MECHANIZING INDUCTION

Ronald Ortner and Hannes Leitgeb

In this chapter we will deal with "mechanizing" induction, i.e. with ways in which theoretical computer science approaches inductive generalization.

In the field of *Machine Learning*, algorithms for induction are developed. Depending on the form of the available data, the nature of these algorithms may be very different. Some of them combine geometric and statistical ideas, while others use classical reasoning based on logical formalism. However, we are not so much interested in the algorithms themselves, but more on the philosophical and theoretical foundations they share. Thus in the first of two parts, we will examine different approaches and inductive assumptions in two particular learning settings.

While many machine learning algorithms work well on a lot of tasks, the interpretation of the learned hypothesis is often difficult. Thus, while e.g. an algorithm surprisingly is able to determine the gender of the author of a given text with about 80 percent accuracy [Argamon and Shimoni, 2003], for a human it takes some extra effort to understand on the basis of which criteria the algorithm is able to do so. With that respect the advantage of approaches using logic are obvious: If the output hypothesis is a formula of predicate logic, it is easy to interpret. However, if decision trees or algorithms from the area of inductive logic programming are based purely on classical logic, they suffer from the fact that most universal statements do not hold for exceptional cases, and classical logic does not offer any convenient way of representing statements which are meant to hold in the "normal case". Thus, in the second part we will focus on approaches for *Nonmonotonic Reasoning* that try to handle this problem.

Both Machine Learning and Nonmonotonic Reasoning have been anticipated partially by work in philosophy of science and philosophical logic. At the same time, recent developments in theoretical computer science are expected to trigger further progress in philosophical theories of inference, confirmation, theory revision, learning, and the semantic and pragmatics of conditionals. We hope this survey will contribute to this kind of progress by building bridges between computational, logical, and philosophical accounts of induction.

1 MACHINE LEARNING AND COMPUTATIONAL LEARNING THEORY

1.1 Introduction

Machine Learning is concerned with algorithmic induction. Its aim is to develop algorithms that are able to generalize from a given set of examples. This is quite

Handbook of the History of Logic. Volume 10: Inductive Logic. Volume editors: Dov M. Gabbay, Stephan Hartmann and John Woods.
General editors: Dov M. Gabbay, Paul Thagard and John Woods.

a general description, and Machine Learning is a wide field. Here we will confine ourselves to two exemplary settings, viz. *concept learning* and *sequence prediction*.

In concept learning, the learner observes examples taken from some instance space X together with a *label* that indicates for each example whether it has a certain property. The learner's task then is to generalize from the given examples to new, previously unseen examples or to the whole instance space X. As each property of objects in X can be identified with the subset $C \subseteq X$ of objects that have the property in question, this *concept* C can be considered as a *target concept* to be learned.

EXAMPLE 1. Consider an e-mail program that allows the user to classify incoming e-mails into various (not necessarily distinct) categories (e.g. spam, personal, about a certain topic, etc.). After the user has done this for a certain number of e-mails, the program shall be able to do this classification automatically.

Sequence prediction works without labels. The learner observes a finite sequence over an instance set (alphabet) X and has to predict its next member.

EXAMPLE 2. A stock broker has complete information about the price of a certain company share in the past. Her task is to predict the development of the price in the future.

In the following, we will consider each of the two mentioned settings in detail. Concerning concept learning we also would like to refer to the chapter on *Statistical Learning Theory* of von Luxburg and Schölkopf in this volume, which deals with similar questions in a slightly different setting.

1.2 Concept Learning

The Learning Model

We start with a detailed description of the learning model. Given a basic set of instances X, the learner's task is to learn a subset $C \subseteq X$, called a *concept*. Learning such a *target concept* C means learning the characteristic function 1_C on X. That is, for each $x \in X$ the learner shall be able to predict whether x is in C or not.

EXAMPLE 3. For learning the concept "cow" one may e.g. consider X to be the set of all animals, while C would be the set of all cows. The concept would be learned if the learner is able to tell of each animal whether it is a cow.

In order to enable the learner to learn a concept C, she is provided with some *training examples*, that is, some instances taken from X together with the information whether each of these is in C or not. Thus the learner's task is to generalize from such a set of *labeled* training examples

$$\Big\{\big(x_1, 1_C(x_1)\big), \big(x_2, 1_C(x_2)\big), \ldots, \big(x_n, 1_C(x_n)\big)\Big\}$$

with $x_i \in X$ to a hypothesis $h : X \to \{0, 1\}$. If the learner's hypothesis coincides with 1_C she has successfully learned the concept C.

A special case of this general setting is the *learning of Boolean functions* where the task is to learn a function $f(p_1, p_2, \ldots, p_n)$ of Boolean variables p_i that takes values in $\{0, 1\}$. Obviously, any Boolean function can be represented by a formula of propositional logic (and vice versa) if the values of the variables p_i and the value of the function f are interpreted as truth values. Each training example for the learner consists of an assignment of values from $\{0, 1\}$ to the variables p_i together with the respective value of f. The task of the learner is to identify the function f. As each assignment of values to the p_i uniquely corresponds to a vector from $X := \{0, 1\}^n$, learning a Boolean function f is the same as learning the concept of all vectors x in X for which $f(x) = 1$.

No-Free-Lunch Theorems

Unfortunately, the space of possible concepts is the whole power set 2^X, so that without any further inductive assumptions learning is an impossible task (except for the trivial case, in which each instance in X is covered by the training examples), cf. [Mitchell, 1990]. Mitchell in his introduction to Machine Learning [1997, p.23] postulates as desired inductive assumption the following *inductive learning hypothesis*:

> "Any hypothesis found to approximate the target function well over a sufficiently large set of training examples will also approximate the target function well over other unobserved examples."

Although this is of course what we want to have, it does not really help us in achieving it. Given only the training examples, each possible hypothesis in 2^X which correctly classifies the training examples seems to be as likely.

On the other hand, machine learning literature provides a lot of algorithms starting from *decision trees* to *neural networks* and *support vector machines* (for an introduction to all this see e.g. [Mitchell, 1997]) that seem to work well on a lot of problems (and also with some theoretical results complementing the picture). How do these algorithms resolve the problem of induction?

Indeed, each algorithm (at least implicitly) defines its own hypothesis class. For a given set of training examples the algorithm will output a hypothesis. By doing this, the algorithm obviously has to prefer this hypothesis to all other possible hypotheses. As remarked before, the training sample helps only to a limited extent as there are a lot of *consistent* hypotheses in 2^X which classify the training examples correctly. Thus each learning algorithm is biased towards some hypotheses in 2^X in order to be able to make a decision at all. On the other hand, if all hypotheses in 2^X are equally likely, no learning algorithm will be able to perform better than another one *in general*. This is basically the content of so-called "no-free-lunch theorems" for supervised learning of which there exist various versions [Rao *et al.*, 1995; Schaffer, 1994; Wolpert, 1996b; Wolpert, 1996a]. For a discussion see also [von Luxburg and Schölkopf, 2009].

THEOREM 4 No-free-lunch, [Wolpert, 2001]. *Averaged over all possible target concepts, the performance of* any *learning algorithm on previously unseen test examples is that of a random guesser.*

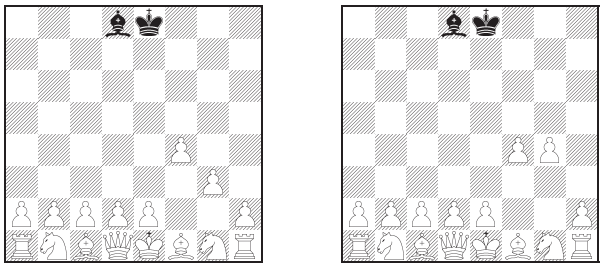
Thus, in order to save the possibility of learning, one either has to accept that each learning problem can be only tackled by suitable algorithms that on the other hand will fail on other learning problems. Alternatively, one may adopt an inductive assumption that denies that all possible target concepts in 2^X are equally likely.

ASSUMPTION 5 [Giraud-Carrier and Provost, 2005]. The process that presents us with learning problems induces a non-uniform probability distribution on possible target concepts $C \subseteq 2^X$.

Actually, this is not enough, as one also has to avoid all probability distributions on target concepts where good performance on some problems is exactly counterbalanced by poor performance on all other problems, see [Rao *et al.*, 1995] for details.

Most learning algorithms use that the instance set X will not be unstructured but (as e.g. in the common case where $X \subseteq \mathbb{R}^n$) provides a suitable distance metric $d : X \times X \to \mathbb{R}$ which can be used as a kind of similarity measure between instances. The inductive assumption in this case is that two similar (i.e. close) instances will have the same label. Thus, if in an object recognition problem two images differ in only a few pixels they will get the same label, if two e-mails differ only in some letters one will be spam only if the other is as well, and so on.[1] Although this is a very natural assumption, note that it won't be true for all learning problems whatsoever, as the following example shows.

EXAMPLE 6. Consider the two chess positions in the following diagram with Black to move in both of them. While the two positions will be considered to



be very close by any ordinary distance measure (as there is only a single white pawn placed differently), the first position is clearly in favor of White due to the enormous material advantage, while in the second diagram Black can checkmate

[1] This observation leads to the simple *k-nearest neighbor* algorithm, which classifies an instance according to the labels of the k nearest training examples. This algorithm not only works well in practice, it also has some advantageous theoretical properties. See [von Luxburg and Schölkopf, 2009].

immediately by moving the bishop to h4. Of course, it may well be that there is some other natural metric on the space of chess positions that works well, yet it is by no means clear what such a metric may look like. Thus, from a practical point of view for many learning problems it is important to find in advance a proper representation space together with a suitable metric for the data.

In the favorable case where there is a suitable distance metric on X, the only difficulty is to determine the boundary between positively and negatively labeled instances. Unfortunately, in general this boundary between positive and negative training examples also will not be uniquely determined by the training examples. Thus, the availability of a distance metric does not really solve our original problem, so that again any algorithm must specify some preference relation. (Accordingly, in Section 2 on Nonmonotonic Reasoning we will see that the standard semantics for nonmonotonic logics is based on preference relations over worlds or states.)

Occam's Razor

A common preference relation on the whole hypothesis space is to prefer in the spirit of Occam's razor simple hypotheses over complicated ones. Thus — to stay with the previous example — when choosing a boundary between positive and negative training examples, a hyperplane is e.g. preferred over a non-differentiable surface. Especially, in the presence of noise (i.e. when the labels of the training data may be wrong with some probability) Occam's razor is often used to avoid the danger of *overfitting* the training data, that is, to choose a hypothesis that perfectly fits the training data but is very complex and hence often does not generalize well.

There has been some discussion on the validity of Occam's razor (and also of the more or less synonymous *overfitting avoidance*) also in the machine learning community.[2] While Occam's razor often remains a rather vague principle, there are some theoretical results (some of which will be mentioned below) and attempts to clarify what Occam's razor in machine learning exactly is. Thus, it has been argued [Domingos, 1998] that the term "Occam's razor" is actually used for two different principles in the machine learning literature.

POSTULATE 7 Occam's first razor. Given two models with the same error on the whole instance space X, choose the simpler one.

POSTULATE 8 Occam's second razor. Given two models with the same error on the training sample, choose the simpler one.

While it is evidently easier to argue for Occam's first razor (although its validity is also not clear), only the second razor is of any use in machine learning. However, finding convincing arguments for this latter version is obviously more difficult. Basically there are two ways of argument for a theoretical justification of Occam's

[2]For a related discussion concerning the trade-off between *estimation error* and *approximation error* see [von Luxburg and Schölkopf, 2009].

second razor. First, there are some theoretical results from so-called PAC learning which can loosely be interpreted as support for Occam's second razor (see the section on *PAC learning* below). Second, there is the Bayesian argument which serves as a base for a lot of learning algorithms, the best-known of which is the MDL (minimum description length) principle introduced by Rissanen [1978].[3]

POSTULATE 9 MDL Principle. Choose the model that minimizes the total number of bits needed to encode the model and the data (given the model).[4]

There are some theoretical results that support the claim that Occam's second razor in the realization of the MDL principle indeed is the best strategy in almost all cases [Vitányi and Li, 2000]. However, these results consider an idealized MDL principle that (due to the use of *Kolmogorov complexity* [Li and Vitányi, 1997]) is uncomputable in practice.[5] On the other hand, although approximations of an idealized MDL approach are often successful in practice, the empirical success of Occam's second razor is controversial, too [Webb, 1996; Schaffer, 1993]. Of course, part of the reason for this is that practical MDL approaches (just as any other learning algorithm) cannot evade the earlier mentioned no-free-lunch results. That MDL has particular problems when the training sample size is small (so that the chosen hypothesis fits the data, but is too simple) is neither surprising nor a real defect of the approach: with insufficient training data provided, also complex models are likely to fail.

Metalearning

Some people try to evade the no-free-lunch theorems by lifting the problem to a meta-level (see e.g. [Baxter, 1998; Vilalta *et al.*, 2005]). Thus, instead of solving all problems with the same algorithm, it is attempted to assign each problem a suitable algorithm. Of course, from the theoretical point of view this does not help, as the no-free-lunch theorems obviously also hold for any meta-algorithm that consists of several individual algorithms. However, from the practical point of view this approach makes sense. In particular, it is e.g. certainly useful to apply an algorithm that is able to make use of any additional assumptions one has on the learning problem at hand. A similar theory of induction that prefers *local* induction over *global* induction has recently been proposed in [Norton, 2003].

After these general considerations we turn to more theoretical models of learning together with some results that have been achieved.

[3]Actually, MDL does not see itself as a Bayesian method. For a discussion of the relation of MDL to Bayesian methods (and a general introduction to MDL) see Chapters 1 and 17 of [Grünwald, 2007]. More detailed descriptions of MDL as well as Bayesian approaches can also be found in [von Luxburg and Schölkopf, 2009].

[4]As such, this basic idea of MDL does not look Bayesian at all, as there seem to be no probabilities involved. Indeed, these come into play by the observation that there is a correspondence between encodings and probability distributions (cf. Section 1.3, in particular footnote 20).

[5]We do not give any details here but refer to the section on *Solomonoff's theory of induction* below, which is closely related to idealized MDL.

PAC Learning

The simplest way to make learning feasible is to consider a setting where the space of possible concepts is restricted to a certain *concept class* known to the learner.

DEFINITION 10. Let X be an arbitrary (possibly infinite) set of instances. Then any subset $\mathcal{C}$ of the power set 2^X of X is called a *concept class over* X.

If the learner knows the concept class $\mathcal{C}$, she will choose her hypothesis h from $\mathcal{H} = \{1_C \mid C \in \mathcal{C}\}$. Thus our first assumption to make learning possible in this framework will be the following.

ASSUMPTION 11. The learner has access to a set of possible hypotheses $\mathcal{H} \subset 2^X$ that also contains a hypothesis corresponding to the target concept.

Of course, it will depend on the size of the concept class and the given training examples to which extent Assumption 11 will be helpful to the learner.

EXAMPLE 12. Assume $X = \{a, b, c, \ldots, z\}$ and $\mathcal{C} = \big\{\{a, b\}, \{b, c\}\big\}$. Then the learner will be able to identify a target concept taken from $\mathcal{C}$ if and only if either a or c is among the training examples.

EXAMPLE 13. Let $X = \{a, b, c\}$ and $\mathcal{C} = \big\{\{a\}, \{b\}, \{c\}, \{a, b\}, \{b, c\}\big\}$. It is easy to check that unlike in Example 12 two distinct training examples are needed in any case in order to identify a target concept taken from $\mathcal{C}$.

In general, the number of distinct training examples that are necessary (in the best as well as in the worst case) in order to identify a concept will depend on the combinatorial structure of the concept class.

DEFINITION 14. For $Y \subseteq X$ we set $\mathcal{C} \cap Y := \{C \cap Y \mid C \in \mathcal{C}\}$. Such a subset $Y \subseteq X$ is said to be *shattered by* $\mathcal{C}$, if $\mathcal{C} \cap Y = 2^Y$.

If $Y \subseteq X$ is shattered by a concept class $\mathcal{C}$ then it is easy to see that in the worst case, learning a concept in $\mathcal{C}$ will take $|Y|$ distinct training examples. Thus, the following definition provides an important combinatorial parameter of a concept class.

DEFINITION 15. The *VC-dimension*[6] of a concept class $\mathcal{C} \subseteq 2^X$ is the cardinality of a largest $Y \subseteq X$ that is shattered by $\mathcal{C}$.

Subsequent results will emphasize the significance of the VC-dimension. For a more detailed account on why the VC-dimension matters see [von Luxburg and Schölkopf, 2009].

EXAMPLE 16. The concept class given in Example 12 has VC-dimension 1, as only the sets $\{a\}$ and $\{c\}$ are shattered. In Example 13 we find that $\mathcal{C}$ shatters $\{a, b\}$ and has VC-dimension 2.

REMARK 17. In the *agnostic* learning setting (see the respective section below) where the learner does not know the concept class from which the target concept

[6]VC stands for Vapnik-Chervonenkis. Vapnik and Chervonenkis [1971] were together with [Sauer, 1972] and [Shelah, 1972] the first to consider the VC-dimension of concept classes.

is taken, the learner has to choose a hypothesis class $\mathcal{H}$ by herself. If there is no hypothesis in $\mathcal{H}$ that suits the training examples, the hypothesis class $\mathcal{H}$ can be considered to be *falsified* by the examples. The number of examples that is necessary in the worst case to falsify a class $\mathcal{H}$ of VC-dimension d is simply $d+1$. This can be easily verified, as the examples in a set Y shattered by $\mathcal{H}$ cannot falsify $\mathcal{H}$, independent of their labels. Hence, the VC-dimension can be used to measure the degree of falsifiability of hypothesis classes, as noted in [Corfield *et al.*, 2005]. Popper [1969, Chapter VI] had similar ideas, yet his measure for the degree of falsifiability in general does not coincide with the VC-dimension [Corfield *et al.*, 2005]. See also the discussion in [von Luxburg and Schölkopf, 2009].

Example 12 shows that if two concepts are close to each other (i.e. their characteristic functions coincide on almost all instances in X), then finding the target concept may take a lot of training examples (in the worst case). In the *PAC-framework* this problem is handled by weakening the task for the learner in that she need not *identify* the target concept, but that it is sufficient to *approximate* it. That is, the learner's hypothesis shall be correct on *most* instances in X.

As the learner usually will not be able to choose the training examples by herself,[7] one assumes that the training examples are drawn independently according to some fixed probability distribution $\mathcal{P}$ on X that is unknown to the learner. The learning success will of course depend on the concrete sample presented to the learner, so that the number of training examples the learner needs in order to approximate the target concept well will be a random variable depending on the distribution $\mathcal{P}$. Thus, with some (usually small) probability the training examples may be not representative for the target concept (e.g. if the same example is repeatedly drawn from X). Hence, it is reasonable to demand from the learner to approximate the target concept only *with high probability*, that is, to learn *probably approximately correct*, which is what PAC stands for.

However, learning still may be impossible, if the distribution $\mathcal{P}$ has support[8] on a proper subset $Y \subset X$, so that some instances will never be sampled. Thus, one measures the performance of the learner's hypothesis not uniformly over the whole instance space X, but according to the same distribution $\mathcal{P}$ that also generates the training examples. That is, the error of the learner's hypothesis $h : X \to \{0,1\}$ with respect to a target concept C and a distribution $\mathcal{P}$ on X is defined as

$$\mathrm{er}_{C,\mathcal{P}}(h) := \mathcal{P}\Big(\big\{x \mid h(x) \neq 1_C(x)\big\}\Big).$$

One may consider this as the error expected for a randomly drawn *test example*, where one makes the *inductive assumption* that this test example is drawn according to the same distribution $\mathcal{P}$ that generated the training sample.

