Journal of Applied Logic ••• (••••) •••–•••

Contents lists available at ScienceDirect

Journal of Applied Logic

www.elsevier.com/locate/jal

# Qualitative probabilistic inference under varied entropy levels

Paul D. Thorn *, Gerhard Schurz

*Heinrich-Heine-University, Institute for Philosophy, Universitaetsstr. 1, 40225 Duesseldorf, Germany*

ARTICLE INFO

*Article history:*
Available online xxxx

*Keywords:*
Ampliative inference
Default reasoning
Non-monotonic reasoning
Probability logic

ABSTRACT

In previous work, we studied four well known systems of qualitative probabilistic inference, and presented data from computer simulations in an attempt to illustrate the performance of the systems. These simulations evaluated the four systems in terms of their tendency to license inference to accurate and informative conclusions, given incomplete information about a randomly selected probability distribution. In our earlier work, the procedure used in generating the unknown probability distribution (representing the true stochastic state of the world) tended to yield probability distributions with moderately high entropy levels. In the present article, we present data charting the performance of the four systems when reasoning in environments of various entropy levels. The results illustrate variations in the performance of the respective reasoning systems that derive from the entropy of the environment, and allow for a more inclusive assessment of the reliability and robustness of the four systems.

## 1. Balancing reward and risk in LP-reasoning in environments of different entropy

Systems of *logico-probabilistic* (LP) reasoning characterize inferences from conditionals that are interpreted as expressing high conditional probabilities. We formalize these conditionals using object-language statements of the form, A ⇒ B, which assert that P(B|A) is ‘high’. The probabilistic interpretation of uncertain conditionals was suggested in philosophy by Adams [1], and within the AI community by Pearl ([10], ch. 10; [11]), Lehmann and Magidor [8], and Goldszmidt and Pearl [3]. In the latter three papers, the focus lay on the interpretation of uncertain conditionals as corresponding to conditional probabilities taking values that are arbitrarily close to 1. In contrast, our work assumes a non-infinitesimal interpretation according to which the relevant conditional probabilities should be high (e.g., $\geq 0.9$) but not necessarily ‘extremely high’. Non-infinitesimal probability semantics is based on the “improbability-sum rule”, which is a less known element of Adams’ work, which has been elaborated in Schurz [13,14], and Schurz and Thorn [15].

[*] Corresponding author.
*E-mail addresses:* thorn@phil-fak.uni-duesseldorf.de (P.D. Thorn), schurz@phil-fak.uni-duesseldorf.de (G. Schurz).

http://dx.doi.org/10.1016/j.jal.2016.05.004
1570-8683/© 2016 Elsevier B.V. All rights reserved.

Reasoning with uncertain conditionals involves a balance between reward and risk. An LP system that licenses a greater number of inferences offers the opportunity of deriving more true and informative conclusions. But with this possible *reward* comes the *risk* of drawing more false conclusions, and more true but uninformative conclusions. In this article, we investigate four LP systems, the systems **O**, **P**, **Z**, and **QC**, which differ strongly in their tendency to make risky inferences. System **O** is the most cautious of the four systems, and infers a conclusion only if the conditional probability associated with that conclusion is guaranteed to be at least as high as the minimum conditional probability of the premises needed in drawing the conclusion. The other systems draw increasingly risky conclusions, with increasing risk as one proceeds from system **P**, to system **Z**, and finally to system **QC**.

In a previous paper [15], we investigated the performance of these reasoning systems in randomly chosen environments, with the unambiguous result that among the four systems, system **Z** achieved the best balance of reward and risk. In this article, we investigate the degree to which the performance of the four systems depends on the *entropy* of the environment which is reasoned about. Entropy (defined in the usual probabilistic way, see below) is a measure of the degree of the un-orderliness or irregularity of the environment. The lower the entropy, the more the probability function deviates from a uniform distribution, resulting in more high-probability regularities in the environment. We here present some interesting, and surprising results, concerning variations in the performance of respective LP-reasoning systems that derive from the entropy of the environment.

## 2. LP reasoning: systems O, P, Z, and QC

We represent the four LP systems considered here using a simple propositional language L, with the usual connectives $\neg$, $\wedge$, $\vee$, and $\supset$, and A, B, C, etc. as meta-logical variables ranging over arbitrary sentences of L. Our interest here will be in extensions of L by means of a default (or uncertain) conditional operator: $\Rightarrow$. We will concern ourselves exclusively with extensions of L by *simple* uncertain conditionals of the form $\mathrm{A} \Rightarrow \mathrm{B}$. Throughout the paper, $\alpha$ and $\beta$ will serve as meta-variables ranging over such simple conditional formulas, while $\Gamma$ ranges over sets of them. "$\vdash$" is used to denote derivability in classical logic, and "$\perp$" to denote an arbitrary contradiction. The four LP systems that we consider are ordered in terms of the number of inferences they license: $\mathbf{O} \subset \mathbf{P} \subset \mathbf{Z} \subset \mathbf{QC}$. We proceed by describing the weakest system first.

### 2.1. System O

System **O** is of interest because of its close connection to the following consequence relation:

(1) *Strict Preservation:* $\mathrm{A}_1 \Rightarrow \mathrm{B}_1, \ldots, \mathrm{A}_\mathrm{n} \Rightarrow \mathrm{B}_\mathrm{n} \Vdash_{\mathrm{s.p.}} \mathrm{C} \Rightarrow \mathrm{D}$ *iff* for all probability functions P (over L): $\mathrm{P}(\mathrm{D}|\mathrm{C}) \geq \min(\{\mathrm{P}(\mathrm{B}_\mathrm{i}|\mathrm{A}_\mathrm{i}) : 1 \leq \mathrm{i} \leq \mathrm{n}\})$.

System **O** was developed by Hawthorne [4] and Hawthorne and Makinson [5] as an inferential calculus for $\Vdash_{\mathrm{s.p.}}$. Throughout the present article, "$\vdash_{\mathbf{O}}$" denotes the syntactical notion of derivability in system **O**.

**System O** (after Hawthorne):

REF (reflexivity): $\vdash_{\mathbf{O}} \mathrm{A} \Rightarrow \mathrm{A}$
LLE (left logical equivalence): if $\vdash (\mathrm{A} \supset \mathrm{B}) \wedge (\mathrm{B} \supset \mathrm{A})$, then $\mathrm{A} \Rightarrow \mathrm{C} \vdash_{\mathbf{O}} \mathrm{B} \Rightarrow \mathrm{C}$
RW (right weakening): if $\vdash \mathrm{B} \supset \mathrm{C}$, then $\mathrm{A} \Rightarrow \mathrm{B} \vdash_{\mathbf{O}} \mathrm{A} \Rightarrow \mathrm{C}$
VCM (very cautious monotony): $\mathrm{A} \Rightarrow \mathrm{B} \wedge \mathrm{C} \vdash_{\mathbf{O}} \mathrm{A} \wedge \mathrm{B} \Rightarrow \mathrm{C}$
XOR (exclusive Or): if $\vdash \neg(\mathrm{A} \wedge \mathrm{B})$, then $\mathrm{A} \Rightarrow \mathrm{C}, \mathrm{B} \Rightarrow \mathrm{C} \vdash_{\mathbf{O}} \mathrm{A} \vee \mathrm{B} \Rightarrow \mathrm{C}$
WAND (weak And): $\mathrm{A} \Rightarrow \mathrm{B}, \mathrm{A} \wedge \neg\mathrm{C} \Rightarrow \perp \vdash_{\mathbf{O}} \mathrm{A} \Rightarrow \mathrm{B} \wedge \mathrm{C}$

It is easy to see that all of the rules of system **O** are correct with respect to $\Vdash_{\text{s.p.}}$, i.e., $\Gamma \vdash_{\mathbf{O}} A \Rightarrow B$ implies $\Gamma \Vdash_{\text{s.p.}} A \Rightarrow B$. While Hawthorne and Makinson [5] conjectured that $\vdash_{\mathbf{O}}$ might also be complete with respect to $\Vdash_{\text{s.p.}}$, Paris and Simmonds [9] have shown that this is not the case.

