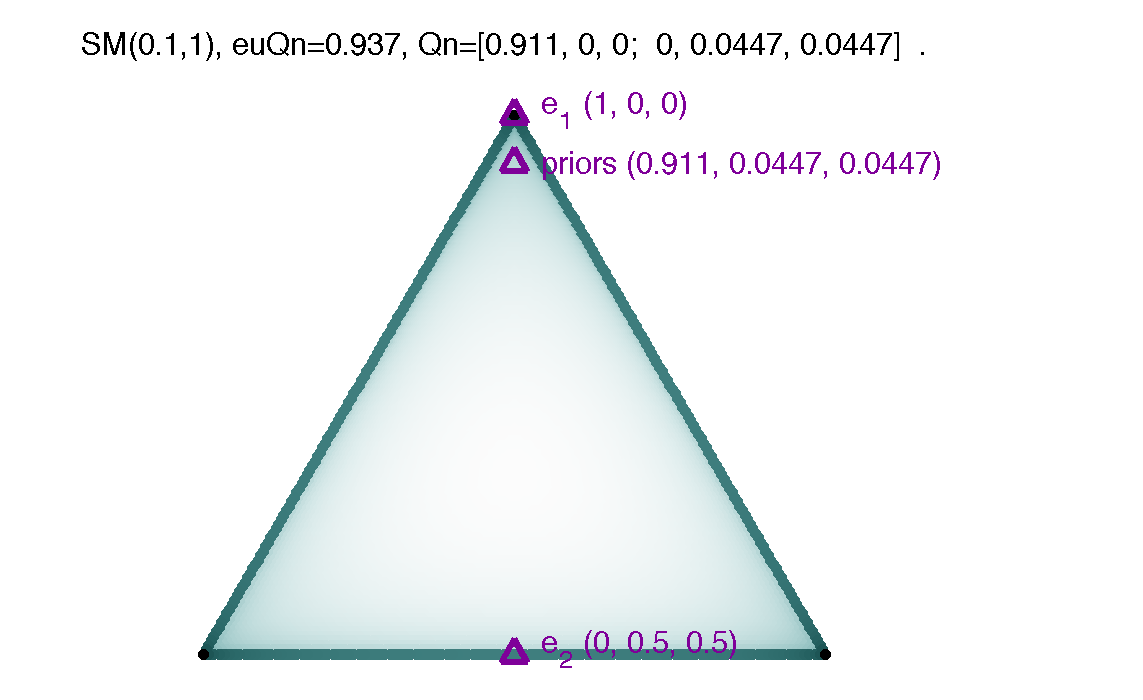
What is the most useful question, given SM(ord=0.2,deg=1), for binary features and a three-way category variable? I depict that below. Note that a binary question E={e1,e2} can be uniquely depicted by three points, which correspond to the prior probabilities, the probabilities if the question takes value e1, and the probabilities if the question takes value e2. According to this entropy metric (if we believe my code; there are no guarantees) starts with a prior distribution of about p(k1)=0.8310, p(k2)=0.0846, and p(k3)=0.0846. About 83% of the time, the answer is e1, and you know for sure that the category is k1. About 17% of the time, the answer is e2, k1 is eliminated, and k2 and k3 have equal posterior probability.

If you are like me, you may immediately object that this question is not very useful, because there is little information to be gained where the initial uncertainty is so low. But this is a question of viewpoint. From the perspective of error entropy, or Shannon entropy, the initial uncertainty is low. But from the perspective of SM(ord=0.2,deg=1), the uncertainty in the initial distribution is barely greater than maximal. Moreover, the chances of obtaining the e1 answer, which leads to zero uncertainty, are high for this question.

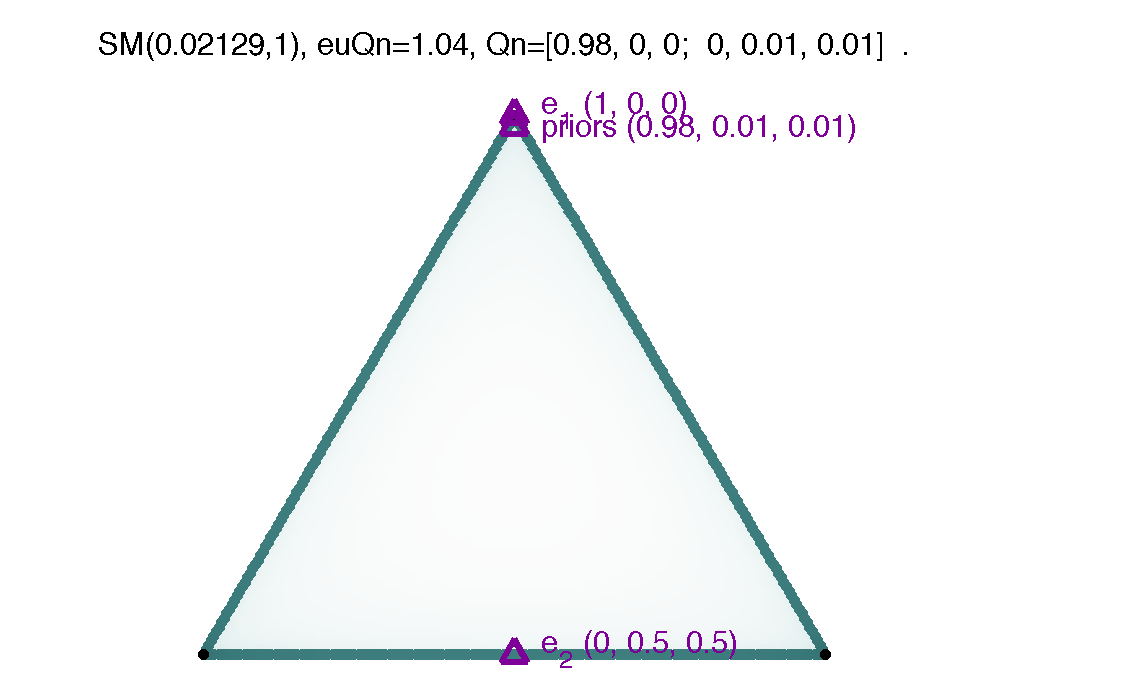


Note that I use the following unique representation of a question: the most probable outcome is e1, followed by e2. The most probable category is defined as k1, followed by k2, then k3. Of course, for symmetric entropy metrics this labeling is arbitrary; however, it helps when visualizing things.

