Chapter 12

From Statistics to Beliefs

In Section 11.3 we saw that, for first-order reasoning about probability, it is possible to put a probability both on the domain and on the set of possible worlds. Putting a probability on the domain is appropriate for "statistical" reasoning, while putting a probability on the set of possible worlds can be viewed as capturing an agent's subjective beliefs. Clearly the two should, in general, be related. That is, if an agent's knowledge base includes statistical information, his subjective probabilities should reflect this information appropriately. Relating the two is quite important in practice. We already saw in example of this in the introduction. Recall that, in this example, a doctor with a patient Eric can see that Eric has jaundice, no temperature, and red hair. His medical textbook includes the statistical information that 90% of people with jaundice have hepatitis and 80% of people with hepatitis have a temperature. What should the doctor's degree of belief be that Eric has hepatitis? This degree of belief is important because it forms the basis of the doctor's future decision regarding the course of treatment.

Unfortunately, there is no definitive "right" way for relating statistical information to degrees of belief. In this chapter, I consider one approach for doing this that has some remarkable properties (unfortunately, not all of them good). It is closely related to maximum entropy (at least, in the case of first-order language with only unary predicates) and gives insight into default reasoning as well. For definiteness, I focus on probabilistic reasoning in this chapter. Many of the ideas presented here should be extendible to other representations of uncertainty, but to date there has been no work on this topic. I also assume that we are dealing with only one agent.

12.1 Reference Classes

Before going into the technical details of the approach, it is worth clarifying some properties that we would like it to have. This is perhaps best done by considering the traditional approach to the problem of going from statistics to beliefs, which uses what are called *reference classes*. To simplify matters, assume for the purposes of this discussion that the agent's knowledge base consists of two types of statements: statistical assertions of the form "90% of people with jaundice have hepatitis" and "80% of people with hepatitis have a temperature" and information about one individual (such as Eric). The problem is to determine appropriate degrees of belief regarding events concerning that individual, given the statistical information and the information about the individual.

The idea of the reference-class approach is to equate the degree of belief in propositions about an individual with the statistics from a suitably chosen *reference class* (i.e., a set of domain individuals that includes the individual in question) about which statistics are known. For example, if the doctor is interested in ascribing a degree of belief to the proposition "Eric has hepatitis", he would first try to find the most suitable reference class for which he has statistics. Since all the doctor knows about Eric is that Eric has jaundice, then the set of people with jaundice seems like a reasonable reference class to use. Intuitively, the reference class is thought of as a set of individuals of which Eric is a "typical member". To the extent that this is true, then Eric ought to be just as likely to satisfy a property as any other member of the reference class. Since someone chosen at random from the set of people with jaundice has probability .9 of having hepatitis, the doctor assigns a degree of belief of .9 to Eric's having hepatitis.

While this seems like a reasonable approach (and not far from what people seem to do in similar cases), it is often difficult to apply in practice. For example, what if the doctor also knows that Eric is a baby and only 10% of babies with jaundice have hepatitis. What reference class should he use in that case? More generally, what should be done if there are competing reference classes? And what counts as a legitimate reference class?

To understand these issues, consider the following examples. To start with, consider the situation where Eric is a baby and only 10% of babies with jaundice have hepatitis. In this case, the standard response is that the doctor should prefer the more *specific* reference class—technically, this means the doctor should use the smallest reference class for which he has statistics. Since the set of babies is a subset of the set of people, this heuristic suggests the doctor ascribe degree of belief .1 to Eric's having hepatitis, rather than .9.

But the preference for the more specific reference class must be taken with a grain of salt, as the following example shows.

Example 12.1.1 Consider again the first knowledge base, where the doctor does not know that Eric is a baby. In that case, it seems reasonable for the doctor to take the appropriate reference class to consist of all people with jaundice and ascribe degree of belief .8 to Eric's having hepatitis. But Eric is also a member of the reference class consisting of jaundiced patients without hepatitis together with Eric. If there are quite a few jaundiced patients without hepatitis (for example, babies), then there are excellent statistics for the proportion of patients in this class with hepatitis: it is approximately 0%. Eric is the only individual in the class that may have hepatitis! Moreover, this reference class is clearly more specific (i.e., a subset of) the reference class of all people with jaundice. Thus, a naive preference for the more specific reference class results in the doctor ascribing degree of belief 0 (or less than ϵ for some very small ϵ) to Eric's having hepatitis! Clearly there is something fishy about considering the reference class consisting of jaundiced patients that do not have hepatitis together with Eric, but exactly what makes this reference class so fishy? ∎

There are other problems with the reference class approach. Suppose that the doctor also knows that Eric has red hair, but has no statistics for the fraction of jaundiced people with red hair that have hepatitis. Intuitively, the right thing to do in this case is ignore the fact that Eric has red hair, and continue to ascribe degree of belief .8 to Eric's having hepatitis. Essentially, this means treating having red hair as irrelevant. But what justifies this? Clearly not all information about Eric is irrelevant; for example, discovering that Eric is a baby is quite relevant.

This discussion of irrelevance should seem reminiscent of the discussion of irrelevance in the context of default reasoning (Section 7.3). This is not an accident. It turns out that much the same issues when trying to ascribe degrees of belief based on statistical information as in default reasoning. This issue is discussed in more detail in Section 12.4.

Going back to Eric, while it seems reasonable to prefer the more specific reference class (assuming that the problems of deciding what counts as a reasonable reference class can be solved), what should the doctor do if he has two competing reference classes? For example, suppose that the doctor knows that 10% of babies with jaundice have hepatitis but 90% of Caucasians with jaundice have hepatitis, and that Eric is a Caucasian baby with jaundice. Now the doctor has two competing reference classes: Caucasians and babies. Neither is more specific than the other. In this case, it seems reasonable to somehow weight the 10% and 90%, but how? The reference-class approach is silent on that issue. More precisely, its goal is to discover a single most appropriate reference class and use the statistics for that reference class to determine the degree of belief. If there is no single most appropriate reference class, it does not attempt to ascribe degrees of belief at all.

The random-worlds approach that I am about to present makes no attempt to identify a single relevant reference class. Nevertheless, it agrees with the reference-class approach when there is an obviously "most-appropriate" reference class. Moreover, it continues to make sense even when no reference class stands out as being the obviously most appropriate one to choose.

12.2 The Random-Worlds Approach

The basic idea behind the random-worlds approach is easy to explain and understand. Fix a finite vocabulary $\mathcal{T}$ and a domain D_N of size N; for simplicity, take $D_N = \{1, \ldots, N\}$. Since $\mathcal{T}$ is finite, there is only finitely many possible relational $\mathcal{T}$-structures with domain D_N. (Since "relational $\mathcal{T}$-structures with domain D_N" is a bit of a mouthful, in the remainder of this chapter, I call them simply a D_N-$\mathcal{T}$-structures.)

- If $\mathcal{T}$ consists of the unary predicate P, there are 2^N D_N-$\mathcal{T}$-structures: for each subset U of D_N, there is a D_N-$\mathcal{T}$ structure $\mathcal{A}$ such that $P^{\mathcal{A}_U} = U$.

- If $\mathcal{T}$ consists of the unary predicate P and the constant symbol c, then there are $2^N N$ D_N-$\mathcal{T}$ structures; these can be characterized by pairs (U, i), where $U \subseteq D_N$ is the interpretation of P and $i \in D_n$ is the interpretation of c.

- If $\mathcal{T}$ consists of the binary predicate B, then there are 2^{2N} D_N-$\mathcal{T}$-structures, one for each subset of $D_N \times D_N$.

Given a D_N-$\mathcal{T}$ structure $\mathcal{A}$, let μ_N^{unif} be the uniform probability measure on D_N, which gives each element of D_N probability $1/N$. Then $(\mathcal{A}, \mu_N^{unif})$ is a statistical $\mathcal{T}$-structure and can be used to determine the truth of all sentences in $\mathcal{L}^{QU,stat}(\mathcal{T})$. The pairs $(\mathcal{A}, \mu_N^{unif})$ can be viewed as possible worlds in a probability structure. The degree of belief assigned to a formula $\varphi \in \mathcal{L}^{QU,stat}(\mathcal{T})$ given a knowledge base KB consisting of formulas in $\mathcal{L}^{QU,stat}(\mathcal{T})$ is then defined as the fraction of worlds satisfying KB that also satisfy φ. This is just the conditional probability of φ given KB, assuming a uniform measure on these worlds.

The intuition behind this approach is not hard to explain. If all worlds are originally equally likely (which seems reasonable, in the absence of any other information), then the degree of belief that the agent ascribes to φ

upon learning KB should be the conditional probability that φ is true, given that KB is true. Put another way, the degree of belief that the agent ascribes to φ is just the probability of choosing a world (relational structure) at random that satisfies φ out of all the interpretations that satisfy KB. That is why this is called the *random-worlds* approach.

There are two details I need to fill in to make this completely formal. I started by assuming a fixed domain size of N. But where did N come from? Why is a particular choice of N the right choice? In fact, there is no obvious choice of N. To sidestep this issue, N will be treated as "large" by taking the limiting conditional probability as N grows to infinity.

The other issue that needs to be dealt with involves some problematic aspects related to the use of the language $\mathcal{L}^{QU,stat}(\mathcal{T})$. To understand the issue, consider a formula such as $\|Hep(x)|Jaun(x)\|_x = .9$, which says that 90% of people with jaundice have hepatitis. Notice, however, that is impossible for exactly 90% of people with jaundice to have hepatitis unless the number of people with jaundice is a multiple of 10. The statistical assertion was almost certainly not intended to have as a consequence such a statement about the number of people with jaundice. Rather, what was intended was almost certainly something like "*approximately* 90% of people with jaundice have hepatitis". Intuitively, this says that the proportion of jaundiced patients with hepatitis is close to 90%: i.e., within some tolerance τ of .9. To capture this, I consider a language that uses approximate equality and inequality, rather than equality and inequality. The language has an infinite family of connectives $\approx_i$ and $\preceq_i$, for $i = 1, 2, 3\ldots$ ("i-approximately equal" or "i-approximately less than or equal"). The statement "80% of jaundiced patients have hepatitis" then becomes, say, $\|Hep(x)|Jaun(x)\|_x \approx_1 .8$. The intuition behind the semantics of approximate equality is that each comparison should be interpreted using some small tolerance factor to account for measurement error, sample variations, and so on. The appropriate tolerance may differ for various pieces of information, so the logic allows different subscripts on the "approximately equals" connectives. A formula such as $\|Flies(x)|Bird(x)\|_x \approx_1 1 \wedge \|Flies(x)|Bat(x)\|_x \approx_2 1$ says that both $\|Flies(x)|Bird(x)\|_x$ and $\|Flies(x)|Bat(x)\|_x$ are approximately 1, but the notion of "approximately" may be different in each case. (Note that the actual choice of subscripts is irrelevant here, as long as different notions of "approximately" are denoted by different subscripts.)

The formal definition of the language $\mathcal{L}^{\approx}(\mathcal{T})$ is identical to that of $\mathcal{L}^{QU,stat}(\mathcal{T})$, except instead of statistical likelihood formulas, inequality formulas of the form $\|\varphi|\psi\|_X \sim \alpha$ are used, where $\sim$ is either $\approx_i$, $\preceq_i$, or $\succeq_i$, ,, for $i = 1, 2, 3, \ldots$. (The reason for using *conditional* statistical likelihood terms, rather than just unconditional ones as in $\mathcal{L}^{QU,stat}$, will shortly become clear. The results in this section and the next still hold even with polynomial statistical likelihood terms, but allowing only these simple inequality formulas simplifies the exposition.) Of course, a formula such as $||\varphi||_X \approx_i \alpha$ is an abbreviation for $\|\varphi|true\|_X \approx_i \alpha$. Call the resulting language $\mathcal{L}^{\approx}(\mathcal{T})$. As usual, I suppress the $\mathcal{T}$ if it does not play a significant role in the discussion.

The semantics for $\mathcal{L}^{\approx}$ must include some way of interpreting $\approx_i$, $\preceq_i$, and $\succeq_i$. This is done by using a *tolerance vector* $\vec{\tau} = \langle \tau_1, \tau_2, \ldots \rangle$, $\tau_i > 0$. Intuitively $\zeta \approx_i \zeta'$ if the values of ζ and ζ' are within τ_i of each other. (For now there is no need to worry about where the tolerance vector is coming from.) A *statistical-approximation $\mathcal{T}$-structure* is a tuple $(\mathcal{A}, \vec{\tau})$, where $\mathcal{A}$ is a relational $\mathcal{T}$-structure and $\vec{\tau}$ is a tolerance vector. Let $\mathcal{M}^{\approx}(\mathcal{T})$ consist of all statistical-approximation $\mathcal{T}$-structures.

