A Representation Theorem and Applications

Manfred Jaeger

Max-Planck-Institut für Informatik
Stuhlsatzenhausweg 85, 66123 Saarbrücken
jaeger@mpi-sb.mpg.de

Abstract. We introduce a set of transformations on the set of all probability distributions over a finite state space, and show that these transformations are the only ones that preserve certain elementary probabilistic relationships. This result provides a new perspective on a variety of probabilistic inference problems in which invariance considerations play a role. Two particular applications we consider in this paper are the development of an equivariance-based approach to the problem of measure selection, and a new justification for Haldane's prior as the distribution that encodes prior ignorance about the parameter of a multinomial distribution.

1 Introduction

Many rationality principles for probabilistic and statistical inference are based on considerations of indifference and symmetry. An early expression of such a principle is Laplace's principle of insufficient reason: *"One regards two events as equally probable when one can see no reason that would make one more probable than the other, because, even though there is an unequal possibility between them, we know not which way, and this uncertainty makes us look on each as if it were as probable as the other"*(Laplace, Collected Works vol. VIII, cited after [3]). Principles of indifference only lead to straightforward rules for probability assessments when the task is to assign probabilities to a finite number of different alternatives, none of which is distinguished from the others by any information we have. In this case all alternatives will have to be assigned equal probabilities. Such a formalization of indifference by equiprobability becomes notoriously problematic when from state spaces of finitely many alternatives we turn to infinite state spaces: on countably infinite sets no uniform probability distributions exist, and on uncountably infinite sets the concept of uniformity becomes ambiguous (as evidenced by the famous Bertrand's paradox [6, 19]).

On (uncountably) infinite state spaces concepts of uniformity or indifference have to be formalized on the basis of certain transformations of the state space: two sets of states are to be considered equiprobable, if one can be transformed into the other using some natural transformation t. This, of course, raises the sticky question what transformations are to be considered as natural and probability-preserving. However, for a given state space, and a given class of probabilistic inference tasks, it often is possible to identify natural transformation, so that the solution to the inference tasks (which, in particular, can be probability assessments) should be invariant under the

transformations. The widely accepted resolution of Bertrand's paradox, for example, is based on such considerations of invariance under certain transformations.

In this paper we are concerned with probabilistic inference problems that pertain to probability distributions on finite state spaces, which are by far the most widely used type of distributions used for probabilistic modelling in artificial intelligence. As indicated above, when dealing with finite state spaces there does not seem to be any problem of capturing indifference principles with equiprobability. However, even though the underlying space of alternatives may be finite, the object of our study very often is the infinite set of probability distributions on that space, i.e. for the state space $S = \{s_1, \ldots, s_n\}$ the $(n-1)$-dimensional probability polytope

$$\Delta^n = \{(p_1, \ldots, p_n) \in \mathbb{R}^n \mid p_i \in [0,1], \sum_i p_i = 1\}.$$

The objective of this paper now can be formulated as follows: we investigate what natural transformations there exist of Δ^n, such that inference problems that pertain to Δ^n should be solved in a way that is invariant under these transformations. In section 2 we identify a unique class of transformations that can be regarded as most natural in that they alone preserve certain relevant relationships between points of Δ^n. In sections 3 and 4 we apply this result to the problems of measure selection and choice of Bayesian priors, respectively.

An extended version of this paper containing the proofs of theorems is available as [9].

2 Representation Theorem

The nature of the result we present in this section can best be explained by an analogy: suppose, for the sake of the argument, that the set of probability distributions we are concerned with is parameterized by the whole Euclidean space $\mathbb{R}^n$, rather than the subset Δ^n. Suppose, too, that all inputs and outputs for a given type of inference problem consist of objects (e.g. points, convex subsets, $\ldots$) in $\mathbb{R}^n$. In most cases, one would then probably require of a rational solution to the inference problem that it does not depend on the choice of the coordinate system; specifically, if all inputs are transformed by a translation, i.e. by adding some constant offset $\boldsymbol{r} \in \mathbb{R}^n$, then the outputs computed for the transformed inputs should be just the outputs computed for the original inputs, also translated by $\boldsymbol{r}$:

$$sol(\boldsymbol{i} + \boldsymbol{r}) = sol(\boldsymbol{i}) + \boldsymbol{r}, \tag{1}$$

where $\boldsymbol{i}$ stands for the inputs and *sol* for the solution of an inference problem. Condition (1) expresses an *equivariance principle*: when the problem is transformed in a certain way, then so should be its solution (not to be confused with *invariance principles* according to which certain things should be unaffected by a transformation).

The question we now address is the following: what simple, canonical transformations of the set Δ^n exist, so that for inference problems whose inputs and outputs are objects in Δ^n one would require an equivariance property analogous to (1)? Intuitively,

we are looking for transformations of Δ^n that can be seen as merely a change of coordinate system, and that leave all relevant geometric structures intact. The following definition collects some key concepts we will use.

