TECHNICAL REPORT
R-144

A Maximum Entropy Approach to Nonmonotonic Reasoning*

Moisés Goldszmidt
moises@cs.ucla.edu
Cognitive Systems Lab.
University of California
Los Angeles, CA 90024

Paul Morris
morris@intellicorp.com
Intellicorp
1975 El Camino Real West
Mountain View, CA 94040

Judea Pearl
judea@cs.ucla.edu
Cognitive Systems Lab.
University of California
Los Angeles, CA 90024

Abstract

This paper describes a probabilistic approach to nonmonotonic reasoning which combines the principle of infinitesimal probabilities with that of maximum entropy, and which sanctions inferences similar to those produced by the principle of minimizing abnormalities. The paper provides a precise formalization of the consequences entailed by a defeasible knowledge base, develops the computational machinery necessary for deriving these consequences, and compares the behavior of the maximum entropy approach to those of ε-semantics ([Pearl 89a]) and rational closure ([Lehmann 89]).

1 Introduction

An approach to defeasible reasoning based on probabilities arbitrarily close to 1 (see [Geffner & Pearl 88], [Pearl 88]) produces a semi-monotonic logic that properly respects specificity-based preferences but often appears to be too conservative. This system, called ε-semantics, was proposed as a common core for all nonmonotonic formalisms, but, in itself, turns out too weak to capture many aspects of common sense reasoning such as chaining, contraposition, and respecting irrelevancies ([Pearl 89a]). Pearl has proposed to increase the inferential power of ε-semantics using the independence assumptions embedded in distributions of maximum entropy, and has shown that when applied to knowledge bases containing a small number of rules, maximum entropy yields patterns of reasoning which are rather pervasive in common discourse (see [Pearl 88] chapter 10).

This paper explores a system based on infinitesimal probabilities augmented by maximum entropy considerations. Given a set $\mathcal{R}$ of rules and a set $\mathcal{P}_{\varepsilon,\mathcal{R}}$ of probability distributions that satisfy each of the rules in $\mathcal{R}$ to within ε, we single out a distinguished distribution $P^*_{\varepsilon,\mathcal{R}}$ having the greatest entropy: $-\sum_\omega P(\omega)\log P(\omega)$. We then define the notion of a *plausible conclusion* of $\mathcal{R}$ in terms of a collection of such distributions, parameterized by ε.

This system is related to those based on minimizing abnormalities (e.g. circumscription [McCarthy 86]), in that inferences are sanctioned if they hold in a model that minimizes a weighted count of rule violation. We discuss the computational and behavioral aspects of the ME approach, indicating improvements over ε-semantics and the rational closure of [Lehmann 89].

The paper is organized as follows: Section 2 introduces the language and some basic definitions and theorems. Section 3 is concerned with the formalism of parameterized probability distributions (PPDs). In section 4 the necessary machinery is developed for computing the maximum entropy distribution and deciding whether an arbitrary conditional sentence is a plausible conclusion of a given knowledge base. Section 5 provides a summary and examples, and Section 6 evaluates the main results. Proofs to all theorems can be found in the full paper [Goldszmidt, Morris & Pearl 90].

2 Notation and Preliminaries.

Let $\mathcal{L}$ be a closed set of well formed propositional formulas, built in the usual way from a *finite* set of propositional variables and the connectives "$\vee$" and "$\neg$". The letters A, B, C, D will be used to denote formulas in $\mathcal{L}$.

A world w is an assignment of truth values to the propositional variables in $\mathcal{L}$. The satisfaction of a formula by a world is defined as usual, and will be written as $w \models A$. Note that if there are n propositional variables in $\mathcal{L}$ there will be 2^n worlds. Let $\mathcal{U}$ stand for the set of worlds.

A more complete treatment of the concepts summarized below can be found in [Goldszmidt & Pearl 89].

Using the binary connective "$\rightarrow$" and two formulas A and B from $\mathcal{L}$ we can construct the defeasible rule $A \rightarrow B$. We will use $\mathcal{R}$ to denote a set of such rules. A

*The participation of the second author was funded by a joint project of NASA and DARPA under contract F30602-88-C-0045. The work of the first and third authors was supported in part by National Science Foundation grant #IRI-88-21444 and Naval Research Laboratory grant #N00014-89-J-2007.

rule $A \rightarrow B$ is said to be *verified* by w, if $w \models A \wedge B$. The same rule is said to be *falsified* or *violated* by w, if $w \models A \wedge \neg B$. If $w \not\models A$, the rule is considered as neither verified nor falsified.

A rule r is *tolerated* by a set $\mathcal{R}$ if we can find a world w that verifies r while no other sentence in $\mathcal{R}$ is falsified by w. We will say that a non-empty set $\mathcal{R}$ of rules is *confirmable* if we can find a rule $r \in \mathcal{R}$ that is tolerated by $\mathcal{R}$.

Given a positive real number ε, we will say that a probability measure P ε-satisfies the rule $A \rightarrow B$ if $P(B|A) \geq 1 - \varepsilon$. Given a set $\mathcal{R}$ of rules, we will use $\mathcal{P}_{\varepsilon,\mathcal{R}}$ to denote the set of probability distributions that ε-satisfy $\mathcal{R}$. We will say that a probabilty measure P is *proper* for $\mathcal{R}$, if $P(A) > 0$ for all A such that $A \rightarrow B \in \mathcal{R}$. A rule will be considered proper if its antecedent is satisfiable.