ASSUMPTION 18. The training as well as the test examples are drawn from the instance set X according to an unknown but fixed distribution $\mathcal{P}$.

[7] See however the *active learning* setting described below.

[8] The *support* of a probability distribution is basically the set of instances that have positive probability.

Summarizing, a learner *PAC learns* a concept class, if for ε, $\delta > 0$ there is a number $m = m(\varepsilon, \delta)$ of training examples that are sufficient to approximate the target concept with high probability. That is, with probability at least $1 - \delta$ the output hypothesis has error smaller than ε. More precisely, this leads to the following definition.[9]

DEFINITION 19. A concept class $\mathcal{C} \subseteq 2^X$ is called *PAC learnable*, if for all ε, $\delta \in (0, 1)$ there is an $m = m(\varepsilon, \delta)$, such that for all probability distributions $\mathcal{P}$ on X and all $C \in \mathcal{C}$: when learning C from m examples, the output hypothesis h has error $\mathrm{er}_{C,\mathcal{P}}(h) > \varepsilon$ with probability smaller than δ (in respect to the m examples drawn independently according to $\mathcal{P}$ and labeled by $\mathcal{C}$).

This framework has been introduced by Valiant [1984]. For a comparison of this learning model to alternative models see [Haussler *et al.*, 1991] or also [Wolpert, 1995] where the PAC learning model is embedded into a Bayesian framework. A lot of earlier work in computational learning dealt with learning a target concept in the limit (i.e. when the number of training examples goes to infinity). Research in this direction (with some links to recursion theory) goes back to [Gold, 1967]. For an overview see [Angluin and Smith, 1983]; [Osherson and Weinstein, 2009] also deals with that approach.

Valiant [1984, p.1142] remarks that an interesting consequence of his learning model is that when a population has successfully learned a concept based on the same underlying probability distribution, there still may be significant differences on the learned concept. In particular, examples that appear only with very small probability are irrelevant for learning, so that

> "thought experiments and logical arguments involving unnatural hypothetical situations may be meaningless activities."

It is a natural question to ask which concept classes are PAC learnable. Obviously, this will also depend on the learning algorithm. Choosing e.g. a stupid algorithm that even classifies most of the training examples wrongly will obviously prevent learning. Thus, one often turns the attention to *consistent* learners that always choose a hypothesis h that is consistent with the training sample,[10] i.e. for a target concept C and training examples $x_1, \ldots, x_n$ one has $h(x_i) = 1_C(x_i)$ for $1 \leq i \leq n$.[11] Such consistent learners then are able to PAC learn finite concept

[9] Usually, there are also some considerations about the run-time complexity of an algorithm that PAC learns a concept class. For now, we will neglect this for the sake of simplicity, and come back to the question of *efficient* PAC learning when discussing *Occam algorithms* and *polynomial learnability* below.

[10] While at first sight it may look foolish to consider hypotheses that are not consistent, this certainly makes sense if there is noise in the training data. Furthermore as mentioned by Kelly [2004b], when also considering questions of computability it may happen that the restriction to *computable* consistent hypotheses prevents learning, see also [Osherson *et al.*, 1988; Kelly and Schulte, 1995].

[11] As we assume that the learner has access to the concept class $\mathcal{C}$ from which the target concept is taken, it is obvious that there is a consistent hypothesis $h \in \mathcal{H} = \{1_C \mid C \in \mathcal{C}\}$.

classes, where the necessary number of examples can be shown to depend on the size of the concept class.

THEOREM 20 [Haussler, 1988]. *Any consistent learning algorithm needs* $O\left(\frac{1}{\varepsilon}(\log\frac{1}{\delta}+\log|\mathcal{C}|)\right)$ *examples for PAC learning any finite concept class* $\mathcal{C}$.

More generally, not the absolute size but the *VC-dimension* of a concept class turns out to be the critical parameter. Thus, concept classes of finite VC-dimension are PAC learnable, and the number of necessary examples for learning can be upper bounded using the VC-dimension as follows.[12]

THEOREM 21 [Blumer *et al.*, 1989]. *Any consistent learning algorithm needs* $O\left(\frac{1}{\varepsilon}(\log\frac{1}{\delta}+d\log\frac{1}{\varepsilon})\right)$ *examples for PAC learning any well-behaved*[13] *concept class of VC-dimension* d.

For finite concept classes this is slightly worse than the result of Theorem 20, as the VC-dimension of a finite class $\mathcal{C}$ may take values up to $\log_2|\mathcal{C}|$, see [Blumer *et al.*, 1989].

Of course, particular learning algorithms may PAC learn concept classes using fewer examples. Here is e.g. an alternative bound for the (consistent) *1-inclusion graph algorithm* of [Haussler *et al.*, 1994].

THEOREM 22 [Haussler *et al.*, 1994]. *The 1-inclusion graph learning algorithm needs* $O\left(\frac{d}{\varepsilon}\log\frac{1}{\delta}\right)$ *examples for PAC learning any well-behaved concept class of VC-dimension* d.

The following lower bound shows that the dependence on the VC-dimension is necessary.

THEOREM 23 [Ehrenfeucht *et al.*, 1989]. *Let* $\mathcal{C}$ *be an arbitrary concept class of VC-dimension* d. *Then there is a probability distribution* $\mathcal{P}$ *such that any consistent learner needs* $\Omega\left(\frac{1}{\varepsilon}(d+\log\frac{1}{\delta})\right)$ *examples for PAC learning* $\mathcal{C}$.

In particular this means that it is impossible to PAC learn concept classes of infinite VC-dimension,[14] so that it becomes an interesting question which concept classes have finite VC-dimension. For certain concept classes the VC-dimension can be easily calculated. Thus, the concept class of all (open or closed) intervals on the real line $\mathbb{R}$ has VC-dimension 2. The class of axis-parallel rectangles in $\mathbb{R}^n$ has VC-dimension $2n$. Half-spaces as well as balls in $\mathbb{R}^n$ have VC-dimension $n+1$. Further concept classes with finite VC-dimension can be found e.g. in [Vapnik and Chervonenkis, 1974], [Dudley, 1984], [Wenocur and Dudley, 1981], [Assouad, 1983], or [Haussler and Welzl, 1987]. Examples for concept classes with infinite VC-dimension are finite unions of intervals or the interiors of Jordan curves in $\mathbb{R}^2$.

[12] Although the results presented by von Luxburg and Schölkopf [2009] concern a slightly different setting, they give some good intuition for why finiteness is important and why a combinatorial parameter like the VC-dimension matters even if hypothesis classes and instance space are infinite or even continuous.

[13] Usually, one has to impose some modest measure-theoretic assumptions when the considered concept classes are infinite, cf. Appendix A1 of [Blumer *et al.*, 1989].

[14] See however the paragraph on *polynomial learnability* below.

In spite of the negative result implied by Theorem 23 it is not hopeless to learn such classes. Algorithms as well as theoretical results may make use of a suitable parametrization of the domain in order to achieve useful results about what is called *polynomial learnability* (Definition 25 below) in [Blumer *et al.*, 1989]. For details see the discussion of *Occam algorithms* below.

For particular classes with finite VC-dimension, beside the PAC bound of Theorem 21 often alternative or sharper, sometimes even optimal bounds can be derived. Thus (in view of Theorem 23) optimal PAC bounds of $O\left(\frac{1}{\varepsilon}(d + \log\frac{1}{\delta})\right)$ can be derived e.g. for axis-parallel hyperrectangles in $\mathbb{R}^d$ [Auer *et al.*, 1998] or classes with certain combinatorial structure [Auer and Ortner, 2007]. However, some of these bounds only hold for special algorithms, while for particular consistent algorithms sharper lower bounds than given in Theorem 23 can be shown. Thus, consider e.g. the concept class $\mathcal{C}_{X,d} := \{C \subseteq X \mid |C| \leq d\}$ of all subsets of X of size at most d. This simple concept class has VC-dimension d and is PAC learnable only from $\Omega\left(\frac{1}{\varepsilon}(d\log\frac{1}{\varepsilon} + \log\frac{1}{\delta})\right)$ examples for the learning algorithm which chooses a *largest* consistent hypothesis from $\mathcal{C}_{X,d}$ [Auer and Ortner, 2007]. It is notable that the algorithm that chooses a *smallest* consistent hypothesis needs only $O\left(\frac{1}{\varepsilon}(d + \log\frac{1}{\delta})\right)$ examples to PAC learn $\mathcal{C}_{X,d}$ [Auer and Ortner, 2007]. This can be seen as a theoretical justification of Occam's razor, as choosing a simple (which in this case means small) hypothesis provides better PAC bounds than choosing a more complex (i.e. larger) hypothesis.

Occam Algorithms and Polynomial Learnability

In fact, there are also other theoretical justifications of Occam's razor. Consider a concept class $\mathcal{C}$ (of arbitrary, possibly infinite VC-dimension) together with some kind of complexity measure on the concepts in $\mathcal{C}$. It would be a straightforward realization of Occam's razor to demand to choose a hypothesis consistent with the training examples which has smallest complexity. However, as it turns out that computing such a hypothesis may be (NP-)hard (cf. e.g. the example given in [Blumer *et al.*, 1989]), one confines oneself to the more modest requirement to choose a hypothesis that is significantly simpler than the training sample. This can be made precise as follows. Let $\mathcal{C}^{\mathcal{A}}_{s,m}$ be the *effective hypothesis space*[15] of an algorithm $\mathcal{A}$ that is presented with m training examples of a concept of complexity $\leq s$.

DEFINITION 24. A learning algorithm $\mathcal{A}$ is an *Occam algorithm* for a concept class $\mathcal{C}$ with complexity measure $s : \mathcal{C} \to \mathbb{Z}^+$, if there is a polynomial $p(s)$ and a constant $\alpha \in [0, 1)$, such that for all $s, m \geq 1$, the VC-dimension of $\mathcal{C}^{\mathcal{A}}_{s,m}$ is upper bounded by $p(s)m^{\alpha}$.

It can be shown that Occam algorithms satisfy a learnability condition that is similar to PAC learning.

DEFINITION 25. A concept class $\mathcal{C} \subseteq 2^X$ with given complexity measure

[15] This is the set of all hypotheses that may be the outcome of the algorithm.

$s : \mathcal{C} \to \mathbb{Z}^+$ is called *polynomially learnable*, if for all ε, $\delta \in (0,1)$ there is an $m = m(\varepsilon, \delta, s)$, such that for all probability distributions $\mathcal{P}$ on X and all $C \in \mathcal{C}$ with $s(C) \leq s$: when learning C from m examples, the output hypothesis h has error $\mathrm{er}_{C,\mathcal{P}}(h) > \varepsilon$ with probability smaller than δ in respect to the m examples drawn independently according to $\mathcal{P}$ and labeled by $\mathcal{C}$.

Thus, unlike in the original PAC setting, polynomial learning of more complex concepts (in terms of the given complexity measure s) is allowed to take more examples.

THEOREM 26 [Blumer *et al.*, 1989]. *Any concept class for which there is an Occam algorithm is polynomially learnable.*

Theorem 26 is a generalization of a similar theorem of [Blumer *et al.*, 1987]. Sample complexity bounds as given for PAC learning can be found in [Blumer *et al.*, 1989]. As already mentioned in [Blumer *et al.*, 1989], Theorem 26 can be considered as showing a relationship between learning and data compression. If an algorithm is able to compress the training data (as Occam algorithms do), it is capable of learning. Interestingly, there are also some results that indicate some validness for the other direction of this implication [Board and Pitt, 1990; Li *et al.*, 2003].

The general idea of connecting learnability and compressability is the base of *Solomonoff's theory of induction*, which will be discussed in Section 1.3 below. In this direction, the definition of Occam algorithms has been adapted using the notion of *Kolmogorov complexity* in [Li and Vitányi, 1997] and [Li *et al.*, 2003], resulting in improved complexity bounds.

Agnostic Learning and Efficient PAC Learning

Obviously, the setting introduced above is not realistic in that usually the learner has no idea what the possible concepts are that may label the training examples. Thus, in general the learner has no access to a concept class that contains the target concept, so that Assumption 11 does not hold. Learning in this restricted setting is called *agnostic*. In the agnostic learning model introduced by Haussler [1992] no assumption about the labels of the examples are made (i.e. there need not be a target concept according to which examples are labeled). Instead of a concept class from which target concept and the learner's hypothesis are taken, there is only a set of possible hypotheses $\mathcal{H}$ from which the learner chooses. As it is not clear a priori whether the learner's hypothesis class is suitable for the problem at hand, the learner's performance is measured not with respect to a perfect label prediction (which may be impossible to achieve with an unsuitable hypothesis space $\mathcal{H}$), but with respect to the best hypothesis in $\mathcal{H}$. That way, an analogous definition of PAC learning (Definition 19 above) can be given. Thus, as above it is assumed that the distribution that produces the training examples is also used for measuring the performance of the learner's hypothesis, that is, Assumption 18 still holds.

Haussler [1992] has shown that in order to achieve positive results on agnostic PAC learnability with respect to some hypothesis class $\mathcal{H}$, it is sufficient to solve the optimization problem of finding for any finite set of labeled examples the hypothesis $h \in \mathcal{H}$ with the minimal number of misclassifications *on these examples*. Unfortunately, this optimization problem is computationally hard for many interesting hypothesis classes. Further, this also concerns *efficient* (i.e. polynomial time) PAC learning [Kearns *et al.*, 1994; Feldman, 2008]. Consequently, there have been some negative results on efficient agnostic PAC learning of halfspaces [Höffgen *et al.*, 1995] or conjunctions of literals [Kearns *et al.*, 1994].[16] Similar negative results about *efficient* non-agnostic PAC learning go back to [Pitt and Valiant, 1988]. These latter results also show that results about efficient (non-)learnability depend on the chosen representation of the concept class. However, for suitable hypothesis classes there are also some positive results for agnostic PAC learning, see e.g. [Kearns *et al.*, 1994; Maass, 1994; Auer *et al.*, 1995].

Online Learning, Transduction and Active Learning

The discussed models are only a small part of the machine learning and computational learning theory literature. In this section, we would like to indicate the existence of other interesting models not mentioned above. For a general overview of computational learning theory models and results see [Angluin, 1992].

The e-mail example (Example 1) shows some peculiarities that we have not considered so far. First, the test examples the program has to classify are not present all at once but have to be classified one after another. This is called *online learning*. This form of learning may have advantages as well as disadvantages. On the one hand, the learner does not have the distribution of the test examples to draw any conclusions from it. On the other hand, if an example is misclassified, the user may intervene and correct the mistake, so that the program gets additional information. For more about online learning see e.g. [Blum, 1998].

A related special feature of Example 1 is that, as there is always only a single example to classify, it is not necessary for the program to generate a hypothesis for the whole space of possible e-mails. It is sufficient to classify each incoming e-mail individually. Of course, this can be done by first generating a global hypothesis from which one infers the label of the example in question. However, as Vapnik [1998], p.477 put it:

> "When solving a problem of interest, do not solve a more general problem as an intermediate step. Try to get the answer that you really need

[16] These results usually hold only relative to an unsolved problem of complexity theory, i.e. they are valid provided that the complexity classes of NP and RP do not coincide. RP is a superclass of P that contains all decision problems which can be solved in *randomized polynomial time*, i.e. in polynomial time by a probabilistic Turing machine. If NP=RP, then it is easy to give an efficient learning algorithm.

> but not a more general one. (...) Do not estimate a function if you only need to estimate its values at given points. (Try to perform direct inference rather than induction.)"

Thus, it is quite natural to try to find the correct label of the single example directly. This has been termed *transduction* (contrary to *induction* which generates general rules) by Vapnik [1995].[17] Although from a conceptual point of view there may be little difference between transduction and induction (after all, if I know how to get a label for each single instance, I have automatically also an inductive rule), practically it may be easier to get the label for a single instance than to label the whole instance space.

Another possibility for our e-mail program may be that it asks the user for the label of an e-mail that is difficult to classify. This is called *active learning* or *query learning*. Usually, this is not considered in an online model (as our e-mail example), but the learner is allowed to choose examples from the instance space by herself and query their labels. As a consequence, the learner may concentrate on "interesting" examples which contain more information, so that she will sometimes be able to learn concepts with fewer examples than in ordinary PAC learning.[18] Geometrically, these interesting examples usually lie very close to the boundary that corresponds to the learner's current hypothesis, that shall separate the positively labeled from the negative examples. For an overview of active learning results see e.g. [Angluin, 1992; Freund *et al.*, 1997; Angluin, 2004].

1.3 Sequence Prediction

Learning Setting

In concept learning, the labels provide the learner with a certain pattern among the training examples. If the learner has to discover such patterns by herself one speaks of *unsupervised learning* (as opposed to *supervised learning*). There are various unsupervised learning settings such as *data mining*, where the learner tries to extract useful information from large data sets. Here we want to consider an *online* setting, where the learner's task is to predict the next entry in a finite sequence of observations. Thus, the learner is presented with a sequence of observations

[17] Actually, already Carnap [1950] distinguished between various forms of inductive inference, two of which are *universal inference* and *singular predictive inference*, the latter corresponding to what Vapnik calls *transduction*.

[18] There are also similar settings where the learner need not identify a concept, but has to check whether the concept at hand has a certain property. This *property testing* scenario is in particular natural when dealing with graphs, that is, when X is the set of all possible edges of a graph with given vertex set V, and the concept class is a set of graphs with common vertex set V. For the relation between property testing and learning see [Goldreich *et al.*, 1998]. There is also some literature on query learning of graphs, as in the graph setting beside *membership* (i.e. edge) queries there are other natural queries one may consider (e.g. edge count, shortest path etc.). See e.g. [Alon and Asodi, 2005; Angluin and Chen, 2008; Bouvel *et al.*, 2005].

$x_1, x_2, \ldots, x_t$ over some instance set (alphabet) X and has to predict the next observation x_{t+1}. After her prediction the true value is revealed.

REMARK 27. Note that this simple setting may deal with seemingly far more general problems. Consider e.g. transductive online concept learning from training examples $(z_1, y_1), \ldots, (z_n, y_n)$ where the z_i are taken from some instance space and the labels $y_i \in \{0, 1\}$ are given according to some target concept. The learner's task is to classify some z_{n+1}. This can be encoded as a sequence prediction problem where the task is to predict the next entry in the sequence $z_1, y_1, \ldots, z_n, y_n, z_{n+1}$.

The learner's performance when predicting an $x'_{t+1} \in X$ is usually evaluated by $\ell(x_{t+1}, x'_{t+1})$ for a suitable *loss function* $\ell : X \times X \to \mathbb{R}$ that measures the distance between the true value x_{t+1} and the prediction x'_{t+1}. A natural loss function that works for arbitrary X is obtained by setting $\ell(x, x') := 0$ if $x = x'$ and 1 otherwise. If $X \subseteq \mathbb{R}$, another common loss function is the squared error

$$(1) \quad \ell(x, x') := (x - x')^2.$$

In any case, loss functions are usually chosen so that the learner will try to minimize the loss.

More generally, the learner will not predict a single instance in X but e.g. a probability distribution on X. Thus, in general it is assumed that the learner makes at time $t + 1$ a decision d_{t+1} taken from some decision space D. The loss function ℓ then maps each pair $(d_{t+1}, x_{t+1}) \in D \times X$ to $\mathbb{R}$.