Following [15], we propose a marriage of system **O**, and a rule for inferring lower probability bounds that corresponds to the correctness of system **O** for $\Vdash_{\text{s.p.}}$. We employ statements of the form "$A \Rightarrow_r B$" to express that $P(B|A) \geq r$, and say that system **O** licenses the (valid) inference to $C \Rightarrow_{\min(\{r_i:1\leq i\leq n\})} D$ from $A_1 \Rightarrow_{r_1} B_1, \ldots, A_n \Rightarrow_{r_n} B_n$, whenever $A_1 \Rightarrow B_1, \ldots, A_n \Rightarrow B_n \vdash_{\mathbf{O}} C \Rightarrow D$. For example, given the premises $A \Rightarrow_{0.8} C$ and $\neg A \Rightarrow_{0.9} C$, we say that system **O** licenses the inference to $A \vee \neg A \Rightarrow_{0.8} C$, since $A \Rightarrow C$ and $\neg A \Rightarrow C \vdash_{\mathbf{O}} A \vee \neg A \Rightarrow C$ (by XOR), and $0.8 = \min\{0.8, 0.9\}$.

A noteworthy fact about system **O** is its weakness compared to standard systems of conditional logic. Segerberg [16], for example, argued the weakest 'reasonable' system of conditional logic includes REF, LLE, and RW, along with the following rule:

(AND): from $A \Rightarrow B$ and $A \Rightarrow C$ infer $A \Rightarrow B \wedge C$.

In comparison with WAND, observe that $A \Rightarrow C$ is **O**-derivable from $A \wedge \neg C \Rightarrow \bot$ (by RW, REF, and XOR), while $A \wedge \neg C \Rightarrow \bot$ is not **O**-derivable from $A \Rightarrow C$. So WAND is weaker than AND.[1] Moreover, by adding AND to the system **O**, one obtains (in one step) the well-known system **P**.

### *2.2. System* **P**

As described in [15], system **P** represents the confluence of a number of different semantic criteria. But the feature of system **P** that is of greatest interest here is its connection with the following consequence relation (cf. [1]):

(2) *Improbability-Sum Preservation:* $A_1 \Rightarrow B_1, \ldots, A_n \Rightarrow B_n \Vdash_{\text{i.s.p.}} C \Rightarrow D$ *iff* for all probability functions over L: $I(D|C) \leq \Sigma\{I(B_i|A_i) : 1 \leq i \leq n\}$, where $I(A|B)$ is defined as $1 - P(A|B)$.

Adams demonstrated that the following calculus (denoted by $\vdash_{\mathbf{P}}$) is correct and complete for $\Vdash_{\text{i.s.p.}}$.

**System P** (after Adams):

$\left.\begin{array}{l}\text{REF}\\ \text{LLE}\\ \text{RW}\end{array}\right\}$ as with system **O**

AND: as above

CC (cautious cut): $A \Rightarrow B,\ A \wedge B \Rightarrow C \vdash_{\mathbf{P}} A \Rightarrow C$

CM (cautious monotony): $A \Rightarrow B,\ A \Rightarrow C \vdash_{\mathbf{P}} A \wedge B \Rightarrow C$

OR: $A \Rightarrow C,\ B \Rightarrow C \vdash_{\mathbf{P}} A \vee B \Rightarrow C$

Following [15], we propose a marriage of system **P**, and a rule for inferring lower probability bounds that corresponds to the correctness of system **P** for $\Vdash_{\text{i.s.p.}}$, and say that system **P** licenses the (valid) inference to $C \Rightarrow_{1-\Sigma\{1-r_i:1\leq i\leq n\}} D$ from $A_1 \Rightarrow_{r_1} B_1, \ldots, A_n \Rightarrow_{r_n} B_n$, whenever $A_1 \Rightarrow B_1, \ldots, A_n \Rightarrow B_n \vdash_{\mathbf{P}} C \Rightarrow D$. For example, given the premises $A \Rightarrow_{0.8} B$, $A \wedge B \Rightarrow_{0.9} C$, we say that system **P** licenses the inference to $A \Rightarrow_{0.7} C$, since $A \Rightarrow B,\ A \wedge B \Rightarrow C \vdash_{\mathbf{P}} A \Rightarrow C$ (by CC), and $0.7 = 1 - ((1 - 0.8) + (1 - 0.9))$.

[1] We also observe that XOR is weaker than the rule OR (introduced below), and that VCM implies CM (below), in the presence of AND, while CM implies VCM (below), in the presence of RW ([5], 251).

### *2.3. System* **Z**

While system **P** sanctions many more inferences than system **O**, there are still many inferences not licensed by **P** that one might reasonably accept. For instance, **P** does not license inference via *subclass inheritance* based on default assumptions of *irrelevance* (or *independence*). For example, if we know that a given animal is a male bird ($\mathrm{B} \wedge \mathrm{M}$) and that birds can normally fly ($\mathrm{B} \Rightarrow \mathrm{F}$), and nothing else of relevance, then we would intuitively draw the conclusion that this male bird can fly (F). However, $\mathrm{B} \wedge \mathrm{M} \Rightarrow \mathrm{F}$ is not **P**-entailed by $\mathrm{B} \Rightarrow \mathrm{F}$, because there are probability distributions in which $\mathrm{P}(\mathrm{F}|\mathrm{B} \wedge \mathrm{M})$ is much smaller than $\mathrm{P}(\mathrm{F}|\mathrm{B})$. If we do infer $\mathrm{B} \wedge \mathrm{M} \Rightarrow \mathrm{F}$ from $\mathrm{B} \Rightarrow \mathrm{F}$, in such cases, then we assume, *by default*, that the additional factor M (in this case the gender of a bird) is *irrelevant* to its ability to fly, or in other words, M and F are assumed to be probabilistically independent given B. A straightforward means of enlarging the set of LP-derivable conditionals, in order to include such default inferences, is to give up the requirement that a reasonable inference be valid for *all possible* probability distributions, and consider only 'normal' probability distributions, i.e., distributions that satisfy the default assumption of irrelevance. An early suggestion for realizing this idea was the *maximum entropy* approach to default inference (cf. [6]; [10], 491–3). By selecting a probability distribution that maximizes entropy, one minimizes probabilistic dependences. Despite having some attractive features, the maximum entropy approach is rather complicated to apply, and has the further disadvantage of being dependent on the partition of possible world over which entropy is assessed (cf. [19]).

System **Z** of Pearl [11] and Goldszmidt and Pearl [3] maintains many of the advantages of the maximum entropy approach, while surmounting its disadvantages. Like the maximum entropy approach, inference in system **Z** proceeds via the construction of a semantic model of the premise conditionals that maximizes (probabilistic) independence. In system **Z**, this is achieved by maximizing the 'degree of normality' of the set of possible 'worlds' represented within a ranked world model, (W, r), consisting of a set of possible worlds W (i.e., maximal consistent sets of propositional formulae), and a ranking function, r, which takes each element of W to a natural number. A ranked world model is said to satisfy a conditional $\mathrm{A} \Rightarrow \mathrm{B}$ *iff* the rank of lowest ranked world that satisfies $\mathrm{A} \wedge \mathrm{B}$ is less than the rank of lowest ranked world that satisfies $\mathrm{A} \wedge \neg\mathrm{B}$. Similarly, a ranked world model is said to satisfy a set of conditionals $\Gamma$ *iff*, for all $\mathrm{A} \Rightarrow \mathrm{B}$ in $\Gamma$, the rank of lowest ranked world that satisfies $\mathrm{A} \wedge \mathrm{B}$ is less than the rank of lowest ranked world that satisfies $\mathrm{A} \wedge \neg\mathrm{B}$. The *z-model* for a set of conditionals $\Gamma$ is defined as follows (where a set of conditionals $\Gamma$ is called **P**-consistent *iff* $\Gamma$ does not **P**-entail $\neg\bot \Rightarrow \bot$):

(3) *Definition*: $(\mathrm{W}_\Gamma, \mathrm{z}_\Gamma)$ is the *z-model* for a **P**-consistent set of conditionals $\Gamma$ *iff* (i) $\mathrm{W}_\Gamma$ is the set of worlds generated by the propositional atoms appearing in $\Gamma$, (ii) $(\mathrm{W}_\Gamma, \mathrm{z}_\Gamma)$ satisfies $\Gamma$, and (iii) for any ranked world model, $(\mathrm{W}_\Gamma, \mathrm{r})$, that satisfies (ii), we have $\mathrm{z}_\Gamma(\mathrm{w}) \leq \mathrm{r}(\mathrm{w})$, for all w in $\mathrm{W}_\Gamma$ (cf. [11], section 1; [3], 65, fig. 2, 68, def. 15).