A set $\mathcal{R}$ is *probabilistically consistent* if, for every $\varepsilon > 0$, there is a proper probability assignment P such that P ε-satisfies every rule $A \rightarrow B \in \mathcal{R}$. Intuitively, consistency means that it is possible for all rules to be as close to absolute certainty as desired. Alternatively, it means that $\mathcal{P}_{\varepsilon,\mathcal{R}}$ is nonempty for all $\varepsilon > 0$, and hence, the existence of $P^*_{\varepsilon,\mathcal{R}}$ is guaranteed for consistent rule sets. Moreover, if $\mathcal{P}_{\varepsilon,\mathcal{R}}$ is a convex set, $P^*_{\varepsilon,\mathcal{R}}$ will be unique. The next theorem constitutes the basis of a simple procedure for testing consistency:

Theorem 1 (Consistency.) [1] *A set $\mathcal{R}$ is* probabilistically consistent *if and only if every nonempty subset $\mathcal{R}'$ of $\mathcal{R}$ is confirmable.*

In other words, $\mathcal{R}$ is consistent iff we can find a sentence tolerated by $\mathcal{R}'$, in every subset $\mathcal{R}'$ of $\mathcal{R}$.

Corollary 1 ([Goldszmidt & Pearl 89].) *Given a set $\mathcal{R}$, consistency can be tested in $|\mathcal{R}|^2/2$ satisfiability tests by the following simple labeling procedure: construct the set $\mathcal{R}_1$ with all rules tolerated by $\mathcal{R}$, then construct the set $\mathcal{R}_2$ with all rules tolerated by $\mathcal{R} - \mathcal{R}_1$ and so on. If a partition of $\mathcal{R}$ is obtained the set is consistent; otherwise $\mathcal{R}$ is inconsistent.*

Although propositional satisfiability is in general NP-complete, for the case of Horn clauses it is linear on the number of occurrences of literals in $\mathcal{R}$ [Dowling & Gallier 84].

3 Parameterized Probability Distributions

Among the general *laws* that a common sense consequence relation (denoted by $\mid\!\sim$)[2] might be expected to obey, the following have been proposed ([Geffner & Pearl 88], [Kraus et.al. 88], [Makinson 89], [Pearl 89a]):

(Logic) If $P \vdash Q$, then $P \mid\!\sim Q$.
(Cumulativity) If $P \mid\!\sim Q$, then $P \mid\!\sim R$ iff $P \wedge Q \mid\!\sim R$.
(Cases) If $P \mid\!\sim R$ and $Q \mid\!\sim R$, then $P \vee Q \mid\!\sim R$.

Kraus, Lehmann and Magidor [Kraus et.al. 88] introduce the class of preferential models, and show that each preferential model satisfies the three laws given above. Moreover, they show every consequence relation satisfying those laws can be represented as a preferential model. (Kraus, et. al. actually use a slightly different set of laws, but they are easily shown to be equivalent to those above.) Equivalent results were shown in [Lehmann & Magidor 88] with respect to the class of ranked preferential models and the set of rules above augmented by Rational Monotony: If $P \mid\!\sim R$ and $P \not\mid\!\sim \neg Q$ then $P \wedge Q \mid\!\sim R$.

As it stands, ε-semantics does not quite fit within the same framework as preferential models. The basic notion behind ε-entailment is: Given a set $\mathcal{R}$, a new rule $A \rightarrow B$ is ε-entailed, if for all $\delta > 0$, there exists an $\varepsilon > 0$ such that for all P in $\mathcal{P}_{\varepsilon,\mathcal{R}}$ we have $P(B|A) \geq 1 - \delta$. Thus, ε-semantics defines an entailment relation which is essentially that induced by the *class* of preferential models [Lehmann & Magidor 88], but, it presents no direct counterpart to the notion of an *individual* preferential model. Furthermore, in general, ε-semantics does not satisfy rational monotony. This motivates the following reformulation of the idea of ε-semantics:

Definition 1 *A* parameterized probability distribution *(PPD) is a collection $\{P_\varepsilon\}$ of probability measures over a space of worlds, indexed by a parameter ε that ranges over positive real numbers in a neighborhood of zero.*

Definition 2 *Every parameterized probability distribution $\{P_\varepsilon\}$ induces an* consequence relation *on formulas as follows: $A \mid\!\sim B$ iff* $\lim_{\varepsilon \rightarrow 0} P_\varepsilon(B|A) = 1$.

To avoid having to treat some cases separately in the proofs and definitions, it is convenient for the purposes of this section to define $P(B|A) = 1$ when $P(A) = 0$ (thus extending Definition 2 to non-proper distributions.) Under this convention, a PPD consequence relation can now contain instances of the form $A \mid\!\sim$ *false* even when A is logically consistent (see [Adams 66].)

It is easy to show from elementary probability equivalences that each such consequence relation satisfies the Logic, Cumulativity, and Cases laws discussed earlier. Also, as might be expected, there is a close relation between PPDs and ε-semantics:

Theorem 2 *A proper rule is a consequence of a finite probabilistically consistent set of rules with respect to the class of PPDs iff it is ε-entailed.*

We now identify a subclass of PPDs that is of special interest. We will say a PPD $\{P_\varepsilon\}$ is *convergent* if $P_\varepsilon(B|A)$ converges (as $\varepsilon \rightarrow 0$) for each pair of sentences A and B. The following is an important sufficient condition for PPD convergence. We define a PPD to be

[1] This theorem appears initially in [Adams 75] and is extended to mixtures of defeasible and strict rules in [Goldszmidt & Pearl 89].

