85-29

Information and Computer Science

The computation of contingency in classical conditioning

Richard H. Granger, Jr. and Jeffrey C. Schlimmer

Center for the Neurobiology of Learning and Memory
and
Computer Science Dept.
University of California
Irvine, California 92717

TECHNICAL REPORT

UNIVERSITY OF CALIFORNIA IRVINE

The computation of contingency in classical conditioning

Richard H. Granger, Jr. and Jeffrey C. Schlimmer

Center for the Neurobiology of Learning and Memory
and
Computer Science Dept.
University of California
Irvine, California 92717

This research was supported in part by the Office of Naval Research under grants N00014-84-K-0391 and N00014-85-K-0854, by the Army Research Institute under contract MDA903-85-C-0324, and by the National Science Foundation under grants IST-81-20685 and IST-85-12419.

Published in *The Psychology of Learning and Motivation*, Volume 20, 1986.

The computation of contingency in classical conditioning

Richard H. Granger, Jr. and Jeffrey C. Schlimmer

Center for the Neurobiology of Learning and Memory
and
Computer Science Dept.
University of California
Irvine, California 92717

Contents

This research was supported in part by the Office of Naval Research under grants N00014-84-K-0391 and N00014-85-K-0854, by the Army Research Institute under contract MDA903-85-C-0324, and by the National Science Foundation under grants IST-81-20685 and IST-85-12419.

Published in *The Psychology of Learning and Motivation*, Volume 20, 1986.

1 Introduction: Theory and experiment in classical conditioning

Experimental and theoretical work on classical conditioning over the past 20 years includes mathematical formulations of the conditions under which conditioning will and will not occur in animals (Rescorla 1967, 1968; Gibbon, Berryman and Thompson 1974); algorithms that give rise to this behavior (e.g., Rescorla and Wagner 1972, Mackintosh 1975, Pearce and Hall 1980, Wagner 1981); computer simulations of the behavior (e.g., Rescorla and Wagner 1972, Sutton and Barto 1981, Hampson and Kibler 1983); and substrate-level implementations of the circuits that may underlie conditioning (Hawkins and Kandel 1984, Chang and Gelperin 1980, Alkon 1980, Thompson et. al. 1984, Gluck and Thompson 1985).

It is quite difficult, however, to *evaluate*, in a principled way, how all of these experimental results, algorithms, computer models and proposed circuits are related to each other. For instance, how could we go about deciding whether the Rescorla–Wagner (1972) or Mackintosh (1975) algorithms do what the Rescorla (1968) constraint specifies that such algorithms are supposed to do? How might we decide whether a particular experimental result should imply a revision to that constraint?

This paper presents a unified framework within which to view the computations, algorithms and neurobiological implementations underlying classical conditioning. In particular, we present an extensive mathematical analysis of the constraints on classical conditioning, as originally identified by Rescorla (1968); i.e., the precise *contingency* conditions under which mammals will and will not learn a particular association between two events in a classical conditioning situation. In classical conditioning an unconditional stimulus (US), i.e., a cue that is inherently biologically salient to an animal (such as an electric shock) is repeatedly paired with a conditional stimulus (CS), a cue that initially has no special significance to the animal (e.g., a tone or a light); over repeated trials, the animal can learn that the CS is predictive of, or associated with, the US. This phenomenon of associative learning is subject to laws and constraints: an association will be learned to some extent in some conditions, and to a lesser extent (or not at all) in others.

Using Marr's (1982) distinction among the computational (roughly, behavioral), algorithmic (abstract mechanism) and implementation (neurobiological) levels of analysis of psychobiological mechanisms, our computational analysis may be applied as a partial test of the adequacy of a number of proposed algorithm-level and circuit-level mechanisms for classical conditioning.

Our computational analysis is applied to a broad range of issues relating to contingency in classical

conditioning, and a number of results are derived:

- A new class of trial presentation conditions for classical conditioning is identified and distinguished from other presentation conditions. This new class of conditions, which we term *'partial warning'* , is simply the reciprocal of the well-known 'partial reinforcement' condition: where partial reinforcement intersperses spurious (unpaired) CS trials with CS-US pairings (with no spurious USs), the partial warning condition intersperses spurious USs with no spurious CSs; both of these 'partial' conditions are differentiated from the 'composite' class of presentation conditions, in which combinations of *both* spurious CS and spurious US trials are added to CS-US pairings. The new condition has been mentioned only rarely in the literature, and we show how comparative analysis of these conditions may prove fruitful in evaluating proposed algorithms and circuits for contingency.

- A number of new predictions are generated which may be tested experimentally; in particular, the computational analysis of contingency predicts that learning of a positive CS-US association should occur in even the extreme cases of the partial warning condition, as it does in extreme partial reinforcement conditions, but *not* in extreme composite conditions.

- It is shown that the standard predictions of contingency-based associative learning in classical conditioning (from Rescorla 1968), depend critically on strong assumptions about timing. In particular, under different assumptions about the duration of a trial (2 minutes vs. 3 minutes, etc.), the contingency prediction of whether or not a particular CS-US association will be learned, or the extent to which it will be learned, is greatly altered.

- Algorithms presently in the literature are analysed for their adequacy to account for the range of effects predicted by the computational contingency constraint. A new algorithm is proposed, that accounts for the appropriate computational constraints (including the new partial warning prediction), as well as accounting for blocking and providing a coherent account of some learned-irrelevance and latency effects in conditioning.

- Proposed neurobiological circuits for classical conditioning are similarly analyzed for their adequacy to account for these predictions. In particular, Hawkins and Kandel (1984) have offered an analysis of a neurobiological circuit in *Aplysia* as evidence that the operation of this circuit gives rise to associative learning; we address the question of whether the circuit's operation also conforms to the same specific predictions as mammals in classical conditioning situations.

If so then a strong connection between molluscan and mammalian conditioning will have been shown; if not, then it will be possible to rigorously distinguish molluscan and mammalian classical conditioning.

This paper raises a number of theoretical and experimental questions in light of our framework for couching the mechanisms of associative learning. The rest of the paper is divided roughly into two parts: Sections 2 and 3 provide overview, introduction and background to our approach and our results; Sections 4 and 5 then give detailed and in-depth analyses of the questions we have raised. For many of these theoretical questions, no answers are provided *per se*, but, wherever possible, we have attempted to develop explicit experimental predictions from our theoretical work, to ensure that our results are testable and falsifiable.

2 A three-level analysis of classical conditioning

2.1 Characterization of partial vs. composite presentation conditions

Mammals have been tested extensively for their sensitivity to various presentation conditions in classical conditioning (e.g., Rescorla 1968, 1972; Mackintosh 1975; Dickinson 1980; Rescorla and Wagner 1972; Gibbon et. al. 1974). Rescorla (1968) identified the conditions that enable and prevent learning of a particular association over trials: a positive CS-US association will be learned only if the probability of the US occurring, given that the CS has occurred, is greater (over trials) than the probability of the US occurring given that the CS has not occurred, or, formally, $p(US|CS) > p(US|\overline{CS})$.

This new constraint condition on associative learning in classical conditioning, termed contingency by Rescorla (1968), displaced the then-prevalent notion that simple contiguity (i.e., the number of paired presentations of CS and US) was the key factor that determined the level of learning of a CS-US association (Spence, 1936). Rescorla demonstrated that it was this measure of relative conditional probabilities, not number of pairings, that determined whether a particular association would be learned, and the extent of the associative strength that would be perceived between the CS and US.

Analysis of this constraint of relative conditional probabilities shows that learning of positive CS-US associations is enabled in certain categories of presentation conditions, and is prevented in other conditions. For instance, animals will readily learn a positive CS-US association in a *'perfect pairings'* condition (i.e., repeated CS-US pairing trials, with no misinformation presented). From the statement of relative conditional probabilities, it can be readily predicted that animals will also learn the association

to some extent even in extreme *'partial reinforcement'* conditions[1] (perfect pairings with many spurious CSs mixed in), but that learning of the positive association will be severely degraded in *'composite misinformation'* conditions, where *both* spurious CSs *and* spurious USs are mixed in with presentations of pairings.[2] This is because the above conditional probability inequality holds throughout the perfect pairings and partial reinforcement conditions, but does not necessarily hold in composite conditions.

(FIGURE 1 GOES ABOUT HERE)

Hence, based on the contingency constraint of relative conditional probabilities, we can rigorously distinguish between characteristics of learning in partial reinforcement conditions vs. in composite misinformation conditions. In the partial condition, as more and more spurious (unpaired) CS trials are mixed in with paired CS-US trials, learning will degrade only very mildly. If the level of associative CS-US correlation is plotted against the percentage of presented spurious CS trials in partial reinforcement (top curve in Figure 1), learning of the association degrades very gently until the percentage of spurious CSs is up around 90%, and only goes to zero when there are 100% spurious CS trials. This means that there will be some learning of the positive CS-US association no matter how many spurious CSs are added in a partial reinforcement condition, up to but not including 100% spurious CSs, and, furthermore, that the level of learning of the CS-US association will barely be degraded at all unless trials consist of more than 90% spurious CSs overall.

In contrast to the almost imperceptible, gentle degradation in the partial reinforcement condition, in the composite misinformation condition learning will be severely degraded with the addition of more and more spurious CS and spurious US trials are mixed in with paired trials. The lower line in Figure 1 plots the strength of CS-US learning against the percentage of spurious CS and spurious US trials in the composite condition; in this case, learning of the association severely degrades down to zero association with 50% spurious trials, and as the percentage of spurious trials increases over 50%, the *inverse* of the association is increasingly learned; i.e., the CS is learned to be a 'safety signal' indicating that the US will not occur. The difference between the partial and composite cases can be clearly seen: as the

[1]This condition has of course been extensively tested and confirmed in the literature (e.g., Fitzgerald (1963), Rescorla (1968)).

[2]There are a number of subcategories of the composite misinformation condition: for very few spurious CSs and USs, the animal will still learn the positive association; as these are increased, the animal will increasingly fail to learn the positive association, and will increasingly tend to learn the *inverse* of the association: i.e., that the CS is a 'safety signal' indicating that the US is *not* about to occur. Section 4 presents a precise specification of learning in these conditions and the implications thereof.

percentage of spurious trials increases past about 30% or 40%, learning will be almost unimpaired in the partial condition, but will be severely degraded in the composite condition.

2.2 The four trial-presentation conditions of classical conditioning

We have distinguished the characteristics of learning in the perfect pairings (no spurious trials) condition, the partial reinforcement class of conditions (spurious CS trials, but no spurious USs), and the composite misinformation condition class (both spurious CS and spurious US trails): the composite condition exhibits very severe degradation of the learning of a CS-US association, while the partial condition yields only very gentle degradation of learning.

When presented in this way, a fourth logical class of testing categories comes to light: what will happen in a situation in which spurious USs, but no spurious CSs, are mixed in with CS-US pairings? We term this fourth category the *'partial warning'* condition: in this condition, not all of the USs (e.g., shocks) are preceded by a CS (tone) warning. In this light, 'perfect pairings' represent both perfect reinforcement (of the tone CS) and perfect warning (of the shock US); the two 'partial' conditions correspond to misinformation along one and only one of these two dimensions, while the 'composite' condition presents misinformation in both ways (reinforcement and warning). The partial warning condition seems not to have been extensively tested in the literature;[3] in particular, it is unclear from the literature whether gentle or severe degradation of associative learning learning occurs in this class of testing conditions.

It is important to note that these four classes of trial-presentation conditions simply represent subdivisions of the continuum of all possible such conditions — they are not discrete, discontinuous categories, but rather are particular sub-areas of the overall 'space' of possibilities. This 'contingency space' will be introduced in Section 2.3, and then will be explored in some depth in Section 4 of this paper. However, continued reference to these four categories of trial-presentation conditions will enable us to afford a clear discussion of the particular characteristics of learning in each condition.

A number of researchers have experimented with spurious USs, but apparently only in combination with spurious CSs (thereby forming a composite misinformation condition). For example, Rescorla (1968, 1972), who was the first to perform a systematic exploration of contingency effects in conditioning,

[3] Gibbon et. al. (1974) identified this category of presentations, which they termed the " 'CS–implies–US implication' ... CS implies US but USs occur with some probability in ~CS [the absence of a CS] also." They reported then that "This ... implication represents another case of partial schedules that has not been investigated" This seems still to be true more than a decade later; we are in the process of testing this condition in our laboratory.

	No spurious (unpaired) USs	Spurious USs
No spurious (unpaired) CSs	Perfect Pairings (PP)	Partial Warning (PW)
Spurious CSs	Partial Reinforcement (PR)	Composite Condition (C)

Table 1: Categories of trial presentations in classical conditioning

proceeds by first testing partial reinforcement conditions, i.e., CS-US pairs and spurious CSs, showing that gentle degradation of learning occurs with the addition of spurious CSs. He then adds spurious US trials to the spurious CSs, generating composite misinformation cases, and demonstrates that learning becomes severely decremented as the percentage of spurious (CS and US) trials is increased. Careful reading of Rescorla (1966, 1968, 1972) shows clearly that he does *not* report testing the effects of spurious USs without spurious CSs, i.e., the partial warning condition. In the same vein, work on contingency following Rescorla's (e.g., Gamzu and Williams 1971; Hearst and Franklin 1977; Mackintosh 1983; Dickinson 1980) has concentrated on the partial reinforcement and composite misinformation conditions — we have been unable to find any report of systematic testing of CS-US pairs plus spurious USs, without spurious CSs, in the animal or human learning literature.

The mathematical analysis of contingency presented here predicts that only gentle degradation of learning should occur throughout the partial warning condition, just as it does in the partial reinforcement condition. To test this prediction, we are conducting an experiment replicating Rescorla (1968, exp. 2) on the partial reinforcement (PR) condition (0.4–0), composite (C) condition (0.4–0.4), the null (N) condition (0–0) (a control in which no CSs or USs are presented to the animal), and adding a partial warning (PW) condition (1.0–0.4).

2.3 Significance of the new findings: A three-level analysis

2.3.1 The three levels: computation, algorithm, implementation

This new partial warning condition is potentially just as integral a part of classical conditioning as is the well-known partial reinforcement condition; the four conditions (PP, PW, PR and C) taken together constitute complete coverage of the possible testing conditions for classical conditioning. Using the new analysis presented here of the *contingency constraint*, i.e., the computational specification of the conditions under which classical conditioning will and will not occur, we have been able to define this

presentation condition formally and clearly, and to generate the prediction that learning should occur throughout this condition, just as it does in partial reinforcement.

Using Marr's (1982) distinction among the computational, algorithmic and implementation levels of analysis of psychobiological mechanisms, we propose a set of related analyses of classical conditioning at all three levels. Any complete theory of any complex phenomenon can usefully be divided into these three separate levels. Animals can be said to perform identifiable computations, that is, to transform inputs to outputs in a principled way. For instance, in order to learn which of many possible cues (tone, light, air puff) reliably predict the occurrence of some other salient stimulus (e.g., shock, food), a rat in a classical conditioning situation must 'compute' the relative predictiveness of each of the possible cues with respect to the occurrence of shock.

At the *computational* level, we may speak simply of these computations that must somehow be performed, without reference to *how* those computations may be carried out. The *algorithm* level constitutes the level of mathematical function that performs the necessary computations. Finally, these algorithms may be instantiated in a substrate (e.g., neurons, wires, computer bits) at the *implementation* level. These three levels are not wholly independent. In particular, the algorithm level must conform to the constraints provided by each of the other two levels: it must compute all and only those things that have been identified as actually occurring in animal learning (at the computational level) by making use of only those tools provided in the substrate (at the implementation level).

It is at the computational level that the target behavior is formally and precisely stated, so this level prescribes the characteristics of the object of study. This allows us to speak of the degree of "correctness" of algorithms that are proposed to calculate the behavior. Any algorithm, no matter how fast, elegant or efficient, is a *correct* algorithm for, say, contingency in classical conditioning, only to the extent that it gives rise to the precise target behavior; i.e., it learns or fails to learn a specific CS-US association in precisely those conditions that the computational level specifies. Similarly, any neurobiological circuit actually instantiates classical conditioning only to the extent that its operation gives rise to those correct target behaviors specified in the computational-level analysis. Once such a computational analysis has been performed (in this case, by Rescorla, 1968), then using the computational level as 'arbiter' of the adequacy of proposed algorithms and circuits enables us to narrow the search for valid mechanisms of associative learning.

By the same token, however, a given computation is correct only to the extent that it is actually computable with the mechanisms provided by the substrate. The relevant neurobiology therefore es-

tablishes an equally crucial constraint, in the sense that it is the substrate that (somehow) gives rise to the target behavior. Just as algorithms for learning must conform to the computational constraints of the target behaviors to be explained, so must algorithms conform to the implementational constraints of the substrate. For instance, any proposed algorithm must be able to be run in a parallel, associative network of neurons, since that is the nature of the substrate. The problem is that it is often possible to experimentally identify a precise characterisation of a target behavior long before the relevant neurobiology is identified; this is clearly the case, for instance, with classical conditioning.

These three are distinct, and in most respects independent (although they interact); it is often quite unclear just what a particular algorithm or implementation computes. For instance, 'connectionist' models of learning (e.g., Anderson et. al. (1977), Hampson and Kibler (1983)) consist of large numbers of distributed, parallel nodes and links that cooperatively and competitively perform individual calculations; analysing what the overall system computes quite often turns out to be a mathematically difficult or intractable task.

The computational constraint of contingency in classical conditioning can be stated loosely as the fact that the positive CS-US association will be learned in the PP, PR and PW conditions described above, and will not be learned in particular C conditions. The mathematical formulation of this set of results enables us to re-cast the existing analyses of contingency into a larger framework. This analysis may then be used to determine which of many proposed mathematical algorithms, and neurobiological circuits, conform to the appropriate constraint.