Given a tolerance vector $\vec{\tau}$, a formula $\varphi \in \mathcal{L}^{\approx}$ can be translated to a formula $\varphi^{\vec{\tau}} \in \mathcal{L}^{QU,stat}$. The idea is that a formula such as $\|\varphi|\psi\|_X \preceq_i \alpha$ becomes $\|\varphi|\psi\|_X \le \alpha + \tau_i$; multiplying out the denominator, this is $||\varphi \wedge \psi||_X \le (\alpha + \tau_i)||\psi||_X$. Formally, the translation is defined inductively as follows:

- $\varphi^{\vec{\tau}} = \varphi$ if $\varphi \in \mathcal{L}^{fo}$,
- $(\varphi_1 \wedge \varphi_2)^{\vec{\tau}} = \varphi_1^{\vec{\tau}} \wedge \varphi_2^{\vec{\tau}}$,
- $(\neg\varphi)^{\vec{\tau}} = \neg(\varphi^{\vec{\tau}})$,
- $(\|\varphi|\psi\|_X \preceq_i \alpha)^{\vec{\tau}} = ||(\varphi \wedge \psi)^{\vec{\tau}}||_X \le (\alpha + \tau_i)||\psi^{\vec{\tau}}||_X$,
- $(\|\varphi|\psi\|_X \succeq_i \alpha)^{\vec{\tau}} = ||(\varphi \wedge \psi)^{\vec{\tau}}||_X \ge (\alpha - \tau_i)||\psi^{\vec{\tau}}||_X$,
- $(\|\varphi|\psi\|_X \approx_i \alpha)^{\vec{\tau}} = (\alpha - \tau_i)||\psi^{\vec{\tau}}||_X \le ||(\varphi \wedge \psi)^{\vec{\tau}}||_X \le (\alpha + \tau_i)||\psi^{\vec{\tau}}||_X$.

This translation shows why conditional statistical terms are taken as primitive in $\mathcal{L}^{\approx}$, rather than taking them to be abbreviations for the expressions that result by clearing the denominator. Suppose that the knowledge base *KB* says

$$(||Penguin(x)||_x \approx_1 0) \wedge (\|Flies(x)|Penguin(x)\|_x \approx_2 0);$$

that is, the proportion of penguins is very small but the proportion of fliers among penguins is also very small. Clearing the denominator naively results in the knowledge base

$$KB' = (||Penguin(x)||_x \approx_1 0) \wedge (||Flies(x) \wedge Penguin(x)||_x \approx_2 0 \cdot ||Penguin(x)||_x),$$

which is equivalent to

$$(||Penguin(x)||_x \approx_1 0) \wedge (||Flies(x) \wedge Penguin(x)||_x \approx_2 0).$$

This last formula simply asserts that the proportion of penguins and the proportion of flying penguins are both small, but says nothing about the proportion of fliers among penguins. In fact, the world where all penguins fly is consistent with KB'. Clearly, the process of multiplying out across an approximate connective does not preserve the intended interpretation of the formulas.

In any case, using the translation, it is straightforward to give semantics to formulas in $\mathcal{L}^{\approx}$. For a formula $\varphi \in \mathcal{L}^{\approx}$

$$(\mathcal{A}, \vec{\tau}) \models \varphi \text{ iff } (\mathcal{A}, \mu_N^{unif}) \models \varphi^{\vec{\tau}}.$$

It remains to assign degrees of belief to formulas. Let $W_N(\mathcal{T})$ consist of all D_N-$\mathcal{T}$ structures; let $\#worlds_N^{\vec{\tau}}(\varphi)$ be the number of worlds $\mathcal{A} \in W_N(\mathcal{T})$ such that $(\mathcal{A}, \vec{\tau}) \models \varphi$. Then define the degree of belief in φ given KB with respect to W_N and $\vec{\tau}$ to

$$\mu_N^{\vec{\tau}}(\varphi|KB) = \frac{\#worlds_N^{\vec{\tau}}(\varphi \wedge KB)}{\#worlds_N^{\vec{\tau}}(KB)}.$$

If $\#worlds_N^{\vec{\tau}}(KB) = 0$, the degree of belief is undefined.

Strictly speaking, I should write $\#worlds_N^{\mathcal{T},\vec{\tau}}(\varphi)$ rather than $\#worlds_N^{\vec{\tau}}(\varphi)$, since the number also depends on the choice of $\mathcal{T}$. The degree of belief, however, does not depend on the vocabulary. It is not hard to show that if both $\mathcal{T}$ and $\mathcal{T}'$ contain all the symbols that appear in φ and KB, then

$$\frac{\#worlds_N^{\mathcal{T},\vec{\tau}}(\varphi \wedge KB)}{\#worlds_N^{\vec{\tau}}(KB)} = \frac{\#worlds_N^{\mathcal{T}',\vec{\tau}}(\varphi \wedge KB)}{\#worlds_N^{\vec{\tau}}(KB)}$$

(Exercise 12.1).

Typically, we know neither N nor $\vec{\tau}$ exactly. As I suggested earlier, I want to treat N as "large"; similarly, I treat $\vec{\tau}$ as "small". This suggests taking limits, that is, taking the degree of belief in φ given KB to be $\lim_{\vec{\tau} \to \vec{0}} \lim_{N \to \infty} \mu_N^{\vec{\tau}}(\varphi|KB)$. Notice that the limit is taken first over N for each fixed $\vec{\tau}$ and then over $\vec{\tau}$. This order is important. If the limit $\lim_{\vec{\tau} \to \vec{0}}$ appeared last, then nothing would be gained by using approximate equality, since the result would be equivalent to treating approximate equality as exact equality (Exercise 12.2).

This limit may not exist, for a number of reasons. One obvious one is that $\mu_N^{\vec{\tau}}(\varphi|KB)$ is undefined if $\#worlds_N^{\vec{\tau}}(KB) = 0$. It actually is not important if $\#worlds_N^{\vec{\tau}}(KB) = 0$ for finitely many values of N; in the limit, this is irrelevant. However, what if KB includes a conjunct such as FIN_{100}, which is true only if $N \leq 100$? In that case, $\#worlds_N^{\vec{\tau}}(KB) = 0$ for all

$N > 100$, and the limit will certainly not exist. Of course, if the agent is fortunate enough to know the domain size, then this approach (without taking limits) can be applied to domains of that size. However, in this chapter I will be interested in the case that there are no known upper bounds on the domain size for any given tolerance. More precisely, I focus on knowledge bases KB that are *eventually consistent*, in that there exists $\vec{\tau}^*$ such that for all $\vec{\tau} < \vec{\tau}^*$ ($\vec{\tau} < \vec{\tau}^*$ means that $\vec{\tau}_i < \vec{\tau}_i^*$ for all i) there exists $N_{\vec{\tau}}$ such that $\#worlds_N^{\vec{\tau}}(KB) > 0$ for all $N > N_0$.

Even if KB is eventually consistent, the limit may not exist. For example, it may be the case that for some i, $\mu_N^{\vec{\tau}}(\varphi|KB)$ oscillates between $\alpha + \tau_i$ and $\alpha - \tau_i$ as N gets large. In this case, for any particular $\vec{\tau}$, the limit as N grows does not exist. However, it seems as if the limit as $\vec{\tau}$ grows small "should", in this case, be α, since the oscillations about α go to 0. Such problems can be avoided by considering the *lim sup* and *lim inf*, rather than the limit. The *lim inf* of a sequence is the limit of the infimums; that is,

$$\liminf_{N\to\infty} a_N = \lim_{N\to\infty} (\inf\{a_i : i > N\}).$$

The *lim sup* is defined analogously, using sup instead of inf. Thus, for example, the lim inf of the sequence $0, 1, 0, 1, 0, \ldots$ is 0; the lim sup is 1. The limit clearly does not exist. The lim inf exists for any sequence bounded from below, even if the limit does not; similarly, the lim sup exists for any sequence bounded from above (Exercise 12.3).

The lim inf and lim sup of a sequence are equal iff the limit of the sequence exists and is equal to each of them; that is, $\liminf_{N\to\infty} a_n = \limsup_{N\to\infty} a_n = a$ iff $\lim_{N\to\infty} a_n = a$. Thus, using lim inf and lim sup to define the degree of belief gives a definition that generalizes the one give earlier in terms of limits. Moreover, since, for any $\vec{\tau}$, the sequence $\mu_N^{\vec{\tau}}(\varphi|KB)$ is always bounded from above and below, the lim sup and lim inf always exist. With this motivation, the formal definition of degree of belief given a knowledge base KB can be given.

Definition 12.2.1 If

$$\lim_{\vec{\tau}\to\vec{0}} \liminf_{N\to\infty} \mu_N^{\vec{\tau}}(\varphi|KB) \quad \text{and} \quad \lim_{\vec{\tau}\to\vec{0}} \limsup_{N\to\infty} \mu_N^{\vec{\tau}}(\varphi|KB)$$

both exist and are equal, then the *degree of belief in φ given KB*, written $\Pr_\infty(\varphi|KB)$, is defined as the common limit; otherwise $\Pr_\infty(\varphi|KB)$ does not exist.

Even using this definition, there are many cases where the degree of belief does not exist. This is not necessarily bad. It simply says that the information provided in the knowledge base does not allow the agent to come up with a well-defined degree of belief. There are certainly cases where it is better to recognize that the information is inconclusive rather than trying to create a number. (See Example 12.3.9 for a concrete illustration.)

Definitions cannot be said to be right or wrong; we can, however, try to see whether they are interesting or useful and to what extent they capture our intuitions. In the next four sections, I prove a number of properties of the random-worlds approach to obtaining a degree of belief given a knowledge base consisting of statistical and first-order information, as captured by Definition 12.2.1. The next three sections illustrate some attractive features of the approach; Section 12.6 considers some arguably unattractive features.

12.3 Properties of Random Worlds

Any reasonable method of ascribing degrees of belief given a knowledge base should certainly assign the same degrees of belief to a formula φ given two equivalent knowledge bases. Not surprisingly, random worlds satisfies this property.

Proposition 12.3.1 *If* $\mathcal{M}^{\approx} \models KB \Leftrightarrow KB'$*, then* $\Pr_\infty(\varphi|KB) = \Pr_\infty(\varphi|KB')$ *for all formulas* φ*. (*$\Pr_\infty(\varphi|KB) = \Pr_\infty(\varphi|KB')$ *means that either both degrees of belief exist and have the same value, or neither exists. A similar convention is used in other results.)*

Proof By assumption, precisely the same set of worlds satisfy KB and KB'. Therefore, for all N and $\vec{\tau}$, $\mu_N^{\vec{\tau}}(\varphi|KB)$ and $\mu_N^{\vec{\tau}}(\varphi|KB')$ are equal. Therefore, the limits are also equal. ∎

What about more interesting examples; in particular, what about the examples considered in Section 12.1? First consider the perhaps the simplest case, where there is a single reference class that is precisely the "right one". For example, if KB says that 90% of people with jaundice have hepatitis and Eric has hepatitis, that is, if $KB = \|Hep(x|)\|_{Jaun(x)} x \approx_i .9 \wedge Jaun(Eric)$, then we would certainly hope that $\Pr_\infty(Hep(Eric)|KB) = .9$. (Note that the degree of belief assertion uses equality while the statistical assertion uses approximate equality.) More generally, suppose that the formula $\psi(c)$ represents all the information in the knowledge base about the constant c. In this case, every individual x satisfying $\psi(x)$ agrees with c on all properties for which there is information about c in the knowledge base. If there is statistical information in the knowledge base about the fraction of individuals satisfying ψ that also satisfy φ, then clearly ψ is the most appropriate reference class to use for assigning a degree of belief in $\varphi(c)$.

The next result says that the random-worlds approach satisfies this desideratum. It essentially says that if KB is of the form $\psi(c) \wedge \|\varphi(x)|\psi(x)\|_x \approx_i \alpha \wedge KB'$, and $\psi(c)$ is all the information in KB about c, then $\Pr_\infty(\varphi(c)|KB) = \alpha$. Here, KB' is simply intended to denote the rest of the information in the knowledge base, whatever it may be. But what does it mean that "$\psi(c)$ is all the information in KB about c"? For the purposes of this result, it means (a) that c does not appear in either $\varphi(x)$ or $\psi(x)$ and (b) c does not appear in KB'. To understand why c cannot appear in $\varphi(x)$, suppose that $\varphi(x)$ is $Q(x) \vee x = c$, $\psi(x)$ is *true* and KB is the formula $\|\varphi(x)|\mathit{true}\|_x \approx_1 .5$. If the desired result held without the requirement that c not appear in $\varphi(x)$, it would lead to the erroneous conclusion that $\Pr_\infty(\varphi(c)|KB) = .5$. But since $\varphi(c)$ holds tautologically, it follows that $\Pr_\infty(\varphi(c)|KB) = 1$. To see why the constant c cannot appear in $\psi(x)$, suppose that $\psi(x)$ is $(P(x) \wedge x \neq c) \vee \neg P(x)$, $\varphi(x)$ is $P(x)$, and the KB is $\psi(c) \wedge \|P(x)|\psi(x)\|_x \approx_2 .5$. Again, if the result held without the requirement that c not appear in $\psi(x)$, it would lead to the erroneous conclusion that $\Pr_\infty(P(c)|KB) = .5$. But $\psi(c)$ is equivalent to $\neg P(c)$, so in fact $\Pr_\infty(P(c)|KB) = 0$.