It is demonstrable that there is a unique such ranked model ([11], 123–5, Eq. 5, 6, and 10). **Z**-entailment is defined as follows:

(4) *Definition*: For all $\Gamma$ : $\Gamma \mathrel{\|\!\!\sim\!\!\sim}_{\mathrm{Z}} \mathrm{C} \Rightarrow \mathrm{D}$ ($\Gamma$ **Z**-entails $\mathrm{C} \Rightarrow \mathrm{D}$) *iff* either (a) $\Gamma$ is **P**-inconsistent, or (b) $(\mathrm{W}_\Gamma, \mathrm{z}_\Gamma)$ satisfies $\mathrm{C} \Rightarrow \mathrm{D}$.[2]

The z-model of a set of conditionals $\Gamma$ maximizes independence between propositions by minimizing the rank of each world, while adhering to the constraint that the z-model satisfies $\Gamma$. As immediate consequence, **Z**-entailment validates *subclass inheritance* by default, i.e., $\{\mathrm{A} \Rightarrow \mathrm{B}\} \mathrel{\|\!\!\sim\!\!\sim}_{\mathrm{Z}} \mathrm{A} \wedge \mathrm{C} \Rightarrow \mathrm{B}$. Such entailments hold

[2] A more detailed explication of z-entailment, which tracks the manner in which z entailments were computed within our simulations, is found in [15].

Please cite this article in press as: P.D. Thorn, G. Schurz, Qualitative probabilistic inference under varied entropy levels, J. Appl. Log. (2016), http://dx.doi.org/10.1016/j.jal.2016.05.004

by default, since (i) the respective z-model satisfies $A \Rightarrow B$, so that the rank of lowest ranked world that satisfies $A \wedge B$ is less than the rank of lowest ranked world that satisfies $A \wedge \neg B$, and (ii) minimizing the rank of worlds, ensures, by default, that all worlds satisfying $A \wedge B$ receive the same rank, so that the rank of lowest ranked world that satisfies $A \wedge B \wedge C$ is less than the rank of lowest ranked world that satisfies $A \wedge \neg B \wedge C$. For similar reasons, **Z**-entailment validates inference by *default contraposition* (i.e., $\{A \Rightarrow B\} \|\!\sim\!\sim_{Z} \neg B \Rightarrow \neg A$). To be more precise, these inferences are upheld under the condition that the conditional knowledge base does not contain further conditionals that are $\varepsilon$-inconsistent[3] with the conclusions of these inferences (cf. [1]). The relation $\|\!\sim\!\sim_{Z}$ is thus *non-monotonic*, since, for example, whether $\Gamma \cup \{A \Rightarrow B\} \|\!\sim\!\sim_{Z} \neg B \Rightarrow \neg A$, depends on whether $\Gamma \cup \{A \Rightarrow B\} \cup \{\neg B \Rightarrow \neg A\}$ is $\varepsilon$-inconsistent.

One disadvantage of **Z**-entailment is that (in the absence of further assumptions) it does not automatically provide information concerning probabilistic reliability, such as provided by the improbability-sum semantics for system **P**. However, in [15] it is shown how to obtain this desideratum (based on work in [12]):

**Theorem.** *If* $A_1 \Rightarrow B_1, \ldots, A_n \Rightarrow B_n \|\!\sim\!\sim_{Z} C \Rightarrow D$ *holds, then improbability-sum preservation* ($I(D|C) \leq \Sigma\{I(B_i|A_i) : 1 \leq i \leq n\}$) *holds for all probability functions* $P$ *that satisfy the default assumption* $P(A_i \supset B_i|C) \geq P(B_i|A_i)$*, for all* $1 \leq i \leq n$.

**Proof.** See [15], theorem 4 (5). □

We proceed here as if the default assumptions specified in the above theorem hold, and say that system **Z** licenses the inference to $C \Rightarrow_{1-\Sigma\{1-r_i:1\leq i\leq n\}} D$ from $A_1 \Rightarrow_{r_1} B_1, \ldots, A_n \Rightarrow_{r_n} B_n$, in cases where $A_1 \Rightarrow B_1, \ldots, A_n \Rightarrow B_n \|\!\sim\!\sim_{Z} C \Rightarrow D$. As with the evaluations conducted in [15], a central question concerns whether, and in which environments, inference in accordance with the preceding principle tends to yield accurate conclusions.

### *2.4. System* **QC**

**Z**-entailment is not the strongest (minimally reasonable) inference calculus for risky default inference among uncertain conditionals. An even stronger, and quite simple, calculus is *quasi-classical* reasoning. Here one reasons with uncertain conditionals as if they were material implications[4]:

(5) $\Gamma \vdash_{\mathbf{QC}} C \Rightarrow D$ *iff* $\{A \supset B : A \Rightarrow B \in \Gamma\} \vdash C \supset D$.

Improbability-sum preservation holds for inferences between material conditionals, or more generally, between formulas of propositional logic, as was shown by Suppes ([17], 54). In particular, $\{A_1, \ldots, A_n\} \vdash B$ *iff* it holds for all probability distributions that $I(B) \leq \Sigma\{I(A_i) : 1 \leq i \leq n\}$. Beyond the result of Suppes, it is possible to formulate probabilistic conditions under which **QC**-reasoning approximately satisfies improbability-sum preservation. In particular, it is shown in [15], sec. 2.5, (13) that a **QC** inference from a given set of premises is guaranteed to preserve probability in the manner of system **P** *iff* the improbability-sum of the premises is very small, and some decimal powers smaller than the probability of the conclusion's antecedent. Following [15], we proceed as if these conditions hold, and say that system **QC** licenses the inference to $C \Rightarrow_{1-\Sigma\{1-r_i:1\leq i\leq n\}} D$ from $A_1 \Rightarrow_{r_1} B_1, \ldots, A_n \Rightarrow_{r_n} B_n$, in cases where $A_1 \Rightarrow B_1, \ldots, A_n \Rightarrow B_n \vdash_{\mathbf{QC}} C \Rightarrow D$. Once again, the question remains of whether inference in accordance with the preceding principle tends to yield accurate conclusions.

[3] A set of conditionals is $\boldsymbol{\varepsilon}$-consistent just in case the corresponding conditional probabilities can be simultaneously made arbitrarily close to 1.

[4] System **QC** is obtained in one step if one adds the unrestricted monotonicity rule ($A \Rightarrow B / A \wedge C \Rightarrow B$) to system **P**.

Please cite this article in press as: P.D. Thorn, G. Schurz, Qualitative probabilistic inference under varied entropy levels, J. Appl. Log. (2016), http://dx.doi.org/10.1016/j.jal.2016.05.004

## 3. The simulations

Following [15], our simulations operate over a simple language with four two-valued variables: a, b, c, and d.[5] Similarly, we assume a probability distribution over the sixteen possible worlds, or more precisely states of the world, describable in this language. For all of our simulations, we generated a probability distribution over these worlds by setting the values of the following fifteen independently variable probabilities: $P(a)$, $P(b|a)$, $P(b|\neg a)$, $P(c|a \wedge b)$, $P(c|a \wedge \neg b)$, $P(c|\neg a \wedge b)$, $P(c|\neg a \wedge \neg b)$, $P(d|a \wedge b \wedge c)$, $P(d|a \wedge b \wedge \neg c)$, $P(d|a \wedge \neg b \wedge c)$, $P(d|a \wedge \neg b \wedge \neg c)$, $P(d|\neg a \wedge b \wedge c)$, $P(d|\neg a \wedge b \wedge \neg c)$, $P(d|\neg a \wedge \neg b \wedge c)$, and $P(d|\neg a \wedge \neg b \wedge \neg c)$ (which permits the calculation of $P(\neg a)$, $P(\neg b|a)$, $P(\neg b|\neg a)$, etc.).[6] We then generated the full probability distribution, P, by the chain rule, which is familiar from the study of Bayes-networks. In particular, where appending "±" to a propositional atom, v, generates a variable ranging over $\{v, \neg v\}$, the probability of possible worlds defined over our four atoms is given by the instances of $P(\pm a \wedge \pm b \wedge \pm c \wedge \pm d) = P(\pm a) \cdot P(\pm b | \pm a) \cdot P(\pm c| \pm a \wedge \pm b) \cdot P(\pm d| \pm a \wedge \pm b \wedge \pm c)$ [cf. [10], p. 123].