Similar to the case of concept learning, without further assumptions the learner will stand little chance to learn a given sequence. Even in the simplest, binary case with $X = \{0, 1\}$ each prediction of the learner can be thwarted by a suitable sequence (see [Dawid, 1985]).[19] The strategy to make *concept learning* possible has been twofold. On the one hand, one assumes that not all concepts are equally likely (Assumption 5), on the other hand one restricts the space of possible hypotheses (which e.g. in the PAC learning setting was done by giving the learner access to a concept class that contains the target concept). While in the setting of (PAC) concept learning both of these measures are taken, in sequence prediction either assumption leads to a different framework. Thus, we either assume that the sequence is generated by an unknown probability distribution (the *probabilistic setting*), or we consider a fixed sequence and restrict the possible hypotheses for prediction (*deterministic setting*). Thus, there is some duality between these two settings concerning the made assumptions. This duality is particularly strong when the chosen loss function is the *self-information loss* function (sometimes also called *log-loss* function)

$$(2) \quad \ell(d, x) = -\log_2 d(x),$$

[19] Another argument of this kind with emphasis on computability has been produced by Putnam [1963] in order to criticize Carnap's inductive logic. Putnam showed that no computable prediction algorithm will work on all computable sequences (cf. also the discussion in [Kelly, 2004b]).

where d is a probability distribution on X and $x \in X$. Beside some technical advantages of this function, it also establishes a relation between prediction and coding, as $\log_2 d(x)$ gives the *ideal* (binary) *code length* of x with respect to the probability distribution d.[20] For details on the duality between probabilistic and deterministic setting under the self-information loss function see [Merhav and Feder, 1998], which also gives a general overview of sequence prediction results.

In the probabilistic setting, good prediction performance need not be measured by a loss function. If the learner's decision is a probability distribution on X that shall approximate the real underlying distribution, there are various requirements the learner's probability distribution should have in order to be regarded as good. Dawid's *prequential analysis* [Dawid, 1984; Dawid and Vovk, 1999] deals with these questions of testing statistical models. Interestingly, this question again is closely related to the self-information loss function (see [Merhav and Feder, 1998]).

In the following, we pick two particularly interesting topics, on the one hand *Solomonoff's theory of induction* in the probabilistic setting, and on the other hand, *prediction with expert advice* in the deterministic setting.

Solomonoff's Theory of Induction

We have already met the idea that learning is related to compression (see the part on *Occam algorithms* above), which leads to the application of information theoretic ideas to learning. Ray Solomonoff's theory of induction [Solomonoff, 1964a; Solomonoff, 1964b] reduces prediction to data compression. The idea is summarized in the following postulate, which is evidently an implementation of Occam's razor that identifies simplicity with compressability.

POSTULATE 28. Given a (finite) sequence σ over an alphabet X, predict the $x \in X$ that minimizes the difference between the length of the shortest program that outputs σx (i.e. the sequence σ followed by x) and the length of the shortest program that outputs σ.[21]

There seems to be a fundamental problem with this postulate, as it looks as if it depended on the chosen programming language. However, as was shown independently by Solomonoff [1964a], Kolmogorov [1965], and Chaitin [1969], asymptotically the length of two equivalent computer programs in different *universal* programming languages differs by at most an additive constant (stemming from the length of a compiler that translates one language into the other). Under this *invariance theorem* it makes sense to consider the length $K(\sigma)$ of the shortest program that outputs a sequence σ, which is basically the *Kolmogorov complexity*

[20] That is, if one wants to encode words over the alphabet X where the probability of a letter $x \in X$ is $d(x)$, then an optimal *binary* encoding (to keep the average word length as small as possible) will assign the letter x a code of length about $\log_2 d(x)$, see [Shannon, 1948] and [Rissanen, 1976].

[21] Actually, one may also predict more than a single element, so that in general x may be a (finite) sequence over X as well.

of the sequence.[22] It can be shown that prediction under Postulate 28, which chooses x so that $K(\sigma x) - K(\sigma)$ is minimized, works well in the limit for a large majority of sequences, provided that the sequence is binary (i.e. $X = \{0, 1\}$) and the underlying distribution generating the sequence satisfies some benign technical assumptions.

THEOREM 29 [Vitányi and Li, 2000]. *Let $\mathcal{P}$ be a distribution on the set $\{0,1\}^\infty$ of all possible infinite sequences over $\{0,1\}$ that generates an infinite sequence ω.*[23] *Assume that $\mathcal{P}$ is a recursive measure*[24] *and that ω is a $\mathcal{P}$-random*[25] *sequence. Then the x that maximizes $\mathcal{P}(x|\sigma)$ minimizes $K(\sigma x) - K(\sigma)$ with $\mathcal{P}$-probability converging to 1, as the length of σ tends to infinity.*

Solomonoff was not only interested in prediction. His motivation was to determine the degree of confirmation[26] that a sequence σ is followed by x. Thus, the aim is to obtain a respective probability distribution on all possible continuations of the sequence σ. Solomonoff uses a Bayesian approach to achieve this. For the general problem of choosing a suitable prior probability distribution, a *universal* distribution $\mathcal{U}$ (which is also closely related to the notion of Kolmogorov complexity) is defined which prefers simple continuations of σ and exhibits some favorable properties [Solomonoff, 1978].[27] In particular, the universal distribution converges fast to the real underlying probability distribution (under similar assumptions as in Theorem 29).

THEOREM 30 Gács [Li and Vitányi, 1997]. *Given that $\mathcal{P}$ is a positive recursive measure over $\{0,1\}^\infty$ that generates a $\mathcal{P}$-random binary infinite sequence,*

$$\frac{\mathcal{U}(x|\sigma)}{\mathcal{P}(x|\sigma)} \to 1$$

with $\mathcal{P}$-probability 1, when the length of σ tends to infinity.

Moreover, the sum over the expected squared errors is basically bounded by

[22] Actually there are various (Kolmogorov) complexity variants that coincide up to an additive constant (that depends on the sequence σ). For the sake of simplicity we are not going to distinguish them here. See Section 4.5 of [Li and Vitányi, 1997] for details.

[23] Note that the assumption made in this setting is different from the PAC learning setting. Whereas in PAC learning it is assumed that each single example is drawn according to a fixed distribution, here the distribution is over all possible infinite sequences, which is actually more general.

[24] This is a modest assumption on the computability of the distribution function, see Chapter 4 of [Li and Vitányi, 1997].

[25] A sequence $\sigma = \sigma_1\sigma_2\ldots$ is *$\mathcal{P}$-random* if $\sup_n \mathcal{U}(\sigma_1\ldots\sigma_n)/\mathcal{P}(\sigma_1\ldots\sigma_n) < \infty$, where $\mathcal{U}$ is the universal prior distribution (cf. below). In the set of all infinite sequences, the $\mathcal{P}$-random sequences have $\mathcal{P}$-measure 1, that is, almost all considered sequences will be $\mathcal{P}$-random, see Section 4.5 of [Li and Vitányi, 1997].

[26] A student of Carnap, he explicitly refers to Carnap's [1950]. For more about Carnap's role see [Solomonoff, 1997].

[27] It has been argued [Kelly, 2004a] that Solomonoff's theory of induction only provides a circular argument for Occam's razor, as the chosen prior already prefers short descriptions. However, this neglects that the prior distribution itself is a good predictor as the subsequent results show.

the Kolmogorov complexity $K(\cdot)$ of the underlying distribution [Li and Vitányi, 1997].[28]

THEOREM 31 [Solomonoff, 1978]. *If $\mathcal{P}$ is a recursive measure over $\{0,1\}^\infty$ that generates a binary sequence, then*

$$\sum_n \sum_{|\sigma|=n-1} \mathcal{P}(\sigma)\big(\mathcal{U}(0|\sigma) - \mathcal{P}(0|\sigma)\big)^2 \;\leq\; \frac{K(\mathcal{P})\ln 2}{2},$$

where $|\sigma|$ denotes the length of the sequence σ.

Unfortunately, neither Kolmogorov complexity nor the universal prior distribution are computable [Li and Vitányi, 1997]. Thus, while Solomonoff's framework of *algorithmic probability* may offer a theoretical solution to the problem of induction, it cannot be directly applied to practical problems. However, on the other hand there are some principal theoretical limitations on the computability of prediction algorithms, cf. e.g. [Putnam, 1963][29] and [V′yugin, 1998].[30] In fact, it has been argued that there is a strong analogy between uncomputability and the problem of induction [Kelly, 2004c].

Another problem is that the mentioned constant of the invariance theorem in general will be quite large so that for short sequences the theoretical results are worthless, while the approach may not work well in practice. In spite (or maybe because of) these two deficiencies, Solomonoff's research has ignited a lot of research that on the one hand improved over theoretical results [Li and Vitányi, 1997; Hutter, 2001; Hutter, 2004; Hutter, 2007], while on the other hand, many practical approaches can be considered as approximations to his uncomputable algorithm. In particular, the MDL approach mentioned in Section 1.2 (see Postulate 9) emanated from Solomonoff's work. For a closer comparison of the two frameworks see Chapter 17 of [Grünwald, 2007].

Prediction with Expert Advice

In the deterministic setting where the underlying sequence that shall be predicted is considered to be fixed, it will be necessary to compete with the best hypothesis in a confined hypothesis space, as it is impossible to compete with perfect prediction in general (similarly to the no-free-lunch theorem). On the other hand, it is obviously futile to predict deterministically. For each deterministic prediction there is a sequence where the prediction will be wrong (and more generally, will maximize the loss function). Thus, the learner has to maintain a probability distribution on the possible predictions. Note that this probability distribution is used for randomization. Unlike that, in the probabilistic setting, the learner often uses a probability distribution to approximate the real underlying distribution.

[28] For the definition of the Kolmogorov complexity of a distribution we refer to Chapter 4 of [Li and Vitányi, 1997].

[29] Cf. footnote 19.

[30] For further theoretical limitations due to Gödel's incompleteness results see [Legg, 2006].

Usually (cf. also the concept learning setting), the set of hypotheses will be large in order to guarantee that there is a hypothesis that predicts the sequence well. For more about this setting see [Merhav and Feder, 1998]. Here we will consider the setting where the hypothesis space $\mathcal{H}$ is finite. The hypotheses in $\mathcal{H}$ are usually referred to as *experts* that serve as reference forecasters. We assume that the experts' predictions are taken from the same decision space D the learner chooses her prediction from. (In special cases D may equal X.) Note that as it is not known how the experts determine their predictions, there is no assumption about this. The learner may use these experts' advice to determine her own prediction. Note that in general the learner's prediction will not coincide with one of the expert's prediction: all the experts may suggest a deterministic prediction, while we have already seen that it only makes sense for the learner to predict randomly according to some distribution. The learner's goal is to compete with the best expert. That is, the learner will suffer a loss of $\ell(d_t, x_t)$ at time t for her decision $d_t \in D$. Similarly, at time t each expert $E \in \mathcal{H}$ has loss $\ell(d_t^E, x_t)$ for his decision d_t^E. Competing with the best expert then means that the learner will try to keep the *regret* with respect to the best expert

$$R_T \;:=\; \min_{E\in\mathcal{H}} \sum_{t=1}^{T} \ell(d_t^E, x_t) - \sum_{t=1}^{T} \ell(d_t, x_t)$$

as low as possible. Surprisingly, under some mild technical assumptions, one can show that for some learning algorithms the average regret (over time) tends to 0 for each individual sequence, when T approaches infinity.

THEOREM 32 [Auer *et al.*, 2002; Cesa-Bianchi and Lugosi, 2006]. *Assume that the decision set D is a convex subset of $\mathbb{R}^n$ and consider some expert set $\mathcal{H}$. Further let ℓ be a loss function that takes values only in the interval $[0,1]$ and is convex in the first argument,*[31] *i.e. for each $x \in X$, $\lambda \in [0,1]$ and $d, d' \in D$:*

$$\ell\big(\lambda d + (1-\lambda)d', x\big) \;\le\; \lambda\ell(d,x) + (1-\lambda)\ell(d',x).$$

Then the regret R_T of the exponentially weighted forecasting algorithm (as specified on pp.14 and 17 in [Cesa-Bianchi and Lugosi, 2006]) can be bounded as

$$R_T \;\le\; 2\sqrt{\frac{T}{2}\ln|\mathcal{H}|} + \sqrt{\frac{\ln|\mathcal{H}|}{8}}$$

for each $T > 0$ and each individual sequence over X.

What is remarkable about this theorem is that it does not need any inductive assumptions. The reason why the theorem holds for any sequence is that, intuitively speaking, by considering the loss with respect to the best expert only the difference to this best expert matters, so that the underlying sequence in some

[31] The convexity condition holds e.g. for the square loss function (1) and the logarithmic loss function (2).

sense is not important anymore. Of course, practically the theorem only has impact if there is at least one expert whose predictions are good enough to keep the loss with respect to the underlying sequence low. Increasing the number of experts $|\mathcal{H}|$ to guarantee this of course deteriorates the bound.

For more results in the expert advice setting see [Cesa-Bianchi and Lugosi, 2006] that also deals with applications to game theory.

2 NONMONOTONIC REASONING

2.1 Introduction

In order to cope successfully with the real world, AI applications need to reproduce patterns of everyday commonsense reasoning. As theoretical computer scientists began to realize in the late 1970s, such patterns of inference are hard, if not impossible, to formalize in standard first-order logic. New proof-theoretic and semantic mechanisms were sought-after by which conclusions could be inferred in all "normal" cases in which the premises were true, thus trying to capture the way in which human agents fill knowledge gaps by means of default assumptions, in particular, *conditional* default assumptions of an 'if...then...' form.

EXAMPLE 33. Assume you want to describe what happens to your car when you turn the ignition key: 'If the ignition key is turned in my car, then the car starts.' seems to be a proper description of the situation. But how shall we represent this claim in a first-order language? The standard way of doing it, according to classical AI, would be in terms of universal quantification and material implication, i.e. by means of something of the form $\forall t(\varphi[t] \rightarrow \psi[t])$, where φ, ψ are complex formulas, t is a variable for points of time, and $\rightarrow$ is the material conditional sign. But what if the gas tank is empty? You better improve your description by adding a formalization of '...and the gas tank is not empty' to φ. However, the resulting statement could still be contradicted by a potato that is clogging the tail pipe, or by a failure of the battery, or by an extra-terrestrial blocking your engine, and so forth. The possible exceptions to universally quantified material conditionals are countless, heterogeneous, and unclear. Nevertheless we seem to be able to communicate and reason rationally with the original information 'If the ignition key is turned in my car, then the car starts.', and the same should be true of intelligent computers.

How are human agents able to circumvent this problem? The key to an answer is to understand that we do not actually take 'If the ignition key is turned in my car, then the car starts.' as expressing that at any point of time it is not the case that the ignition key is turned and the car does not start — after all, what is negated here might indeed be the case in exceptional circumstances — but rather that *normally* given the ignition key is turned at a time, the car starts. Instead of trying to enumerate a possibly indefinite class of exceptions, we tacitly or explicitly qualify 'If the ignition key is turned in my car, then the car starts.' as saying

something about normal or likely circumstances, whatever these circumstances may look like. As a consequence, the logic of such everyday if-then claims differs from the logic of (universally quantified) material conditionals in first-order logic. In particular, while *Monotonicity* (or Strengthening of the Antecedent), i.e.

$$\frac{\varphi \rightarrow \psi}{\varphi \wedge \rho \rightarrow \psi}$$

is logically valid for material $\rightarrow$ (whether in the scope of universal quantifiers or not), the acceptance of the conditional 'If Tweety is a bird, then [normally] Tweety is able to fly.' does not seem to rationally necessitate the acceptance of any of the following conditionals: 'If Tweety is a penguin bird, then [normally] Tweety is able to fly.'; 'If Tweety is a dead bird, then [normally] Tweety is able to fly.'; 'If Tweety is a bird with his feet set in concrete, then [normally] Tweety is able to fly.'. So computer scientists found themselves in need of expressing formally if-then statements on the basis of which computers should be able to draw justified inferences about everyday matters, but where these statements do not logically obey Monotonicity; hence their speaking of 'nonmonotonic' conditionals or inference. This is the subject matter of *Nonmonotonic Reasoning*, without doubt one of the most vibrant areas of theoretical computer science in the last 30 years.

Nonmonotonic reasoning systems become *inductive* reasoners in the sense of the *Machine Learning* part of this article by the following move: assume the complete information that a database contains is the factual information

$$\varphi_1, \ldots, \varphi_m$$

together with the conditional information

$$\alpha_1 \Rightarrow \beta_1, \ldots, \alpha_n \Rightarrow \beta_n$$

where '$\Rightarrow$' is a new conditional connective which expresses 'if... then normally ...'. From the conditionals that are stored in the database the reasoning system now aims to derive some further conditionals the antecedents of which exhaust the complete factual information in the database, i.e. conditionals of the form

$$\varphi_1 \wedge \ldots \wedge \varphi_m \Rightarrow \psi$$

For every conditional that can be derived in this way, the factual information ψ is inferred by the system. Since the conditionals involved only express what holds in normal circumstances, this is an *inductive* inference from $\varphi_1, \ldots, \varphi_m$ to ψ under the tacit assumption that the reasoning system does not face an abnormal situation. The antecedents of the conditionals which the system aims to derive have to consist of the total factual information that is accessible to the system, as it would be invalid to strengthen weaker antecedents by means of the Monotonicity rule. On the methodological side, the main question to be answered at this point is: Which

rules of inference may the reasoner apply in order to derive further "normality conditionals" from its given "normality conditionals"? We are going to deal with this question in detail further down below.

Here are some brief pointers to the literature: Although Ginsberg [1987] is outdated as a collection of articles, it still proves to be useful if one wants to see where Nonmonotonic Reasoning derives from historically. Brewka, Dix, and Konolige [1997] and Makinson [2005] give excellent and detailed overviews of Nonmonotonic Reasoning. Schurz and Leitgeb [2005] is an informative compendium of papers dealing with some of the more empirical and philosophical aspects of Nonmonotonic Reasoning; for more references to psychological investigations into nonmonotonic reasoning see Oaksford, Chater, and Hahn [2009]. Finally, the *Stanford Encyclopedia of Philosophy* includes two nice entries on "Non-monotonic Logic" and "Defeasible Reasoning" which can be accessed online.

2.2 Nonmonotonic Reasoning: The KLM Approach

There are, broadly speaking, two approaches of how to formalize statements such as 'If the ignition key is turned in my car, then the car starts.' or 'If Tweety is a bird, then Tweety is able to fly.' in terms of nonmonotonic conditionals: Either exceptional circumstances are represented explicitly as those which contradict certain explicitly made claims, or they are left implicit, simply by not mentioning them at all.

The paradigm case of the first type of formalization is default logic (see Reiter [1980]) in which e.g. the Tweety case is handled by a so-called default rule which expresses: *If you know that Tweety is a bird, and nothing you know is inconsistent with Tweety being able to fly, then you are allowed to conclude that Tweety is able to fly.* Such consistency-based approach dominated the scene in the 1980s.

According to the other approach — which goes back to Shoham [1987], but which is exemplified most famously by the KLM approach, i.e. Kraus, Lehmann, and Magidor [1990], which really took off in the 1990s — the conditional 'If Tweety is a bird, then Tweety is able to fly.' is left unchanged syntactically, but the 'if'-'then' connective that it contains is understood as: *In the most normal or preferred circumstances in which Tweety is a bird, it is the case that Tweety is able to fly.* In the following, we will concentrate exclusively on the second, preferential approach, which turned out to be the dominant one as far as the *logical* aspects of nonmonotonic reasoning are concerned. Although the KLM account was anticipated by theories in philosophical logic, philosophy of language, and inductive logic — as we will highlight in the later sections — it is still widely unknown outside of computer science. So summarizing its main achievements proves to be useful even though the original sources (mainly, KLM [1990] and Lehmann and Magidor [1992]) are themselves clear, self-contained and extensive. The KLM approach also led to new logical treatments of inference in neural networks, which we will discuss briefly as well.

