On the Applicability of the 'Number of Possible States' Argument in Multi–Expert Reasoning

Martin Adamčík*

Martin de Tours School of Management and Economics, Assumption University, 10540 Samut Prakan, Thailand

Abstract

The aim of this paper is to explore the applicability of the 'number of possible states' argument in inferential problems in multi–expert reasoning. The argument is Bayesian and it is similar in spirit to the one used to derive the maximum entropy inference process. Under certain conditions a particular way of applying it surprisingly suggests that the weighted arithmetic mean should be used in meta–analysis with unexplained heterogeneity.

Keywords: Kullback–Leibler divergence, discrete probability function, entropy, multi–expert reasoning, meta–analysis, inferential statistics

2010 MSC: 05A20, 52A20

1. Introduction

1.1. Bayesian inference

The origins of the problem which we address in this paper go back to 1713 and to the posthumous publication [1] of J. Bernoulli. While the concept of probability had been established as the number of favourable cases over the total number of equally probable cases due to the contemporary popularity of gambling, Bernoulli asked how the concept of probability could be applied in nature, medicine and science. The assumption that there is a population from which we can choose members with the same likelihood appeared to be

*Corresponding author

Email address: maths38@gmail.com (Martin Adamčík)

only wishful thinking considering that, for example in medicine, we may never determine things such as the number of diseases. So the conclusion was that the probability must be somehow indirectly inferred from observations [2, Chapter A].

This problem of inferring some theoretical probability was historically called the problem of inverse probabilities but currently it is known as inferential statistics. Meta–analysis is an example of inferential statistics which in general merges partitions observed by several studies investigating the same problem [3, Chapter 9]. One simple way how to perform a meta–analysis of several studies where each has observed a partition on a sample is to merge those partitions by the weighted arithmetic mean with weights being the corresponding sample sizes. This inferential solutions is due to sampling theorists and frequentists (or as we call them now statisticians) who consider probabilities as measurable frequencies. In this approach, developed by such major figures as Fisher and Pearson, a theoretical probability such as the weighted arithmetic mean mentioned above is assumed and then assessed if the deviation of the frequency measured in a sample from this theoretical probability could be due to chance or if there is some underlying process that needs to be explained.

Now, it is widely accepted that any inferential solution is wrong if there are some underlying processes affecting the observations; either because the research objectives of studies are different or because they investigate the same objective by dissimilar methods [4, Chapter 7]. On the other hand, there are often similar studies that produce observations that differ in a manner that is unlikely to be only due to chance. This is called statistically detected heterogeneity. In this paper statistically detected heterogeneity that has no obvious explanation will be called simply unexplained heterogeneity.

A number of statistical methods that are jointly called random–effect meta–analysis were developed to deal with unexplained heterogeneity [3, Chapter 12]. These methods assume that various studies are in effect observing different populations due to discrepancies being caused by unknown factors operating in the process of observation. If a small study differs significantly from larger

studies its weight relatively to other studies is increased. In extreme cases the weighting due to sample sizes vanishes and we are left with the un–weighted arithmetic mean. Nevertheless, this frequentist approach has been criticised. As was argued in [5], the disagreement on its own should not mean that smaller studies more accurately represent the spectrum of measured populations than a single large study.

In this paper, we consider statistical inference from a more traditional point of view than is the popular frequentist approach mentioned above. It was Thomas Bayes who in 1763 was the first to answer Bernoulli's problem although it was Pierre–Simon Laplace who later gave a rigorous formulation on how an equiprobability assumption on some prior states can generate an inferred probability. This is not a new idea. The Bayesian approach to meta–analysis [3, Chapters 8 and 16] is currently being intensively studied and advocated as a tool of inference, nevertheless in the present context the author is not aware of any work deriving the foundations of some particular method of inference directly from first principles. And that basic argument used in this paper once led us to the much celebrated maximum entropy principle: the 'number of possible states' argument.

By applying the 'number of possible states' argument to a specific interpretation of meta–analysis as drawing with replacement we show in particular that the weighted arithmetic mean in the case of unexplained heterogeneity actually provides the overwhelmingly most likely estimate of the true partition in the population under certain conditions. In other words, surprisingly the same inferential approach should be used regardless of whether or not statistical heterogeneity is present in our specific modelling of meta–analysis. Considering the increasing importance of meta–analysis particularly in medicine, this result could lead to reconsideration of the current statistical approach to unexplained heterogeneity.

1.2. Entropy

As we have already mentioned, our main tool in obtaining the result above

is related to the notion of entropy. Entropy was first developed for thermodynamics in the nineteenth century as a measure of disorder. In 1871 Boltzmann [6, 7, 8] gave entropy a 'number of possible states' basis by quantifying the number of micro–states of molecules of a gas that lead to a single thermodynamic macro–state and thus showed the maximum entropy principle to be a powerful tool which provides simple answers to seemingly difficult problems. E. T. Jaynes [9] informally described the usefulness of this tool for solving problems as follows.

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

In other words, if we can guess all the relevant processes operating in an experiment then we can safely ignore many others that make the problem difficult and entropy will simply provides us with the macro–state that is overwhelmingly the most likely to be observed. For example, in 1948 Shannon [10] trying to encode messages in English into binary digits in the most efficient way suggested that we do not need to consider all statistical properties of English in his inferential problem:

> "... we may choose to use some of our statistical knowledge of English in constructing a code, but not all of it. In such a case we consider the source with the maximum entropy subject to the statistical conditions we wish to retain. The entropy of this source determines the channel capacity which is necessary and sufficient."

A similar approach was adopted in 1989 by Paris and Vencovská [11]. In the specific setting of the uncertain reasoning as defined in [12] they contemplated how an expert should determine a partition on a set of examples given there are only several specific constraints on possible partitions available to the expert. The set of examples here refers to an abstract population; for example if the

expert is a physician it could consist of patients and partitions could consist of groups of patients with mutually exclusive and exhaustive sets of symptoms. For Paris and Vencovská the examples were however not real; they were only a means of describing the state of knowledge of the expert, only a way of representing his or her experience.

Paris and Vencovská then assumed that everything relevant to the problem of determining the true partition is given by the constraints on possible partitions and any additional processes that may affect what the partition could be were ignored. The resulting 'number of possible states' argument gives the most entropic solution subject only to these constraints. This has justified a specific inference process of uncertain reasoning called the maximum entropy inference process. Later Wilmers [13, Section 5.2] asked a natural question as to whether there is a similar argument justifying any inference process in a multi–expert context.

1.3. Our result

In this paper we extend the setting of examples considered in [11] to a context where a large number of examples were individually observed by several experts or studies, however the proportion of collectively observed examples is insignificant in comparison to the overall number of existing examples. Unlike Paris and Vencovská we consider these examples as more tangible quantities, for example they could be patients investigated in some medical studies. The above mentioned process of observation is modelled by drawing with replacement. Furthermore, we assume that there are unknown processes operating in the process of observation that make the observations reported by each expert (or study) in the form of admissible partitions more different from each other than expected from the statistical sampling theory (i.e. there is statistical heterogeneity). Given that we do not know those processes we aim to employ the 'number of possible states' argument subject only to the reported observations. In the spirit of Bayesian inference, we adopt an equiprobability assumption concerning the process of observation and determine an inferred partition by

the 'number of possible states' argument rather than assuming a theoretical partition as in a frequentist approach.

We will show that if we count the states similarly as in [11] then an overwhelming number of states generate the uniform partition on the set of all examples and a partition that is generated by the process which combines the weighted arithmetic mean with the maximum entropy inference process on the set of collectively observed examples. Furthermore, we will show that if we confine the 'number of possible states' argument only to the observation process then an overwhelming number of states generate partitions that we can obtain from the process which combines the weighted arithmetic mean with the Kullback-Leibler divergence on both sets; the set of all examples and the set of collectively observed examples. In particular, if the observations made by each expert (or study) determine a unique partition both combinations above are reduced to the weighted arithmetic mean of these partitions; a surprising result considering that current statistical methods would suggest (to some degree) an un–weighted mean.

It might be concerning that we consider two ways of counting the states and do not advocate a unified approach to both single–expert and multi–expert reasoning. Nevertheless, we do believe that these are profoundly different problems that should not be mixed together. The prior should deal with the states of Nature that result in different partitions and the latter should deal with processes that cause differences in observing a single partition already existing in Nature.

It is important to stress that the presented argument is no ultimate approach to multi–expert inference in general and meta–analysis with unexplained heterogeneity in particular but rather an argument applicable in a narrow setting where our model of observing examples by drawing with replacement under all assumptions outlined above is appropriate. On the other hand, the present result could be significant if those assumptions can be satisfied in practice. We will discuss whether this could be the case at the end of the paper.

2. Framework

We create our model of drawing with replacement for multi–expert reasoning in general and meta–analysis in particular in a classical setting of uncertain reasoning as defined in [12]. In this setting we consider an enumeration of atomic sentences in some fixed order $\alpha_1, \ldots, \alpha_J$. The set $\{\alpha_1, \ldots, \alpha_J\}$ of all atoms will be denoted by At. The atoms of At could be thought of as mutually exclusive and exhaustive elementary events.

A probability function $\mathbf{w}$ over those atoms is defined by a function $\mathbf{w} : \mathrm{At} \to [0, 1]$ such that

$$\sum_{j=1}^{J} \mathbf{w}(\alpha_j) = 1.$$

For the sake of simplicity we will often write w_j instead of $\mathbf{w}(\alpha_j)$. We will denote the set of all probability functions with non–zero coordinates (i.e. $w_j > 0,\ 1 \leq j \leq J$) by $\mathbb{D}^J$ and in this paper we confine ourselves only to such restricted probability functions. Later we will define partitions mentioned in the introduction as special probability functions.

We talked informally about entropy in the introduction. Formally, entropy is a function $H : \mathbb{D}^J \to \mathbb{R}$ defined by

$$H(\mathbf{w}) = -\sum_{j=1}^{J} w_j \log w_j,$$

where in our consideration log stands for the natural logarithm but everything in this paper could be performed just as well with $\log_2$, which is a more common choice in information theory (e.g. the so called Shannon entropy [10]).

H is a strictly concave function and it is well–known and easy to prove that it has a unique maximum over any closed, convex and non–empty subset in $\mathbb{D}^J$. We will use such closed convex sets to represent partitions admissible according to an expert or a study. The maximum entropy principle states that given a closed, convex and non–empty set W in $\mathbb{D}^J$ we should choose the most entropic point in this set, i.e. $\arg\max_{\mathbf{w} \in W} H(\mathbf{w})$, as its representative. We denote such

a point by $\mathbf{ME}(W)$. As we have mentioned in the introduction, this choice was justified by Paris and Vencovská in [11] by the 'number of possible states' argument and we will discuss it in detail in Section 3.

In respect to entropy there is a closely connected notion of an asymmetric 'distance' between two probability functions known as the Kullback–Leibler divergence and defined by

$$\mathrm{KL}(\mathbf{w}\|\mathbf{v}) = \sum_{j=1}^{J} w_j \log \frac{w_j}{v_j}.$$

Since $\mathrm{KL}(\cdot\|\cdot)$ is strictly convex in the first argument, given a closed, convex and non–empty set $W \subseteq \mathbb{D}^J$ we may define the KL-projection of $\mathbf{v}$ into W as that unique point which minimises $\mathrm{KL}(\mathbf{w}\|\mathbf{v}) = \sum_{j=1}^{J} w_j \log \frac{w_j}{v_j}$ subject to $\mathbf{w} \in W$ as illustrated in Figure 1.

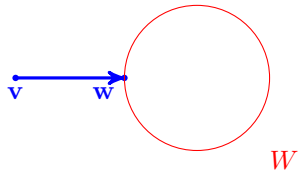

Figure 1: An illustration of a KL-projection.

It is easy to see that the KL-projection of the uniform probability function $\mathbf{u}$, defined by $u_j = \frac{1}{J}$ for all $1 \leq j \leq J$, into W is the most entropic point in W; i.e. $\mathbf{ME}(W)$.

The maximum entropy inference process says something about choosing a probability function in a single set which is closed, convex and non–empty. One possible way of choosing a probability function in several closed convex sets, say $W_1, \ldots, W_m \subseteq \mathbb{D}^J$ with respective weights $\lambda_1, \ldots, \lambda_m \in (0, 1)$, $\sum_{i=1}^{m} \lambda_i = 1$, involves minimising a sum of KL-divergences:

$$V = \left\{ \arg\min_{\mathbf{v}\in\mathbb{D}^J} \sum_{i=1}^{m} \lambda_i \,\mathrm{KL}(\mathbf{w}^{(i)}\|\mathbf{v}); \text{ subject to } \mathbf{w}^{(1)} \in W_1, \ldots, \mathbf{w}^{(m)} \in W_m \right\}.$$

The closed convex sets $W_1, \ldots, W_m$ can be used to represent admissible partitions observed by an expert or a study. Weighting factors $\lambda_1, \ldots, \lambda_m$ allow us to assign different reliability to experts in multi–expert reasoning in general and can represent different sample sizes in studies when this is applied in meta–analysis of some studies in particular.

The set V consists of all possible minimisers of the sum subject to the conditions. There is not necessary a unique minimiser, but the set V is known to be closed, convex and non–empty (see [14, Section 3.2] in the case when weighting is uniform and [15, Theorem 12] for a more detailed and general proof). Thanks to this property the combination of the arithmetic mean with the Kulback-Leibler divergence was considered as a well-defined probabilistic merging operator and called the linear entropy operator in [14]. We will use the same name for our operator employing a weighted arithmetic mean. The operator from [14] is equal to what we call the linear entropy operator if the weighting is uniform.