The entropy of a probability distribution, P, over a finite set of possible worlds, W, is defined as $\mathrm{Ent}(P) = -\sum_i P(w_i) \cdot \log_2(P(w_i))$ (for $w_i \in W$). So in the case where W contains sixteen worlds (as is the case in our simulations) Ent(P) will be in $[0, 4]$, where $\mathrm{Ent}(P) = 4$ means that P is a uniform probability distribution over W, and $\mathrm{Ent}(P) = 0$ means that P is a standard valuation function, assigning the value 1 to exactly one world, and the value 0 to all others. To illustrate the notion of entropy further, assume that our 'toy' worlds are formed by combinations of the four binary states ±a, ±b, ±c and ±d, and assume that the set of possible states is reduced, given that one or more *strict laws* (of the form $x \rightarrow y$, where $x, y \in \{a, b, c, d\}$) obtain (where $a \rightarrow b$ means that there are no worlds in which $a \wedge \neg b$). Finally, assume that the probability distribution over the possible states which are not excluded by these laws is uniform.[7] If n is the number of such possible states, then the entropy of the resulting probability distribution is $E(n) = n \cdot (1/n) \cdot \log_2(n) = \log_2(n)$. Using this formula, the following examples illustrate the manner in which the regularity of the state space, as characterized by strict laws, corresponds to its entropy level:

One law: $a \rightarrow b$: $n = 12$, $E \approx 3.6$.
Two laws $a \rightarrow b$, $c \rightarrow d$: $n = 9$, $E \approx 3.2$.
Three laws: $a \leftrightarrow b$, $c \rightarrow d$: $n = 6$, $E \approx 2.6$.
Four laws: $a \leftrightarrow b$, $c \leftrightarrow d$: $n = 4$, $E = 2$.
Eight laws (the maximal number): $a \leftrightarrow b \leftrightarrow c \leftrightarrow d$: $n = 2$, $E = 1$.
($E < 1$ is only possible if the probability of admitted worlds is non-uniform.)

The preceding illustrates that entropy levels between 3 and 3.5 correspond to moderate levels of orderliness, and entropy-levels between 3 and 2 correspond to high degrees of orderliness. Entropy levels below 2 correspond to unrealistically high degrees of orderliness, while entropy levels below 1 correspond to degrees of orderliness that are stronger than what can be enforced by 'strict laws'.[8]

Within [15], the probability distributions over the sixteen possible worlds were selected for each simulation, by setting the above fifteen conditional probabilities according to a uniform probability distribution on $[0, 1]$. Probability distributions selected in the method of [15] have a mean entropy level of about 2.88.

[5] The simulations in [15] were programmed in Visual Basic .NET 2010. We adapted that code in order to run the simulations described here.

[6] We used the RandomClass constructor within the .NET Framework in order to generate these values. The constructor generates pseudo random numbers according to an algorithm based on Donald E. Knuth's subtractive random number generator algorithm [7], with a time-dependent seed value which is determined by the system clock.

[7] This implies that entropy is maximized under the constraints imposed by the strict laws, which implies that, absent the present assumption, the entropy values described within our illustration express upper bounds on the entropy of the corresponding distributions.

[8] If the laws are not strict but merely high-probability laws, then the corresponding entropy levels are slightly smaller than the ones given.

By the above considerations, this is a moderate level of entropy, not too high and not too low, allowing for regularities, without excluding too many possible states. Since we averaged our results in [15] over many simulations, each starting with a randomly selected probability distribution, the results presented there reflect the performance of the four reasoning systems in moderate entropy level environments. Diverging from [15], we now control the entropy level of the probability distributions over the sixteen worlds. For each simulation, we chose a particular entropy level $\delta$. Our program then proceeded by generating probability distributions in the manner of [15] until a distribution was generated whose entropy resided in the interval $[\delta - 0.001, \delta + 0.001]$.

To manage the search space in assessing the four LP systems, we restricted our attention to conditionals whose antecedent and consequent are conjunctions of literals. We also assumed that no propositional atom appears twice in any *premise conditional* or *inferred conditional*. These restrictions effectively limited the language under consideration to 464 conditionals (cf. [15]). We call the language composed of this set of 464 conditionals "$\mathrm{L}_4$". Drawing from $\mathrm{L}_4$, we assumed that a small number of conditionals, so-called *premise conditionals*, together with their associated probabilities, were known to the reasoning systems. We further required that the probability associated with each premise conditional was at least 0.9. In each simulation, the three premise conditionals were selected at random from among the sentences of our language ($\mathrm{L}_4$) whose probability was at least 0.9. We then allowed each LP reasoning system to infer, from the given premise conditionals, all of the conditionals, $\mathrm{C} \Rightarrow_{\mathrm{r}} \mathrm{D}$, that follow according to the respective systems. For systems **P**, **Z**, and **QC**, the value r, for each inferred conditional, was set to be one minus the sum of the improbabilities of the premise conditionals needed in deriving the conclusion. For system **O**, r was set to be the probability value of the least probable premise conditional needed for the derivation of $\mathrm{C} \Rightarrow \mathrm{D}$ in **O**.

The restriction of our simulations to cases where the systems are provided with three possible premise conditionals expressed within $\mathrm{L}_4$ *partly* limits the scope of our results. However, we believe that our choice to limit our attention to conditionals whose antecedent and consequent are conjunctions of literals, without repetition, is not overly limiting. The prohibition of repeated literals is sensible, given the restriction to conditionals whose antecedent and consequent are conjunctions of literals. Indeed, repetition in this case can generate only redundancy, or conditionals whose probability is one, zero, or undefined. The restriction of our attention to conditionals whose antecedent and consequent are conjunctions of literals omits conditionals whose antecedent and/or consequent is a disjunction of conjunctions of literals. The restriction excludes instances of OR and XOR, along with some instances of right weakening, along with some instances of inheritance reasoning (as licensed by systems **Z** and **QC**). The order of excluded instances of OR and XOR in comparison to the set of inferences still included given the imposed restrictions is doubtless small. On the other hand, the number of excluded instances of right weakening and inheritance reasoning that are excluded is non-negligible. However, instances of right weakening and inheritance reasoning are amply represented within our simulations. In any case, extending the range of possible antecedents and consequents in order to include all propositional formulae in a four atom language is not practicable, as it would leave us with a language of $2^{17}$ conditionals. Moreover, it is unclear that evaluating the four systems according to such a language would provide a better assessment of the four systems, as such an evaluation would place too much weight in the value of informationally irrelevant instances of right weakening, e.g., instances that introduce irrelevant disjuncts, such as: $\mathrm{A} \Rightarrow \mathrm{C}$ / $\mathrm{A} \Rightarrow \mathrm{C} \vee \mathrm{D}$.

The limitation of our simulations to a propositional language formed from four atoms also suggests limitations in the scope of our results. Were we to increase the number of atoms in our propositional language, while holding fixed the number conditionals available as premises, the effect would be to decrease the prevalence of repeated atoms within the available premises, with the consequence that fewer multiple premise inferences are drawn. On the other hand, we could increase the number of multiple premise inferences under such conditions by increasing the number of available premises. The preceding facts suggest that it would be worthwhile to disaggregate the performance of the systems according to the number of premises used in drawing inference (e.g., one premise inferences, two premise inferences, etc.). This is something that

Please cite this article in press as: P.D. Thorn, G. Schurz, Qualitative probabilistic inference under varied entropy levels, J. Appl. Log. (2016), http://dx.doi.org/10.1016/j.jal.2016.05.004

has not yet been done, though we conjecture, based on results presented in [15] and [18], which controlled for the number of available premises, that the ordinal rankings for the four systems across inferences based on varied numbers of premises would not change.

After determining which conclusions were inferred by the four systems, each system was assigned numeric scores for each of the conclusions that it inferred, as a measure of accuracy and informativeness. Since there are no established and uncontroversial measures for scoring the accuracy and informativeness of such conclusions (i.e., lower probability bounds), we devised several scoring methods that have principled motivations and are pertinent to assessing accuracy and informativeness (despite the fact that other plausible measures are conceivable).

