On Efficient Decision Preserving Translations of Score Systems into Probabilistic Systems

Manfred Schramm
Institut für Informatik
Technische Universität München
schramma@in.tum.de

Bertram Fronhöfer
Institut für Informatik
Ludwig-Maximilians-Universität München
fronhoefer@pit-systems.de

Abstract

Due to their simple applicability score systems are in widespread use as a tool for decision taking. As it is well-known that probabilistic systems are more powerful than Score systems, we would like to have automated translations into probabilistic systems, which overcome some of their limitations. Using a straightforward translation (as in [KR00], [FS01]) reveals some properties of score systems, but leads to an exponential number of probabilistic rules. The aim of this paper is to define two different translations into probabilistic systems, which keep the simplicity of a score system (i.e. they use the same amount of rules as the score system). Moreover, the resulting probabilistic systems show their structure more explicitely than score systems, and they are also open to the addition of further knowledge.

Keywords: Score Systems, Probabilistic Reasoning, Maximum Entropy, Automated Diagnosis, Independence, Bayesian Networks

1 Introduction

Medicine, industry and economy abound with problems of diagnosis (and decision). Exemplified by an example from medicine, such problems consist of a knowledge base of propositions about the problem domain (e.g. 'Appendicitis is usually accompanied by strong stomach aches'), the values of certain symptoms in an actual (diagnosis) case (e.g. 'The patient has stomach aches'), a wanted diagnosis (e.g. 'Does the patient suffer from appendicitis?'), and finally, a decision ('surgical intervention?').

To solve such problems, **Score Systems** are frequently developed in research labs, e.g. in medicine ([SP89, OYF95]), or are in wide-range use, e.g. in economics ([KR00]), whenever uncertain knowledge plays an important role with the kind of problem to be solved.

A Score System is based on a set of **attributes** (or **variables**) which have each a set of possible (attribute or variable) **values**.

> For instance, in the medical domain, there may be the symptom/attribute 'body temperature' with (discrete) values 'low', 'normal', 'high' and 'very high'. To each attribute value a numerical value — its **weight** or **score**— is assigned (see Table 1 for an example).

When applying a score system to a concrete case the scores corresponding to the observed attribute values are added up. If the obtained sum falls in a certain **score interval**, the decision associated to this interval is proposed.

> Thus, for instance, a proposal for a medical treatment is established on the basis of symptoms found with a patient and which are represented by a list of attribute values.

When applying a score system to an actual case of decision finding, we feel intuitively that a score system seems to make some kind of assumption about the unrelatedness of the

Example (Score System):

symptom / attribute	score, if yes
tenderness in RLQ	4.5
rebound tenderness	2.5
no micturition	2.0
continuous type of pain	2.0
number of leucocytes ≥ 10000	1.5
age < 50 years	1.5
relocation of pain to RLQ	1.0
rigidity	1.0

Table 1: 'Ohmann Score' [OFY+95] for the diagnosis of appendicitis: In case of negative answers the scores are zero. Patients are diagnosed as having appendicitis if score sum ≥ 12, they are interned in case of $6 - 12$, and are sent home in case of ≤ 6. (RLQ: right lower quadrant of abdomen, seen from the patient.)

variables, because it provides no means to adapt the scores assigned to the values of one variable in dependence of the selected values of other variables. In [FS01] we discussed some of their underlying assumptions. Here our aim is different: After defining the technical background (in Section 2), we recall (in Section 3) the straightforward translation ([KR00, FS01]) of score systems into probabilistic systems. In the two sections following thereafter we present two translations of a score system into a probabilistic system. The resulting probabilistic systems yield **equivalent decisions** (see section 2.7) and have a minimal set of rules, thus meeting the simplicity of score systems. Our first translation (section4) generates probabilistic rules of the form

'Given a person showing a symptom value s, the probability of the illness d is p'.

We prove the translation's consistency and equivalence (in terms of decisions on complete symptom vectors) to the original score system. Section 5 reconsiders these issues for a different translation, based on rules like

'If a person has illness d, the probability for him to show symptom value s is p'.

The paper concludes with a comparison of the properties of these translations, including a comparative view of the difference of score systems and probabilistic systems, where the latter, of course, are much less limited in representing complex relations between symptoms and diagnosis.

Examples in this paper are taken from the field of medical diagnosis. Often our general explanations are tinged by medical language, which — in our opinion — better adds to clarity, than a more neutral, but less suggestive language.

2 Technical Preliminaries

2.1 Variables

In order to express knowledge about the evidence of a disease in view of certain symptoms we use a finite set of **variables** — a set of symptom variables and a variable for the diagnosis — with each having a finite set of **values**.

The **symptom variables** describe properties / symptoms / attributes relevant for the diagnosis task, e.g. examination results in the medical case. We denote symptom variables by S_i ($1 \leq i \leq m$), and we identify them with the set of their values: $S_i = \{s_{ij} | 1 \leq j \leq k_i\}$. We denote by $\vec{s}$ a tuple of values (of different symptoms) $\langle s_1, \ldots, s_m \rangle$ with $s_i \in S_i$. (Theoretically, we will always assume $S_i \cap S_j = \emptyset$ for $i \neq j$, which however, is not handy in practical examples.) We further refer to values of variable S_i by s_i, s_i' or by the additional index j (like in s_{ij}).

The values of a **diagnosis variable** define a classification of the possible diagnostic results, e.g. in kinds of diseases, based on the values of the symptom variables. For our purpose it is sufficient to consider a single binary variable D with values $\{d, \overline{d}\}$. If the probabilistic system is extended, we can, of course, increase the number of values or add other diagnosis variables to the system (which is not possible in score systems, where D is implicit and therefore basically one-valued)[1].

2.2 Events

Symptoms and diagnosis define our tuple space $\Omega := S_1 \times \cdots \times S_m \times D$.

In order to define arbitrary subsets of events in this space, we consider Ω as an **event space** with its power set as **set of events** or **event algebra**.

Consequently, an expression $\langle s_1, \ldots, s_m, \delta \rangle$ with $s_i \in S_i$ and $\delta \in D$ is an **elementary event**[2] in Ω (All general **events** are sets of elementary events.).