Qualitatively similar findings obtain for other Renyi entropy metrics with low-but-greater-than-zero order. For instance, for SM(0.1, 1), the distributions conditional on e1 (1,0,0) and e2 (0,0.5,0.5) are the same, but the prior probabilities are closer to the situation where k1 has 100% prior probability.



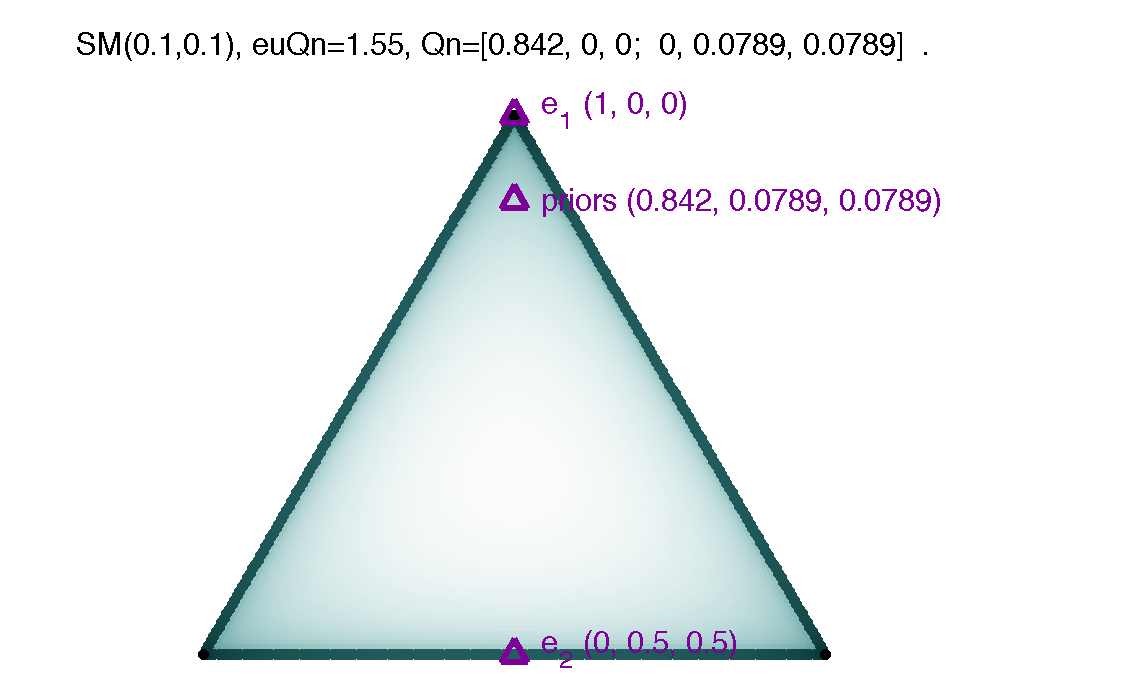
Now for the question I emailed a while back. Matthias solved it. Remember, I asked what the best binary question is, for SM(ord=0.02129, deg=1). To three decimal places, the best question rounds to the following. Probabilities conditional on e1 (1,0,0) and e2 (0,0.5,0.5) are the same as the other low-order Renyi entropies discussed above. However, the prior probability of k1 is even higher, at 98%.



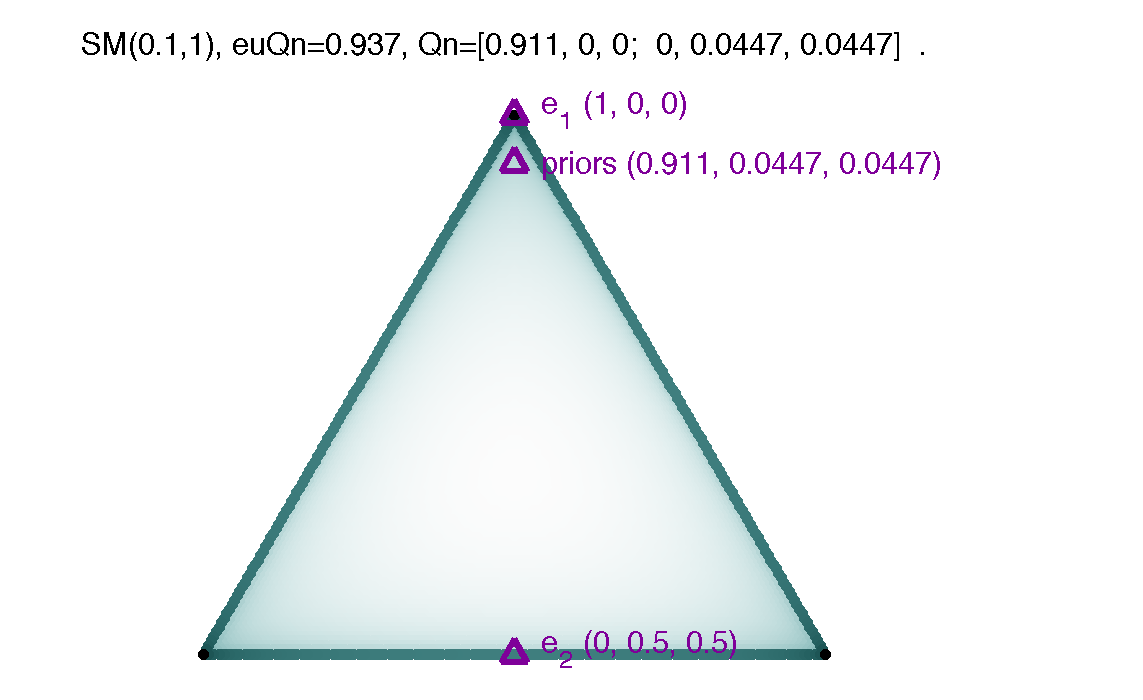
It is hard to visualize due to overlapping markers and text on the chart, but we can infer that for extremely low orders, the most useful question for SM(ord,1) is [(1-eps), 0, 0; 0 eps/2 eps/2], for some extremely small eps value.