In general, a probabilistic merging operator is a mapping which maps a finite collection of closed convex nonempty subsets of $\mathbb{D}^J$, say $W_1, \ldots, W_n$, to a single closed convex nonempty subset of $\mathbb{D}^J$. Using the notation from [15, 16], where the linear entropy operator was denoted by $\Theta^{\mathrm{KL}}_{\vec{\lambda}}$, we will write $\Theta^{\mathrm{KL}}_{\vec{\lambda}}(W_1, \ldots, W_n)$ in the place of V above. The concept of a probabilistic merging operator is related both to the notion of a propositional merging operator of Konieczny and Pino–Pérez [17] and to the Paris–Vencovská notion of an inference process, and may be considered as a natural generalisation of the former to the probabilistic context. Such a probabilistic merging operator combines the individually consistent probabilistic knowledge of several experts, which may however be collectively inconsistent, into a single closed convex set of probability functions. In contrast to the case of a (single expert) inference processes of Paris and Vencovská, the result of the application of a probabilistic merging operator is thus not in general a single probability function, but rather a convex set of such functions. The effect of the operator is thus to remove any prior collective inconsistency in the knowledge, but not to provide a unique probabilistic belief function. The combination is not arbitrary but it is supposed to satisfy certain

natural principles.

The linear entropy operator $\Theta^{\mathrm{KL}}_{\vec{\lambda}}$ is only one of many probabilistic merging operators that have appeared in the literature. Changing the direction of the KL-projection (i.e. defining the projection of $\mathbf{v}$ into W as that unique point which minimises $\mathrm{KL}(\mathbf{v}\|\mathbf{w})$ subject to $\mathbf{w} \in W$) gives rise to a different probabilistic merging operator [14, Section 3.1] and we can even replace the KL-divergence by a more general convex Bregman divergence that is strictly convex in its second argument [15, 16]. And there are other methods different in flavour.

In this paper we however argue that if we confine the 'number of possible states' argument only to the observation process then there is an interesting argument in favour of the particular representation $\Theta^{\mathrm{KL}}_{\vec{\lambda}}$, which combines the weighted arithmetic mean with the Kullback–Leibler divergence. On the other hand, if we directly extend the argument from [11] to our multi–expert setting then we can argue in favour of the uniform probability function $\mathbf{u}$ as a universal representation (which however appears counter–intuitive).

3. The classical 'number of possible states' argument for the maximum entropy inference process

The original justification for the maximum entropy inference process from [11] is quite complicated and although we could attempt to simplify it (one obvious way would be to use the non-standard analysis), such an attempt would bring another burden to our consideration. Our aim here is however to explain the argument for the maximum entropy inference process as simply as possible and this is what we attempt to do in this section. Although everything in this section appeared in some form previously in the literature, the underlying idea actually in 1871 by Boltzmann [6, 7, 8], and we could get away with just references, we consider this section as an excellent opportunity to introduce our notation and review previous ideas on something the reader might be familiar with.

To this end, let us denote by U a fixed universe of examples observed by

an expert or investigated in a study (a set whose members are called examples) and assume that $|U| = N \in \mathbb{N}$. The reader may think of examples as patients observed by a physician or investigated in a medical study. The mapping

$$M_U : U \to \mathrm{At}$$

is called a model of U. A model M_U completely characterises the set of examples U. That means that for any particular example $u \in U$ there is precisely one atom $\alpha_j \in \mathrm{At}$ which is true for u according to M_U.

For given M_U we define a mapping $\mathbf{w}_{M_U} : \mathrm{At} \to \mathbb{Q}$ by

$$\mathbf{w}_{M_U}(\alpha_j) = \frac{|u \in U : M_U(u) = \alpha_j|}{|U|},$$

for all $\alpha_j \in \mathrm{At}$. So $\mathbf{w}_{M_U}$ of α_j returns the proportion of examples that according to the model M_U satisfy the atom α_j. We denote the set of all such mappings for fixed N by $\mathbb{P}^N$ and we call them partitions. Note that if $\mathbf{w} \in \mathbb{P}^N$ then also $\mathbf{w} \in \mathbb{P}^{nN}$ for every $n \in \mathbb{N}$. (This introduction of the parameter n is a means of expanding the universe of examples without any need to redefine previously introduced partitions.) Also notice that $\sum_{j=1}^{J} \mathbf{w}_{M_U}(\alpha_j) = 1$ so consequently $\mathbf{w}_{M_U}$ is a probability function. If we assume that no coordinate of $\mathbf{w}_{M_U}$ is zero (as we will often do) we may write $\mathbf{w}_{M_U} \in \mathbb{D}^J$.

Now consider a partition $\mathbf{w} \in \mathbb{P}^N$ generated by some model M_U with non–zero coordinates. It is apparent that there may be more models generating the same partition $\mathbf{w}$. Actually, there are

$$\sharp_N(\mathbf{w}) := \frac{N!}{\prod_{j=1}^{J}(w_j N)!}$$

models over the set of examples U of cardinality N that have the partition $\mathbf{w}$.

Looking at the situation from a different perspective, using only partitions to characterise the set U, we are not able to distinguish between two models if they produce the same partition. If we call such models isomorphic then $\sharp_N(\mathbf{w}_{M_U})$ denotes the number of models isomorphic with M_U.

Here comes a natural assumption of equiprobability of models which was used by Paris and Vencovská in [11] and which together with prior equiprobability of partitions gives the following argument: The more isomorphisms there are associated with a particular partition, the more likely that partition is. Note that 'number of models' corresponds to 'number of possible states' we talked about in the introduction. This reasoning is a version of ingenious Boltzmann's idea that the more ways molecules of a gas can arrange themselves in a particular distribution in space, the more likely that distribution is.

This is a purely combinatorial viewpoint, which seemingly misses some processes that operate in the problem, but it stands right behind the maximum entropy inference process as we will shortly see. Jaynes in [2, Chapter A] explained that Botlzmann "put into his equation only the dynamical information that happened to be relevant to the question he was asking". There are so many factors that affect the distribution of molecules of a gas, but by ignoring them we can make correct predictions in a few lines of calculations (because they would in fact turn out to be irrelevant if we performed the analysis in such details) [2, Chapter A]. Shannon exploited this when he applied this idea in the information theory [10] while Paris and Vencovská exploited this when they applied this idea in the field of inference processes [11] as detailed below. Only now we do not work with molecules of a gas but with examples (say patients) observed by an expert.

After explaining the idea, we proceed to actual mathematics.

Lemma 1. *Let $\mathbf{w}, \mathbf{v} \in \mathbb{P}^N$. Then there is a bounded real function c defined later by Formula (2) such that*

$$\frac{\sharp_{nN}(\mathbf{v})}{\sharp_{nN}(\mathbf{w})} = c(n, \mathbf{v}, \mathbf{w}) \cdot e^{-nN \cdot (H(\mathbf{w}) - H(\mathbf{v}))}.$$

Proof. Consider

$$\frac{\sharp_{nN}(\mathbf{v})}{\sharp_{nN}(\mathbf{w})} = \frac{\prod_{j=1}^{J}(w_j nN)!}{\prod_{j=1}^{J}(v_j nN)!}.$$

By Stirling's formula for every natural number n there is $\xi \in [0,1]$ such that

$$n! = \left(\frac{n}{e}\right)^n \sqrt{2\pi n} \cdot e^{\frac{\xi}{4n}}.$$

Hence

$$\frac{\sharp_{nN}(\mathbf{v})}{\sharp_{nN}(\mathbf{w})} = \frac{\prod_{j=1}^{J}(w_j nN)^{w_j nN} e^{-w_j nN}(2\pi w_j nN)^{\frac{1}{2}} e^{\frac{\xi_j}{4w_j nN}}}{\prod_{j=1}^{J}(v_j nN)^{v_j nN} e^{-v_j nN}(2\pi v_j nN)^{\frac{1}{2}} e^{\frac{\zeta_j}{4v_j nN}}}.$$

Since $\sum_{j=1}^{J} w_j = 1$ and $\sum_{j=1}^{J} v_j = 1$ equalities such as $\prod_{j=1}^{J}(nN)^{w_j nN} = (nN)^{nN}$ hold so the above is equal to

$$\frac{\prod_{j=1}^{J}(w_j)^{w_j nN}(w_j)^{\frac{1}{2}} e^{\frac{\xi_j}{4w_j}}}{\prod_{j=1}^{J}(v_j)^{v_j nN}(v_j)^{\frac{1}{2}} e^{\frac{\zeta_j}{4v_j}}} = \frac{\prod_{j=1}^{J}(w_j)^{w_j nN}}{\prod_{j=1}^{J}(v_j)^{v_j nN}} \cdot c(n,\mathbf{v},\mathbf{w}), \tag{1}$$

where

$$c(n,\mathbf{v},\mathbf{w}) := \frac{\prod_{j=1}^{J}(w_j)^{\frac{1}{2}} e^{\frac{\xi_j}{4w_j}}}{\prod_{j=1}^{J}(v_j)^{\frac{1}{2}} e^{\frac{\zeta_j}{4v_j}}}. \tag{2}$$

Now, $c(n,\mathbf{v},\mathbf{w})$ is bounded from up and down. This is because in Stirling's formula ξ depends on n but it varies only within $[0,1]$ and so do ξ_j and ζ_j, $1 \leq j \leq J$, above. Furthermore, w_j and v_j, $1 \leq j \leq J$, are by the definition non–zero constants.

Finally, by applying the identity $x = e^{\log x}$ we can rewrite the right-hand side of Formula (1) as

$$c(n,\mathbf{v},\mathbf{w}) \cdot e^{-nN\cdot(H(\mathbf{w})-H(\mathbf{v}))}.$$

□

It follows that if $H(\mathbf{w}) > H(\mathbf{v})$ then for a sufficiently big n the number of models isomorphic with a model generating the partition $\mathbf{w}$ is higher than the number of models isomorphic with a model generating $\mathbf{v}$. It actually could be

a much higher number of models; no matter the fixed difference $H(\mathbf{w}) - H(\mathbf{v})$, we can choose n sufficiently big to achieve that.

In the following we go back to the maximum entropy inference process by considering a closed convex non–empty set $W \subseteq \mathbb{D}^J$. We assume that the expert knows only that the partition on the set of all examples U is among those in W. For simplicity we often say that the expert observed these admissible partitions in W although he or she in fact established this set by observing examples. How exactly this was established is not discussed in this paper as this can be done in several ways. Paris and Vencovská for example assumed that expert observed constraints on possible partitions.

Furthermore, we assume that the probability functions in W are fixed in no coordinates to irrational coefficients. This means that if $\mathbf{w} \in W$ has for some j the value w_j irrational then there is another $\mathbf{v} \in W$ with $v_j \neq w_j$. Nevertheless, even if there were fixed irrational coefficients the results obtained in this section would not be different. Indeed, if W is such that all $\mathbf{w} \in W$ have the same value on some coordinate then the problem of maximising

$$H(\mathbf{w}) = -\sum_{j=1}^{J} w_j \log w_j$$

subject to $\mathbf{w} \in W$ is not affected by this coordinate and we may ignore it.

To introduce some examples, $W_1 = \{(\frac{1}{\sqrt{2}}, \frac{\sqrt{2}-1}{\sqrt{2}} - 2x, x, x), x \in [0.01, \frac{\sqrt{2}-1}{2\sqrt{2}} - 0.01]\}$ fixes the first coordinate to the irrational number $\frac{1}{\sqrt{2}}$ but $W_2 = \{(\frac{1}{\sqrt{2}} - x, \frac{\sqrt{2}-1}{\sqrt{2}} - x, x, x), x \in [0.01, \frac{\sqrt{2}-1}{\sqrt{2}} - 0.01]\}$ fix no coordinate to irrational coefficients. Note the way we have avoided zero coordinates so that the set is closed.

Lemma 2. *Let $W \subseteq \mathbb{D}^J$ be a closed convex non–empty set of probability functions that are fixed in no coordinates to irrational coefficients,*

$$\mathbf{x} = \arg\max_{\mathbf{w} \in W} H(\mathbf{w})$$

and $\epsilon > 0$ be a real number. Then there are a natural number N, a real number

$\delta > 0$ and $\mathbf{w} \in \mathbb{P}^N \cap W$ such that for all $\mathbf{y} \in W$ having

$$|x_j - y_j| > \epsilon \text{ for some } 1 \leq j \leq J \tag{3}$$

we have

$$|H(\mathbf{w}) - H(\mathbf{y})| > \delta.$$

Proof. For a contradiction, say that for all $N > N_0$, for all $\delta > 0$ and for all $\mathbf{w}^{(\delta)} \in \mathbb{P}^N \cap W$ there is $\mathbf{y}^{(\delta)} \in W$ with Property (3) such that $|H(\mathbf{w}^{(\delta)}) - H(\mathbf{y}^{(\delta)})| \leq \delta$. Consider an infinite sequence $\{\mathbf{y}^{(\delta_1)}, \mathbf{y}^{(\delta_2)}, \ldots\}$ where $\{\delta_i\}_{i=1}^{\infty} \to 0$ and $\{\mathbf{w}^{(\delta_i)}\}_{i=1}^{\infty} \to \mathbf{x}$. The latter convergence is possible since there are no fixed irrational coefficients in W. Since W is compact the sequence $\{\mathbf{y}^{(\delta_1)}, \mathbf{y}^{(\delta_2)}, \ldots\}$ has its limit $\mathbf{y}$ with Property (3) inside W. Hence $H(\mathbf{y}) = H(\mathbf{x})$ and we have a contradiction with the maximality of H uniquely at $\mathbf{x}$. □

We denote the set of all $\mathbf{y} \in W$ with Property (3) for some $1 \leq j \leq J$ by $W'_\epsilon(\mathbf{x})$.