By $\vec{s}$ we denote the event $\langle s_1, \ldots, s_m \rangle$ with $s_i \in S_i$, which corresponds to the set $\{s_1\} \times \ldots \times \{s_m\} \times D$, and it will be called an **elementary symptom event**. In addition, we denote by $\langle \vec{s}, \delta \rangle$ the elementary event $\langle s_1, \ldots, s_m, \delta \rangle$ with $\delta \in D$.

2.3 Conditional Judgment

We also write $E \mathrel{-\!\!\triangleright} E'$ for the **conditional event** $E'|E$ due to its similarity with the common sense implication 'if-then'. We make the convention to drop the parentheses if simple events occur in conditional events. Additional Convention: For sake of simplicity we will just write $\vec{s} \mathrel{-\!\!\triangleright} \delta$ instead of $\vec{s} \mathrel{-\!\!\triangleright} \langle \vec{s}, \delta \rangle$.

In addition to the event space, we require a **method of judgment** to be given, e.g. a **judgment function** on (all) the events.

Example: For the standard application of a diagnostic system we have to make a judgment of $\vec{s} \mathrel{-\!\!\triangleright} \delta$ which has (for the medical context) the common reading of :

> 'If I know a patient showing the symptom values $\vec{s}$, what can I say
> — in view of this knowledge — about his risk of having the illness δ ?'

2.4 Score Systems

Formally, a Score System can be defined as follows:

For each variable S_i exists a set $W_i = \{w_{i1}, \ldots, w_{ik_i}\}$ of nonnegative **weights** or **scores** and a bijective **score function** w_i with $w_i(s_{ij}) \in W_i$. We also have a **(global) score function** w defined as $w(\vec{s}) := \sum_{i=1}^{n} w_i(s_i)$.

[1] Of course, score systems can distinguish different degrees of **one** illness via their border values, but they cannot cope with really different diseases. (See e.g. [SP89] for attacking this problem via combining different score systems.)

[2] Sometimes an elementary event is also called a **full conjunction**.

<table>
<tr><th colspan="4">With $w_i(s_i) = i$ and $w_i(\overline{s_i}) = 0$ we receive for $w(\vec{s})$ the following values:</th></tr>
<tr><td>$w(\langle \overline{s_1}, \overline{s_2}, \overline{s_3} \rangle) = 0$</td><td>$w(\langle s_1, \overline{s_2}, \overline{s_3} \rangle) = 1$</td><td>$w(\langle \overline{s_1}, s_2, \overline{s_3} \rangle) = 2$</td><td>$w(\langle \overline{s_1}, \overline{s_2}, s_3 \rangle) = 3$</td></tr>
<tr><td>$w(\langle s_1, s_2, \overline{s_3} \rangle) = 3$</td><td>$w(\langle s_1, \overline{s_2}, s_3 \rangle) = 4$</td><td>$w(\langle \overline{s_1}, s_2, s_3 \rangle) = 5$</td><td>$w(\langle s_1, s_2, s_3 \rangle) = 6$</td></tr>
</table>

Table 2: An example of the global weighting function $w(\vec{s})$ for a score system with 3 binary variables S_i (whose values we denote as $S_i = \{s_i, \overline{s_i}\}$) together with the 3 score functions w_i.

Additionally, there are **score intervals** given by a set of **border values** $b_1 < \cdots < b_{k_T}$, a **decision variable** T with values $\{t_1, \ldots, t_{k_T}\}$ and a **decision function** t which maps a sum of scores $w(\vec{s})$ to t_i iff $b_{i-1} < w(\vec{s}) \leq b_i$ (with $b_0 := 0$). W.r.t.g we assume the smallest $w_i(s_{ij})$ to be zero (as otherwise we can subtract a certain amount from the weight of the symptom S_i and subtract this amount also from the border values[3]).
Tab. 2 introduces a simple example of a score system with 3 binary variables, which we will reuse in the following. For an example including a decision function see Tab. 1 above.

2.5 Probabilistic Systems

Probabilistic systems use a **P-measure** P, which (in our finite case) can be specified by mapping every elementary event to a nonnegative real number such that the sum of function values over all elementary events is equal to 1. Since every event E is a (unique) union of elementary events $e_1, \ldots, e_n$ we define $P(E) := \sum_{i=1}^{n} P(e_i)$ and use the conditional probability for the conditional statement (i.e. $P(s \mathrel{-\!\!\triangleright} \delta) := P(\langle s, \delta \rangle)/P(s)$).

2.6 Extending Probabilistic Systems by Indifference, Independence and Maximum Entropy

In many cases (including our translations) we will have incomplete knowledge about P and therefore in general an infinite number of probabilistic models (P-models) which fulfill our constraints. We then use additional principles to determine a unique P, since such a unique P-measure is necessary in order to obtain unique probabilistic decisions. More exactly, we choose the well-known **Maximum Entropy Method** [PV90, KI97], which extends the principles of **Indifference** and **Independence** ([SG95]). For short, the Maximum Entropy Method chooses a probability model with maximal entropy[4] from all probability measures, which fulfill the given constraints.
Notation:
As we know have knowledge from different sources, we want to denote where our rules come from. We therefore use

- the sign *'c'* to mark a (quantitative) statement as constraint, obtained via a translation from a score, and collected in a certain knowledge base (see e.g. eq. (4)).

[3] Assuming $b_1 \geq \min w(\vec{s})$ guarantees that the border values remain ≥ 0. This assumption is very natural as a border value $< \min w(\vec{s})$ makes no sense.

[4] The (Shannon) Entropy on probability vectors v is defined as $H(v) = -\sum_i v_i \cdot \log v_i$. (The base of the logarithm does not matter, in most cases ln is taken) . For constraints, linear in v (as in our cases here), the maximum of $H(v)$ is known to be unique.

- the sign *'me'* to mark a (qualitative) statement, valid if the Maximum Entropy Method is applied to a certain knowledge base (see e.g. eq. 3). If this kind of knowledge is used in a proof, we refer to the number of the corresponding equation (see. e.g. eq.(11)).
- the sign * to mark (quantitative) statements, which are valid under the maximum entropy distribution in an example for a certain knowledge base (see e.g. eq. Table 4 (right)).