**Understanding our reference entropy metrics, via the most (and on occasion, least) useful questions.**

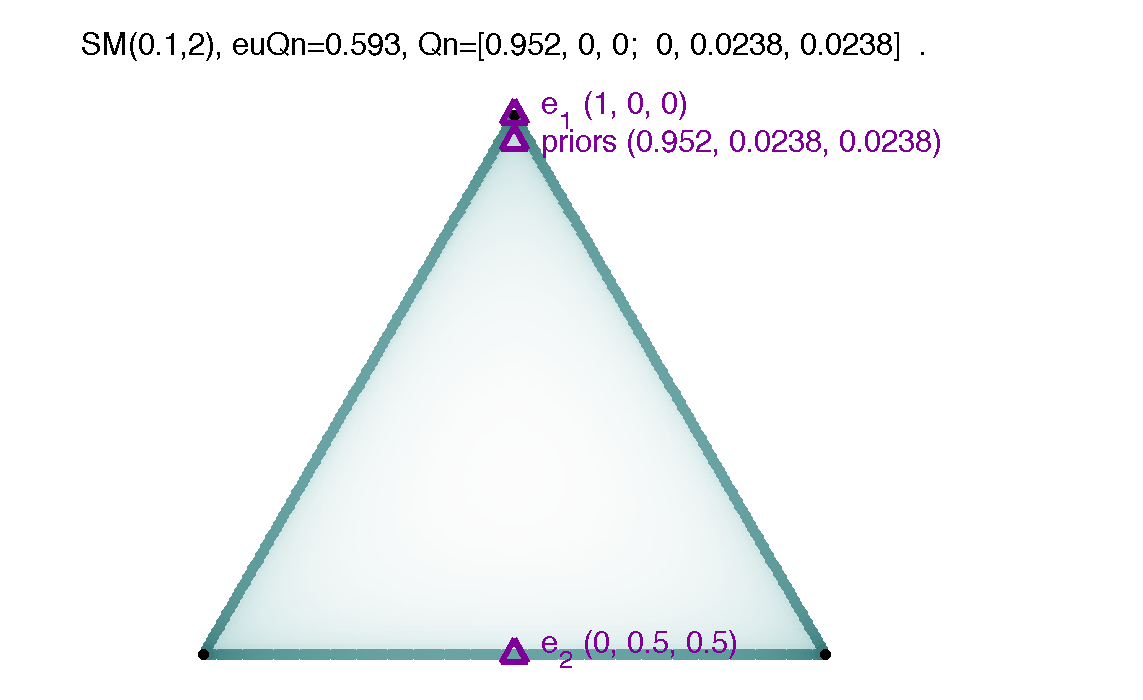
Approaching origin entropy. I show results for SM(0.1,0.1). The best question has similar form as the entropy metrics near to Hartley entropy. [I believe that the entropy metrics near the origin, and near Hartley, might diverge in cases with more than three categories.]



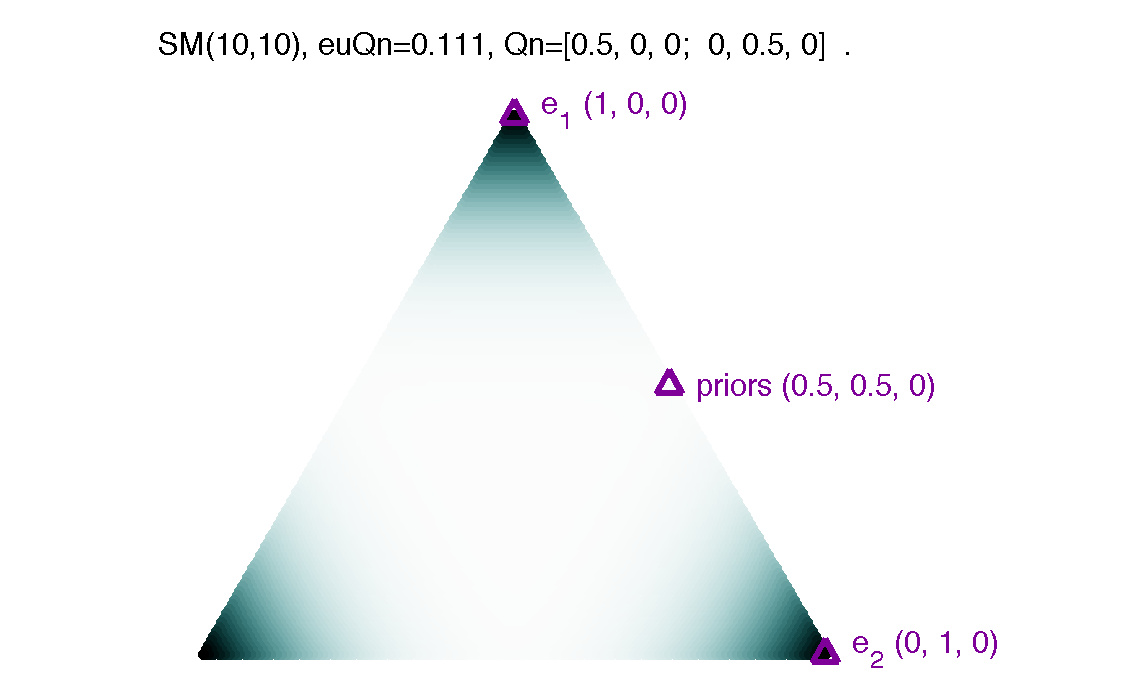
Almost Hartley entropy. I show SM(0.1,1), as from above:



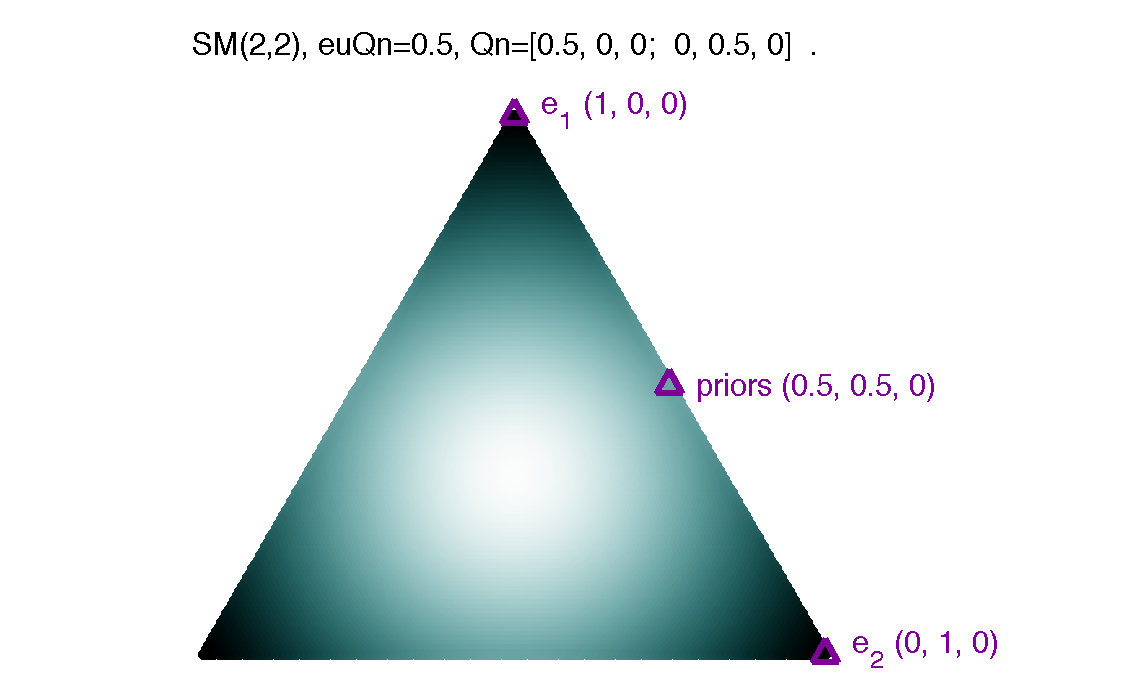
We can follow this up to SM(0.1,2), e.g. starting along the fractionalization entropy metrics, as well. It is the same overall type of question, but it starts from somewhat higher certainty as we increase the degree of the entropy metric.



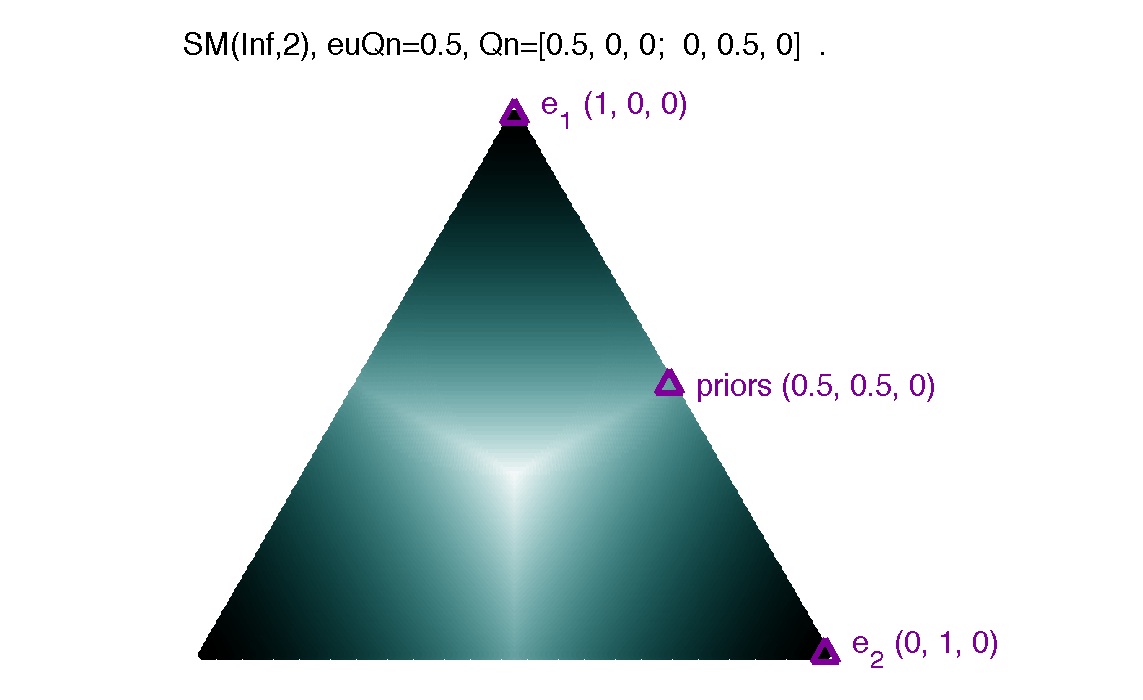
Tsallis(10) entropy. This question may come as a surprise, as the optimizations chose not to use k3 at all, even though they were allowed to. What’s going on? I think Tsallis(10) is not very sensitive to the exact probability distribution; rather, it mostly cares whether you know with certainty or not. Priors of (0,0.5,0.5) are close enough to maximal uncertainty. The upshot of this kind of question is that you can have certainty no matter what answer is obtained. Higher-degree (even some not-very-high-degree) entropy metrics prefer this question to a question that involves some chance of each of the three categories.



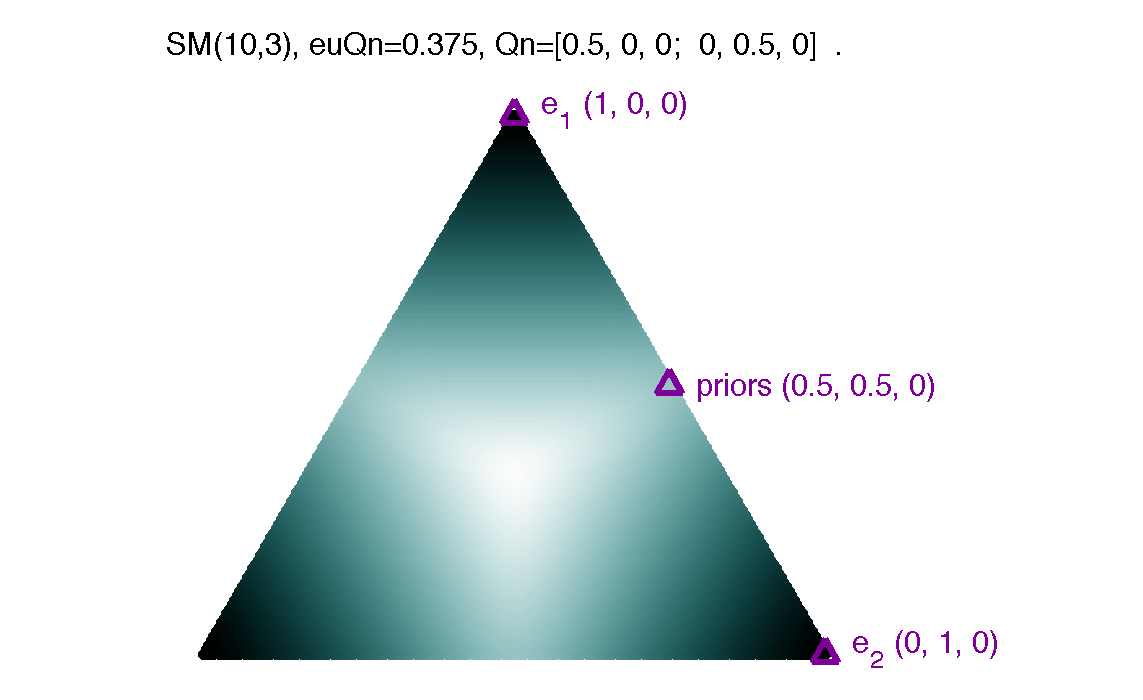
Quadratic entropy. Quadratic entropy, SM(2,2), is easy to describe; it prefers the same question as Tsallis(10).