Section 4 of this paper describes all this in detail; the following is a brief introductory presentation of our computational, algorithmic and implementation-level analyses.

2.3.2 The computational analysis of contingency

Rescorla's (1967, 1968) original characterisation of the contingency computation was that rats are able to learn a positive CS-US association only if the probability of the shock outcome (the US) given the occurrence of the conditional stimulus feature (the CS, e.g., the tone) is greater than the probability of the outcome occurring without that feature having occurred, or, stated in terms of conditional probability, $p(US|CS) > p(US|\overline{CS})$.[4]

[4]It is crucial to note that *conditional* probabilities (e.g., $p(US|CS)$ — the probability of the US *given* occurrence of the CS) are distinct from *joint* probabilities (e.g., $p(US, CS)$ — the probability of the US and CS together; these are related: $p(US|CS) = p(US, CS)/p(CS)$); and both are in turn distinct from simple *marginal probabilities* (e.g., the percentage of USs or CSs over trials). These differences are gone into in detail in Section 4.

This constraint can be translated into a three-dimensional graph, in which the three axes correspond to the joint probabilities of the CS and US occurring, the CS but not US occurring, and the US but not the CS occurring[5]. In other words, the three axes correspond to the probability of CS-US pairs, the probability of spurious (unpaired) CSs, and the probability of spurious (unpaired) USs. Figure 2 shows two rotated views of these axes, and plots the above Rescorla conditional probability boundary surface, which translates into a saddle-shaped surface (hyperbolic paraboloid) in this space. (The mathematical derivation of the equation plotted here is given in Appendix A.)

(FIGURE 2 GOES ABOUT HERE)

Each point in this space corresponds to a specific set of classical conditioning trials, with the probabilities of the CS and US occurring together determined by the point's location in the space. Points on (or in the immediate vicinity of) the saddle surface correspond to those presentation conditions in which the presented CS-US association will not be learned; 'inside' (on the Z-axis side of) the surface, the positive association will be learned to some extent (i.e., the CS signals the US), while 'outside' the surface (the side away from the Z-axis), the negative association will be learned (i.e., that the CS is a "safety signal" indicating that the US will not occur). All proposed mechanisms (mathematical or biological) purported to perform classical conditioning can be tested for their adequacy to account for contingency by measuring their (ideal) performance against the criteria represented by this curve. (Section 4 goes more deeply into this computational analysis and its implications).

The computational contingency constraint itself is not inviolate; though it arose from systematic behavioral testing (Rescorla 1968, 1969, 1972), and has been replicated and extended (Rescorla and Wagner 1972, Mackintosh 1975, Pearce and Hall 1980), there are testable predictions from the formula that have not yet been tested, which, if in conflict with (future) experimental results, would require modification of the constraint. In other words, the computational level constraint is experimentally testable and falsifiable; for instance, the partial warning condition will provide a test of a specific class of predictions of the theory (see Section 4) that have not yet been subjected to systematic testing.

A particular question that arises about the contingency formula is that the calculation of conditional probabilities depends on an explicit assumption about the duration of time that is deemed to constitute a trial. Different choices of trial duration can change the values of the conditional probabilities for any single set of trials. This means that perceived trial duration will alter perceived conditional probabilities,

[5]The fourth logical possibility, the probability that neither CS nor US occur, is uniquely determined at each point in this three-space, and hence is not a separate independent axis.

and so will determine in part which of several potential CS cues the animal will associate with the US, and how strongly that association will be learned. This leads inexorably to the assumption either that (a) particular animals have fixed 'trial window' durations (possibly different fixed durations for different classes of CSs), or (b) that animals have a way of choosing a trial window duration based on some characteristic of the trials, such as the duration of the CS. It is interesting to note that in models and simulations of classical conditioning (e.g., Rescorla and Wagner 1972), as well as in animal experiments (e.g., Rescorla 1968), that the trial window duration is assumed to be set equal to the CS duration. This is by no means the only possible such assumption, and in fact other assumptions can drastically change the predicted behavior of subjects, and the performance of simulations. In sum, the assumption of trial window duration must be added as an explicit assumption applied to the interpretation of these experiments and simulations. These issues will be discussed in more detail in Section 5.5.2.

2.3.3 The contingency algorithm

This computational constraint of when animals will and will not learn an association can be translated into an *algorithm* or abstract mechanism that gives rise to that computation. A number of researchers (including Rescorla and Wagner (1972), Dickinson (1980), Pearce and Hall (1980), Wagner (1981) and Mackintosh (1983)) have developed algorithm-level theories of learning to capture the major effects of contingency in classical conditioning; we briefly review aspects of these theories in Sections 4.2 and 5.

We propose an algorithm based on Bayes' rule of induction (Bayes 1763, Pearl 1983): the algorithm makes use of precisely the inputs that the animal gets in a classical conditioning situation, together with the animal's *expectations* of what will occur, and, in a natural trial-by-trial fashion, assigns incrementally-changing associative strengths to various candidate CS-US pairings. (Indeed, Bayes' centuries-old rule corresponds closely to Rescorla's original (1968) characterisation of the computational constraint for contingency in rats; i.e., that the positive association should be learned only if the probability of the US given the occurrence of the CS is greater than the probability of the US occurring without the CS: $p(US|CS) > p(US|\overline{CS})$. It is compelling to note that the two were arrived at entirely independently, yet both to account for inductive learning; one in mathematical philosophy and one in animal learning.)

Section 4.2 shows that this algorithm performs as it should: i.e., it learns in the appropriate presentation conditions. Furthermore, the algorithm yields the Kamin (1968) blocking phenomenon in a very natural way, as a side effect of its operation (see Section 5.1 on page 34). Finally, the algorithm requires no counterintuitive calculations on the part of the animal; rather, it is a very plausible and

simple calculation to imagine neural circuits to be performing.

2.3.4 The neurobiological implementation of contingency

These new computational and algorithm-level characterisations in turn pose a set of necessary constraints that must be satisfied by candidate biological mechanisms that are proposed to underlie classical conditioning; hence, the characterisation may aid in narrowing the neurobiological search for such candidate mechanisms (at the *implementation* level). The characterisation could in principle have been derived (in a bottom-up fashion) from the known circuitry, but so far analysis of the operation of neurobiological circuits has not given rise to computational constraints for contingency; the laws of animal contingency learning nonetheless constitute a necessary condition for a complete test of the validity of any proposed circuit for classical conditioning.

Using this three-level analysis thus gives us a tool for distinguishing among viable and non-viable candidates for biological mechanisms underlying classical conditioning, and, furthermore, for potentially distinguishing among possible different variations of classical conditioning that may occur in different taxonomic categories ('taxa') of animals (e.g., different orders, classes, phyla, species). For example, Hawkins and Kandel (1984) present evidence that invertebrate *Aplysia* perform associative learning (i.e., response to the CS is altered by the CS being paired with a US), which raises the tantalising possibility that this molluscan associative learning may be equivalent to mammalian classical conditioning. If so, then the *Aplysia* circuit (and the intact preparation) should exhibit only gently-degraded learning in the partial (PW and PR) conditions specified above, but severely-degraded learning in composite misinformation (C) conditions.

If, however, this set of constraints does not hold in *Aplysia*, then there this would indicate that there exist important differences between molluscan and mammalian conditioning. This in turn would suggest that these invertebrates may be performing some related, but distinct, algorithm for associative learning.[6] The results of this analysis could potentially have strong implications that might limit the usefulness of certain taxa (e.g., phyla, classes, orders) of animals as valid models of higher mammalian learning phenomena, by rigorously distinguishing between characteristics of associative learning in different taxa of animals.

[6]The other alternative is, of course, that the Rescorla (1968) computational constraint is in error; since this constraint has been extensively tested (e.g., Rescorla (1967, 1968, 1972), Rescorla and Wagner (1972), Gibbon et. al. (1974), Mackintosh (1974)), we assume for now that the constraint is correct. Complete experimental validation of the constraint will depend on the testing of the partial warning case.

2.4 Partial summary

The computational analysis of contingency gives rise to a mathematical distinction between true contingency and partial approximations of it; identification of a new experimental condition (partial warning) in classical conditioning, and an experimentally-testable prediction about its characteristics (i.e., that learning should occur throughout this case); as well as a new method of analysis of neurobiological circuits proposed to underlie classical conditioning.

The key point here is that in the absence of this kind of computational analysis there would be no principled way to tell whether any particular proposed algorithm, theory or circuit for classical conditioning is correct or not. Using the analysis, we can now specify what constraints any such proposal must satisfy in order to be an adequate candidate mechanism for contingency in classical conditioning.

3 Background: Historical perspective on contingency

3.1 Experimental results on contingency

An examination of the development of the contingency constraint indicates that its roots lie in the notion of contiguity: when two events follow each other closely, animals tend to form excitatory associations. Spence (1936) details a somewhat more fleshed-out approach, explaining that an association between two events (e.g., a CS and a US) is a function of the number of times they occur together versus the number of times they do not; the former *strengthens* while the latter *weakens* an excitatory association between the two events. In terms of joint (not conditional) probabilities, this constraint on learning is: $p(CS, US) > p(CS, \overline{US}) + p(\overline{CS}, US)$.

The next development in the contingency constraint was a sequence of experiments revealing that, under partial reinforcement conditions, animals form excitatory associations similar to those elicited by a contingency based on perfect pairings (Fitzgerald, 1963; Wagner *et. al.* 1964; Thomas and Wagner, 1964; Brimer and Dockrill, 1966).

A few years later, experiments explicitly aimed at exploring the space of possible contingencies led Rescorla to form the characterization that $p(US|CS) > p(US|\overline{CS})$ for excitatory conditioning to occur, that $p(US|CS) < p(US|\overline{CS})$ for inhibitory conditioning, and that $p(US|CS) = p(US|\overline{CS})$ prevents either type of conditioning (Rescorla, 1966, 1967, 1968, 1969). This newly-formulated constraint of contingency supplanted the existing notion that simple contiguity (i.e., the number of CS-US pairings) was the measure of associative learning in classical conditioning. This had far-reaching implications

for the proper control procedures in classical conditioning (Rescorla 1967), and for the possible mechanisms that animals could be using to calculate the associative predictiveness of various cues in classical conditioning situations.

Rescorla's seminal experiments studied a wide range of $p(US|CS)$-to-$p(US|\overline{CS})$ values, denoted by a pair of numbers (n–m) corresponding to the values of $p(US|CS)$ and $p(US|\overline{CS})$. Presentation conditions tested by Rescorla (1966, 1967, 1968) were: 0.0–0.0, 0.0–0.2, 0.0–0.4, 0.0–0.8, 0.1–0.0, 0.1–0.1, 0.2–0.0, 0.2–0.1, 0.2–0.2, 0.4–0.0, 0.4–0.1, 0.4–0.2, 0.4–0.4. The partial reinforcement, composite, and inhibitory contingencies were well explored by Rescorla and others (Hammond, 1967; Gamzu and Williams, 1971; Hearst and Franklin, 1977), confirming the contingency characterization. However, none of these experiments, by Rescorla or others, systematically tested partial warning contingencies, i.e., those in which there are *no* spurious CSs, but there are spurious USs mixed with CS-US pairs.

It is useful to observe some attributes of conditional probabilities in the presentation conditions of classical conditioning. For example, in all partial reinforcement conditions, $p(US|\overline{CS}) = 0$, since USs never occur without CSs in this condition; or, in other words, there are no spurious US presentations. It is the value of $p(US|CS)$ that may be varied. Hence, all partial reinforcement conditions are of the form N–0 (e.g., 0.8–0.0, 0.4–0.0, 0.2–0.0). All such points lie on the X–Z plane of the contingency space. Reciprocally, in all partial warning conditions, $p(US|CS) = 1$, since the US always occurs if the CS has; i.e., there are no unpaired (spurious) CSs in this condition. This means that all partial warning values are of the form 1–N (e.g., 1.0–0.8, 1.0–0.4, 1.0–0.2). These points all lie on the Y–Z plane. Composite values (which lie in the space between the three planes) may be of the form N–M for any N and M ($0 < N < 1$, $0 < M < 1$); those values for which associations will not be learned are those for which N=M (these lie on the saddle surface itself). Finally, the perfect pairings case is 1.0–0.0 ($p(US|CS) = 1$ and $p(US|\overline{CS}) = 0$; these points lie on the Z axis. We list explicit $p(US|CS)$-to-$p(US|\overline{CS})$ values in this section in order to illustrate clearly which categories of conditions have been tested and which have not.

In human experimental settings, contingency has been studied mostly in response outcome situations where $p(O|R)$ (the probability of the outcome given the response) and $p(O|\overline{R})$ lie within the partial reinforcement and composite contingency conditions. Allan and Jenkins (1980) found that when subjects were presented with response-1, response-2, and no-response alternatives, the subjects estimated the actual contingency accurately, provided there was a contingency ($p(O|R_1) \neq p(O|R_2)$). In the absence of a contingency between response and outcome, subjects' estimations were found to be related to the

overall probability of the outcome. The contingencies they investigated included ($p(O|R_1)$-to-$p(O|R_2)$) values equaling: 0.1–0.3, 0.1–0.5, 0.1–0.9, 0.2–0.8, 0.5–0.8, 0.5–0.9, 0.7–0.9.

Wasserman, Chatlosh, and Neunaber (1983) studied the effects of discrete versus continuous responses and temporal regularity in contingency perception during free operant procedures. They investigated the nine combinations of $p(O|R) = 0.125, 0.500, 0.875$ by $p(O|\overline{R}) = 0.125, 0.500, 0.875$ finding that subjects' rating of the contingency was strongly correlated to the actual contingency presented.

Shanks (1985) found that contingency judgements increased toward a positive asymptote when the actual contingency was positive and vice-versa. He investigated $p(O|R)$-to-$p(O|\overline{R})$ = 0.25–0.25, 0.25–0.75, 0.75–0.25, 0.75–0.75. These contingencies lie within the composite condition, asserting the validity of the contingency characterisation within that condition, though they add no new data to the partial reinforcement or partial warning cases.

3.2 Theoretical results on contingency

Rescorla and Wagner (1972), Wagner and Rescorla (1972), Mackintosh (1975), and Pearce and Hall (1980) have described algorithms for the computation of these effects. Each of these models is based on parameters (such as the innate salience of each cue in the environment, and the innate salience of the US) which are used to describe the change in associative strength between a CS and US as a result of repeated pairings. The Rescorla-Wagner (1972) model assumes that each US can support only a limited association strength for which co-occurring CSs compete, and the effectiveness of the US in promoting conditioning is inversely proportional to the degree to which it is predicted by the stimuli occurring on a given trial. An effective signal can greatly reduce the effectiveness of the US and thereby result in blocking. In contrast, attentional models such as those of Mackintosh (1975) and Pearce and Hall (1980) assume that conditionability or salience of the CS varies proportionately to the degree to which the US is predicted. Blocking results in a variation in CS processing rather than reduction in US processing. These features allow attentional models to account for more of the data on blocking and latent inhibition.

Several researchers have designed mathematical models at the implementational level which address the contingency constraint. Sutton and Barto (1981) utilize a neuron-like element which computes an output based on a function of its weighted inputs. The process of adjusting the weights is designed to allow the model to replicate the general characteristics of the reported data on contingency, the effect of the interstimulus interval on conditioning, blocking, and higher–order conditioning. Their work includes

a sizable discussion of the inherent mathematical and implementational constraints on the design of any model at this level.

Other representative mathematical, implementational models are based on the work of perceptrons (Rosenblatt, 1962), also simple neuron-like elements. For example, Hampson and Kibler (1983) demonstrate how a small, layered network of these elements may compute any arbitrary boolean function of its inputs. They present completeness and correctness results and explain how such a model may account for the main effects of contingency learning and blocking.

Alkon (1980), Hawkins and Kandel (1984) and Chang and Gelperin (1980) have all investigated the neural substrates of associative learning in invertebrate preparations; various claims about the extent to which these circuits and preparations actually perform classical conditioning in the mammalian sense have been made. In particular, Hawkins and Kandel (1984) speculate about the ways in which aspects of conditioning might emerge from lower-level processes. They do not distinguish between (a) explicit constraints of contingency-based classical conditioning vs. (b) simple associative learning, in which response to a CS is altered by its pairings with a US. For example, they claim (page 387) that "if unannounced [i.e., spurious] USs occur between pairing trials, the ability of the CS to predict the US is reduced and learning degenerates.... (Rescorla, 1968)". This does not enable us to tell whether the reported degradation of learning corresponds to the gentle degradation of partial conditions or the severe degradation of composite conditions. We go into this and some related problems in more detail in section 4.3.3.

It is with these results in mind that we have attempted to seek a uniform way of evaluating how these human and animal behaviors, mathematical algorithms, neurobiological circuits and computer models are related to each other. Our intent is to provide a both rigorous and understandable account of some major aspects of the computational, algorithmic and implementation attributes of contingency in classical conditioning. The following sections provide a somewhat detailed view of our progress so far.

4 Detail: The contingency computation, algorithm and implementation

4.1 The contingency computation

4.1.1 The theoretical formulation

As already described, Rescorla's (1968) computational constraint of contingency is that a specific presented positive CS-US association will be learned only if the probability of that US given that CS is greater than the probability of the US without the CS, or, formally, $p(US|CS) > p(US|\overline{CS})$. Reciprocally, "safety signals", i.e., CSs denoting the absence of a US, are learned only if $p(US|CS) < p(US|\overline{CS})$.