2.7 Equivalence of decisions

We will, depending on the translation, calculate $P(\vec{s} \mathrel{-\!\!\triangleright} d)$ and prove this probability to be a strictly monotonic increasing function $f : [0, max_{scoresum}] \rightarrow [0,1]$ with $w(\vec{s}) \rightarrow f(w(\vec{s})) = P(\vec{s} \mathrel{-\!\!\triangleright} d)$ and $max_{scoresum} := \sum_i w_{ik_i}$. Given such a function[5], we have

$$w(\vec{s}) > w(\vec{s}\,') \Longleftrightarrow P(\vec{s} \mathrel{-\!\!\triangleright} d) > P(\vec{s}\,' \mathrel{-\!\!\triangleright} d) \tag{1}$$

The definition of the border values $\vec{b'}$ for the obtained probabilistic systems (given the border values $\vec{b}$ of the score system and f) is then straightforward, as we define $b'_z := f(b'_z)$. Using these border values, the probabilistic system yields the same decisions as the score system (given a complete symptom vector).

3 Translation $\mathcal{T}_{\vec{s},d}$

3.1 Rule Base

3.1.1 The Translation Rule

Translation $\mathcal{T}_{\vec{s},d}$ generates probabilistic rules for all elementary symptom events, i.e. all possible combinations of symptom values. As such a combination is expressed by $\vec{s}$, the rule base contains for every possible $\vec{s}$ the probabilistic rule[6]

$${}^{c}P_{\vec{S},D}(\vec{s} \mathrel{-\!\!\triangleright} d) = f(w(\vec{s})) \tag{2}$$

For f we choose a strictly monotonic function. We e.g. may define[7] $f(w(\vec{s})) := \frac{w(\vec{s})}{\widehat{w}_{max}}$ or[8] $f(w(\vec{s})) := \frac{2^{w(\vec{s})}}{2^{\widehat{w}_{max}}}$ where $\widehat{w}_{max} := \max\{w(\vec{s}) | \vec{s} \in \Sigma\}$.

[5] The extension to (partially or completely) unknown symptom values will not be discussed here, as this extension is implicit in probabilistic systems, e.g.

$$(P(s_1 \mathrel{-\!\!\triangleright} d) = P(s_1 \wedge s_2 \mathrel{-\!\!\triangleright} d) \cdot P(s_1 \mathrel{-\!\!\triangleright} s_2) + P(s_1 \wedge \overline{s_2} \mathrel{-\!\!\triangleright} d) \cdot P(s_1 \mathrel{-\!\!\triangleright} \overline{s_2}))$$

and score systems do normally not discuss this topic. (But see [FS01] for the consequences of introducing a weighted sum of all the values of a symptom in a probabilistic context).

[6] The c is for constraint on $P_{\vec{S},D}$, and the index $\vec{s},d$ is to identify different P-measures of different translations.

[7] thus implying a standard difference rule $P(\vec{s} \mathrel{-\!\!\triangleright} d) - P(\vec{s}_{[s_i \rightarrow s'_i]} \mathrel{-\!\!\triangleright} d) = const_{[s_i \rightarrow s'_i]}$ (see [FS01])

[8] thus implying a logarithmic difference rule $P(\vec{s} \mathrel{-\!\!\triangleright} d)/P(\vec{s}_{[s_i \rightarrow s'_i]} \mathrel{-\!\!\triangleright} d) = const_{[s_i \rightarrow s'_i]}$

3.1.2 Rules valid under the use of Maximum Entropy for $\mathcal{T}_{\vec{s},d}$

The following property (which is necessary to guarantee, that Eq. 2 is always defined) is valid if we complete the specification by the Maximum Entropy principle (and therefore not necessary to mention in our knowledge base):

$$^{me}P_{\vec{S},D}(\vec{s}) > 0 \quad \text{(Positivity)} \tag{3}$$

3.1.3 Size of the Rule Base

While a score system has a linear number of rules ($\sum_i k_i$), the rules base of $\mathcal{T}_{\vec{s},d}$ contains an **exponential** number of rules ($\prod_i k_i$) and can therefore not be recommended for practical use. But as shown in e.g. [FS01], we cannot exploit any general independence relations in this translation (which is also true if the specification is completed via Maximum Entropy) to reduce the amount of rules. We therefore propose to give up this translation and look for different ones (still equivalent according to their decisions), which allow to reduce the number of necessary rules via exploiting their independence relations.

3.2 Example

Continuing our Example from Table 2 for the rules of Eq. (2), we get $P_{D,S}(\vec{s} \longrightarrow\!\!\triangleright\, d)$ the probabilities shown in Table 3 (additionally related to $w(\vec{s})$). [9]

$^cP_{\vec{S},D}(\vec{s} \longrightarrow\!\!\triangleright\, d)$	$=$	$(\, w(\vec{s})\,)/\widehat{w}_{max}$	$w(\vec{s})$
$^cP_{\vec{S},D}(\langle \overline{s_1}, \overline{s_2}, \overline{s_3}\rangle \longrightarrow\!\!\triangleright\, d)$	$=$	0/6	0
$^cP_{\vec{S},D}(\langle s_1, \overline{s_2}, \overline{s_3}\rangle \longrightarrow\!\!\triangleright\, d)$	$=$	1/6	1
$^cP_{\vec{S},D}(\langle \overline{s_1}, s_2, \overline{s_3}\rangle \longrightarrow\!\!\triangleright\, d)$	$=$	2/6	2
$^cP_{\vec{S},D}(\langle s_1, s_2, \overline{s_3}\rangle \longrightarrow\!\!\triangleright\, d)$	$=$	3/6	3
$^cP_{\vec{S},D}(\langle \overline{s_1}, \overline{s_2}, s_3\rangle \longrightarrow\!\!\triangleright\, d)$	$=$	3/6	3
$^cP_{\vec{S},D}(\langle s_1, \overline{s_2}, s_3\rangle \longrightarrow\!\!\triangleright\, d)$	$=$	4/6	4
$^cP_{\vec{S},D}(\langle \overline{s_1}, s_2, s_3\rangle \longrightarrow\!\!\triangleright\, d)$	$=$	5/6	5
$^cP_{\vec{S},D}(\langle s_1, s_2, s_3\rangle \longrightarrow\!\!\triangleright\, d)$	$=$	6/6	6

Table 3: Knowledge Base (resp. Queries) in case of $\mathcal{T}_{\vec{s},d}$

3.3 Consistency of $\mathcal{T}_{\vec{s},d}$

Consistency means that there are probability measures, which fulfill all the constraints in our rule base. As our probabilistic rule base is a (partial) specification of a Bayesian Network with marginally independent symptoms as parent nodes and the illness node D as child node (see figure 1 (right)) (with all the rules from Eq. 2 inside D), the specification is known to be consistent.