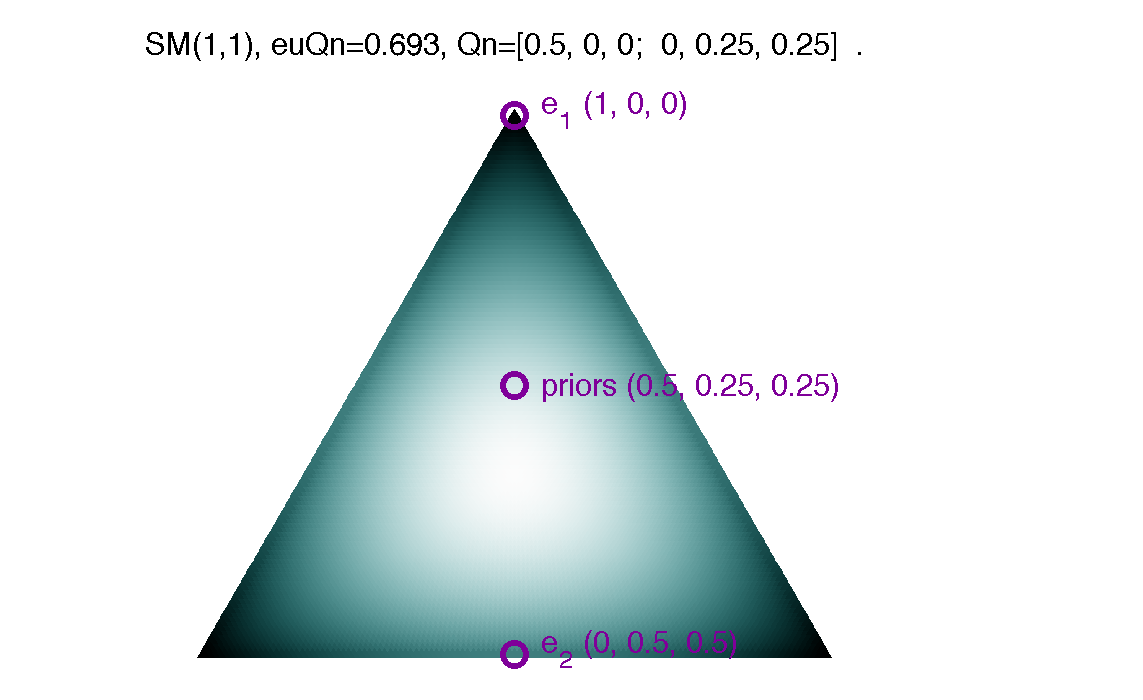
Error entropy. Error entropy, SM(inf,2), prefers the same question as Tsallis(10) and Quadratic entropy.

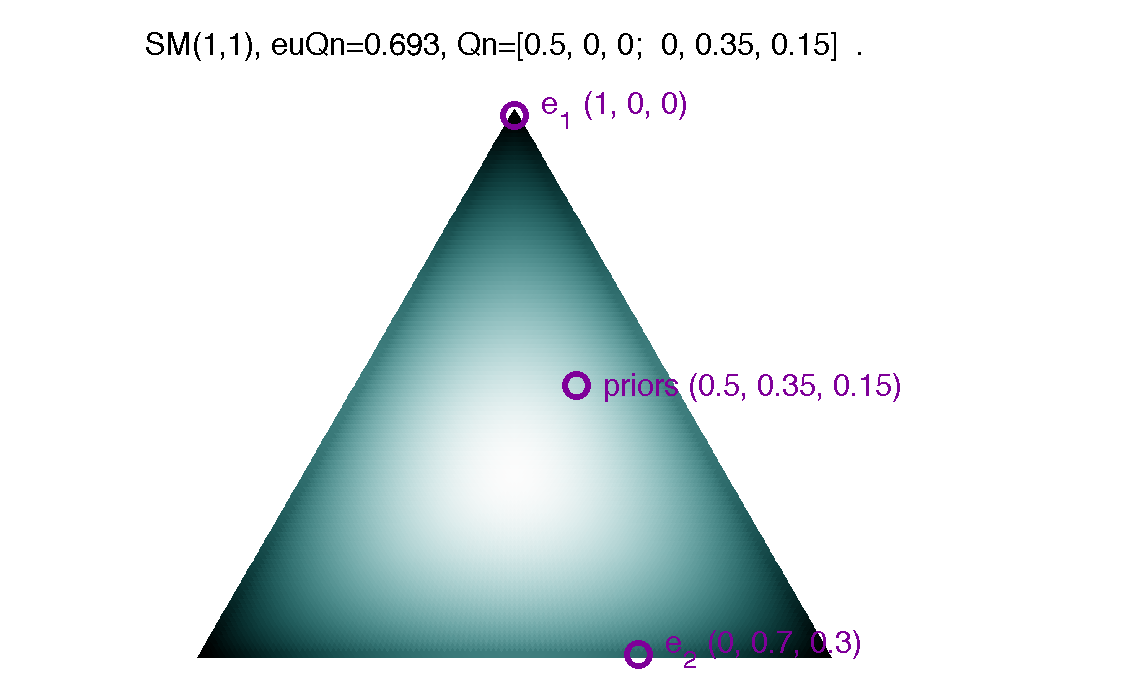


What about a moderate order, higher-than-Arimoto-degree, entropy metric? Here I show SM(10,3). This is an entropy metric that is plausible for describing what people do in words-and-numbers (planet Vuma)-type scenarios. It prefers the same question as the previous several metrics.