Definition 1. *A* transformation *of a set S is any bijective mapping t of S onto itself. We often write ts rather than $t(s)$. For a probability distribution $\boldsymbol{p} = (p_1, \ldots, p_n) \in \Delta^n$ the set $\{i \in \{1, \ldots, n\} \mid p_i > 0\}$ is called the* set of support *of $\boldsymbol{p}$, denoted* support($\boldsymbol{p}$). *A transformation t of Δ^n is said to*

- preserve cardinalities of support *if for all $\boldsymbol{p}$:* $|\,\text{support}(\boldsymbol{p})\,| = |\,\text{support}(t\boldsymbol{p})\,|$
- preserve sets of support *if for all $\boldsymbol{p}$:* $\text{support}(\boldsymbol{p}) = \text{support}(t\boldsymbol{p})$.

A distribution $\boldsymbol{p}$ is called a mixture *of $\boldsymbol{p}'$ and $\boldsymbol{p}''$ if there exists $\lambda \in [0, 1]$ such that $\boldsymbol{p} = \lambda\boldsymbol{p}' + (1 - \lambda)\boldsymbol{p}''$ (in other words, $\boldsymbol{p}$ is a convex combination of $\boldsymbol{p}'$ and $\boldsymbol{p}''$). A transformation t is said to*

- preserve mixtures *if for all $\boldsymbol{p}, \boldsymbol{p}', \boldsymbol{p}''$: if $\boldsymbol{p}$ is a mixture of $\boldsymbol{p}'$ and $\boldsymbol{p}''$, then $t\boldsymbol{p}$ is a mixture of $t\boldsymbol{p}'$ and $t\boldsymbol{p}''$.*

The set of support of a distribution $\boldsymbol{p} \in \Delta^n$ can be seen as its most fundamental feature: it identifies the subset of states that are to be considered as possible at all, and thus identifies the relevant state space (as opposed to the formal state space S, which may contain states s_i that are effectively ruled out by $\boldsymbol{p}$ with $p_i = 0$). When the association of the components of a distribution $\boldsymbol{p}$ with the elements of the state space $S = \{s_1, \ldots, s_n\}$ is fixed, then $\boldsymbol{p}$ and $\boldsymbol{p}'$ with different sets of support represent completely incompatible probabilistic models that would not be transformed into one another by a natural transformation. In this case, therefore, one would require a transformation to preserve sets of support.

A *permutation* of Δ^n is a transformation that maps $(p_1, \ldots, p_n)$ to $(p_{\pi(1)}, \ldots, p_{\pi(n)})$, where π is a permutation of $\{1, \ldots, n\}$. Permutations preserve cardinalities of support, but not sets of support. Permutations of Δ^n are transformations that are required to preserve the semantics of the elements of Δ^n after a reordering of the state space S: if S is reordered according to a permutation π, then $\boldsymbol{p}$ and $\pi\boldsymbol{p}$ are the same probability distribution on S. Apart from this particular need for permutations, they do not seem to have any role as a meaningful transformation of Δ^n.

That a distribution $\boldsymbol{p}$ is a mixture of $\boldsymbol{p}'$ and $\boldsymbol{p}''$ is an elementary probabilistic relation between the three distributions. It expresses the fact that the probabilistic model $\boldsymbol{p}$ can arise as an approximation to a finer model that would distinguish the two distinct distributions $\boldsymbol{p}'$ and $\boldsymbol{p}''$ on S, each of which is appropriate in a separate context. For instance, $\boldsymbol{p}'$ and $\boldsymbol{p}''$ might be the distributions on $S = \{\textit{jam}, \textit{heavy traffic}, \textit{light traffic}\}$ that represent the travel conditions on weekdays and weekends, respectively. A mixture of the two then will represent the probabilities of travel conditions when no distinction is made between the different days of the week.

That a transformation preserves mixtures, thus, is a natural requirement that it does not destroy elementary probabilistic relationships. Obviously, preservation of mixtures immediately implies preservation of convexity, i.e. if t preserves mixtures and A is a convex subset of Δ^n, then tA also is convex.

We now introduce the class of transformations that we will be concerned with in the rest of this paper. We denote with $\mathbb{R}^+$ the set of positive real numbers.

Definition 2. *Let* $\boldsymbol{r} = (r_1, \ldots, r_n) \in (\mathbb{R}^+)^n$. *Define for* $\boldsymbol{p} = (p_1, \ldots, p_n) \in \Delta^n$

$$t_{\boldsymbol{r}}(\boldsymbol{p}) := (r_1 p_1, \ldots, r_n p_n) / \sum_{i=1}^{n} r_i p_i.$$

Also let $T_n := \{t_{\boldsymbol{r}} \mid \boldsymbol{r} \in (\mathbb{R}^+)^n\}$.

Note that we have $t_{\boldsymbol{r}} = t_{\boldsymbol{r}'}$ if $\boldsymbol{r}'$ is obtained from $\boldsymbol{r}$ by multiplying each component with a constant $a > 0$. We can now formulate our main result.

Theorem 1. *Let* $n \geq 3$ *and* t *be a transformation of* Δ^n.