Church (1969) and Gibbon et. al. (1974) diagram the "space" of contingency-based learning by first plotting those areas in a plane corresponding to an association being learned and the association not being learned, according to this formula.

(FIGURE 3 GOES ABOUT HERE)

The two axes in figure 3 denote the likelihood that the US will occur given the CS (Y axis) vs. the likelihood that the US will occur given no CS (X axis). Above the diagonal line through the plane, the association will be learned (e.g., a tone CS signals a shock US); below the line, the opposite of the association will be learned (e.g., the tone is a "safety signal" that the shock will *not* occur). In both cases, the relative conditional probability constraint holds. On the diagonal itself, the probability of the US given the CS is *equal* to the probability of the US given no CS, so presentation conditions along that line will prevent the animal from learning any positive or negative CS-US association.

In this plane, we may also represent points corresponding to particular trial presentation conditions.[7] For example, in Figure 3, four points are presented, corresponding to an example of a partial reinforcement condition (point 1), composite misinformation (point 2) and partial warning (point 3) conditions, and the null condition (no CSs or USs presented; point 4). (No 'perfect pairings' condition is labeled). Points 1,2 and 4 correspond to those trial conditions used by Rescorla (1968); our experiment in progress includes a replication of those three points, and the addition of point 3, which represents a partial warning condition. In this Church–Gibbon plane, the perfect pairings condition is the point at (1,0) (upper

[7]Note that all the points represented in the figure are above the non-contingent line, simply denoting that the four conditions illustrated here were positive-association conditions; i.e., conditions in which the CS indicates that the US is coming, as opposed to 'safety-signal' conditions, which would appear below the line — such conditions have been tested, but they are not illustrated here.

left corner), the partial reinforcement condition corresponds to the left vertical axis, partial warning corresponds to the top (horisontal) axis, and the rest of the square corresponds to the class of composite conditions.

In the same paper described above, Gibbon et. al. (1974) expand their analysis to a three-space. Building on this work, we re-present and extend this analysis[8] by mapping the contingency results into a cartesian three-space (figure 4) in which the three axes correspond to joint (not conditional) probabilities: the Z axis is the probability of the CS and US both occurring (i.e., the probability of CS-US pairs: $p[CS, US]$), X is the probability of the CS and not US occurring (the probability of spurious (unpaired) CS trials mixed in: $p[CS, \overline{US}]$), and Y is the probability of the US and not CS occurring (the probability of spurious (unpaired) USs: $p[\overline{CS}, US]$).[9] These three joint probabilities must add to a total probability ≤ 1, so the overall space used to represent all possible sets of trial presentation conditions corresponds to the truncated cube bounded by the Z, X and Y axes and the plane $X + Y + Z = 1$.

Using this three-space, we can diagram the true contingency constraint, which appears as a 'saddle' shape in the space (Figure 4).

(FIGURE 4 GOES ABOUT HERE)

4.1.2 Interpretation of the contingency space

This contingency space can be broken down into regions that correspond to the four presentation conditions identified earlier (section 2.2).

(FIGURE 5 GOES ABOUT HERE)

The 'perfect pairings' condition is that in which no spurious (unpaired) CSs or USs occur: this corresponds to the Z axis itself (figure 5a). The partial warning condition is the plane defined by the

[8] Our 'saddle' graph of contingency was developed before we had seen the derivation by Gibbon et. al. (1974); we are gratified that we have independently arrived at compatible sets of results.

[9] The three axes of this space represent *joint* probabilities; the Rescorla constraint plotted in the space represents a comparison between two *conditional* probabilities (probability of US *given* the CS greater than the probability of US *given* the absence of the CS, or $p(US|CS) > p(US|\overline{CS})$). These two types of probabilities are distinct from each other, and are related as follows: $p(B|A) = p(A, B)/p(A)$. Furthermore, both of these types of probabilities are distinct from *marginal probabilities*, e.g., the percentages of of CSs and USs over trials. It is quite possible, for example, to change the percentage of CSs and USs in a set of trials without changing either the joint probability of CSs and USs ($p(CS, US)$) nor the conditional probability of a US given a CS ($p(US|CS)$). Similarly, two *different* values of the conditional probability of the US given the CS ($p(US|CS)$) could correspond to a *single* value of their joint probability ($p(US, CS)$). In general, varying the *number* of CSs or USs will *not* necessarily change the conditional or joint *probabilities*.

Z and Y axes (along the 'left side' of the space), since this is the set of cases in which both perfect pairings (Z) and spurious USs (Y) are included, but no spurious CSs are included, so X must have a value of 0 (figure 5b). The partial reinforcement condition is the plane defined by the Z and X axes (the 'right side'), since this condition includes pairs and spurious CSs, but no spurious USs (figure 5c). Finally, the composite-misinformation condition is all of the space between these planes. (The 'bottom' plane, defined by the X and Y axes with a Z value of 0 (figure 5d) would correspond to a 'completely unpaired' condition: i.e., no pairings, only presentations of spurious USs and spurious CSs. This special case of the larger composite misinformation category is one in which the negative 'safety-signal' interpretation of the CS will be readily learned, though the positive CS-US association will not). The actual area in which learning of a positive CS-US association is predicted to occur (by the Rescorla (1968) computational constraint of relative conditional probabilities) is 'behind' the saddle surface, i.e, within the area bounded by the surface and the Z axis. Within this area, the probability of the US given the CS is greater than the probability of the US without the CS ($p(US|CS) > p(US|\overline{CS})$). In front of the saddle, the opposite of the CS-US association will be learned; since $p(US|CS) < p(US|\overline{CS})$, the CS is learned to be a safety signal, indicating that the US will *not* occur. The saddle surface itself corresponds to the points at which $p(US|CS) = p(US|\overline{CS})$: directly on (and in the immediate vicinity of) the surface, the CS will be learned to be unassociated with the US (this corresponds to a 'truly random control' procedure, as discussed by Rescorla 1967, 1972).

Recall that each *point* on the Church–Gibbon plane corresponds to a different potential testing condition: the four points on Figure 3 correspond to a partial reinforcement condition (point 1), composite misinformation (point 2), partial warning (point 3) and the null condition (no CSs or USs presented; point 4). These points are typically denoted by the values of the two conditional probabilities to be compared: the probability of the US given the CS, and the probability of the US in the absence of the CS. for instance, point 1 corresponds to 0.4–0, i.e., $p(US|CS) = 0.4$ and $p(US|\overline{CS}) = 0$. Similarly, point 2 corresponds to 0.4–0.4, point 3 corresponds to to 1.0–0.4, and point 4 is 0–0.

A notable aspect of the three-dimensional saddle graph is the way in which it corresponds to the Church–Gibbon contingency plane: all the above presentation conditions, which are points in the plane, correspond to line segments in the contingency three-space. This is because, by the laws of conditional probability,

$$p(US|CS) = \frac{p(US,CS)}{p(CS)} = \frac{p(US,CS)}{p(CS,\overline{US}) + p(CS,US)}$$

(since the marginal probability $p(CS)$ in the denominator is simply equal to the sum of its joint

probabilities with or without the US). Now each of the three joint probabilities in the resulting equation corresponds directly to a value in the 3–space, so we have

$$p(US|CS) = \frac{Z}{X+Z}$$

Similarly,

$$p(US|\overline{CS}) = \frac{p(US,\overline{CS})}{p(\overline{CS})} = \frac{p(US,\overline{CS})}{p(\overline{CS},\overline{US}) + p(\overline{CS},US)} = \frac{Y}{(1-X-Y-Z)+Y} = \frac{Y}{1-X-Z}$$

Setting either of these two values, say $Z/(X+Z)$, to a particular constant value such as 0.4 defines a plane segment in the contingency space. Similarly, setting $Y/(1-X-Z) = 0$ defines another plane segment; the intersection of these two planes is a line segment. Individual Church–Gibbon squares also correspond to plane segments in this space, and the intersection of a particular Church–Gibbon square with the 0.4–0 line segment corresponds to a point. The set of all such points in the square make up the 0.4–0 line segment in the space. Different Church–Gibbon squares in the space correspond to different settings of $p(\overline{CS},\overline{US})$ — the probability of non-presentations of either the CS or US (see sections 4.2 and 5.5). (Hence, this standard method of specifying trial conditions is *underspecified*; a single specification such as 0.4–0 refers to a large number of different trial conditions. This leads to some counterintuitive predictions that are discussed further in section 5.5).

(FIGURES 6 AND 7 GO ABOUT HERE)

Figure 6 illustrates the line segments in the contingency space comprising the saddle surface that corresponds to the diagonal in the Church–Gibbon square. Each individual line segment represents a point on the Church–Gibbon diagonal. Figure 7 illustrates some of the presentation conditions tested by Rescorla (1968), with one partial warning condition (1–0.4) added. The conditions in which $p(US|CS) = p(US|\overline{CS})$ (0.4–0.4, 0.2–0.2, 0.1–0.1) are those that lie directly on the saddle surface. The 1.0–0.4 condition lies entirely within the partial warning plane. The 0.4–0, 0.2–0 and 0.1–0 conditions all lie entirely on the partial reinforcement surface. A 0.4–0.1 case would be in the space, on the 'inside' of the saddle surface (since this condition enables learning of the positive association); a 0.1–0.4 case would lie 'outside' the surface (in the negative-association area of the composite space). There are interesting consequences of these lines; section 5.5 explores these issues.

4.2 The contingency algorithm

4.2.1 Inputs and outputs of contingency algorithms

An algorithm for contingency must account for an animal's transformation of the inputs that are presented (e.g., trial sequences) into output categorisations of stimuli. The output categorisations can be thought of as differential assignments of associative strengths to different candidate CS-US pairs, where higher strengths would indicate that the animal 'believes' that the CS leads to the US (i.e., has acquired the association), lower strengths indicate the association has not been learned (or, perhaps more accurately, that the CS has been learned to be uncorrelated with the US), and negative strengths indicate that a negative association has been learned (the CS reliably signals the absence of the US). (Furthermore, we will show in Section 4.2.5 that 'context' cues can be formally distinguished from other learned correlated and uncorrelated stimuli. Essentially, the result presented there shows that an animal should be able to tell whether a particular stimulus behaves like a context cue with respect to some particular US, by determining not only the level of correlation of the cues but also noting that the context cue occurs extremely often, i.e., that $p(CS) \approx 1$ for context cues.)

In summary, the logical categories of output relationships that the animal can learn to discern are positive predictions, negative predictions and uncorrelated cues; and context cues can be distinguished from other cues (see Section 4.2.5). The inputs available to the animal are occurrences of features in the environment.[10] For simplicity, we can categorize the logically possible pairwise combinations of two arbitrary feature events F1 and F2 (which, for classical conditioning, correspond to the CS and US, respectively). Either:

(a) F1 occurs and then F2 occurs (which we will term a *successful prediction*),

(b) F1 occurs and then F2 does not occur (error of *commission* – i.e., the environment has 'committed' an 'erroneous' prediction of the F1-F2 sequence by the occurrence of an F1 event without F2 following it),

(c) F1 does not occur and then F2 does occur (error of *omission* – F1 is omitted from the F1-F2 sequence), or

[10] Note that we make the simplifying assumption that event occurrences may be described in terms of discrete time and trials. This is a common assumption in the learning literature (Rescorla and Wagner, 1972; Mackintosh, 1975; Pearce and Hall, 1980); see Section 5.5.2 for a discussion of some of the implications of this assumption.

(d) neither F1 nor F2 occurs (which we refer to as a *non-prediction*, or *non-presentation*).[11]

	F2 present	F2 absent
F1 present	++ Successful Prediction (s)	+− Error of Commission (c)
F1 absent	−+ Error of Omission (o)	−− Non-prediction (n)

Table 2: Possible combinations of F1 and F2

Predictions and non-presentations (non-predictions) both have the effect of strengthening the predictive value, or association, between F1 and F2 (since they either appeared together or failed to appear together), while errors of commission and omission weaken the association. This implies that learning occurs in part in the absence of stimuli, since a non-presentation is the absence of either CS or US. Some implications of this are discussed further in Section 5.5.

4.2.2 Rescorla's interpretation of contingency

Rescorla (1968) offers an algorithmic interpretation of the contingency data by suggesting that two separate, opposing processes are at work: an excitatory association develops as a result of CS-US pairings, and an inhibitory association grows with each spurious US. In a partial reinforcement situation (CS-US pairs with spurious CSs), the excitatory association is formed due to the presence of CS-US pairs, but no inhibitory association is formed, due to the lack of any spurious USs. In the composite condition (pairs, spurious CSs *and* spurious USs), the occurrence of spurious USs results in an inhibitory association which can cancel the excitatory association. This account fails to attribute any effect to spurious CSs.

The predicted net association resulting from a partial warning contingency (pairs plus spurious USs, with no spurious CSs) would therefore be a strongly inhibitory one: some excitatory association from CS-US pairs, but a potentially large inhibitory association arising from spurious USs. Simply, since

[11] Non-predictions (non-presentations) are simply the absence of the two features; if all such non-predictions were counted, there would be a huge, ongoing number. All algorithms must systematically undercount the 'true' number of non-predictions. The method proposed as part of our algorithm (Section 4.2.5) is to only consider a non-prediction to have happened when F2 has been predicted but did not occur. The issue of the role of non-predictions in contingency algorithms remains a crucial one, since the conditional probabilities at the heart of the computational constraint cannot be said to have been calculated without non-predictions being taken into account (see section 5.5).

spurious CSs are hypothesized by this account to have little or no effect on the outcome of conditioning, no distinction is made between the composite and partial warning conditions. Learning is predicted to be severely degraded in both cases, as spurious US trials are introduced. This contradicts the constraint that $p(US|CS) > p(US|\overline{CS})$, which predicts severe degradation in the composite condition, but only gentle degradation in partial warning (as in partial reinforcement).

4.2.3 The Rescorla–Wagner algorithm and partial warning

Rescorla and Wagner (1972) propose an algorithm based on the idea that a single associative strength changes incrementally over trials. That is, as the result of a particular trial, the total associative strength of each of the components (A and X) of a stimulus compound (AX) is increased or decreased by an amount proportional to the size of the combined associative strength of A and X:

$$\Delta V_A = \alpha_A \beta (\lambda - V_{AX})$$

and

$$\Delta V_X = \alpha_X \beta (\lambda - V_{AX})$$

where α and β correspond to salience measures of the CS (A or X) and US, respectively, and λ is the highest (asymptotic) level of associative strength that the particular US is assumed to be capable of supporting (it is assumed that different USs will yield different λ levels).

A crucial assumption underlying Rescorla and Wagner's algorithm is that all potential CSs which could be conditioned to a single US are *competing* against each other for their share of the total available associative strength (λ). Rescorla and Wagner (1972) argue that this competition effect gives rise to a number of desirable features of the model, such as Kamin blocking (1968).

The general line of reasoning in the analysis is that as one stimulus increases in predictive power over other competing ones, the associative strength of the competitors is stolen by the associative strength of the predictive stimulus. The context (e.g., the cage) is thought of here as yet another competing cue, and hence spurious USs could be thought of as strengthening the associative strength of the context, since treating the context as a candidate CS allows the view that 'spurious' USs are occurring in the presence of the context CS.

Rescorla and Wagner offer the argument that in a partial reinforcement condition, the US never occurs in the presence of the context without the CS also being present, so the context has no chance to 'steal' associative strength from the CS. In contrast, in the composite misinformation condition, the US

sometimes occurs in the presence of the context and sometimes in the presence of the CS plus context, and hence the context has opportunities to decrement the strength of the CS-US association.

As before, problems arise when we try to apply this account to the partial warning condition. The US occurs often in the presence of the context with no CS, just as in the composite case, which should lead to the same strengthening of the context as in the composite case; at the same time, there are no more CS-US pairs in this condition than in either the partial reinforcement or composite conditions. This implies that the algorithm will not learn the CS-US association in the partial warning case; yet in this case conditioning *is* predicted by the contingency constraint, since $p(US|CS) > p(US|\overline{CS})$ in all partial warning conditions (1.0–0.6, 1.0–0.4, etc). In fact, the only difference between this partial warning case and the composite case is the lack of spurious CSs in the former. Hence, an explanation of why learning occurs in one condition and not in the other can only rest on an account of how the existence of extra unpaired CSs can either strengthen the association between the context and the US or weaken the association between the CS and the US — and yet these unpaired CSs must not have this effect in the partial reinforcement condition! In other words, the authors' account of the operation of this algorithm offers no way to provide a consistent explanation of why conditioning to the context should prevent learning in the composite condition, but not in the partial reinforcement or partial warning conditions.

It is still possible that this algorithm will predict learning correctly in the partial warning condition; this question may be quantitatively tested regardless of problems of interpretation of the qualitative account. We have performed simulations of the Rescorla–Wagner algorithm with a range of parameter settings (see Appendix B) which show that, in the partial warning condition, learning of the CS-US association is severely degraded with the addition of spurious USs. This indicates that, under the conditions we have tested (and have reported in Appendix B), the algorithm is predicting that the the partial warning condition behaves like the composite condition, rather than like the partial reinforcement condition, in contradiction to the Rescorla contingency constraint, which predicts the same gentle degradation in partial warning as in partial reinforcement. Since Rescorla and Wagner also offer a derivation showing that the algorithm should compute the precise Rescorla constraint, there appears to be an important discrepancy; further investigation of the relationship between this algorithm (representing the mechanism) and computation (which is its intended output) is called for.