Conditional Theories (Nonmonotonic Inference Relations)

We are now going to deal with various systems of nonmonotonic reasoning which have been introduced by KLM [1990]. In contrast with KLM, we will not present these systems in terms of so-called inference or consequence relations, i.e. as binary relations $\mid\!\sim$ on a propositional language $\mathcal{L}$ (cf. Makinson [1994] and [1989]), but rather, more syntactically minded, as conditional theories, i.e. as sets of conditionals closed under rules of nonmonotonic reasoning. So instead of saying that $\alpha \mid\!\sim \beta$, we will say that $\alpha \Rightarrow \beta \in \mathcal{TH}_{\Rightarrow}$, where $\mathcal{TH}_{\Rightarrow}$ is a theory of conditionals, and $\Rightarrow$ is a new conditional connective which we will use to express nonmonotonic conditionals. In this way, it will be easier to compare the logical systems of nonmonotonic reasoning with systems of conditional logic studied in other areas. Furthermore, calling the relations $\mid\!\sim$ *consequence* relations typically leads to confusion on the side of philosophers: These are not meant to be relations of *logical* consequence; rather they have a similar methodological status as theories, i.e. they are meant to support plausible inferences within some intended domain of application. But mainly this is all just a matter of presentation; conditional theories in our sense can still be viewed as being *nothing but* inference relations.

$\mathcal{L}$ will always be some language of propositional logic that is based on finitely many propositional variables, with connectives $\neg$, $\wedge$, $\vee$, $\rightarrow$, $\leftrightarrow$, $\top$ (for tautology), and $\bot$ (for contradiction). $\mathcal{L}_{\Rightarrow}$ will be the set of all formulas of the form $\alpha \Rightarrow \beta$ for $\alpha, \beta \in \mathcal{L}$, with $\Rightarrow$ being the new nonmonotonic conditional sign. Note that $\mathcal{L}_{\Rightarrow}$ does not allow for nestings of nonmonotonic conditionals nor for the application of propositional operators to nonmonotonic conditionals.

Finally, whenever we will refer to a theory $\mathcal{TH}_{\rightarrow}$ (rather than $\mathcal{TH}_{\Rightarrow}$), we mean a deductively closed set of formulas in $\mathcal{L}$; each such set is going to entail deductively a set of *material* conditionals. We will always consider our conditional theories $\mathcal{TH}_{\Rightarrow}$ of "soft" or "defeasible" conditionals, such as $bird \Rightarrow fly$, as extending "hard" material conditionals, such as $penguin \rightarrow bird$, which are entailed by some given theory $\mathcal{TH}_{\rightarrow}$; the corresponding notion of 'extending' is made precise by the rules of Left Equivalence and Right Weakening stated below. We will leave the question open at this point whether the formulas of $\mathcal{L}$ ought to be regarded as open formulas or rather as sentences, and whether formulas of the forms $\alpha \rightarrow \beta$ and $\alpha \Rightarrow \beta$ ought to be regarded as tacitly quantified in some way or not. We will return to this point later.

In our presentation of systems of nonmonotonic logic, we will follow the (more detailed) presentation given by Leitgeb [2004], part III.

DEFINITION 34.

1. A *conditional C-theory* extending $\mathcal{TH}_{\rightarrow}$ is a set $\mathcal{TH}_{\Rightarrow} \subseteq \mathcal{L}_{\Rightarrow}$ with the property that for all $\alpha \in \mathcal{L}$ it holds that

$$\alpha \Rightarrow \alpha \in \mathcal{TH}_{\Rightarrow} \text{ (Reflexivity)}$$

and which is closed under the following rules:

(a) $\dfrac{\mathcal{TH}_{\rightarrow} \vdash \alpha \leftrightarrow \beta,\ \alpha \Rightarrow \gamma}{\beta \Rightarrow \gamma}$ (Left Equivalence)

(b) $\dfrac{\mathcal{TH}_{\rightarrow} \vdash \alpha \rightarrow \beta,\ \gamma \Rightarrow \alpha}{\gamma \Rightarrow \beta}$ (Right Weakening)

(c) $\dfrac{\alpha \wedge \beta \Rightarrow \gamma, \alpha \Rightarrow \beta}{\alpha \Rightarrow \gamma}$ (Cautious Cut)

(d) $\dfrac{\alpha \Rightarrow \beta, \alpha \Rightarrow \gamma}{\alpha \wedge \beta \Rightarrow \gamma}$ (Cautious Monotonicity)

We refer to the axiom scheme and the rules above as the system C (see KLM [1990], pp.176–180). The rules are to be understood as follows: E.g. by Cut, if $\alpha \wedge \beta \Rightarrow \gamma \in \mathcal{TH}_{\Rightarrow}$ (where propositional connectives such as $\wedge$ always bind more strongly than $\Rightarrow$) and $\alpha \Rightarrow \beta \in \mathcal{TH}_{\Rightarrow}$, then $\alpha \Rightarrow \gamma \in \mathcal{TH}_{\Rightarrow}$.

2. A conditional C-theory $\mathcal{TH}_{\Rightarrow}$ (extending whatever set $\mathcal{TH}_{\rightarrow}$) is *consistent* iff $\top \Rightarrow \bot \notin \mathcal{TH}_{\Rightarrow}$.

3. A *conditional CL-theory* $\mathcal{TH}_{\Rightarrow}$ extending $\mathcal{TH}_{\rightarrow}$ is a conditional C-theory extending $\mathcal{TH}_{\rightarrow}$, which is also closed under the following rule:

   $\dfrac{\alpha_0 \Rightarrow \alpha_1, \alpha_1 \Rightarrow \alpha_2, \ldots, \alpha_{j-1} \Rightarrow \alpha_j, \alpha_j \Rightarrow \alpha_0}{\alpha_r \Rightarrow \alpha_{r'}}$ (Loop)

   (r, r' are arbitrary members of $\{0, \ldots, j\}$).

   We refer to C+Loop as the system CL (see KLM [1990], pp.187).

4. A *conditional P-theory* $\mathcal{TH}_{\Rightarrow}$ extending $\mathcal{TH}_{\rightarrow}$ is a conditional CL-theory extending $\mathcal{TH}_{\rightarrow}$, which is closed under the additional rule:

   $\dfrac{\alpha \Rightarrow \gamma, \beta \Rightarrow \gamma}{\alpha \vee \beta \Rightarrow \gamma}$ (Or)

   We refer to CL+Or as the system P (see KLM [1990], pp.189–190; there it is also shown that Loop can actually be derived from the other rules in P).

5. A *conditional R-theory* $\mathcal{TH}_{\Rightarrow}$ extending $\mathcal{TH}_{\rightarrow}$ is a conditional P-theory extending $\mathcal{TH}_{\rightarrow}$, which has the following property (this is a so-called non-Horn condition: see Makinson [1994], Section 4.1, for further details):

   If $\alpha \Rightarrow \gamma \in \mathcal{TH}_{\Rightarrow}$, and $\alpha \Rightarrow \neg\beta \notin \mathcal{TH}_{\Rightarrow}$, then $\alpha \wedge \beta \Rightarrow \gamma \in \mathcal{TH}_{\Rightarrow}$ (Rational Monotonicity).

   We refer to P+Rational Monotonicity as the system R (see Lehmann and Magidor [1992], pp.16–48).

Each of these rules is meant to apply for arbitrary α, β, γ, α_0, α_1, …, $\alpha_j \in \mathcal{L}$.

REMARK 35.

- It is easy to see that a conditional C-theory $\mathcal{TH}_{\Rightarrow}$ is consistent iff $\mathcal{TH}_{\Rightarrow}$ is non-trivial, i.e. $\mathcal{TH}_{\Rightarrow} \neq \mathcal{L}_{\Rightarrow}$ (use Right Weakening and Cautious Monotonicity).
- If a conditional C-theory $\mathcal{TH}_{\Rightarrow}$ extending $\mathcal{TH}_{\rightarrow}$ is consistent, then also $\mathcal{TH}_{\rightarrow}$ is consistent, i.e. $\mathcal{TH}_{\rightarrow} \nvdash \bot$ (use Reflexivity and Right Weakening).

Cumulativity, i.e. Cautious Cut and Cautious Monotonicity taken together, has been suggested by Gabbay [1984] as a valid closure property of plausible reasoning. The stronger system P, which extends cumulativity by a rule for disjunction, has become the standard system of nonmonotonic logic and can be proved sound and complete with respect to many different semantics of nonmonotonic logic (some of them are collected in Gabbay, Hogger, and Robinson [1994]; see also Gärdenfors and Makinson [1994], Chapter 4.3 in Fuhrmann [1997], Benferhat, Dubois, and Prade [1997], Benferhat, Saffiotti, and Smets [2000], Goldszmidt and Pearl [1996], Pearl and Goldszmidt [1997], Halpern [2001b]). We are going to deal with the most influential semantics for the logical systems introduced above — the preferential semantics of KLM [1990] — below.

Derivable Rules

LEMMA 36. *(KLM [1990], pp.179–180) The following rules are derivable in C, i.e. if the premises of the following rules are members of a conditional C-theory $\mathcal{TH}_{\Rightarrow}$, then one can prove that the same holds for their conclusions:*

1. $\dfrac{\alpha \Rightarrow \beta, \alpha \Rightarrow \gamma}{\alpha \Rightarrow \beta \wedge \gamma}$ *(And)*

2. $\dfrac{\alpha \Rightarrow \beta, \beta \Rightarrow \alpha, \alpha \Rightarrow \gamma}{\beta \Rightarrow \gamma}$ *(Equivalence)*

3. $\dfrac{\alpha \Rightarrow (\beta \rightarrow \gamma), \alpha \Rightarrow \beta}{\alpha \Rightarrow \gamma}$ *(Modus Ponens in the Consequent)*

4. $\dfrac{\alpha \vee \beta \Rightarrow \alpha, \alpha \Rightarrow \gamma}{\alpha \vee \beta \Rightarrow \gamma}$

5. $\dfrac{\mathcal{TH}_{\rightarrow} \vdash \alpha \rightarrow \beta}{\alpha \Rightarrow \beta}$ *(Supra-Classicality)*

LEMMA 37. *(KLM [1990], p.191) The following rules are derivable in P:*

1. $\dfrac{\alpha \wedge \beta \Rightarrow \gamma}{\alpha \Rightarrow (\beta \rightarrow \gamma)}$ *(S)*

2. $\dfrac{\alpha \wedge \beta \Rightarrow \gamma, \alpha \wedge \neg\beta \Rightarrow \gamma}{\alpha \Rightarrow \gamma}$ *(D)*

By means of any of the semantics for these systems of nonmonotonic logic, it is easy to prove that neither Contraposition nor Transitivity nor Monotonicity (for $\Rightarrow$) is derivable in any of them. The following examples from everyday reasoning show that this is exactly as it ought to be:

EXAMPLE 38.

- $\dfrac{\text{If } a \text{ is a human, then normally } a \text{ is not a diabetic.} \checkmark}{\text{If } a \text{ is a diabetic, then normally } a \text{ is not human. ???}}$ (Contraposition)

- $\dfrac{\begin{array}{l}\text{If } a \text{ is from Munich, then normally } a \text{ is a German.} \checkmark \\ \text{If } a \text{ is a German, then normally } a \text{ is not from Munich.} \checkmark\end{array}}{\text{If } a \text{ is from Munich, then normally } a \text{ is not from Munich. ???}}$ (Transitivity)

- $\dfrac{\text{If } a \text{ is a bird, then normally } a \text{ is able to fly.} \checkmark}{\text{If } a \text{ is a penguin bird, then normally } a \text{ is able to fly. ???}}$ (Monotonicity)

Derivability of Conditionals from Conditional Knowledge Bases

The notion of derivability of a conditional from a set of conditionals (in AI terms: from a *conditional knowledge base*) is defined in analogy with derivability for formulas of classical propositional logic, with the exception of the system R.

DEFINITION 39. Let $KB_{\Rightarrow} \subseteq \mathcal{L}_{\Rightarrow}$:

1. A *C-derivation* (rel. to $\mathcal{TH}_{\rightarrow}$) of $\varphi \Rightarrow \psi$ from $KB_{\Rightarrow}$ is a finite sequence

$$\langle \alpha_1 \Rightarrow \beta_1, \ldots, \alpha_k \Rightarrow \beta_k \rangle$$

   where $\alpha_k = \varphi$, $\beta_k = \psi$, and for all $i \in \{1, \ldots, k\}$ at least one of the following conditions is satisfied:

   - $\alpha_i \Rightarrow \beta_i \in KB_{\Rightarrow}$.
   - $\alpha_i \Rightarrow \beta_i$ is an instance of Reflexivity.
   - $\alpha_i \Rightarrow \beta_i$ is the conclusion of one of the rules of C, such that the conditional premises of that rule are among $\{\alpha_1 \Rightarrow \beta_1, \ldots, \alpha_{i-1} \Rightarrow \beta_{i-1}\}$, and in the case of Left Equivalence and Right Weakening the derivability conditions concerning $\mathcal{TH}_{\rightarrow}$ are satisfied.

2. $KB_{\Rightarrow} \vdash_{C}^{\mathcal{TH}_{\rightarrow}} \varphi \Rightarrow \psi$ ($\varphi \Rightarrow \psi$ is *C-derivable* rel. to $\mathcal{TH}_{\rightarrow}$ from $KB_{\Rightarrow}$)
   iff there is a C-derivation of $\varphi \Rightarrow \psi$ rel. to $\mathcal{TH}_{\rightarrow}$ from KB.

3. $Ded_C^{\mathcal{TH}_\rightarrow}(KB_\Rightarrow) = \left\{\varphi \Rightarrow \psi \,\middle|\, KB_\Rightarrow \vdash_C^{\mathcal{TH}_\rightarrow} \varphi \Rightarrow \psi\right\}$
   (the *conditional C-closure* of $KB_\Rightarrow$ rel. to $\mathcal{TH}_\rightarrow$).
4. $\varphi \Rightarrow \psi$ is *C-provable* (rel. to $\mathcal{TH}_\rightarrow$) iff $\varnothing \vdash_C^{\mathcal{TH}_\rightarrow} \varphi \Rightarrow \psi$.

Analogous concepts can be introduced for the systems CL and P.

REMARK 40.

1. As in the case of deductive derivability, it follows that
   (a) $KB_\Rightarrow \subseteq Ded_C^{\mathcal{TH}_\rightarrow}(KB_\Rightarrow)$.
   (b) If $KB_\Rightarrow \subseteq KB'_\Rightarrow$ then $Ded_C^{\mathcal{TH}_\rightarrow}(KB_\Rightarrow) \subseteq Ded_C^{\mathcal{TH}_\rightarrow}(KB'_\Rightarrow)$.
   (c) $Ded_C^{\mathcal{TH}_\rightarrow}(Ded_C^{\mathcal{TH}_\rightarrow}(KB_\Rightarrow)) = Ded_C^{\mathcal{TH}_\rightarrow}(KB_\Rightarrow)$.
2. $\mathcal{TH}_\Rightarrow$ is a conditional C-theory extending $\mathcal{TH}_\rightarrow$ iff
   $Ded_C^{\mathcal{TH}_\rightarrow}(\mathcal{TH}_\Rightarrow) = \mathcal{TH}_\Rightarrow$.
   Since $Ded_C^{\mathcal{TH}_\rightarrow}(Ded_C^{\mathcal{TH}_\rightarrow}(KB_\Rightarrow)) = Ded_C^{\mathcal{TH}_\rightarrow}(KB_\Rightarrow)$, $Ded_C^{\mathcal{TH}_\rightarrow}(KB_\Rightarrow)$ is a conditional C-theory extending $\mathcal{TH}_\rightarrow$ for arbitrary $KB_\Rightarrow$. In particular, $Ded_C^{\mathcal{TH}_\rightarrow}(\varnothing)$ (the set of formulas which are C-provable rel. to $\mathcal{TH}_\rightarrow$) is a conditional C-theory extending $\mathcal{TH}_\rightarrow$.
3. $Ded_C^{\mathcal{TH}_\rightarrow}(KB_\Rightarrow)$ is the smallest conditional C-theory extending $\mathcal{TH}_\rightarrow$ which contains $KB_\Rightarrow$.

EXAMPLE 41. Assume $KB_\Rightarrow$ to consist of

1. $bird \Rightarrow fly$,
2. $penguin \Rightarrow \neg fly$.

Suppose $\mathcal{TH}_\rightarrow$ contains

3. $penguin \rightarrow bird$.

By an application of Supra-Classicality to 3, one can derive

4. $penguin \Rightarrow bird$.

Applying Cautious Monotonicity to 4 and 2 yields

5. $penguin \wedge bird \Rightarrow \neg fly$.

This can be interpreted as follows: Since by conditional 3 the penguin information is at least as specific as a mere bird information, conditional 2 overrides conditional 1: penguin birds are derived to be unable to fly.

In the case of R, derivability has to be defined differently due to the presence of a non-Horn rule, i.e. Rational Monotonicity:

DEFINITION 42.

1. $KB_{\Rightarrow} \vdash_R^{\mathcal{TH}_{\rightarrow}} \varphi \Rightarrow \psi$ ($\varphi \Rightarrow \psi$ is *R-derivable* rel. to $\mathcal{TH}_{\rightarrow}$ from $KB_{\Rightarrow}$) iff $\varphi \Rightarrow \psi$ is a member of $\bigcap\{\mathcal{TH}_{\Rightarrow} \mid \mathcal{TH}_{\Rightarrow} \supseteq KB_{\Rightarrow}, \mathcal{TH}_{\Rightarrow}$ is a cond. R-theory extend. $\mathcal{TH}_{\rightarrow}\}$.

2. $Ded_R^{\mathcal{TH}_{\rightarrow}}(KB_{\Rightarrow}) = \left\{\varphi \Rightarrow \psi \,\middle|\, KB_{\Rightarrow} \vdash_R^{\mathcal{TH}_{\rightarrow}} \varphi \Rightarrow \psi\right\}$

   (the *conditional R-closure* of $KB_{\Rightarrow}$ rel. to $\mathcal{TH}_{\rightarrow}$).

3. $\varphi \Rightarrow \psi$ is *R-provable* (rel. to $\mathcal{TH}_{\rightarrow}$) iff $\varnothing \vdash_R^{\mathcal{TH}_{\rightarrow}} \varphi \Rightarrow \psi$.

$Ded_R^{\mathcal{TH}_{\rightarrow}}$ satisfies the same closure conditions as stated above. In particular, note that the deductive closure operator of each of these systems of nonmonotonic logic is *monotonic* (see 1b in Remark 40 above); so these logics are nonmonotonic only in the sense that they are logical systems for conditionals which are not monotonic with respect to their antecedents, i.e. which do not obey Monotonicity as a logically valid rule. In other words: the term 'nonmonotonic reasoning' is ambiguous — it can either refer to 'inference by means of nonmonotonic conditionals' (this is what we have considered so far) or to 'nonmonotonic deductive closure/entailment' (this is what we will deal with in Subsection 2.2 below) or to both.