Theorem 12.3.2 *Let KB be a knowledge base of the form*

$$\psi(c) \wedge \|\varphi(x)|\psi(x)\|_x \approx_i \alpha \wedge KB',$$

where $\alpha \in [0,1]$, KB is eventually consistent, and c does not appear in KB', $\varphi(x)$, or $\psi(x)$. Then $\Pr_\infty(\varphi(c)|KB) = \alpha$.

Proof Since KB is eventually consistent, there exist some N_0 and $\vec{\tau}^*$ such that $\#worlds_N^{\vec{\tau}}(KB) > 0$ for all $N > N_0$ and $\vec{\tau} < \vec{\tau}^*$. Fix $N > N_0$ and $\vec{\tau} < \vec{\tau}^*$. The proof strategy is to partition the worlds in $\#worlds_N^{\vec{\tau}}(KB)$ into disjoint clusters and prove that, within each cluster, the fraction of worlds satisfying $\varphi(c)$ is between $\alpha - \tau_i$ and $\alpha + \tau_i$. From this it follows that the fraction of worlds in $\#worlds_N^{\vec{\tau}}(KB)$ satisfying $\varphi(c)$—i.e., the degree of belief in $\varphi(c)$—must also be between $\alpha - \tau_i$ and $\alpha + \tau_i$. The result then follows by letting $\vec{\tau}$ go to 0.

Here are the details: Partition the worlds in $\#worlds_N^{\vec{\tau}}(KB)$ so that two worlds are in the same cluster if and only if they agree on the denotation of all symbols in $\mathcal{T}$ other than c. Let W' be one such cluster. Since ψ does not mention c, the set of individuals $d \in D_N$ such that $\psi(d)$ holds is the same at all the relational structures in W'. That is, given a world $\mathcal{A} \in W'$, let $D_{\mathcal{A},\psi} = \{d \in D_N : (\mathcal{A}, V[x/d]) \models \psi(x)\}$. Then $D_{\mathcal{A},\psi} = D_{\mathcal{A}',\psi}$ for all $\mathcal{A}, \mathcal{A}' \in W'$, since the denotation of all the symbols in $\mathcal{T}$ other than c is the same in $\mathcal{A}$ and $\mathcal{A}'$, and c does not appear in ψ (Exercise 11.3). I write $D_{W',\psi}$ to emphasize the fact that the set of domain elements satisfying ψ

is the same at all the relational structures in W'. Similarly, let $D_{W',\varphi\wedge\psi}$ be the set of domain elements satisfying $\varphi \wedge \psi$ in W'.

Since the worlds in W' all satisfy $KB^{\vec{\tau}}$, they must satisfy $\|\varphi(x)|\psi(x)\|_x \approx_i \alpha$. Thus, $(\tau_i - \alpha)|D_{W',\psi}| \leq |D_{W',\varphi\wedge\psi}| \leq (\tau_i + \alpha)|D_{W',\psi}|$. Since the worlds in W' all satisfy $\psi(c)$, it must be the case that $c^{\mathcal{A}} \in D_{W',\psi}$ for all $\mathcal{A} \in W'$. Moreover, since c is not mentioned in KB except for the statement $\psi(c)$, the denotation of c does not affect the truth of $\|\varphi(x)|\psi(x)\|_x \approx_i \alpha \wedge KB'$. Thus, for each $d \in D_{W',\psi}$ there must be exactly one world $\mathcal{A}_d \in W'$ such that $c^{\mathcal{A}_d} = d$. That is, there is a one-to-one correspondence between the worlds in W' and $D_{W',\psi}$. Similarly, there is a one-to-one correspondence between the worlds in W' satisfying $\varphi(c)$ and $D_{W',\varphi\wedge\psi}$. Therefore, the fraction of worlds in W' satisfying $\varphi(c)$ is in $[\alpha - \epsilon, \alpha + \epsilon]$.

The fraction of worlds in $\#worlds^{\vec{\tau}}_N(KB)$ satisfying $\varphi(c)$ (which is $\mu^{\vec{\tau}}_N(\varphi|KB)$, by definition) is a weighted average of the fraction within the individual clusters. More precisely, if $f_{W'}$ is the fraction of worlds in W' satisfying $\varphi(c)$, then $\mu^{\vec{\tau}}_N(\varphi|KB) = \sum_{W'} f_{W'}|W'|/\#worlds^{\vec{\tau}}_N(KB)$, where the sum is taken over all clusters W' (Exercise 12.4). Since $f_{W'} \in [\alpha - \tau_i, \alpha + \tau_i]$ for all clusters W', it immediately follows that $\mu^{\vec{\tau}}_N(\varphi|KB) \in [\alpha - \tau_i, \alpha + \tau_i]$.

This is true for all $N > N_0$, so $\liminf_{N\to\infty} \mu^{\vec{\tau}}_N(\varphi(c)|KB)$ and $\limsup_{N\to\infty} \mu^{\vec{\tau}}_N(\varphi(c)|KB)$ are both also in the range $[\alpha - \tau_i, \alpha + \tau_i]$. Since this holds for all $\vec{\tau} < \vec{\tau}^*$, it follows that

$$\lim_{\vec{\tau}\to\vec{0}} \liminf_{N\to\infty} \mu^{\vec{\tau}}_N(\varphi(c)|KB) = \lim_{\vec{\tau}\to\vec{0}} \limsup_{N\to\infty} \mu^{\vec{\tau}}_N(\varphi|KB) = \alpha.$$

Thus, $\Pr_\infty(\varphi(c)|KB) = \alpha$. ∎

Theorem 12.3.2 can be generalized in several ways; see Exercise 12.5. However, even this version suffices for a number of interesting conclusions.

Example 12.3.3 Suppose that the doctor sees a patient Eric with jaundice and his medical textbook says that 90% of people with jaundice have hepatitis, 80% of people with hepatitis have a fever, and 5% of people have hepatitis. Let

$$\begin{aligned} KB_{hep} &= Jaun(Eric) \wedge \|Hep(x)|Jaun(x)\|_x \approx_1 .9 \text{ and} \\ KB'_{hep} &= ||Hep(x)||_x \preceq_2 .05 \ \wedge \|Fever(x)|Hep(x)\|_x \approx_3 .8. \end{aligned}$$

Then $\Pr_\infty(Hep(Eric)|KB_{hep} \wedge KB'_{hep}) = .9$ as desired; all the information in KB'_{hep} is ignored. Other kinds of information would also be ignored. For example, if the doctor had information about other patients and other statistical information, this could be added to KB'_{hep} without affecting the conclusion, as long as it did not mention Eric.

Preference for the more specific reference class also follows from Theorem 12.3.2.

Corollary 12.3.4 *Let KB be a knowledge base of the form*

$$\psi_1(c) \land \psi_2(c) \land \|\varphi(x)|\psi_1(x) \land \psi_2(x)\|_x \approx_i \alpha_1 \land \|\varphi(x)|\psi_1(x)\|_x \approx_j \alpha_2 \land KB',$$

where $\alpha_1, \alpha_2 \in [0,1]$*, KB is eventually consistent, and c does not appear in* KB'*,* $\psi_1(x)$*,* $\psi_2(x)$*, or* $\varphi(x)$*. Then* $\Pr_\infty(\varphi(c)|KB) = \alpha_1$*.*

Proof Set $KB'' = \|\varphi(x)|\psi_1(x)\|_x \approx_j \alpha_2 \land KB'$ and let $\psi(x)$ be $\psi_1(x) \land \psi_2(x)$. Observe that $KB = \psi_1(c) \land \psi_2(c) \land \|\varphi(x)|\psi_1(x) \land \psi_2(x)\|_x \approx_i \alpha_1 \land KB''$ and that c does not appear in KB'', so the result follows immediately from Theorem 12.3.2. ∎

As an immediate consequence of Corollary 12.3.4, if the doctor also knows all the facts in knowledge base $KB_{hep} \land KB'_{hep}$ of Example 12.3.3 and, in addition, knows that Eric is a baby and only 10% of babies with jaundice have hepatitis, then the doctor would ascribe degree of belief .1 to Eric's having hepatitis.

Preference for the more specific reference class sometime comes in another guise, that makes it even clearer that the more specific reference class is the smaller one.

Corollary 12.3.5 *Let KB be a knowledge base of the form*

$$\psi_1(c) \land \psi_2(c) \land \forall x(\psi_1(x) \Rightarrow \psi_2(x)) \land \|\varphi(x)|\psi_1(x)\|_x \approx_i \alpha_1 \land \|\varphi(x)|\psi_2(x)\|_x \approx_j \alpha_2 \land KB',$$

where $\alpha_1, \alpha_2 \in [0,1]$*, KB is eventually consistent, and c does not appear in* KB'*,* $\psi_1(x)$*,* $\psi_2(x)$*, or* $\varphi(x)$*. Then* $\Pr_\infty(\varphi(c)|KB) = \alpha_1$*.*

Proof Let KB_1 be identical to KB except without the conjunct $\psi_2(c)$. KB is equivalent to KB_1, since $\models (\psi_1(c) \land \forall x(\psi_1(x) \Rightarrow \psi_2(x))) \Rightarrow \psi_2(c)$. Thus, by Proposition 12.3.1, $\Pr_\infty(\varphi(c)|KB) = \Pr_\infty(\varphi(c)|KB_1)$. The fact that $\Pr_\infty(\varphi(c)|KB_1) = \alpha_1$ is an immediate consequence of Theorem 12.3.2, since $\forall x(\psi_1(x) \Rightarrow \psi_2(x)) \land \|\varphi(x)|\psi_2(x)\|_x \approx_j \alpha_2$ does not mention c, so can be incorporated into KB'. ∎

Example 12.1.1 shows that a preference for the more specific reference class can sometimes be problematic. Why does the random-worlds not encounter this problem? The following example suggests one answer.

Example 12.3.6 Let $\psi(x) =_{\rm def} \mathit{Jaun}(x) \land (\neg \mathit{Hep}(x) \lor x = \mathit{Eric})$. Fix a small $\epsilon > 0$ and let $KB''_{hep} = KB_{hep} \land \|\mathit{Hep}(x)|\psi(x)\|_x < \epsilon$. Now, given KB''_{hep}, $\psi(x)$ is more specific than $\mathit{Jaun}(x)$. That is, $\models KB''_{hep} \Rightarrow$

$\forall x(\psi(x) \Rightarrow Jaun(x))$. Corollary 12.3.5 seems to suggest that the doctor's degree of belief that Eric has hepatitis should be less than ϵ. However, this is not the case; Corollary 12.3.5 does not apply because $\psi(x)$ mentions *Eric*. This observation suggests that what makes the reference class used in Example 12.1.1 fishy is that it mentions Eric. A reference class that explicitly mentions Eric should not be used in deriving a degree of belief regarding Eric, even if very good statistics are available for that reference class. (In fact, it can be shown that $\Pr_\infty(Hep(Eric)|KB''_{hep}) = \Pr_\infty(Hep(Eric)|KB_{hep}) = .9$, since in fact $\Pr_\infty(\|Hep(x)|\psi(x)\|_x < \epsilon|KB_{hep}) = 1$: the new information in KB''_{hep} holds in almost all worlds that satisfy KB_{hep}, so does not really add anything. However, a proof of this fact is beyond the scope of this book.) ∎

In Theorem 12.3.2, the knowledge base is assumed to have statistics for precisely the right reference class to match the knowledge about the individual(s) in question. Unfortunately, in many cases, the available statistical information is not detailed enough for Theorem 12.3.2 to apply. Consider the knowledge base KB_{hep} from the hepatitis example, and suppose that the doctor also knows that Eric has red hair; that is, his knowledge is characterized by $KB_{hep} \wedge Red(Eric)$. Since the knowledge base does not include statistics for the frequency of hepatitis among red-haired individuals, Theorem 12.3.2 does not apply. We would like to be able to ignore $Red(Eric)$. But what entitles us to ignore $Red(Eric)$ and not $Jaun(Eric)$? To solve this problem in complete generality would require a detailed theory of irrelevance, perhaps using the ideas of conditional independence from Chapter 5. Such a theory is not yet available. Nevertheless, the next theorem shows that, if irrelevance is taken to mean "uses only symbols not mentioned in the relevant statistical likelihood formula", the random-worlds approach gives the desired result. Roughly speaking, the theorem says that if the KB includes the information $\|\varphi(x)|\psi(x)\|_x \approx_i \alpha \wedge \psi(c)$, and perhaps a great deal of other information (including possibly information about c), then the degree of belief in $\varphi(c)$ is still α, provided that the other information about c does no involve symbols that appear in φ and whatever other statistics are available about φ in the knowledge base are "subsumed" by the information $\|\varphi(x)|\psi(x)\|_x \approx_i \alpha$. "Subsumed" here means that for any other statistical term of the form $\|\varphi(x)|\psi'(x)\|_x$ either $\forall x(\psi(x) \Rightarrow \psi'(x))$ or $\forall x(\psi(x) \Rightarrow \neg\psi'(x))$ follows from the knowledge base.