4.2.4 Contingency vs. strengthening-and-weakening

As in the case of possible discrepancies between Rescorla-Wagner (1972) and Rescorla (1968), it is often not obvious just what a particular algorithm will compute, so that it is often difficult to tell whether a particular algorithm conforms to the computational constraint of contingency. A number of algorithms proposed to simulate aspects of learning in general (though not classical conditioning in particular) do so by variants of a basic mechanism that *strengthens* an association upon successful pairings of the CS and US, and *weakens* the association on unsuccessful pairings, i.e., when either the CS or US occurs unpaired (e.g., Spence 1936, Anderson, 1983; Langley et. al., 1983). We will show that this intuitively natural mechanism cannot be made to conform to the computational constraint of classical conditioning, and cannot be an algorithm for this particular form of learning.

Any account that depends on a linear strengthening/weakening algorithm (henceforth 'S/W' algorithm) will correspond to an equation in which the incremental change in associative strength of a stimulus A (ΔV_A) changes as an additive function of the three axes of the contingency space. That is, all S/W algorithms yield an equation of the form:

$$\Delta V_A = \alpha(\gamma Z + \delta X + \sigma Y + \rho(1 - X - Y - Z))$$

Any such additive equation in this space will always and only give rise to a planar surface denoting the 'non-contingent' boundary, i.e., the boundary between learning positive and negative associations. Since this boundary will be a plane in the space for S/W algorithms, it can never be more than a planar approximation of the (non-planar) saddle surface. For much of the space, the plane can be placed in such a way that is a reasonable approximation of part of the saddle. This is only true, however, as long as either the partial reinforcement condition or the partial warning condition is ignored. These two conditions correspond to the areas of the saddle surface which curve away from the S/W plane. As long as these algorithms do not try to account for both partial reinforcement and partial warning, it is possible to present models of conditioning that generate planar approximations of either the left or the right portion of the true contingency saddle surface, and correspondingly will approximate the predictions of either partial reinforcement or partial warning learning, but not both.

(FIGURE 8 GOES ABOUT HERE)

(Figures 8a and 8b show two different placements of a 'strengthening/weakening' plane that approximate the partial warning and partial reinforcement portions of the contingency surface, respectively). Once both the partial reinforcement and partial warning conditions together are taken into account,

it will be seen that there can be no placement of the S/W plane that will serve as even the roughest approximation of the contingency saddle surface. The reason is simply that S/W algorithms do not differentiate between spurious CSs and spurious USs; all additions of misinformation to these algorithms are viewed as composite misinformation. It is by distinguishing among the types of misinformation (unpaired CSs, unpaired USs, composites) that the correct contingency computation can be achieved.

4.2.5 A new algorithm for contingency

Bayesian statistics (Bayes 1763, Pearl 1983, Skyrms 1966) provide formulae for the calculation of two values in inductive logic: *Logical Sufficiency (LS)*, which indicates the extent to which the presence of one event predicts, or increases the expectation of, another particular event; and, reciprocally, *Logical Necessity (LN)*, which represents the extent to which the *absence* of an event *decreases* expectation or prediction of the second event. LS and LN are defined to be:

$$\text{LS} = \frac{p(F2|F1)}{p(F2|\overline{F1})} \qquad\qquad \text{LN} = \frac{p(\overline{F2}|F1)}{p(\overline{F2}|\overline{F1})}$$

If we consider F2 to be a US and F1 a CS, then note that when $LS > 1$, it is also true that $p(US|CS) > p(US|\overline{CS})$, and vice versa. Additionally, $LS = 1$ if and only if $p(US|CS) = p(US|\overline{CS})$, and $LS < 1$ iff $p(US|CS) < p(US|\overline{CS})$.[12] The values of LS and LN may be calculated by a pair of simple formulae composed of precisely the four possible input categories of pairwise features occurrences given in section 4.2.1 above:

$$\text{LS} = \frac{s(n+o)}{o(s+c)} \qquad\qquad \text{LN} = \frac{c(n+o)}{n(s+c)}$$

where s is the count of successful predictions, c is errors of commission, o is errors of omission, and n denotes non-predictions (non-presentations).

For each biologically-salient cue (i.e., US) that the animal has learned about, the animal is assumed to be maintaining simple memories of these counts of successes, omissions, commissions and non-presentations[13], from which LS and LN are derived as shown above. At any given time, (e.g., on a particular trial), the animal calculates its *level of expectation* that the US might occur, based on these stored values. The actual algorithm is as follows: assume a number of candidate CSs (e.g., CS_1, CS_2, CS_3) that have been experienced in conjunction with a particular US; then there are existing counts

[12] Since $p(\overline{US}|CS) = 1 - p(US|CS)$ and $p(\overline{US}|\overline{CS}) = 1 - p(US|\overline{CS})$. Furthermore, it can be shown that $LS > 1$ if and only if $LN < 1$. However, it is not true in general that $|LN - 1| = |LS - 1|$ or that $LS = LN$

[13] With the proviso given in Section 4.2.1 that non-presentations will be systematically undercounted.

in memory for the associations of each of these CS_is with this particular US. At a given trial, assume some number of cues actually occur (e.g., CS_1 and a new, as-yet-unobserved cue, CS_4). Then the level of expectation of the US is calculated by multiplying the LS values of those cues that occurred (in this case, CS_1 and CS_4 with the LN values of those cues that did not occur (but had been seen before: CS_2, CS_3). This has the effect of combining the extent to which the cues that are present increase expectation of the US (LS) with the extent to which the cues that are absent decrease expectation of the US (LN).

This illustrates the reason that we make use of separate values for LS and LN, rather than only maintaining a single 'associative strength' for each cue, as, for example, Rescorla and Wagner do: there is somewhat different information being learned about the effect of the absence of a cue than the information learned about the effect of its presence. It can also now be seen that this algorithm is not a 'competitive' one in the sense that the Rescorla–Wagner algorithm is: as an individual cue gains associative strength with respect to the US in the Rescorla–Wagner algorithm, that cue is 'stealing' associative strengths of other, competing cues. In our algorithm, the LS and LN values of each cue progress independently of each other with respect to a US, and then all such values are used 'cooperatively' to compute the level of expectation of the US at any given time.[14]

This use of LS and LN to compute levels of expectation of the US can be viewed in terms of the extent to which individual cues are being 'categorized' by the animal as positive or negative predictive cues, as context cues, or as uncorrelated cues. LS values range from 0 to ∞, with high LSs corresponding to a particular feature (CS_i) strongly predicting a second feature (the US), since high LS implies a high ratio of successes to errors of commission; and very low LSs corresponding to the case where the CS implies that the US will *not* occur (low ratio of successes to commissions). Hence, for a high LS value, CS_i is a positive predictor of the US; for low LS, CS_i is a negatively predictive cue, i.e., the presence of this CS predicts that the US will not occur. An LN value near 1 indicates that the absence of a cue may be ignored, while a low LN value (near zero) indicates that presence of the cue is necessary for prediction.

When the value of LS is approximately 1, i.e., neither very high nor very low, then the cue CS is uncorrelated. A context cue, i.e., one that that occurs with an extremely high frequency, may be identified by simply computing $p(CS)$: when this is approximately equal to 1, the cue is appearing almost all the time (in every trial), and is a candidate context cue. Calculation of this probability is

[14] Our 'cooperative' algorithm also proposes new boolean combinations (conjunctions and disjunctions) of features as independent cues; these composite cues then build up their own LS and LN values independently of their constituents. This is described in some detail in Granger, Schlimmer and Young (1985) and Granger and Schlimmer (1985).

straightforward: $p(CS) = (s + c)/(s + c + o + n)$.

Again, what is being described is a way in which the use of LS and LN for calculating levels of expectation of a US can be viewed as an approximate categorisation of cues by their predictiveness. This view can be summarised as follows:

Positive cue	LS >> 1, LN << 1
Negative cue	LS << 1, LN >> 1
Uncorrelated	LS ≈ LN ≈ 1
Context	$p(CS) \approx 1$

These categories roughly capture how the animal's behavior will reflect its internal LS and LN values (and hence its level of expectation of the US). It is not the case that any given cue is necessarily categorised 'all-or-none' as either, say, a positive cue vs. a context cue. Any given cue is more usefully viewed as having attributes of a number of these categories, so that a particular cue may be viewed as, for instance, a weak positive predictor (say .4) and a somewhat stronger context cue (.6). It is the actual levels of expectation calculated from LS and LN values by the algorithm that are the true internal measurements of what has been learned.

4.2.6 Gathering evidence: Incremental operation of the algorithm

We have constructed a computer simulation of the algorithm to illustrate its operation; this section describes that simulation (Granger, Schlimmer and Young (1985) and Granger and Schlimmer (1985) contain further discussion of the algorithm and the computer simulation). This section provides a brief overview of the operation of the program.

All counts in memory are initially set to 1.[15] These counts are updated only when a memory trace (corresponding to a feature complex) is triggered by matching cues in the environment, at which point the matched trace becomes the source of predictions of what will happen and what behaviors are associated with these predictions.

This trace is matched against new events. When a prediction succeeds, the success scores of matched features in the environment are incremented. Cues failing to match receive incremented omission scores. If a prediction fails, each cue feature that matched the environment scores a commission; each cue feature that was absent from the environment, a non-prediction. Novel features present in the environment are

[15] In fact, any Bayesian algorithm must start with some arbitrarily-chosen initial probability values; the choice of values will not change the overall operation of the algorithm, though it may affect the initial learning of a novel stimulus.

	s ++	*c* +−	*o* −+	*n* −−	LS	LN
Cage	52	11	1	1	1.65	0.35
Tone	52	7	1	5	5.29	0.14
Light	52	8	1	4	4.33	0.17
Buzz	19	4	34	8	1.02	0.91
Whirr	43	10	10	2	0.97	1.13
And[Tone,Light]	48	3	1	5	5.65	0.07

Table 3: Positive Contingency

added with an initial score of 1 commission, 1 prediction, 1 omission, and 1 non-prediction.

Assume a situation where tones, lights, noises, and shocks are occurring. The program's task is to construct a memory record which will allow it to predict the occurrence of the shock accurately (presumably in order to avoid it). Specifically, given a positive contingency situation, i.e., one in which the shock is reliably preceded by a conjunction of features (e.g., tone and light), a table representing a portion of memory about the shock will look similar to table 3. (Note that successes are indicated by '++', commissions by '+−', omissions by '−+', and non-predictions by '−−'. The figures in table 3 are taken from runs of an early version of our computer model.)

To reiterate, the LS (logical sufficiency) value indicates the degree to which a cue is sufficient to cause expectation of a result feature, with values greater than 1 indicating a positive contribution to expectation. The LN (logical necessity) value indicates the degree to which absence of a cue precludes expectation of a result feature. An LN value near 1 indicates that absence of a cue may be ignored, while an LN value near zero indicates that a cue is necessary for expectation. (An interesting sidelight is that the *conjunction* of light and tone has been proposed by the program itself: see discussion in Granger, Schlimmer and Young (1985); Granger and Schlimmer (1985)).

This chart illustrates important differences between contingency learning on the one hand and strengthening/weakening algorithms (based on number of pairings) on the other. Cage and tone receive the same number of pairings with shock, but tone is a much better predictor of shock. Moreover, tone was involved in a greater number of mistaken predictions (errors of commission) than was buzz, but tone is still recognized as the better predictor.

4.2.7 Summary: Performance of the algorithms

We have discussed Rescorla and Wagner's (1972) algorithm, the class of strengthening/weakening (S/W) algorithms and our new proposed contingency algorithm based on the calculation and use of sufficiency and necessity (LS and LN) values. We have performed simulations of all three categories of algorithms, and summarize here our findings on their performance. Appendix B contains a set of results of simulations using all three algorithms.

S/W algorithms will learn a positive CS-US association appropriately in the perfect pairings (PP) presentation condition and performance will fall off severely (again appropriately) in the composite misinformation (C) condition. However, degradation of learning in the partial warning (PW) and partial reinforcement (PR) cases is indistinguishable from the composite case for S/W algorithms; this of course contradicts the contingency constraint, which predicts severe degradation in the composite condition but very gentle degradation in both partial conditions (see Section 2.1).

The Rescorla-Wagner algorithm learns appropriately in the PP and PR conditions, and is severely degraded (appropriately) in the C condition. However, in our simulations of the Rescorla-Wagner algorithm on the PW case, learning is just as severely degraded as in the C condition, not gently as in the partial reinforcement condition. Further investigation and interpretation of these results are required.

Our algorithm is based directly on the contingency constraint, and so it will learn appropriately in all four presentation categories: it shows severe composite degradation and only gentle partial degradation. Like the other two algorithms, it requires no complex or counterintuitive calculations on the part of the animal; the correct constraint arises naturally from a set of simple operations. The algorithm also accounts naturally for blocking, and provides an account of aspects of learned irrelevance, latency and tracking of changes in the environment (see section 5). We are continuing to apply the algorithm to a range of conditioning phenomena, to test its breadth and range of usefulness.

4.3 Circuits for contingency

4.3.1 The evaluation of the adequacy of proposed circuits

The neurobiology of learning and memory involves the search for biological mechanisms that underlie, and, by their operation, give rise to overt learning behavior. Associative learning is an area in which a great deal of recent progress has uncovered a number of competing candidates for the biological

mechanism underlying the class of phenomena comprising classical conditioning (e.g., Hawkins and Kandel (1984), Alkon (1980), Chang and Gelperin (1980), Thompson et. al. 1984)), in addition to a number of mathematical and computational models of these proposed mechanisms (e.g., Sutton and Barto (1981), Gluck and Thompson (1985), Hampson and Kibler (1983)). Models of this kind have focused primarily on the temporal constraints on classical conditioning, e.g., the interstimulus and intertrial intervals (ISI and ITI); and most have also attended to the constraint of conditional probability in contingency.

The problem to be addressed here is: how can we determine which (if any) of these proposed mechanisms may be correct ones for classical conditioning — or, in other words, how can competing mechanisms be evaluated against each other, and against the (behavioral-level) classical conditioning data. In order to determine which might be valid candidate classical conditioning mechanisms, each proposed mechanism must be tested to see that its performance conforms (at least) to the known attributes of classical conditioning, such as range of interstimulus and intertrial intervals, blocking effects, and conditional probability (contingency) effects.

4.3.2 Categories of mechanisms

Without conforming precisely to known computational constraints, any given candidate mechanism may turn out to be a mechanism for *some* form of associative conditioning, but not *the* particular set of algorithms that mammals use to perform associative learning in classical conditioning situations.

What might it mean for a mechanism to conform to many, but not all, of the constraints of contingency in mammalian classical conditioning? Imagine a proposed biological mechanism that exhibits behaviors resembling mammalian classical conditioning (MCC), but is not identical to them, and so cannot be *the* complete mechanism that underlies such learning. We may distinguish among three categories of proposed biological mechanism for mammalian classical conditioning:

(a) *Insufficient (or incomplete) mechanisms* are those which do not successfully give rise to the phenomena of mammalian classical conditioning, either because the mechanism is incorrect or, possibly, is only one component of some larger, as-yet-undiscovered mechanism.

(b) *Taxon-specific mechanisms* are those which accurately reflect the associative learning abilities of some particular taxonomic category (e.g., class, order, or phylum) of animal, but in which that animal's classical conditioning behavior can be shown to be distinct from mammalian learning in some specific identifiable fashion. Such a mechanism is a correct classical conditioning mechanism,

but is not a correct mammalian classical conditioning mechanism. For example, if it turns out that, although mammals learn in the partial warning presentation condition, *Aplysia* does not do so, then it still may be the case that the proposed Hawkins and Kandel (1984) mechanism for *Aplysia* classical conditioning might indeed be the circuit that performs classical conditioning in *Aplysia*, but it would not then be the case that that same circuit mechanism is the one that underlies conditioning in mammals.

(c) *Mammalian classical conditioning* (*MCC*) mechanisms are those biological mechanism(s) that underlie the performance of actual mammalian conditioning.

	animal x does not do *MCC*	animal x does *MCC*
mechanism is insufficient for *MCC* in animal x	incorrect or incomplete mech	incorrect or incomplete mech
mechanism is sufficient for *MCC* in animal x	taxon–specific mechanism	*MCC* mechanism

Table 4: Evaluation of proposed biological mechanisms

Of these three, (a) (incorrect or incomplete) simply represents the class of mechanisms that cannot be shown to perform the right behaviors, and calls for further exploration; (b) (taxon–specific) represents the possibility that different groups of animals perform associative learning differently — this is a sensible possibility in that the "point" of associative learning is to note and learn about regularities in the environment, and there may be many differing mechanisms that have evolved to instantiate different versions of this regularity-detecting ability. From a computational point of view, certain taxon–specific (e.g., phylum–specific or class–specific) mechanisms may be useful approximations of a true mammalian classical conditioning mechanism, but from a biological point of view, such a phylum–specific mechanism cannot indiscriminately be considered to be the same as mammalian classical conditioning: differences that appear almost insignificant may likely point to biological differences that are crucially important. It is more useful to identify both the similarities and differences among distinct animal phyla, rather than simply using one as a convenient approximation of another as though the differences were not important. Finally, class (c) (mammalian classical conditioning) represents those mechanisms that may actually underlie classical conditioning in mammals – there may still be differences among mechanisms across species, or even within a single individual.