Due to the lemma above we can consider two sets of partitions from $\mathbb{P}^N$. The first set contains a single partition $\mathbf{w}$, which is an approximation of the most entropic point $\mathbf{x}$ in W, i.e. $x = \mathbf{ME}(W)$. The second set contains all partitions $\mathbf{v}$ from $\mathbb{P}^N \cap W'_\epsilon(\mathbf{x})$; note that $\mathbf{w} \notin \mathbb{P}^N \cap W'_\epsilon(\mathbf{x})$. The following theorem says that for a sufficiently big universe the number of models isomorphic to a model generating $\mathbf{w}$ is overwhelmingly higher than the number of models generating the partitions from the second set.

Theorem 3. *Let $W \subseteq \mathbb{D}^J$ be a closed convex non–empty set of probability functions that are fixed in no coordinates to irrational coefficients, $\mathbf{x}$ be the most entropic point in W and ϵ be a fixed positive real number. Let a natural number N and a real number $\delta > 0$ be such that there is $\mathbf{w} \in \mathbb{P}^N \cap W$ such that for all $\mathbf{y} \in W'_\epsilon(\mathbf{x})$ we have $|H(\mathbf{w}) - H(\mathbf{y})| > \delta$. Then*

$$\lim_{n \to \infty} \frac{\sum_{\mathbf{v} \in \mathbb{P}^{nN} \cap W'_\epsilon(\mathbf{x})} \sharp_{nN}(\mathbf{v})}{\sharp_{nN}(\mathbf{w})} = 0.$$

Proof. First, if $\mathbf{x}$ is the most entropic point in W and $\epsilon > 0$ then by Lemma 2 there are a natural number N, a real number $\delta > 0$ and $\mathbf{w} \in \mathbb{P}^N \cap W$ such that for all $\mathbf{y} \in W'_\epsilon(\mathbf{x})$ we have $|H(\mathbf{w}) - H(\mathbf{y})| > \delta$. So the assumptions of the theorem can be satisfied. The rest of the argument is about establishing that Lemma 1 shows much stronger predominance of models isomorphic to a model generating $\mathbf{w}$ than the number of all models generating partitions from $\mathbb{P}^{nN} \cap W'_\epsilon(\mathbf{x})$ could be. Indeed, there are certainly less than $(nN)^J$ partitions over a universe containing nN elements. Hence

$$0 \leq \lim_{n\to\infty} \frac{\sum_{\mathbf{v}\in\mathbb{P}^{nN}\cap W'_\epsilon(\mathbf{x})} \sharp_{nN}(\mathbf{v})}{\sharp_{nN}(\mathbf{w})} =$$

$$= \lim_{n\to\infty} \Big[\sum_{\mathbf{v}\in\mathbb{P}^{nN}\cap W'_\epsilon(\mathbf{x})} c(n,\mathbf{v},\mathbf{w}) \cdot e^{-nN\cdot(H(\mathbf{w})-H(\mathbf{v}))} \Big] \leq \lim_{n\to\infty} \Big[(nN)^J e^{-nN\cdot\delta} \Big] = 0.$$

The second inequality above is due to the fact that $c(n,\mathbf{v},\mathbf{w})$ is bounded. However, it is slightly more complicated to see this than in Lemma 1. There, $\mathbf{v}$ and $\mathbf{w}$ were both fixed constants with non–zero coordinates. Here $\mathbf{v}$ can change but it must lie in W. Since W is a closed convex subset of $\mathbb{D}^J$ every coordinate of $\mathbf{v}$ is bounded from zero by a constant. Looking at $c(n,\mathbf{v},\mathbf{w})$ (Formula (2)) with this in mind we can easily see that it is bounded. □

So, in the present setting, if an expert knows only that the partition on the set of all examples U is among those in W then this partition is overwhelmingly the most likely close to $\mathbf{ME}(W)$. Furthermore, in absence of any restriction on the partition the overwhelmingly most likely partition is close to the uniform probability function $\mathbf{u}$.

The result above is weaker than the one established by Paris and Vencovská in [11]. They dealt also with probability functions containing zero coordinates, they did not simply ignore fixed irrational coordinates and they introduced a technique which naturally translates linear constraints with real coefficients into rational models. Giving such a precise argument in favour of the maximum entropy inference process was certainly an achievement. It is however an

unfortunate attribute of the argument due to Paris and Vencovská that it gets lost in technical details. Nonetheless, we do hope that we have explained the underlying idea as clearly as possible to the reader so in the next section we can recreate it in multi–expert reasoning.

4. The 'number of possible states' argument in multi–expert reasoning

In this section we will modify the 'number of possible states' argument described in the previous section to a multi–expert setting. First, we will explain necessary changes to the framework introduced in the previous section. Second, we justify and explain the assumptions that were briefly mentioned in the introduction. Finally, given the assumptions, we prove the result.

4.1. Modifications to the framework

In this new setting we consider that m experts observed their own sets of examples $U_1 \subseteq U, \ldots, U_m \subseteq U$. We will assume that cardinalities of collectively observed examples are significantly smaller than the cardinality of U. This is a necessary assumption and should be viewed as a limitation of the argument. It also means that we do not simply recover the analysis performed in Section 3 by taking $m = 1$. We denote the relative proportion of observed examples by $\epsilon\lambda_i = \frac{|U_i|}{|U|}$, $1 \leq i \leq m$; where $\lambda_1, \ldots, \lambda_m \in (0, 1)$, $\sum_{i=1}^{m} \lambda_i = 1$, allow us to talk about relative weights of expert observations; indeed, the more examples an expert observed, the higher weight we ought to put to his or her observation.

In the previous section we called $M_U : U \to \mathrm{At}$ a model. Recall that M_U generates a partition $\mathbf{x}$ on U and that there could be more models generating the same partition. In this section our language is richer and to every member of the universe U we need to assign also experts that observed it. Therefore, we will call a mapping generating $\mathbf{x}$ above a representative of $\mathbf{x}$ and in this section a model is an ordered pair (M_U, R_U), where $R_U : U \to \mathcal{P}(1, \ldots, m)$ maps every element of the universe to the power set of indexes denoting m experts

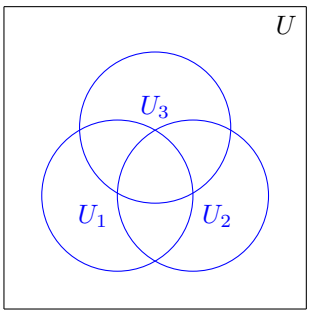

Figure 2: The illustration of the multi–expert setting for $m = 3$.

respectively. Such a model (M_U, R_U) induces partitions $\mathbf{x} \in \mathbb{P}^N$, $\mathbf{x}^{(1)} \in \mathbb{P}^{\epsilon\lambda_1 N}$, $\ldots$, $\mathbf{x}^{(m)} \in \mathbb{P}^{\epsilon\lambda_m N}$ over U, $U_1, \ldots, U_m$ respectively. Note that experts could have observed the same examples and an illustration of this setting for $m = 3$ is in Figure 2.

Now, fixing a representative M_U inducing the partition $\mathbf{x}$, given fixed sizes of $U_1 \subseteq U, \ldots, U_m \subseteq U$, we are interested in the number of ways we can choose $U_1, \ldots, U_m$ from U that the resulting partitions will be the same m–tuple $\mathbf{x}^{(1)} \in \mathbb{P}^{\epsilon\lambda_1 N}$, $\ldots$, $\mathbf{x}^{(m)} \in \mathbb{P}^{\epsilon\lambda_m N}$. We shall formally refer to a particular m–tuple $U_1, \ldots, U_m \subseteq U$ generating $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$ given a fixed representative M_U and fixed sizes of $U_1, \ldots, U_m$ as a way of observing $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$ from $\mathbf{x}$. In this paper we are not concerned with how precisely the real–world process of observing examples leads to a particular partition (e.g. constraints on possible partitions as used by Paris and Vencovská) but instead we identify the number of ways the experts could have made their observations $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$ if $\mathbf{x}$ is the true partition on U with the number of ways we can choose $U_1, \ldots, U_m$ above. Since we can choose the examples belonging to U_{i_1} independently on those belonging to U_{i_2} the number of ways is given by the formula

$$\sharp_N^{\vec{\lambda}_\epsilon}(\mathbf{x}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}) := \prod_{j=1}^{J} \prod_{i=1}^{m} \frac{(x_j N)!}{(x_j^{(i)} N\epsilon\lambda_i)!(x_j N - x_j^{(i)} N\epsilon\lambda_i)!} =$$

$$= \prod_{j=1}^{J} \prod_{i=1}^{m} \binom{x_j N}{x_j^{(i)} N \epsilon \lambda_i}. \tag{4}$$

Similarly as in the previous section, we say that two models are isomorphic if we are not able to distinguish between them by means of partitions that they generate on the sets $U_1, \ldots, U_m$ and U. Therefore, given a model (M_U, R_U) generating $\mathbf{x} \in \mathbb{P}^N$ on U and $\mathbf{x}^{(i)} \in \mathbb{P}^{\epsilon\lambda_i N}$ on U_i, $1 \leq i \leq m$, the number of models isomorphic to this model is

$$\sharp_N^{\vec{\lambda}\epsilon}(\mathbf{x}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}) \cdot \frac{N!}{\prod_{j=1}^{J}(x_j N)!}. \tag{5}$$

Now we ask a question as to whether the assumption of equiprobability of models similar in spirit to the one adopted in the previous section is sensible in this new setting and what solution it gives. We will explore this in the next section while mathematical results will be given in Section 4.3.

4.2. Assumptions

Before we proceed further we shall define our mathematical modelling of multi–expert reasoning in general and meta–analysis in particular using drawing with replacement. Imagine that there is a huge box full of balls coloured with J colours (J is known). Each of m experts selected a large number of balls from the box but this number is insignificant in comparison to the overall number of balls and then he or she returned them back to the box. We are given only the observations of the experts; m partitions of observed balls according to the colour and the relative numbers of observed balls $\lambda_1, \ldots, \lambda_m$. Based only on this information we would like to infer the true partition of balls in the box.

Assume for now that we continue with the assumption of equiprobability of models. First, recall that there are far more representatives generating partitions on U close to the uniform probability function $\mathbf{u}$, defined by $u_j = \frac{1}{J}$ for all $1 \leq j \leq J$, than those generating other partitions altogether. It turns out that, given m partitions $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$ observed respectively by experts, there is an overwhelming number of models that are isomorphic to a model which induces

these partitions $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$ on the sets of observed examples $U_1, \ldots, U_m$ respectively and which induces a partition as close as possible to the uniform probability function on the set of all examples U. So we would be compelled to argue that, regardless of the partitions observed by experts, overwhelmingly the most likely partition of balls in the box is the uniform probability function. And although there was nothing obviously wrong with the argument prior to this further analysis, after seeing the result of directly extending the argument due to Paris and Vencovská to our multi–expert setting we argue that such an extension is counter–intuitive if we focus only on the set of all examples U. In Section 6 we will show that the partition generated on the set of collectively observed examples $U_1 \cup \ldots \cup U_m$ in fact makes sense.

Nevertheless, here it appears that by imposing equiprobability on models we actually take away prior equiprobability from partitions on U. To obtain equiprobability of partitions we fix a representative for each partition $\mathbf{x} \in \mathbb{P}^N$ on U. The number of ways we could have observed given $\mathbf{x}^{(i)} \in \mathbb{P}^{\epsilon\lambda_i N}$ on U_i, $1 \leq i \leq m$, from $\mathbf{x} \in \mathbb{P}^N$ on U is given by Formula (4). We assume that the probability that those m partitions were observed in a particular way is the same across all representatives. Unfortunately, the idea of fixed representatives is here adopted *ad hoc* to obtain eqiprobability of partitions although it may seem natural if we focus only on the process of observation.

More rigorously, for every $\mathbf{x} \in \mathbb{P}^N$ let $M_U^{\mathbf{x}}$ be a representative generating $\mathbf{x}$. A way of observing $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$ from $\mathbf{x}$ is an m–tuple $U_1, \ldots, U_m \subseteq U$ generating partitions $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$ respectively where U is characterised by $M_U^{\mathbf{x}}$. The union of all these m–tuples across all $\mathbf{x} \in \mathbb{P}^N$ forms our probabilistic space over which we employ the equiprobability assumption. In other words, we first set equiprobability of partitions by assigning only one representative to each and then we set equiprobability of all ways of observing $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$ from every partition $\mathbf{x} \in \mathbb{P}^N$.

Note that the probability of a way of observing above is not a frequency measurable in the real–world. We use the concept of probability differently; in the same way as Shannon did when he wrote about the probability of receiving a

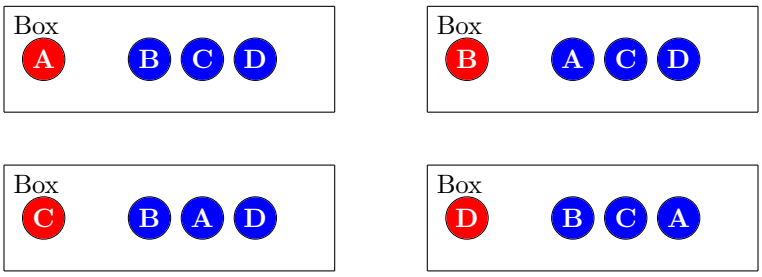

Figure 3: Four possible representatives of the partition 'one red and three blue'.

certain message via a communication channel. Jaynes [2, Chapter A] explained it in the following way.