The first scoring measure that we applied is called the *advantage-compared-to-guessing* measure (cf. [15]). The idea behind this measure derives from the fact that the mean difference between two random choices of two real values r and s from the unit interval is, provably, 1/3. Based on this fact, we assessed each system by counting a judged lower probability bound that differs from the true probability by more than one-third *negatively*, and counting a judged lower probability bound that differs from the true probability by less than one-third *positively*. We scored the judged lower probability bounds by a simple *linear* measure of their distance from the true probabilities:

(6) The *advantage-compared-to-guessing* (ACG) score for derived conditionals:

$$\mathrm{Score_{ACG}}(\mathrm{C} \Rightarrow_{\mathrm{r}} \mathrm{D}, \mathrm{P}) := 1/3 - \left|\mathrm{r} - \mathrm{P}(\mathrm{D}|\mathrm{C})\right|.$$

The ACG score measures the (linear) accuracy of inferred lower probability bounds. But the ACG measure does not distinguish between *underestimations*, i.e., bounds being much lower than the true probability, which are merely uninformative but not incorrect, and *overestimations*, i.e., bounds that are greater than the true probability, which count as errors. Note that only the systems **Z** and **QC**, but not the systems **O** and **P,** can make such errors. Systems **O** and **P** are protected from error, at the cost of inferring fewer conclusions.[9]

In order to take a broader perspective on the evaluation of LP-reasoning, we considered a second measure which punishes errors. We call this second measure the *subtle-price-is-right* measure (cf. [15]).[10] This measure assigns a positive score to any inferred lower probability bound that does not exceed the true probability, but penalizes inferred bounds that exceed the true probability by a negative score, where in both cases, the score received decreases linearly with the distance of the lower bound from the true probability:

(7) The *subtle-price-is-right* score for derived conditionals:

$$\begin{aligned}\mathrm{Score_{sPIR}}(\mathrm{C} \Rightarrow_{\mathrm{r}} \mathrm{D}, \mathrm{P}) &:= \mathrm{r}, \quad \text{if } \mathrm{r} \le \mathrm{P}(\mathrm{D}|\mathrm{C}),\\ &:= \mathrm{P}(\mathrm{D}|\mathrm{C}) - \mathrm{r}, \quad \text{otherwise.}\end{aligned}$$

Both the ACG and the sPIR score measure the accuracy of the lower probability bound of an inferred conditional; sPIR punishes errors while ACG does not. However, neither measure takes account of the *applicability* of an inferred conditional $\mathrm{A} \Rightarrow_{\mathrm{r}} \mathrm{B}$, which is proportional to the probability of its antecedent P(A), inasmuch as predictions and/or decisions based on a conditional $\mathrm{A} \Rightarrow_{\mathrm{r}} \mathrm{B}$ are ordinarily of consequence only if the conditional's antecedent is true. The accuracy and the applicability of a conditional are two factors

[9] System **O**, unlike system **P**, is also protected against drawing uninformative conclusions (at the cost of drawing very few inferences).

[10] The name of the measure derives from the long running American game show where contestants must guess the price of items, and succeed by having the most accurate guess that does not exceed the price of the relevant item.

that both determine the *utility* of an inferred conditional. In order to take into account the applicability of inferred conditionals, we introduced a third scoring measure in [18], which we call the *expected utility* measure. After some mathematical transformations (see below), this measure can be defined as follows:

(8) The *expected utility* score for derived conditionals:

$$\mathrm{Score}_{\mathrm{EU}}(\mathrm{C} \Rightarrow_{\mathrm{r}} \mathrm{D}, \mathrm{P}) := \left(\mathrm{P}(\mathrm{D}|\mathrm{C})^2 - \left(\mathrm{P}(\mathrm{D}|\mathrm{C}) - \mathrm{r}\right)^2\right) \cdot \mathrm{P}(\mathrm{C})/2.$$

The EU measure scores an inferred conditional, $\mathrm{C} \Rightarrow_{\mathrm{r}} \mathrm{D}$, by evaluating the expected value of the decisions licensed by the acceptance of such a conditional (i.e., a conditional whose content is $\mathrm{P}(\mathrm{D}|\mathrm{C}) \geq \mathrm{r}$). We thereby assumed that a judged greatest lower conditional probability bound has the following behavioral import: If r is the greatest lower probability bound that a given agent accepts for D given C, then (if she is prudent and has sufficient wealth) she will purchase all wagers on D, conditional on C, at price \$s, so long as $\mathrm{s} < \mathrm{r}$, and refuse to accept such wagers for $\mathrm{s} \geq \mathrm{r}$. Given this behavioral interpretation of inferred conditionals, we considered an environment in which a respective agent is offered repeated opportunities to purchase wagers on D conditional on C with a stake s, where s is determined at random, according to a uniform probability distribution over the interval $[0, 1]$. In that environment, the expected value of accepting the greatest lower probability bound r on $\mathrm{P}(\mathrm{D}|\mathrm{C})$ is *provably*: $(\mathrm{P}(\mathrm{D}|\mathrm{C})^2 - (\mathrm{P}(\mathrm{D}|\mathrm{C}) - \mathrm{r})^2) \cdot \mathrm{P}(\mathrm{C})/2$ (cf. [18]). So calculated, the expected utility of inferring a conditional, $\mathrm{C} \Rightarrow_{\mathrm{r}} \mathrm{D}$, is proportional to the product of an applicability factor, $\mathrm{P}(\mathrm{C})/2$ (where $\mathrm{P}(\mathrm{C})$ is the probability of the conditional's antecedent), and an 'extended' non-linear accuracy factor $(\mathrm{P}(\mathrm{D}|\mathrm{C})^2 - (\mathrm{P}(\mathrm{D}|\mathrm{C}) - \mathrm{r})^2)$, which takes into account the squared accuracy of the inferred conditional, along with the squared magnitude of the true conditional probability of the inferred conditional.

## 4. The results

Since the four LP systems that we consider are ordered in terms of the number of inferences they license ($\mathbf{O} \subset \mathbf{P} \subset \mathbf{Z} \subset \mathbf{QC}$), our focus here is on the 'new' inferences licensed by each system as one proceeds from system **O** to system **QC**, i.e., the inferences licensed by system **O**, the inferences licensed by system **P** that are not licensed by system **O** (**P**–**O**), the inferences licensed by system **Z** that are not licensed by system **P** (**Z**–**P**), and the inferences licensed by system **QC** that are not licensed by system **Z** (**QC**–**Z**). Table 1 lists the mean number of conclusions inferred by each (sub)system, across varied entropy levels, and the mean number of erroneous inferences among the **Z**–**P** and **QC**–**Z** inferences (i.e., those instances where the inferred lower bound exceeded the actual probability), along with standard error rates for the mean values "±S.E.M."[11] For example, the very first entry of Table 1 tells us that at entropy level 3.5 system **O** draws on average only 3.016 conclusions out of the 464 possible conclusions expressible within $\mathrm{L}_4$. Further entries to the right tell us that **P**–**O** draws only 0.07 additional conclusions, on average, while **Z**–**P** draws a further 10.0 additional conclusions, where 8.1 are in error (over-estimates), on average. The mean values reported here are based on a sample of one thousand simulations at each listed entropy level.

The most obvious pattern exhibited in Table 1 is that the number of inferences drawn by each system is a decreasing function of the entropy level. This pattern was expected, since lower entropy levels imply a less evenly distributed probability function, and in turn a greater number of possible premise conditionals with multiple conjuncts in their consequents. Such conditionals support a greater number of inferences for all of the systems considered. The number of errors among the new inferences drawn by system **Z** (**Z**–**P**) and system **QC** (**QC**–**Z**) also increase with decreasing entropy level. Concerning the relatively high error rates for **Z**–**P** and **QC**–**Z**, it should be observed that a system that makes persistent errors may well be regarded

[11] We adopt the convention for reporting significant digits proposed in [2].