4.3.3 Computational analysis of *Aplysia*

Hawkins and Kandel (1984), on p. 387, briefly discuss a trial-presentation condition that directly corresponds to the partial warning condition, and suggest how learning may proceed in this condition. They begin by stating that "In classical conditioning, animals do not simply learn that the CS precedes the US (contiguity), but they also learn the contingency or correlation between the CS and US"; they go on to say that

> ... if unannounced [i.e., spurious] USs occur between pairing trials, the ability of the CS to predict the US is reduced and learning degenerates. In the limit, if the probability of unannounced USs is the same as the probability of announced (paired) USs so that there is zero contingency, animals do not learn to associate the CS and US despite the fact that they are paired together many times (Rescorla, 1968).
>
> Rescorla and Wagner (1972) proposed that this effect could be explained by an extension of the argument they advanced for blocking ....
>
> In [a] hypothetical example the addition of unpredicted USs would not only cause a decrease in the difference between the strengths of the CS^+ and CS^-, but would also cause a decrease in the absolute strength of the CS^+. Results similar to those shown ... have recently been obtained in *Aplysia* in an experiment ... (Hawkins, Carew, & Kandel, 1983). [Hawkins and Kandel, 1984, pp. 387-388]

These statements deserve careful examination in light of our computational analysis of contingency effects in classical conditioning. First of all, Hawkins and Kandel state that as spurious US trials are added (presumably to CS-US pairs), learning degenerates. As we have seen, learning is predicted to degenerate to some extent in all conditions that have spurious trials, but the key difference between true contingency and other possible classical conditioning mechanisms (such as strengthening/weakening algorithms, section 4.2.4) is that in contingency-based conditioning, the composite condition severely degrades learning of the CS-US association, while the two partial conditions (PW and PR) only gently degrade this learning. It is, therefore, this distinction between partial and composite conditions that must be tested experimentally in order to determine what this circuit (and animal) is actually computing.

Hawkins and Kandel go on to state that "in the limit", learning should be degraded to zero with the addition of enough spurious US trials. Since they seem clearly to be describing a partial warning condition, with no spurious CS trials, this limit will only be reached when $p(US|CS) = p(US|\overline{CS}) = 1$,

which can only happen when there is a US presented in *every trial*. Assuming that this is not what Hawkins and Kandel meant, this again calls for the crucial distinction to be made between the partial warning versus composite cases: in the latter, 50% spurious US trials will degrade the learning to zero, since this is the severe-degradation case, but in partial warning, it takes 100% USs to degrade learning to zero.

Hawkins and Kandel cite Rescorla (1968) and Rescorla and Wagner (1972) for explanations of the degradation of learning, but again it is the case that these cited papers explain the *difference* between partial reinforcement and composite conditions; furthermore, neither paper mentions the partial warning condition, i.e., any condition in which spurious USs but no spurious CSs are added to pairing trials.

Finally, the initial experiments referred to by Hawkins and Kandel demonstrate degradation of learning in *Aplysia*, but it is the distinction between the gentle degradation of the partial conditions versus the severe degradation of the composite condition that should be experimentally tested.[16] In the absence of testing for this distinction, it cannot be determined whether the *Aplysia* circuit is performing contingency-based classical conditioning (as mammals do), or some form of learning that is distinct from this contingency-based conditioning.

5 Breadth of the theory: blocking, latency, tracking, learned irrelevance

5.1 Blocking

The failure of an animal subject to form an association with the novel component of a compound stimulus following successful classical conditioning to the familiar component is called *blocking*. Kamin (1968) originally demonstrated this effect by first training animals to associate a noise with a shock. Then animals were repeatedly presented a compound of light and noise followed by a shock. Upon testing, the animals demonstrated little or no conditioning between the light and the shock; the previous effective conditioning of the noise to the shock 'blocked' subsequent conditioning to the light.

[16] Gluck and Thompson (1985, 1986) have constructed a computer simulation of Hawkins and Kandel's (1984) *Aplysia* circuit mechanism. While it is often quite difficult to test a wide range of behaviors in a circuit preparation, Gluck and Thompson's simulation of the circuit may be analyzed to see how it actually behaves under various circumstances. We are currently collaborating with Gluck and Thompson to test whether the model satisfies the behavioral constraints identified above.

All accounts of this effect concur that expectation on the part of the animal is crucial, for the light offers no new information about the onset of the shock. Rescorla and Wagner (1972) offer an account in which stimuli *compete* for a limited amount of associative strength. A single stimulus may acquire the complete amount; subsequent stimuli compounded with this previously conditioned cue must compete for associative strength with the completely effective cue, and thus acquire no association. Mackintosh (1975) explains that the animal may instead be learning not to pay attention to the redundant stimulus. The animal then simply does not modify associative strengths for the new stimuli since no unexpected US occurred.

Our account is similar to Mackintosh's, in that there is no competition for a limited resource. Like each of the two other accounts, learning only occurs when expectation fails: either the shock is not expected and it is received (an error of expectation *omission*) or the shock is expected and is not received (error of expectation *commission*). When one stimulus comes to predict the US completely, no additional associational modifications are made until that stimulus is no longer accurate. A rough differential prediction may be made between our account and Mackintosh's: in Mackintosh's account the lack of attention to the redundant cue is a residue of the blocking experiment; in our algorithm, when the contingencies of the experimental setup change, we would predict that animals would demonstrate little hesitancy to form an association with the previously redundant cue.

5.2 Latency

Another characteristic of the classically-conditioned animal is the delay between the onset of the CS and the animal's response. A salient feature of this latency is that it tends to be as big as the delay from the onset of the CS to the onset of the US in classical conditioning, but for the same animals in an instrumental conditioning task, the response latency tends to be quite short. A representative experiment performed by Wahlsten and Cole (1972) demonstrates just this difference. Subjects were divided into classically and instrumentally conditioned groups. For both groups a CS signaled an aversive US: in the classical group, the US was unavoidable; in the instrumental group, the US was terminated by the CR of the animal. Subjects in the classical group waited until just before the onset of the US before responding, whereas subjects in the instrumental group originally waited as long as the classical animals did, but then began to make the response immediately following the onset of the CS; the animals are making a response as early as is effective. This could be accounted for by assuming that the animal "experimented" with smaller response latencies. For the classical subjects, this would prove

useless because the US is unavoidable. The instrumental subjects, however, would initially just lessen the impact of the US, but through continued shortening of their response latency would come to avoid it altogether. Further details of this theoretical viewpoint and a simulation may be found in Granger, Schlimmer and Young (1985).

5.3 Tracking changes in the environment

Subjects adapt to changes in their environment over time. For instance, the fox adapts to the seasonal coat of his prey, and a one-legged bird will learn to change its landing behavior. This ability to track changes over time is another computational constraint which may be used to test proposed learning algorithms. Rescorla and Wagner (1972) and Mackintosh (1975) utilise a formula which allows a reversal of the sign corresponding to the increment of an association's strength (ΔV). This enables the algorithm to switch from strengthening a previously successful association to weakening it when it is no longer effective. Our algorithm is not based on a formula describing a change in associative strength, but on the calculation of associativity based on a history of a cue's effectiveness. As that history reflects changes in the environment, the associative strength assigned to a concept changes as well. For instance, as the environment changes over time, some previously predictive cue might become non-predictive, in which case predictions would start failing, and the ongoing count of successful predictions would slowly be overtaken by the growing counts of commissions and omissions. Reciprocally, if a previously unpredictive cue becomes predictive, it will get re-introduced as a potential cue, and its successful predictions will allow its LS value to grow. Similarly, tracking changes in boolean feature combinations follows naturally from a thresholding effect associated with the formation of those combinations (Granger and Schlimmer, 1985).

5.4 Learned Irrelevance

The reluctance of animals to form different associations between a previously associated CS and US includes results from *learned irrelevance*. A set of experiments by Siegel and Domjan (1971) tested five conditions where the subjects were pre-exposed to the CS, to the US, to an uncorrelated presentation of the CS and the US, to a backward pairing of the US and CS, and given no pre-exposure. These animals were then placed in a standard excitatory contingency situation. They found that the rate at which subjects acquired the new association was ordered from greatest to smallest as follows: animals with no pre-exposure learned most quickly, followed by pre-exposure to the CS or to the US, uncorrelated

pre-exposure to the CS and US, and finally the backward pairing group, which was the slowest to form an association. Learned irrelevance refers to the difference between (a) the effect of pre-exposure to the CS or to the US, and (b) the effects of receiving pre-exposure to an uncorrelated presentation of the CS and US. In the latter condition, the CS is initially learned to be irrelevant to the US, while in the former condition no such relationship is present.

Mackintosh's (1975) model of selective attention would account for this in terms of a gradual reduction of the stimulus-specific learning parameters which represent attention. After an uncorrelated presentation of the CS and US, little attention would be paid to the CS and subsequent excitatory conditioning would be inhibited. The Rescorla and Wagner (1972) model might account for learned irrelevance if an association were formed between the context and the US in the uncorrelated condition. This context association might then block the further acquisition of association on the part of the CS during the excitatory conditioning. While conditioning to the context certainly does occur, this model would predict that no subsequent learning to the CS would be demonstrated. Our model explains the difference between the pre-exposure to the CS or US group and the pre-exposure to the uncorrelated presentation group by specifying that the associative calculations on the part of the animal are based on the history of association between the CS and US. By retaining the counts of event types, the computation is not based solely on the present association as it is in the delta models of Rescorla–Wagner and Mackintosh, but rather on the resultant of the previous values of these measures. In other words, all three models (ours, Mackintosh's and Rescorla and Wagner's) provide accounts of blocking and tracking changes over time in the environment. Our algorithm, however, sometimes *resists* tracking a change in accordance with learned-irrelevance data, while the Rescorla–Wagner algorithm will sometimes tend to track changes in the environment "too well", i.e., their algorithm will change more readily than animals will (according to learned-irrelevance data) in response to environmental changes.

5.5 Time, background and probability

5.5.1 Underspecification of trial conditions

In section 4.1.2 it was shown that trial presentation conditions (e.g., 0.4–0, 1.0–0.4, etc.) correspond not to points but to line segments in the contingency space (figures 6 and 7). This fact implies that this standard method[17] of specifying a testing condition is *underspecified*: there are multiple different testing conditions that would all be describable as, say 0.4–0.2. The contingency constraint means

[17] Used extensively by Rescorla (1967, 1968, 1972), Rescorla and Wagner (1972), Mackintosh (1983), etc.

that excitatory conditioning should hold in all 0.4–0.2 conditions, but any attempt at *replication* of an experimental condition that is only described as 0.4–0.2 may be confounded by lack of information about *which* 0.4–0.2 condition is meant.

(FIGURE 9 GOES ABOUT HERE)

Imagine two different testing conditions A and B that both lie along the 0.4–0.2 line segment in contingency space (Figure 9); just what are the differences between these two points? What is it that is changing as we travel along the line from point A to point B?

Point A contains fewer CS-US pairs (since its Z value is lower), fewer spurious CSs (since its X value is lower), and slightly more spurious USs (since its Y value is a bit higher) than point B There are not enough extra spurious USs to make up for the smaller number of pairs and spurious CSs; what is substituted are more 'non-presentations' (see section 4.2.1) at point A than at point B. That is, the set of trial conditions described by point A contain more 'events' in which neither the CS nor US occurs than that described by point B. Since a certain amount of time is allocated to the overall set of trials, these events are translated into 'empty' time durations. For purposes of replication, then, a complete specification of a trial–presentation condition would require more information than just the two conditional probabilities of contingency. An alternative formulation would offer these two conditional probabilities as well as the number of CS-US pairs and the total number of trials or total amount of time allocated for presentations to the animal. The trial–presentation condition corresponding to point A might be specified as [0.4–0.2;25/100(8hr)], denoting that $p(US|CS) = 0.4$, $p(US|\overline{CS}) = 0.2$, with 25 pairings presented over 100 total trials (for a total duration of 8 hours),[18] thereby specifying the joint probability of pairings being presented $p(CS, US) = 25/100 = 0.25$. Similarly, then, the condition corresponding to point B might be specified as (0.4–0.2;35/100(8hr)), denoting that in this condition there were 35 pairs over 100 trials: $p(CS, US) = 0.35$.

These additional numbers are required because for complete (replicable) specification of a trial–presentation condition we need to know each of the marginal or joint probabilities corresponding to the X, Y and Z axis values ($p(CS, \overline{US})$, $p(\overline{CS}, US)$ and $p(CS, US)$). By the laws of conditional probability, we know that $p(US|CS) = p(CS, US)/p(CS)$ and $p(US|\overline{CS}) = p(\overline{CS}, US)/p(\overline{CS})$. In the new proposed specification, we have $p(CS, US)$ (the Z axis value) directly derivable as the ratio of the number of pairings and the total number of trials (or the total amount of time for trials times the amount of time per trial). For point A, the Z value is 0.25; for point B, it is 0.35.

[18] This 'total-time' value is redundant with that of the total number of trials *if* the time allocated per trial is specified.

Then we can compute $p(CS) = p(CS,US)/p(US|CS)$, and $p(\overline{CS}) = 1 - p(CS)$, and thereby compute the Y axis value $p(\overline{CS},US) = p(US|\overline{CS})p(\overline{CS})$. The Y value for point A is 0.075, and for point B it is 0.025. Finally, all that is left to compute is the value of the X axis by the equation $p(CS,\overline{US}) = p(CS) - p(CS,US)$, to completely constrain the point in the space (for point A, $X = 0.375$, for point B, $X = 0.525$).

In summary then, reporting the two conditional probabilities, the number of pairings, and the total number of trials (or total trial time) is sufficient to completely specify the training conditions.

The theoretical formulation of contingency in fact requires that these 'non-presentations' or 'empty trials' be taken into account; different theories have handled this in different ways. Rescorla and Wagner (1972) presume that all such 'empty' trials are in fact exposures to the context; this is another way of viewing what it means for the context to compete with other CSs for associative strength in their theory. Most recent theories of conditioning (e.g., Mackintosh 1975; Dickinson 1980; Pearce and Hall 1980) adopt variations of this idea.

In contrast, our theory represents the context as an independent candidate CS like all the others; the difference is that we explicitly identify context cues, since, as we showed in Sections 4.2.1 and 4.2.5, we can mathematically distinguish between context cues and other types of predictive and uncorrelated cues. The implication is that the animal is capable of learning the extent to which particular cues are predictive cues (normal CS+, either positive or negative safety signals), are uncorrelated (CS-) or are context cues.

Rescorla and Wagner, therefore, deal with time *implicitly* by interpreting 'non-presentations' as exposures to the context or background cues. We attempt to deal with time *explicitly* by counting non-presentations; we deal with context cues as initially being candidate CSs competing as possible predictive cues, and over trials becoming learned to be a separate category of cues that are neither predictive nor uncorrelated. Gibbon (1977, 1984) presents a theory based in part on timing that also attempts to treat time as an independent entity.

5.5.2 The trial-window duration assumption

Even given the above complete, replicable specification of a particular set of trials, there is a problem that confounds both the theoretical formulation and experimental testing of contingency: the assumption of the duration of a particular trial. The calculation of conditional probabilities (and therefore the prediction of when a particular CS-US association should or should not be learned, and the predicted

strength of its learning) is dependent on the assumption that the experimenter (or theorizer) makes about the duration of a trial. This is not an idle issue: different assumptions can lead to drastically different conditional probability calculations. Figure 10 illustrates this: given a particular layout of cue presentations (in the figure, T indicates tone, L indicates light and S indicates shock), then the values of $p(US|CS)$ and $p(US|\overline{CS})$ are given under three different assumptions about the trial window duration: 2 minutes, 3 minutes and 4 minutes.

(FIGURE 10 GOES ABOUT HERE)

First of all, ignoring the tone CS and simply looking at the predicted associativity of the light and the shock, these three different assumptions render this set of trials as 0.5–0.2 when the trial size is assumed to be two minutes, 0.25–0.6 when it is assumed to be three minutes, and 0.75–0.33 when it is assumed to be four minutes. Under the first and last assumptions, the light CS is predicted to be strongly learned (since $p(US|CS) > p(US|\overline{CS})$), while under the 3-minute assumption, the opposite prediction is made: the light CS should be strongly learned to be a safety signal, indicating that the shock will not occur. This is strongly counterintuitive, and indeed rests on an example that was crafted explicitly to give rise to such a result, but nonetheless, by the strict rule of contingency, these are the correct predictions under these three different trial window assumptions.

Furthermore, the predicted *ordering* of the two cues (tone and light) will be reversed in this example: the tone and light CSs in Figure 10 will be about equally predictive of the shock under the two-minute assumption ((0.5–0.25 for tone, 0.5–0.2 for light); but the tone will be more predictive of the shock than the light is under the three-minute assumption (0.5–0.43 for tone, 0.25–0.6 for light); and finally, the tone will be much less predictive of the shock under the four-minute assumption (0.5–0.6 for tone, 0.75–0.33 for light). Were we to run an animal experiment using these trial data, our prediction of whether the tone or light or both would be learned to be associated with shock would depend directly on our assumption about the trial window duration.

It seems intuitively clear that animals do not make judgments about pairings on the basis of something so artificial and arbitrary as a 'time window'; rather, if a tone CS is followed closely by a shock US (within, say, 5 seconds), then a pairing is perceived by the animal, independent of whether or not a time-window boundary should ideally have fallen between the CS and US. This logically implies that the ideal contingency constraint, as it currently stands, is in need of revision or extension. Subjects must be choosing trial windows at least in part on the basis of the cues and events that are perceived; yet the very perception of the nature of those events seems to be dependent in part on the choice of trial win-

dows. One possible extension to the theory can be based on this apparent par the animal may first determine the salient cues in the environment, and may acquire information about their durations, and then that information may be used in part to incrementally calculate the associativity or predictiveness of various cues (via some algorithm). Indeed, Rescorla (1968) and Rescorla and Wagner (1972) have made the assumption that trial window duration was equal to CS duration, and have shown that this assumption leads to consistently successful experimental testing and successful predictive simulations of contingency. However, it has not been made clear in this literature how the animal may come to choose the CS duration as the perceived trial window duration. Assuming that CS duration is somehow used as approximate trial window duration by animals, then it is possible that it is the rapid conditioning of nonspecific response systems (e.g., heartrate, galvanic skin response, respiration) that is used to select candidate cues and to identify their durations, and then that these cue durations are used as candidate trial window durations as part of the process of determining associativity of cues. Experimentation with this theoretical line of thinking may clarify the relationship between rapid acquisition of nonspecific responses and slower learning of complex skeletal responses in associative learning.