> "... an example where absurdity of a frequency interpretation is so obvious that no one can fail to see it; but the usefulness of the probability approach was equally clear. The probabilities assigned to individual messages are not measurable frequencies; they are only a means of describing a state of knowledge; ... ."

To illustrate our new approach deviating from the one used by Paris and Vencovská, assume that there are four balls A, B, C and D in the box and there are two colours: red and blue. Assume that the first expert observed one ball and that it was blue and the second expert observed one ball which was also blue. For the partition 'one red and three blue' there are many possible representatives; for example, in one representation the ball A could be red and in another one the ball B could be red. These two representatives are isomorphic if we only distinguish by partitions on balls according to the colour. Say that we select the first representative. Note that there are $\frac{4!}{1!3!} = 4$ representatives isomorphic with this representative. See Figure 3 for an illustration.

Another partition could be 'two red and two blue'. Again we select one possible representative; say A and B have the colour red while C and D have the colour blue. This representative is not isomorphic to the previously selected representative (where only A was red) but there are $\frac{4!}{2!2!} = 6$ other representatives

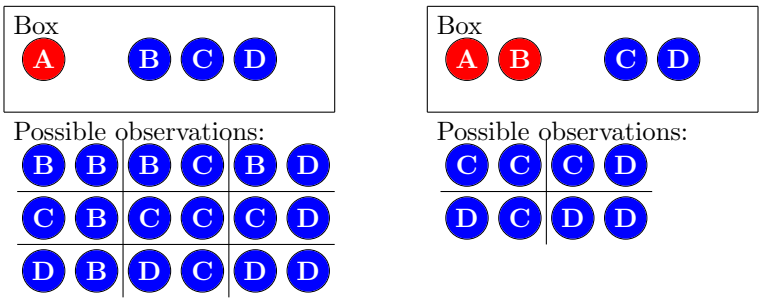

Figure 4: Possible observations from representatives of two different partitions.

isomorphic with it.

Should the first representative (where only A was red) be real, experts could have observed two blue balls in $3 \cdot 3 = 9$ ways. On the other hand, if the second representative was real then experts could have observed two balls in $2 \cdot 2 = 4$ ways. See Figure 4 for an illustration. Similarly, the partition 'no red and four blue' generates 16 ways of observing, the partition 'three red and one blue' generates 1 way of observing and the partition 'four red and no blue' generates 0 ways of observing.

We assume that, given we only know that each expert observed a blue ball, each of the ways of observing regardless of the representative is *a priori* equally probable. Now, the partition 'no red and four blue' generates the highest number of ways of observing, 16. To argue rigorously in favour of this partition, say $\mathbf{x}$, let us establish the probability that it is true given the observations of experts, say $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}$, by the Bayes formula:

$$P(\mathbf{x}|\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \frac{P(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}|\mathbf{x}) \cdot P(\mathbf{x})}{P(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})} = \frac{P(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}|\mathbf{x}) \cdot P(\mathbf{x})}{\sum_{\mathbf{y} \in \mathbb{P}^4} P(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}|\mathbf{y}) \cdot P(\mathbf{y})},$$

where $P(\mathbf{x})$ is the prior probability that $\mathbf{x}$, i.e. 'no red and four blue', is the true partition of balls in the box, $P(\mathbf{y})$ is the prior probability that the partition $\mathbf{y}$ is the actual partition, $P(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}|\mathbf{x})$ and $P(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}|\mathbf{y})$ are the probabilities that each experts selects a blue ball given the partition is 'no red and four blue' and $\mathbf{y}$ respectively, and finally $P(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$ is the probability that each expert

selects a blue ball without assuming any particular partition.

Now, establishing $P(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}|\mathbf{y})$ for every possible partition $\mathbf{y}$ by counting the number of ways of observing we obtain the following numbers: $\frac{16}{16+0+0+0}$ for the partition 'no red and four blue', $\frac{9}{9+3+3+1}$ for 'one red and three blue', $\frac{4}{4+4+4+4}$ for 'two red and two blue', $\frac{1}{1+3+3+9}$ for 'three red and one blue' and $\frac{0}{0+0+0+16}$ for 'four red and no blue'. Finally, due to the assumption of prior equiprobability of partitions, the posterior probability is

$$P(\mathbf{x}|\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \frac{P(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}|\mathbf{x})}{\sum_{\mathbf{y}\in\mathbb{P}^4} P(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}|\mathbf{y})} = \frac{16}{16+9+4+1+0} = \frac{16}{30}.$$

It appears rational to conclude that the most likely partition of balls in the box has four blue balls even there is only one possible representative of the partition 'no red and four blue'. Rather than arguing in favour of multiplying 9 ways of observing four times since there are 4 possible representatives of the partition 'one red and three blue', multiplying 4 ways of observing six times since there are 6 possible representatives of the partition 'two red and two blue' and multiplying 1 way of observing four times since there are 4 possible representatives of the partition 'three red and one blue' (see Table 1), which is exactly what we did when we employed the original approach of equibrobability of models (i.e. Formula (5)). Note however that, unlike in this illustration containing only four balls, we do assume a large number of balls and significantly smaller sets of observed examples that are nevertheless still large.

Partition	**No. Ways**	**No. Representatives**	**No. Models**
'no red and four blue'	16	1	16
'one red and three blue'	9	4	36
'two red and two blue'	4	6	24
'three red and one blue'	1	4	4
'four red and no blue'	0	1	0

Table 1: The comparison between the mumber of ways of observing, the number of representatives and the number of models for different partitions in our four–ball example.

The computation that we have just illustrated on a simple example above can be performed in full generality: We wish to establish the posterior probability that a partition $\mathbf{x}$ is true given partitions $\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}$ observed by experts respectively, which is denoted here by $P(\mathbf{x}|\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)})$, using the Bayes formula:

$$P(\mathbf{x}|\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}) = \frac{P(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}|\mathbf{x})\cdot P(\mathbf{x})}{\sum_{\mathbf{y}\in\mathbb{P}^N} P(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}|\mathbf{y})\cdot P(\mathbf{y})}, \tag{6}$$

where

$$P(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}|\mathbf{y}) = \frac{\sharp_N^{\vec{\lambda}\epsilon}(\mathbf{y}|\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)})}{\sum_{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}} \sharp_N^{\vec{\lambda}\epsilon}(\mathbf{y}|\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)})}.$$

The latter formula, where in the denominator we sum over all possible partitions $\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}$ that experts could observe when each expert i selected a fixed number $\lambda_i\epsilon N$ of balls, is well defined due to the assumption of equiprobability of ways of observing.

Now, we will establish that $\sum_{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}} \sharp_N^{\vec{\lambda}\epsilon}(\mathbf{y}|\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)})$ is a constant independent on a particular selection of $\mathbf{y}$ (this in fact can be observed in our illustrative example above). The key to proving this is Vandermonde's identity:

$$\binom{N}{s} = \sum_{k=0}^{s}\left[\binom{N-m}{s-k}\cdot\binom{m}{k}\right],$$

which can be inductively generalised to

$$\binom{N}{s} = \sum_{k_1+\ldots+k_{J-1}=0}^{s}\left[\binom{N-m_1-\ldots-m_{J-1}}{s-k_1-\ldots-k_{J-1}}\cdot\binom{m_{J-1}}{k_{J-1}}\cdot\ldots\cdot\binom{m_1}{k_1}\right],$$

and expressed as

$$\binom{N}{s} = \sum_{x_1^{(i)},\ldots,x_J^{(i)}}\left[\binom{y_1}{x_1^{(i)}}\cdot\binom{y_2}{x_2^{(i)}}\cdot\ldots\cdot\binom{y_J}{x_J^{(i)}}\right],$$

where $\sum_{j=1}^{J} y_j = N$ and $\sum_{j=1}^{J} x_j^{(i)} = s$. Now, put $s = \lambda_i\epsilon N$. Then it follows that

$$\sum_{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}} \sharp_N^{\vec{\lambda}\epsilon}(\mathbf{y}|\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}) = \prod_{i=1}^{n} \binom{N}{\lambda_i \epsilon N}.$$

With this result and assuming prior equiprobability of partitions Formula (6) becomes

$$P(\mathbf{x}|\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}) = \frac{\sharp_N^{\vec{\lambda}\epsilon}(\mathbf{x}|\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)})}{\sum_{\mathbf{y}\in\mathbb{P}^N} \sharp_N^{\vec{\lambda}\epsilon}(\mathbf{y}|\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)})}.$$

In other words, the higher the number of ways of observing, the higher the posterior probability that $\mathbf{x}$ is the true partition.

To conclude our ball observations problem with this new approach, given m observed partitions $\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(m)}$ with proportions of observed balls $\lambda_1,\ldots,\lambda_m$ respectively, it will turn out that overwhelmingly the most likely partition $\mathbf{x}$ of balls in the box is just the weighted arithmetic mean of $\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(m)}$ with weights $\lambda_1,\ldots,\lambda_m$ respectively, i.e.

$$w_j = \sum_{i=1}^{m} \lambda_i w_j^{(i)},\ 1 \leq j \leq J.$$

A rigorous proof for the claim above will be given in Section 6.1. Note that if we change our equiprobability assumption we obtain a different result as seen in Table 2.

Assumption: Equiprobability of	**Overwhelmingly Most Likely Partition**
Ways of Observing	Weighted Arithmetic Mean of Partitions
Models	Uniform Partition

Table 2: Overwhelmingly the most likely partition on the set of all examples given each expert observed a unique partition and under different equiprobability assumptions.

Considering studies investigating patients in place of experts drawing balls we obtain the argument in favour of using the weighted arithmetic mean in meta–analysis with unexplained heterogeneity. Furthermore, this result will

turn out to be a special case in our more general multi–expert consideration where we are given only that the partitions on examples observed by experts are among particular closed convex sets of probability functions rather than specifically which single partitions they are.

The presented idea of counting 'number of ways' (we talked about 'number of possible states' in the introduction) mathematically corresponds to original Boltzmann's idea, which was spelled out in the previous section in terms of models, but in this setting it is not without criticism. To explain this further, there is actually a single true partition of balls in the box. If we subsequently apply the statistical sampling theory then we can establish that in the present setting it is unlikely that each observed partition in each coordinate differ more than by $1/\sqrt{\text{'sample size'}}$ from this true partition. However, what if we actually get from experts observations that are much more different than statistically expected (i.e. there is unexplained heterogeneity)? From our position, knowing only the observations, we may argue that the statistical approach is unsatisfactory because of the unknown processes operating in the real–world process of observation, but at the same time we have no means how to learn about those unknown processes. Therefore, in our mathematical model we represent the variability of the unknown processes by the variability of ways of observing. As in reality any process could be affecting the observations so in our mathematical model every way of observing is equally possible. Finally, we deduce that the way of observing that was actually realised most likely originates in that partition which produces the most ways of observing. And of course, this argument breaks down at the moment we learn what particular process causes the difference (explained heterogeneity), but in absence of such information treating the ways of observing differently would include information we do not posses.

Finally, the actual existence of a true partition of balls in the box should point out to the reason why we might like to refrain from the assumption of equiprobability of models but instead assume equiprobability of ways of observing. The observation process is where unknown processes are in reality causing higher difference than statistically expected. Processes that operated when the

partition was created might be considered irrelevant from the point of view of multi–expert reasoning, they should be the object of study of single–expert reasoning. So in our opinion single–expert reasoning should not be considered as a special case of multi–expert reasoning and analysing a study should not be considered as a special case of analysing multiple studies. Putting it in another way, if we want to uncover an unknown proportion of patients with some symptom from several observations as in meta–analysis, we are interested in what went wrong with the observation process. We do not contemplate at that stage in how many ways Nature could have created the partition. The partition is already physically present in our world and if we had resources we could have identified it by examining every single patient in the world.

Our interpretation of multi–expert reasoning as drawing with replacement and in particular the interpretation of unknown processes affecting the real process of observation by ways of observing is indeed only an abstraction of reality and it is up to the reader whether he or she finds the reasoning explained in this section convincing. In fact we do encourage reader to only accept our conclusions in this very case and not because he or she too finds the answer produced by counting the 'number of models' counter–intuitive. We cite George Wilmers [18] on this matter:

> "Sometimes one can see that a model is wrong because it gives results that don't seem reasonable, but one can't immediately see what is wrong with the original intuition. On the other hand it is in my opinion a waste of energy to construct artificially a supposed heuristic model which justifies a particular technical method, if that model lacks a clear intuitive basis. The field we are working in has methodologically much in common with pure science or applied mathematics, involving a dialectical interplay between mathematics and intuition. One proceeds back and forth between the two, improving our models at each stage. However, in order to do this efficiently one needs at each stage to be ruthlessly critical in appraising both the

consequences of the particular mathematical model and the inherent plausibility of the intuition generating it."

Furthermore, on the set of collectively observed examples the result provided by counting 'number of models' seems rational. We will take a look at this issue in Section 6.1.