[9] Please note that overlining is part of our succinct notation of values of binary variables and does not refer to a general concept of set complement. For the general case consider the translation rule (2).

3.4 Preservation of decisions

Using a strictly increasing monotonic function implies that Eq. (1) is valid. In case of e.g. $f(w(\vec{s})) := \frac{w(\vec{s})}{\widehat{w}_{max}}$ the probabilities of the diagnoses are the old scores normed by $\widehat{w}_{max}$. With $t_i/\widehat{w}_{max}$ as new border values a decision function of the score system can be easily adapted and we have equivalence of the decisions proposed by the score system and those proposed by a probabilistic system resulting from translation $\mathcal{T}_{\vec{s},d}$.

4 Translation $\mathcal{T}_{d,s}$

4.1 Rule Base

4.1.1 The Translation Rule

Translation $\mathcal{T}_{d,s}$ generates for each symptom value $s_{ij} \in S_i$ the following constraint [10]:

$$ {}^{c}P_{D,S}(d \longrightarrow\!\!\triangleright\, s_i) \;=\; \frac{2^{w_i(s_i)}}{\sum_{s_{ij}\in S_i} 2^{w_i(s_{ij})}} \tag{4}$$

With these rules the P-measure $P_{D,S}$ is of course not uniquely determined and we need additional principles such as Maximum Entropy[11] for completing the specification.

4.1.2 Rules valid under the use of Maximum Entropy

The following rules are valid if we complete the specification by Maximum Entropy (and therefore will not be mentioned in our knowledge base):

- ${}^{me}P_{D,S}(d) > 0$ (Positivity)
  which is necessary to guarantee, that Eq. 4 is always defined.

- As the rules of Eq. (4) introduce a difference between $P_{D,S}(\langle d, s_{ij}\rangle)$ and $P_{D,S}(\langle d, \overline{s_{ij}}\rangle)$, but no difference between $P_{D,S}(\langle \overline{d}, s_{ij}\rangle)$ and $P_{D,S}(\langle \overline{d}, \overline{s_{ij}}\rangle)$, the Maximum Entropy P-model obtained from the knowledge base will show 'conditional' indifference of the k_i values s_{ij}, i.e.
  $$ {}^{me}P_{D,S}(\overline{d} \longrightarrow\!\!\triangleright\, s_{ij}) = 1/k_i \tag{5}$$

- As the rule base does not contain a rule including more than one symptom at the same time (and thus does not draw a direct link between two or more symptoms in the dependence graph) the Maximum Entropy P-model obtained from the knowledge base (derived via Eq. (4)) satisfies (see e.g. [Sch96]) **conditional independence** of the symptom variables, given a value for D. Formally this means for $s_i \in S_i$

$$\begin{aligned} {}^{me}P_{D,S}(d \longrightarrow\!\!\triangleright\, \vec{s}) &= \prod_{i=1}^{m} P_{D,S}(d \longrightarrow\!\!\triangleright\, s_i) \\ &=_{(4)} \frac{2^{\sum_{i=1}^{m} w_i(s_i)}}{\prod_{i=1}^{m}\sum_{s_{ij}\in S_i} 2^{w_i(s_{ij})}} = \frac{2^{w(\vec{s})}}{\prod_{i=1}^{m}\sum_{s_{ij}\in S_i} 2^{w_i(s_{ij})}} \end{aligned} \tag{6}$$

[10] The proposed translation rule may be modified by varying the base value without affecting the following.

[11] See e.g. [PIT] for theory and implementation of the Maximum Entropy Method.

and for $\overline{d}$

$$ {}^{me}P_{D,S}(\overline{d} \mathrel{-\!\!\triangleright} \vec{s}) \quad = \quad \prod_{i=1}^{m} P_{D,S}(\overline{d} \mathrel{-\!\!\triangleright} s_i) =_{(5)} 1/\prod_{i=1}^{m} k_i \tag{7} $$

4.1.3 Size of Rule Base

When using $\mathcal{T}_{d,s}$ we obtain k_i rules for each variable S_i. Since each of these rules can be derived from the $k_i{-}1$ other rules, only $k_i{-}1$ rules must be specified for variable S_i, which means that we need to state $\sum_{i=1}^{m}(k_i{-}1)$ rules in case of m variables (which makes m rules for the case of m binary symptom variables as in our example of Tab. 4).

4.2 Example

Continuing our Example of Table 2, we receive (using Eq. (4)) the knowledge base of Table 4 (left). The probabilities for the queries (calculated by Maximum Entropy) $P^*_{D,S}(\vec{s} \mathrel{-\!\!\triangleright} d)$ are shown in Table 4 (right), additionally related to $w(\vec{s})$. [12] With Maximum Entropy we get in addition $P^*_{D,S}(\overline{d} \mathrel{-\!\!\triangleright} s_i) = P^*_{D,S}(\overline{d} \mathrel{-\!\!\triangleright} \overline{s_i}) = 0.500$ $(1 \leq i \leq 3)$ and $P^*_{D,S}(d) \approx 0.356$.

Knowledge Base

${}^{c}P_{D,S}(d \mathrel{-\!\!\triangleright} s_1) = \frac{2^1}{2^1+2^0}) = \frac{2}{3} = 0.667$
${}^{c}P_{D,S}(d \mathrel{-\!\!\triangleright} s_2) = \frac{2^2}{2^2+2^0}) = \frac{4}{5} = 0.800$
${}^{c}P_{D,S}(d \mathrel{-\!\!\triangleright} s_3) = \frac{2^3}{2^3+2^0}) = \frac{8}{9} = 0.889$

(For the number of rules see (4.1.2))