Theorem 12.3.7 *Let KB be a knowledge base of the form*

$$\psi(c) \wedge \|\varphi(x)|\psi(x)\|_x \approx_i \alpha \wedge KB',$$

where

(a) $\alpha \in [0,1]$,

(b) *KB is eventually consistent,*

(c) *c does not appear in* $\varphi(x)$,

(d) *none of the symbols in* $\mathcal{T}$ *that appear in* $\varphi(x)$ *appear in* $\psi(x)$ *or* KB', *except possibly in statistical expressions of the form* $\|\varphi(x)|\psi'(x)\|_x$; *moreover, for any such expression, either* $\mathcal{M}^{\approx} \models \forall x(\psi(x) \Rightarrow \psi'(x))$ *or* $\mathcal{M}^{\approx} \models \forall x(\psi(x) \Rightarrow \neg\psi'(x))$.

Then $\Pr_\infty(\varphi(c)|KB) = \alpha$.

Proof Just as in the proof of Theorem 12.3.2, the key idea involves partitioning the worlds in $\#worlds_N^{\vec{\tau}}(KB)$ appropriately. The details are left to Exercise 12.6. ∎

Note how Theorem 12.3.7 differs from Theorem 12.3.2. In Theorem 12.3.2, c cannot appear in $\psi(x)$ or KB'. In Theorem 12.3.7, c is allowed to appear in $\psi(x)$ and KB', but no symbol in $\mathcal{T}$ that appears in $\varphi(x)$ may appear in $\psi(x)$ or KB'. Thus, if $\varphi(x)$ is $P(x)$, then $\psi(x)$ cannot be $(P(x) \wedge x \neq c) \vee \neg P(x)$, because P cannot appear in $\psi(x)$.

From Theorem 12.3.7, it follows immediately that $\Pr_\infty(\mathit{Hep}(\mathit{Eric})|KB_{hep} \wedge \mathit{Red}(\mathit{Eric})) = .9$. The degree of belief would continue to be .9 even if other information about Eric were added to KB_{hep}, such as Eric has a fever and Eric is a baby, as long as the information did not involve the predicate *Hep*.

I now consider a different issue: competing reference classes. In all the examples I have considered so far, there is an obviously "best" reference class. In practice, this will rarely be the case. It seems difficult to completely characterize the behavior of the random-worlds approach on arbitrary knowledge bases (although the connection between random worlds and maximum entropy described in Section 12.5 certainly gives some insight). Interestingly, if there are competing reference classes that are essentially disjoint, Dempster's Rule of Combination can be used to compute the degree of belief.

For simplicity, assume that the knowledge base consists of exactly two pieces of statistical information, both about a unary predicate P—$\|P(x)|\psi_1(x)\|_x \approx_i \alpha_1$ and $\|P(x)|\psi_2(x)\|_x \approx_j \alpha_2$—and, in addition, the knowledge base says that there is exactly one individual satisfying both $\psi_1(x)$ and $\psi_2(x)$; that is, the knowledge base includes the formula $\exists! x(\psi_1(x) \wedge \psi_2(x))$. (See Exercise 12.7 for the precise definition of $\exists! x\varphi(x)$.) The two statistical likelihood formulas can be viewed as providing evidence in favor of P to degree α_1 and α_2, respectively. Consider two probability measures μ_1 and μ_2 on a two-point space $\{0,1\}$ such that $\mu_1(1) = \alpha_1$ and $\mu_2(1) = \alpha_2$. (Think of

$\mu_1(1)$ as describing the degree of belief that $P(c)$ is true according to the evidence provided by the statistical formula $\|P(x)|\psi_1(x)\|_x$ and $\mu_2(1)$ as describing the degree of belief that $P(c)$ is true according to $\|P(x)|\psi_2(x)\|_x$.) As we saw in Section 3.3, according to Dempster's Rule of Combination, $\mu_1 \oplus \mu_2 = \frac{\alpha_1\alpha_2}{\alpha_1\alpha_2+(1-\alpha_1)(1-\alpha_2)}$. The next theorem shows that this is also the degree of belief ascribed by the random-worlds approach.

Theorem 12.3.8 *Let KB be a knowledge base of the form*

$$\|P(x)|\psi_1(x)\|_x \approx_i \alpha_1 \land \|P(x)|\psi_2(x)\|_x \approx_j \alpha_2 \land \psi_1(c) \land \psi_2(c) \land \exists! x(\psi_1(x) \land \psi_2(x)),$$

where KB is eventually consistent, P is a unary predicate, neither P nor c appear in $\psi_1(x)$ or $\psi_2(x)$, and either $\alpha_1 < 1$ and $\alpha_2 < 1$ or $\alpha_1 > 0$ and $\alpha_2 > 0$. Then $\Pr_\infty(P(c)|KB) = \frac{\alpha_1\alpha_2}{\alpha_1\alpha_2+(1-\alpha_1)(1-\alpha_2)}$.

Proof Again, the idea is to appropriately cluster the worlds in $\#worlds_N^{\vec{\tau}}(KB)$. See Exercise 12.8. ∎

This result can be generalized to allow more than two pieces of statistical information; Dempster's Rule of Combination still applies (Exercise 12.9). It is also not necessary to assume that there is a unique individual satisfying both ψ_1 and ψ_2. It suffices that the set of individuals satisfying $\psi_1 \land \psi_2$ is "small" relative to the set satisfying ψ_1 and the set satisfying ψ_2, although the technical details are beyond the scope of this book.

The following example illustrates Theorem 12.3.8.

Example 12.3.9 Suppose that we are interested in assigning a degree of belief to the assertion "Nixon is a pacifist". Assume that the knowledge base consists of the information that Nixon is both a Quaker and a Republican, and there is statistical information for the proportion of pacifists within both classes. More formally, assume that KB_{Nixon} is

$$\begin{array}{l} \|Pac(x)|Quak(x)\|_x \approx_1 \alpha \land \\ \|Pac(x)|Repub(x)\|_x \approx_2 \beta \land \\ Quak(Nixon) \land Repub(Nixon) \land \\ \exists! x\,(Quak(x) \land Repub(x)), \end{array}$$

and that φ is $Pac(Nixon)$. The degree of belief $\Pr_\infty(\varphi|KB_{Nixon})$ takes different values, depending on the values α and β for the two reference classes. If $\{\alpha, \beta\} \neq \{0, 1\}$, then $\Pr_\infty(\varphi|KB_{Nixon})$ always exists and its value is equal to $\frac{\alpha\beta}{\alpha\beta+(1-\alpha)(1-\beta)}$. If, for example, $\beta = .5$, so that the information for Republicans is neutral, then $\Pr_\infty(\varphi|KB_{Nixon}) = \alpha$: the data for Quakers is used to determine the degree of belief. If the evidence given by the two reference classes is conflicting—$\alpha > .5 > \beta$—then $\Pr_\infty(\varphi|KB_{Nixon}) \in [\alpha, \beta]$: some intermediate value is chosen. If, on the other hand, the two reference classes provide evidence in the same direction, then the degree of belief is greater than both α and β. For example, if $\alpha = \beta = .8$, then the degree of belief is about .94. This has a reasonable explanation: if there two independent bodies of evidence both supporting φ, then their combination should provide even more support for φ.

Now assume that $\alpha = 1$ and $\beta > 0$. In that case, it follows from Theorem 12.3.8 that $\Pr_\infty(\varphi|KB_{Nixon}) = 1$. Intuitively, an extreme value dominates. But what happens if the extreme values conflict? For example, suppose that $\alpha = 1$ and $\beta = 0$. This says that almost all Quakers are pacifists and almost no Republicans are. In that case, Theorem 12.3.8 does not apply. In fact, it can be shown that the degree of belief does not exist. This is because the value of the limit depends on the way in which the tolerances $\vec{\tau}$ tend to 0. More precisely, if $\tau_1 \ll \tau_2$, so that the "almost all" in the statistical interpretation of the first conjunct is much closer to "all" than the "almost none" in the second conjunct is closer to "none", then the limit is 1. Symmetrically, if $\tau_1 \gg \tau_2$, then the limit is 0. On the other hand, if $\tau_1 = \tau_2$, then the limit is 1/2. (In particular, this means that if the subscript 1 were used for the $\approx$ in both statistical assertions, then the degree of belief would be 1/2.)

There are good reasons for the limit not to exist in this case. The knowledge base simply does not say what the relationship between τ_1 and τ_2 is. (It would certainly be possible, of course, to consider a richer language that allows such relationships to be expressed.) ∎

12.4 Random Worlds and Default Reasoning

One of the most attractive features of the random-worlds approach is that it provides a well-motivated system of default reasoning, with a number of desirable properties. Recall that at the end of Chapter 11 I observed that if "birds typically fly" is interpreted as a statistical assertion and "Tweety flies" is interpreted as a statement about a (high) degree of belief, then in order to do default reasoning and, in particular, conclude that Tweety the bird flies from the fact that birds typically fly, we need to have some way of connecting statistical assertions with statements about degrees of belief. The random-worlds approach provides precisely such a connection.

The first step in exploiting this connection is to find an appropriate representation for "birds typically fly". The intuition here goes back to that presented in Chapter 7: "birds typically fly" should mean that birds are very likely to fly. Probabilistically, this should mean that the probability that given bird flies is very high. As we saw in Section 7.2.1, there are problems deciding how high is high enough: it will not work (in the sense of not giving us the KLM properties) to take "high" to be "with probability greater than $1 - \epsilon$" for some fixed ϵ. One way of dealing with that problem that was presented in Section 7.2.1 involved using sequences of probabilities. The language $\mathcal{L}^{\approx}$ is expressive enough to provide another approach—using approximate equality. "Birds typically fly" becomes $\|Flies(x)|Bird(x)\|_x \approx_i 1$. (The exact choice of subscript on $\approx$ is not important, although if there are several defaults, it may be important to use different subscripts for each one; I return to this issue below.)

This way of expressing defaults can be used to express far more complicated defaults than can be represented in propositional logic, as the following examples show.

Example 12.4.1 Consider the fact that people who have at least one tall parent are typically tall. This default can be expressed in as

$$\|\mathit{Tall}(x)|\exists y\,(\mathit{Child}(x,y) \land \mathit{Tall}(y))\|_x \approx_i 1. \quad \blacksquare$$

Example 12.4.2 Typicality statements can have nesting. For example, consider the nested default "typically, people who normally go to bed late normally rise late". This can be expressed using nested statistical assertions. The individuals who normally rise late are those who rise late most days; these are the individuals x satisfying $\|\mathit{Rises\text{-}late}(x,y)|\mathit{Day}(y)\|_y \approx_1 1$. Similarly, the individuals who normally go to bed late are those satisfying $\|\mathit{To\text{-}bed\text{-}late}(x,y')|\mathit{Day}(y')\|_{y'} \approx_2 1$. Thus the default can be captured by saying most individuals x that go to bed late also rise late:

$$\Big\|\|\mathit{Rises\text{-}late}(x,y)|\mathit{Day}(y)\|_y \approx_1 1 \;\Big|\; \|\mathit{To\text{-}bed\text{-}late}(x,y')|\mathit{Day}(y')\|_{y'} \approx_2 1\Big\|_x \approx_3 1.$$

On the other hand, the related default that "Typically, people who go to bed late rise late (the next morning)" can be expressed as:

$$\Big\|\mathit{Rises\text{-}late}(x, \mathit{Next\text{-}day}(y)) \;\Big|\; \mathit{Day}(y) \land \mathit{To\text{-}bed\text{-}late}(x,y)\Big\|_{x,y} \approx_1 1. \quad \blacksquare$$

Representing typicality statements is only half the battle. What about a conclusion such as "Tweety flies"? This corresponds to a degree of belief of 1. More precisely, given a knowledge base KB (which, for example, may include $\|Flies(x)|Bird(x)\|_x \approx_i 1$), the default conclusion "Tweety flies" follows from KB if $\Pr_\infty(\varphi|KB) = 1$.

The formula φ is a *default conclusion from KB*, written $KB \mathrel{|\!\!\sim}_{rw} \varphi$, if $\Pr_\infty(\varphi|KB) = 1$. Note that it follows immediately from Theorem 12.3.2

that $\|Flies(x)|Bird(x)\|_x \approx_i 1 \mathrel{|\!\!\sim}_{rw} Flies(Tweety)$. That is, the conclusion "Tweety flies" does indeed follow from "Birds typically fly". Moreover, if Tweety is a penguin then it follows that Tweety does not fly. That is, if

$$\begin{aligned} KB_1 = \quad & \|Flies(x)|Bird(x)\|_x \approx_1 1 \wedge \|Flies(x)|Penguin(x)\|_x \approx_2 0 \wedge \\ & \forall x(Penguin(x) \Rightarrow Bird(x)) \wedge Penguin(Tweety), \end{aligned}$$

then it is immediate from Theorem 12.3.2 that

$$KB_1 \mathrel{|\!\!\sim}_{rw} \neg Flies(Tweety).$$

(The same conclusion would also hold if $\forall x(Penguin(x) \Rightarrow Bird(x))$ were replaced by $\|Penguin(x)|Bird(x)\|_x \approx_3$; the latter formula is closer to what was used in Section 7.3, but the former better represents the actual state of affairs.)

In fact, the theorems of Section 12.3 show that quite a few other desirable conclusions follow. Before getting into them, let's first establish that the relation $\mathrel{|\!\!\sim}_{rw}$ satisfies the axioms of **P** described in Section 7.1, since these are considered the core properties of default reasoning.