[2] We reserve the symbol $\vdash$ for classical derivability

analytic if, for every event E, $P_\varepsilon(E)$ has an extension to a function over complex values of ε that is analytic in a neighborhood of 0. (This implies that it possesses derivatives of all orders, all of which converge as ε approaches 0.)

Theorem 3 *Every analytic PPD is a convergent PPD.*

The proof is a direct consequence of the fact that any given analytic PPD can be expanded as a Taylor series about zero. Either $P_\varepsilon(E)$ is identically zero, or at least one of the coefficients must be non-zero. In the latter case, as ε approaches 0, the series is dominated by the first term whose coefficient is non-zero.

Besides the three laws considered earlier, a convergent PPD consequence relation satisfies Rational Monotony. The following theorem is an easy consequence of the results and methods in [Lehmann & Magidor 88]. A similar result has been independently obtained by Satoh [Satoh 90].

Theorem 4 *Every convergent PPD entailment relation can be represented as a ranked preferential model,*[3] *and every ranked preferential model with a finite non-empty state space can be represented as a convergent PPD entailment relation.*

There is also a connection between PPDs and preferential models: since the entailment relation of a PPD satisfies the laws of Logic, Cumulativity, and Cases, it can be represented as a preferential model [Kraus et.al. 88]. The following result shows that the converse is also true, for finite systems.

Theorem 5 *Every PPD entailment relation may be represented as a preferential model, and every preferential model with finite non-empty state space may be represented as a PPD theory.*

The basic idea in the converse part of the proof is to consider the total order extensions of the partial order that determines the preferential model. Each of these corresponds to a ranked preferential model, which by theorem 4 can be expressed as a convergent PPD. Interleaving the individual PPDs then gives a single PPD that represents the original preferential model.

Preference logics were originally introduced as a generalization of circumscription. One might ask where circumscriptive theories fit in the framework discussed above. The simplest form of circumscription is one that minimizes a single finite abnormality predicate, letting everything else vary. Thus, it is characterized by a preference for worlds that satisfy minimal subsets of a finite set of abnormality propositions. We will call a system of axioms together with such a preference a *finite abnormality model*. Clearly, every such system can be represented as a preferential model. The following is a partial converse.

Theorem 6 *Every preferential model with a finite propositional language and finite state space can be represented as a finite abnormality model.*

4 Maximizing the Entropy.

As mentioned earlier, given any ruleset $\mathcal{R}$, there is a distinguished PPD $\{P^*_{\varepsilon,\mathcal{R}}\}$ where $P^*_{\varepsilon,\mathcal{R}}$ is defined as the distribution of maximum entropy that ε-satisfies each rule in $\mathcal{R}$. This suggests the following definition (we assume proper probability distributions and proper rules):

Definition 3 (ME-plausible conclusions.) *Given a consistent set $\mathcal{R}$, we say that $A \rightarrow B$ is a ME-plausible conclusion of $\mathcal{R}$ iff $lim_{\varepsilon\rightarrow 0} P^*_{\varepsilon,\mathcal{R}}(B|A) = 1$. Equivalently, we say that $A \mid\!\sim B$ is in the ME-consequence relation of $\mathcal{R}$, denoted $C_{ME}(\mathcal{R})$*

While plausible conclusions in ε-semantics are required to attain arbitrarily high probabilities in **all** probability distributions in $\mathcal{P}_{\varepsilon,\mathcal{R}}$, the requirement for ME-plausible conclusions concerns only one distinguished distribution, that having the maximum entropy among those in $\mathcal{P}_{\varepsilon,\mathcal{R}}$. In this section we develop the machinery for deciding whether a rule $A \rightarrow B$ is an ME-plausible conclusion of $\mathcal{R}$ in accordance with the definition above.

Let $\mathcal{R}$ be a set of defeasible rules $r_i : A_i \rightarrow B_i$, $1 \leq i \leq n$. Each of these rules imposes the constraint:

$$P(B_i|A_i) \geq 1 - \varepsilon \tag{1}$$

on the space of distributions. Using elementary probability theory we can rewrite Eq. (1) as:

$$\frac{\varepsilon}{1-\varepsilon} \times P(B_i, A_i) \geq P(\bar{B}_i, A_i) \tag{2}$$

where $\bar{B}_i$ denotes the complement of B_i. Note that the term $P(B_i, A_i)$ equals the sum of the probabilities of the worlds in which $A_i \rightarrow B_i$ is verified and similarly $P(\bar{B}_i, A_i)$ equals the sum of the probabilities of the worlds in which $A_i \rightarrow B_i$ is falsified. Writing $W^+_{r_i}$ as a shorthand for the set of worlds in which r_i is verified, and $W^-_{r_i}$ for the set of worlds that falsify r_i, Eq. (2) can be written as:

$$P(W^-_{r_i}) - \frac{\varepsilon}{1-\varepsilon} \times P(W^+_{r_i}) \leq 0 \tag{3}$$

where $P(W^-_{r_i}) = \sum_{\omega \in W^-_{r_i}} P(\omega)$ and $P(W^+_{r_i}) = \sum_{\omega \in W^+_{r_i}} P(\omega)$.