Possible Queries : $P_{D,S}(\vec{s} \mathrel{-\!\!\triangleright} d)$		$w(\vec{s})$
$P^*_{D,S}(\langle \overline{s_1}, \overline{s_2}, \overline{s_3} \rangle \mathrel{-\!\!\triangleright} d)$	≈ 0.032	0
$P^*_{D,S}(\langle s_1, \overline{s_2}, \overline{s_3} \rangle \mathrel{-\!\!\triangleright} d)$	≈ 0.061	1
$P^*_{D,S}(\langle \overline{s_1}, s_2, \overline{s_3} \rangle \mathrel{-\!\!\triangleright} d)$	≈ 0.116	2
$P^*_{D,S}(\langle s_1, s_2, \overline{s_3} \rangle \mathrel{-\!\!\triangleright} d)$	≈ 0.207	3
$P^*_{D,S}(\langle \overline{s_1}, \overline{s_2}, s_3 \rangle \mathrel{-\!\!\triangleright} d)$	≈ 0.207	3
$P^*_{D,S}(\langle s_1, \overline{s_2}, s_3 \rangle \mathrel{-\!\!\triangleright} d)$	≈ 0.344	4
$P^*_{D,S}(\langle \overline{s_1}, s_2, s_3 \rangle \mathrel{-\!\!\triangleright} d)$	≈ 0.512	5
$P^*_{D,S}(\langle s_1, s_2, s_3 \rangle \mathrel{-\!\!\triangleright} d)$	≈ 0.677	6

Table 4: Knowledge Base and Comparison of Score Sums and Probabilities in case of $\mathcal{T}_{d,s}$

4.3 Consistency of $\mathcal{T}_{d,s}$

The set of constraints is a partial specification of a simple Bayesian Network (expressing conditional independence) as drawn in Fig. 1 (left). The node D contains a probability for d (where any choice from $(0,1)$ is admissible), a node S_i contains the $k_i - 1$ rules ${}^{c}P_{D,S}(d \mathrel{-\!\!\triangleright} s_{ij})$, automatically completed by Maximum Entropy (see e.g. [Luk00]) or by the rules of Eq. (5). As every Bayesian Network defines a distribution, the set of constraints is consistent. ■

[12] Please note that overlining is part of our succinct notation of values of binary variables and does not refer to a general concept of set complement. For the general case consider the translation rule (4).

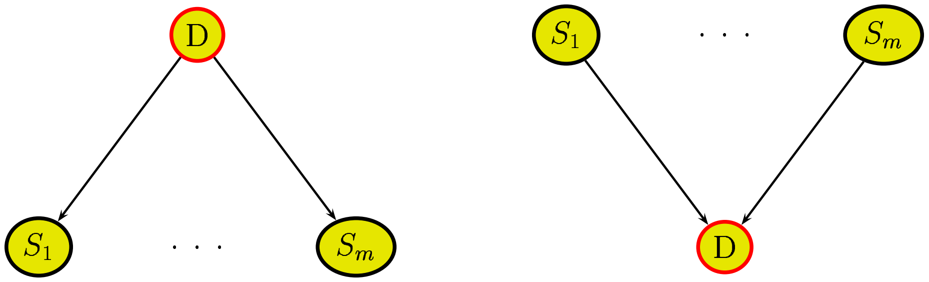

Figure 1: Conditional Independence of the variables S_i (given a value for D) (left) and Marginal Independence of the variables S_i (right) in Bayesian Network notation'

4.4 Preservation of decisions

In order to get the same decisions as with the score system, we have to show that the equivalence (1) holds for $P_{D,S}$ as well. This is proved as follows:
We first compute $P_{D,S}(\vec{s} \rightarrowtriangle d)$ for an arbitrary elementary symptom event $\vec{s}$.

$$\begin{aligned}
P_{D,S}(\vec{s} \rightarrowtriangle d) &= \frac{P_{D,S}(\langle \vec{s}, d\rangle)}{P_{D,S}(\langle \vec{s}, d\rangle) + P_{D,S}(\langle \vec{s}, \overline{d}\rangle)} && (8)\\
&= \frac{\dfrac{P_{D,S}(\langle \vec{s}, d\rangle)}{P_{D,S}(d)} \cdot P_{D,S}(d)}{\dfrac{P_{D,S}(\langle \vec{s}, d\rangle)}{P_{D,S}(d)} \cdot P_{D,S}(d) + \dfrac{P_{D,S}(\langle \vec{s}, \overline{d}\rangle))}{P_{D,S}(\overline{d})} \cdot P_{D,S}(\overline{d})} && (9)\\
&= \frac{P_{D,S}(d \rightarrowtriangle \vec{s})}{P_{D,S}(d \rightarrowtriangle \vec{s}) + P_{D,S}(\overline{d} \rightarrowtriangle \vec{s}) \cdot \dfrac{P_{D,S}(\overline{d})}{P_{D,S}(d)}} && (10)
\end{aligned}$$

With $c := P_{D,S}(\overline{d} \rightarrowtriangle \vec{s}) \cdot \dfrac{P_{D,S}(\overline{d})}{P_{D,S}(d)} =_{(7)} (\prod_{i=1}^{m} \dfrac{1}{k_i}) \cdot \dfrac{P_{D,S}(\overline{d})}{P_{D,S}(d)}$

being a constant value (for any choice of $P_{D,S}(d)$) and with

$$P_{D,S}(d \rightarrowtriangle \vec{s}) =_{(6)} \frac{2^{w(\vec{s})}}{\prod_{i=1}^{m}(\sum_{s_{ij} \in S_i} 2^{w_i(s_{ij})})} \qquad (11)$$

where $c' := \prod_{i=1}^{m}(\sum_{s_{ij} \in S_i} 2^{w_i(s_{ij})})$ is constant, we continue:

$$P_{D,S}(\vec{s} \rightarrowtriangle d) = \frac{\frac{2^{w(\vec{s})}}{c'}}{\frac{2^{w(\vec{s})}}{c'} + c} = \frac{2^{w(\vec{s})}}{2^{w(\vec{s})} + c \cdot c'}$$

which is sufficient for Eq. (1) to hold. ■

5 Translation $\mathcal{T}_{s,d}$

5.1 Rule Base

5.1.1 The Translation Rule

With translation $\mathcal{T}_{s,d}$ we get for each symptom value $s_i \in S_i$ the following rule[13]:

$$^{c}P_{S,D}(s_i \mathrel{-\!\!\triangleright} d) \;=\; \frac{2^{w_i(s_i)}}{1 + 2^{w_i(s_i)}} \tag{12}$$

With these rules the P-measure $P_{S,D}$ is not uniquely determined and we need additional principles such as Maximum Entropy[14] for completing the specification.