Theorem 12.4.3 *The relation $\mathrel{|\!\!\sim}_{rw}$ satisfies the axioms of* **P**. *More precisely, the following properties hold if KB and KB' are eventually consistent:*

LLE. *If $\mathcal{M}^{\approx} \models KB \Leftrightarrow KB'$, then $KB \mathrel{|\!\!\sim}_{rw} \varphi$ iff $KB' \mathrel{|\!\!\sim}_{rw} \varphi$.*

RW. *If $\mathcal{M}^{\approx} \models \varphi \Rightarrow \varphi'$, then $KB \mathrel{|\!\!\sim}_{rw} \varphi$ implies $KB \mathrel{|\!\!\sim}_{rw} \varphi'$.*

REF. *$KB \mathrel{|\!\!\sim}_{rw} KB$.*

AND. *If $KB \mathrel{|\!\!\sim}_{rw} \varphi$ and $KB \mathrel{|\!\!\sim}_{rw} \psi$, then $KB \mathrel{|\!\!\sim}_{rw} \varphi \wedge \psi$.*

OR. *If $KB \mathrel{|\!\!\sim}_{rw} \varphi$ and $KB' \mathrel{|\!\!\sim}_{rw} \varphi$, then $KB \vee KB' \mathrel{|\!\!\sim}_{rw} \varphi$.*

CM. *If $KB \mathrel{|\!\!\sim}_{rw} \varphi$ and $KB \mathrel{|\!\!\sim}_{rw} \varphi'$, then $KB \wedge \varphi \mathrel{|\!\!\sim}_{rw} \varphi'$.*

Proof LLE follows immediately from (indeed, is just a restatement of) Proposition 12.3.1. RW is immediate from the observation that $\mu_N^{\vec{\tau}}(\varphi|KB) \geq \mu_N^{\vec{\tau}}(\varphi'|KB)$ if $\mathcal{M}^{\approx} \models \varphi \Rightarrow \varphi'$ (provided that $\#worlds_N^{\vec{\tau}}(KB) \neq 0$). REF is immediate from the fact that $\mu_N^{\vec{\tau}}(KB|KB) = 1$, provided that $\#worlds_N^{\vec{\tau}}(KB) \neq 0$. I leave the proof of AND, OR, and CM to the reader (Exercise 12.10). ∎

Not only does $\mathrel{|\!\!\sim}_{rw}$ satisfy the axioms of **P**, it can go well beyond **P**. Let KB_1 be the knowledge base described earlier, which says that birds typically fly, penguins don't, and Tweety is a penguin. Then the following are all immediate consequences of Theorems 12.3.2 and 12.3.7:

- red penguins do not fly:

$$KB_1 \land \mathit{Red}(\mathit{Tweety}) \mathrel{|\!\!\sim}_{rw} \neg \mathit{Flies}(\mathit{Tweety});$$

- if birds typically have wings, then both robins and birds have wings:

$$KB_1 \land \|\mathit{Winged}(x)|\mathit{Bird}(x)\|_x \approx_3 1 \land \forall x(\mathit{Robin}(x) \Rightarrow \mathit{Bird}(x)) \land \mathit{Robin}(\mathit{Sweety}) \\ \mathrel{|\!\!\sim}_{rw} \mathit{Winged}(\mathit{Sweety});$$

- if yellow things are typically easy to see, then yellow penguins are easy to see:

$$KB_1 \land \|\mathit{Easy\text{-}to\text{-}See}(x)|\mathit{Yellow}(x)\|_x \approx_4 1 \land \mathit{Yellow}(\mathit{Tweety}) \\ \mathrel{|\!\!\sim}_{rw} \mathit{Easy\text{-}to\text{-}See}(\mathit{Tweety}).$$

Thus, the random-worlds approach gives all the results that we were hoping to get in Section 7.3.

The next two examples show how **P** can be combined with Theorems 12.3.2 and 12.3.7 to give further results.

Example 12.4.4 Suppose that the predicates $\mathit{LUsable}$, $\mathit{LBroken}$, $\mathit{RUsable}$, $\mathit{RBroken}$ indicate, respectively, that the left arm is usable, the left arm is broken, the right arm is usable, and the right arm is broken. Let KB'_{arm} consist of the statements

- $||\mathit{LUsable}(x)||_x \approx_1 1$, $\|\mathit{LUsable}(x)|\mathit{LBroken}(x)\|_x \approx_2 0$ (left arms are typically usable, but not if they are broken),
- $||\mathit{RUsable}(x)||_x \approx_3 1$, $\|\mathit{RUsable}(x)|\mathit{RBroken}(x)\|_x \approx_4 0$ (right arms are typically usable, but not if they are broken).

Now, consider

$$KB_{arm} = (KB'_{arm} \land (\mathit{LBroken}(\mathit{Eric}) \lor \mathit{RBroken}(\mathit{Eric})));$$

the last conjunct KB_{arm} just says that at least one of Eric's arms is broken (but does not specify which one or ones). From Theorem 12.3.2 it follows that

$$KB'_{arm} \land \mathit{LBroken}(\mathit{Eric}) \mathrel{|\!\!\sim}_{rw} \neg \mathit{LUsable}(\mathit{Eric}).$$

From Theorem 12.3.7, it follows that

$$KB'_{arm} \land \mathit{LBroken}(\mathit{Eric}) \mathrel{|\!\!\sim}_{rw} \mathit{RUsable}(\mathit{Eric}).$$

The AND rule gives

$$KB'_{arm} \land \mathit{LBroken}(\mathit{Eric}) \mathrel{|\!\!\sim}_{rw} \mathit{RUsable}(\mathit{Eric}) \land \neg \mathit{LUsable}(\mathit{Eric})$$

and RW then gives

$$KB'_{arm} \land LBroken(Eric) \mathrel{|\!\!\sim}_{rw} (\neg LUsable(Eric) \land RUsable(Eric)) \lor (\neg RUsable(Eric) \land LUsable(Eric)).$$

Similar reasoning shows that

$$KB'_{arm} \land RBroken(Eric) \mathrel{|\!\!\sim}_{rw} (\neg LUsable(Eric) \land RUsable(Eric)) \lor (\neg RUsable(Eric) \land LUsable(Eric)).$$

The *Or* rule then gives

$$KB_{arm} \mathrel{|\!\!\sim}_{rw} (\neg LUsable(Eric) \land RUsable(Eric)) \lor (\neg RUsable(Eric) \land LUsable(Eric)).$$

That is, by default it follows from KB_{arm} that exactly one of Eric's arms is usable, but we can draw no conclusions as to which one it is. This seems reasonable: given that arms are typically not broken, knowing that at least arm is broken should lead to the conclusion that exactly one is broken, but not which one it is. ∎

Example 12.4.5 Recall that Example 12.4.2 showed how the nested typicality statement "typically, people who normally go to bed late normally rise late" can be expressed by the knowledge base KB_{late}:

$$\Big\| \|Rises\text{-}late(x,y) | Day(y)\|_y \approx_1 1 \;\Big|\; \|To\text{-}bed\text{-}late(x,y') | Day(y')\|_{y'} \approx_2 1 \Big\|_x \approx_3 1.$$

Let KB'_{late} be

$$KB_{late} \;\land\; \|To\text{-}bed\text{-}late(Alice,y') | Day(y')\|_{y'} \approx_2 1 \land Day(Tomorrow).$$

Taking $\psi(x)$ to be $\|To\text{-}bed\text{-}late(x,y') | Day(y')\|_{y'} \approx_2 1$ and applying Theorem 12.3.2, it follows that Alice typically rises late. That is,

$$KB'_{late} \mathrel{|\!\!\sim}_{rw} \|Rises\text{-}late(Alice,y) | Day(y)\|_y \approx_1 1.$$

By Theorem 12.3.2 again, it follows that

$$KB'_{late} \land \|Rises\text{-}late(Alice,y) | Day(y)\|_y \approx_1 1 \mathrel{|\!\!\sim}_{rw} Rises\text{-}late(Alice, Tomorrow).$$

The CUT Rule (Exercise 12.12) says that if $KB \mathrel{|\!\!\sim}_{rw} \varphi$ and $KB \land \varphi \mathrel{|\!\!\sim}_{rw} \psi$ then $KB \mathrel{|\!\!\sim}_{rw} \psi$. Thus, $KB'_{late} \mathrel{|\!\!\sim}_{rw} Rises\text{-}late(Alice, Tomorrow)$: by default, Alice will rise late tomorrow (and every other day, for that matter). ∎

Finally, consider the lottery paradox from Example 11.4.3, which caused so many problems in Section 11.4.

Example 12.4.6 The knowledge base corresponding to the lottery paradox is just

$$KB_{lottery} = \exists x\, Winner(x) \land ||\, Winner(x)||_x \approx_1 0.$$

This knowledge base is clearly eventually consistent. Moreover, it immediately follows from Theorem 12.3.2 that $KB_{lottery} \mathrel{|\!\sim}_{rw} \neg\, Winner(c)$ for any particular individual c. From RW, it is also immediate that $KB_{lottery} \mathrel{|\!\sim}_{rw} \exists x Winner(x)$. The expected answer drops right out.

An objection to the use of the random worlds approach here might be that it depends on the domain size growing unboundedly large. Suppose that, for example, we know that N tickets to the lottery are sold. To simplify the analysis, further suppose that exactly one person wins the lottery. winner and in order to win one must purchase a lottery ticket. Let $Ticket(x)$ denote that x purchased a lottery ticket and let

$$KB'_{lottery} = \exists! x\ Winner(x) \land \forall x\, (Winner(x) \Rightarrow Ticket(x)) \land Ticket(c).$$

With no further assumptions, it is not hard to show that $KB'_{lottery} \mathrel{|\!\sim}_{rw} \neg\, Winner(c)$, i.e., $\Pr_\infty(\,Winner(c)|KB'_{lottery}) = 0$ (Exercise 12.13(a)).

Now let $KB''_{lottery} = KB'_{lottery} \land \exists^N x\ Ticket(x)$, where $\exists^N x\ Ticket(x)$ is the formula stating that there are precisely N ticket holders. (This assertion can easily be expressed in first-order logic—see Exercise 12.7.) Then it is easy to see that $\Pr_\infty(\,Winner(c)|KB''_{lottery}) = 1/N$. That is, the degree of belief that any particular individual c wins the lottery is $1/N$. This numeric answer seems just right: it simply says that the lottery is fair. Note that this conclusion is not part of the knowledge base. Essentially, the random-worlds approach is concluding fairness in the absence of any other information. ∎

12.5 Random Worlds and Maximum Entropy

The entropy function has been used in a number of context in reasoning about uncertainty. As mentioned in the notes to Chapter 4, it was originally introduced in the context of information theory, where it was viewed as the amount of "information" in a probability measure. Intuitively, a uniform measure, which has high entropy, gives less information about the actual situation than a distribution that puts probability 1 on a single point (which has the lowest possible entropy, 0). The entropy function, specifically maximum entropy, was used in Section 7.3 to define a probability sequence that had some desirable properties for default reasoning. Another common usage of entropy is in the context of trying to pick a single probability measure among a set of possible probability measures characterizing a situation, defined by some constraints. The *principle of maximum entropy*, first espoused by Jaynes, suggests choosing the measure with the maximum entropy (provided that there is in fact a unique such measure), because it incorporates in some sense the "least additional information" above and beyond the constraints that characterize the set.

No explicit use of maximum entropy is made by the random-worlds approach. Indeed, although they are both tools for reasoning about probabilities, the types of problems considered by the random-worlds approach and maximum entropy techniques seem unrelated. Nevertheless, it turns out that there is a surprising and very close connection between the random-worlds approach and maximum entropy provided that the vocabulary consists only of *unary* predicates and constants. In this section I briefly describe this connection, without going into technical details.

Suppose that the vocabulary $\mathcal{T}$ consists of the unary predicate symbols $P_1, \ldots, P_k$ together with some constant symbols. (Thus, $\mathcal{T}$ includes neither function symbols nor higher-arity predicates.) Consider the 2^k *atoms* that can be formed from these predicate symbols, namely, the formulas of the form $Q_1 \wedge \ldots \wedge Q_k$, where each Q_i is either P_i or $\neg P_i$. (Strictly speaking, I should write $Q_i(x)$ for some variable x, not just Q_i. I omit the parenthesized x here, since it just adds clutter.) The knowledge base KB can be viewed as simply placing constraints on the proportion of domain elements satisfying each atom. For example, the formula $\|P_1(x)|P_2(x)\|_x = .6$ says that the fraction of domain elements satisfying some atoms containing both P_1 and P_2 as conjuncts is .6 times the fraction satisfying atoms containing P_1 as a conjunct. For unary languages (only) it can be shown that every formula can be rewritten in a canonical form from which constraints on the possible proportions of atoms can be simply derived. For example, if $\mathcal{T} = \{c, P_1, P_2\}$, there are four atoms: $A_1 = P_1 \wedge P_2$, and $A_2 = P_1 \wedge \neg P_2$, then $\|P_1(x)|P_2(x)\|_x = .6$ is equivalent to $||A_1(x) \vee A_2(x)||_x = .6||A_1(x)||_x$. (Using $\approx$ instead of $=$ would make things somewhat more complicated here.)