The entropy associated with a probability distribution P is defined as:

$$H[P] = -\sum_{\omega} P(\omega) \log P(\omega) \tag{4}$$

The problem of computing the maximum entropy distribution $P^*_{\varepsilon,\mathcal{R}}$ reduces to the problem of maximizing the entropy expression (Eq. (4)) subject to the set of constraints Eq. (3) and the normalization constraint $\sum_{\omega} P(\omega) = 1$.

[3] We remark that the proof of theorem 3 shows that for analytic PPDs, the ranks are well-ordered. This stands in contrast to example 4.1 in [Lehmann & Magidor 88].

One of the more powerful techniques for solving such optimization problems is that of Lagrange multipliers [Aoki 71]. This technique associates a factor α with each constraint (rule), and yields a distribution $P^*(\omega)$ that is expressible as a product of these factors ([Cheeseman 83]). We will show that, under the infinitesimal approximation, $P^*(\omega)$ will be proportional to the product of the factors (α) associated only with rules falsified in ω [4].

At the point of maximum entropy, the status of a constraint such as (3) can be one of two types: *active*, when the constraint is satisfied as an equality, and *passive*, when the constraint is satisfied as a strict inequality. Passive constraints do not affect the point of maximum entropy and can be ignored (see [Aoki 71]). The task of identifying the set of *active* constraints is discussed at the end of this section. We will first assume that all constraints are active.

An application of the Lagrange multiplier technique on a set of n active constraints yields the following expression for each term $P(\omega)$ (see the appendix in [Goldszmidt, Morris & Pearl 90] for a step by step derivation):[5]

$$P(w) = \alpha_0 \times \prod_{r_i \in R^-_\omega} \alpha_{r_i} \times \prod_{r_j \in R^+_\omega} \alpha_{r_j}^{(-\frac{\varepsilon}{1-\varepsilon})} \qquad (5)$$

where R^-_ω denotes the set of rules falsified in ω and R^+_ω denotes the set of rules verified in ω. Motivated by Theorem 3, we look for an asymptotic solution where each α_{r_i} is proportional to ε^{κ_i} for some non-negative integer κ_i,[6] namely, each term of the form $\alpha_{r_j}^{(-\frac{\varepsilon}{1-\varepsilon})}$ will tend to 1 as ε tends to 0 . The term α_0 is a normalization constant that will be present in each term of the distribution and thus can be safely ignored. Using P' to denote the unnormalized probability function, and taking the limit as ε goes to 0, equation (5) yields:

$$P'(\omega) \approx \begin{cases} 1 & R^-_\omega = \emptyset \\ \prod_{r_i \in R^-_\omega} \alpha_{r_i} & \text{otherwise} \end{cases} \qquad (6)$$

Thus, the probability of a given world ω depends only on the rules that are falsified in that world. Once the α-factors are computed, we can construct the desired probability distribution and determine which new rules are plausible conclusions of $\mathcal{R}$.

In order to compute the α-factors we substitute the expression for each $P'(\omega)$ (Eq. (6)) in each of the the active constraints equations (Eq. (3)), and obtain:

$$\sum_{\omega \in W^-_{r_i}} [\prod_{r_k \in R^-_\omega} \alpha_{r_k}] = \frac{\varepsilon}{1-\varepsilon} \times \sum_{\omega \in W^+_{r_i}} [\prod_{r_j \in R^-_\omega} \alpha_{r_j}] \qquad (7)$$

where $1 \leq i \leq n$. A few observations are in order: First, Eq. (7) constitutes a system of n equations (one for each active rule) with n unknowns (the α-factors, one for each active rule). Unfortunately, each summation might range over an exponentially large number of worlds. Second, by our assumption, $\alpha_{r_i} \approx a_i \varepsilon^{\kappa_i}$ where κ_i is a nonnegative integer. This implies $\log \alpha_{r_i} \approx \log a_i + \kappa_i \log \varepsilon \approx \kappa_i \log \varepsilon$, and $\frac{\varepsilon}{1-\varepsilon} \approx \varepsilon$. Thus, each probability term $P'(\omega)$ is determined once the values of the κ's are computed (see Eq. (6)). We can rewrite Eq. (7) in terms of the κ's, by replacing the summations in Eq. (7) by the *min* operation since the highest order term (the term with minimum κ) will be the most significant one as ε approaches 0. Taking the *log* on both sides of Eq. (7) yields:

$$\min_{\omega \in W^-_{r_i}} [\sum_{r_k \in R^-_\omega} \kappa_k] = 1 + \min_{\omega \in W^+_{r_i}} [\sum_{r_j \in R^-_\omega} \kappa_j] \quad 1 \leq i \leq n \qquad (8)$$

Each κ_i can be regarded as the cost added to a world ω that violates rule r_i; since such violation causes $\log(P'(\omega))$ to decrease by κ_i.