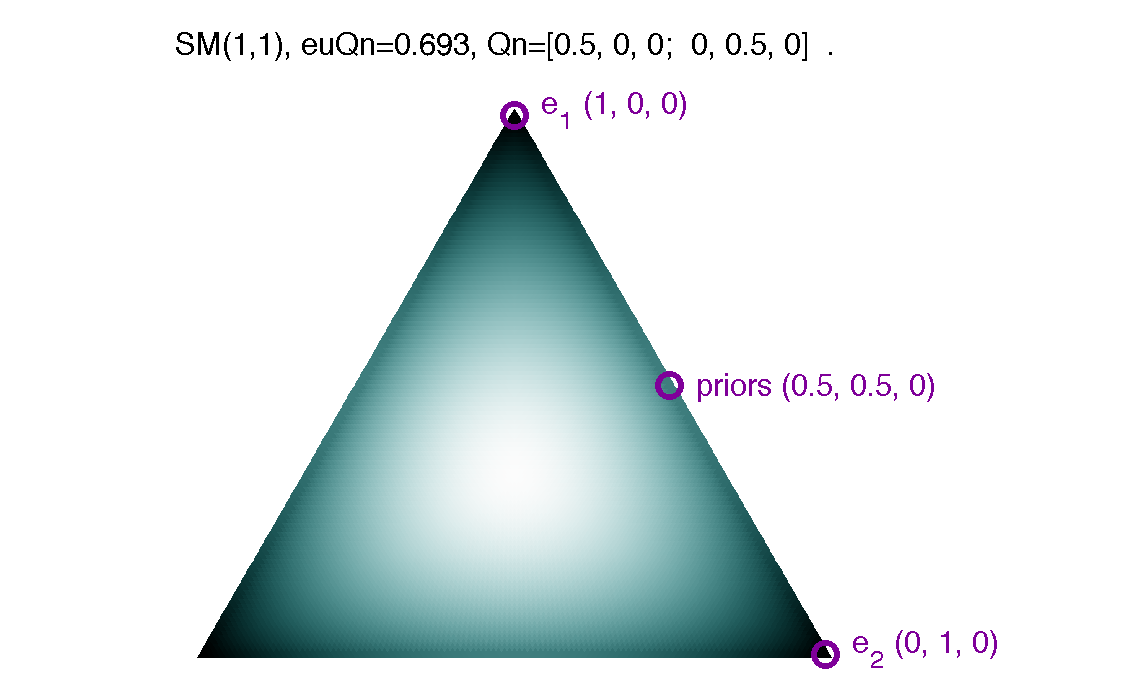


Shannon entropy. Shannon entropy does not uniquely prefer a single question. For instance, the questions [.5 0 0; 0 .25 .25], [.5 0 0; 0 .35 .15], and [.5 0 0; 0 0.5 0] all have value of one bit, i.e. ln(2) nats. What Shannon entropy seems to be saying is the following: give me a question where P(e1)=0.5, the category probabilities given e1 are (1,0,0), and where P(k1|e2)=0. How you assign 50% of the joint answer-category distribution among P(e2&k2) and P(e2&k3) is immaterial. You will get exactly 1 bit of information in any case, and that is the best that you can do. I illustrate these equivalent (tied for maximally useful) questions below.





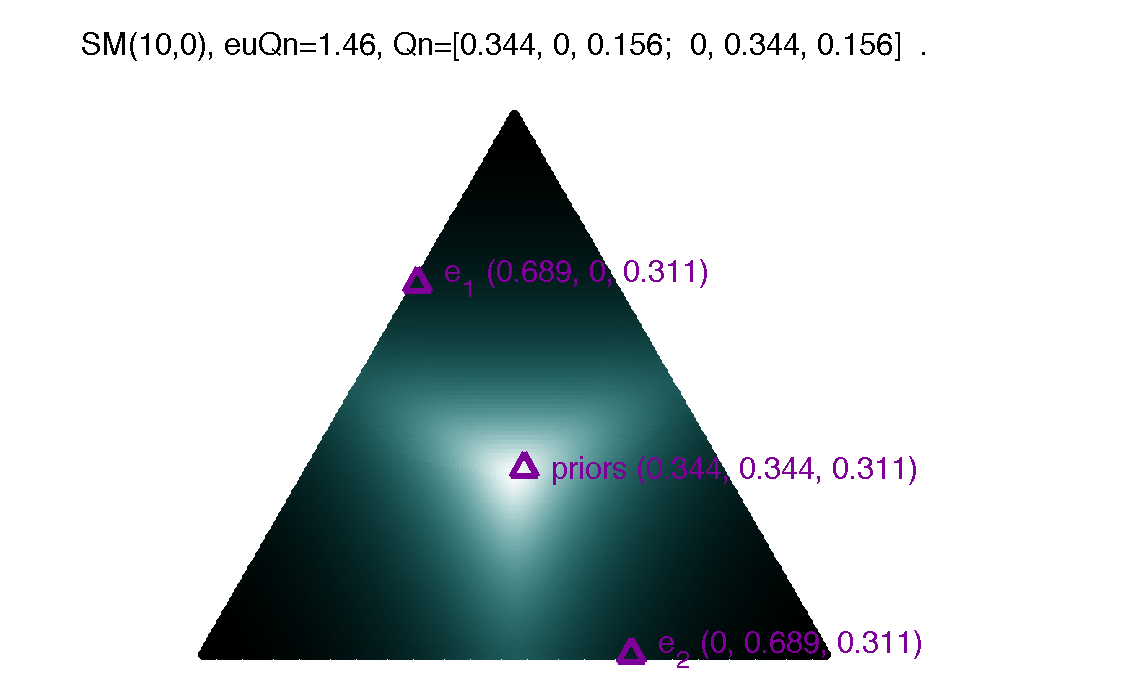
I believe that when we understand this, we will have a better idea about how Shannon entropy is unique (for better or worse) among the Sharma-Mittal entropy metrics.