Obviously, the system C is weaker than CL in terms of derivability, and the system CL is weaker than P, where the weaker-than relations in question are nonreversable. More surprisingly, P and R are equally strong in terms of derivability:

THEOREM 43. *(See Lehmann and Magidor [1992], pp.24f, for the semantic version of this result.)*
$KB_{\Rightarrow} \vdash_P^{\mathcal{TH}_{\rightarrow}} \alpha \Rightarrow \beta$ *iff* $KB_{\Rightarrow} \vdash_R^{\mathcal{TH}_{\rightarrow}} \alpha \Rightarrow \beta$.

With respect to provability, i.e. derivability from the empty knowledge base, all of the systems dealt with above turn out to be equally strong and in fact just as strong as classical logic (if $\Rightarrow$ is replaced by $\rightarrow$).

Preferential Semantics

Next we follow KLM [1990] and Lehmann and Magidor [1992] by introducing preferential or ranked models for conditional theories, where the intended interpretation of the preference relations or rankings that are part of such models is in terms of "degrees of normality". (Such ranked models are also closely related to Spohn's [1987] so-called ordinal conditional functions or ranking functions.) For each of these types of models, we presuppose a non-empty set W of possible worlds which we consider to be given antecedently and we think of as representing an agent's "hard" knowledge. Each world $w \in W$ is assumed to stand in a standard satisfaction relation with respect to the formulas of $\mathcal{L}$.

DEFINITION 44.

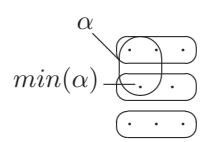
1. A *cumulative model* $\mathfrak{M}$ is a triple $\langle S, l, \prec \rangle$ with
   (a) a non-empty set S of so-called "states",
   (b) a labeling $l : S \to 2^W \setminus \{\varnothing\}$ of states,
   (c) a normality "order", or preference relation, $\prec \subseteq S \times S$ between states; if $s_1 \prec s_2$, we say that s_1 is more normal than s_2 (note that $\prec$ is not necessarily a strict order relation);
   (d) such that, $\mathfrak{M}$ satisfies the Smoothness Condition (see below).
2. Factual formulas $\alpha \in \mathcal{L}$ are made true by states $s \in S$ in the following way:
   $s \models\!\!\!\models \alpha$ iff $\forall w \in l(s)$: $w \models \alpha$
   (in such a case we also say that s is an α-state).
3. For every $\alpha \in \mathcal{L}$ let $\hat{\alpha} = \{s \in S \,|\, s \models\!\!\!\models \alpha\}$.
4. For every $\alpha \in \mathcal{L}$: $s \in \hat{\alpha}$ is *minimal in* $\hat{\alpha}$ iff $\neg\exists s' \in \hat{\alpha}$: $s' \prec s$.
5. The *Smoothness Condition*: Every state that makes α true is either itself most normal among the states which make α true, or there is a more normal state that makes α true and which is also most normal among the states that make α true; i.e.:
   $\forall \alpha \in \mathcal{L}$, $\forall s \in \hat{\alpha}$: s is minimal in $\hat{\alpha}$ or $\exists s' \prec s$, such that s' is minimal in $\hat{\alpha}$.
6. Relative to a cumulative model $\mathfrak{M} = \langle S, l, \prec \rangle$, we can define:
   $$\mathfrak{M} \models \alpha \Rightarrow \beta$$
   iff $\forall s \in S$: if s is minimal in $\hat{\alpha}$, then $s \models\!\!\!\models \beta$
   (i.e.: the most normal states among those that make α true also make β true, or: normal α are β).
7. Let $\mathcal{TH}_{\Rightarrow}(\mathfrak{M}) = \{\alpha \Rightarrow \beta \,|\, \mathfrak{M} \models \alpha \Rightarrow \beta\}$:
   $\mathcal{TH}_{\Rightarrow}(\mathfrak{M})$ is the *conditional theory corresponding to* $\mathfrak{M}$.
8. $\alpha \Rightarrow \beta$ is *cumulatively valid* iff
   for every cumulative model $\mathfrak{M}$: $\mathfrak{M} \models \alpha \Rightarrow \beta$.
9. Let $\mathfrak{M}$ be a cumulative model:
   $\mathfrak{M} \models KB_{\Rightarrow}$ iff for every $\alpha \Rightarrow \beta \in KB_{\Rightarrow}$ it holds that $\mathfrak{M} \models \alpha \Rightarrow \beta$.
10. We say that
    $$KB_{\Rightarrow} \models_c \alpha \Rightarrow \beta$$
    ($KB_{\Rightarrow}$ *cumulatively entails* $\alpha \Rightarrow \beta$) iff
    for every cumulative model $\mathfrak{M}$: if $\mathfrak{M} \models KB_{\Rightarrow}$, then $\mathfrak{M} \models \alpha \Rightarrow \beta$.

The additional types of models we will study are:

DEFINITION 45.

1. A *cumulative-ordered model* $\mathfrak{M}$ is a cumulative model $\langle S, l, \prec \rangle$, such that $\prec$ is a strict partial order, i.e. irreflexive and transitive.
2. A *preferential model* $\mathfrak{M}$ is a cumulative-ordered model $\langle S, l, \prec \rangle$, such that $\forall s \in S$: $l(s)$ is a singleton, i.e. $l(s) = \{w\}$ for some $w \in W$.
3. A *ranked model* $\mathfrak{M}$ is a preferential model $\langle S, l, \prec \rangle$, where for some $k \in \mathbb{N}$ there is a surjective mapping $rk : S \to \{0, \ldots, k\}$, such that for all $s_1, s_2 \in S$: $s_1 \prec s_2$ iff $rk(s_1) < rk(s_2)$
   ($rk(s)$ is called the 'rank' of s under rk).

Here is a diagram of what a typical ranked model looks like:



This model consists of three layers of worlds of equal rank. Within the set of α-states, the minimal ones are singled out, as they are taken to minimize "abnormality"; if these minimal α-states are all β-states, then $\alpha \Rightarrow \beta$ is considered satisfied by the model.

For each of these classes of models, the corresponding notions of satisfaction, determined conditional theories, validity (cumulative-ordered-valid, preferentially valid, rank-valid), and entailment ($\vDash_{co}$, $\vDash_p$, $\vDash_r$, i.e. cumulative-ordered-entails, preferentially entails, rank-entails) can be introduced in analogy with the case of cumulative models. The definition of ranked models in Lehmann and Magidor [1992] is actually more complex than ours, but our definition is equivalent for the case of a *finite* set W of worlds, and it is certainly more handy.

Obviously, the various kinds of entailment defined above come with strictly increasing strength, except for (see Lehmann and Magidor [1992]):

THEOREM 46.
$KB_{\Rightarrow}$ preferentially entails $\alpha \Rightarrow \beta$ iff $KB_{\Rightarrow}$ rank-entails $\alpha \Rightarrow \beta$.

As for validity, all notions of validity corresponding to the classes of models defined above coincide; indeed, they coincide with validity for *material* conditionals $\alpha \to \beta$.

REMARK 47. $\{\alpha \Rightarrow \beta \in \mathcal{L}_{\Rightarrow} \,|\, \alpha \to \beta \in \mathcal{TH}_{\to}(W)\}$ is a conditional theory of any of the defined types.

KLM [1990] and Lehmann and Magidor [1992] show the following soundness and completeness properties of systems of nonmonotonic logic with respect to preferential semantics:

THEOREM 48.

1. *(KLM [1990], pp.184–185)*

   $\mathcal{TH}_{\Rightarrow} \subseteq \mathcal{L}_{\Rightarrow}$ *is a consistent conditional C-theory extending* $\mathcal{TH}_{\rightarrow}$ *iff*

   there is a cumulative model $\mathfrak{M}$ *based on the set* W *of worlds satisfying* $\mathcal{TH}_{\rightarrow}$*, such that* $\mathcal{TH}_{\Rightarrow} = \mathcal{TH}_{\Rightarrow}(\mathfrak{M})$.

2. *(KLM [1990], p.189)*

   $\mathcal{TH}_{\Rightarrow} \subseteq \mathcal{L}_{\Rightarrow}$ *is a consistent conditional CL-theory extending* $\mathcal{TH}_{\rightarrow}$ *iff*

   there is a cumulative-ordered model $\mathfrak{M}$ *based on the set* W *of worlds satisfying* $\mathcal{TH}_{\rightarrow}$*, such that* $\mathcal{TH}_{\Rightarrow} = \mathcal{TH}_{\Rightarrow}(\mathfrak{M})$.

3. *(KLM [1990], p.196)*

   $\mathcal{TH}_{\Rightarrow} \subseteq \mathcal{L}_{\Rightarrow}$ *is a consistent conditional P-theory extending* $\mathcal{TH}_{\rightarrow}$ *iff*

   there is a preferential model $\mathfrak{M}$ *based on the set* W *of worlds satisfying* $\mathcal{TH}_{\rightarrow}$*, such that* $\mathcal{TH}_{\Rightarrow} = \mathcal{TH}_{\Rightarrow}(\mathfrak{M})$.

4. *(Lehmann and Magidor [1992], pp.21–23)*

   $\mathcal{TH}_{\Rightarrow} \subseteq \mathcal{L}_{\Rightarrow}$ *is a consistent conditional R-theory extending* $\mathcal{TH}_{\rightarrow}$ *iff*

   there is a ranked model $\mathfrak{M}$ *based on the set* W *of worlds satisfying* $\mathcal{TH}_{\rightarrow}$*, such that* $\mathcal{TH}_{\Rightarrow} = \mathcal{TH}_{\Rightarrow}(\mathfrak{M})$.

THEOREM 49. *Let* $\mathcal{TH}_{\rightarrow}$ *be the set of formulas satisfied by every world in the given set* W *of worlds.*

It holds:

1. $KB_{\Rightarrow} \vdash_{C}^{\mathcal{TH}_{\rightarrow}} \alpha \Rightarrow \beta$ *iff* $KB_{\Rightarrow} \vDash_{c} \alpha \Rightarrow \beta$.
2. $KB_{\Rightarrow} \vdash_{CL}^{\mathcal{TH}_{\rightarrow}} \alpha \Rightarrow \beta$ *iff* $KB_{\Rightarrow} \vDash_{co} \alpha \Rightarrow \beta$.
3. $KB_{\Rightarrow} \vdash_{P}^{\mathcal{TH}_{\rightarrow}} \alpha \Rightarrow \beta$ *iff* $KB_{\Rightarrow} \vDash_{p} \alpha \Rightarrow \beta$.
4. $KB_{\Rightarrow} \vdash_{R}^{\mathcal{TH}_{\rightarrow}} \alpha \Rightarrow \beta$ *(iff* $KB_{\Rightarrow} \vdash_{P}^{\mathcal{TH}_{\rightarrow}} \alpha \Rightarrow \beta$*) iff*
   $KB_{\Rightarrow} \vDash_{r} \alpha \Rightarrow \beta$.

Nonmonotonic Deductive Closure/Entailment

As we have seen in Subsection 2.2, deductive closure in nonmonotonic logic as understood above is actually monotonic, and by the results in the last subsection the same is true of the relations of logical entailment introduced above, i.e.: if $KB_{\Rightarrow} \vDash \alpha \Rightarrow \beta$ and $KB_{\Rightarrow} \subseteq KB'_{\Rightarrow}$, then also $KB'_{\Rightarrow} \vDash \alpha \Rightarrow \beta$ (with $\vDash$ being one of these entailment relations). However, there are also some strengthenings of logical entailment which are even nonmonotonic in the entailment sense:

E.g. Lehmann's and Magidor's [1992] rational closure operator (which is virtually identical to Pearl's [1990] so-called system Z) strengthens entailment by demanding truth preservation not in every ranked model in which a given conditional knowledge base is satisfied but only in those ranked models which maximize cautiousness and normality, in a sense that is made precise in Lehmann and Magidor [1992]. Goldszmidt, Morris, and Pearl [1993] maximum entropy approach and Lehmann's [1995] lexicographic entailment are further methods of nonmonotonic closure.

Some Complexity Considerations

While some of the consistency-based approaches to nonmonotonic reasoning, according to which exceptions to conditional defaults are stated explicitly (recall the introductory part of Subsection 2.2), do have nice implementations in terms of PROLOG or logic programs, nonmonotonic reasoning in the preferential KLM style is implemented in very much the same manner as standard systems of modal logic, and most of the complexity considerations concerning the latter (see standard textbooks on modal logic) carry over to the former.

Let φ be the conjunction of all entries in a (finite) factual knowledge base. It is to be decided whether $\varphi \Rightarrow \psi$ is entailed by the conditional knowledge base in one of the senses explained. One can show that this decision problem is co-NP-complete for preferential entailment and hence just as hard as the unsatisfiability decision problem for propositional logic (see Lehmann and Magidor [1992], p.16). However, as Lehmann and Magidor prove as well, the decision problem is polynomial in the case of Horn assertions. Lehmann and Magidor [1992], p.41, show that the decision procedure for rational closure is essentially as complex as the satisfiability problem for propositional logic. An excellent overview of such results can be found in Eiter and Lukasiewicz [2000]. One of the lessons to be drawn from these results is this: While progress in Nonmonotonic Reasoning has added to the expressive power of symbolic knowledge *representation*, it has not increased accordingly the *inferential power* of symbolic reasoning mechanisms by finding ways of improving their computational efficiency significantly.

The Interpretation of Conditionals Reconsidered

Computer scientists rarely address the question of what the exact interpretation of default conditionals of the form $\varphi \Rightarrow \psi$ ought to be. In particular, the following two sets of locutions are often not distinguished properly: on the one hand,

- *if φ then normally ψ*
- *if φ then it is very likely that ψ*

and, on the other,

- *normal φ are ψ*

- *by far most of the φ are ψ*

In the first set, φ and ψ are to be replaced by sentences such as 'Tweety is a bird.' and 'Tweety is able to fly.', whereas in the second set φ and ψ are to be substituted by generics such as 'birds' and 'flyers' (or, in a more formalized context, by open formulas such as 'x is a bird.' and 'x is able to fly.'). In the first set, $\Rightarrow$ is a sentential operator, while in the second set it is actually a generalized quantifier. (See van Benthem [1984] for more on this correspondence between conditionals and quantifiers; see Peters and Westerstahl [2006] for an extensive treatment of generalized quantifiers.) As far as preferential semantics is concerned, the set W of "possible worlds" is not so much a set of possible worlds in the second case but rather a universe of "possible objects" which are ordered by the normality of their occurrence. Accordingly, if a member of the first set were intended to express something probabilistic, then the probability measure in question should be a subjective probability measure by which rational degrees of belief are attributed to propositions, whereas in the case of the members of the second set, the corresponding probability measure should be a statistical one by which (limit) percentages are attributed to properties. For both sets of interpretation, the systems of nonmonotonic logic studied above are valid, but the application of these systems to actual reasoning tasks is still sensitive to the intended interpretation of $\varphi \Rightarrow \psi$.

2.3 Bridges

Now we are turning to formalisms and theories which are, in a sense to be explained, closely related to Nonmonotonic Reasoning.

The Bridge to the Logic of Counterfactuals

Amongst conditionals in natural language, usually the following distinction is made (this famous example is due to Ernest Adams):

1. If Oswald had not killed Kennedy, then someone else would have.

2. If Oswald did not kill Kennedy, then someone else did.

Sentence 2 is accepted by almost everyone, whilst we do not seem to know whether sentence 1 is true. This invites the following classification: A conditional such as sentence 2 is called *indicative*, a conditional like sentence 1 is called *subjunctive*. In conversation, the antecedents of subjunctive conditionals are often assumed or presupposed to be false: in such cases, one speaks of these subjunctive conditionals as *counterfactuals*. Subjunctive and indicative conditionals may have the same antecedents and consequents while differing only in their conditional connectives, i.e. their 'if'-'then' occurrences have different meanings. What both occurrences of 'if'-'then' in these examples have in common, however, is that they are nonmonotonic: E.g. the indicative 'If it rains, I will give you an umbrella.' does

not seem to logically imply "If it rains and I am in prison, I will give you an umbrella.', nor does the subjunctive 'If it rained, I would give you an umbrella.' seem to logically imply "If it rained and I were in prison, I would give you an umbrella.'. Accordingly, add e.g. '...and Kennedy in fact survived all attacks on his life.' to the antecedent of 'If Oswald did not kill Kennedy, then someone else did.' and the resulting conditional does not seem acceptable anymore. Therefore, philosophical logicians started to investigate new logical systems in which Monotonicity or Strengthening of the Antecedent is not logically valid. For a nice and recent introduction into this topic, presented from the viewpoint of the philosophy of language, see Bennett [2003].

We will consider the logic of indicative conditionals in our subsection on Probabilistic Logic below, but for now we are going to focus on subjunctive conditionals. D. Lewis [1973a], [1973b] famously introduced a semantics for subjunctive conditionals which we will state more or less precisely (compare Stalnaker's related semantics in Stalnaker [1991]). Reconsider 'If Oswald had not killed Kennedy, then someone else would have.': according to Lewis, this counterfactual says something — in this case: something false — about the world: If the world had been such that Oswald had not killed Kennedy, *but otherwise it would have been as similar as possible to what our actual world is like*, then someone else would have killed Kennedy in that world. However, if we consider all the possible worlds in which Oswald did not kill Kennedy, and if we focus just on those worlds among them which are maximally similar to our actual world, then it seems we only end up with worlds in which no one killed Kennedy at all — that is exactly why we tend to think that 'If Oswald had not killed Kennedy, then someone else would have' is false.

Now let us make this intuition about subjunctive conditionals formally precise. Lewis' semantics does so by introducing the following "ingredients":

- We focus on a language $\mathcal{L}$ that is closed under the following syntactic rules:
  - If A is in $\mathcal{L}$, then $\neg A$ is in $\mathcal{L}$.
  - If A is in $\mathcal{L}$ and B is in $\mathcal{L}$, then $(A \vee B)$ is in $\mathcal{L}$.
  - If A is in $\mathcal{L}$ and B is in $\mathcal{L}$, then $(A \wedge B)$ is in $\mathcal{L}$.
  - If A is in $\mathcal{L}$ and B is in $\mathcal{L}$, then $(A \rightarrow B)$ is in $\mathcal{L}$.
  - If A is in $\mathcal{L}$ and B is in $\mathcal{L}$, then $(A > B)$ is in $\mathcal{L}$.

  (The last clause is for subjunctive conditionals.)

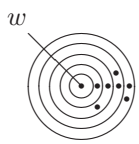
- We choose a non-empty set W, which we call the set of possible worlds.

- We assume that we can "measure" the closeness or similarity of worlds to any world w in W. Formally, this can be done by assuming that for every world w there is a sphere system $\mathfrak{S}_w$ of "spheres" around w, i.e. a class $\mathfrak{S}_w$ of subsets of W, such that the following two conditions are satisfied:

- $\{w\}$ is a sphere in $\mathfrak{S}_w$,
- if X and Y are spheres in $\mathfrak{S}_w$, then either X is a subset of Y or Y is a subset of X.

(Lewis considers further conditions on systems of spheres, but we restrict ourselves just to the most relevant ones.)

So, a sphere system $\mathfrak{S}_w$ around w looks like this:



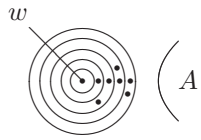
Intuitively, such a sphere system is meant to express:

- If X is a sphere in $\mathfrak{S}_w$ and w' is a member of X, then w' is closer or more similar to w than all those worlds in W that are not in X.
- If w' is not a member of any sphere around w — formally: w' is not a member of the union $\bigcup \mathfrak{S}_w$ of all spheres around w — then w' is not possible relative to w.