Since rule r_i is falsified in each world on the left-hand-side of equation (8), κ_i will appear in each one of the $\sum$-terms inside the *min* operation and can be isolated:

$$\kappa_i + \min_{\omega \in W^-_{r_i}} [\sum_{\substack{r_k \in R^-_\omega \\ k \neq i}} \kappa_k] = 1 + \min_{\omega \in W^+_{r_i}} [\sum_{r_j \in R^-_\omega} \kappa_j] \qquad (9)$$

Even with these simplifications, it is not clear how to compute the values for the κ's in the most general case. We now introduce a class of rule sets $\mathcal{R}$ for which a simple greedy strategy can be used to solve the set of equations above:

Definition 4 (Minimal Core Sets.) *A set $\mathcal{R}$ is a* minimal core (MC) set *iff for each rule $r_i : A_i \rightarrow B_i \in \mathcal{R}$, its* negation *$A_i \rightarrow \neg B_i$ is tolerated by $\mathcal{R} - \{r_i\}$. Equivalently, for each rule r_i there is a world that falsifies r_i and none other rule in $\mathcal{R}$.*

Clearly, to decide whether a set $\mathcal{R}$ is an MC set takes $|\mathcal{R}|$ satisfiability tests. Note also that the MC property excludes sets $\mathcal{R}$ that contain redundant rules, namely, rules r that are already ε-entailed by $\mathcal{R} - \{r\}$. This is so because the toleration requirement of MC sets guarantees that the negation of each rule r_i is consistent with respect to the rest of the rules in $\mathcal{R}$ and it is known (see [Goldszmidt & Pearl 89]) that a rule r_i is ε-entailed by $\mathcal{R} - \{r_i\}$ if and only if its negation is inconsistent with $\mathcal{R} - \{r_i\}$. For example, consider the rule set $\mathcal{R}_{scb} = \{s \rightarrow c, s \rightarrow b, s \rightarrow c \wedge b\}$[7]. This set is redundant because the third rule is ε-entailed by the first two, and vice versa. Indeed, $\mathcal{R}_{scb}$ does not meet the MC requirements (Def. 4); any world falsifying $s \rightarrow c$ (or $s \rightarrow b$) must also falsify $s \rightarrow c \wedge b$.

[4] We drop the subscript "$\varepsilon, \mathcal{R}$" for notation clarity.

[5] In equation (5) $\alpha_0 = e^{(\lambda_0+1)}$ and $\alpha_{r_k} = e^{\lambda_k}$, where λ_0 and λ_k are the actual Lagrange multipliers.

[6] We use a "bootstrapping" approach: if this assumption yields a solution, then the uniqueness of P^* will justify this assumption. Note that this amounts to the assumption that there is no world whose probability depends exponentially on ε.

[7] A possible interpretation for this set could be: "typically, Swedes are civilized", "typically, Swedes are blond", "typically, Swedes are civilized and blond".

The MC property guarantees that for each rule $r_i \in \mathcal{R}$ there is a world ω_i, in which only that rule is falsified. Thus, From Eq. (6), $P'(\omega_i) \approx \alpha_{r_i} \approx \varepsilon^{\kappa_i}$. Note that in the equation representing the constraint imposed by rule r_i (Eq. (9)), the *min* operation on the left-hand-side ranges over all worlds ω in which r_i is falsified. Clearly, the minimum of such worlds is ω_i, and the constraint equations for an MC set can be further simplified to be:

$$\kappa_i \;=\; 1 + \min_{\omega \in W^+_{r_i}} [\sum_{r_j \in R^-_\omega} \kappa_j] \quad 1 \leq i \leq n \qquad (10)$$

We now describe a greedy strategy for solving Eq. (10). Let us assume that we are given a consistent MC set $\mathcal{R}$ and let $\{\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3, \ldots\}$ be the partition of $\mathcal{R}$ that results from applying the labeling procedure described in Section 2 (Corollary 1). For every rule $r_i \in \mathcal{R}_1$ there is a world ω'_i for which the set $R^-_{\omega'_i}$ is empty, no rule is falsified by ω'_i and consequently $\sum_{r_j \in R^-_{\omega'_i}} \kappa_j = 0$. It follows that for every $r_i \in \mathcal{R}_1$ we must have $\kappa_i = 1$ (see Eq. (10)). We can now use these values to compute an initial upper bound for the rest of the κ's. We set to infinity all κ's associated with rules in $\mathcal{R} - \mathcal{R}_1$, and evaluate the right-hand-side of Eq. (10) associated with rules in $\mathcal{R}_2$. These evaluations will produce an upper bound for the κ's associated with these rules. Using these upper bounds we repeat this process and compute upper bounds for the κ's associated with rules in $\mathcal{R}_3$ and so on. By following the ordering induced by the labeling procedure, we are assured that none of this upper bounds for the κ's will be infinity. By definition, for each rule $r_i \in \mathcal{R}_n$ there is at least one world ω_i in which r_i is verified and the rules that are falsified must belong to $\mathcal{R}_m$ where $1 \leq m < n$. Thus, the *min* operator ranging over ω_i must produce a $\sum$-term which is no greater than that associated with that particular world ω_i, and this term is clearly finite. Once these initial upper bounds are computed we can divide the κ's into two disjoint sets: the first set includes the κ's for which a precise value is known (initially only the κ's associated with rules in $\mathcal{R}_1$), and the second set includes those for which only an upper bound is known. Let FINAL denote the first set, and let BOUNDED denote the second set. The objective is to compute precise values for the κ's in BOUNDED and transfer them into FINAL. Thus, until BOUNDED is empty, we repeatedly perform the following steps: (1) identify those κ^* in BOUNDED with minimal upper bound, (2) remove them from BOUNDED, (3) include them in FINAL and update the bounds of the remaining κ's in BOUNDED.