(i) *t preserves sets of support and mixtures iff $t \in T_n$.*
(ii) *t preserves cardinalities of support and mixtures iff $t = t' \circ \pi$ for some permutation π and some $t' \in T_n$.*

The statements (i) and (ii) do not hold for $n = 2$: Δ^2 is just the interval $[0, 1]$, and every monotone bijection of $[0, 1]$ satisfies (i) and (ii). A weaker form of this theorem was already reported in [8]. The proof of the theorem closely follows the proof of the related representation theorem for collineations in projective geometry. The following example illustrates how transformations $t \in T_n$ can arise in practice.

Example 1. In a study of commuter habits it is undertaken to estimate the relative use of buses, private cars and bicycles as a means of transportation. To this end, a group of research assistants is sent out one day to perform a traffic count on a number of main roads into the city. They are given count sheets and short written instructions. Two different sets of instructions were produced in the preparation phase of the study: the first set advised the assistants to make one mark for every bus, car, and bicycle, respectively, in the appropriate column of the count sheet. The second (more challenging) set of instructions specified to make as many marks as there are actually people travelling in (respectively on) the observed vehicles. By accident, some of the assistants were handed instructions of the first kind, others those of the second kind.

Assume that on all roads being watched in the study, the average number of people travelling in a bus, car, or on a bicycle is the same, e.g. 10, 1.5, and 1.01, respectively. Also assume that the number of vehicles observed on each road is so large, that the actually observed numbers are very close to these averages.

Suppose, now, that we are more interested in the relative frequency of bus, car and bicycle use, rather than in absolute counts. Suppose, too, that we prefer the numbers that would have been produced by the use of the second set of instructions. If, then, an assistant hands in counts that were produced using the first set of instructions, and that show frequencies $\boldsymbol{f} = (f_1, f_2, f_3) \in \Delta^3$ for the three modes of transportation, then we obtain the frequencies we really want by applying the transformation $t_{\boldsymbol{r}}$ with $\boldsymbol{r} = (10, 1.5, 1.01)$. Conversely, if we prefer the first set of instructions, and are given frequencies generated by the second, we can transform them using $\boldsymbol{r}' = (1/10, 1/1.5, 1/1.01)$.

This example gives rise to a more general interpretation of transformations in T_n as analogues in discrete settings to rescalings, or changes of units of measurements, in a domain of continuous observables.

3 Equivariant Measure Selection

A fundamental probabilistic inference problem is the problem of *measure selection*: given some incomplete information about the true distribution $\boldsymbol{p}$ on S, what is the best rational hypothesis for the precise value of $\boldsymbol{p}$? This question takes on somewhat different aspects, depending on whether $\boldsymbol{p}$ is a statistical, observable probability, or a subjective degree of belief. In the first case, the "true" $\boldsymbol{p}$ describes actual long-run frequencies, which, in principle, given sufficient time and experimental resources, one could determine exactly. In the case of subjective probability, the "true" $\boldsymbol{p}$ is a rational belief state that an ideal intelligent agent would arrive at by properly taking into account all its actual, incomplete knowledge.

For statistical probabilities the process of measure selection can be seen as a prediction on the outcome of experiments that, for some reason, one is unable to actually conduct. For subjective probabilities measure selection can be seen as an introspective process of refining one's belief state. A first question then is whether the formal rules for measure selection should be the same in these two different contexts, and to which of the two scenarios our subsequent considerations pertain.

Following earlier suggestions of a frequentist basis for subjective probability [16, 1], this author holds that subjective probability is ultimately grounded in empirical observation, hence statistical probability [7]. In particular, in [7] the process of subjective measure selection is interpreted as a process very similar to statistical measure selection, namely a prediction on the outcome of hypothetical experiments (which, however, here even unlimited experimental resources may not permit us to carry out in practice). From this point of view, then, formal principles of measure selection will have to be the same for subjective and statistical probabilities, and our subsequent considerations apply to both cases. We note, however, that Paris [12] holds an opposing view, and sees no reason why his rationality principles for measure selection, which were developed for subjective probability, should also apply to statistical probability. On the other hand, in support of our own position, it may be remarked that the measure selection principles Shore and Johnson [18] postulate are very similar to those of Paris and Vencovská [15], but they were formulated with statistical probabilities in mind.

There are several ways how incomplete information about $\boldsymbol{p}$ can be represented. One common way is to identify incomplete information with some subset A of Δ^n: A is then regarded as the set of probability distributions $\boldsymbol{p}$ that are to be considered possible candidates for being the true distribution. Often A is assumed to be a closed and convex subset of Δ^n. This, in particular, will be the case when the incomplete information is given by a set of linear constraints on $\boldsymbol{p}$. In that case, A is the solution set of linear constraints, i.e. a polytope.

Example 2. (continuation of example 1) One of the research assistants has lost his count sheet on his way home. Unwilling to discard the data from the road watched by this assistant, the project leader tries to extract some information about the counts that the assistant might remember. The assistant is able to say that he observed at least 10 times as many cars as buses, and at least 5 times as many cars as buses and bicycles combined. The only way to enter the observation from this particular road into the study, however, is in the form of accurate relative frequencies of bus, car, and bicycle use. To this end,

the project leader has to make a best guess of the actual frequencies based on the linear constraints given to him by the assistant.