The set of constraints generated by KB (with $\approx$ replaced by $=$) defines a subset of $[0,1]^{2^k}$ called $S(KB)$. That is, each vector in $S(KB)$, say $\vec{p} = \langle p_1, \ldots, p_{2^k}\rangle$, is a solution to the constraints defined by KB (where p_i is the proportion of atom i). For example, if $\mathcal{T} = \{c, P_1, P_2\}$, and $KB = \|P_1(x)|P_2(x)\|_x = .6$ as above, then the only constraint is that $p_1 + p_2 \leq .6p_1$. That is, $S(KB) = \{\langle p_1, \ldots, p_4\rangle \in [0,1]^4 : p_1 + p_2 \leq 6.p_1\}$.

For another example, suppose that $KB' = \forall x\, P_1(x) \wedge ||P_1(x) \wedge P_2(x)||_x \preceq_1 .3$. The first conjunct of KB' clearly constrains both p_3 and p_4 (the proportion of domain elements satisfying atoms A_3 and A_4) to be 0. The second conjunct forces p_1 to be (approximately) at most .3. Thus, $S(KB') = \{\langle p_1, \ldots, p_4\rangle \in [0,1]^4 : p_1 \leq .3, p_3 = p_4 = 0, p_1 + p_2 = 1\}$.

The connection between maximum entropy and the random-worlds approach is based on the following observations. Every world w can be associated with the vector $\vec{p}^w$, where p_i^w is the fraction of domain elements in world w satisfying the atom A_i. Each vector $\vec{p}$ can be viewed as a probability measure on the space of atoms $A_1, \ldots, A_{2^k}$; therefore, each such vector $\vec{p}$ has an associated entropy, $H(\vec{p}) = -\sum_{i=1}^{2^k} p_i \log p_i$ (where, as before, $p_i \log p_i$ is taken to be 0 if $p_i = 0$). Define the *entropy* of w to be $H(\vec{p}^w)$. Now, consider some point $\vec{p} \in S(KB)$. What is the number of worlds $w \in W_N$ such that $\vec{p}^w = \vec{p}$? Clearly, for those $\vec{p}$ where some p_i is not an integer multiple of $1/N$, the answer is 0. However, for those $\vec{p}$ which are "possible", this number can be shown to grow asymptotically as $e^{NH(\vec{p})}$ (Exercise 12.15). Thus, there are vastly more worlds w for which $\vec{p}^w$ is "near" the maximum entropy point of $S(KB)$ than there are worlds elsewhere. The following result then follows: If, for all sufficiently small $\vec{\tau}$, a formula θ is true in all worlds around the maximum entropy point(s) of $S(KB)$, then $\Pr_\infty(\theta|KB) = 1$.

In the examples above, the maximum-entropy point in $S(KB)$ is $(.25, .25, .25, .25)$: this is the maximum entropy point without any constraints, and it is compatible with the constraints. The maximum entropy point of $S(KB')$ is $(.3, .7, 0, 0)$. (It must be the case that the last two components are 0 since this is true in all of $S(KB')$; the first two components are "as equal as possible" subject to the constraints, and this maximizes entropy (cf. Exercise 4.29).) But now consider KB' some small fixed ϵ and the formula $\theta^\epsilon = ||P_2(x)||_x \in [.3-\epsilon, .3+\epsilon]$. Since this formula certainly holds at all worlds w where $\vec{p}^w$ is sufficiently close to $\vec{p}^*$, it follows that $\Pr_\infty(\theta^\epsilon|KB') = 1$. The generalization of Theorem 12.3.2 given in Exercise 12.5 implies that $\Pr_\infty(P_2(c)|KB' \wedge \theta^\epsilon) \in [.3-\epsilon, .3+\epsilon]$. It follows from Exercise 12.12 that $\Pr_\infty(\psi|KB' \wedge \theta^\epsilon) = \Pr_\infty(\psi|KB')$ for all formulas ψ and, hence, in particular, for $P_2(c)$. Since $\Pr_\infty(P_2(c)|KB') \in [.3-\epsilon, .3+\epsilon]$ for all sufficiently small ϵ, it follows that $\Pr_\infty(P_2(c)|KB') = .3$, as desired. That is, the degree of belief in $P_2(c)$ given KB' is the probability of P_2 (i.e., the sum of the probabilities of the atoms that imply P_2) in the measure of maximum entropy satisfying the constraints determined by KB'. A similar argument shows that $\Pr_\infty(P_2(c)|KB) = .5$. Again, the degree of belief in $P_2(c)$ given KB is the probability of $P_2 = A_1 \vee A_2$ in the measure of maximum entropy satisfying the constraints determined by KB.

Thus, the random-worlds approach can be viewed as providing justification for the use of maximum entropy, at least when only unary predicates are involved. Indeed, random worlds can be viewed as a generalization of maximum entropy to cases where there are non-unary predicates.

These results connecting random worlds to maximum entropy also shed light on the maximum-entropy approach to default reasoning considered in Section 7.3. Indeed, the maximum-entropy approach can be embedded in the random worlds approach. Let Σ be a collection of propositional defaults (i.e., formulas of the form $\varphi \rightarrow \psi$) which mention the primitive propositions $\{p_1, \ldots, p_n\}$. Let $\{P_1, \ldots, P_n\}$ be unary predicates. Convert each default $\theta = \varphi \rightarrow \psi \in \Sigma$ to the formula $\theta^r = \|\psi^*(x)|\varphi^*(x)\|_x \approx_1 1$, where ψ^* and φ^* are obtained by replacing each occurrence of a primitive proposition p_i by $P_i(x)$. Thus, the translation treats a propositional default statement as a statistical assertion about sets of individuals. Note that all the formulas θ^r use the same approximate equality relation $\approx_1$. This is essentially because the maximum-entropy approach treats all the defaults in Σ as having the same strength (in the sense of Example 12.3.9). This comes out in the maximum-entropy approach in the following way: Recall that in the probability sequence $(\mu_1^{me}, \mu_2^{me}, \ldots)$, the kth probability measure μ_k^{me} is the measure of maximum entropy among all those satisfying Σ^k, where Σ^k is the result of replacing each default $\varphi \rightarrow \psi \in \Sigma$ by the $\mathcal{L}^{QU}$ formula $\ell(\psi|\varphi) \geq 1 - 1/k$. That is, $1 - 1/k$ is used for all defaults (as opposed to choosing a possible different number close to 1 for each default). I return to this issue again shortly.

Let $\Sigma^r = \{\theta^r : \theta \in \Sigma\}$. The following theorem, whose proof is beyond the scope of this book, captures the connection between the random-worlds approach and the maximum-entropy approach to default reasoning.

Theorem 12.5.1 *Let c be a constant symbol. Then $\Sigma \mid\!\approx^{me} \varphi \rightarrow \psi$ iff*

$$\Pr_\infty(\psi^*(c)|\Sigma^r \wedge \varphi^*(c)) = 1.$$

Note that the translation used in the theorem converts the default rules in Σ to statistical statements about individuals but converts the left-hand and right-hand sides of the conclusion, φ and ψ, to statements about a particular individual (whose name was arbitrarily chosen to be c). This is in keeping with the typical use of default rules. Knowing that birds typically fly, we want to conclude something about a particular bird, Tweety or Opus.

Theorem 12.5.1 can be combined with Theorem 12.3.7 to provide a formal characterization of some of the inheritance properties of $\mid\!\approx^{me}$. For example, it follows that not only does $\mid\!\approx^{me}$ satisfy all the properties of **P**, but that it is able to ignore irrelevant information and to allow subclasses to inherit properties from superclasses, as discussed in Section 7.3.

The assumption that the same approximate equality relation is used for every formula θ^r is crucial in proving the equivalence in Theorem 12.5.1.

For suppose that Σ consists of the two rules $p_1 \land p_2 \to q$ and $p_3 \to \neg q$. Then $\Sigma \not\approx^{me} p_1 \land p_2 \land p_3 \to q$. This seems reasonable, as there is evidence for q ($p_1 \land p_2$) and against q (p_3), and neither piece of evidence is more specific than the other. However, suppose that Σ' is Σ together with the rule $p_1 \to \neg q$. Then it can be shown that $\Sigma' \approx^{me} p_1 \land p_2 \land p_3 \to q$. This behavior seems counterintuitive, and is a consequence of the use of the same ϵ for all the rules. Intuitively, what is occurring here is that prior to the addition of the rule $p_1 \to \neg q$, the sets $P_1(x) \land P_2(x)$ and $P_3(x)$ are of comparable size. The new rule forces $P_1(x) \land P_2(x)$ to be a factor of ϵ smaller than $P_1(x)$, since almost all P_1's are $\neg Q$'s, whereas almost all $P_1 \land P_2$'s are Q's. The size of the set $P_3(x)$, on the other hand, is unaffected. Hence, the default for the ϵ-smaller class $P_1 \land P_2$ now takes precedence over the class P_2.

If different approximate equality relations are used for each default rule, each one corresponding to a different ϵ, then this conclusion no longer follows. An appropriate choice of τ_i can make the default $\|\neg Q(x)|P_3(x)\|_x \approx_i 1$ so strong that the number of Q's in the set $P_3(x)$, and hence the number of Q's in the subset $P_1(x) \land P_2(x) \land P_3(x)$, is much smaller than the size of the set $P_1(x) \land P_2(x) \land P_3(x)$. In this case, the rule $p_3 \to \neg q$ takes precedence over the rule $p_1 \land p_2 \to q$. More generally, with no specific information about the relative strengths of the defaults, the limit in the random-worlds approach does not exist, so no conclusions can be drawn, just as in Example 12.3.9. On the other hand, if all the approximate equality relations are known to be the same, the random-world approach will conclude $Q(c)$, just as the maximum-entropy approach of Section 7.3. This example shows how the added expressive power of allowing different approximate equality relations can play a crucial role in default reasoning.

It is worth stressing that, although this section shows that there is a deep connection between the random-worlds approach and the maximum-entropy approach, this connection holds only if the vocabulary is restricted to unary predicates and constants. The random-worlds approach makes perfect sense (and the theorems proved in Sections 12.3 and 12.4 apply) to arbitrary vocabularies. However, there seems to be no obvious way to relate random worlds to maximum entropy once there is even a single binary predicate in the vocabulary. Indeed, there seems to be no way of even converting formulas in a knowledge base that involves binary predicates to constraints on probability measures so that maximum entropy can be applied.

12.6 Problems With the Random-Worlds Approach

The previous sections have shown that the random-worlds approach has many desirable properties. This section presents the flip side, and shows that the random-worlds approach also suffers from some serious problems. I focus on two of them here: *representation dependence* and *learning*.

Suppose that the only predicate in our language is *White* and *KB* is *true*. Then $\Pr_\infty(\mathit{White}(c)|\mathit{KB}) = 1/2$. On the other hand, if $\neg\mathit{White}$ is refined by adding *Red* and *Blue* to the vocabulary and KB' asserts that $\neg\mathit{White}$ is the disjoint union of *Red* and *Blue* (i.e., KB' is $\forall x((\neg\mathit{White}(x) \Leftrightarrow (\mathit{Red}(x) \lor \mathit{Blue}(x)) \land \neg(\mathit{Red}(x) \land \mathit{Blue}(x)))$, then it is not hard to show that $\Pr_\infty(\mathit{White}(c)|\mathit{KB}') = 1/3$ (Exercise 12.16). The fact that simply expanding the language and giving a definition of an old notion ($\neg\mathit{White}$) in terms of the new notions (*Red* and *Blue*) can affect the degree of belief seems to be a serious problem.

This kind of representation dependence seems to be unavoidable in probabilistic reasoning if we want to obtain conclusions beyond logical consequences. How bad is that? In some cases, lack of representation dependence may indicate something about our knowledge base. For example, suppose that we know that only about half of birds can fly, Tweety is a bird, and Opus is some other individual (who may or may not be a bird). One obvious way to represent this information is to have a language with predicates *Bird* and *Flies*, and take the knowledge base *KB* to consist of the statements $\|\mathit{Flies}(x)|\mathit{Bird}(x)\|_x \approx_1 .5$ and $\mathit{Bird}(\mathit{Tweety})$. It is easy to see that $\Pr_\infty(\mathit{Flies}(\mathit{Tweety})|\mathit{KB}) = .5$ and $\Pr_\infty(\mathit{Bird}(\mathit{Opus})|\mathit{KB}) = .5$. But suppose that instead we use a vocabulary with predicates *Bird* and *FlyingBird*. Let KB' consist of the statements $\|\mathit{FlyingBird}(x)|\mathit{Bird}(x)\|_x \approx_2 .5$, $\mathit{Bird}(\mathit{Tweety})$, and $\forall x(\mathit{FlyingBird}(x) \Rightarrow \mathit{Bird}(x))$. KB' seems to be expressing the same information as *KB*. But $\Pr_\infty(\mathit{FlyingBird}(\mathit{Tweety})|\mathit{KB}') = .5$ and $\Pr_\infty(\mathit{Bird}(\mathit{Opus})|\mathit{KB}') = 2/3$. The degree of belief that Tweety flies is .5 in both cases, although the degree of belief that Opus is a bird changes. Arguably, the fact that our degree of belief that Opus is a bird is language dependent is a direct reflection of the fact that the knowledge base does not contain sufficient information to assign it a single "justified" value. This suggests that it would be useful to characterize those queries that are language independent, while recognizing that not all queries will be.