6 Summary: Limitations and contributions of the theory

6.1 Status of our progress

We have attempted to provide in this paper an analysis of the effects of contingency in classical conditioning, and the implications of that analysis to predicted experimental outcomes, proposed algorithms and the evaluation of neurobiological circuits underlying conditioning. We are in the process of testing some of our theoretical predictions in our experimental laboratory (including the partial warning prediction and aspects of different trial-window duration assumptions). We intend the empirical results of these experimental studies to provide support or falsification for specific aspects of the theory. We are continuing a research program of extending our results to a broader range of phenomena of learning and memory, though we feel that classical conditioning clearly represents a reasonable paradigm for testing the limits of the way in which animals learn observed associations in their natural environments. It is probably the case that, in this regard, instrumental conditioning represents a still more 'natural' set of experimental procedures; our investigation has led us toward an integrative view of classical and instrumental conditioning (Granger, Schlimmer and Young 1985) which we intend to pursue. Similarly, there are a number of well-known associative and non-associative effects, especially extinction phenomena,

sensitization and habituation, and their relation to conditioning.

6.2 Interdependence of the three levels

The key question we have addressed here is: how can we evaluate proposed theories, algorithms, circuits or models of learning and memory in a principled way? The answer offered is that constraints on learning arise from both the computational level (where the precise defining features of the behavior are established) and the implementation level (where the biophysical mechanisms that underlie the behavior are identified). Since these two levels rarely meet each other, most theories are mediated through the algorithm level. Mechanistic descriptions of circuit operations are the bottom-up contribution, and derivations of behavioral-level constraints are the top-down contributions to a theory.

A computational-level analysis of the target behavior establishes the range of conditions that define (and thereby constrain) the behavior under study (such as classical conditioning). A complete theory must also be constrained by the physical attributes of the substrate system in which it is embedded; the neurobiological basis of classical conditioning is crucial. In principle, if we had a perfect implementation-level characterization of classical conditioning in, say, a circuit, then we would be able to determine the computational constraint (bottom-up) from the operation of that circuit. In the absence of such information (at least in the case of classical conditioning), the computational constraint was derived instead from animal experiments (Rescorla 1968); this of course still constitutes a 'bottom-up' derivation, as all such derivations must initially be. Once the computational constraint is in place, however, then the target behavior is defined, and all proposed theories, algorithms or circuits must conform to the constraint.

We cannot be sure that any given computational constraint is perfect or 'finished', either; for instance, if it turned out that a positive CS-US association was not learned in a partial warning presentation condition, then that would imply that the Rescorla (1968) constraint would require refinement. More complex counterintuitive predictions of the computation (such as the dependence on assumptions about trial window duration, section 5.5.2) also may give rise to experimentally-testable questions about the validity, extent and accuracy of the theory. Furthermore, the constraint only refers to effects of contingency in classical conditioning, yet the overall learning and memory capabilities of mammals certainly have more complex and far-reaching computational characterizations than just this constraint; the contingency constraint can be viewed, then, as one element of a larger class of constraints.

Our aim has been to attempt to analyze and clarify the contingency constraint; to apply it to generate

useful predictions (such as learning in the partial warning condition); and to provide a uniform way of evaluating proposed algorithms, behavioral predictions, circuits and models. We hope that theoretical and experimental investigators will continue to work together towards testing and refinement of the contingency constraint. We further hope that the analysis of contingency presented here will be used as a tool for researchers to test their own theories and experiments, and even as a 'measuring stick' to keep us on track in our evaluation of what is and is not contingency in associative learning.

Acknowledgments

Our thanks to Donald H. Perkel for his help with our analysis of contingency and development of the saddle graph; to Mark A. Gluck and Nelson Donegan for their extremely helpful comments on earlier drafts of this paper; to Michal T. Young for his extensive collaboration with us, especially in the development of the LS–LN contingency algorithm; to Lynn Nadel, Jeff Willner and Lisa Kurz for their helpful discussions about the Rescorla–Wagner and competing algorithms, and to them and Frank Schottler for help with our experimental setup; to David Benjamin for his help in designing and implementing the computer software for our animal experiments; to Norman M. Weinberger, Gary S. Lynch and James L. McGaugh for many helpful discussions; to Stacey Murren Granger, Donna Stephens and Charles L. Post, who are running our experiment-in-progress testing the partial warning condition; and last, but far from least, thanks to Stacey and Joyce, for their tolerance and support.

References

1. Alkon, D.L. 1980. Membrane depolarization accumulates during acquisition of an associative behavioral change. *Science, 210,* 1375–1376.
2. Allan, L.G. and Jenkins, H.M. 1980. The judgment of contingency and the nature of the response alternatives. *Canad. J. Psychol./Rev. Canad. Psychol.* **34,** 1–11.
3. Anderson, J.A., Silverstein, J.W., Ritz, S.A. and Jones, R.S. 1977. Distinctive Features, Categorical Perception and Probability Learning: Some Applications of a Neural Model. *Psychological Review, 84,* 413–451.
4. Anderson, J.R. 1983. *The Architecture of Cognition.* Cambridge: Harvard University Press.
5. Bayes, 1763. An Essay Towards Solving a Problem in the Doctrine of Chances by the late Rev. Mr. Bayes. *Phil. Trans. of Royal Soc.*
6. Brimer, C.J. and Dockrill, F.J. (1966). Partial reinforcement and the CER. *Psychonometric Science,* 5, 185–186.
7. Chang, J.J. and Gelperin, A. 1980. Rapid taste aversion learning by an isolated molluscan central nervous system. *Proceedings of the National Academy of Sciences, 77,* 6204.
8. Church, R.M. 1969. Response Suppression. In *Punishment and Aversive Behavior,* B. A. Campbell and R. M. Church (eds), Conference on Punishment, Princeton, NJ, 1967. New York: Appleton-Century-Crofts.
9. Dickinson, A. 1980. *Contemporary Animal Learning Theory.* Cambridge, London: Cambridge University Press.
10. Fitzgerald, R.D. (1963). Effects of partial reinforcement with acid on the classically conditioned salivary response in dogs. *Journal of Comparative and Physiological Psychology,* **56,** 1056–1060.

11. Gamsu, E. and Williams, D.R. (1971). Classical conditioning of a complex skeletal response. *Science*, **171**, 923–925.
12. Gibbon, J. 1977. Scalar Expectancy Theory and Weber's Law in Animal Timing. *Psychological Review*, *84*, 279–325.
13. Gibbon, J., Church, R.M. and Meck, W.H. 1984. Scalar Timing in Memory. In: *Timing and Time Perception*, J. Gibbon and L. Allan (Eds.), New York: The New York Academy of Sciences.
14. Gibbon, J., Berryman, R. and Thompson, R.L. 1974. Contingency spaces and measures in classical and instrumental conditioning. *Journal of the Experimental Analysis of Behavior*, *21* 585–605.
15. Gluck, M.A. and Thompson, R.F. 1985. A computer model of the neural substrates of classical conditioning in the *Aplysia*. In *Proceedings of the Seventh Annual Conference of the Cognitive Science Society*, 36–42.
16. Gluck, M.A. and Thompson, R.F. 1986. Modeling the neural substrates of associative learning and memory: A computational approach. *Psychological Review*, in press.
17. Granger, R.H. and Schlimmer, J.C. 1985. Learning salience among features through contingency in the CEL framework. In *Proceedings of the Seventh Annual Conference of the Cognitive Science Society*, 65–79.
18. Granger, R.H., Schlimmer, J.C. and Young, M.T. 1985. Contingency and Latency in Associative Learning: Computational, Algorithmic and Implementation Analyses. Department of Computer Science Technical Report 85-10, University of California, Irvine. In *Brain Structures, Learning and Memory*, Eds. Davis, J., Wegman, E. and Newburg, R. (to appear).
19. Grossberg, S. 1982. Processing of expected and unexpected events during conditioning and attention: A psychophysiological theory. *Psychological Review, 89*, 529–572.
20. Hammond, L.J. (1967). A traditional demonstration of the active properties of Pavlovian inhibition using differential CER. *Psychonometric Science*, **9**, 65–66.
21. Hampson, S. and Kibler, D. 1983. A Boolean Complete Neural Model of Adaptive Behavior. *Biological Cybernetics, 49*, 9–19.
22. Hawkins, R.D., Carew, T.J. and Kandel E.R. 1983. Effects of interstimulus interval and contingency on classical conditioning in *Aplysia*. *Society for Neuroscience Abstracts*, 9, 168.
23. Hawkins, R.D. and Kandel, E.R. 1984. Is there a cell-biological alphabet for simple forms of learning? *Psychological Review, 91*, 376–391.
24. Hearst, E. and Franklin, S.R. (1977). Positive and negative relations between a signal and food: approach-withdrawal behavior to the signal. *Journal of Experimental Psychology: Animal Behavioral Processes*, **3**, 37–52.
25. Kamin, L. J. 1968. Predictability, surprise, attention, and conditioning. In *Miami Symposium on the Prediction of Behavior, Aversive Stimulation*, M. R. Jones (ed), Coral Gables, Florida: University of Miami Press.
26. Langley, P.W., Zytkow, J.M., Simon, H.A., and Bradshaw, G.L. 1983. Mechanisms for qualitative and quantitative discovery. *Proceedings of the International Machine Learning Workshop*, University of Illinois, Urbana–Champaign, 121–132.
27. Mackintosh, N.J. 1974. *The Psychology of Animal Learning*. New York: Academic Press.
28. Mackintosh, N.J. 1975. A theory of attention: Variations in the associability of stimulus with reinforcement. *Psychological Review, 82*, 276–298.
29. Mackintosh, N.J. 1983. *Conditioning and Associative Learning*. New York: Oxford University Press.
30. Marr, D. 1982. *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W. H. Freeman.
31. Pearce, J. M. and Hall, G. 1980. A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review* **87**, 532–52.
32. Pearl, J. 1982. Reverend Bayes on Inference Engines: A Distributed Hierarchical Approach. *Proceedings of the National Conference on Artificial Intelligence*, 133–136.
33. Rescorla, R. 1966. Predictability and number of pairings in Pavlovian fear conditioning. *Psychonomic Science* **4**, 383–384.

34. Rescorla, R. 1967. Pavlovian conditioning and its proper control procedures. *Psychological Review* **74**, 71–80.

35. Rescorla, R. 1968. Probability of shock in the presence and absence of CS in fear conditioning. *J. Comparative and Physiological Psychology* **66**, 1–5.

36. Rescorla, R. 1972. Informational variables in Pavlovian conditioning. *The Psychology of Learning and Motivation* **6**, 1–46.

37. Rescorla, R. and Wagner, A. R. 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In *Classical Conditioning II: Current Research and Theory*, A. H. Black and W. F. Prokasy (eds). New York: Appleton-Century-Crofts, 1972.

38. Rosenblatt, F. (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms.* Washington, D.C: Spartan Books.

39. Shanks, D.R. 1985. Continuous monitoring of human contingency judgment across trials. *Memory and Cognition*, **13**, 158–167.

40. Siegel, S. and Domjan, M. (1971). Backward conditioning as an inhibitory procedure. *Learning and Motivation*, **2**, 1–11.

41. Skyrms, B. (1966). *Choice and Chance: An Introduction to Inductive Logic.* Belmont, California: Dickenson Publishing Company, Inc.

42. Spence, K.W. (1936). The nature of discrimination learning in animals. *Psychological Review, 43*, 427–449.

43. Sutton, R.S. and Barto, A.G. 1981. Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review, 88*, 135–170.

44. Thomas, E. and Wagner, A.R. (1964). Partial reinforcement of the classically conditioned eyelid response in the rabbit. *Journal of Comparative and Physiological Psychology*, **58**, 157–158.

45. Thompson, R.F., Clark, G.A., Donegan, N.H., Lavond, D.G., Madden, J., Mamounas, L.A., Mauk, M.D. and McCormick, D.A. (1984). Neuronal substrates of basic associative learning. In L. Squire and N. Butters (Eds.), *Neuropsychology of Memory.* New York: Guilford Press.

46. Wagner, A.R. and Rescorla, R.A. (1972). Inhibition in Pavlovian conditioning: application of a theory. In R.A. Boakes and M.S. Halliday (Eds.). *Inhibition and Learning.* London: Academic Press.

47. Wagner, A.R., Siegel, S., Thomas, E., and Ellison, G.D. (1964). Reinforcement history and the extinction of a conditioned salivary response. *Journal of Comparative and Physiological Psychology*, **58**, 354–358.

48. Wagner, A.R. (1981). SOP: a model of automatic memory processing in animal behavior. In *Information processing in animals: memory mechanisms*, N.E. Spear and R.R. Miller (eds), pp. 5–47, Erlbaum, Hillsdale, NJ.

49. Wahlsten, D. L. and Cole, M. 1972. Classical and avoidance training of leg flexion in the dog. In *Classical Conditioning II: Current Research and Theory*, A.H. Black and W.F. Prokasy (eds), New York: Appleton-Century-Crofts.

50. Wasserman, E.A., Chatlosh, D.L. and Neunaber, D.J. 1983. Perception of causal relations in humans: Factors affecting judgments of response-outcome contingencies under free-operant procedures. *Learning and Motivation*, **14**, 406–432.

Appendix A: Derivation of contingency surface

If it is assumed that trials are discrete, independent, and randomized then we may consider each of the four possible stimulus combinations:

X	=	$P(CS, \overline{US})$	CS alone
Y	=	$P(\overline{CS}, US)$	US alone
Z	=	$P(CS, US)$	CS followed by US
$1-X-Y-Z$	=	$P(\overline{CS}, \overline{US})$	Empty trial

Consider a Cartesian coordinate system in a Euclidian three-dimensional space. All possible stimulus combination points (defined above) lie within a right triangular prism within the unit cube bounded by the X–Y, X–Z and Y–Z planes and by a truncating slanted plane passing through the points $(X, Y, Z) = (1,0,0)$, $(0,1,0)$, and $(0,0,1)$, since $X + Y + Z \leq 1$.

The contingency characterization states that conditioning does not occur when $P(US|CS) = P(US|\overline{CS})$, and that this equality therefore defines the boundary between learning of positive and negative associations. By the definition of conditional probabilities we have:

$$P(US|CS) = P(US, CS)/P(CS) \qquad P(US|\overline{CS}) = P(US, \overline{CS})/P(\overline{CS})$$

The marginal probabilities are directly derived as:

$$\begin{aligned} P(CS) &= X + Z \\ P(\overline{CS}) &= 1 - X - Z \\ P(US) &= Y + Z \\ P(\overline{US}) &= 1 - Y - Z \end{aligned}$$

Substituting, we have

$$P(US|CS) = Z/(X + Z) \qquad P(US|\overline{CS}) = Y/(1 - X - Z)$$

Substituting these expressions in the contingency boundary equation, we have:

$$Y = \frac{Z(1 - X - Z)}{(X + Z)}$$

which describes a hyperbolic paraboloid. It is illustrated within the truncated unit cube in Figure 1.

Appendix B: Comparative analysis of performance of contingency algorithms

Rescorla and Wagner

The Rescorla and Wagner (1972; Wagner and Rescorla, 1972) model was simulated under a pair of conditions. In the first set of simulations $P(CS,US) = 0.10$, that is, one of every ten trials was a reinforced presentation of the CS. The parameters were chosen following those in the original presentation (Rescorla and Wagner, 1972, page 82). Specifically, $\alpha_{CS} = \alpha_{context} = 1.0$, $\lambda_{reinforced} = 1$, $\lambda_{nonreinforced} = 0$, $\beta_{reinforced} = \beta_{nonreinforced} = 0.15$. The last parameter is larger than one originally used and was chosen to allow asymptotic learning in 25 trials ($= 0.10 \times 250$). The asymptotic associative strengths for the CS and context in the presence of various amounts of spurious cues are presented below.

$P(CS,US) = 0.10$, 250 Trials Total
(Figures are single samples from an arbitrary, uniform ordering.)

%	Type	$P(US\|CS)$	$P(US\|\overline{CS})$	V_{CS}	$V_{context}$
0	—	1.00	0.00	0.98	0.00
25	CS	0.25	0.00	0.29	0.01
25	US	1.00	0.28	0.71	0.26
25	CS,US	0.44	0.16	0.34	0.20
50	CS	0.17	0.00	0.16	0.01
50	US	1.00	0.56	0.43	0.61
50	CS,US	0.29	0.38	-0.10	0.46
75	CS	0.12	0.00	0.14	0.02
75	US	1.00	0.83	0.18	0.74
75	CS,US	0.21	0.71	-0.53	0.67

A second set of simulations were performed, this time with the $P(CS,US) = 0.20$ and the exact set of parameters used in Rescorla and Wagner (1972, p. 88), $\alpha_{CS} = 0.5$, $\alpha_{context} = 0.1$, $\lambda_{reinforced} = 1$, $\lambda_{nonreinforced} = 0$, $\beta reinforced = 0.1$, $\beta nonreinforced = 0.05$. The asymptotic associative strengths for the CS and context are presented below.

$P(CS,US) = 0.20$, 250 Trials Total
(Figures are a single sample from an arbitrary, uniform ordering.)