Common formulations of the measure selection problem now are: define a selection function *sel* that maps closed and convex subsets A of Δ^n (or, alternatively: polytopes in Δ^n; or: sets of linear constraints on $\boldsymbol{p}$) to distributions $\mathit{sel}(A) \in A$.

The most widely favored solution to the measure selection problem is the *entropy maximization* rule: define $\mathit{sel}_{me}(A)$ to be the distribution $\boldsymbol{p}$ in A that has maximal entropy (for closed and convex A this is well-defined). Axiomatic justifications for this selection rule are given in [18, 15]. Both these works postulate a number of formal principles that a selection rule should obey, and then proceed to show that entropy maximization is the only rule satisfying all the principles. Paris [13] argues that all these principles in essence are just expressions of one more general underlying principle, which is expressed by an informal statement (or slogan) by van Fraassen [19]: *Essentially similar problems should have essentially similar solutions*.

In spite of its mathematical sound derivation, entropy maximization does exhibit some behaviors that appear counterintuitive to many (see [8] for two illustrative examples). Often this counterintuitive behavior is due to the fact that the maximum entropy rule has a strong bias towards the uniform distribution $\boldsymbol{u} = (1/n, \ldots, 1/n)$. As $\boldsymbol{u}$ is the element in Δ^n with globally maximal entropy, $\boldsymbol{u}$ will be selected whenever $\boldsymbol{u} \in A$. Consider, for example, figure 1 (i) and (ii). Shown are two different subsets A and A' of Δ^3. Both contain $\boldsymbol{u}$, and therefore $\mathit{sel}_{me}(A) = \mathit{sel}_{me}(A') = \boldsymbol{u}$. While none of Paris' rationality principles explicitly demands that $\boldsymbol{u}$ should be selected whenever possible, there is one principle that directly implies the following for the sets depicted in figure 1: assuming that $\mathit{sel}(A) = \boldsymbol{u}$, and realizing that A' is a subset of A, one should also have $\mathit{sel}(A') = \boldsymbol{u}$. This is an instance of what Paris [14] calls the *obstinacy principle*: for any A, A' with $A' \subseteq A$ and $\mathit{sel}(A) \in A'$ it is required that $\mathit{sel}(A') = \mathit{sel}(A)$. The intuitive justification for this is that additional information (i.e. information that limits the previously considered distribution A to A') that is consistent with the previous default selection (i.e. $\mathit{sel}(A) \in A'$) should not lead us to revise this default selection. While quite convincing from a default reasoning perspective (in fact, it is a version of Gabbay's [2] *restricted monotonicity* principle), it is not entirely clear that this principle is an expression of the van Fraassen slogan. Indeed, at least from a geometric point of view, there does seem to exist little similarity between the two problems given by A and A', and thus the requirement that they should have similar solutions (or even the same solution) hardly seems a necessary consequence of the van Fraassen slogan.

An alternative selection rule that avoids some of the shortcomings of sel_{me} is the *center of mass* selection rule sel_{cm}: $\mathit{sel}_{cm}(A)$ is defined as the center of mass of A. With sel_{cm} one avoids the bias towards $\boldsymbol{u}$, and, more generally, the bias of sel_{me} towards points on the boundary of the input set A is reversed towards an exclusive preference for points in the interior of A. A great part of the intuitive appeal of sel_{cm} is probably owed to the fact that it satisfies (1), i.e. it is translation-equivariant.

Arguing that translations are not the right transformations to consider for Δ^n, however, we would prefer selection rules that are T_n-equivariant, i.e. for all A for which *sel*

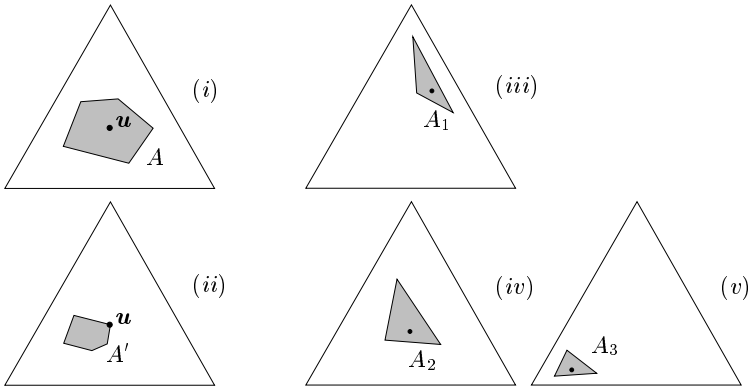

Fig. 1. Maximum Entropy and T_n- equivariant selection

is to be defined, and all $t_{\boldsymbol{r}} \in T_n$:

$$sel(t_{\boldsymbol{r}} A) = t_{\boldsymbol{r}} sel(A). \tag{2}$$

This, we would claim, is the pertinent (and succinct) formalization of the van Fraassen slogan for the measure selection problem. In fact, van Fraassen [19], after giving the informal slogan, proceeds to explain it further as a general equivariance principle of the form (1) and (2). The question, thus, is not so much whether this slogan is best captured as an equivariance requirement, but with which class of transformations the equivariance principle is to be instantiated. Interpreting theorem 1 as an identification of the transformations in T_n as the most "similarity preserving" transformations of Δ^n, we arrive at our answer that T_n-equivariance is the principle we require.