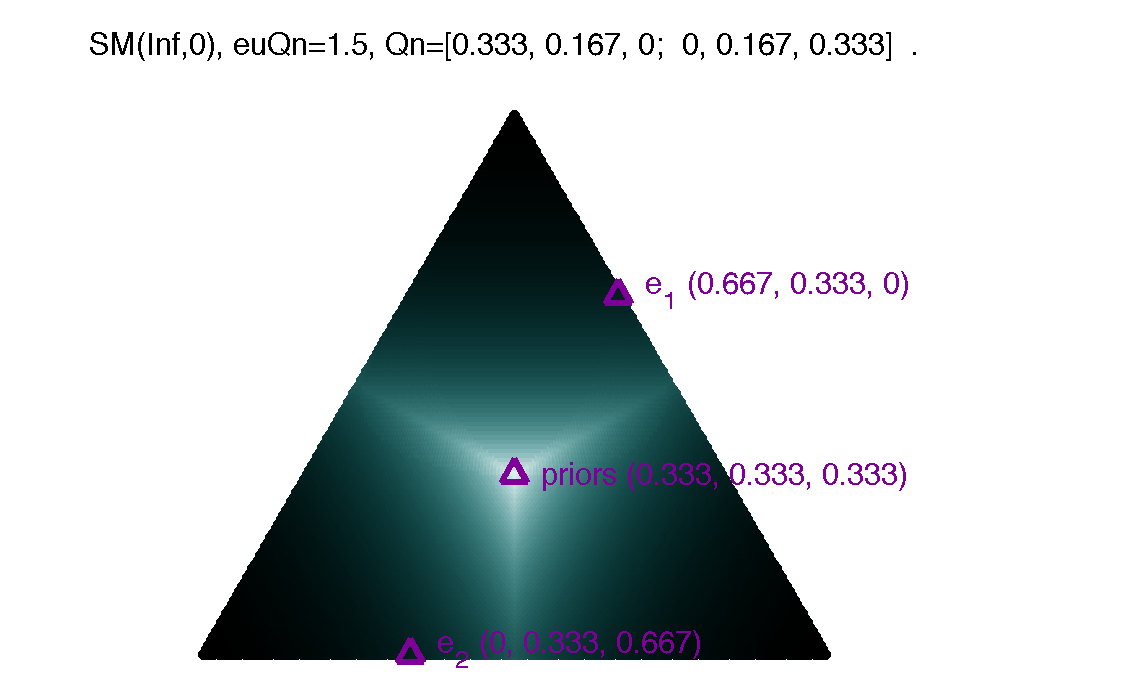


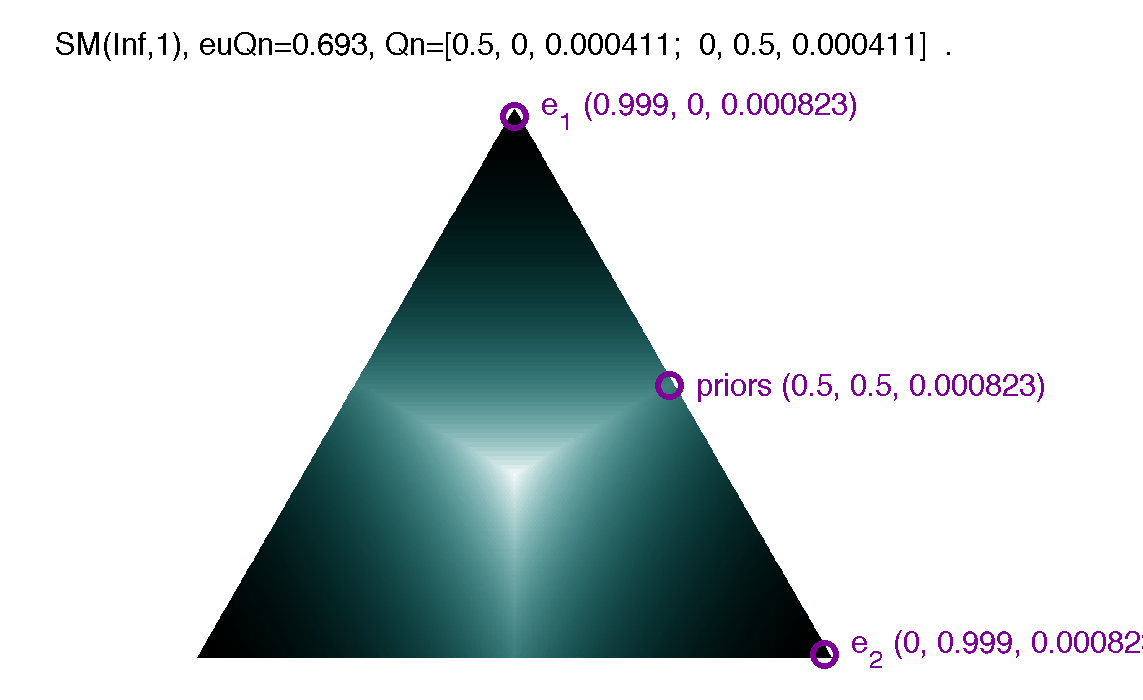
Near-to-Shannon entropy metrics. I ran my code for order={0.99, 1, 1.01} × degree={0.99, 1, 1.01}. What resulted? For degree=1.01, the optimizations consistently found [0.5 0 0; 0 0.5 0] as the best question. For degree=0.99, the precise numeric values varied, but the form was always [k 0 0; 0 (1-k)/2 (1-k)/2], where k decreased with increasing order. For order=0.99, k=0.501; for order=1.00, k=0.499; for order=1.01, k=0.497. For degree 1 (Renyi near Shannon), things are interesting: for SM(0.99,1), the best question was [0.502 0 0; 0 0.249 0.249]. Note that the pattern of questions from Hartley entropy, with order increasing up to but not including Shannon entropy, follow a similar pattern. Shannon entropy itself appears to be unique in (at least this region of) the Sharma-Mittal space, in not having a unique best question, but only the apparent requirement that the question be of the form [0.5 0 0; 0 x x]. For Renyi entropy with order slightly greater than 1, questions typically followed the form [0.50 0 0; 0 0.5 0], but in some instances optimizations found a question of the form [0.5 eps 0; 0 0.5 0]. It may be that the question with a small eps (e.g. around 1^-10) in it is actually a bit better; this remains to be verified. See results for SM(inf,1), the maximum order Renyi entropy, as well (below).

The bottom line is that two entropy metrics can be visually indistinguishable, but have very different implications for what questions are deemed most useful. One extreme case is if we fix order=1.00, and change degree from 0.99 to 1.00 to 1.01. For SM(1,0.99), it seems that [0 0.251 0.251; 0.499 0 0] is the unique best question; for SM(1,1) the question found varies, but always follows the form [0.5 0 0; 0 x x]; for SM(1,1.01) the best question is [0.5 0 0; 0 0.5 0].

Nonconcave entropies. What about the highly nonconcave entropy metrics? Here we consider the best (and worst) questions found by SM(10,0), SM(inf,0), and SM(inf,1).