5.1.2 Rules valid under the use of Maximum Entropy for $\mathcal{T}_{s,d}$

The following properties are valid if we complete the specification by the Maxent principle (and will not be mentioned in our knowledge base):

- $^{me}P_{S,D}(s_i) > 0.$ (Positivity)
  which is necessary to guarantee, that Eq. 12 is always defined.

- Similar to $\mathcal{T}_{d,s}$, no rule from (12) connects different symptoms. So we have again in the Maxent Model — obtained from the knowledge base — **Conditional Independence** of the symptom variables given a value for d:

  $$^{me}P_{S,D}(d \mathrel{-\!\!\triangleright} \vec{s}) = \prod_{i=1}^{m} P_{S,D}(d \mathrel{-\!\!\triangleright} s_i) \tag{13}$$

  for $s_i \in S_i$ (and analogously for $\overline{d}$).

5.1.3 Size of the Rule Base

When using $\mathcal{T}_{s,d}$ we obtain k_i rules for each variable S_i , which means that we need to state $\sum_{i=1}^{m} k_i$ rules in case of m variables.

5.2 Example

For the score system from Tab. 2 we obtain with translation $\mathcal{T}_{s,d}$ the Tab. 5 showing the rule base (left) and (with Maximum Entropy) the values for our possible queries $P^{*}_{S,D}(\vec{s} \mathrel{-\!\!\triangleright} d)$ ((right), in comparism to the values of $w(\vec{s})$). [15]
For this example, Maximum Entropy chooses a value $P^{*}_{S,D}(d) \approx 0.607$, where this value was free to choose from $(0.5, \frac{2}{3})$ (see 5.3 for the explanation).

[13] The proposed translation rule may be modified by varying the base value without affecting the following. For adjusting such a system in practice, also a shift to negative weigths is possible.

[14] See [PIT] for theory and implementation of the Maximum Entropy Method.

[15] Please note that overlining is part of our succinct notation of values of binary variables. For many valued variables (more than two values) consider the translation rule (12).

Knowledge Base

${}^cP_{S,D}(s_1 \mathrel{-\!\!\rhd} d) = \frac{2^1}{1+2^1} = \frac{2}{3} = 0.667$
${}^cP_{S,D}(s_2 \mathrel{-\!\!\rhd} d) = \frac{2^2}{1+2^2} = \frac{4}{5} = 0.800$
${}^cP_{S,D}(s_3 \mathrel{-\!\!\rhd} d) = \frac{2^3}{1+2^3} = \frac{8}{9} = 0.889$
${}^cP_{S,D}(\overline{s_1} \mathrel{-\!\!\rhd} d) = \frac{2^0}{1+2^0} = \frac{1}{2} = 0.500$
${}^cP_{S,D}(\overline{s_2} \mathrel{-\!\!\rhd} d) = \frac{2^0}{1+2^0} = \frac{1}{2} = 0.500$
${}^cP_{S,D}(\overline{s_3} \mathrel{-\!\!\rhd} d) = \frac{2^0}{1+2^0} = \frac{1}{2} = 0.500$

Possible Queries : $P_{S,D}(\vec{s} \mathrel{-\!\!\rhd} d)$		$w(\vec{s})$
$P^*_{S,D}(\langle \overline{s_1}, \overline{s_2}, \overline{s_3} \rangle \mathrel{-\!\!\rhd} d)$	≈ 0.296	0
$P^*_{S,D}(\langle s_1, \overline{s_2}, \overline{s_3} \rangle \mathrel{-\!\!\rhd} d)$	≈ 0.456	1
$P^*_{S,D}(\langle \overline{s_1}, s_2, \overline{s_3} \rangle \mathrel{-\!\!\rhd} d)$	≈ 0.627	2
$P^*_{S,D}(\langle s_1, s_2, \overline{s_3} \rangle \mathrel{-\!\!\rhd} d)$	≈ 0.771	3
$P^*_{S,D}(\langle \overline{s_1}, \overline{s_2}, s_3 \rangle \mathrel{-\!\!\rhd} d)$	≈ 0.771	3
$P^*_{S,D}(\langle s_1, \overline{s_2}, s_3 \rangle \mathrel{-\!\!\rhd} d)$	≈ 0.871	4
$P^*_{S,D}(\langle \overline{s_1}, s_2, s_3 \rangle \mathrel{-\!\!\rhd} d)$	≈ 0.931	5
$P^*_{S,D}(\langle s_1, s_2, s_3 \rangle \mathrel{-\!\!\rhd} d)$	≈ 0.964	6

Table 5: Knowledge Base and Comparison of Score Sums and Probabilities

5.3 Consistency of $\mathcal{T}_{s,d}$

As the events $\langle s_i, d \rangle$ are conditionally independent given a value for D (see eq. (13) and compare fig. (1,left)), it is sufficient to choose a common $P_{S,D}(d)$ and then show the consistency of a single 'node' S_i, containing all the rules for S_i and the common value (shared by all nodes) for $P_{S,D}(d)$.
Every rule (from eq. 12) fixes the **relation** between $P_{S,D}(\langle s_{ij}, d \rangle)$ and $P_{S,D}(\langle s_{ij}, \overline{d} \rangle)$ for a certain s_{ij}, but leaves open the value $P_{S,D}(s_{ij})$, (beside respecting ${}^cP_{S,D}(s_{ij}) > 0$).
Now lets consider a certain node S_i which 'contains' k_i rules. Which set of values $P_{S,D}(d)$ can be reached by varying $P_{S,D}(s_{ij})$ in this node ?
As for every probability model we have the 'weighing' rule
$P(d) = \sum_{j=1}^{k_i} P(s_{ij} \mathrel{-\!\!\rhd} d) \cdot P(s_{ij})$, the subsystem of node S_i can assume every value for $P_{S,D}(d)$ between (not including) the lowest and the highest value of ${}^cP_{S,D}(s_{ij} \mathrel{-\!\!\rhd} d)$. Let x_i be this maximal value in a certain node S_i and $\vec{x}$ the vector of these values x_i for all nodes S_i. As the lowest probability is (w.r.t.g, see above) 0.5 for every symptom S_i, choose $P_{S,D}(d) \in (0.5, x_*)$, where x_* is the lowest value in $\vec{x}$ and every node S_i can configure itself to this value. By this choice every single node is consistent and furthermore, as the only common variable D has the same distribution in every node, the whole system is consistent.