Figure 1 (iii)-(v) illustrates the T_n-equivariance principle: shown are three different transformations A_1, A_2, A_3 of a polytope defined by three linear constraints, and the corresponding transformations $\boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_3$ of one distinguished element inside the A_i. T_n-equivariance now demands that $sel(A_1) = \boldsymbol{p}_1 \Leftrightarrow sel(A_2) = \boldsymbol{p}_2 \Leftrightarrow sel(A_3) = \boldsymbol{p}_3$.

Example 3. (continuation of example 2) Assume that the unlucky assistant in example 2 was given instructions of the first type, and that he collected his data accordingly. If, instead, he had been given instructions of the second type, then the frequencies on the lost count sheet would have been frequencies $\boldsymbol{f}' = t_{\boldsymbol{r}} \boldsymbol{f}$, where $\boldsymbol{f}$ are the actual frequencies on the lost sheet, and $t_{\boldsymbol{r}}$ is as in example 1. The partial information he would then have been able to give also would have taken a different form. For instance, he might then have stated that he observed at least 6 times as many cars as buses, and at least 4.5 times as many cars as buses and bicycles combined.

One can show [8] that under very natural modelling assumptions, there corresponds to the transformation $t_{\boldsymbol{r}}$ on Δ^n a dual transformation $\bar{t}_{\boldsymbol{r}}$ on the space of linear constraints, such that stating a constraint c for $\boldsymbol{p}$ corresponds to stating the constraint $\bar{t}_{\boldsymbol{r}} c$ for $t_{\boldsymbol{r}} \boldsymbol{p}$. The crucial assumption is *consistency preservation*, which, in our example, means that a constraint c the research assistant will state when the frequencies on the

lost count sheet are $\boldsymbol{f}$ is consistent for $\boldsymbol{f}$ (i.e. satisfied by $\boldsymbol{f}$) iff the constraint c' he would give for frequencies $\boldsymbol{f}'$ is consistent for $\boldsymbol{f}'$. The transformation $\bar{t}_{\boldsymbol{r}}$ can also be characterized by the condition: for all sets of constraints $\boldsymbol{c}$

$$\mathit{Sol}(\bar{t}_{\boldsymbol{r}}\boldsymbol{c}) = t_{\boldsymbol{r}}\mathit{Sol}(\boldsymbol{c}),$$

where *Sol* denotes the solution set.

When the project leader uses a T_n-equivariant selection rule for reconstructing the true frequencies from the information he is given, then the following two approaches will lead to the same solution, whatever set of instructions this particular assistant was using: 1: first infer the actual frequencies observed by the assistant by applying the selection rule to the given constraints, and then transform to the preferred type of frequencies. 2: first transform the given constraints so as to have them refer to the preferred type of frequencies (knowing that this should be done by applying the $\bar{t}_{\boldsymbol{r}}$ transformation), and then apply the selection rule.

T_n-equivariance imposes no restriction on what $\mathit{sel}(A_i)$ should be for any single A_i in figure 1. It only determines how the selections for the different A_i should be related. It thus is far from providing a unique selection rule, like the rationality principles of Paris and Vencovská [15]. On the other hand, we have not yet shown that T_n-equivariant selection rules even exist. In the remainder of this section we investigate the feasibility of defining T_n-equivariant selection rules, without making any attempts to find the best or most rational ones.

From (2) one immediately derives a limitation of possible T_n-equivariant selection rules: let $A = \Delta^n$ in (2). Then $t_{\boldsymbol{r}}A = A$ for every $t_{\boldsymbol{r}} \in T_n$, and equivariance demands that $t_{\boldsymbol{r}}\mathit{sel}(A) = \mathit{sel}(A)$ for all $t_{\boldsymbol{r}}$, i.e. $\mathit{sel}(A)$ has to be a fixpoint under all transformations. The only elements of Δ^n that have this property are the n vertices $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$, where $\boldsymbol{v}_i$ is the distribution that assigns unit probability to $s_i \in S$. Clearly a rule with $\mathit{sel}(\Delta^n) = \boldsymbol{v}_i$ for any particular i would be completely arbitrary, and could not be argued to follow any rationality principles (more technically, such a rule would not be *permutation equivariant*, which is another equivariance property one would demand in order to deal appropriately with reorderings of the state space, as discussed in section 2).