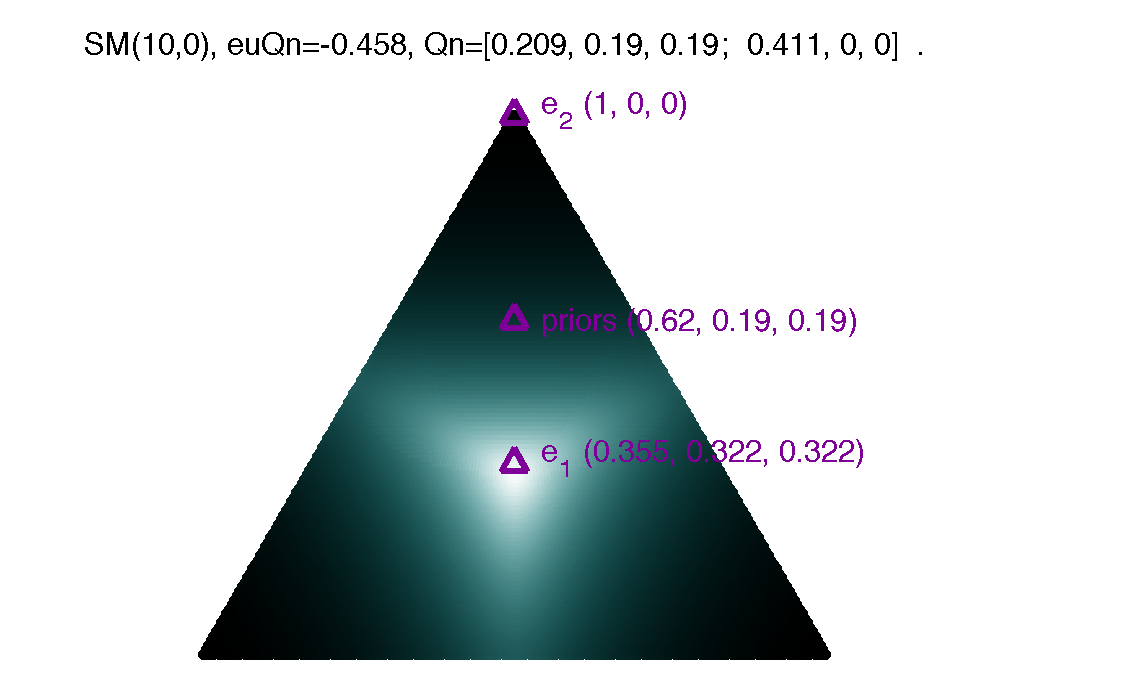






The worst questions for the nonconcave entropy metrics. For entropy metrics on or above the Arimoto curve, no questions have negative expected information gain (relevance). However, especially for the strongly nonconcave (degree zero, high order) metrics, some questions have negative expected information gain, e.g. are less useful than asking an uninformative question. The least useful questions, e.g. with the most strongly negative relevance, that my optimizations found, are given for SM(10,0), SM(inf,0), and SM(inf,1), below.

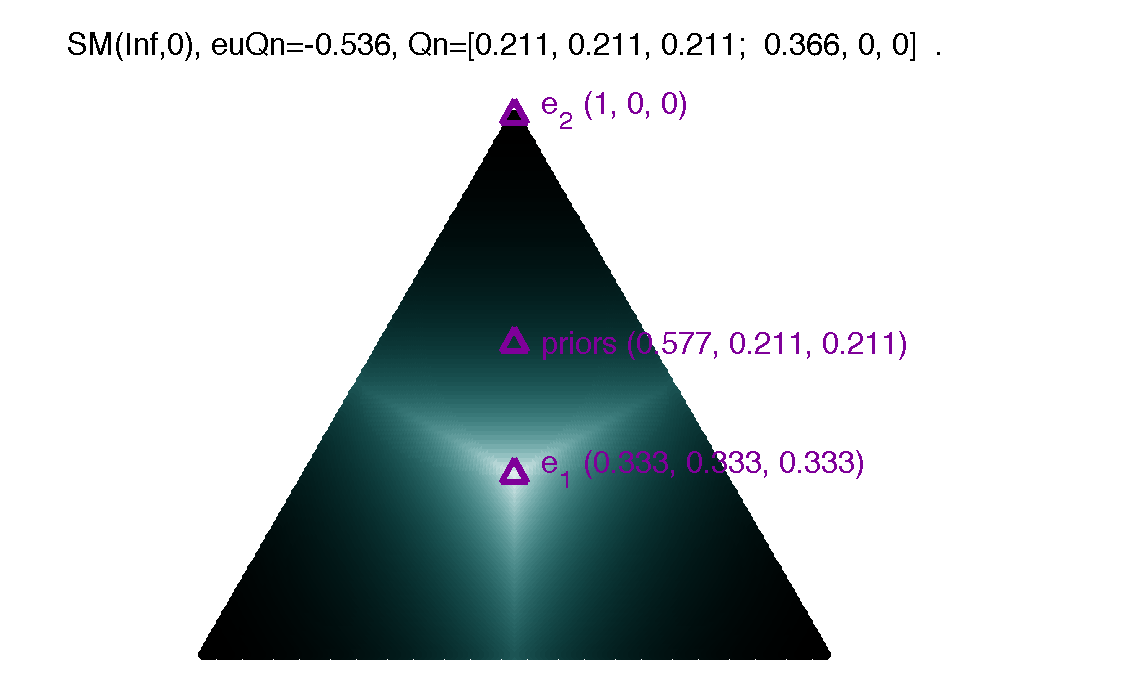
SM(10,0) thinks the worst question ends near to maximal uncertainty most of the time, with certainty about 41% of the time:



SM(inf,0) gives the most strongly negative expected information gain question that I have seen. Interestingly, the most frequent outcome is maximal uncertainty; this occurs about 63% of the time. Otherwise the question gives certainty. I think this is the only case in the binary-feature, three-way-category optimizations that I have found where maximal uncertainty across the three categories is obtained in one of the answers. I think this case is important so I paste the full (non-rounded) numbers, as well:

Qn=[0.21132486754265,0.21132486754265,0.21132486754265; 0.36602539737206,0,0]

R^SM(inf,0) Qn = -0.53589838486225.



If we look at the corresponding most nonconcave Renyi entropy, SM(inf,1), the scenario is quite close to that of SM(inf,0), but the negativity is much less extreme.