**Table 1**
Mean number of inferences and errors.

| Entropy level | Mean number of inferences | | | | Mean number of errors | |
|---|---|---|---|---|---|---|
| | **O** | **P–O** | **Z–P** | **QC–Z** | **Z–P** | **QC–Z** |
| 3.5 | 3.016 ± 0.006 | 0.07 ± 0.01 | 10.0 ± 0.1 | 3.3 ± 0.2 | 8.1 ± 0.1 | 3.2 ± 0.2 |
| 3 | 3.12 ± 0.02 | 0.2 ± 0.03 | 23.2 ± 0.2 | 20.9 ± 0.6 | 16.0 ± 0.2 | 20.2 ± 0.6 |
| 2.5 | 3.26 ± 0.03 | 0.45 ± 0.05 | 30.4 ± 0.3 | 34.1 ± 0.8 | 21.1 ± 0.2 | 32.8 ± 0.7 |
| 2 | 3.63 ± 0.05 | 0.77 ± 0.08 | 35.2 ± 0.3 | 43 ± 1 | 24.8 ± 0.2 | 41.7 ± 0.8 |
| 1.5 | 4.09 ± 0.07 | 1.5 ± 0.1 | 38.5 ± 0.3 | 50 ± 1 | 27.0 ± 0.2 | 47.7 ± 0.9 |
| 1 | 4.6 ± 0.1 | 1.8 ± 0.2 | 40.2 ± 0.4 | 53 ± 1 | 29.8 ± 0.2 | 50.8 ± 0.9 |
| 0.5 | 4.8 ± 0.1 | 2.4 ± 0.2 | 41.5 ± 0.4 | 55 ± 1 | 31.4 ± 0.2 | 53.1 ± 0.9 |

**Table 2**
Mean ACG scores.

| Entropy level | Mean ACG scores | | | |
|---|---|---|---|---|
| | **O** | **P–O** | **Z–P** | **QC–Z** |
| 3.5 | 1.005 ± 0.002 | 0.019 ± 0.004 | 2.37 ± 0.04 | −0.43 ± 0.02 |
| 3 | 1.036 ± 0.006 | 0.055 ± 0.007 | 4.44 ± 0.07 | −3.62 ± 0.09 |
| 2.5 | 1.079 ± 0.009 | 0.12 ± 0.01 | 3.8 ± 0.1 | −6.5 ± 0.2 |
| 2 | 1.19 ± 0.02 | 0.21 ± 0.02 | 2.6 ± 0.1 | −8.6 ± 0.2 |
| 1.5 | 1.33 ± 0.02 | 0.41 ± 0.03 | 1 ± 0.1 | −10.2 ± 0.2 |
| 1 | 1.50 ± 0.03 | 0.51 ± 0.04 | −0.7 ± 0.1 | −11.1 ± 0.2 |
| 0.5 | 1.57 ± 0.03 | 0.73 ± 0.05 | −2.5 ± 0.2 | −12.3 ± 0.2 |

**Table 3**
Mean sPIR scores.

| Entropy level | Mean sPIR Scores | | | |
|---|---|---|---|---|
| | **O** | **P–O** | **Z–P** | **QC–Z** |
| 3.5 | 2.833 ± 0.005 | 0.06 ± 0.01 | 0.84 ± 0.07 | −1.48 ± 0.06 |
| 3 | 2.94 ± 0.02 | 0.17 ± 0.02 | 3.7 ± 0.1 | −9.9 ± 0.2 |
| 2.5 | 3.1 ± 0.03 | 0.40 ± 0.04 | 2.6 ± 0.2 | −16.6 ± 0.4 |
| 2 | 3.45 ± 0.05 | 0.69 ± 0.06 | 1.0 ± 0.2 | −21.3 ± 0.4 |
| 1.5 | 3.90 ± 0.07 | 1.3 ± 0.1 | −1.6 ± 0.2 | −24.5 ± 0.4 |
| 1 | 4.42 ± 0.09 | 1.7 ± 0.1 | −4.0 ± 0.2 | −26.5 ± 0.5 |
| 0.5 | 4.7 ± 0.1 | 2.3 ± 0.2 | −6.4 ± 0.2 | −28.6 ± 0.5 |

**Table 4**
EU scores.

| Entropy level | Mean EU scores | | | |
|---|---|---|---|---|
| | **O** | **P–O** | **Z–P** | **QC–Z** |
| 3.5 | 0.227 ± 0.003 | 0.005 ± 0.001 | 0.379 ± 0.008 | −0.0001 ± 0.0002 |
| 3 | 0.469 ± 0.005 | 0.033 ± 0.004 | 1.29 ± 0.02 | −0.025 ± 0.001 |
| 2.5 | 0.651 ± 0.009 | 0.1 ± 0.01 | 1.85 ± 0.03 | −0.044 ± 0.002 |
| 2 | 0.893 ± 0.02 | 0.22 ± 0.02 | 2.4 ± 0.04 | −0.060 ± 0.002 |
| 1.5 | 1.23 ± 0.03 | 0.51 ± 0.04 | 2.87 ± 0.05 | −0.070 ± 0.003 |
| 1 | 1.64 ± 0.05 | 0.73 ± 0.06 | 3.22 ± 0.06 | −0.079 ± 0.004 |
| 0.5 | 1.93 ± 0.05 | 1.11 ± 0.07 | 3.68 ± 0.08 | −0.056 ± 0.002 |

as highly accurate and informative. Take, for example, a miscalibrated oracle that invariably overestimates the probability of contingent propositions by a value of 0.0001, and is thereby always in error.

We now consider the average scores earned by the respective systems for the full set of conclusions drawn within a single simulation. Tables 2, 3, and 4 list the results.

Examining Tables 2, 3, and 4, we observe three major results:

**Result 1.** We see that the **QC–Z** inferences earn negative scores at every entropy level, according to all three scoring rules. This provides a relatively good reason for concluding that we should not reason in accordance with system **QC**, independently of the entropy level of the environment, as long as we are reasoning with

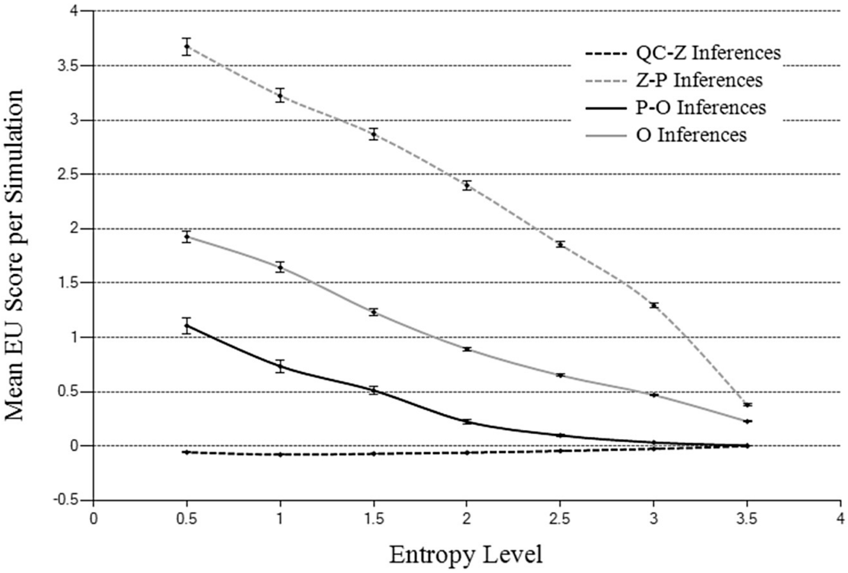

**Fig. 1.** Mean EU scores, with standard error bars.

conditionals of moderate uncertainty (high but not extremely high probability), and our concern is to draw conclusions that are accurate and useful. This result is not surprising if we consider that the *new* inferences drawn by system **QC** are default inferences drawn in the face of known exceptions: for example, cases where one infers $A \wedge C \Rightarrow B$ from $A \Rightarrow B$ by **QC**, and the inference is not permitted by system **Z**, are cases where a further premise asserts $C \Rightarrow \neg B$.

**Result 2.** We also see that **O** and **P**–**O** inferences earn positive scores at every entropy level, according to all three scoring rules. This strongly suggests, again independently of the entropy level of the environment, that it is always reasonable to reason in accord with the rules of systems **O** and **P**. This conclusion is unsurprising given conditions (1) and (2) above, which characterize the ability of systems **O** and **P** to preserve premise probability.

**Result 3.** It is only when we turn to evaluate the quality of **Z**–**P** inferences that the data from Tables 1, 2, and 3 is equivocal. When considering the ACG and sPIR scores for the **Z**–**P** inferences, we observe a peak in performance, when the entropy level of the underlying probability distribution is relatively high. Thereafter, decreasing entropy correlates with decreasing ACG and sPIR scores, which are nevertheless positive, so long as the entropy is not extremely low – for the ACG measure as low as 1, and for the sPIR measure as low as 1.5.[12] On the other hand, decreasing entropy generally correlates with increasing EU scores for the **Z**–**P** inferences. Fig. 1 provides a graphical representation of that pattern.