Similar problems arise whenever *sel* is to be applied to some $A \subseteq \Delta^n$ that is invariant under some transformations of T_n. To evade these difficulties, we focus in the following on sets that are not fixpoints under any transformations $t_{\boldsymbol{r}}$ (this restriction can be lifted by allowing selection rules that may also return subsets of A, rather than unique points in A). Let $\mathcal{A}$ denote the class of all $A \subseteq \Delta^n$ with $t_{\boldsymbol{r}}A \neq A$ for all $t_{\boldsymbol{r}} \in T_n$. One can show that $\mathcal{A}$ contains (among many others) all closed sets A that lie in the interior of Δ^n, i.e. $\mathit{support}(\boldsymbol{p}) = \{1, \ldots, n\}$ for all $\boldsymbol{p} \in A$. In the following example a T_n-equivariant selection rule is constructed for all convex $A \in \mathcal{A}$. This particular rule may not be a serious candidate for a best or most rational equivariant selection rule. However, it does have some intuitive appeal, and the method by which it is constructed illustrates a general strategy by which T_n-equivariant selection rules can be constructed.

Example 4. Let $\mathcal{A}^c$ denote the set of all convex $A \in \mathcal{A}$. On $\mathcal{A}^c$ an equivalence relation $\sim$ is defined by

$$A \sim A' \quad \Leftrightarrow \quad \exists t_{\boldsymbol{r}} \in T_n : \; A' = t_{\boldsymbol{r}}A.$$

The equivalence class $\mathit{orb}(A) := \{A' \mid A' \sim A\}$ $(= \{t_{\boldsymbol{r}} A \mid t_{\boldsymbol{r}} \in T_n\})$ is called the *orbit* of A (these are standard definitions). It is easy to verify that for $A \in \mathcal{A}$ also $\mathit{orb}(A) \subseteq \mathcal{A}$, and that for every $A' \in \mathit{orb}(A)$ there is a unique $t_{\boldsymbol{r}} \in T_n$ with $A' = t_{\boldsymbol{r}} A$ (here transformations are unique, but as observed above, this does not imply that the parameter $\boldsymbol{r}$ representing the transformation is unique).

Suppose that $\mathit{sel}(A) = \boldsymbol{p} = (p_1, \ldots, p_n)$. With $\boldsymbol{r} = (1/p_1, \ldots, 1/p_n)$ then $t_{\boldsymbol{r}} \boldsymbol{p} = \boldsymbol{u}$, and by equivariance $\mathit{sel}(t_{\boldsymbol{r}} A) = \boldsymbol{u}$. It follows that in every orbit there must be some set A' with $\mathit{sel}(A') = \boldsymbol{u}$. On the other hand, if $\mathit{sel}(A') = \boldsymbol{u}$, then this uniquely defines $\mathit{sel}(A)$ for all A in the orbit of A': $\mathit{sel}(A) = \boldsymbol{p}$, where $\boldsymbol{p} = t_{\boldsymbol{r}} \boldsymbol{u}$ with $t_{\boldsymbol{r}}$ the unique transformation with $t_{\boldsymbol{r}} A' = A$. One thus sees that the definition of an equivariant selection rule is equivalent to choosing for each orbit in $\mathcal{A}^c$ a representative A' for which $\mathit{sel}(A') = \boldsymbol{u}$ shall hold.

One can show that for each $A \in \mathcal{A}^c$ there exists exactly one $A' \in \mathit{orb}(A)$ for which $\boldsymbol{u}$ is the center of mass of A'. Combining the intuitive center-of-mass selection rule with the principle of T_n-equivariance, we thus arrive at the T_n-equivariant center-of-mass selection rule: $\mathit{sel}_{\mathit{equiv\text{-}cm}}(A) = \boldsymbol{p}$ iff $A = t_{\boldsymbol{r}} A'$, $\boldsymbol{u}$ is the center of mass of A', and $\boldsymbol{p} = t_{\boldsymbol{r}} \boldsymbol{u}$.

4 Noninformative Priors

Bayesian statistical inference requires that a prior probability distribution is specified on the set of parameters that determines a particular probability model. Herein lies the advantage of Bayesian methods, because this prior can encode domain knowledge that one has obtained before any data was observed. Often, however, one would like to choose a prior distribution that represents the absence of any knowledge: an ignorant or noninformative prior. The set Δ^n is the parameter set for the multinomial probability model (assuming some sample size N to be given). The question of what distribution on Δ^n represents a state of ignorance about this model has received much attention, but no conclusive answer seems to exist.

Three possible solutions that most often are considered are: the uniform distribution, i.e. the distribution that has a constant density c with respect to Lebesgue measure, Jeffreys' prior, which is given by the density $c \prod_i p_i^{-1/2}$ (where c is a suitable normalizing constant), and Haldane's prior, given by density $\prod_i p_i^{-1}$. Haldane's prior (so named because it seems to have first been suggested in [4]) is an improper prior, i.e. it has an infinite integral over Δ^n. All three distributions are Dirichlet distributions with parameters $(1, \ldots, 1)$, $(1/2, \ldots, 1/2)$, and $(0, \ldots, 0)$, respectively (in the case of Haldane's distribution, the usual definition of a Dirichlet distribution has to be extended so as to allow the parameters $(0, \ldots, 0)$). Schafer [17] considers all Dirichlet distributions with parameters $(\alpha, \ldots, \alpha)$ for $0 \leq \alpha \leq 1$ as possible candidates for a noninformative prior.