Theorem 7 *Given a consistent MC set $\mathcal{R}$ the above procedure computes a solution to Eq. (10), and requires no more than $|\mathcal{R}| - |\mathcal{R}_1|$ iterations.*

Once the κ-values are computed we have a complete model, and a new rule $r : A \rightarrow B$ is ratified as an ME-plausible conclusion if and only if the following equation is satisfied:

$$\min_{\omega \in W^+_r} [\sum_{r_k \in R^-_\omega} \kappa_k] < \min_{\omega \in W^-_r} [\sum_{r_j \in R^-_\omega} \kappa_j] \qquad (11)$$

Note that under the approximations described, the satisfaction of this equation will guarantee the satisfaction of Def. 3.

We conclude this section with a discussion of the issue of recognizing the active constraints in non-MC knowledge bases. The Lagrange multipliers method treats all constraints as equalities, and finds local maxima on the boundaries defined by these constraints. The problem with blindly assuming that all rules in a set $\mathcal{R}$ are active is that the system may become overspecified, and the technique we have been using might find spurious solutions which do not satisfy all the constraints. Such constraints violations cannot be detected by the infinitesimal analysis presented here since the coefficients of ε were ignored. Some passive constraints could, in principle, be detected before the maximization process begins, since they do not participate in delimiting the feasible region. For example consider:[8]

$$P(b|s) \geq 1 - \varepsilon \qquad (12)$$
$$P(w|s) \geq 1 - \varepsilon \qquad (13)$$
$$P(b, w|s) \geq 1 - \varepsilon \qquad (14)$$

Since the third statement implies the first two, the point of maximum entropy must lie in the region defined solely by Eq. (14). The first two constraints are completely irrelevant (and will be satisfied by strict inequalities). The main problem are those constraints that do constrict the feasible region, but do not influence the maximum value of the entropy. These constraints represent rules which already belong to the maximum entropy closure of the active set of rules. We know of no effective method of identifying these passive constraints in advance, and are currently exploring ways of detecting these constraints within the κ equations since, being passive, they should obtain a κ value of 0.

We remark that the task of identifying the passive constraints will be performed only once, during the construction of the model from $\mathcal{R}$, and can be *amortized* over many queries as long as $\mathcal{R}$ remains fixed. This optimistic note however, should be further qualified by the fact that the minimization required by Eq. (11) is NP-complete even for Horn expressions[9].

5 Summary and Illustration

The proposed method of infinitesimal ME-analysis computes a ranking function Θ on worlds, where

[8] Note that these are the probability constraints imposed by $\mathcal{R}_{scb}$ above.

[9] Rachel Ben-Eliyahu, personal communication.

$\Theta(\omega) = \log(P'(\omega))$ corresponds to the lowest exponent of ε in the expansion of $P^*_{\varepsilon,\mathcal{R}}(w)$ into a power series in ε. This ranking function is encoded parsimoniously by assigning an integer weight κ to each rule $r \in R$, and letting $\Theta(\omega)$ be the sum of the weights associated with the rules falsified by ω. Thus, worlds of lower Θ are considered more "normal" than those of higher Θ. The weight κ, in turn, reflects the "cost" we must add to each ω that falsifies the associated rule $A \rightarrow B$, so that the resulting ranking function would satisfy the constraint conveyed by $\mathcal{R}$, namely,

$$\begin{aligned}\min\{\ \Theta(\omega) \ &\mid \ \omega \models (A_i \wedge \neg B_i)\ \} > \\ \min\{\ \Theta(\omega) \ &\mid \ \omega \models (A_i \wedge B_i)\ \} \quad \forall r_i \in \mathcal{R} \quad (15)\end{aligned}$$

These considerations led to a set of $|\mathcal{R}|$ nonlinear equations for the weights κ which under certain conditions can be solved by iterative methods. The criterion for deciding whether an arbitrary rule $P \rightarrow Q$ is a ME-plausible conclusion of $\mathcal{R}$ is:

$$\hat{\Theta}(P \wedge Q) < \hat{\Theta}(P \wedge \neg Q) \qquad (16)$$

where for any formula E, $\hat{\Theta}(E)$ is defined as $\hat{\Theta}(E) = \min\{\Theta(\omega)|\omega \models E\}$. In other words, a rule is ME-plausible iff the ranking associated with the minimal world falsifying the rule is higher than the ranking associated with the minimal world verifying the rule.

As an example consider the MC set $\mathcal{R}_{pw} = \{r_1 : p \rightarrow \neg f,\ r_2 : p \rightarrow b,\ r_3 : b \rightarrow f,\ r_4 : b \rightarrow w\}$[10]. Both r_3 and r_4 are tolerated by $\mathcal{R}_{pw}$ (they belong to the first set in the partition of $\mathcal{R}_{pw}$) hence $\kappa_3 = \kappa_4 = 1$. The equations for κ_1 and κ_2 are:

$$\begin{aligned}\kappa_1 &= 1 + \min[\kappa_3, \kappa_2, (\kappa_3 + \kappa_4)] \\ \kappa_2 &= 1 + \min[\kappa_3, \kappa_1, (\kappa_3 + \kappa_4), (\kappa_1 + \kappa_4)]\end{aligned}$$

giving $\kappa_1 = \kappa_2 = 2$. These κ_i-values, $1 \leq i \leq 4$, completely specify the ranking Θ.