Before we try to give an explanation for this remarkable result concerning **Z**–**P** inferences, it is helpful to look at the average scores earned for single inferences across varied entropy levels. Tables 5, 6, and 7 list the results, and Fig. 2 provides a graphic presentation of the information presented in Table 7.

Our main remaining concern is to evaluate the quality of **Z**–**P** inferences. Tables 5 and 6 show that **Z**–**P** inferences lead to bounds that are relatively close to the true probabilities, when entropy is high. However, when the entropy level is very low, the distance between the judged bounds and the true probability tends to be rather large. For example, when the entropy of the underlying distribution is 0.5, the mean difference between a lower bound inferred by **Z**–**P** and the true probability is about 0.4. Similarly, while **Z**–**P** inferences

[12] The difference derives from the fact that sPIR, but not ACG, punishes erroneous **Z**–**P** inferences.

**Table 5**
Mean ACG scores per inference.

| Entropy level | Mean ACG score per inference | | | |
|---|---|---|---|---|
| | **O** | **P–O** | **Z–P** | **QC–Z** |
| 3.5 | 0.33318 ± 0.00005 | 0.275 ± 0.007 | 0.236 ± 0.001 | −0.130 ± 0.005 |
| 3 | 0.3319 ± 0.0002 | 0.275 ± 0.003 | 0.191 ± 0.001 | −0.173 ± 0.002 |
| 2.5 | 0.3308 ± 0.0002 | 0.277 ± 0.002 | 0.125 ± 0.001 | −0.191 ± 0.002 |
| 2 | 0.3276 ± 0.0003 | 0.276 ± 0.002 | 0.074 ± 0.001 | −0.197 ± 0.001 |
| 1.5 | 0.3256 ± 0.0003 | 0.280 ± 0.001 | 0.026 ± 0.002 | −0.203 ± 0.001 |
| 1 | 0.3239 ± 0.0003 | 0.280 ± 0.001 | −0.018 ± 0.002 | −0.209 ± 0.001 |
| 0.5 | 0.3267 ± 0.0002 | 0.3032 ± 0.0006 | −0.060 ± 0.002 | −0.222 ± 0.001 |

**Table 6**
Mean sPIR scores per inference.

| Entropy level | Mean sPIR score per inference | | | |
|---|---|---|---|---|
| | **O** | **P–O** | **Z–P** | **QC–Z** |
| 3.5 | 0.9394 ± 0.0005 | 0.850 ± 0.005 | 0.0837 ± 0.005 | −0.446 ± 0.006 |
| 3 | 0.9427 ± 0.0005 | 0.871 ± 0.005 | 0.159 ± 0.004 | −0.471 ± 0.003 |
| 2.5 | 0.9500 ± 0.0005 | 0.879 ± 0.003 | 0.087 ± 0.003 | −0.487 ± 0.002 |
| 2 | 0.9513 ± 0.0005 | 0.890 ± 0.002 | 0.027 ± 0.003 | −0.490 ± 0.002 |
| 1.5 | 0.9550 ± 0.0004 | 0.902 ± 0.001 | −0.040 ± 0.003 | −0.490 ± 0.002 |
| 1 | 0.9574 ± 0.0004 | 0.910 ± 0.001 | −0.099 ± 0.003 | −0.496 ± 0.002 |
| 0.5 | 0.9719 ± 0.0003 | 0.9524 ± 0.0007 | −0.155 ± 0.003 | −0.516 ± 0.002 |

**Table 7**
Mean EU scores per inference.

| Entropy level | Mean EU score per inference | | | |
|---|---|---|---|---|
| | **O** | **P–O** | **Z–P** | **QC–Z** |
| 3.5 | 0.0753 ± 0.0008 | 0.075 ± 0.002 | 0.0378 ± 0.0003 | 0.0000 ± 0.0001 |
| 3 | 0.150 ± 0.002 | 0.166 ± 0.003 | 0.0558 ± 0.0004 | −0.00119 ± 0.00006 |
| 2.5 | 0.200 ± 0.002 | 0.219 ± 0.004 | 0.0611 ± 0.0005 | −0.00130 ± 0.00004 |
| 2 | 0.246 ± 0.002 | 0.290 ± 0.003 | 0.0682 ± 0.0006 | −0.00138 ± 0.00004 |
| 1.5 | 0.301 ± 0.002 | 0.346 ± 0.002 | 0.0745 ± 0.0007 | −0.00140 ± 0.00004 |
| 1 | 0.356 ± 0.002 | 0.401 ± 0.002 | 0.0801 ± 0.0008 | −0.00148 ± 0.00004 |
| 0.5 | 0.401 ± 0.002 | 0.459 ± 0.001 | 0.0887 ± 0.0009 | −0.00101 ± 0.00003 |

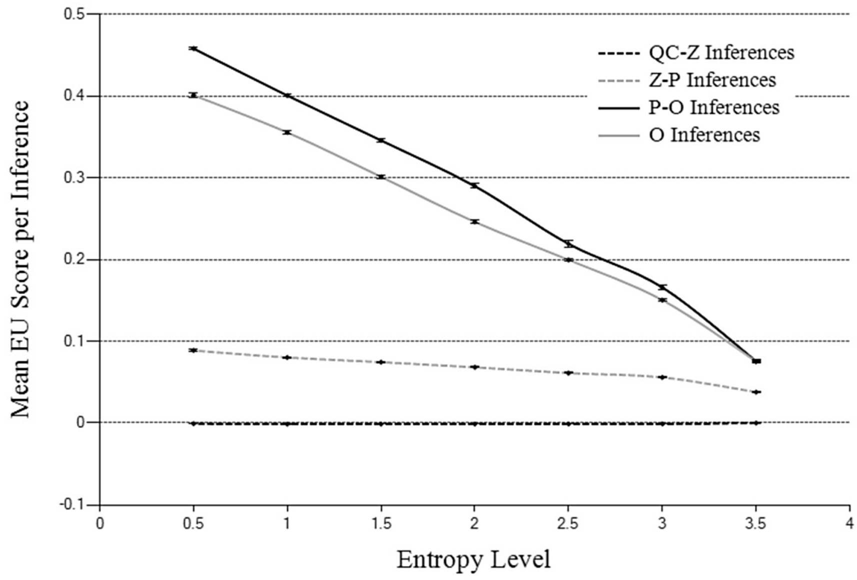

**Fig. 2.** Mean EU scores per inference (standard errors negligible).

Please cite this article in press as: P.D. Thorn, G. Schurz, Qualitative probabilistic inference under varied entropy levels, J. Appl. Log. (2016), http://dx.doi.org/10.1016/j.jal.2016.05.004

are expected to yield relatively high sPIR scores, when an inferred bound is not in error (ranging from about 23 to 38 percent of cases, depending on the entropy level), we see that when the entropy level is low, a typical erroneous inferred bound exceeds the true probability by a significant margin.

The results in the case of the ACG and sPIR measures can be explained by observing that accurate default reasoning in system **Z** requires two things: first that the conditional probabilities associated with the premises be sufficiently high (which correlates with low entropy), and second, that the entropy for the marginal probability distributions over worlds of identical rank is high (tending toward uniformity). Taken together, these two 'opposed effects' explain why the accuracy of the **Z**–**P** inferences, as measured by ACG and sPIR, first increases with entropy levels falling from high to moderately high, and then decreases with entropy levels falling from moderate to very low (where the average **Z**–**P** scores eventually become negative). For example, if one infers $A \wedge C \Rightarrow B$ from $A \Rightarrow B$ in system **Z**, then one assumes by default that $A \Rightarrow \neg C$ does not hold (since by rational monotony the inference from $A \Rightarrow B$ to $A \wedge C \Rightarrow B$ or $A \Rightarrow \neg C$ is **P**-valid). This means that the probability of $A \wedge B \wedge \neg C$ should not be much greater than that of $A \wedge B \wedge C$. But at entropy levels of 1 or so, these two probabilities will always be highly dissimilar, so that every second default inference of this sort will go wrong.