The justifications for identifying any particular distribution as the appropriate noninformative prior are typically based on invariance arguments: generally speaking, ignorance is argued to be invariant under certain problem transformations, and so the noninformative prior should be invariant under such problem transformations. There are different types of problem transformations one can consider, each leading to a different concept of invariance, and often leading to different results as to what constitutes a

noninformative prior (see [5] for a systematic overview). In particular, there exist strong invariance-based arguments both for Jeffreys' prior [11], and for Haldane's prior [10, 20]. In the following, we present additional arguments in support of Haldane's prior.

Example 5. (continuation of example 3) Assume that the true, long-term relative frequencies of bus, car, and bicycle use are the same on all roads at which the traffic count is conducted (under both counting methods). Then the counts obtained in the study are multinomial samples determined by a parameter $\boldsymbol{f}_1^* \in \Delta^3$ if the first set of instructions is used, and $\boldsymbol{f}_2^* \in \Delta^3$ if the second set of instructions is used. Suppose the project leader, before seeing any counts, feels completely unable to make any predictions on the results of the counts, i.e. he is completely ignorant about the parameters $\boldsymbol{f}_i^*$.

When the samples are large (i.e. a great number of vehicles are observed on every road), then the observed frequencies $\boldsymbol{f}$ obtained using instructions of type i are expected to be very close to the true parameter $\boldsymbol{f}_i^*$. The prior probability *Pr* assigned to a subset $A \subseteq \Delta^n$ then can be identified with a prior expectation of finding in the actual counts relative frequencies $\boldsymbol{f} \in A$. If this prior expectation is to express complete ignorance, then it must be the same for both sampling methods: being told by the first assistant returning with his counts that he had been using instructions of type 2 will have no influence on the project leader's expectations regarding the frequencies on this assistant's count sheet. In particular, merely seeing the counts handed in by this assistant will give the project leader no clue as to which instructions were used by this assistant.

The parameters $\boldsymbol{f}_i^*$ are related by $\boldsymbol{f}_2^* = t_{\boldsymbol{r}}\boldsymbol{f}_1^*$, where $t_{\boldsymbol{r}}$ is as in example 1. Having the same prior belief about $\boldsymbol{f}_2^*$ as about $\boldsymbol{f}_1^*$ means that for every $A \subseteq \Delta^3$ one has $Pr(A) = Pr(t_{\boldsymbol{r}}A)$. A noninformative prior, thus, should be invariant under the transformation $t_{\boldsymbol{r}}$. As the relation between $\boldsymbol{f}_1^*$ and $\boldsymbol{f}_2^*$ might also be given by some other transformation in T_n, this invariance should actually hold for all these transformations.

This example shows that invariance under T_n-transformations is a natural requirement for a noninformative prior. The next theorem states that this invariance property only holds for Haldane's prior. In the formulation of the theorem a little care has to be taken in dealing with the boundary of Δ^n, where the density of Haldane's prior is not defined. We therefore restrict the statement of the theorem to the prior on the interior of Δ^n, denoted *int*Δ^n.

Theorem 2. *Let* Pr *be a measure on* intΔ^n *with* $\Pr(\text{int}\Delta^n) > 0$ *and* $\Pr(A) < \infty$ *for all compact subsets* A *of* intΔ^n. Pr *is invariant under all transformations* $t_{\boldsymbol{r}} \in T_n$ *iff* Pr *has a density with respect to Lebesgue measure of the form* $c\prod_i p_i^{-1}$ *with some constant* $c > 0$.

It is instructive to compare the justification given to Haldane's prior by this theorem with the justification given by Jaynes [10]. Jaynes gives an intuitive interpretation of a noninformative prior as a distribution of beliefs about the true value of $\boldsymbol{p}$ that one would find in "a population in a state of total confusion": an individual i in the population believes the true value of $\boldsymbol{p}$ to be $\boldsymbol{p}_i \in \Delta^n$. The mixture of beliefs one finds in a population whose individuals base their beliefs on "different and conflicting information" corresponds to a noninformative prior on Δ^n. Supposing, now, that to all members of this population a new piece of evidence is given, and each individual changes its belief about

$\boldsymbol{p}$ by conditioning on this new evidence, then a new distribution of beliefs is obtained. By a suitable formalization of this scenario, Jaynes shows that a single individual's transition from an original belief θ to the new belief θ' is given by $\theta' = a\theta/(1 - \theta + a\theta)$ (Jaynes only considers the binary case, where $\theta \in [0, 1]$ takes the role of our $\boldsymbol{p} \in \Delta^n$). This can easily be seen as a transformation from our group T_2. Jaynes' argument now is that a collective state of total confusion will remain to be one of total confusion even after the new evidence has been assimilated by everyone, and so the belief distribution about θ in the population must be invariant under the transformation $\theta \mapsto \theta'$.