In any case, in general, it seems that the best that we can do is to accept representation dependence and, indeed, declare that it is (at times) justified. The choice of an appropriate vocabulary is indeed a significant one, which does encode some of the information at our disposal. In the example with colors, the choice of vocabulary can be viewed as reflecting the bias of the reasoner with respect to the partition of the world into colors. Researchers in machine learning and the philosophy of induction have long realized that bias is an inevitable component of effective inductive reasoning. So we should not be completely surprised if it turns out that the related problem of finding degrees of belief should also depend on the bias. Of course, if this is the case, then we would hope to have a good intuitive understanding of how the degrees of belief depend on the bias. In particular, we would like to give the knowledge base designer some guidelines to selecting the "appropriate" representation. Unfortunately, such guidelines do not exist (for random worlds or any other approach) to the best of my knowledge.

To understand the problem of learning, note that so far I have taken the knowledge base as given. But how does an agent come to "know" the information in the knowledge base? For some assertions, like "Tom has red hair", it seems reasonable that the knowledge comes from direct perceptions, which agents typically accept as reliable. But under what circumstances should a statement such as $\|\mathit{Flies}(x)|\mathit{Bird}(x)\|_x \approx_i .9$ be included in a knowledge base? Although I have viewed statistical assertions as objective statements about the world, it is unrealistic to suppose that anyone could examine all the birds in the world and count how many of them fly. In practice, it seems that this statistical statement would appear in KB if someone inspects a (presumably large) sample of birds and about 90% of the birds in this sample fly. Then a leap is made: the sample is assumed to be typical, and the statistics in the sample are taken to be representative of the actual statistics.

Unfortunately, the random-worlds method by itself does not support this leap, at least not if sampling is represented in the most obvious way. Suppose that an agent starts with no information other than that Tweety is a bird. In that case, the agent's degree of belief that Tweety flies according to the random-worlds approach is, not surprisingly, .5. That is, $\Pr_\infty(\mathit{Flies}(\mathit{Tweety})|\mathit{Bird}(\mathit{Tweety})) = .5$ (Exercise 12.17(a)). In the absence of information, this seems quite reasonable. But the agent then starts observing birds. In fact, the agent observes N birds (think of N as large), say $c_1, \ldots, c_N$, and the information regarding which of them fly is recorded in the knowledge base. Let $\mathit{Bird}(\mathit{Tweety}) \land KB'$ be the resulting knowledge base. Thus, KB' has the form

$$\mathit{Bird}(c_1) \land \mathit{Flies}_1(c_1) \land \mathit{Bird}(c_2) \land \neg\mathit{Flies}_2(c_2) \land \ldots \land \mathit{Bird}(c_N) \land \mathit{Flies}_N(c_N),$$

where $\mathit{Flies}_i(c_i)$ is either $\mathit{Flies}(c_i)$ or $\neg\mathit{Flies}(c_i)$. We might hope that if most (say 90%) of the N birds observed by the agent fly, then the agent's belief that Tweety flies increases. Unfortunately, it doesn't; $\Pr_\infty(\mathit{Flies}(\mathit{Tweety}|\mathit{Bird}(\mathit{Tweety}) \land KB') = .5$ (Exercise 12.17(b)).

What if instead the sample is represented using a predicate S? To simplify matters, suppose that we even know that α% of the domain elements were sampled. The fact that 90% of sampled birds fly then becomes $\|Flies(x)|Bird(x) \wedge S(x)\|_x \approx_1 .9$. This helps, but not much. If KB'' consists of this fact, $||S(x)||_x \approx \alpha$, and $Bird(Tweety)$, we might hope that $\Pr_\infty(Flies(Tweety)|KB'') = .9$, but it is not. In fact, $\Pr_\infty(Flies(Tweety)|KB'') = .9\alpha + .5(1-\alpha)$ (Exercise 12.17(c)). Random worlds treats the birds in S and those outside S as two unrelated populations; it maintains the default degree of belief (1/2) that a bird not in S will fly. (This follows using from maximum entropy considerations, along the lines discussed in Section 12.5.) Intuitively, random worlds is not treating S as a *random* sample.

Of course, the failure of the obvious approach does not imply that random worlds is incapable of learning statistics. Perhaps another representation can be found that will do better (although none has been found yet).

So where does this leave us? The random-worlds approach has many attractive features, but has some serious flaws as well. There are variants of the approach that deal well with some of the problems, but not with others. (See, for example, Exercise 12.18.) Perhaps the best lesson that can be derived from this discussion is that it may be impossible to come up with a generic method for ascribing degrees of belief from statistical information that does the "right" thing in all possible circumstances. There is no escape from the need to understand the details of the application.

Exercises

12.1 Show that if both $\mathcal{T}$ and $\mathcal{T}'$ contain all the symbols that appear in φ and KB, then

$$\frac{\#worlds_N^{\mathcal{T},\vec{\tau}}(\varphi \wedge KB)}{\#worlds_N^{\vec{\tau}}(KB)} = \frac{\#worlds_N^{\mathcal{T}',\vec{\tau}}(\varphi \wedge KB)}{\#worlds_N^{\vec{\tau}}(KB)}.$$

12.2 Let $\varphi^=$ be the result of replacing all instances of approximate equality ($\approx_i$, for any i) in φ by equality ($=$). Show that

$$\lim_{\vec{\tau}\to\vec{0}} \mu_N^{\vec{\tau}}(\varphi|KB) = \mu_N^{\vec{\tau}}(\varphi^=|KB^=).$$

Thus, if the order of the limits in the definition of $\Pr_\infty(\varphi|KB)$ were reversed, then all the advantages of using approximate equality would be lost.

12.3 Show that if $a_0, a_1, \ldots$ is a sequence of real numbers bounded from below then $\liminf_{n\to\infty} a_n$ exists. Similarly, show that if $a_0, a_1, \ldots$ is bounded from above then $\limsup_{n\to\infty} a_n$ exists.

12.4 Show that, in the proof of Theorem 12.3.2, $\mu_N^{\vec{\tau}}(\varphi|KB) = \sum_{W'} f_{W'}|W'|/\#worlds_N^{\vec{\tau}}(KB)$, where the sum is taken over all clusters W'.

* **12.5** Theorem 12.3.2 can be generalized in several ways. In particular,

(a) it can be applied to more than one individual at a time,

(b) it applies if there are bounds on statistical information, not just in the case where the statistical information is approximately precise, and

(c) the statistical information does not actually have to be in the knowledge base; it just needs to be a logical consequence of it for sufficiently small tolerance vectors.

To make this precise, let $X = \{x_1, \ldots, x_k\}$ and $C = \{c_1, \ldots, c_k\}$ be sets of distinct variables and distinct constants, respectively. I write $\varphi(X)$ to indicate that all of the free variables in the formula φ are in X; $\varphi(C)$ denotes the new formula obtained by replacing each occurences of x_i in φ by c_i. (Note that φ may contain other constants not among the c_i's; these are unaffected by the substitution.) Prove the following generalization of Theorem 12.3.2:

> Let KB be a knowledge base of the form $\psi(C) \wedge KB'$ and assume that, for all sufficiently small tolerance vectors $\vec{\tau}$,
>
> $$\mathcal{M}^{\approx} \models KB^{\vec{\tau}} \Rightarrow \alpha \leq \|\varphi^{\vec{\tau}}(X)|\psi^{\vec{\tau}}(X)\|_X \leq \beta.$$
>
> If no constant in C appears in KB', in $\varphi(X)$, or in $\psi(X)$, then $\Pr_\infty(\varphi(C)|KB) \in [\alpha, \beta]$, provided the degree of belief exists.

(Note that the degree of belief may not exist since $\lim_{\vec{\tau}\to\vec{0}} \liminf_{N\to\infty} \mu_N^{\vec{\tau}}(\varphi|KB)$ may not be equal to $\lim_{\vec{\tau}\to\vec{0}} \limsup_{N\to\infty} \mu_N^{\vec{\tau}}(\varphi|KB)$. However, it follows from the proof of the theorem that both of these limits lie in the interval $[\alpha, \beta]$. This is why the limit does exist if $\alpha = \beta$, as in Theorem 12.3.2.)

* **12.6** Prove Theorem 12.3.7. (Hint: for each domain size N and tolerance vector $\vec{\tau}$, partition the worlds of size N satisfying $KB^{\vec{\tau}}$ into clusters, where each cluster W' is a maximal set satisfying the following four conditions:

(a) All worlds in W' agree on the denotation of every vocabulary symbol except possibly those appearing in $\varphi(x)$ (so that, in particular, they agree on the denotation of the constant c).

(b) All worlds in W' also agree as to which elements satisfy $\psi_0(x)$; let this set be A_0.

(c) The denotation of symbols in φ must also be constant, except possibly when a member of A_0 is involved. More precisely, let $\overline{A_0}$ be the set of domain elements $\{1, \ldots, N\} - A_0$. Then for any predicate symbol R or function symbol f of arity r appearing in $\varphi(x)$, and for all worlds $w, w' \in W'$, if $d_1, \ldots, d_r, d_{r+1} \in \overline{A_0}$ then $R(d_1, \ldots, d_r)$ holds in w iff it holds in w, and $f(d_1, \ldots, d_r) = d_{r+1}$ in w iff $f(d_1, \ldots, d_r) = d_{r+1}$ in w'. In particular, this means that for any constant symbol c' appearing in $\varphi(x)$, if it denotes $d' \in \overline{A_0}$ in w, then it must denote d' in w'.

(d) All worlds in the cluster are isomorphic with respect to the vocabulary symbols in φ. (That is, if w and w' are two worlds in the cluster, then there is a bijection on $\{1, \ldots, n\}$ such that for each vocabulary symbol P in φ, P^w is isomorphic to $P^{w'}$ under f. For example, if P is a constant symbol d, then $f(d^w) = d^{w'}$; similarly, if P is a binary predicate, the $(d, d') \in P^w$ iff $(f(d), f(d')) \in P^{w'}$.)

Then show that, within each cluster W', the probability of $\varphi(c)$ is within τ_i of α.)

12.7 First-order logic can express not only that there exists an individual that satisfies the formula $\varphi(x)$, but that there exists a unique individual that satisfies $\varphi(x)$. Let $\exists! x \varphi(x)$ be an abbreviation for

$$\exists x \varphi(x) \land \forall y (\varphi(y) \Rightarrow y = x).$$

(a) Show that $(\mathcal{A}, V) \models \exists! x \varphi(x)$ iff there is a unique $d \in \mathrm{dom}(\mathcal{A})$ such that $(\mathcal{A}, V[x/d]) \models \varphi(x)$.

(b) Generalize this to find a formula $\exists^N \varphi(x)$ that expresses the fact that exactly N individuals in the domain satisfy φ.

* **12.8** Prove Theorem 12.3.8. (Hint: suppose that $\alpha_1, \alpha_2 > 0$. Consider $\vec{\tau}$ such that $\alpha_i - \tau_i > 0$. Let $\beta_i = \min(\alpha_i + \tau_i, 1)$. For each domain size N, partition the worlds of size N satisfying $KB^{\vec{\tau}}$ into clusters where each cluster W' is a maximal set satisfying the following three conditions:

(a) All worlds in W' agree on the denotation of every vocabulary symbol except for P. In particular, they agree on the denotations of c, ψ_1, and ψ_2. Let A_i be the denotation of ψ_i in W' (that is, $A_i = \{d \in D : w \models \psi(d)\}$ for $w \in W'$) and let $n_i = |A_i|$).

(b) All worlds in W' have the same denotation of P for elements in $\overline{A} = \{1, \ldots, N\} - (A_1 \cup A_2)$.

(c) For $i = 1, 2$, all worlds in W' must have the same number of elements r_i satisfying P within each set A_i.

Note that, since all worlds in W' satisfy $KB^{\vec{\tau}}$, it follows that $\beta_i = r_i/n_i \in [\alpha_i - \tau_i, \alpha_i + \tau_i]$ for $i = 1, 2$. Show that the number of worlds in W' satisfying $P(c)$ is $\binom{n_1-1}{r_1-1}\binom{n_2-1}{r_2-1}$ and the number of worlds satisfying $\neg P(c)$ is $\binom{n_1-1}{r_1}\binom{n_2-1}{r_2}$. Conclude from this that the fraction of worlds satisfying $P(c)$ is $\frac{\beta_1\beta_2}{\beta_1\beta_2+(1-\beta_1)(1-\beta_2)}$.)

* **12.9** State and prove a generalized version of Theorem 12.3.8 that allows more than two pieces of statistical information.