4.3. The result

After we have successfully established the multi–expert framework we work in (Section 4.1) and rigorously formulated our assumptions (Section 4.2) it is finally time to prove the main result of this paper. Namely, in our special framework with all those assumptions the linear entropy operator $\Theta^{\mathrm{KL}}_{\vec{\lambda}}$ defined in Section 2 produces partitions that are such that each of them is overwhelmingly more likely than all the other partitions not produced by this operator together.

We will prove this result in a similar fashion to how we have established our weaker version of the results due to Paris and Vencovská in Section 3. First, in Lemma 4 we will show that there is a connection between the Kullback–Leibler divergence and the number of ways of observing similarly as there is a connection between entropy and the number of models. Second, we formally define what we mean by all the other partitions not produced by the linear entropy operator. Finally, we prove our main result: Theorem 5. There, similarly as in Section 3, we show that Lemma 4 gives much stronger predominance of certain ways of observing.

Lemma 4. *Let* $\lambda_1, \ldots, \lambda_m \in (0,1)$ *be such that* $\sum_{i=1}^m \lambda_i = 1$*, let* $\mathbf{w}, \mathbf{v} \in \mathbb{P}^N$*,* $\mathbf{w}^{(i)}, \mathbf{v}^{(i)} \in \mathbb{P}^{\epsilon\lambda_i N}$*, where* ϵ *is such that* $\epsilon\lambda_i N \in \mathbb{N}$ *for all* $1 \leq i \leq m$*, and*

$$d = \sum_{j=1}^{J}\sum_{i=1}^{m}\Big(w_j \log\Big(1 - \frac{w_j^{(i)}\epsilon\lambda_i}{w_j}\Big) + w_j^{(i)}\epsilon\lambda_i \log\frac{w_j^{(i)}\lambda_i}{w_j - w_j^{(i)}\epsilon\lambda_i}\Big) -$$

$$- \sum_{j=1}^{J}\sum_{i=1}^{m}\Big(v_j \log\Big(1 - \frac{v_j^{(i)}\epsilon\lambda_i}{v_j}\Big) + v_j^{(i)}\epsilon\lambda_i \log\frac{v_j^{(i)}\lambda_i}{v_j - v_j^{(i)}\epsilon\lambda_i}\Big). \tag{7}$$

Then there is a bounded real function c define later by Formula (9) such that

$$\frac{\sharp_{nN}^{\vec{\lambda}_\epsilon}(\mathbf{v}|\mathbf{v}^{(1)},\dots,\mathbf{v}^{(n)})}{\sharp_{nN}^{\vec{\lambda}_\epsilon}(\mathbf{w}|\mathbf{w}^{(1)},\dots,\mathbf{w}^{(n)})} = c(n,\vec{\mathbf{v}},\vec{\mathbf{w}})\cdot e^{nN\cdot d}.$$

675 *Proof.* Consider

$$\frac{\sharp_{nN}^{\vec{\lambda}_\epsilon}(\mathbf{v}|\mathbf{v}^{(1)},\dots,\mathbf{v}^{(n)})}{\sharp_{nN}^{\vec{\lambda}_\epsilon}(\mathbf{w}|\mathbf{w}^{(1)},\dots,\mathbf{w}^{(n)})}. \tag{8}$$

By Stirling's formula, similarly as in Lemma 1, there are constants $\xi_j \in [0,1]$, $\zeta_j \in [0,1]$, $\xi_j^{(i)} \in [0,1]$ and $\zeta_j^{(i)} \in [0,1]$, $1 \le i \le m$, such that Formula (8) is equal to

$$\prod_{j=1}^{J}\prod_{i=1}^{m}\Bigg(\frac{(v_jnN)^{v_jnN+\frac{1}{2}}e^{\frac{\zeta_j}{4nNv_j}}e^{-v_jnN}}{(w_jnN)^{w_jnN+\frac{1}{2}}e^{\frac{\xi_j}{4nNw_j}}e^{-w_jnN}}$$

$$\frac{(w_j^{(i)}nN\epsilon\lambda_i)^{w_j^{(i)}nN\epsilon\lambda_i+\frac{1}{2}}e^{\frac{\xi_j^{(i)}}{4w_j^{(i)}nN\epsilon\lambda_i}}e^{-w_j^{(i)}nN\epsilon\lambda_i}}{(v_j^{(i)}nN\epsilon\lambda_i)^{v_j^{(i)}nN\epsilon\lambda_i+\frac{1}{2}}e^{\frac{\zeta_j^{(i)}}{4v_j^{(i)}nN\epsilon\lambda_i}}e^{-v_j^{(i)}nN\epsilon\lambda_i}}$$

680

$$\frac{(w_jnN-w_j^{(i)}nN\epsilon\lambda_i)^{w_jnN-w_j^{(i)}nN\epsilon\lambda_i+\frac{1}{2}}e^{\frac{\xi_j^{(i)}}{4(w_jnN-w_j^{(i)}nN\epsilon\lambda_i)}}e^{-w_jnN+w_j^{(i)}nN\epsilon\lambda_i}}{(v_jnN-v_j^{(i)}nN\epsilon\lambda_i)^{v_jnN-v_j^{(i)}nN\epsilon\lambda_i+\frac{1}{2}}e^{\frac{\zeta_j^{(i)}}{4(v_jnN-v_j^{(i)}nN\epsilon\lambda_i)}}e^{-v_jnN+v_j^{(i)}nN\epsilon\lambda_i}}\Bigg).$$

Since $\sum_{j=1}^{J} v_j = 1$, $\sum_{j=1}^{J} w_j = 1$, $\sum_{j=1}^{J} w_j^{(i)} = 1$ and $\sum_{j=1}^{J} v_j^{(i)} = 1$, $1 \le i \le m$, the above is equal to

$$\prod_{j=1}^{J}\prod_{i=1}^{m}\Bigg(\frac{(v_j)^{v_jnN+\frac{1}{2}}e^{\frac{\zeta_j}{v_j}}(w_j^{(i)}\lambda_i)^{w_j^{(i)}nN\epsilon\lambda_i+\frac{1}{2}}e^{\frac{\xi_j^{(i)}}{w_j^{(i)}}}}{(w_j)^{w_jnN+\frac{1}{2}}e^{\frac{\xi_j}{w_j}}(v_j^{(i)}\lambda_i)^{v_j^{(i)}nN\epsilon\lambda_i+\frac{1}{2}}e^{\frac{\zeta_j^{(i)}}{v_j^{(i)}}}}$$

$$\left.\frac{(w_j - w_j^{(i)}\epsilon\lambda_i)^{w_j nN - w_j^{(i)} nN\epsilon\lambda_i + \frac{1}{2}} e^{\frac{\xi_j^{(i)}}{w_j - w_j^{(i)}\epsilon\lambda_i}}}{(v_j - v_j^{(i)}\epsilon\lambda_i)^{v_j nN - v_j^{(i)} nN\epsilon\lambda_i + \frac{1}{2}} e^{\frac{\zeta_j^{(i)}}{v_j - v_j^{(i)}\epsilon\lambda_i}}}\right).$$

Taking

$$c(n, \vec{\mathbf{v}}, \vec{\mathbf{w}}) := \prod_{j=1}^{J}\prod_{i=1}^{m} \frac{(v_j)^{\frac{1}{2}} e^{\frac{\zeta_j}{v_j}} (w_j^{(i)}\lambda_i)^{\frac{1}{2}} e^{\frac{\xi_j^{(i)}}{w_j^{(i)}}} (w_j - w_j^{(i)}\epsilon\lambda_i)^{\frac{1}{2}} e^{\frac{\xi_j^{(i)}}{w_j - w_j^{(i)}\epsilon\lambda_i}}}{(w_j)^{\frac{1}{2}} e^{\frac{\xi_j}{w_j}} (v_j^{(i)}\lambda_i)^{\frac{1}{2}} e^{\frac{\zeta_j^{(i)}}{v_j^{(i)}}} (v_j - v_j^{(i)}\epsilon\lambda_i)^{\frac{1}{2}} e^{\frac{\zeta_j^{(i)}}{v_j - v_j^{(i)}\epsilon\lambda_i}}} \tag{9}$$

we obtain

$$c(n, \vec{\mathbf{v}}, \vec{\mathbf{w}}) \cdot \prod_{j=1}^{J}\prod_{i=1}^{m} \left(\frac{(v_j)^{v_j nN} (w_j^{(i)}\lambda_i)^{w_j^{(i)} nN\epsilon\lambda_i} (w_j - w_j^{(i)}\epsilon\lambda_i)^{w_j nN - w_j^{(i)} nN\epsilon\lambda_i}}{(w_j)^{w_j nN} (v_j^{(i)}\lambda_i)^{v_j^{(i)} nN\epsilon\lambda_i} (v_j - v_j^{(i)}\epsilon\lambda_i)^{v_j nN - v_j^{(i)} nN\epsilon\lambda_i}} \right),$$

where c is a real function bounded from up and down (for the same reason as in Lemma 1). By applying the identity $x = e^{\log x}$ we can rewrite the above as

$$c(n, \vec{\mathbf{v}}, \vec{\mathbf{w}}) \cdot e^{nN \cdot d},$$

where d was defined in the statement of the lemma. □

Let $W_1, \ldots, W_m \subseteq \mathbb{D}^J$ be closed convex non–empty sets of probability functions with respective weights $\lambda_1, \ldots, \lambda_m \in (0, 1)$, $\sum_{i=1}^{m} \lambda_i = 1$. We assume that each expert 'i' knows only that the partition on the set of examples U_i observed by him is among those in W_i. Furthermore, we assume that probability functions in $W_1, \ldots, W_m$ are fixed in no coordinates to irrational coefficients and $\mathbf{x}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$ is an $(m+1)$–tuple of probability functions which minimises

$$\sum_{i=1}^{m} \lambda_i \, \mathrm{KL}(\mathbf{x}^{(i)} \| \mathbf{x})$$

subject to $\mathbf{x}^{(1)} \in W_1, \ldots, \mathbf{x}^{(m)} \in W_m$. Since such an $(m+1)$–tuple is not necessary unique, we cannot look at all other $(m+1)$–tuples that differ from this

one in some coordinates by some positive constant as we did in the case of the argument for the maximum entropy inference process. Also, we cannot simply ignore coordinates that are fixed to irrational coefficients as we could before. Indeed, if some W_{i_1} fixes a particular coordinate it still could vary for another W_{i_2}. And although we could try to create a machinery for approximating such sets using partitions, this could obscure the main argument and we are trying to present the idea as simply as possible. For these reasons we derive the following result, which is weaker than the one established by Paris and Vencovská in the singe–expert context, but on the bright side it is in a more complex multi–expert setting.

First, we take all $(m+1)$–tuples $\mathbf{y} \in \mathbb{D}^J, \mathbf{y}^{(1)} \in W_1, \ldots, \mathbf{y}^{(m)} \in W_m$ such that

$$\sum_{i=1}^{m} \lambda_i \operatorname{KL}(\mathbf{y}^{(i)} \| \mathbf{y}) - \sum_{i=1}^{m} \lambda_i \operatorname{KL}(\mathbf{x}^{(i)} \| \mathbf{x}) > \delta$$

for some $\delta > 0$ and denote this set by W'_δ. Then we consider two sets of $(m+1)$–tuples in $\mathbb{P}^N \times \mathbb{P}^{\epsilon\lambda_1 N} \times \ldots \times \mathbb{P}^{\epsilon\lambda_m N}$. The first consists of a single $(m+1)$–tuple $\mathbf{w} \in \mathbb{D}^J, \mathbf{w}^{(1)} \in W_1, \ldots, \mathbf{w}^{(m)} \in W_m$ such that

$$\sum_{i=1}^{m} \lambda_i \operatorname{KL}(\mathbf{w}^{(i)} \| \mathbf{w}) - \sum_{i=1}^{m} \lambda_i \operatorname{KL}(\mathbf{x}^{(i)} \| \mathbf{x}) < \frac{\delta}{2}. \tag{10}$$

The second consists of all $(m+1)$–tuples $\mathbf{v}, \mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(m)}$ which are also members of W'_δ. Note that

$$\sum_{i=1}^{m} \lambda_i \operatorname{KL}(\mathbf{v}^{(i)} \| \mathbf{v}) - \sum_{i=1}^{m} \lambda_i \operatorname{KL}(\mathbf{w}^{(i)} \| \mathbf{w}) > \frac{\delta}{2}.$$

It is not hard to see that for a sufficiently large N and a sufficiently small ϵ those two sets are non–empty: Since there are no fixed coordinates to irrational coefficients then either $\mathbf{x}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$ have only rational coefficients or there is another (in irrational coordinates) distinct $(m+1)$–tuple $\mathbf{z} \in \mathbb{D}^J, \mathbf{z}^{(1)} \in W_1, \ldots, \mathbf{z}^{(m)} \in W_m$ such that on the line joining those two points (convexity) we can form a sequence indexed by N of $(m+1)$–tuples each in $\mathbb{P}^N \times \mathbb{P}^{\epsilon\lambda_1 N} \times \ldots \times$

$\mathbb{P}^{\epsilon\lambda_m N}$ converging to $\mathbf{x}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$ as we increase N and adjust ϵ accordingly. Since $\mathrm{KL}(\cdot\|\cdot)$ is a continuous function over $\mathbb{D}^J \times \mathbb{D}^J$ we can find a sufficiently large N and $\mathbf{w}, \mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(m)}$ with Property (10). Finally, the second set is non–empty for whatever large δ since we can easily choose $\mathbf{v} \in \mathbb{D}^J$ having one coordinate sufficiently close to zero which in turn makes the divergence sufficiently big.