5.4 Preservation of decisions

In order to get the same decisions as with the score system, we have to show that the equivalence (1) holds for $P_{S,D}$ as well. This is proved as follows:
We will compute $P_{S,D}(\vec{s} \mathrel{-\!\!\rhd} d)$ for an arbitrary elementary symptom event $\vec{s} = \langle s_1, \ldots, s_m \rangle$.

To prepare this calculation, we first recall that the diagnosis variable D is two valued here, so we have $P_{S,D}(s_i \mathrel{-\!\!\rhd} \overline{d}) = 1 - P_{S,D}(s_i \mathrel{-\!\!\rhd} d) = \frac{1}{1+2^{w(s_i)}}$. We state further that

$$\frac{P_{S,D}(\langle d, s_i \rangle)}{P_{S,D}(\langle \overline{d}, s_i \rangle)} = \frac{\dfrac{P_{S,D}(\langle d, s_i \rangle)}{P_{S,D}(s_i)}}{\dfrac{P_{S,D}(\langle \overline{d}, s_i \rangle)}{P_{S,D}(s_i)}} = \frac{{}^cP_{S,D}(s_i \mathrel{-\!\!\rhd} d)}{{}^cP_{S,D}(s_i \mathrel{-\!\!\rhd} \overline{d})} = 2^{w(s_i)} \tag{14}$$

We now are ready to calculate

$$\begin{aligned}\frac{P_{S,D}(\langle d,\vec{s}\rangle)}{P_{S,D}(\langle \bar{d},\vec{s}\rangle)} &= \frac{\frac{P_{S,D}(\langle d,\vec{s}\rangle)}{P_{S,D}(d)}}{\frac{P_{S,D}(\langle \bar{d},\vec{s}\rangle)}{P_{S,D}(\bar{d})}} \cdot \frac{P_{S,D}(d)}{P_{S,D}(\bar{d})} =_{(13)} \frac{(\prod_{j=1}^{m} \frac{P_{S,D}(\langle d,s_i\rangle)}{P_{S,D}(d)}) \cdot P_{S,D}(d)}{(\prod_{j=1}^{m} \frac{P_{S,D}(\langle \bar{d},s_i\rangle)}{P_{S,D}(\bar{d})}) \cdot P_{S,D}(\bar{d})} \\ &= \left(\prod_{j=1}^{m} \frac{P_{S,D}(\langle d,s_i\rangle)}{P_{S,D}(\langle \bar{d},s_i\rangle)}\right) \cdot \frac{P_{S,D}(\bar{d})^{m-1}}{P_{S,D}(d)^{m-1}} =_{(14)} 2^{\sum_{j=1}^{m} w(s_i)} \cdot c \quad := p_{d,\bar{d}}\end{aligned} \tag{15}$$

with a constant $c := \frac{P_{S,D}(\bar{d})^{m-1}}{P_{S,D}(d)^{m-1}}$. We now use this equation to continue with

$$\begin{aligned}P_{S,D}(\vec{s} \relbar\joinrel\rhd d) = \frac{P_{S,D}(\vec{s} \relbar\joinrel\rhd d)}{P_{S,D}(\vec{s} \relbar\joinrel\rhd \bar{d})} \cdot P_{S,D}(\vec{s} \relbar\joinrel\rhd \bar{d}) &= \frac{\frac{P_{S,D}(\langle \vec{s},d\rangle)}{P_{S,D}(\vec{s})}}{\frac{P_{S,D}(\langle \vec{s},\bar{d}\rangle)}{P_{S,D}(\vec{s})}} \cdot P_{S,D}(\vec{s} \relbar\joinrel\rhd \bar{d}) \\ &= \frac{P_{S,D}(\langle \vec{s},d\rangle)}{P_{S,D}(\langle \vec{s},\bar{d}\rangle)} \cdot (1 - P_{S,D}(\vec{s} \relbar\joinrel\rhd d))\end{aligned}$$

Solving this equation for $P_{S,D}(\vec{s} \relbar\joinrel\rhd d)$ yields

$$P_{S,D}(\vec{s} \relbar\joinrel\rhd d) = \frac{p_{d,\bar{d}}}{1 + p_{d,\bar{d}}} = \frac{c \cdot 2^{\sum_{i=1}^{m} w(s_i)}}{1 + c \cdot 2^{\sum_{i=1}^{m} w(s_i)}} = \frac{2^{\sum_{i=1}^{m} w(s_i)}}{\frac{1}{c} + 2^{\sum_{i=1}^{m} w(s_i)}} \tag{16}$$

which is a strictly monotonic increasing function of $w(\vec{s})$ and therefore sufficient for equation (1) to hold. ■

6 Discussion and Conclusion

By presenting 3 translations of a score system into a probabilistic system, we want to emphasize the superiority of probabilistic systems, which allow for a richer representation of the available knowledge while avoiding unnecessary complexity. Especially we find the following features worth mentioning:

- **Size of the Representation:** Probabilistic systems offer a rich language to relate symptoms and diagnoses. Of course, not every problem needs a complex and large set of rules. The simplicity of score systems can be reestablished in probabilistic systems of type $\mathcal{T}_{d,s}$ or $\mathcal{T}_{s,d}$: While $\mathcal{T}_{\vec{s},d}$ uses an exponential number of rules to map the decisions of a corresponding score system, $\mathcal{T}_{d,s}$ and $\mathcal{T}_{s,d}$ use a linear number. For simplicity of the rules, $\mathcal{T}_{d,s}$ and $\mathcal{T}_{s,d}$ do not reflect the 'difference property' of score systems [16], but use conditional independence, to reduce the necessary number of rules.