12.10 Complete the proof of Theorem 12.4.3.

* **12.11** This exercise considers to what extent *Rational Monotonicity* holds in the random-worlds approach. Recall that rational monotonicity is characterized by axiom C5 in AX^{cond} (see Section 7.4). Roughly speaking, it holds if the underlying likelihood measure is totally ordered. Since probability is totally ordered, it would seem that something like Rational Monotonicity should hold for the random-worlds approach, and indeed it does. Rational Monotonicity in the random worlds framework becomes:

RM. If $KB \mathrel{|\!\sim}_{rw} \varphi$ and $KB \mathrel{|\!\not\sim}_{rw} \neg\theta$, then $KB \wedge \theta \mathrel{|\!\sim}_{rw} \varphi$.

Show that the random-worlds approach satisfies the following weakened form of RM: if $KB \mathrel{|\!\sim}_{rw} \varphi$ and $KB \mathrel{|\!\not\sim}_{rw} \neg\theta$, then $KB \wedge \theta \mathrel{|\!\sim}_{rw} \varphi$ provided that $\Pr_\infty(\varphi|KB \wedge \theta)$ exists. Moreover, a sufficient condition for $\Pr_\infty(\varphi|KB \wedge \theta)$ to exist is that $\Pr_\infty(\theta|KB)$ exists.

12.12 The CUT property was introduced in Exercise 7.2 and shown to follow from **P**. In the setting of this chapter, CUT becomes

CUT. If $KB \mathrel{|\!\sim}_{rw} \varphi$ and $KB \wedge \varphi \mathrel{|\!\sim}_{rw} \psi$ then $KB \mathrel{|\!\sim}_{rw} \varphi$.

Show directly that CUT holds in the random-worlds approach. In fact, show that the following stronger result holds: If $\Pr_\infty(\varphi|KB) = 1$, then $\Pr_\infty(\psi|KB) = \Pr_\infty(\psi|KB \wedge \varphi)$ (where equality here means that either neither limit exists or both do and are equal).

* **12.13** (a) Show that $KB'_{lottery} \mathrel{|\!\!\sim}_{rw} \neg \mathit{Winner}(c)$, where $KB'_{lottery}$ is defined in Example 12.4.6. (Hint: fix a domain size N. Cluster the worlds according to the number of ticket holders. That is, let W_k consist of all worlds with exactly k ticket holders. Observe that $|W_k| = k\binom{N}{k}$ (since the winner must be one of the k ticket holders). Show that the fraction of worlds in W_k in which c wins is $1/k$. Next, observe that

$$|\cup_{k \leq N/4} W_k| = \sum_{k=1}^{N/4} k\binom{N}{k} \leq (N/4)\sum_{k=1}^{N/4}\binom{N}{N/4} = (N/4)^2\binom{N}{N/4}.$$

Similarly

$$|\cup_{k > N/4} W_k| = \sum_{k=N/4+1}^{N} k\binom{N}{k} > (N/2)\binom{N}{N/2}$$

(since $(N/2)\binom{N}{N/2}$ is just one term in the sum). Show that

$$\lim_{N\to\infty} \frac{(N/4)^2\binom{N}{N/4}}{(N/2)\binom{N}{N/2}} = 0;$$

that is, for N sufficiently large, in almost all worlds there are at least $N/4$ ticket holders. The desired result now follows easily.)

(b) Show that $\Pr_\infty(\mathit{Winner}(c) | KB''_{lottery}) = 1/N$. (This actually follows easily from the first part of the analysis of part (a).)

12.14 Show that the random-worlds approach takes different constants to denote different individuals, by default. That is, show that if c and d are distinct constants, then $\mathit{true} \mathrel{|\!\!\sim}_{rw} c \neq d$. The assumption that different individuals are distinct has been called the *unique names assumption* in the literature. This shows that the unique names assumption holds by default in the random-worlds approach.

* **12.15** Consider a vocabulary $\mathcal{T}$ consisting of k unary predicates $P_1, \ldots, P_k$. Let $\vec{p} = (N_1/N, \ldots, N_n/N)$.

(a) Show that there are

$$\binom{N}{N_1, \ldots, N_k} = \frac{N!}{N_1!N_2!\ldots N_k!}$$

D_N-$\mathcal{T}$ structures $\mathcal{A}$ such that such that there are N_i domain elements satisfying P_i (i.e., $|P_i^{\mathcal{A}}| = N_i$).

(b) *Stirling's approximation* says that

$$m! = \sqrt{2\pi m}\, m^m e^{-m}(1 + O(1/m)).$$

Using Stirling's approximation show that there exist constants L and U such that

$$\frac{L}{U^K N^K}\frac{N^N \prod_{i=1}^K e^{N_i}}{e^N \prod_{i=1}^K N_i^{N_i}} \le \frac{N!}{N_1!N_2!\dots N_K!} \le \frac{UN}{L^K}\frac{N^N \prod_{i=1}^K e^{N_i}}{e^N \prod_{i=1}^K N_i^{N_i}}.$$

(c) Let $\vec{p} = (p_1, \dots, p_n)$, where $p_i = N_i/N$. Show that

$$\frac{N^N \prod_{i=1}^K e^{N_i}}{e^N \prod_{i=1}^K N_i^{N_i}} = e^{-N\sum_{i=1}^K u_i \ln(u_i)} = e^{NH(\vec{p})}.$$

(d) Conclude from (c) and (d) that

$$\frac{h(N)L}{U^K N^K} e^{NH(\vec{p})} \le |\{w \in W_N : \vec{p}^{\,w} = \vec{p}\} \le N^{|\mathcal{C}|} h(N)\frac{UN}{L^K} e^{NH(\vec{p})}.$$

12.16 Show that $\Pr_\infty(\mathit{White}(c)|\mathit{true}) = .5$ and that

$$\Pr_\infty(\mathit{White}(c)|\forall x(\neg\mathit{White}(x) \Leftrightarrow ((\mathit{Red}(x)\vee\mathit{Blue}(x))\wedge\neg(\mathit{Red}(x)\wedge\mathit{Blue}(x))))) = 1/3.$$

12.17 (a) Show that $\Pr_\infty(\mathit{Flies}(\mathit{Tweety})|\mathit{Bird}(\mathit{Tweety})) = .5$.

(b) Show that $\Pr_\infty(\mathit{Flies}(\mathit{Tweety})|\mathit{Bird}(\mathit{Tweety}) \wedge \mathit{KB}') = .5$ if KB' has the form

$$\mathit{Flies}_1(c_1) \wedge \mathit{Bird}(c_2) \wedge \mathit{Flies}_2(c_2) \wedge \dots \wedge \mathit{Bird}(c_N) \wedge \mathit{Flies}_N(c_N),$$

where Flies_i is either Flies or $\neg\mathit{Flies}$ for $i = 1, \dots, N$.

(c) Show that if

$$\mathit{KB}'' = \|\mathit{Flies}(x)|\mathit{Bird}(x) \wedge S(x)\|_x \approx_1 .9 \wedge ||S(x)||_x \approx \alpha \wedge \mathit{Bird}(\mathit{Tweety})$$

then $\Pr_\infty(\mathit{Flies}(\mathit{Tweety})|\mathit{KB}'') = .9\alpha + .5(1-\alpha)$.

12.18 Suppose that $\mathcal{T} = \{P_1, \dots, P_m, c_1, \dots, c_n\}$. That is, the vocabulary consists only of unary predicates and constants. Fix a domain size N. For each tuple $(k_1, \dots, k_m)$ such that $0 \le k_i \le N$, let $W_{(k_1,\dots,k_m)}$ consist of all structures $\mathcal{A} \in W_N$ such that $|P_i^{\mathcal{A}}| = k_i$, for $i = 1, \dots, N$. Note that there are m^{N+1} sets of the form $W_{(k_1,\dots,k_m)}$. Let μ_N be the probability measure on W_N such that $\mu_N(W_{(k_1,\dots,k_m)}) = 1/m^{N+1}$ and all the worlds in $W_{(k_1,\dots,k_m)}$ are equally likely.

(a) Let $\mathcal{A} \in W_N$ be such that $|P_i^{\mathcal{A}}| = 0$ (that is, no individual satisfies any of $P_1, \ldots, P_N$ in $\mathcal{A}$). What is $\mu_N(\mathcal{A})$?

(b) Assume that N is even and let $\mathcal{A} \in W_N$ be such that $|P_i^{\mathcal{A}}| = N/2$ for $i = 1, \ldots, N$. What is $\mu_N(\mathcal{A})$?

You should get different answers for (a) and (b). Intuitively, μ_N does not make all worlds in W_N, but it does make each possible cardinality of $P_1, \ldots, P_N$ equally likely. For $\varphi, KB \in \mathcal{L}^{fo}(\mathcal{T})$, define $\Pr'_\infty(\varphi|KB)$ to be the common limit of

$$\lim_{\vec{\tau} \to \vec{0}} \liminf_{N \to \infty} \mu_N(\varphi|KB) \quad \text{and} \quad \lim_{\vec{\tau} \to \vec{0}} \limsup_{N \to \infty} \mu_N(\varphi|KB),$$

if the limit exists. $\Pr'_\infty(\varphi|KB)$ gives a different way of obtaining degrees of belief from statistical information.

(c) Show that the following simplified version of Theorem 12.3.2 holds for $\Pr'_\infty$:

$$\Pr'_\infty(\varphi(c) | \|\varphi(x)|\psi(x)\|_x \approx_i \alpha \wedge \psi(c)) = \alpha.$$

Actually, the general version of Theorem 12.3.2 also holds. Moreover, learning from samples works for $\Pr'_\infty$:

$$\Pr'_\infty(\mathit{Flies}(\mathit{Tweety}) | \mathit{Bird}(\mathit{Tweety}) \wedge \|\mathit{Flies}(x)|\mathit{Bird}(x) \wedge S(x)\|_x \approx_i .9) = .9,$$

although the proof of this (which requires maximum entropy techniques) is beyond the scope of the book.

Notes

The earliest sophisticated attempt at clarifying the connection between objective statistical knowledge and degrees of belief, and the basis for most subsequent proposals involving reference classes, is due to Reichenbach [1949]. A great deal of further work has been done on reference classes, perhaps most notably by Kyburg [1983, 1974] and Pollock [1990]; this work mainly elaborates the way in which the reference class should be chosen in case there are competing reference classes.

The random-worlds approach is due to Bacchus, Grove, Koller and me [1996]. However, the key ideas in the approach are not new. Many of them can be found in the work of Johnson [1932] and Carnap [1950, 1952], although these authors focus on knowledge bases that contain only first-order information, and for the most part restrict their attention to unary predicates. More recently, Chuaqui [1991] and Shastri [1989] approaches have presented similar in spirit to the random-worlds approach.

Much of the discussion in this chapter is taken from [Bacchus, Grove, Halpern, and Koller 1996]. Stronger versions of Theorems 12.3.2, 12.3.7, and 12.3.8 are proved in the paper (cf. Exercises 12.5 and 12.9). More discussion of dealing with approximate equality can be found in [Koller and Halpern 1992].

Example 12.3.9 is taken from [Reiter and Criscuolo 1981]; it is called the *Nixon Diamond* and is one of the best-known examples in the default reasoning literature of the difficulties in dealing with conflicting information. Example 12.4.4 is due to Poole [1989]; he presents it as an example of problems that arise in Reiter's [1980] *default logic*, which would conclude that both arms are usable.

The connections to maximum entropy discussed in Section 12.5 are explored in more detail in [Grove, Halpern, and Koller 1994], where Theorem 12.5.1 is proved. This paper also provides further discussion of the relationship between maximum entropy and random-worlds (and why this relationship breaks down when there are nonunary predicates in the vocabulary). Paris and Venkovska [1989, 1992] use an approach based on maximum entropy to deal with reasoning about uncertainty, although they work at the propositional level. The observation that the maximum-entropy to default reasoning in Section 7.3 leads to some anomalous conclusions as a result of using the same ϵ for all rules is due to Geffner [1992b]. Geffner presents another approach to default reasoning that seems to result in the same conclusions as the random-worlds translation of the maximum-entropy approach if different approximate equality relations are used, however the exact relationship between the two approaches is as yet unknown. Stirling's approximation to $m!$ (which is used in Exercise 12.15) is proved in [Graham, Knuth, and Patashnik 1989].

Problems with the random-worlds approach (including ones not mentioned here) are discussed in [Bacchus, Grove, Halpern, and Koller 1996]. Because of the connection between random worlds and maximum entropy, random worlds inherits some well-known problems of the maximum-entropy approach, such as representation dependence. On the other hand, in [Halpern and Koller 1995] a definition of representation independence in the context of probabilistic reasoning is provided; it is shown that essentially every interesting non-deductive inference procedure cannot be representation independent in the sense of this definition. Thus the problem is not unique to maximum entropy (or random worlds).

A number of variants of the random-worlds approach are presented in

[Bacchus, Grove, Halpern, and Koller 1992]; each of them has its own problems and features. The one presented in Exercise 12.18 is called the *random-propensities* approach. It does allow some learning, at least as long as the vocabulary is restricted to unary predicates. In that case, as shown in [Koller and Halpern 1996], it satisfies analogues of Theorem 12.3.2 and 12.3.7. However, the random-propensities method does not extend too well to non-unary predicates. æ