The following theorem says that for a sufficiently large universe and a sufficiently small proportion of observed examples the number of ways of observing $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(m)}$ from $\mathbf{w}$ is overwhelmingly higher than the number of ways of observing generated in the second set.

Theorem 5. *Let $\mathbf{w}, \mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(m)}$ be as defined in Property (10). Then for any real number $\delta > 0$ there are a natural number N and a real number $\epsilon > 0$ such that*

$$\lim_{n\to\infty} \frac{\sum_{(\mathbf{v},\mathbf{v}^{(1)},\ldots,\mathbf{v}^{(m)})\in(\mathbb{P}^{nN}\times\mathbb{P}^{\epsilon\lambda_1 nN}\times\ldots\times\mathbb{P}^{\epsilon\lambda_m nN})\cap W'_\delta} \sharp^{\vec{\lambda}\epsilon}_{nN}(\mathbf{v}|\mathbf{v}^{(1)},\ldots,\mathbf{v}^{(m)})}{\sharp^{\vec{\lambda}\epsilon}_{nN}(\mathbf{w}|\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(m)})} = 0.$$

Proof. As in Theorem 3 the argument is about showing that Lemma 4 gives much stronger predominance of ways of observing $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(m)}$ from $\mathbf{w}$ than the number of all ways of observing that could be generated in W'_δ. And as before, there are certainly less than $(nN)^J$ partitions over a universe containing nN elements and less than $(\epsilon nN)^J$ partitions over a universe containing ϵnN elements so there are less than $(nN)^J(\epsilon nN)^{Jm}$ different $(m+1)$–tuples.

Now consider d defined by Formula (7). By Taylor's formula $\log(1-\frac{w_j^{(i)}}{w_j}\epsilon\lambda_i) = -\frac{w_j^{(i)}}{w_j}\epsilon\lambda_i + \xi\epsilon^2$ for some constant ξ whenever $0 < \frac{w_j^{(i)}}{w_j}\lambda_i\epsilon < 1$. This holds since $w_j N > w_j^{(i)}\epsilon\lambda_i N > 0$ by the definition. Therefore

$$\sum_{j=1}^{J}\sum_{i=1}^{m}\Big(w_j \log\Big(1-\frac{w_j^{(i)}\epsilon\lambda_i}{w_j}\Big) - v_j\log\Big(1-\frac{v_j^{(i)}\epsilon\lambda_i}{v_j}\Big)\Big) =$$

$$= \sum_{j=1}^{J}\sum_{i=1}^{m}\Big(-w_j\frac{w_j^{(i)}}{w_j}\epsilon\lambda_i + v_j\frac{v_j^{(i)}}{v_j}\epsilon\lambda_i\Big) + \xi\epsilon^2 = \xi\epsilon^2,$$

where ξ is some constant. The last equality above is due to the fact that both $\sum_{j=1}^{J} w_j^{(i)} = 1$ and $\sum_{j=1}^{J} v_j^{(i)} = 1$ for all $1 \leq i \leq m$. Note that the term above is of order ϵ^2 while

$$\sum_{j=1}^{J}\sum_{i=1}^{m}\Big(w_j^{(i)}\epsilon\lambda_i \log\frac{w_j^{(i)}\lambda_i}{w_j - w_j^{(i)}\epsilon\lambda_i} - v_j^{(i)}\epsilon\lambda_i \log\frac{v_j^{(i)}\lambda_i}{v_j - v_j^{(i)}\epsilon\lambda_i}\Big), \tag{11}$$

which is equal to

$$\sum_{j=1}^{J}\sum_{i=1}^{m}\Big(w_j^{(i)}\epsilon\lambda_i \log\frac{w_j^{(i)}}{w_j - w_j^{(i)}\epsilon\lambda_i} - v_j^{(i)}\epsilon\lambda_i \log\frac{v_j^{(i)}}{v_j - v_j^{(i)}\epsilon\lambda_i}\Big), \tag{12}$$

is only of order ϵ. The equality between Formulas (11) and (12) holds since

$$\sum_{j=1}^{J}\sum_{i=1}^{m}\Big(w_j^{(i)}\epsilon\lambda_i \log\lambda_i - v_j^{(i)}\epsilon\lambda_i \log\lambda_i\Big) = 0$$

by $\sum_{j=1}^{J} w_j^{(i)} = 1$ and $\sum_{j=1}^{J} v_j^{(i)} = 1$, $1 \leq i \leq m$. Formula (12) together with

$$\sum_{i=1}^{m}\lambda_i \,\mathrm{KL}(\mathbf{v}^{(i)}\|\mathbf{v}) - \sum_{i=1}^{m}\lambda_i \,\mathrm{KL}(\mathbf{w}^{(i)}\|\mathbf{w}) > \frac{\delta}{2}$$

gives that there is a sufficiently small ϵ such that $d < -\epsilon\frac{\delta}{4} < 0$ and independent of n for all $(\mathbf{v}, \mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(m)}) \in (\mathbb{P}^{nN} \times \mathbb{P}^{\epsilon\lambda_1 nN} \times \ldots \times \mathbb{P}^{\epsilon\lambda_m nN}) \cap W'_\delta$. Then by Lemma 4

$$0 \leq \lim_{n\to\infty} \frac{\sum_{(\mathbf{v},\mathbf{v}^{(1)},\ldots,\mathbf{v}^{(m)})\in(\mathbb{P}^{nN}\times\mathbb{P}^{\epsilon\lambda_1 nN}\times\ldots\times\mathbb{P}^{\epsilon\lambda_m nN})\cap W'_\delta} \sharp_{nN}^{\vec{\lambda}\epsilon}(\mathbf{v}|\mathbf{v}^{(1)},\ldots,\mathbf{v}^{(m)})}{\sharp_{nN}^{\vec{\lambda}\epsilon}(\mathbf{w}|\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(m)})} =$$

$$= \lim_{n\to\infty}\Big[\sum_{(\mathbf{v},\mathbf{v}^{(1)},\ldots,\mathbf{v}^{(m)})\in(\mathbb{P}^{nN}\times\mathbb{P}^{\epsilon\lambda_1 nN}\times\ldots\times\mathbb{P}^{\epsilon\lambda_m nN})\cap W'_\delta} c(n,\vec{\mathbf{v}},\vec{\mathbf{w}})\cdot e^{nN\cdot d}\Big] \leq$$

$$\leq \lim_{n\to\infty}\Big[(nN)^J(\epsilon nN)^{Jm}\cdot e^{nN\cdot(-\epsilon\frac{\delta}{4})}\Big] = 0. \tag{13}$$

The second inequality above is due to the fact that $c(n, \vec{\mathbf{v}}, \vec{\mathbf{w}})$ is bounded. As in Theorem 3 this does not directly follow from Lemma 4 but it can be observed by further considering that we take $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(m)}$ from closed convex sets in $\mathbb{D}^J$, which does not allow zero coordinates. So every coordinate of every possible $\mathbf{v}^{(i)}$ is bounded away from zero by a constant. Finally, $\mathbf{v}$ is bounded away from zero since $v_j > v_j^{(i)} \epsilon \lambda_i$ for all $1 \leq j \leq J$ by the definition. Note that both ϵ and λ_i are non–zero constants. □

For a small δ the theorem above suggests that, in the presented setting, it would be quite unreasonable for experts to collectively choose any $\mathbf{v} \in \mathbb{D}^J$ as a possible partition on the set of all examples U that does not belong to $\Theta^{\mathrm{KL}}_{\vec{\lambda}}(W_1, \ldots, W_n)$, i.e.

$$\left\{ \arg\min_{\mathbf{v} \in \mathbb{D}^J} \sum_{i=1}^{m} \lambda_i \, \mathrm{KL}(\mathbf{w}^{(i)} \| \mathbf{v}); \text{ subject to } \mathbf{w}^{(1)} \in W_1, \ldots, \mathbf{w}^{(m)} \in W_m \right\}, \quad (14)$$

although, since there may be more probability functions in $\Theta^{\mathrm{KL}}_{\vec{\lambda}}(W_1, \ldots, W_n)$, it does not answer a question as to whether a specific point inside $\Theta^{\mathrm{KL}}_{\vec{\lambda}}(W_1, \ldots, W_n)$ could be more optimal than other points in $\Theta^{\mathrm{KL}}_{\vec{\lambda}}(W_1, \ldots, W_n)$. Recall that each expert only knows that the possible partition on the set of examples U_i that he or she observed is among those in W_i and that the proportion of examples in U_i in respect to the set of collectively observed examples $U_1 \cup \ldots \cup U_m$ is λ_i. Furthermore, in absence of any information about admissible partitions from experts the above analysis does not allow us to restrict possible partition on U. This however does not mean that such a restriction is not possible, only that it does not follow from our analysis.

It could be interesting to point out that even if we compute the number of ways of observing $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(m)}$ from $\mathbf{w}$ using the expression

$$\sharp^{\vec{\lambda}\epsilon}_{nN}(\mathbf{w} | \mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(m)}) \cdot (nN)^{k_{\mathbf{w}}},$$

where $k_{\mathbf{w}}$ is some constant, then we still obtain the same result as Formula (13) in the proof above becomes

$$\lim_{n\to\infty}\left[(nN)^J(\epsilon nN)^{Jm}(nN)^k\cdot e^{nN\cdot(-\epsilon\frac{\delta}{4})}\right]=0,$$

where k is a constant. So our equiprobability assumption could be actually changed to be slightly less restrictive (i.e. it is not necessary to strictly assume equiprobability of partitions as far as the proportion of probabilities of any two partitions is a polynomial function of the argument nN).

On the other hand, computing 'the number of models' instead of 'the number of ways of observing' using Formula (5), in Formula (12) in the proof above the term

$$-\sum_{j=1}^{J} v_j \log v_j - \Big(-\sum_{j=1}^{J} w_j \log w_j\Big)$$

appears. Since this term does not involve ϵ in any way, a partition $\mathbf{w}$ having higher entropy than $\mathbf{v}$ for sufficiently small ϵ forces the limit concerned to approach zero regardless of $W_1,\ldots,W_m$. Therefore we would be compelled to argue that the experts should always collectively choose the uniform probability function, the claim we made in Section 4.2, and we reiterate that in our opinion this original approach from [11] would be unreasonable in the case that we are interested in the overwhelmingly most likely partition on the set of all examples U.

5. Relationship to large deviations

In this paper we have addressed a particular problem of inferential statistics; given some frequency observations we infer the probability, in our case a probability function in a simplex $\mathbb{D}^J$. To introduce a similar problem, we may be given several results of rolling a dice and then asked to infer the probability of rolling a particular number. Although, as we have noted in Section 4.2, the frequencies we worked with were not measurable frequencies only a means of describing knowledge.

On the other hand, we may start with a probability and then ask how many times we expect to roll that number out of N trials as statisticians (or frequentists) would do. Our setting of drawing with replacement was extensively studied in the literature from this alternative point of view in the so–called study of large deviations. In particular, if we assume that the probability of drawing a ball of the colour j, $1 \leq j \leq J$, from a box with N balls is v_j then the probability of drawing Nw_j balls of the colour $1 \leq j \leq J$ respectively is given by Bernoulli's formula

$$\frac{N!}{\prod_{j=1}^{J}(Nw_j)!} \cdot \prod_{j=1}^{J}(Nv_j)^{Nw_j} = \sharp_N(\mathbf{w}) \cdot \prod_{j=1}^{J}(Nv_j)^{Nw_j}.$$

Notice that in our terminology $\mathbf{v}$ is a given probability function and $\mathbf{w}$ is a given partition. Now cloning the balls in the box n-times and applying Stirling's formula we obtain fairly similar mathematics to Lemma 1:

$$\sharp_{nN}(\mathbf{w}) \cdot \prod_{j=1}^{J}(nNv_j)^{nNw_j} =$$

$$= \frac{\sqrt{2\pi nN}(nN)^{nN}e^{-nN}e^{\frac{\xi}{4nN}}}{\prod_{j=1}^{J}\sqrt{2\pi nNw_j}(nNw_j)^{nNw_j}e^{-nNw_j}e^{\frac{\xi_j}{4nNw_j}}} \cdot \prod_{j=1}^{J}(nNv_j)^{nNw_j},$$

where $\xi, \xi_j \in [0,1]$, $1 \leq j \leq J$, are constants. By applying $\sum_{j=1}^{J} w_j = 1$ and $\sum_{j=1}^{J} v_j = 1$ we obtain $\prod_{j=1}^{J} nN^{nNw_j} = nN^{nN}$ and $\prod_{j=1}^{J} nN^{nNv_j} = nN^{nN}$ so the above can be rewritten as

$$\frac{\sqrt{2\pi nN}e^{\frac{\xi}{4nN}}}{\prod_{j=1}^{J}\sqrt{2\pi nNw_j}(w_j)^{nNw_j}e^{\frac{\xi_j}{4nNw_j}}} \cdot (nN)^{nN}\prod_{j=1}^{J}(v_j)^{nNw_j} =$$

$$= \frac{(nN)^{nN}\sqrt{2\pi nN}e^{\frac{\xi}{4nN}}}{\prod_{j=1}^{J}\sqrt{2\pi nN}} \cdot e^{-nN\,\mathrm{KL}(\mathbf{w}\|\mathbf{v})-\sum_{j=1}^{J}\frac{\xi_j}{4nNw_j}-\frac{1}{2}\sum_{j=1}^{J}\ln w_j}.$$

In this computation originally performed in 1957 by Sanov [19, Theorem 1] we have established the probability of a certain drawing expressed here by the

partition $\mathbf{w}$ given a probability function $\mathbf{v}$. This is a different problem to the one investigated in this paper where we were given observations and the task is to find $\mathbf{v}$. Nevertheless, the implications seem to be the same.