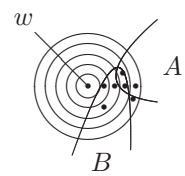
- Finally, we consider a mapping V that maps each formula A in $\mathcal{L}$ and each world w in W to a truth value in $\{0, 1\}$ according to the following semantic rules:
  - $V(\neg A, w) = 1$ if and only if $V(A, w) = 0$.
  - $V(A \vee B, w) = 1$ if and only if $V(A, w) = 1$ or $V(A, w) = 1$.
  - $V(A \wedge B, w) = 1$ if and only if $V(A, w) = 1$ and $V(A, w) = 1$.
  - $V(A \rightarrow B, w) = 1$ if and only if $V(A, w) = 0$ or $V(A, w) = 1$.
  - The truth condition for subjunctive conditionals:

    $V(A > B, w) = 1$ if and only if either of the following two conditions is satisfied:

    * There is no A-world in $\bigcup \mathfrak{S}_w$, i.e. for all worlds w' in $\bigcup \mathfrak{S}_w$: $V(A, w') = 0$.



* There is a sphere X in $\mathfrak{S}_w$, such that (i) for at least one world w' in X it holds that $V(A, w') = 1$, and (ii) for all worlds w' in X it holds that: $V(A \to B, w') = 1$.



- Summing up: We call $\langle W, (\mathfrak{S}_w)_{w \in W}, V \rangle$ a *(Lewis-)spheres model for subjunctive conditionals* if and only if all of the conditions above are satisfied. (By means of '$(\mathfrak{S}_w)_{w \in W}$' we denote the family of sphere systems for worlds w in W within a given spheres model.)

- We call a formula φ in $\mathcal{L}$ *logically true* (according to the spheres semantics) if and only if φ is true at every world in every spheres model. Accordingly, an argument $P_1, \ldots, P_n \therefore C$ is called *logically valid* — equivalently: C *follows logically* from $P_1, \ldots, P_n$ — if and only if $(P_1 \wedge \ldots \wedge P_n) \to C$ is logically true.

Given further constraints on such models, the truth condition for subjunctive conditionals can be simplified:

- $\langle W, (\mathfrak{S}_w)_{w \in W}, V \rangle$ satisfies the *Limit Assumption* if and only if

  for every world w in W, and for every A in $\mathcal{L}$ for which $\bigcup \mathfrak{S}_w$ contains at least one A-world, it holds that there is a least sphere X in $\mathfrak{S}_w$ that includes a world w' for which $V(A, w') = 1$ is the case.
  ('Least' implies that every sphere that is a proper subset of X does not contain any A-world at all.)

- If $\langle W, (\mathfrak{S}_w)_{w \in W}, V \rangle$ satisfies the Limit Assumption, then Lewis' truth condition for $A > B$ reduces to:

  $V(A > B, w) = 1$ if and only if either of the following two conditions is satisfied:

  - There is no A-world in $\bigcup \mathfrak{S}_w$, i.e. for all w' in $\bigcup \mathfrak{S}_w$: $V(A, w') = 0$.
  - If X_{least} is the least A-*permitting* sphere, i.e. the least sphere X in $\mathfrak{S}_w$ for which it is the case that for some world w' in X it holds that $V(A, w') = 1$, then for all worlds w' in X_{least} it is the case that $V(A \to B, w') = 1$.
    (In words: B holds at all closest A-worlds.)

Obviously, Lewis' sphere systems with the Limit Assumption are very similar to ranked models with their Smoothness Assumption that we have discussed above,

and indeed one can be viewed as a notational variant of the other (modulo some minor differences such as the existence of a unique "most normal" world in Lewis' semantics). Accordingly, the satisfaction clause for counterfactuals $A > B$ in the one case mimics the satisfaction clause for nonmonotonic conditionals $\alpha \Rightarrow \beta$ in the other.

Lewis [1973a], [1973b] showed the following soundness and completeness result:

THEOREM 50. *The system VC of conditional logic (see below) is sound and complete with respect to the spheres semantics for subjunctive conditionals.*

- Rules of VC:

  1. Modus Ponens (for $\rightarrow$)
  2. Deduction within subjunctive conditionals: for any $n \geq 1$

$$\frac{\vdash (B_1 \wedge \ldots \wedge B_n) \rightarrow C}{\vdash ((A > B_1) \wedge \ldots \wedge (A > B_n)) \rightarrow (A > C)}$$

  3. Interchange of logical equivalents

- Axioms of VC:

  1. Truth-functional tautologies
  2. $A > A$
  3. $(\neg A > A) \rightarrow (B > A)$
  4. $(A > \neg B) \vee (((A \wedge B) > C) \leftrightarrow (A > (B \rightarrow C)))$

  C1 Weak Centering: $(A > B) \rightarrow (A \rightarrow B)$

  C2 Centering: $(A \wedge B) \rightarrow (A > B)$

'V' stands for 'Variably strict', which reflects that one can think of subjunctive conditionals as strict conditionals $\Box(A \rightarrow B)$, but with variably strict degrees of necessity; 'C' is short for the 'Centering axioms' C1 and C2 (or semantically for assuming that $\{w\}$ is a sphere in $\mathfrak{S}_w$).

Unsurprisingly, if the logical consequence relation is restricted to counterfactuals of the form $A > B$ with A and B not containing $>$, i.e. if only the so-called "flat" fragment of Lewis' logic for counterfactuals is considered, the system P of nonmonotonic logic re-emerges. So, on the formal level, the main difference between the logic of counterfactuals and nonmonotonic logic turns out to be a syntactical one: while the former allows for nestings of conditionals and also for the application of propositional connectives to conditionals, the latter does not.

The Bridge to Belief Revision

AGM [1985] — short for: Alchourrón, Gärdenfors, Makinson — and Gärdenfors [1988] have developed a now well-established theory of belief revision, i.e. a theory which states and justifies rationality constraints on how agents ought to revise their beliefs in the light of new information. In this theory a belief state of an agent is considered as a set of formulas, i.e. the set of formulas the agent believes to be true. Furthermore, agents are assumed to be rational in the sense that their belief sets are deductively closed. So we have:

- Belief set G: a deductively closed set of formulas

Now an agent is taken to receive some new evidence, where evidence is regarded as being given by a formula again:

- Evidence A: a formula

Formally, the agent's revision of her belief set G on the basis of her new evidence A is supposed to lead to a new belief set that is denoted by '$G * A$':

- Revised belief set $G * A$: a deductively closed set of formulas

The corresponding function $*$, which maps a pair of a formula and a set of formulas to a further set of formulas, is called the "revision operator".

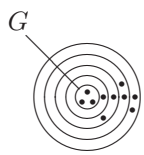
How is this revision process considered to take place? In principle, there are two possible cases to consider:

- (Consistency Case) If A is consistent with G, then it is rational for the agent to simply add A to G, whence $G * A$ will presumably be simply $G \cup \{A\}$ together with all of its logical consequences.

- (Inconsistency Case) If A is inconsistent with G, then in order to revise G by A the agent has to give up some of her beliefs; she does so rationally, or so Quine, Gärdenfors, and others have argued, if she follows a *principle of minimal mutilation*, i.e. she gives up as few of her old beliefs as possible.

This guiding idea does not necessarily determine $G * A$ uniquely, but it yields rational constraints on the belief revision operator $*$ which can be stated as axioms. Such axioms have indeed been suggested in the theory of belief revision, and they have been studied systematically in the last two decades; they are usually referred to as the 'AGM axioms'. We will not state these axioms here, since their standard presentation is too far removed from nonmonotonic inference relations or conditionals: for more information see the references given above (moreover, Hansson [1999] is a recent textbook on belief revision). But even independently of the exact details of the axiomatic treatment of belief revision, it is clear that belief revision operators may be expected not to be monotonic in view of the Inconsistency case from above: if A logically implies B, then $G * A$ is by no means guaranteed to be logically stronger than, i.e. a superset of $G * B$.

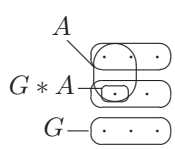
As Grove [1988] has shown, belief revision operators can be characterized semantically by a sphere semantics that is "almost" like Lewis' sphere semantics for subjunctive conditionals and which is more or less identical to the ranked model semantics for nonmonotonic conditionals. Without going into the formal details, this is the main idea: Whereas in Lewis' semantics the innermost sphere of a sphere system around a world w contains exactly one world, namely w (Centering), Grove's sphere systems are not "around" particular worlds at all, and consequently the innermost sphere of a sphere system might contain more than one world. Indeed, the set of formulas which are true in all worlds of the innermost sphere is regarded as the "original" unrevised belief set G in a sphere system:



Intuitively, such a sphere system is meant to express:

- If X is a sphere and w' is a member of X, then w' is more plausible to be a candidate for being the actual world than all those worlds that are not in X. The spheres themselves correspond to epistemic "fallback positions" that are supposed to kick in if new evidence contradicts the current belief set G.
- If w' is not a member of any sphere, then w' is not regarded epistemically possible.

Alternatively, one can use a graphical representation along the lines of ranked models: Instead of proper spheres, one has layers or ranks again; the lowest layer corresponds to the innermost sphere, while taking the union of the lowest layer with the second layer from below corresponds to the next larger sphere, and so forth. G is the set of formulas that are true in all worlds which are members of the lowest layer; $G * A$ is the set of formulas which are satisfied by all those worlds that have minimal rank among the worlds that satisfy A:



For every such sphere system $\mathfrak{S}$ in Grove's sense, i.e. every class $\mathfrak{S}$ of subsets of W satisfying Lewis' assumptions on spheres except for the "centeredness on worlds" assumption, a corresponding belief revision operator $*_{\mathfrak{S}}$ can be defined in much the same way as the truth conditions for subjunctive conditionals are

determined in Lewis' semantics and as the satisfaction conditions for nonmonotonic conditionals are stated in preferential semantics:

$B \in G *_{\mathfrak{S}} A$ if and only if either of the following two conditions is satisfied:

- There is no A-world in the union $\bigcup \mathfrak{S}$ of spheres in $\mathfrak{S}$, i.e. for all worlds w' in $\bigcup \mathfrak{S}$: A is false in w'.
- There is a sphere X in $\mathfrak{S}$, such that (i) for at least one world w' in X it holds that A is true in w', and (ii) for all worlds w' in X it holds that: $A \to B$ is true in w'.

Grove [1988] proved the following theorem:

THEOREM 51.

- *For every sphere system $\mathfrak{S}$ in Grove's sense (with W being the set of all truth value assignments over $\mathcal{L}$), the corresponding operator $*_{\mathfrak{S}}$ is a belief revision operator, i.e. it satisfies the AGM axioms.*
- *For every belief revision operator $*$ satisfying the AGM axioms, there is a sphere system $\mathfrak{S}$ in Grove's sense (with W being the set of all truth value assignments over $\mathcal{L}$), such that for all $A, B \in \mathcal{L}$:*

$$B \in G * A \text{ iff } B \in G *_{\mathfrak{S}} A$$

Clearly, the semantics of belief revision operators, Lewis' semantics of subjunctive conditionals, and ranked model semantics of nonmonotonic logic share a lot of formal structure. Accordingly, there are translation results from belief revision into nonmonotonic logic (as well as Lewis' logic) and *vice versa*; see e.g. Gärdenfors and Makinson [1994]. Expressions of the form

$$B \in G * A$$

are mapped thereby to expressions of the form

$$\alpha \Rightarrow \beta \in \mathcal{TH}_{\Rightarrow}$$

However, the intended philosophical interpretations of these logical frameworks differ of course: In particular, counterfactuals are meant to express something *ontic*, belief revision operators are meant to be *epistemic*, and nonmonotonic conditionals are best regarded open to both understandings.

The Bridge to Probabilistic Logic

Since the nonmonotonicity phenomenon was already well known in probability theory — a conditional probability $P(Y|X)$ being high does not entail the conditional probability $P(Y|X \cap Z)$ being high — it is not surprising that some of the modern accounts of nonmonotonic conditionals turn out to rely on a probabilistic

semantics. Let us go back to Adams' example of an indicative conditional, i.e. 'If Oswald did not kill Kennedy, then someone else did.'. According to Adams [1975], asserting such an indicative conditional aims at expressing that one's subjective conditional probability of 'Someone other than Oswald killed Kennedy.' given that 'Oswald did not kill Kennedy.' is high. Adams famously developed a non-truth-conditional semantics along these lines, which we will sketch below. A more recent introduction to this type of probabilistic logic is given by Adams [1998]; Pearl [1988] nicely builds on, and extends, Adams' original theory.

Let $\mathcal{L}$ be the language of propositional logic. We state a probabilistic semantics for two types of formulas: (i) formulas $A, B, C, D, E, F, \ldots$ of $\mathcal{L}$, (ii) formulas of the form $B \Rightarrow C$, where B and C are members of $\mathcal{L}$ (so we disregard again both nestings of conditionals and propositional constructions from conditionals). The formulas in (ii) are meant to represent indicative conditionals.

By a probability measure on $\mathcal{L}$ we mean the following:

DEFINITION 52. A *probability measure on* $\mathcal{L}$ is a function P with the following properties:

1. $P : \mathcal{L} \to [0, 1]$, i.e.: P maps each sentence in $\mathcal{L}$ to a real number x, such that $0 \leq x \leq 1$.

2. For all $A, B \in \mathcal{L}$: If A is logically equivalent to B, then $P(A) = P(B)$.

3. For all $A, B \in \mathcal{L}$: If $A \wedge B \vDash \bot$, then $P(A \vee B) = P(A) + P(B)$. That is: If two sentences are inconsistent with each other, then the probability of their disjunction equals the sum of their probabilities.

4. For all $A \in \mathcal{L}$: If A is logically true, then $P(A) = 1$.

(The axioms are not meant to be independent of each other.)

Additionally, conditional probabilities can be introduced by means of the so-called "Ratio Formula":

- For all $A \in \mathcal{L}$ with $P(A) > 0$,

$$P(B|A) = \frac{P(B \wedge A)}{P(A)}$$

where the 'P' on the left hand side denotes the conditional probability measure that belongs to, or corresponds to, the unconditional probability measure that is denoted by 'P' on the right hand side.

Now we consider arguments of either of the following two forms:

$$\begin{array}{l} A_1 \\ \vdots \\ A_m \\ B_1 \Rightarrow C_1 \\ \vdots \\ B_n \Rightarrow C_n \\ \hline D \end{array} \qquad\qquad \begin{array}{l} A_1 \\ \vdots \\ A_m \\ B_1 \Rightarrow C_1 \\ \vdots \\ B_n \Rightarrow C_n \\ \hline E \Rightarrow F \end{array}$$

According to Adams' semantics, such arguments are called *probabilistically valid* if and only if for all infinite sequences $P_1, P_2, P_3, \ldots$ of subjective probability measures on $\mathcal{L}$ the following is the case:

If		If	
$P_i(A_1)$	tends to 1 for $i \to \infty$,	$P_i(A_1)$	tends to 1 for $i \to \infty$,
$\vdots$		$\vdots$	
$P_i(A_m)$	tends to 1 for $i \to \infty$,	$P_i(A_m)$	tends to 1 for $i \to \infty$,
$P_i(C_1\|B_1)$	tends to 1 for $i \to \infty$,	$P_i(C_1\|B_1)$	tends to 1 for $i \to \infty$,
$\vdots$		$\vdots$	
$P_i(C_n\|B_n)$	tends to 1 for $i \to \infty$,	$P_i(C_n\|B_n)$	tends to 1 for $i \to \infty$,
then		then	
$P_i(D)$	tends to 1 for $i \to \infty$	$P_i(F\|E)$	tends to 1 for $i \to \infty$

where if $P_i(\varphi) = 0$ then $P_i(\psi|\varphi)$ is regarded to be equal to 1.

REMARK 53. It is possible to omit this last extra clause if conditional probability measures — so-called *Popper functions* — are used from the start, rather than having conditional probabilities determined by absolute probabilities through the standard ratio formula. More about this may be found in McGee [1994], Hájek [2003], and Halpern [2001a].

So put in a slogan: An argument is valid according to Adams' probabilistic semantics if and only if the more certain the premises, the more certain the conclusion.

Adams [1975] showed the following soundness and completeness result:

THEOREM 54. *The following list of rules is sound and complete with respect to probabilistic validity:*

- *In case A logically implies B:*

  $$\frac{}{A \Rightarrow B} \text{ (Supraclassicality)}$$

- *In case A is logically equivalent with A':*

  $$\frac{A \Rightarrow B}{A' \Rightarrow B} \text{ (Left Logical Equivalence)}$$

- $\dfrac{\top \Rightarrow A}{A}$ *(Trivial Antecedent 1)*

  where $\top$ *is any propositional tautology*

- $\dfrac{A}{\top \Rightarrow A}$ *(Trivial Antecedent 2)*

  where $\top$ *is any propositional tautology*

- $\dfrac{\begin{array}{c} A \Rightarrow B \\ A \Rightarrow C \end{array}}{A \wedge B \Rightarrow C}$ *(Cautious Monotonicity)*

- $\dfrac{\begin{array}{c} A \Rightarrow B \\ A \wedge B \Rightarrow C \end{array}}{A \Rightarrow C}$ *(Cautious Cut)*

- $\dfrac{\begin{array}{c} A \Rightarrow C \\ B \Rightarrow C \end{array}}{A \vee B \Rightarrow C}$ *(Disjunction)*

These rules are to be understood in the way that if one has derived the premises of any of these rules from a set of factual or conditional assumptions, then one may also derive the conclusion of that rule from the same set of assumptions.

Once again, one can show that neither Contraposition nor Transitivity nor Monotonicity is probabilistically valid. Indeed, Adams' logic of indicative conditionals is nothing else but the system P of nonmonotonic logic that has been discussed above. This probabilistic style of doing nonmonotonic reasoning has become quite prominent in the meantime (see e.g. Lukasiewicz [2002]) and connects Nonmonotonic Reasoning to an area that is sometimes referred to as 'Uncertain Reasoning' (see Paris [1994] for an excellent introduction into all formal aspects of uncertain reasoning).

As Adams himself has observed (see also Snow [1999]), an equivalent probabilistic semantics can be given in terms of so-called "probabilistic orders of magnitude" which replace the qualitative degrees of normality in preferential semantics. (See Schurz [2001] for a philosophical investigation into the conceptual differences between qualitative and statistical notions of normality.) Lehmann and Magidor [1992], pp.48–53, suggest a probabilistic semantics for their system R in terms of probability measures which allow for nonstandard real number values. See McGee [1994], Hawthorne [1996], Bamber [2000], Halpern [2001a], Arlo-Costa and Parikh [2005] for further probabilistic accounts of conditionals, nonmonotonic inference relations, and even nonmonotonic deductive closure or entailment.

The Bridge to Neural Network Semantics

Interpreted dynamical systems — the paradigm instances of which are *artificial neural networks* that come with a logical interpretation — may also be used to

yield a semantics for nonmonotonic conditionals. Here are some relevant references: d'Avila Garcez, Lamb, and Gabbay [2008] give a general overview of connectionist non-classical logics, including connectionist (i.e. neural networks-related) nonmonotonic logic, as well as lots of references to their own original work. Balkenius [1991], Blutner [2004], and Leitgeb [2001], [2004], [2005] are important primary references. The main idea behind all of these theories is that if classical logic is replaced by some system of nonmonotonic reasoning, then a logical description or characterization of neural network states and processes becomes possible. The following exposition will introduce Leitgeb's approach which yields a neural network semantics for KLM-style systems; the presentation will follow the more detailed introduction to neural network semantics for conditionals in Leitgeb [2007].