In contrast, the EU scores for **Z**–**P** inferences are generally increasing with decreasing entropy of the underlying probability distribution. This can be explained by the fact that the EU score for an inferred conditional depends on its applicability, i.e., the probability of the antecedent of the inferred conditional. Assume, once again, that one infers $A \wedge C \Rightarrow B$ from $A \Rightarrow B$ in system **Z**. If this inference 'goes wrong', then the conditional probability $P(\neg C|A)$ is high, whence the probability $P(A \wedge C)$ must be low, much lower than that of $P(A)$. On the other hand, if the inference to $A \wedge C \Rightarrow B$ from $A \Rightarrow B$ 'goes right', then the conditional probability $P(C|A)$ is not low, but at least moderate, whence $P(A \wedge C)$ is also not much lower than $P(A)$. Now since the probability $P(A \wedge C)$ is just the applicability of the inferred conditional $A \wedge C \Rightarrow B$, it follows that inaccurate inferences of the present sort (which are typical **Z**–**P** inferences) are less frequently applicable than in cases where inferences of this form yield accurate conclusions. This explains why the expected utility of **Z**–**P** inferences increases with decreasing entropy, even at very low entropy levels.

The fact that the EU scores for **Z**–**P** inferences are generally positive marks a positive sign in favor of their quality. Indeed, the present fact reflects a significant capacity of **Z**–**P** inferences to generate *informative and applicable* conclusions about the environment. Indeed, if we consider plausible *aprioristic* methods of assigning lower probability bounds, such as the ones considered in [18], i.e., methods of assigning lower probability bounds to the elements of our language ($L_4$) without exploiting the information that was supplied to the four LP systems (in the form of premise conditionals), then we see that the EU scores earned for **Z**–**P** inferences tend to be much higher than the scores earned by aprioristic methods. For example, the most successful aprioristic method considered in [18] assigned the lower bounds 1/2, 1/4, and 1/8, respectively, to conditionals with one, two, or three conjuncts in their consequent (as the values 1/2, 1/4, and 1/8 are the average probabilities for conditionals with the corresponding number of conjuncts in their consequents). In the case where entropy is not controlled, this aprioristic method earned an EU score of about 0.02044 per inference, which is far lower than the average scores earned by **Z**–**P** inferences, across all entropy levels.[13]

## 5. Conclusions

We found that, independently of the entropy level of the environment, it is reasonable to accept the conclusions of **O** and **O**–**P** inferences, so long as our goal is to accept accurate, informative, and useful probability statements. We also found that, independently of the entropy level of the environment, one

[13] We also found no significant difference in the mean EU scores received by inferences made by this method when we varied the entropy level of the underlying probability function (using samples of 1000 simulations for each of the entropy levels studied here).

should not accept the conclusions of **QC–Z** inferences as long as one is reasoning from uncertain conditionals whose associated conditional probabilities are merely high, and not extremely high.

The difficult choice is whether, and in which situations, one should accept the conclusions of **Z–P** inferences. The correct choice depends, not only on the environment, but also on one's goals in forming conclusions about one's environment. If one's interest is in inferring accurate conditionals (or accurate and correct ones), then it is reasonable to shoulder the risk inherent in accepting the conclusions of **Z–P** inferences as long as one can reasonably assume that the entropy of the environment is not very low. However, if one is interested in drawing conclusions that may serve as a basis for action, and one is interested in making the best inference one can, given the information one has, then the tendency of **Z–P** inferences to deliver significant positive EU scores (even when the entropy of the underlying distribution is very low) indicates the value of these inferences in such circumstances.

In considering the risks involved in accepting the conclusions of **Z–P** inferences, it is worth considering whether there are alternatives that would support better probability judgments. Since we know that our present method of associating lower probability bounds with **Z–P** inferences is prone to overestimation, we conjecture that a more optimal method would make a downward correction to these assigned bounds. It would also make sense to vary the size of this correction, in cases where the entropy of the underlying distribution is known. The exploration of this idea is left to future work.

One general point that is illustrated by the results presented here is that the choice of an appropriate system of defeasible reasoning may be context dependent. On the other hand, some systems of defeasible reasoning are very robust, delivering reasonable results over a wide range of contexts. Our main conclusion, in this article, is that system **Z** falls in this category. In particular, we think that the results presented here go some way in establishing the value of reasoning by system **Z** in a wide range of contexts, and also in a wide range of situations where one is uncertain about the stochastic features of one's environment and/or about the manner in which those features were determined. Within the simulations presented here, the environment was generated by determinate procedures. Given knowledge of these procedures, there is little doubt that we could have devised reasoning strategies, using the full inferential machinery of probability theory, to draw more accurate conclusions than the ones licensed by system **Z**. However, given considerably less information about the stochastic environment, or the means of its generation, such an approach becomes, in some cases, intractable, and, in others, impossible. It is in such circumstances where systems of reasoning, such as system **Z**, may be of great utility.

## Acknowledgements

Work on this paper was supported by the DFG-Grant SCHU1566/5-1 as part of the EuroCores LogiCCC project *The Logic of Causal and Probabilistic Reasoning in Uncertain Environments* (LcpR), and the DFG-Grant SCHU1566/9-1 as part of the SPP 1516 priority program "New Frameworks of Rationality". For valuable comments, we are indebted to David Makinson, Hannes Leitgeb, James Hawthorne, Angelo Gilio, Donald Bamber, Gerhard Brewka, Wolfgang Spohn, Gernot Kleiter, Christian Eichhorn, Gabriele Kern-Isberner, Matthias Unterhuber, and in memoriam to Ernest Adams.

## References

[1] E.W. Adams, The Logic of Conditionals, Reidel, Dordrecht, 1975.
[2] D. Bindel, J. Goodman, Principles of scientific computing, manuscript, 2009.
[3] M. Goldszmidt, J. Pearl, Qualitative probabilities for default reasoning, belief revision and causal modeling, Artif. Intell. 84 (1996) 57–112.
[4] J. Hawthorne, On the logic of non-monotonic conditionals and conditional probabilities, J. Philos. Log. 25 (1996) 185–218.
[5] J. Hawthorne, D. Makinson, The quantitative/qualitative watershed for rules of uncertain inference, Stud. Log. 86 (2007) 247–297.
[6] E. Jaynes, Prior probabilities, IEEE Trans. Syst. Sci. Cybern. 4 (3) (1968) 227–241.

[7] D. Knuth, The Art of Computer Programming, Vol. 2: Seminumerical Algorithms, Addison–Wesley, Reading, MA, 1981.

[8] D. Lehmann, M. Magidor, What does a conditional knowledge base entail? Artif. Intell. 55 (1992) 1–60.

[9] J.B. Paris, R. Simmonds, O is not enough, Rev. Symb. Log. 2 (2) (2009) 298–309.

[10] J. Pearl, Probabilistic Reasoning in Intelligent Systems, Morgan Kaufmann, Santa Mateo, California, 1988.

[11] J. Pearl, System Z, in: Proceedings of Theoretical Aspects of Reasoning about Knowledge, Santa Mateo, California, 1990, pp. 121–135.

[12] G. Schurz, Probabilistic default reasoning based on relevance and irrelevance assumptions, in: D. Gabbay, et al. (Eds.), Qualitative and Quantitative Practical Reasoning, in: Lect. Notes Artif. Intell., vol. 1244, Springer, Berlin, 1997, pp. 536–553.

[13] G. Schurz, Probabilistic Semantics for Delgrande's conditional logic and a counterexample to his default logic, Artif. Intell. 102 (1) (1998) 81–95.

[14] G. Schurz, Non-monotonic reasoning from an evolution-theoretic perspective: ontic, logical and cognitive foundations, Synthese 146 (1–2) (2005) 37–51.

[15] G. Schurz, P. Thorn, Reward versus risk in uncertain inference: theorems and simulations, Rev. Symb. Log. 4 (2) (2012) 574–612.

[16] K. Segerberg, Notes on conditional logic, Stud. Log. 48 (1989) 157–168.

[17] P. Suppes, Probabilistic inference and the concept of total evidence, in: J. Hintikka, P. Suppes (Eds.), Aspects of Inductive Logic, North-Holland Publ. Comp., Amsterdam, 1966, pp. 49–65.

[18] P. Thorn, G. Schurz, A utility based evaluation of logico-probabilistic systems, Stud. Log. 102 (4) (2014) 867–890.

[19] B. Van Fraassen, Laws and Symmetry, Oxford University Press, 1989.

Please cite this article in press as: P.D. Thorn, G. Schurz, Qualitative probabilistic inference under varied entropy levels, J. Appl. Log. (2016), http://dx.doi.org/10.1016/j.jal.2016.05.004