The higher the divergence $\mathrm{KL}(\mathbf{w}\|\mathbf{v})$ the smaller the probability of observing the partition $\mathbf{w}$ given n is sufficiently large. This is because the terms in the exponent other than the divergence are not getting further away from zero with increasing n and the term by which the exponentiation is multiplied is independent of $\mathbf{w}$. Now taking $\mathbf{v} = \mathbf{u}$, the uniform probability function, minimising the KL–divergence is nothing other than maximising the entropy H.

Furthermore, considering the probability of observing successively several partitions $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(m)}$ we obtain

$$\Big[\frac{(nN)^{nN}\sqrt{2\pi nN}e^{\frac{\xi}{4nN}}}{\prod_{j=1}^{J}\sqrt{2\pi nN}}\Big]^{m} \cdot e^{-nN\sum_{i=1}^{m}\mathrm{KL}(\mathbf{w}^{(i)}\|\mathbf{v})-\sum_{j=1}^{J}\sum_{i=1}^{m}\left[\frac{\xi_j^{(i)}}{4nNw_j^{(i)}}-\frac{1}{2}\ln w_j^{(i)}\right]}.$$

We could conclude that smaller the sum $\sum_{i=1}^{m}\mathrm{KL}(\mathbf{w}^{(i)}\|\mathbf{v})$ the higher the probability so the probability functions produced by the linear entropy operator have the highest possible probability. This is a rather neat result. There is symmetry between the expected frequencies given a probability and inferred probability given observed frequencies. Nevertheless, this is not something that we should take for granted.

As we have seen in the previous section, if we change the equiprobability assumptions in our Bayesian–styled argument in a specific way then we infer instead the uniform probability function starting from any observations. It appears that a Bayesian inference with its variety of prior probabilities allows us to deal with a larger range of problems. This perhaps somehow suggests that we may consider doing what E. T. Jaynes in [2, Chapter A] said people should have been doing after Laplace introduced his solution to Bernoulli's inverse probability problem although they instead turned to the sampling theory:

> "One might have thought, particularly in view of the great pragmatic success achieved by Laplace ..., that ... (the) next order of business

should have been seeking new and more general principles for determining prior probabilities, thus extending the range of problems where probability theory is useful ... ."

6. Relationship to multi–expert reasoning

In this section we point out how are the results obtained in Section 4 related to some existing observations and approaches in multi–expert reasoning. First, we will talk about special cases. Second, we will show what would happen should we let the weighting factors to vary freely. Finally, we will make a conclusion whether a single probabilistic merging operator could be ever preferred as optimal in general.

6.1. Special cases

First, consider that we are given a single closed convex set $W_1 \subseteq \mathbb{D}^J$ which fixes no coordinates to irrational coefficients and a partition (i.e. a probability function with rational coordinates) $\mathbf{w}$ of examples in a universe U. Now we can ask, given that the expert observed only a tiny fraction U_1 of all examples, what is actually the most likely partition of those examples in U_1 among those partitions that are also in W_1. A similar argument to the one presented in Section 4 suggests that we should choose as close to $\arg\min_{\mathbf{w}^{(1)} \in W_1} \mathrm{KL}(\mathbf{w}^{(1)} \| \mathbf{w})$, which is the KL-projection of $\mathbf{w}$ into W_1, as possible. This accords with the result by Paris [12] who used the maximum entropy inference process to justify such a choice in a similar setting. It is perhaps not surprising that we can argue in favour of the KL-projection using the mathematics behind the 'number of possible states' argument directly, without an intermediate step in the form of the maximum entropy inference process.

Second, in [15, Theorem 9] the linear entropy operator $\Theta^{\mathrm{KL}}_{\vec{\lambda}}$ was characterised as the set of all fixed points of a mapping $F : \mathbb{D}^J \to \mathbb{D}^J$ where $F(\mathbf{v})$ was defined as the weighted arithmetic mean (or the so called **LinOp**-pooling operator) with weights $\lambda_1, \ldots, \lambda_m$ respectively of the KL-projections of $\mathbf{v}$ into $W_1, \ldots, W_m$ respectively.

In general, a pooling operator is a mapping which maps m–tuple of probability functions $\mathbf{w}^{(1)} \in \mathbb{D}^J, \ldots, \mathbf{w}^{(m)} \in \mathbb{D}^J$ into a single probability function in $\mathbb{D}^J$. The **LinOp**-pooling operator with a weighting $\vec{\lambda} = (\lambda_1, \ldots, \lambda_m)$ is the weighted arithmetic mean of $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(m)}$, i.e

$$w_j = \sum_{i=1}^{m} \lambda_i w_j^{(i)},\ 1 \leq j \leq J.$$

We denote $\mathbf{w}$ defined above by $\mathbf{LinOp}_{\vec{\lambda}}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(m)})$.

It is easy to prove, e.g. a modification of [14, Lemma 4.4] or [15, Theorem 4], that, given fixed $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)} \in \mathbb{D}^J$, that $\mathbf{x}$ which minimises

$$\sum_{i=1}^{m} \lambda_i \, \mathrm{KL}(\mathbf{x}^{(i)} \| \mathbf{x})$$

is the weighted arithmetic mean of $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$, i.e. $\mathbf{LinOp}_{\vec{\lambda}}(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)})$. Applying Theorem 5 to the situation where each expert knows precisely the partition $\mathbf{w}^{(i)}$ on the set of examples U_i he or she observed thus proves the claim we made in connection to the ball observation problem at the beginning of Section 4.2 that the overwhelmingly the most likely partition of balls in the box is $\mathbf{LinOp}_{\vec{\lambda}}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(m)})$. Furthermore, if the proportion of collectively observed examples in comparison to the overall number is quite small (recall that this was expressed by the constant ϵ) then the proportion of examples observed by more than one expert is small as well in comparison to the cardinality of the set of collectively observed examples (in fact this ratio is ϵ). Thus the partition within the collectively observed examples $U_1 \cup \ldots \cup U_m$ is also close to the weighted arithmetic mean $\mathbf{LinOp}_{\vec{\lambda}}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(m)})$.

Finally, considering both observations above, we could argue that if we are given closed convex sets $W_1, \ldots, W_m$ which are fixed in no coordinates to irrational coefficients with respective weights $\lambda_1, \ldots, \lambda_m \in (0,1)$, $\sum_{i=1}^{m} \lambda_i = 1$, and establish $\mathbf{w} \in \Theta^{\mathrm{KL}}_{\vec{\lambda}}(W_1, \ldots, W_n)$ as a likely partition on U then a likely partition on the set of collectively observed examples $U_1 \cup \ldots \cup U_m$ is $\mathbf{LinOp}_{\vec{\lambda}}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(m)})$, where $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(m)}$ are the KL-projections of $\mathbf{w}$ into $W_1, \ldots, W_m$ respectively. And by the result [15, Theorem 9] mentioned

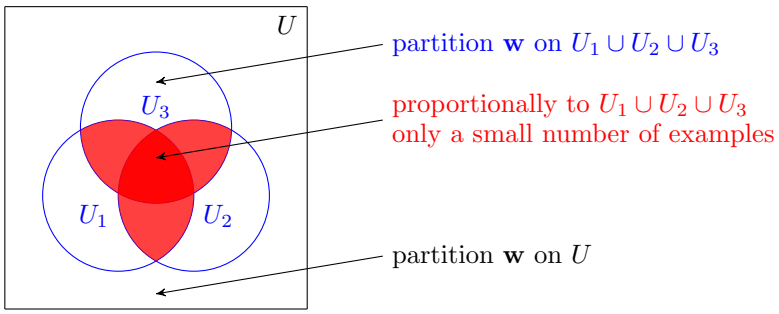

Figure 5: The intutition of our modification of the 'number of possible states' argument.

above this $\mathbf{LinOp}_{\vec{\lambda}}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(m)})$ is actually equal to $\mathbf{w}$, the partition on the set of all examples U. Our modification of the 'number of possible states' argument thus suggests the same partition on the set of collectively observed examples and on the set of all examples and this seems rational. See Figure 5 for an illustration.

The above holds if Formula (4) counting the number of ways of observing is chosen. A reader who finds Formula (5) counting the number of models more appealing (i.e. the approach similar to [11]) could be interested which partition on the set of collectively observed examples $U_1 \cup \ldots \cup U_m$ would be overwhelmingly the most likely. We already know that we would be compelled to assume the uniform probability function $\mathbf{u}$ on the set of all examples U. Therefore, going through our argumentation again, in Formula (12) $\mathbf{w}$ and $\mathbf{v}$ would be fixed to $\mathbf{u}$ which in turn for sufficiently small ϵ would give predominance of partitions $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(m)}$ which are such that each $\mathbf{w}^{(i)}$ is as close as possible to the most entropic point in W_i, i.e. $\mathbf{ME}(W_i)$. Since we can choose the proportion of examples observed by more than one expert in comparison to the set of collectively observed examples to be arbitrarily small (the constant ϵ), we could argue that the resulting partition on the set of collectively observed examples is overwhelmingly the most likely as close to $\mathbf{LinOp}_{\vec{\lambda}}(\mathbf{ME}(W_1), \ldots, \mathbf{ME}(W_m))$ as possible. It is worth mentioning that this corresponds to the result of a special probabilistic merging operator in the class of obdurate merging operators. Informally,

an obdurate merging operators first represent each of the sets $W_1, \ldots, W_m$ by a single probability function and then they apply a pooling operator on them. These operators unfortunately do not satisfy the so called consistency principle, a property that we would expect from a natural merging operator to satisfy; see [15, Section 2.2] for more details.

6.2. Variable weights

Throughout the paper we considered the relative proportions of observed examples $\lambda_1, \ldots, \lambda_m$ as fixed constants. If we however consider them as unknowns, still using Formula (5), in Formula (11) we must keep λ-s removed to get Formula (12); i.e. we must consider in the proof the following formula.

$$\sum_{j=1}^{J}\sum_{i=1}^{m}\Big(w_j^{(i)}\epsilon\lambda_i \log\frac{w_j^{(i)}\lambda_i}{w_j - w_j^{(i)}\epsilon\lambda_i} - v_j^{(i)}\epsilon\lambda_i \log\frac{v_j^{(i)}\lambda_i}{v_j - v_j^{(i)}\epsilon\lambda_i}\Big).$$

Since by Formula (5) we are compelled to assume that both $\mathbf{w}$ and $\mathbf{v}$ above are uniform probability functions similarly as in the proof of Theorem 5 by taking ϵ to be sufficiently small we can argue for predominance of partitions $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(m)}$ and weighting $\lambda_1, \ldots, \lambda_m$ that are as close as possible to the unique global maximisers of

$$-\sum_{j=1}^{J}\sum_{i=1}^{m} w_j^{(i)}\lambda_i \log(w_j^{(i)}\lambda_i)$$

subject to $\mathbf{w}^{(1)} \in W_1, \ldots, \mathbf{w}^{(m)} \in W_m$ and $\sum_{i=1}^{m}\lambda_i = 1$. Overwhelmingly the most likely partition on the set of collectively observed examples would be then as close to $\mathbf{LinOp}_{\vec{\lambda}}(\mathbf{ME}(W_1), \ldots, \mathbf{ME}(W_m))$ as possible, where

$$\lambda_i = \frac{e^{H(\mathbf{ME}(W_i))}}{\sum_{k=1}^{m} e^{H(\mathbf{ME}(W_k))}},$$

$1 \leq i \leq m$ [20, Theorem 4.4.1], and it would correspond to the result of the merging operator due to Kern–Isberner and Rödder [21]. Kern–Isberner and

Rödder derived this operator by applying the maximum entropy inference process directly to a particular setting which resembles very much our set of collectively observed examples if the partition on the set of all examples is uniform. It is perhaps not surprising that we have obtained the same operator by using the mathematics behind the 'number of possible states' argument directly. However we should note that this popular probabilistic merging operator is also obdurate [15, Section 2.2].

Nevertheless, the result of applying the original 'number of models' approach on the set of collectively observed examples is actually rather sensible, whether we fix the weights or not. We could argue that the uniformly distributed set of all examples is not our actual population; it is only a potential population. The actual population should have simply the same partition as is the partition on the set of collectively observed examples.

6.3. Conclusion

We have shown how applying the original argument from [11] in our multi–expert setting can produce some of the previously considered obdurate merging operators. Note that an obdurate merging operator produces always a singleton (a set containing a single probability function) and this modification involved first the single–expert 'number of models' argument and then a multi–expert consideration in form of the set of collectively observed examples. On the other hand, justification of the linear entropy operator involved only the multi–expert 'number of ways' argument and the representation $\Theta^{\mathrm{KL}}_{\vec{\lambda}}(W_1, \ldots, W_m)$ is in general a set. Note that if there is only one expert we have that $\Theta^{\mathrm{KL}}_{\vec{\lambda}}(W_1) = W_1$; we could say that the multi–expert 'number of ways' argument does not do anything here. One would perhaps suggest that we could reintroduce a single–expert argument as a second stage to choose a single point inside $\Theta^{\mathrm{KL}}_{\vec{\lambda}}(W_1, \ldots, W_m)$. This is basically the idea of a two–stage operator pioneered by Wilmers in [13]. In [15, Section 4.1] it has been suggested that a natural second stage for the linear entropy operator is the $\mathbf{CM}^{\infty}$ inference process, see [12] for a detailed explanation of this inference process. However, this two–stage operator does not

satisfy another natural requirement called the disagreement principle, which we would expect from a probabilistic merging operator to satisfy [15, Section 4.1]. And while the linear entropy operator satisfies both the consistency principle (the proof is trivial) and the disagreement principle (the proof is a simple modification of the proof for [20, Theorem 4.1.5] combined with [15, Theorem 9]) it does not in general produce a singleton. In [15, Theorem 22] it was shown that there is in fact no probabilistic merging operator producing always singletons and satisfying both the disagreement and consistency principles.