[16] in standard $P(\vec{s} \relbar\joinrel\rhd d) - P(\vec{s}_{[s_i \to s'_i]} \relbar\joinrel\rhd d) = const_{[s_i \to s'_i]}$ or a logarithmic form $P(\vec{s} \relbar\joinrel\rhd d)/P(\vec{s}_{[s_i \to s'_i]} \relbar\joinrel\rhd d) = const_{[s_i \to s'_i]}$ (see e.g. [FS01])

- **Ability to add additional information**
  - Probabilistic systems allow to state **complex relations between symptoms and illnesses**. Therefore, e.g. the well-known 'exclusive or' [17] is no problem for probabilistic systems (but for Score Systems)[18]
    $P(\langle s_1, s_2\rangle \mathrel{-\!\!\rhd} d) = P(\langle \overline{s_1}, \overline{s_2}\rangle \mathrel{-\!\!\rhd} d) = \mathit{high}$ but
    $P(\langle \overline{s_1}, s_2\rangle \mathrel{-\!\!\rhd} d) = P(\langle s_1, \overline{s_2}\rangle \mathrel{-\!\!\rhd} d) = \mathit{low}$.
    By the way of translation, most of this ability is lost in $\mathcal{T}_{\vec{s},d}$. On the other side, $\mathcal{T}_{d,s}$ and $\mathcal{T}_{s,d}$ allow the specification of e.g. 'contrary' knowledge.
    Example: Both knowledge bases can be extended by the rule

    $$P(\langle s_1, s_2, s_3\rangle \mathrel{-\!\!\rhd} d) = 0.2;$$

    without becoming inconsistent.
  - All three translations[19] offer a choice of the (a priori) **probability of the disease** d — which may be different at different locations — and the freedom to add information about the dependence of symptoms. Of course, the conditional independence, which is now present in $\mathcal{T}_{d,s}$ and $\mathcal{T}_{s,d}$, will get lost if we add relations between symptoms[20].
  - **Multiple diseases**
    Probabilistic systems are open to (consistently) combine knowledge about different illnesses in one system. As score systems work by adding more or less points, they are able to model within the same system different kinds of severity of a single disease, but are not able to model different diseases with partly similar, partly different symptoms. Moreover, it is not clear how different score systems might be integrated into a combined decision system.
- **Transparency**
  Probabilistic systems are more transparent than score systems, as the elements of their knowledge base may be understood (and checked) in terms of relative frequencies.
- **Adaptivity**
  Probabilistic systems provide a language, with is appropriate to formalize common sense knowledge about frequencies ([Gig96]). While $\mathcal{T}_{d,s}$ uses knowledge similar to 'cause → effect' relations as e.g. available in medical books, $\mathcal{T}_{s,d}$ uses 'symptom → cause' relations, as e.g. available from experienced practitioners or by mining in databases. Moreover, if carefully specified, these two kinds may both be used in one

[17] which asks for a simple, but non linear separating function on the data (in terms of the machine learning community).

[18] For just a few cases this can be treated in score systems via defining a new 'combined' symptom including the appropriate table of weights. But as these tables grow exponentially with the number of symptoms included, they hurt the aim of simplicity of score systems and do not offer a general solution. Probabilistic systems do not suffer from an exponential growth of the knowledge base in such cases.

[19] $\mathcal{T}_{s,d}$ in a restricted way, but remember that we can change the base or shift the scores, by which the full range for $P(d)$ can be reached.

[20] We remember, that using Maximum Entropy means to look for independences (and indifferences) which can be derived from the available knowledge (in the sense that the amount of additional information, which leads to those models, is only minimal)

knowledge base, thus making available different kinds of knowledge from different sources.

But probabilistic systems enhanced by Maximum Entropy methods do not only deliver a theoretically well justified method to cope with incomplete and uncertain knowledge (e.g. [PV90, Sch96]). For several years they have been available in powerful implementations ([PIT, SPI]) and have demonstrated their practical benefit in real world applications (e.g. [LEX99, SER01]).
We therefore recommend the use of probabilistic systems, which may be as simple as score systems, but may also grow and become as complex as necessary for the application.

References

[FS01] B. Fronhöfer and M. Schramm. A Probability Theoretic Analysis of Score Systems. In *Proc. of the Workshop Uncertainty in Artificial Intelligence at KI, Wien*, 2001.

[Gig96] G. Gigerenzer. The psychology of good judment: Frequency formats and simple algorithms. *Medical Decision Making*, 1996.

[KI97] Gabriele Kern-Isberner. Characterizing the principle of minimum cross-entropy within a conditional-logical framework. *AI*, pages 1–40, Oktober 1997.

[KR00] F. Kulmann and W. Rödder. Probabilistische Modellbildung auf der Basis von Scoring-Schemata. In *Symp. on Operations Research (SOR), Dresden.* LNCS, 2000.

[LEX99] LEXMED: Computer Aided Diagnosis of Appendicitis .
http://www.pit-systems.de/Lexmed or http://www.lexmed.de, June 1999.

[Luk00] Thomas Lukaszewicz. Credal networks under maximum entropy. *Uncertainty in AI*, pages 363–370, 2000.

[OFY$^+$95] C. Ohmann, C. Franke, Q. Yang, M.Margulies, M.Chan, P.J. van Elk, F.T. de Dombal, and H.-D. Röher. Diagnosescore für akute Appendizitis. *Der Chirurg*, 66:135–141, 1995.

[OYF95] C. Ohmann, Q. Yang, and C. Franke. Diagnostic scores for acute appendicitis. *Eur. J. Surg.*, 161:273–281, 1995.

[PIT] Homepage of PIT. http://www.pit-systems.de.

[PV90] J.B. Paris and A. Vencovska. A Note on the Inevitability of Maximum Entropy. *International Journal of Approximate Reasoning*, 3:183–223, 1990.

[Sch96] M. Schramm. *Indifferenz, Unabhängigkeit und maximale Entropie: Eine Wahrscheinlichkeitstheoretische Semantik für nichtmonotones Schließen.* Number 4 in Dissertationen zur Informatik. CS Press, München, 1996.

[Sch00] M. Schramm. Simulation von Scoresystemen durch probabilistische Systeme. Technical report, FH Ravensburg-Weingarten, May 2000.

[SER01] Manfred Schramm, Wolfgang Ertel, and Walter Rampf. Bestimmung der Wahrscheinlichkeit einer Appendizitis mit LEXMED. *Biomedical Journal*, 57:9–11, 2001.

[SG95] M. Schramm and M. Greiner. Foundations: Indifference, Independence & Maxent. In J. Skilling, editor, *Maximum Entropy and Bayesian Methods in Science and Engeneering (Proc. of the MaxEnt'94)*. Kluwer Academic Publishers, 1995.

[SP89] Martin F. Sturman and Manuel Perez. Computer-assisted diagnosis of acute abdominal pain. *Comprehensive Therapy*, 15(2):26–35, 1989. Multi Score System.

[SPI] Homepage of SPIRIT. http://www.xspirit.de.