%	Type	$P(US\|CS)$	$P(US\|\overline{CS})$	V_{CS}	$V_{context}$
0	—	1.00	0.00	0.84	0.09
20	CS	0.50	0.00	0.66	0.34
20	US	1.00	0.25	0.66	0.34
20	CS,US	0.67	0.14	0.59	0.21
40	CS	0.33	0.00	0.44	0.06
40	US	1.00	0.50	0.50	0.54
40	CS,US	0.50	0.33	0.37	0.35
60	CS	0.25	0.00	0.34	0.05
60	US	1.00	0.75	0.37	0.72
60	CS,US	0.40	0.60	0.14	0.47
80	CS	0.20	$\perp$	0.27	0.05
80	US	1.00	1.00	0.25	0.86
80	CS,US	0.20	0.80	-0.05	0.61

$\perp \equiv$ undefined.

Granger and Schlimmer

The Granger and Schlimmer model has been similarly tested for the cases where $P(CS,US) = 0.10$ and $P(CS,US) = 0.20$. The LS and LN measures computed for each potential cue stimulus are interpreted first as odds, then then converted to a probability ($p = odds/(1 + odds)$), and then are mapped onto the range $[-1, 1]$ ($V = (p - 0.5)/0.5$) for the purposes of straightforward comparison with the other models presented. The results for varying degrees of spurious CSs, spurious USs, and spurious CSs and USs are presented below.

$P(CS,US) = 0.10$, 250 Trials Total
(Each data point represents an average over 10 orderings.)

%	Type	$P(US\|CS)$	$P(US\|\overline{CS})$	V_{CS}	$V_{context}$
0	—	1.00	0.00	0.99	-0.28
25	CS	0.25	0.00	0.96	-0.28
25	US	1.00	0.28	0.90	-0.13
25	CS,US	0.44	0.16	0.47	-0.22
50	CS	0.17	0.00	0.89	-0.28
50	US	1.00	0.56	0.85	0.11
50	CS,US	0.29	0.38	-0.08	-0.13
75	CS	0.12	0.00	0.65	-0.28
75	US	1.00	0.83	0.64	0.54
75	CS,US	0.21	0.71	-0.47	-0.02

$P(CS,US) = 0.20$, 250 Trials Total
(Each data point represents an average over 5 orderings.)

%	Type	$P(US\|CS)$	$P(US\|\overline{CS})$	V_{CS}	$V_{context}$
0	—	1.00	0.00	0.99	-0.23
20	CS	0.50	0.00	0.97	-0.23
20	US	1.00	0.25	0.94	-0.11
20	CS,US	0.67	0.14	0.64	-0.17
40	CS	0.33	0.00	0.94	-0.23
40	US	1.00	0.50	0.92	0.11
40	CS,US	0.50	0.33	0.19	-0.09
60	CS	0.25	0.00	0.86	-0.23
60	US	1.00	0.75	0.85	0.42
60	CS,US	0.40	0.60	-0.20	0.00
80	CS	0.20	$\perp$	-0.23	-0.23
80	US	1.00	1.00	-0.01	0.98
80	CS,US	0.20	0.80	-0.49	0.11

Strengthening and Weakening

In contrast to those algorithms which compute conditional probability correctly, we simulated an algorithm from a class which computes

$$\Delta V = \alpha(\gamma Z + \delta X + \sigma Y + \rho(1 - X - Y - Z))$$

Specifically, we chose $\alpha = 0.15$, $\gamma = 0.90$, $\delta = 0.10$, $\sigma = 0.40$, and $\rho = 0.00$. The results for $P(CS,US) = 0.10$ are presented below:

$P(CS,US) = 0.10$, 250 Trials Total
(Figures are a single sample from an arbitrary, uniform ordering.)

%	Type	$P(US\|CS)$	$P(US\|\overline{CS})$	V_{CS}	$V_{context}$
0	—	1.00	0.00	1.00	0.00
25	CS	0.25	0.00	0.99	0.00
25	US	1.00	0.28	-0.35	0.97
25	CS,US	0.44	0.16	0.99	0.96
50	CS	0.17	0.00	0.97	0.00
50	US	1.00	0.56	-0.99	1.00
50	CS,US	0.29	0.38	-0.94	1.00
75	CS	0.12	0.00	0.63	0.00
75	US	1.00	0.83	-0.99	1.00
75	CS,US	0.21	0.71	-0.94	1.00

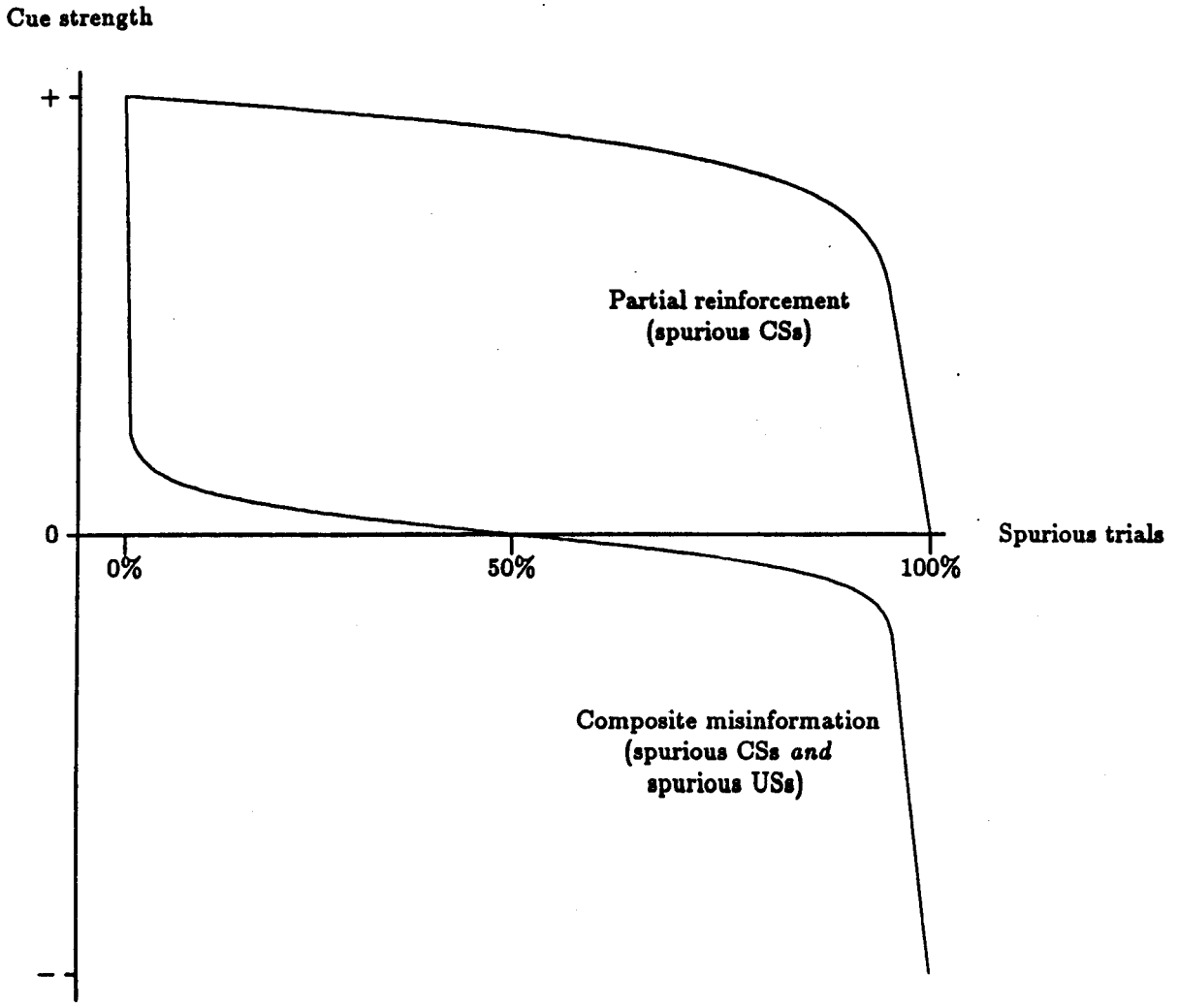


Figure 1: Degradation of learned predictiveness in partial vs. composite conditions.

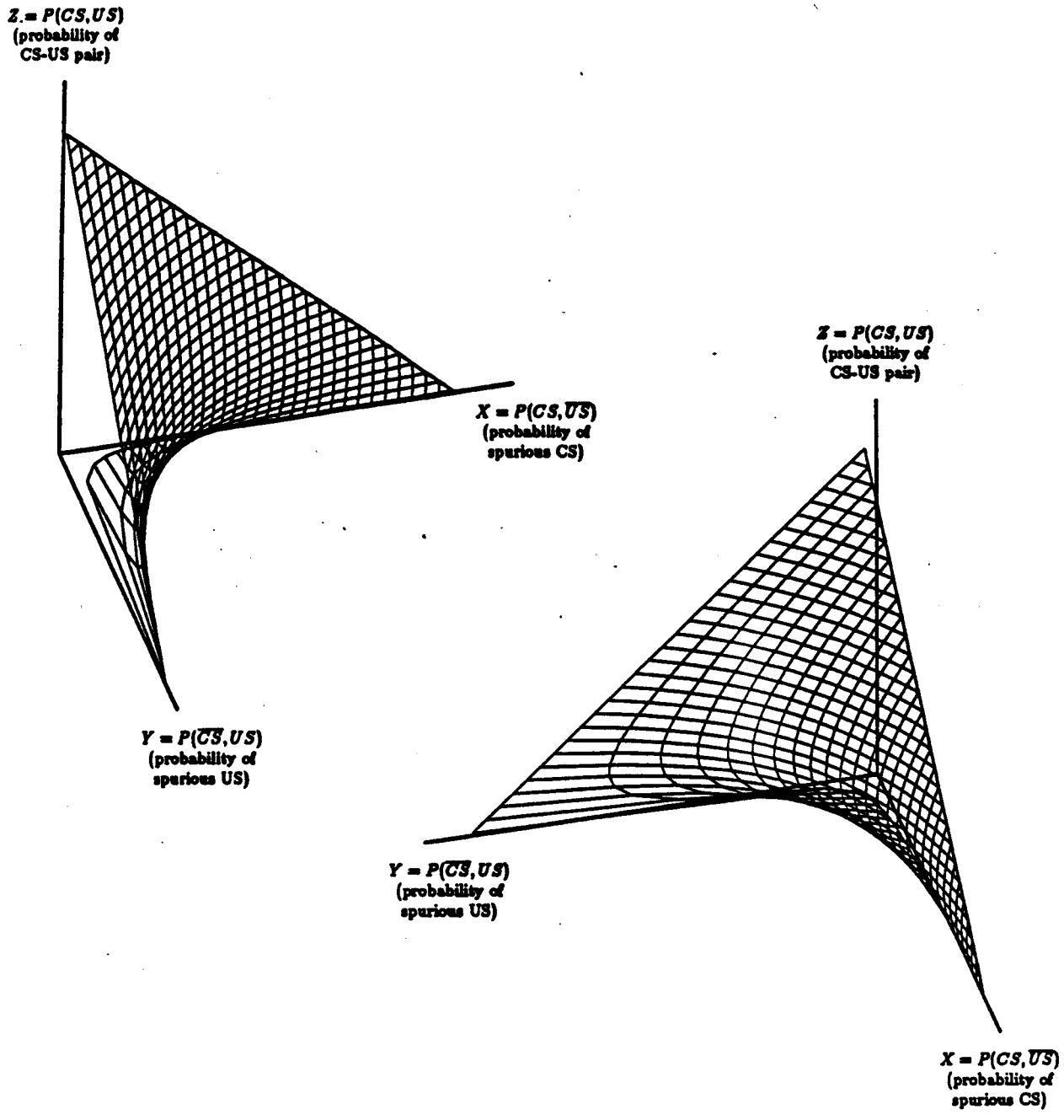


Figure 2: The computational constraint of contingency (330° and 240° rotations)