The goal is to complement the typical description of neural networks as dynamical systems by one according to which cognitive dynamical systems have beliefs, draw inferences, and so forth. Hence, the task is to associate *states* and *processes* of cognitive dynamical systems with *formulas*. Here is what we will presuppose: We deal with discrete dynamical systems with a set S of states. On S, a partial order $\leqslant$ is defined, which we will interpret as an ordering of the amount of information that is carried by states; so $s \leqslant s'$ will mean: s' carries at least as much information as s does. We will also assume that for every two states s and s' there is a uniquely determined state $sup(s, s')$ which (i) carries at least as much information as s, which also (ii) carries at least as much information as s', and which (iii) is the state with the least amount of information among all those states for which (i) and (ii) hold. Formally, such a state $sup(s, s')$ is the *supremum* of s and s' in the partial order $\leqslant$. Finally, an internal next-state function is defined for the dynamical system, where this next-state function is meant to be insensitive to possible external inputs to the system; we will introduce inputs only in the subsequent step.

In this way, we get what is called an 'ordered discrete dynamical system' in Leitgeb [2005]:

DEFINITION 55. An *ordered discrete dynamical system* is a triple $\mathcal{S} = \langle S, ns, \leqslant \rangle$, such that:

1. S is a non-empty set (the set of states).
2. $ns : S \to S$ (the internal next-state function).
3. $\leqslant \subseteq S \times S$ is a partial order (the information ordering) on S,

   such that for all s, $s' \in S$ there is a supremum $sup(s, s') \in S$ with respect to $\leqslant$.

In case an artificial neural network is used, the information ordering on its states, i.e. on its possible patterns of activation, can be defined according to the following idea: the more the nodes are activated in a state, the more information

the state carries. Accordingly, $sup(s, s')$ would be defined as the maximum of the activation patterns that correspond to s and s'; in such a case one might also speak of $sup(s, s')$ as the "superposition of the states s and s'". (But note that this is just one way of viewing neural networks as ordered systems.) The internal dynamics of the network would be captured by the next-state mapping ns that is determined by the pattern of edges in the network.

Next, we add external inputs which are regarded to be represented by states $s^* \in S$ and which are considered to be fixed for a sufficient amount of time. The state transition mapping F_{s^*} can then be defined by taking both the internal next-state mapping and the input s^* into account: The next state of the system is given by the superposition of s^* with the next internal state $ns(s)$, i.e.:

$$F_{s^*}(s) := sup(s^*, ns(s))$$

The dynamics of our dynamical systems is thus determined by iteratively applying F_{s^*} to the initial state. Fixed points s_{stab} of F_{s^*} are regarded to be the "answers" which the system gives to s^*, as it is common procedure in neural network computation. Note that in general there may be more than just one such *stable state* for the state transition mapping F_{s^*} that is determined by the input s^* (and by the given dynamical system), and there may also be no stable state at all for F_{s^*}: in the former case, there is more than just one "answer" to the input, in the latter case there is no "answer" at all. The different stable states may be reached by starting the computation in different initial states of the overall system.

Now formulas can be assigned to the states of an ordered discrete dynamical system. These formulas are supposed to express the content of the information that is represented by these states. For this purpose, we fix a propositional language $\mathcal{L}$. The assignment of formulas to states is achieved by an interpretation mapping $\mathfrak{I}$. If φ is a formula in $\mathcal{L}$, then $\mathfrak{I}(\varphi)$ is the state that carries exactly the information that is expressed by φ, i.e. neither less nor more than what is expressed by φ. So we presuppose that for every formula in $\mathcal{L}$ there is a uniquely determined state the total information of which is expressed by that formula. If expressed in terms of belief, we can say that in the state $\mathfrak{I}(\varphi)$ *all the system believes is that* φ, i.e. the system only believes φ and all the propositions which are contained in φ from the viewpoint of the system. (This relates to Levesque's [1990] modal treatment of the 'all I know' operator.) We will not demand that every state necessarily receives an interpretation but just that every formula in $\mathcal{L}$ will be the interpretation of some state. Furthermore, not just any assignment whatsoever of states to formulas will be allowed, but we will additionally assume certain postulates to be satisfied which will guarantee that $\mathfrak{I}$ is compatible with the information ordering that was imposed on the states of the system beforehand. An ordered discrete dynamical system together with such an interpretation mapping is called an 'interpreted ordered system' (cf. Leitgeb [2005]). This is the definition in detail:

DEFINITION 56. An *interpreted ordered system* is a quadruple $\mathcal{S}_{\mathfrak{I}} = \langle S, ns, \leqslant, \mathfrak{I} \rangle$, such that:

1. $\langle S, ns, \leqslant \rangle$ is an ordered discrete dynamical system.

2. $\mathfrak{I} : \mathcal{L} \to S$ (the interpretation mapping) is such that the following postulates are satisfied:

   (a) Let $\mathcal{TH}_{\mathfrak{I}} = \{\varphi \in \mathcal{L} \,|\text{for all } \psi \in \mathcal{L}\text{: } \mathfrak{I}(\varphi) \leqslant \mathfrak{I}(\psi)\}$:
   then it is assumed that for all $\varphi, \psi \in \mathcal{L}$: if $\mathcal{TH}_{\mathfrak{I}} \vdash \varphi \to \psi$, then $\mathfrak{I}(\psi) \leqslant \mathfrak{I}(\varphi)$.

   (b) For all $\varphi, \psi \in \mathcal{L}$: $\mathfrak{I}(\varphi \wedge \psi) = sup(\mathfrak{I}(\varphi), \mathfrak{I}(\psi))$.

   (c) For every $\varphi \in \mathcal{L}$: there is an $\mathfrak{I}(\varphi)$-stable state.

   (d) There is an $\mathfrak{I}(\top)$-stable state s_{stab}, such that $\mathfrak{I}(\bot) \nleq s_{stab}$.

We say that $\mathcal{S}_{\mathfrak{I}}$ satisfies the uniqueness condition if for every $\varphi \in \mathcal{L}$ there is precisely one $\mathfrak{I}(\varphi)$-stable state.

E.g., postulate 2b expresses that the state that belongs to a conjunctive formula $\varphi \wedge \psi$ ought to be the supremum of the two states that are associated with the two conjuncts φ and ψ: this is the cognitive counterpart of the proposition expressed by a conjunctive sentence being the supremum of the propositions expressed by its two conjuncts in the partial order of logical entailment. For a detailed justification of all the postulates, see Leitgeb [2005].

Finally, we define what it means for a *nonmonotonic* conditional to be satisfied by an interpreted ordered system. We say that a system satisfies $\varphi \Rightarrow \psi$ if and only if whenever the state that is associated with φ is fed into the system as an input, i.e. whenever the input represents a total belief in φ, the system will eventually end up believing ψ in its "answer states", i.e. the state that is associated with ψ is contained in all the states which are stable with respect to this input. Collecting all such conditionals $\varphi \Rightarrow \psi$ which are satisfied by the system, we get what we call the 'conditional theory' that corresponds to the system.

DEFINITION 57. Let $\mathcal{S}_{\mathfrak{I}} = \langle S, ns, \leqslant, \mathfrak{I} \rangle$ be an interpreted ordered system:

1. $\mathcal{S}_{\mathfrak{I}} \vDash \varphi \Rightarrow \psi$ iff for every $\mathfrak{I}(\varphi)$-stable state s_{stab}: $\mathfrak{I}(\psi) \leqslant s_{stab}$.

2. $\mathcal{TH}_{\Rightarrow}(\mathcal{S}_{\mathfrak{I}}) = \{\varphi \Rightarrow \psi \,|\mathcal{S}_{\mathfrak{I}} \vDash \varphi \Rightarrow \psi\}$
   (the *conditional theory corresponding to* $\mathcal{S}_{\mathfrak{I}}$).

Leitgeb [2005] proves the following soundness and completeness theorem:

THEOREM 58.

- *Let $\mathcal{S}_{\mathfrak{I}} = \langle S, ns, \leqslant, \mathfrak{I} \rangle$ be an interpreted ordered system which satisfies the Uniqueness Assumption:*

  Then $\mathcal{TH}_{\Rightarrow}(\mathcal{S}_{\mathfrak{I}})$ is a consistent conditional C-theory extending $\mathcal{TH}_{\mathfrak{I}}$.

- *Let $\mathcal{TH}_{\Rightarrow}$ be a consistent conditional C-theory extending a given classical theory $\mathcal{TH}_{\rightarrow}$:*

  It follows that there is an interpreted ordered system $\mathcal{S}_{\mathfrak{I}} = \langle S, ns, \leqslant, \mathfrak{I} \rangle$, such that $\mathcal{TH}_{\Rightarrow}(\mathcal{S}_{\mathfrak{I}}) = \mathcal{TH}_{\Rightarrow}$, $\mathcal{TH}_{\mathfrak{I}} \supseteq \mathcal{TH}_{\rightarrow}$, and $\mathcal{S}_{\mathfrak{I}}$ satisfies the uniqueness condition.

These results can be extended into various directions. In particular, some interpreted ordered systems can be shown to have the property that each of their states s may be decomposed into a set of substates s_i which can be ordered in a way such that the dynamics for each substate s_i is determined by the dynamics for the substates $s_1, s_2, \ldots, s_{i-1}$ at the previous point of time. Such systems are called 'hierarchical' in Leitgeb [2005]. We will not go into any details, but one can prove soundness and completeness theorems for such *hierarchical* interpreted systems and the system CL. In Leitgeb [2004] further soundness and completeness theorems are proved for more restricted classes of interpreted dynamical systems and even stronger logical systems for nonmonotonic conditionals in the KLM tradition.

As it turns out, if artificial neural networks with an information ordering are extended by an interpretation mapping along the lines explained above, then they are special cases of interpreted ordered systems; moreover, if the underlying artificial neural network consists of layers of nodes, such that the layers are arranged hierarchically and all connections between nodes are only from one layer to the next one, then the corresponding interpreted ordered system is a hierarchical one. Thus, various systems of nonmonotonic logic are sound and complete with respect to various types of neural network semantics. However, so far these results only cover the short-term dynamics of neural networks that is triggered by external input and for which the topology of edges and the distribution of weights over the edges within the network is taken to be rigid. The long-term dynamics of networks given e.g. by supervised learning processes which operate on sequences of input-output pairs is still beyond any logical treatment that is continuous with KLM-style nonmonotonic reasoning. So, the inductive logic of *learning*, rather than inference, within neural networks is still an open research problem (see Leitgeb [2007] for a detailed statement of this research agenda).

The Bridge to Philosophy of Science

In traditional general philosophy of science, the nonmonotonicity phenomenon is well-known from inductive logic and the theory of statistical explanation. In order to cope with it, Carnap introduced his "requirement of total evidence": an inductive argument should only be applied by an agent if its premises comprise the agent's total knowledge; in the nonmonotonic reasoning context we saw this principle at work already in the introduction to Section 2. Hempel improved Carnap's rule by the related "rule of maximal specificity"; for a discussion of both rules see Stegmüller [1969], Chapter IX; for more on Carnap and Hempel see Zabell

[2009] and Sprenger [2009]. In the meantime, progress in Nonmonotonic Reasoning has started to feed back into philosophy of science. E.g.: Flach [2004] argues that the same logics that govern valid commonsense inferences can be interpreted as logics for scientific induction, i.e. for data constituting incomplete and uncertain evidence for empirical hypotheses. His formal account of scientific confirmation relations is modelled after the KLM approach to nonmonotonic inference relations. Schurz [2002] suggests to take system P of nonmonotonic logic to be the logic of *ceteris paribus* laws in science, i.e. laws that are meant to hold only in normal or standard conditions. More such bridges to philosophy of science may be expected to emerge.

ACKNOWLEDGMENTS

Ronald Ortner and Hannes Leitgeb would like to thank each other.

BIBLIOGRAPHY

[Adams, 1975] E. W. Adams. *The Logic of Conditionals*. D. Reidel, Dordrecht, 1975.

[Adams, 1998] E. W. Adams. *A Primer of Probability Logic*. CSLI Publications, Stanford, 1998.

[Alchourrón, Gärdenfors and Makinson, 1985] C. E. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50:510–530, 1985.

[Alon and Asodi, 2005] N. Alon and V. Asodi. Learning a hidden subgraph. *SIAM Journal on Discrete Mathematics*, 18(4):697–712, 2005.

[Angluin and Chen, 2008] D. Angluin and J. Chen. Learning a hidden graph using O(log n) queries per edge. *Journal of Computer and System Sciences*, 74(4):546–556, 2008.

[Angluin and Smith, 1983] D. Angluin and C. H. Smith. Inductive inference: Theory and methods. *ACM Computing Surveys*, 15(3):237–269, 1983.

[Angluin, 1992] D. Angluin. Computational learning theory: Survey and selected bibliography. In *Proceedings of the Twenty Fourth Annual ACM Symposium on Theory of Computing (STOC), 4-6 May 1992, Victoria, British Columbia, Canada*, pages 351–369. ACM, 1992.

[Angluin, 2004] D. Angluin. Queries revisited. *Theoretical Computer Science*, 313(2):175–194, 2004.

[Argamon and Shimoni, 2003] S. Argamon and A. R. Shimoni. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17:401–412, 2003.

[Arló-Costa and Parikh, 2005] H. Arló-Costa and R. Parikh. Conditional probability and defeasible inference. *Journal of Philosophical Logic*, 34:97–119, 2005.

[Assouad, 1983] P. Assouad. Densité et dimension. *Université de Grenoble. Annales de l'Institut Fourier*, 33(3):233–282, 1983.

[Auer and Ortner, 2007] P. Auer and R. Ortner. A new PAC bound for intersection-closed concept classes. *Machine Learning*, 66(2–3):151–163, 2007.

[Auer *et al.*, 1995] P. Auer, R. C. Holte, and W. Maass. Theory and applications of agnostic PAC-learning with small decision trees. In A. Prieditis and S. J. Russell, editors, *Machine Learning, Proceedings of the Twelfth International Conference on Machine Learning (ICML), Tahoe City, California, USA, July 9-12, 1995*, pages 21–29. Morgan Kaufmann, 1995.

[Auer *et al.*, 1998] P. Auer, P. M. Long, and A. Srinivasan. Approximating hyper-rectangles: Learning and pseudorandom sets. *Journal of Computer and System Sciences*, 57(3):376–388, 1998.

[Auer *et al.*, 2002] P. Auer, N. Cesa-Bianchi, and C. Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64(1):48–75, 2002.

[d'Avila Garcez, Lamb, and Gabbay, 2008] A. S. d'Avila Garcez, L. C. Lamb, and D. M. Gabbay. *Neural-Symbolic Cognitive Reasoning*. Cognitive Technologies, Springer, Berlin, 2008.

[Balkenius and Gärdenfors, 1991] C. Balkenius and P. Gärdenfors. Nonmonotonic inferences in neural networks. In J. A. Allen, R. Fikes, and E. Sandewall, editors, *Principles of Knowledge Representation and Reasoning*, pages 32–39. Morgan Kaufmann, San Mateo, 1991.

[Bamber, 2000] D. Bamber. Entailment with near surety of scaled assertions of high conditional probability. *Journal of Philosophical Logic*, 29:1–74, 2000.

[Baxter, 1998] J. Baxter. Theoretical models of learning to learn. In S. Thrun and L. Pratt, editors, *Learning to learn*, pages 71–94. Kluwer Academic Publishers, Norwell, MA, USA, 1998.

[Benferhat, Dubois, and Prade, 1997] S. Benferhat, D. Dubois, and H. Prade. Nonmonotonic reasoning, conditional objects and possibility theory. *Artificial Intelligence*, 92:259–276, 1997.

[Benferhat, Saffiotti, and Smets, 2000] S. Benferhat, A. Saffiotti, and P. Smets. Belief functions and default reasoning. *Artificial Intelligence*, 122:1–69, 2000.

[Bennett, 2003] J. Bennett. *A Philosophical Guide to Conditionals*. Clarendon Press, Oxford, 2003.

[van Benthem, 1984] J. van Benthem. Foundations of conditional logic. *Journal of Philosophical Logic*, 13:303–349, 1984.

[Blum, 1998] A. Blum. On-line algorithms in machine learning. In A. Fiat and G. J. Woeginger, editors, *Online Algorithms, The State of the Art*, *Lecture Notes in Computer Science*, volume 1442, pages 306–325. Springer, 1998.

[Blumer *et al.*, 1987] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Occam's razor. *Information Processing Letters*, 24(6):377–380, 1987.

[Blumer *et al.*, 1989] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

[Blutner, 2004] R. Blutner. Nonmonotonic inferences and neural networks. *Synthese*, 142:143–174, 2004.

[Board and Pitt, 1990] R. A. Board and L. Pitt. On the necessity of Occam algorithms. In *Proceedings of the Twenty Second Annual ACM Symposium on Theory of Computing (STOC), 14-16 May 1990, Baltimore, Maryland, USA*, pages 54–63. ACM, New York, NY, USA, 1990.

[Bouvel *et al.*, 2005] M. Bouvel, V. Grebinski, and G. Kucherov. Combinatorial search on graphs motivated by bioinformatics applications: A brief survey. In D. Kratsch, editor, *Graph-Theoretic Concepts in Computer Science, 31st International Workshop, WG 2005, Metz, France, June 23-25, 2005, Revised Selected Papers*, *Lecture Notes in Computer Science*, volume 3787, pages 16–27. Springer, 2005.

[Brewka, 1996] G. Brewka *Principles of Knowledge Representation*. CSLI Publications, Stanford, 1996.

[Brewka, 1997] G. Brewka, J. Dix, and K. Konolige *Nonmonotonic Reasoning. An Overview*. CSLI Publications, Stanford, 1997.

[Carnap, 1950] R. Carnap. *Logical Foundations of Probability*. University of Chicago Press, Chicago, 1950.

[Carnap, 1962] R. Carnap. The aim of inductive logic. In E. Nagel, P. Suppes, and A. Tarski, editors, *Logic, Methodology and Philosophy of Science*, pages 303–318. Stanford University Press, Stanford, 1962.

[Cesa-Bianchi and Lugosi, 2006] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, 2006.

[Chaitin, 1969] G. J. Chaitin. On the length of programs for computing finite binary sequences: statistical considerations. *Journal of the ACM*, 16:145–159, 1969.

[Corfield *et al.*, 2005] D. Corfield, B. Schölkopf, and V. Vapnik. Popper, falsification and the VC-dimension. Technical Report 145, Max Planck Institute for Biological Cybernetics, Department of Empirical Inference, Tübingen, Germany, November 2005.

[Dawid and Vovk, 1999] A. P. Dawid and V. G. Vovk. Prequential probability: principles and properties. *Bernoulli*, 5(1):125–162, 1999.

[Dawid, 1984] A. P. Dawid. Present position and potential developments: Some personal views. Statistical theory. The prequential approach. *Journal of the Royal Statistical Society. Series A. General*, 147(2):278–292, 1984.

[Dawid, 1985] A. P. Dawid. Comment on the impossibility of inductive inference. *Journal of the American Statistical Association*, 80(390):340–341, 1985.

[Domingos, 1998] P. Domingos. Occam's two razors: The sharp and the blunt. In R. Agrawal, P. E. Stolorz, and G. Piatetsky-Shapiro, editors, *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98), August 27-31, 1998, New York City, New York, USA*, pages 37–43. AAAI Press, 1998.