Assume we wish to inquire whether "green-birds fly", i.e. $(b \wedge g) \mid\!\sim f \in \mathcal{C}_{ME}(\mathcal{R}_{pw})$. Since the propositional variable g [11] does not appear in any rule of $\mathcal{R}_{pw}$, its truth value does not constraint the ranking function Θ (see Eq. (15)). Thus, it must be the case that $\hat{\Theta}(g \wedge b \wedge f) = \hat{\Theta}(b \wedge f)$ and $\hat{\Theta}(g \wedge b \wedge \neg f) = \hat{\Theta}(b \wedge \neg f)$. Since $b \rightarrow f \in \mathcal{R}_{pw}$, $\hat{\Theta}(b \wedge \neg f) > \hat{\Theta}(b \wedge f)$ and $(b \wedge g) \mid\!\sim f$ is indeed in the closure. In general it follows that the ME formalism complies with the intuition that, if nothing is specified in $\mathcal{R}$ about some property g, and $A \rightarrow B$ can be concluded from $\mathcal{R}$, then $(A \wedge g) \rightarrow B$ should also follow from $\mathcal{R}$.

Now consider whether penguins, despite being an exceptional class of birds (with respect to flying) can *inherit* other properties of birds. In particular, we wish to test whether $\mathcal{R}_{pw}$ sanctions that penguins are winged-animals. It is easy to verify that $\hat{\Theta}(p \wedge w) = 1$ while $\hat{\Theta}(p \wedge \neg w) = 2$, and in accordance with Eq. (16), $p \rightarrow w$ is an ME-plausible conclusion of $\mathcal{R}_{pw}$. Such conclusions, representing property inheritance across exceptional classes, are not sanctioned by ε-semantics nor by the rational closure of [Lehmann 89].

6 Discussion

As we saw in the previous section, ME overcomes some of the deficiencies of ε-semantics as well as rational monotony. In particular it properly handles irrelevant properties (a deficiency of ε-semantics), and sanctions property inheritance across exceptional subclasses (a deficiency of both ε-semantics and rational monotony). In fact maximum entropy can be viewed as an extension of these two systems. Like ε-semantics, ME is based on infinitesimal probability analysis, and like rational monotony, ME is based on optimal rankings of models subject to constraints, and a sanctions inferences on the basis of more normal worlds. Rational monotony however, is driven by a different ranking, uniquely determined by the relation of tolerance (see section (2)). In this ranking, called Z-ranking in [Pearl 90], worlds are ranked according to the *most crucial* rule violated in each world, while the rules are ranked according to the partition formed by the consistency test (see section (2)). In contrast ME ranks worlds according to the weighted sum of rule violations, and it is this difference that explains the ability of ME to conclude that "penguins are winged-animals" in the example from the previous section.

Another instance where the ME-ranking proves beneficial is in answering the following question, posed in [Lifschitz 89]: can the fact that we derive $\neg p \vee \neg q$ from $p \vee q$ when p, q are jointly circumscribed be explained in terms of probabilities close to 0 or 1? Translated to the ME formalism, we have $\mathcal{R}_{pq} = \{True \rightarrow \neg p, True \rightarrow \neg q\}$, and we wish to inquire whether $(p \vee q) \mid\!\sim (\neg p \vee \neg q)$ is in fact in $\mathcal{C}_{ME}(\mathcal{R}_{pq})$. Since the minimal world verifying $(p \vee q) \rightarrow (\neg p \vee \neg q)$ violates a subset of the rules violated by the world verifying $(p \vee q) \rightarrow (p \wedge q)$, we see that $\hat{\Theta}((p \vee q) \wedge (p \wedge q)) > \hat{\Theta}((p \vee q) \wedge \neg(p \wedge q))$ which verifies the conclusion. This conclusion is not sanctioned by rational monotony since, given that the two initial rules belong to the same rank (first level of the consistency partition), the ranking on worlds violating one or two rules will be the same. Note however, that had we encoded the information slightly different, e.g. $\mathcal{R}'_{pq} = \{True \rightarrow \neg(p \wedge q)\}$, ME would not yield the expected conclusion. This sensitivity to the format in which rules are expressed seems at odds with one of the basic conventions of traditional logic where $a \rightarrow (b \wedge c)$ is regarded as a "shorthand" for $a \rightarrow b$ and $a \rightarrow c$. However, it might be useful for distinguishing fine nuances in natural dis-

[10] The literals can be taken to mean bird, penguin, fly, and winged-animal, respectively.

[11] We are slightly abusing the language by using g both as a *propositional variable* denoting the property "green" and the *proposition* that "green" is true. The correct meaning of g however, should be clear from the context.

curse, treating q and p as two independent properties if expressed by two rules, and related properties if expressed together. Another pattern of reasoning sanctioned by maximum entropy is contraposition. For example, from $\mathcal{R}_{pw}$ we could conclude that animals with no wings are not birds ($\neg w \mid\!\sim \neg b$), but penguins with no wings are "ambiguous", they may or may not be birds.