This justification, thus, derives a transformation of Δ^2 in a concrete scenario in which it seems intuitively reasonable to argue that a noninformative prior should be invariant under these transformations. This is similar to our argument for the invariance of a noninformative prior under the transformation $t_{\boldsymbol{r}}$ in example 5. Justifications of Haldane's (or any other) prior that are based on such specific scenarios, however, always leave the possibility open that similarly intuitive scenarios can be constructed which lead to other types of transformations, and hence to invariance-based justifications for other priors as noninformative. Theorems 1 and 2 together provide a perhaps more robust justification of Haldane's prior: any justification for a different prior which is based on invariance arguments under transformations of Δ^n must use transformations that do not have the conservation properties of definition 1, and therefore will tend to be less natural than the transformations on which the justification of Haldane's prior is based.

5 Conclusions

Many probabilistic inference problems that are characterized by a lack of information have to be solved on the basis of considerations of symmetries and invariances. These symmetries and invariances, in turn, can be defined in terms of transformations of the mathematical objects one encounters in the given type of inference problem.

The representation theorem we have derived provides a strong argument that in inference problems whose objects are elements and subsets of Δ^n, one should pay particular attention to invariances (and equivariances) under the transformations T_n. These transformations can be seen as the analogue in the space Δ^n of translations in the space $\mathbb{R}^n$.

One should be particularly aware of the fact that it usually does not make sense to simply restrict symmetry and invariance concepts that are appropriate in the space $\mathbb{R}^n$ to the subset Δ^n. A case in point is the problem of noninformative priors. In $\mathbb{R}^n$ Lebesgue measure is the canonical choice for an (improper) noninformative prior, because its invariance under translations makes it the unique (up to a constant) "uniform" distribution. Restricted to Δ^n, however, this distinction of Lebesgue measure does not carry much weight, as translations are not a meaningful transformation of Δ^n. Our results indicate that the choice of Haldane's prior for Δ^n is much more in line with the choice of Lebesgue measure on $\mathbb{R}^n$, than the choice of the "uniform" distribution, i.e. Lebesgue measure restricted to Δ^n.

In a similar vein, we have conjectured in section 3 that some of the intuitive appeal of the center-of-mass selection rule is its equivariance under translations. Again, how-

ever, translations are not the right transformations to consider in this context, and one therefore should aim to construct T_n-equivariant selection rules, as, for example, the T_n-equivariant modification of center-of-mass.

An interesting open question is how many of Paris and Vencovská's [15] rationality principles can be reconciled with T_n-equivariance. As the combination of all uniquely identifies maximum entropy selection, there must always be some that are violated by T_n-equivariant selection rules. Clearly the obstinacy principle is rather at odds with T_n-equivariance (though it is not immediately obvious that the two really are inconsistent). Can one find selection rules that satisfy most (or all) principles except obstinacy?

References

1. R. Carnap. *Logical Foundations of Probability*. The University of Chicago Press, 1950.
2. D. Gabbay. Theoretical foundations for nonmonotonic reasoning in expert systems. In K. Apt, editor, *Logics and Models of Cuncurrent Systems*. Springer-Verlag, Berlin, 1985.
3. I. Hacking. *The Emergence of Probability: a Philosophical Study of Early Ideas About Probability, Induction and Statistical Inference*. Cambridge University Press, 1975.
4. J.B.S. Haldane. A note on inverse probability. *Proceedings of the Cambridge Philosophical Society*, 28:55–61, 1932.
5. J. Hartigan. Invariant prior distributions. *Annals of Mathematical Statistics*, 35(2):836–845, 1964.
6. J. Holbrook and S. S. Kim. Bertrand's paradox revisited. *The Mathematical Intelligencer*, pages 16–19, 2000.
7. M. Jaeger. Minimum cross-entropy reasoning: A statistical justification. In Chris S. Mellish, editor, *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95)*, pages 1847–1852, Montréal, Canada, 1995. Morgan Kaufmann.
8. M. Jaeger. Constraints as data: A new perspective on inferring probabilities. In B. Nebel, editor, *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-01)*, pages 755–760, 2001.
9. M. Jaeger. A representation theorem and applications to measure selection and noninformative priors. Technical Report MPI-I-2003-2-002, Max-Planck-Institut für Informatik, 2003. in preparation.
10. E.T. Jaynes. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, 4(3):227–241, 1968.
11. H. Jeffreys. *Theory of Probability*. Oxford University Press, third edition edition, 1961.
12. J. Paris. On filling-in missing information in causal networks. Submitted to *Knowledge-based Systems*.
13. J. Paris. Common sense and maximum entropy. *Synthese*, 117:75–93, 1999.
14. J. B. Paris. *The Uncertain Reasoner's Companion*. Cambridge University Press, 1994.
15. J.B. Paris and A. Vencovská. A note on the inevitability of maximum entropy. *International Journal of Approximate Reasoning*, 4:183–223, 1990.
16. H. Reichenbach. *The Theory of Probability*. University of California Press, 1949.
17. J. L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC, 1997.
18. J.E. Shore and R.W. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, IT-26(1):26–37, 1980.
19. B. C. van Fraassen. *Laws and Symmetry*. Clarendon, 1989.
20. C. Villegas. On the representation of ignorance. *Journal of the American Statistical Association*, 72:651–654, 1977.