[Dudley, 1984] R. M. Dudley. A course on empirical processes. In *École d'Été de Probabilités de Saint-Flour XII-1982. Lecture Notes in Mathematics 1097*. Springer, New York, 1984.

[Ehrenfeucht *et al.*, 1989] A. Ehrenfeucht, D. Haussler, M. J. Kearns, and L. G. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–261, 1989.

[Eiter and Lukasiewicz, 2000] T. Eiter and T. Lukasiewicz. Default reasoning from conditional knowledge bases: Complexity and tractable cases. *Artificial Intelligence*, 124:169–241, 2000.

[Feldman, 2008] V. Feldman. Hardness of proper learning. In Ming-Yang Kao, editor, *Encyclopedia of Algorithms*. Springer, 2008.

[Flach, 2004] P. A. Flach. Logical characterizations of inductive learning. In D. M. Gabbay and R. Kruse, editors, *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, volume 4, pages 155–196. Kluwer, Dordrecht, 2004.

[Freund *et al.*, 1997] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2–3):133–168, 1997.

[Fuhrmann, 1997] A. Fuhrmann. *An Essay on Contraction*. CSLI Publications, Stanford, 1997.

[Gabbay, 1984] D. M. Gabbay. Theoretical foundations for non-monotonic reasoning in expert systems. In K. R. Apt, editor, *Logics and Models of Concurrent Systems*, pages 439–458. Springer, Berlin, 1984.

[Gabbay, Hogger, and Robinson, 1994] D. M. Gabbay, C. J. Hogger, and J. A. Robinson, editors. *Handbook of Logic in Artificial Intelligence and Logic Programming*. Volume 3, Clarendon Press, Oxford, 2003.

[Gärdenfors, 1988] P. Gärdenfors. *Knowledge in Flux*. The MIT Press, Cambridge, Mass., 1988.

[Gärdenfors and Makinson, 1994] P. Gärdenfors and D. Makinson. Nonmonotonic inference based on expectations. *Artificial Intelligence*, 65:197–245, 1994.

[Ginsberg, 1987] M. L. Ginsberg, editor. *Readings in Nonmonotonic Reasoning*. Morgan Kaufmann, Los Altos, 1987.

[Giraud-Carrier and Provost, 2005] C. G. Giraud-Carrier and F. J. Provost. Toward a justification of meta-learning: Is the no free lunch theorem a show-stopper? In *Proceedings of the ICML-2005 Workshop on Meta-learning*, pages 12–19, 2005.

[Gold, 1967] E. M. Gold. Language identification in the limit. *Information and Control*, 10(5):447–474, 1967.

[Goldreich *et al.*, 1998] O. Goldreich, S. Goldwasser, and D. Ron. Property testing and its connection to learning and approximation. *Journal of the ACM*, 45(4):653–750, 1998.

[Goldszmidt, Morris, and Pearl, 1993] M. Goldszmidt, P. Morris, and J. Pearl. A maximum entropy approach to nonmonotonic reasoning. *Pattern Analysis and Machine Intelligence*, 15:220–232, 1993.

[Goldszmidt and Pearl, 1996] M. Goldszmidt and J. Pearl. Qualitative probabilities for default reasoning, belief revision, and causal modeling. *Artificial Intelligence*, 84:57–112, 1996.

[Grove, 1988] A. Grove. Two modellings for theory change. *Journal of Philosophical Logic*, 17:157–170, 1988.

[Grünwald, 2007] P. D. Grünwald. *The Minimum Description Length Principle*. MIT Press, Cambridge, MA, USA, 2007.

[Hájek, 2003] A. Hájek. What conditional probability could not be. *Synthese*, 137:273–323, 2003.

[Halpern, 2001a] J. Halpern. Lexicographic probability, conditional probability, and nonstandard probability. In J. van Benthem, editor, *Proceedings of the Eighth Conference on Theoretical Aspects of Rationality and Knowledge*, pages 17–30. Morgan Kaufmann, Ithaca, NY, 2001.

[Halpern, 2001b] J. Halpern. Plausibility measures: A general approach for representing uncertainty. In *Proceedings of the 17th International Joint Conference on AI (IJCAI 2001)*, pages 1474–1483. Morgan Kaufmann, Ithaca, NY, 2001.

[Hansson, 1999] S. O. Hansson. *A Textbook of Belief Dynamics*. Kluwer, Dordrecht, 1999.

[Haussler and Welzl, 1987] D. Haussler and E. Welzl. ϵ-nets and simplex range queries. *Discrete & Computational Geometry*, 2(2):127–151, 1987.

[Haussler *et al.*, 1991] D. Haussler, M. Kearns, N. Littlestone, and M. K. Warmuth. Equivalence of models for polynomial learnability. *Information and Computation*, 95(2):129–161, 1991.

[Haussler *et al.*, 1994] D. Haussler, N. Littlestone, and M. K. Warmuth. Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):248–292, 1994.

[Haussler, 1988] D. Haussler. Quantifying inductive bias: AI learning algorithms and Valiant's learning framework. *Artificial Intelligence*, 36(2):177–221, 1988.

[Haussler, 1992] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

[Hawthorne, 1996] J. Hawthorne. On the logic of nonmonotonic conditionals and conditional probabilities. *Journal of Philosophical Logic*, 25:185–218, 1996.

[Höffgen *et al.*, 1995] K.-U. Höffgen, H.-U. Simon, and K. S. Van Horn. Robust trainability of single neurons. *Journal of Computer and System Sciences*, 50(1):114–125, 1995.

[Hutter, 2001] M. Hutter. New error bounds for Solomonoff prediction. *Journal of Computer and System Sciences*, 62(4):653–667, 2001.

[Hutter, 2004] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2004.

[Hutter, 2007] M. Hutter. On universal prediction and Bayesian confirmation. *Theoretical Computer Science*, 384(1):33–48, 2007.

[Kearns *et al.*, 1994] M. J. Kearns, R. E. Schapire, and L. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994.

[Kelly and Schulte, 1995] K. T. Kelly and O. Schulte. The computable testability of theories making uncomputable predictions. *Erkenntnis*, 43(1):29–66, 1995.

[Kelly, 2004a] K. T. Kelly. Justification as truth-finding efficiency: how Ockham's razor works. *Minds and Machines*, 14(4):485–505, 2004.

[Kelly, 2004b] K. T. Kelly. Learning theory and epistemology. In I. Niiniluoto, J. Woleński, and M. Sintonen, editors, *Handbook of Epistemology*, pages 183–204. Kluwer Academic Publishers, Dordrecht, 2004.

[Kelly, 2004c] K. T. Kelly. Uncomputability: the problem of induction internalized. *Theoretical Computer Science*, 317(1-3):227–249, 2004.

[Kolmogorov, 1965] A. N. Kolmogorov. Three approaches to the definition of the concept "quantity of information". *Problemy Peredači Informacii*, 1(vyp. 1):3–11, 1965.

[Kraus, Lehmann, and Magidor, 1990] S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic Reasoning, Preferential Models and Cumulative Logics. *Artificial Intelligence*, 44:167–207, 1990.

[Legg, 2006] S. Legg. Is there an elegant universal theory of prediction? In J. L. Balcázar, P. M. Long, and F. Stephan, editors, *Algorithmic Learning Theory, 17th International Conference, ALT 2006, Barcelona, Spain, October 7-10, 2006, Proceedings*, *Lecture Notes in Computer Science*, volume 4264, pages 274–287. Springer, 2006.

[Lehmann and Magidor, 1992] D. Lehmann and M. Magidor. What does a conditional knowledge base entail? *Artificial Intelligence*, 55:1–60, 1992.

[Lehmann, 1995] D. Lehmann. Another perspective on default reasoning. *Annals of Mathematics and Artificial Intelligence*, 15:61–82, 1995.

[Leitgeb, 2001] H. Leitgeb. Nonmonotonic reasoning by inhibition nets. *Artificial Intelligence*, 128:161–201, 2001.

[Leitgeb, 2004] H. Leitgeb. *Inference on the Low Level. An Investigation into Deduction, Nonmonotonic Reasoning, and the Philosophy of Cognition*. Kluwer, Dordrecht, 2004.

[Leitgeb, 2005] H. Leitgeb. Interpreted dynamical systems and qualitative laws: From inhibition networks to evolutionary systems. *Synthese*, 146:189–202, 2005.

[Leitgeb, 2007] H. Leitgeb. Neural network models of conditionals: An introduction. In X. Arrazola, J. M. Larrazabal et al., editors, *Proceedings of the First ILCLI International Workshop on Logic and Philosophy of Knowledge, Communication and Action*, LogKCA-07, pages 191–223. University of the Basque Country Press, Bilbao, 2007.

[Levesque, 1990] H. Levesque. All I know: A study in autoepistemic logic. *Artificial Intelligence*, 42:263–309, 1990.

[Lewis, 1973a] D. Lewis. Counterfactuals and comparative possibility. *Journal of Philosophical Logic*, 2:418–46, 1973.

[Lewis, 1973b] D. Lewis. *Counterfactuals*. Basil Blackwell, Oxford, 1973.

[Li and Vitányi, 1997] M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, New York, second edition, 1997.

[Li *et al.*, 2003] M. Li, J. Tromp, and P. Vitányi. Sharpening Occam's razor. *Information Processing Letters*, 85(5):267–274, 2003.

[Lukasiewicz, 2002] T. Lukasiewicz. Nonmonotonic probabilistic logics between model-theoretic probabilistic logic and probabilistic logic under coherence. In S. Benferhat and E. Giunchiglia, editors, *Proceedings of the 9th International Workshop on Non-Monotonic Reasoning*, NMR 2002, pages 265–274. Toulouse, 2002.

[Maass, 1994] W. Maass. Efficient agnostic PAC-learning with simple hypothesis. In *Proceedings of the Seventh Annual ACM Conference on Computational Learning Theory (COLT 1994), July 12-15, 1994, New Brunswick, NJ, USA*, pages 67–75. ACM, 1994.

[Makinson, 1994] D. Makinson. General patterns in nonmonotonic reasoning. In D. M. Gabbay, C. J. Hogger, and J. A. Robinson, editors, *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 3, pages 35–110. Clarendon Press, Oxford, 1994.

[Makinson, 1989] D. Makinson. General theory of cumulative inference. In M. Reinfrank et al., editors, *Non-Monotonic Reasoning*, Lecture Notes on Artificial Intelligence, volume 346, pages 1–18. Springer, Berlin, 1989.

[Makinson, 2005] D. Makinson. *Bridges from Classical to Nonmonotonic Logic*. Texts in Computing, volume 5. College Publications, London, 2005.

[McGee, 1994] V. McGee. Learning the impossible. In E. Eells and B. Skyrms, editors, *Probability and Conditionals. Belief Revision and Rational Decision*, pages 177–199. Cambridge University Press, Cambridge, 1994.

[Merhav and Feder, 1998] N. Merhav and M. Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6):2124–2147, 1998.

[Mitchell, 1990] T. M. Mitchell. The need for biases in learning generalizations. Technical Report CBM-TR-117, Rutgers Computer Science Department, May, 1980. Reprinted in J. W. Shavlik and T. G. Dietterich, editors, *Readings in Machine Learning*. Morgan Kaufmann, 1990.

[Mitchell, 1997] T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[Norton, 2003] J. D. Norton. A material theory of induction. *Philosophy of Science*, 70:647–670, 2003.

[Oaksford, Chater, and Hahn, 2009] M. Oaksford, N. Chater, and U. Hahn. Inductive Logic and Empirical Psychology. *This volume*, 2009.

[Osherson *et al.*, 1988] D. N. Osherson, M. Stob, and S. Weinstein. Mechanical learners pay a price for Bayesianism. *Journal of Symbolic Logic*, 53(4):1245–1251, 1988.

[Osherson and Weinstein, 2009] D. N. Osherson and S. Weinstein. Formal Learning Theory in Context. *This volume*, 2009.

[Paris, 1994] J. Paris. *The Uncertain Reasoner's Companion – A Mathematical Perspective*. Cambridge Tracts in Theoretical Computer Science, volume 39. Cambridge University Press, Cambridge, 1994.

[Pearl, 1988] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, 1988.

[Pearl, 1990] J. Pearl. System Z: a natural ordering of defaults with tractable applications to nonmonotonic reasoning. In R. Parikh, editor, *Proceedings of the Third Conference on Theoretical Aspects of Reasoning About Knowledge*, pages 121–135. Morgan Kaufmann, San Mateo, 1990.

[Pearl, 1997] J. Pearl and M. Goldszmidt. Probabilistic Foundations of Reasoning with Conditionals. In [Brewka, 1997], pages 33–68.

[Peters and Westerstahl, 2006] S. Peters and D. Westerstahl. *Quantifiers in Language and Logic*. Oxford University Press, Oxford, 2006.

[Pitt and Valiant, 1988] L. Pitt and L. G. Valiant. Computational limitations on learning from examples. *Journal of the ACM*, 35(4):965–984, 1988.

[Popper, 1969] K. R. Popper. *Logik der Forschung*. Mohr, third edition, 1969.

[Putnam, 1963] H. Putnam. 'Degree of confirmation' and inductive logic. In P. A. Schilpp, editor, *The Philosophy of Rudolf Carnap*, *The library of living philosophers*, volume 11, pages 761–783. Open Court, La Salle, Illinois, 1963. Reprinted in *Mathematics, Matter and Method. Philosophical Papers*, volume 1, pages 270–292. Cambridge, Cambridge University Press, 1975.

[Rao *et al.*, 1995] R. Bharat Rao, D. F. Gordon, and W. M. Spears. For every generalization action, is there really an equal and opposite reaction? In A. Prieditis and S. J. Russell, editors,

Machine Learning, Proceedings of the Twelfth International Conference on Machine Learning (ICML 1995), Tahoe City, California, USA, July 9-12, 1995, pages 471–479. Morgan Kaufmann, 1995.

[Reiter, 1980] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.

[Rissanen, 1976] J. J. Rissanen. Generalized Kraft inequality and arithmetic coding. *IBM Journal of Research and Development*, 20(3):198–203, 1976.

[Rissanen, 1978] J. J. Rissanen. Modelling by the shortest data description. *Automatica*, 14:465–471, 1978.

[Sauer, 1972] N. W. Sauer. On the density of families of sets. *Journal of Combinatorial Theory. Series A*, 13:145–147, 1972.

[Schaffer, 1993] C. Schaffer. Overfitting avoidance as bias. *Machine Learning*, 10(2):153–178, 1993.

[Schaffer, 1994] C. Schaffer. A conservation law for generalization performance. In W. W. Cohen and H. Hirsh, editors, *Machine Learning, Proceedings of the Eleventh International Conference (ICML 1994), Rutgers University, New Brunswick, NJ, USA, July 10-13, 1994*, pages 259–265. Morgan Kaufmann, 1994.

[Schurz, 2001] G. Schurz. What is 'normal'? An evolution-theoretic foundation of normic laws and their relation to statistical normality. *Philosophy of Science*, 68:476–497, 2001.

[Schurz, 2002] G. Schurz. Ceteris paribus laws: Classification and deconstruction. *Erkenntnis*, 57:351–372, 2002.

[Schurz and Leitgeb, 2005] G. Schurz and H. Leitgeb, editors. *Non-Monotonic and Uncertain Reasoning in the Focus of Paradigms of Cognition*. Special volume of *Synthese*, 146/1–2, 2005.

[Shannon, 1948] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[Shelah, 1972] S. Shelah. A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41:247–261, 1972.

[Shoham, 1987] Y. Shoham. A semantical approach to nonmonotonic logics. In J. P. McDermott, editor, *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, pages 388–392. Morgan Kaufmann, San Mateo, 1987.

[Snow, 1999] P. Snow. Diverse confidence levels in a probabilistic semantics for conditional logics. *Artificial Intelligence*, 113:269–279, 1999.

[Solomonoff, 1964a] R. J. Solomonoff. A formal theory of inductive inference. I. *Information and Control*, 7:1–22, 1964.

[Solomonoff, 1964b] R. J. Solomonoff. A formal theory of inductive inference. II. *Information and Control*, 7:224–254, 1964.

[Solomonoff, 1978] R. J. Solomonoff. Complexity-based induction systems: comparisons and convergence theorems. *Institute of Electrical and Electronics Engineers. Transactions on Information Theory*, 24(4):422–432, 1978.

[Solomonoff, 1997] R. J. Solomonoff. The discovery of algorithmic probability. *Journal of Computer and System Sciences*, 55(1):73–88, 1997.

[Spohn, 1987] W. Spohn. Ordinal conditional functions: A dynamic theory of epistemic states. In W. L. Harper and B. Skyrms, editors, *Causation in Decision, Belief Change, and Statistics*, volume 2, pages 105–134. D. Reidel, Dordrecht, 1988.

[Sprenger, 2009] J. Sprenger. Hempel and the Paradoxes of Confirmation. *This volume*, 2009.

[Stalnaker, 1991] R. C. Stalnaker. A theory of conditionals. In N. Rescher, editor, *Studies in Logical Theory*, American Philosophical Quarterly Monograph Series, volume 2, pages 98–112. Blackwell, Oxford, 1991.

[Stegmüller, 1969] W. Stegmüller. *Wissenschaftliche Erklärung und Begründung*. Springer, Berlin, 1969.

[Valiant, 1984] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[Vapnik and Chervonenkis, 1971] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

[Vapnik and Chervonenkis, 1974] V. N. Vapnik and A. Y. Chervonenkis. *Theory of Pattern Recognition (in Russian)*. Nauka, Moscow, 1974.

[Vapnik, 1995] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

[Vapnik, 1998] V. N. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.

[Vilalta *et al.*, 2005] R. Vilalta, C. G. Giraud-Carrier, and P. Brazdil. Meta-learning. In O. Maimon and L. Rokach, editors, *The Data Mining and Knowledge Discovery Handbook*, pages 731–748. Springer, 2005.

[Vitányi and Li, 2000] P. M. B. Vitányi and M. Li. Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Transactions on Information Theory*, 46(2):446–464, 2000.

[von Luxburg and Schölkopf, 2009] U. von Luxburg and B. Schölkopf. Statistical Learning Theory: Models, Concepts, and Results. *This volume*, 2009.

[V′yugin, 1998] V. V. V′yugin. Non-stochastic infinite and finite sequences. *Theoretical Computer Science*, 207(2):363–382, 1998.

[Webb, 1996] G. I. Webb. Further experimental evidence against the utility of Occam's razor. *Journal of Artificial Intelligence Research*, 4:397–417, 1996.

[Wenocur and Dudley, 1981] R. S. Wenocur and R. M. Dudley. Some special Vapnik-Chervonenkis classes. *Discrete Mathematics*, 33(3):313–318, 1981.

[Wolpert, 1995] D. H. Wolpert. The relationship between PAC, the statistical physics framework, the Bayesian framework, and the VC framework. In D. H. Wolpert, editor, *The mathematics of generalization. Proceedings of the SFI/CNLS Workshop on Formal Approaches to Supervised Learning*, pages 117–214. Addison-Wesley Publishing, Reading, MA, 1995.

[Wolpert, 1996a] D. H. Wolpert. The existence of a priori distinctions between learning algorithms. *Neural Computation*, 8(7):1391–1420, 1996.

[Wolpert, 1996b] D. H. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7):1341–1390, 1996.

[Wolpert, 2001] D. H. Wolpert. The supervised learning no-free-lunch theorems. In *Proceedings of the 6th Online World Conference on Soft Computing in Industrial Applications*, 2001.

[Zabell, 2009] S. Zabell. Carnap and the Logic of Induction. *This volume*, 2009.