It seems that whichever approach we adopt there is some concession to make. Therefore, the approach we select should depend on the issues we would like to focus on. This is after all the original spirit of Shannon's approach to entropy; we might like to consider only some processes and find the entropic solution for them. Here we prefer to focus only on unknown processes causing higher than statistically expected difference in reported observations. In the following section we investigate how the linear entropy operator we have obtained by adopting our restricted approach could be applied.

7. Application

In this section we take a closer look on our motivation problem concerning meta–analysis with unexplained heterogeneity. Throughout the paper we have argued that in our specific modelling of this inferential problem the linear entropy operator returns partitions that are such that each is overwhelmingly more likely than all partitions not produced by this operator together. In this section we focus on whether the assumptions we adopted are practically achievable. Recall that those are the following:

1. Only a small proportion of examples have been observed by the studies; but both sets, the set of all examples and the set of collectively observed examples, are large.
2. There are unknown processes operating in the process of observation which

cause a higher than statistically expected difference in the reported observations in the form of admissible partitions.

3. The more ways the reported admissible partitions could have been established by studies from an unknown partition of the set of all examples, the more likely that (unknown) partition is (in absence of further information).

For any practical purposes Assumption 1 above means that the studies should have observed different large samples of an even larger population. For example, in Formula (13), if $J = 16$, $m = 4$, $nN = 7\,000\,000\,000$ and $\frac{\delta}{4} = 0.1$ then to obtain a reasonable predominance (i.e. at least something like 1 to 10^{10}) we will need the cardinality of the set of collectively observed examples ϵnN to be about 10 000 not yet considering any necessary adjustments due to the relationship of δ and ϵ in d and due to the factor $c(n, \vec{\mathbf{v}}, \vec{\mathbf{w}})$. Furthermore, $\frac{\delta}{4} = 0.1$ is for any practical purposes perhaps too high. Smaller delta will require even more observed examples, for example $\frac{\delta}{4} = 0.01$ requires about 120 000 collectively observed examples again not yet considering other issues from the proof.

However, for such numbers alone statistical (or frequentist) techniques offer more powerful results; consider sampling variance for such large samples, which is much narrower than what we obtain with our choice of δ above. Whenever we can use statistics, which is when studies are able to produce (statistically consistent) single probability functions (recall that partitions are special probability functions) to summarise their findings, we should use it rather than the method suggested here. Only when studies are not able to do so (i.e. when Assumption 2 holds), and they are able to produce closed convex non–empty sets of probability functions $W_1, \ldots, W_m \in \mathbb{D}^J$ respectively, then representing their findings by $\Theta^{\mathrm{KL}}_{\vec{\lambda}}(W_1, \ldots, W_n)$, i.e.

$$\Big\{ \arg\min_{\mathbf{v} \in \mathbb{D}^J} \sum_{i=1}^{m} \lambda_i \, \mathrm{KL}(\mathbf{w}^{(i)} \| \mathbf{v}), \text{ subject to } \mathbf{w}^{(1)} \in W_1, \ldots, \mathbf{w}^{(m)} \in W_m \Big\},$$

where $\lambda_1, \ldots, \lambda_m$ are the relative proportion of observed examples, seems as a possibly justified option given Assumptions 1, 2 and 3 above. Even if the

samples are too small to obtain the predominance result we can still argue that Assumption 3 suggests this representation using only Lemma 4.

Note that Assumption 2 is violated if the disagreement between studies could be explained by a known process. In practice, say if we observe patients, this happens when there are different methods of selecting patients or defining when a disease is observed. This is explained heterogeneity and combining such observations is considered at very least questionable if not wrong [4, Chapter 7].

Currently, if there is no heterogeneity among several studies then we summarise them using fixed–effect meta–analysis [3, Chapter 9]; we take the weighted arithmetic mean (with weights computed by a specific statistical technique from sample sizes) of single partitions reported in studies (say of patients according to mutually exclusive and exhaustive sets of symptoms).

If heterogeneity is not explained but it is only statistically detected then techniques of random–effect meta–analysis are usually applied. These rely on manipulating weights of the studies so as to put higher weights to small heterogeneous studies [3, Chapter 9]. The result that we have obtained using the modelling of drawing with replacement under the Assumptions 1, 2 and 3 above suggests that this may not be always justified. The weighted arithmetic mean produces the overwhelmingly the most likely estimate in our model of meta–analysis with unexplained heterogeneity.

Furthermore, the advantage of the representation proposed in this paper is that it can represent a wide range of related studies that can be represented by closed convex non–empty sets of probability functions $W_1, \ldots, W_m \in \mathbb{D}^J$ as opposite to single partitions of fixed–effect meta–analysis. This is the same advantage as the one of the recently developed Bayesian approach to meta–analysis [3, Chapters 8 and 16], which uses similar techniques to those employed in this paper. In particular, it defines a certain prior distribution depending on given problem and computes posterior distribution to represent studies. In this paper we have adopted Assumption 3; the prior distribution for us is that all possible ways in which a study could have obtained its observations are equally probable.

To explain the above mentioned advantage on an example, consider that we are interested in how effective a routine evaluation is in discovering cancer in patients with a specific venous condition. The number of studies directly measuring such a proportion could be fairly small. However, many other studies could provide us with valuable information on a possible proportion of cancer in the considered population, how effective is extensive screening in discovering cancer, etcetera. Those studies do not give us partitions we are interested in but they somehow restrict the possibilities what such a proportion could have been if they measured it. Surely, extensive screening is not going to reveal less cancers than routine evaluation and both would not reveal more cancers than the overall number of cancers in a specific population. This information encoded as a non–empty closed convex set of probability functions can be used in the proposed representation and this encoding could be perhaps performed as suggested in [22, Section 3], which is a modification of the encoding based on constraints due to Paris and Vencovská. Developing a real application would however involve exploring how to decide that statistically detected heterogeneity is unexplained, choosing a particular encoding by closed convex sets and comparing the results with existing methods. Here we at least demonstrate such an encoding on a toy example.

Say that there are two studies and all assumptions specified at the beginning of this section are satisfied. The first study observed that 10% of patients with a specific venous condition have cancer on a sample of 1 000. The second study observed that only 4% of such patients have cancer and this cancer is detected by a routine evaluation in 60% of cases. The sample size was 2 000. One way of representing these observations restricting admissible partitions leads to the following closed convex sets of probability functions: $W_1 = \{(w_1^{(1)}, w_2^{(1)}, w_3^{(1)}, w_4^{(1)}) \in \mathbb{D}^4;\ w_1^{(1)} + w_2^{(1)} = 0.1, w_1^{(1)} > 0.001, w_2^{(1)} > 0.001, w_3^{(1)} > 0.001, w_4^{(1)} > 0.001\}$ and $W_2 = \{(w_1^{(2)}, w_2^{(2)}, w_3^{(2)}, w_4^{(2)}) \in \mathbb{D}^4;\ w_1^{(2)} + w_2^{(2)} = 0.04, w_1^{(2)} = 0.6 \cdot (w_1^{(2)} + w_2^{(2)}), w_1^{(2)} > 0.001, w_2^{(2)} > 0.001, w_3^{(2)} > 0.001, w_4^{(2)} > 0.001\}$ where $w_1^{(i)}$, $w_2^{(i)}$, $w_3^{(i)}$ and $w_4^{(i)}$, $i = 1, 2$, express probabilities assigned to four elementary events 'cancer and detection', 'cancer and no detection', 'no cancer and detection'

and 'no cancer and no detection' respectively. Note that the difference between 10% and 4% cannot be statistically explained given the sample sizes. In this paper we have argued that we should represent the sets above by the set $\left\{ \arg\min_{\mathbf{v}\in\mathbb{D}^4} \left[\frac{1\,000}{3\,000} \cdot \mathrm{KL}(\mathbf{w}^{(1)}\|\mathbf{v}) + \frac{2\,000}{3\,000} \cdot \mathrm{KL}(\mathbf{w}^{(2)}\|\mathbf{v}) \right]; \text{ subject to } \mathbf{w}^{(1)} \in W_1 \text{ and } \mathbf{w}^{(2)} \in W_2 \right\}$.

We however would like point out to a reader who is eager to apply the presented methodology to represent several conflicting studies that it is indeed better to identify the reason why there is higher than statistically expected difference in studies and design a better study. While doing that, if answers are needed say because we need to treat patients, we may consider it as our best guess based on information that is available to us. In any case, before using the presented representation we need to analyse carefully whether it makes sense to do so (unexplained heterogeneity, a large population, big samples, needed answers). We therefore end this paper quoting what Jeff Paris said about the maximum entropy inference process [23] and ask for the same level of caution here.

> "Most of us would surely prefer modes of reasoning which we could follow blindly without being required to make much effort, ideally no effort at all. Unfortunately, Maxent is not such a paradigm; it requires us to understand the assumptions on which it is predicated and be constantly mindful of abusing them."

Acknowledgements

The author is indebted to George Wilmers and Alena Vencovská for many stimulating discussions that in fact inspired and helped to form this paper. Any errors are of course only mine alone. The author is also grateful for the support received from Assumption University of Thailand. Although the (European Community's) Seventh Framework Programme (FP7/2007 – 2013) did not directly support this paper, its creation would not be possible if they did not previously fund the research results from which this project benefits.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] J. Bernoulli, Ars conjectandi, opus posthumum. Accedit Tractatus de seriebus infinitis, et epistola gallicé scripta de ludo pilae reticularis, Thurneysen Brothers, Basel, 1713.

[2] E. T. Jaynes, Where do we stand on maximum entropy?, in: R. D. Levine, M. Tribus (Eds.), The Maximum Entropy Formalism, M.I.T. Press, 1979, pp. 15–118.

[3] Cochrane handbook for systematic reviews of interventions 5.1.0 [updated March 2011], in: J. P. T. Higgins, S. Green (Eds.), The Cochrane Collaboration, 2011.

[4] M. Petticrew, H. Roberts, Systematic reviews in the social sciences: a practical guide, Blackwell Publishing Ltd, Oxford, 2006.

[5] M. M. Al-khalaf, L. Thalib, S. A. R. Doi, Combining heterogenous studies using the random–effects model is a mistake and leads to inconclusive meta–analyses, Journal of Clinical Epidemiology 64 (2011) 119–123.

[6] L. Boltzmann, Über das wärmegleichgewicht zwischen mehratomigen gasmolekülen, Wiener Berichte 63 (1871) 397–418.

[7] L. Boltzmann, Einige allgemeine sätze über wärmegleichgewicht, Wiener Berichte 63 (1871) 679–711.

[8] L. Boltzmann, Analytischer beweis des zweiten haubtsatzes der mechanischen wärmetheorie aus den sätzen über das gleichgewicht der lebendigen kraft, Wiener Berichte 63 (1871) 712–732.

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

[10] C. E. Shannon, A mathematical theory of communication, The Bell System Technical Journal 27 (1948) 379–423, 623–656.

[11] J. B. Paris, A. Vencovská, On the applicability of maximum entropy to inexact reasoning, International Journal of Approximate Reasoning 3 (1989) 1–34.

[12] J. B. Paris, The uncertain reasoner companion, Cambridge University Press, Cambridge, 1994.

[13] G. M. Wilmers, A foundational approach to generalising the maximum entropy inference process to the multi–agent context, Entropy 17 (2015) 594–645.

[14] M. Adamčík, G. M. Wilmers, Probabilistic merging operators, Logique et Analyse 228 (2014) 563–590.

[15] M. Adamčík, The information geometry of Bregman divergences and some applications in multi–expert reasoning, Entropy 16 (2014) 6338–6381.

[16] M. Adamčík, Corrections on Adamčík, M. The information geometry of Bregman divergences and some applications in multi–expert reasoning, Entropy 2014, 16; 6338–6381 (March 2016). doi:10.13140/RG.2.1.2289.3201.

[17] S. Konieczny, R. Pino-Pérez, On the logic of merging, in: A. G. Cohn, L. Schubert, S. C. Shapiro (Eds.), Proceedings of the 6th International Conference on Principles of Knowledge Representation and Reasoning, 1998, pp. 488–498.

[18] G. M. Wilmers, Personal communication (September 2015).

[19] I. N. Sanov, On the probability of large deviations of random variables, Matematicheskii Sbornik 42 (1957) 11–44.

[20] M. Adamčík, Collective reasoning under uncertainty and inconsistency, Phd thesis, University of Manchester (2014).

[21] G. Kern-Isberner, W. Rödder, Belief revision and information fusion on optimum entropy, International Journal of Inteligent System 19 (2004) 837–857.

[22] M. Adamčík, A common point of disjoint closed convex sets and a prototype application in medical reasoning, in: S. Dhompongsa, N. Petrot, S. Plubtieng, S. Suantai (Eds.), Proceedings of the 9th International Conference on Nonlinear Analysis and Convex Analysis, Yokohama publishers, 2016, pp. 1–18.

[23] J. B. Paris, What you see is what you get, Entropy 16 (2014) 6186–6194.