The main weakness of the ME approach is the failure to respond to causal information (see [Pearl 88], pp. 463,519, and [Hunter 89]). This prevents this formalism from properly handling tasks such as the Yale shooting problem [Hanks & McDermott 86], where rules of causal character are given priority over other rules. This weakness may perhaps be overcome by introducing causal operators into the ME formulation, similar to the way causal operators are incorporated within other formalisms of nonmonotonic reasoning (e.g., [Shoham 86], [Geffner 89]).

References

[Adams 66] Adams, E., Probability and The Logic of Conditionals, in *Aspects of Inductive Logic*, ed. J. Hintikka and P. Suppes, Amsterdam: North Holland.

[Adams 75] Adams, E., The Logic of Conditionals, chapter II, Dordrecht, Netherlands: D. Reidel.

[Aoki 71] Aoki, M., *Introduction to Optimization Techniques*, Chapter 5, The Macmillan Company, New York, 1971.

[Cheeseman 83] Cheeseman, P., A Method of Computing Generalized Bayesian Probability Values for Expert Systems, Prc. of Intl. Joint Conf. on AI (IJCAI-83), Karlsruhe, W. Germany, 198—202.

[Dowling & Gallier 84] Dowling, W. and J. Gallier, Linear-Time Algorithms for Testing the Satisfiability of Propositional Horn Formulae, Journal of Logic Programming, 3:267–284, 1984.

[Geffner & Pearl 88] Geffner, H. and J. Pearl, A Framework for Reasoning with Defaults, to appear in *Defeasible Reasoning and Knowledge Representation*, H. Kyburg et. al. (eds.), Kluwer Publishers, 1990.

[Geffner 89] Geffner, H., Default Reasoning: Causal and Conditional Theories, UCLA Cognitive Systems Lab. TR-137, PhD dissertation, December 1989.

[Goldszmidt & Pearl 89] Goldszmidt, M. and J. Pearl, Deciding Consistency of Databases Containing Defeasible and Strict Information, Proceedings of the 5th Workshop on Uncertainty in AI, Windsor, Canada, August 1989, pp. 134–141.

[Goldszmidt & Pearl 90] Goldszmidt, M. and J. Pearl, On the Relation Between System Z and the Rational Closure, to appear in Proceedings of 3rd Intl. Workshop on Nonmonotonic Reasoning, 1990.

[Goldszmidt, Morris & Pearl 90] Goldszmidt, M., P. Morris and J. Pearl, A Maximum Entropy Approach to Nonmonotonic Reasoning, Technical Report, Cognitive Systems Lab., UCLA.

[Hanks & McDermott 86] Hanks, S., and D. McDermott, Default Reasoning, Nonmonotonic Logics, and the Frame Problem, Proc. 5th National Conference on AI (AAAI-86), Philadelphia, pp. 328—333.

[Hunter 89] Hunter, D., Causality and Maximum Entropy Updating, *Intl. Journal of Approximate Reasoning*, 3 (no. 1) pp. 87—114.

[Jaynes 79] Jaynes, E., Where Do We Stand on Maximum Entropy?, in *The Maximum Entropy Formalism*, eds. R. Levine and M. Tribus, Cambridge MIT press, 1979.

[Kraus et.al. 88] Kraus, S., D. Lehmann and M. Magidor, Preferential Models and Cumulative Logics, Technical Report TR 88-15, Dept. of Computer Science, Hebrew University, Jerusalem, Israel, November 1988.

[Lehmann & Magidor 88] Lehmann, D. and M. Magidor, Rational Logics and their Models: A Study in Cumulative Logics, TR-8816 Dept. of Computer Science, Hebrew Univ., Jerusalem, Israel.

[Lehmann 89] Lehmann, D., What Does a Knowledge Base Entail?, Proceedings of First International Conference on Knowledge Representation, Toronto, Canada, 1989, pp. 212–222.

[Lifschitz 89] Lifschitz, V., Open Problems on the Border of Logic and Artificial Intelligence, unpublished manuscript, 1989.

[Makinson 89] Makinson, D., General Theory of Cumulative Inference, Second International Workshop on Non-Monotonic Reasoning, Springer-Verlag, 1989.

[McCarthy 86] McCarthy, J.,Applications of Circumscription to Formalizing Common-Sense Knowledge, *Artificial Intelligence*, 28 (no. 1):89—116.

[Pearl 88] Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, San Mateo: Morgan Kauffman Publishers.

[Pearl 89a] Pearl, J., Probabilistic Semantics for Nonmonotonic Reasoning: A Survey, in Proceedings of the First Intl. Conf. on Principles of Knowledge Representation and Reasoning, Toronto, Canada, May 1989, pp. 505–516.

[Pearl 90] Pearl, J., System Z: A Natural Ordering of Defaults with Tractable Applications to Nonmonotonic Reasoning, in *Theoretical Aspects of Reasoning About Knowledge*, M. Vardi (ed.), Morgan Kaufmann Publishers, 1990, pp. 121–135.

[Satoh 90] Satoh, K., A Probabilistic Interpretation for Lazy Nonmonotonic Reasoning, to appear as ICOT-TR-525, Institute for New Generation Computer Technology, 1990.

[Shoham 86] Shoham, Y., Chronological Ignorance: Time, Necessity, and Causal Theories, Proc. 5th Natl. Conf. on AI (AAAI-86), Philadelphia, pp. 389—393.