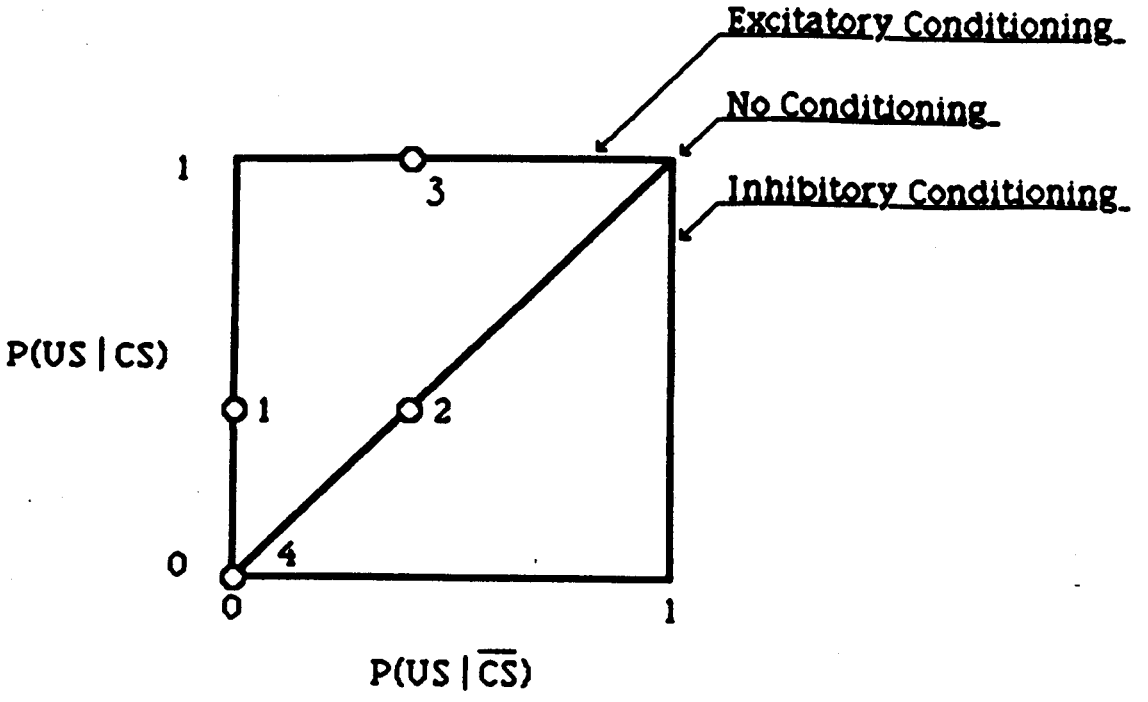


Figure 3: Church–Gibbon contingency plane

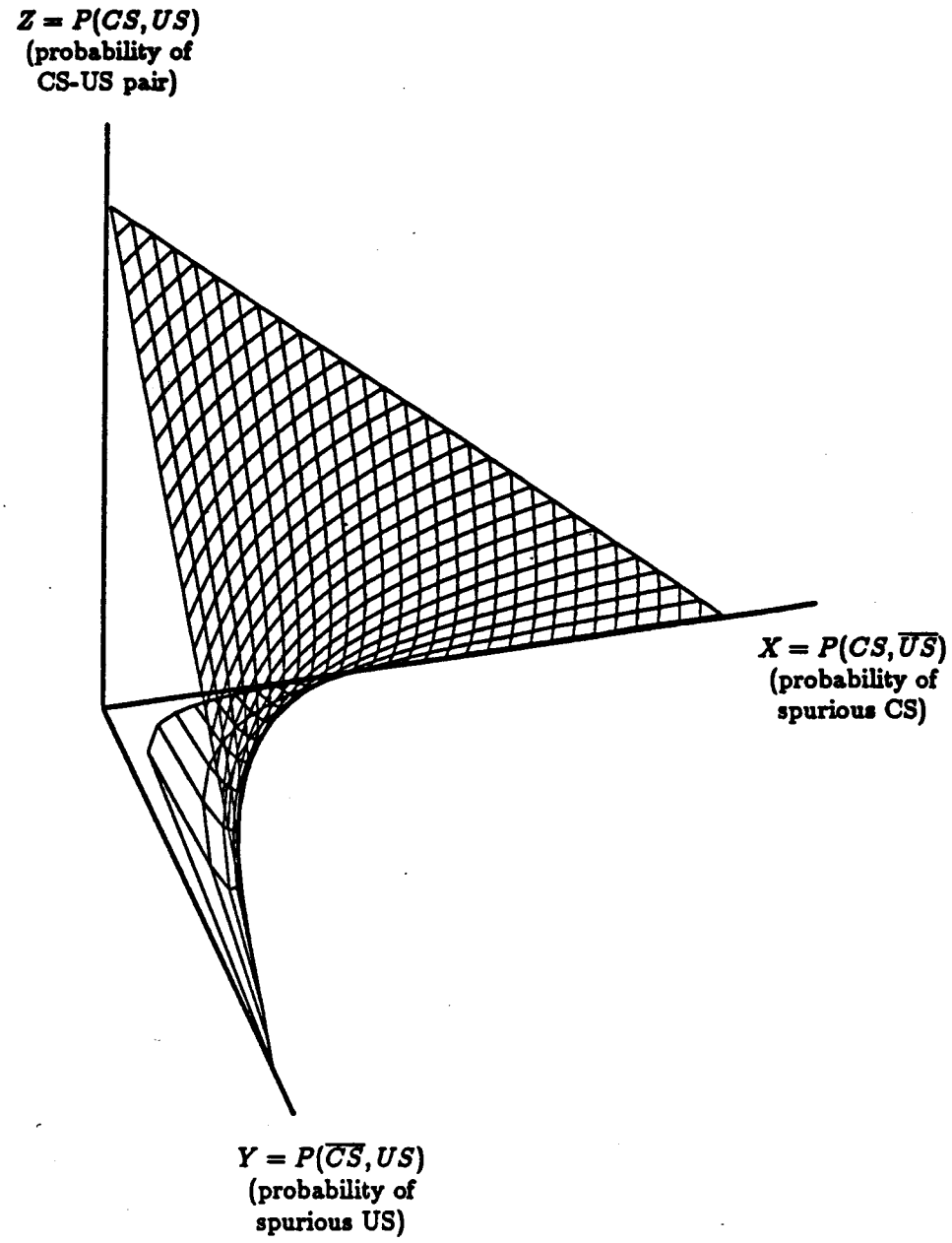
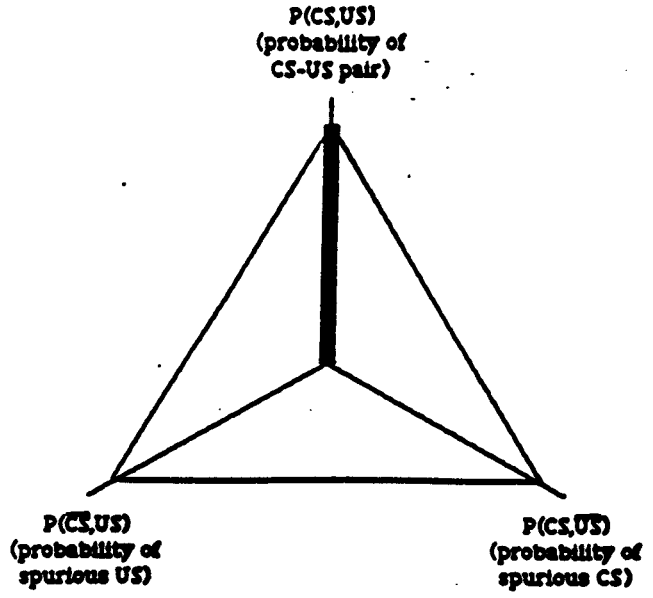
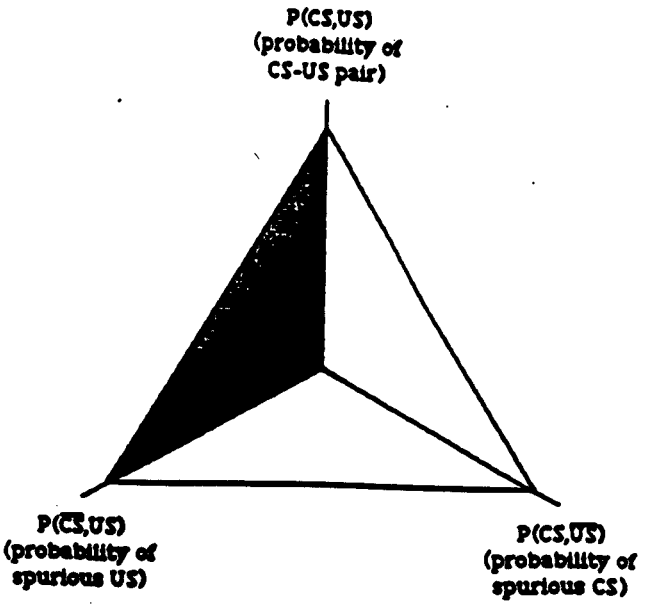


Figure 4: The contingency constraint



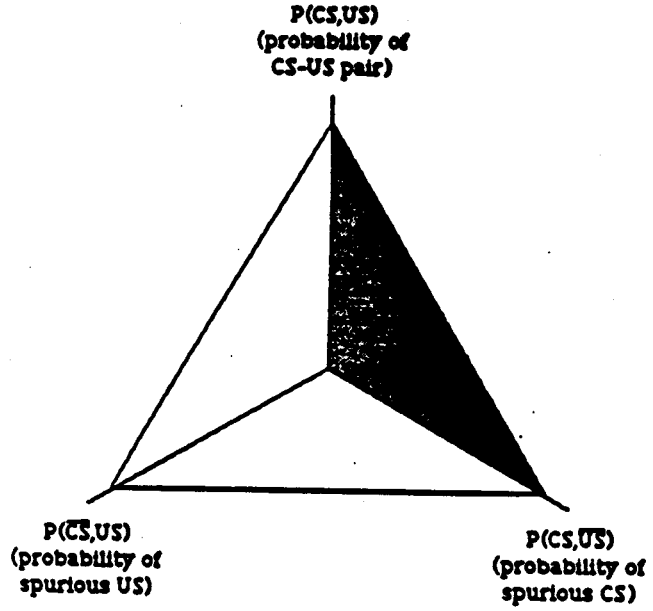
(a)

Perfect Pairings Contingency



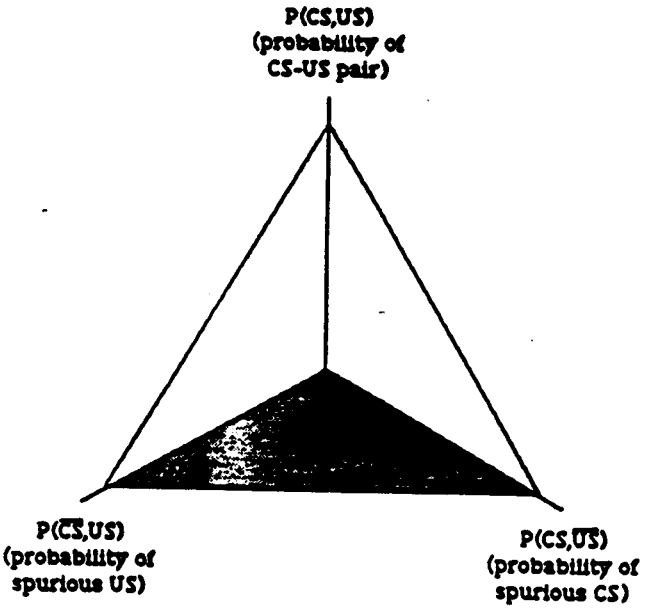
(b)

Partial Warning Contingency



(c)

Partial Reinforcement Contingency



(d)

Completely Unpaired Contingency

Figure 5. Regions of the contingency space

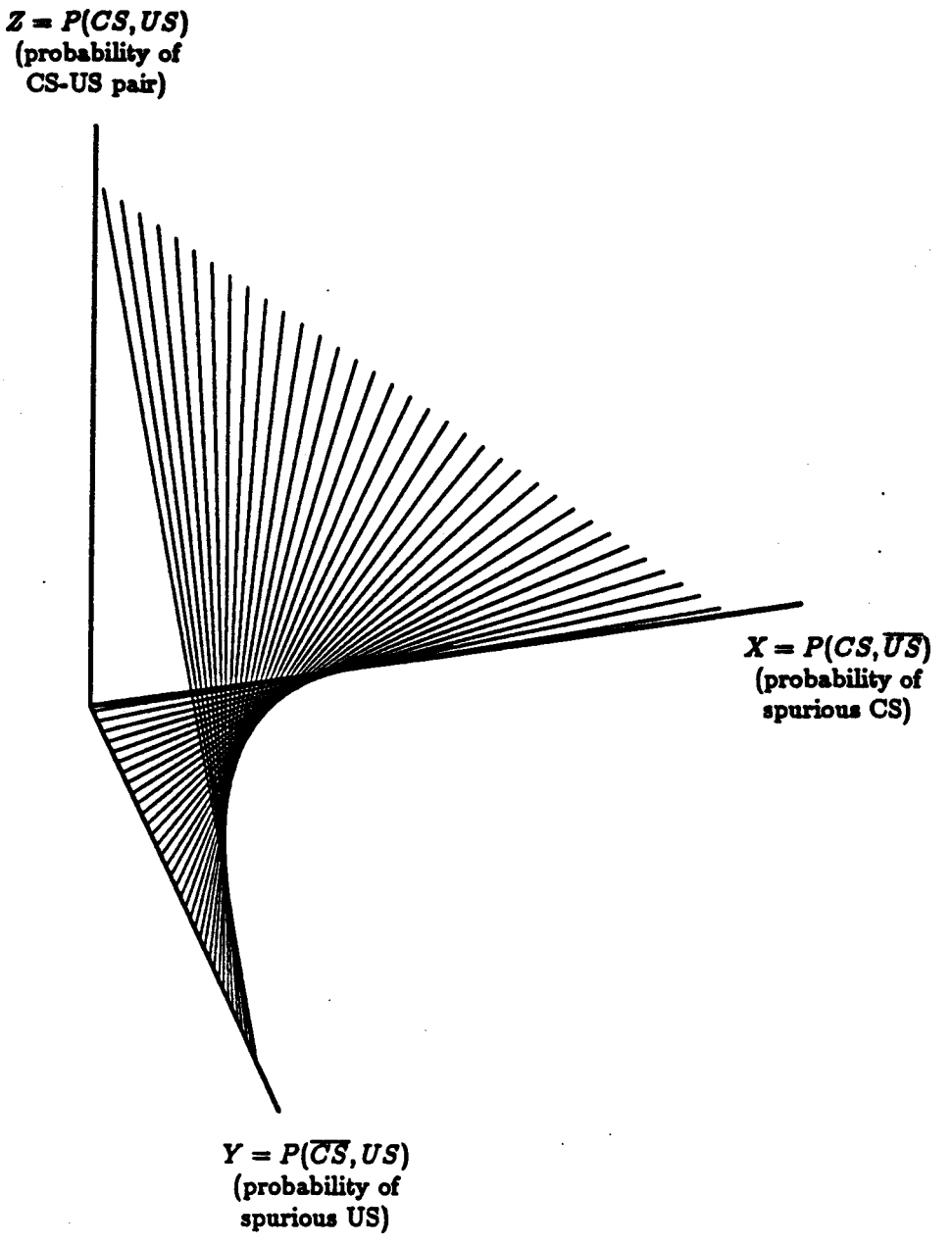


Figure 6: Presentation lines comprising the contingency saddle surface

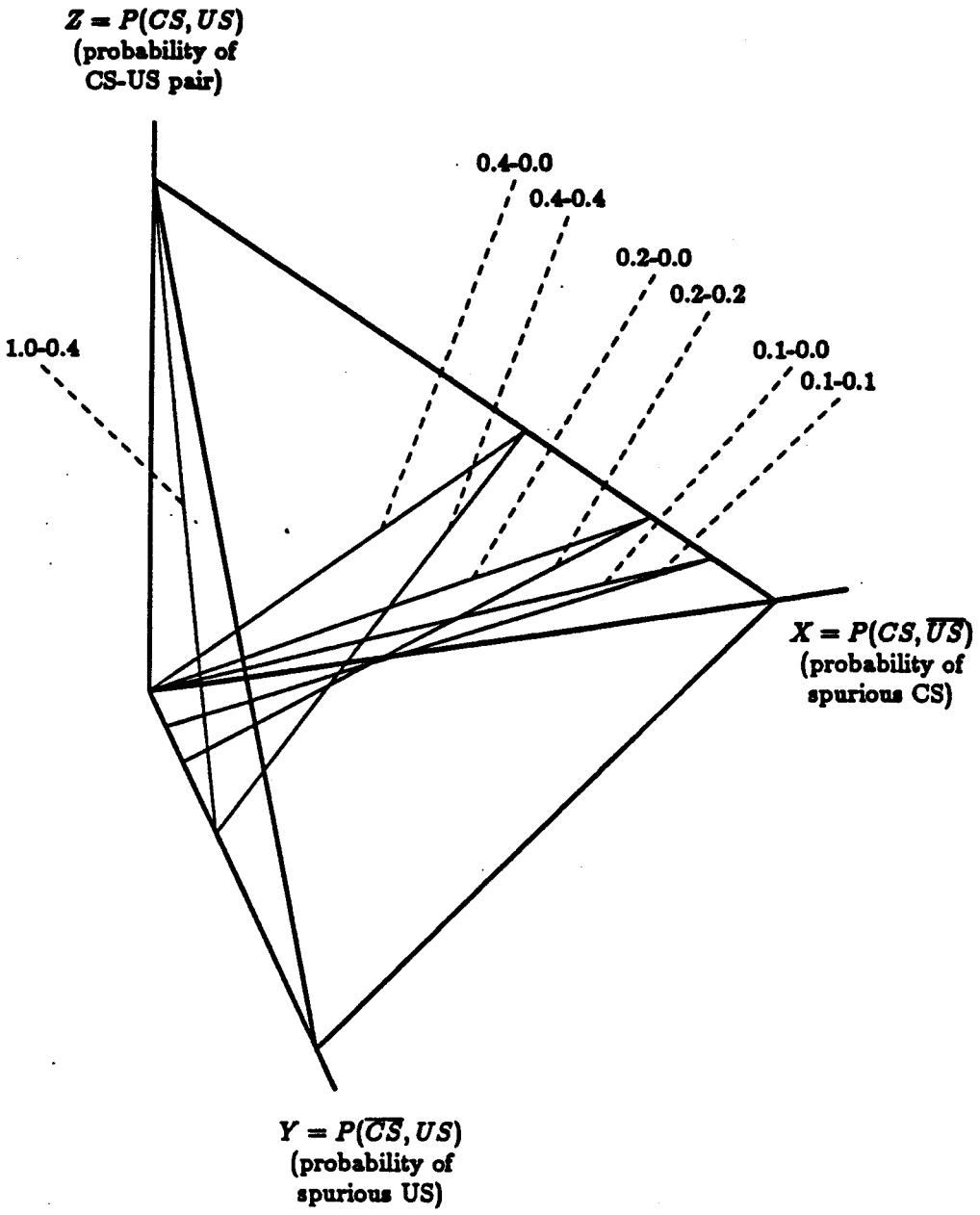


Figure 7: Specific presentation condition lines in contingency space

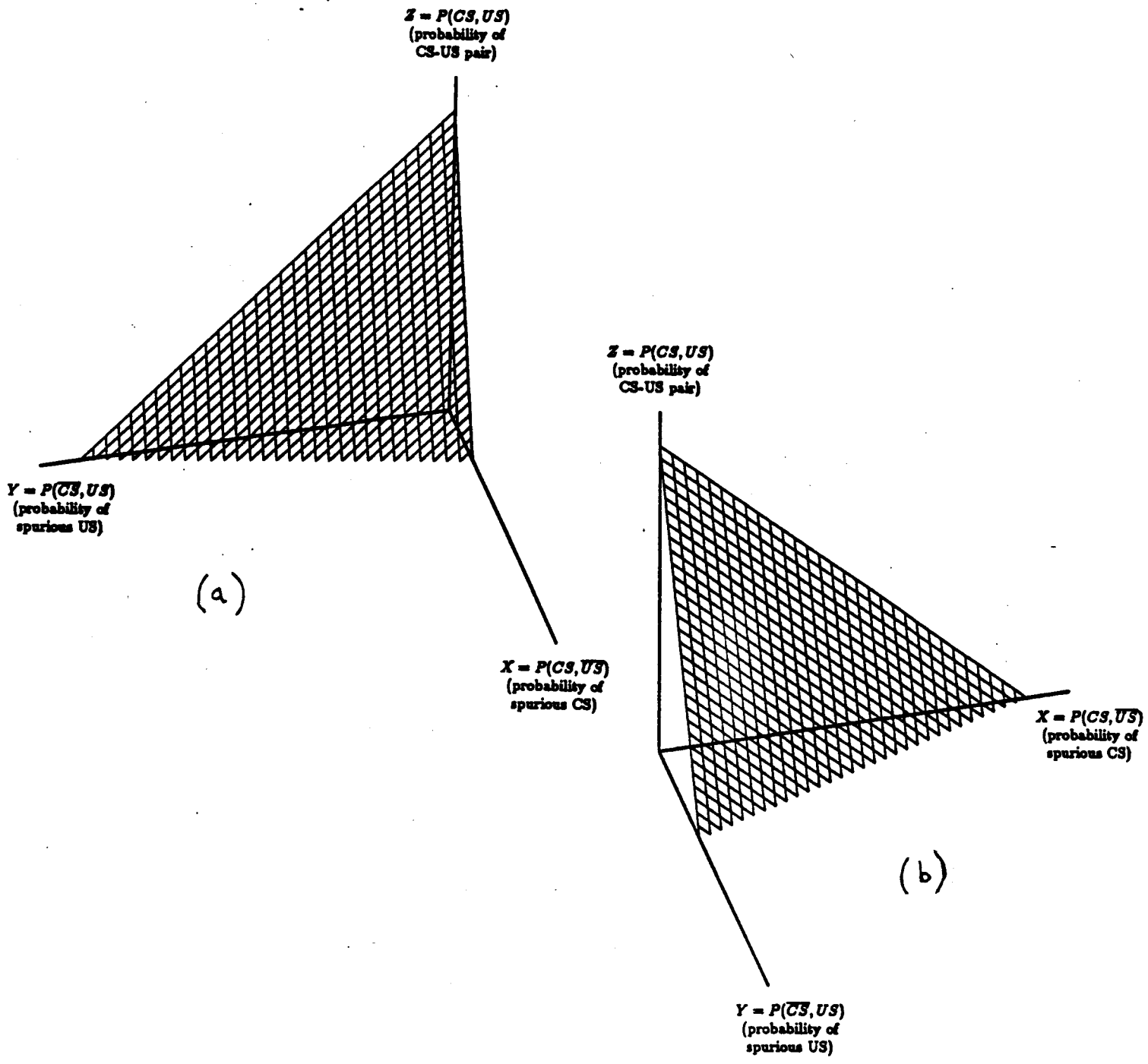


Figure 8: S/W algorithms approximating (a) partial warning; (b) partial reinforcement

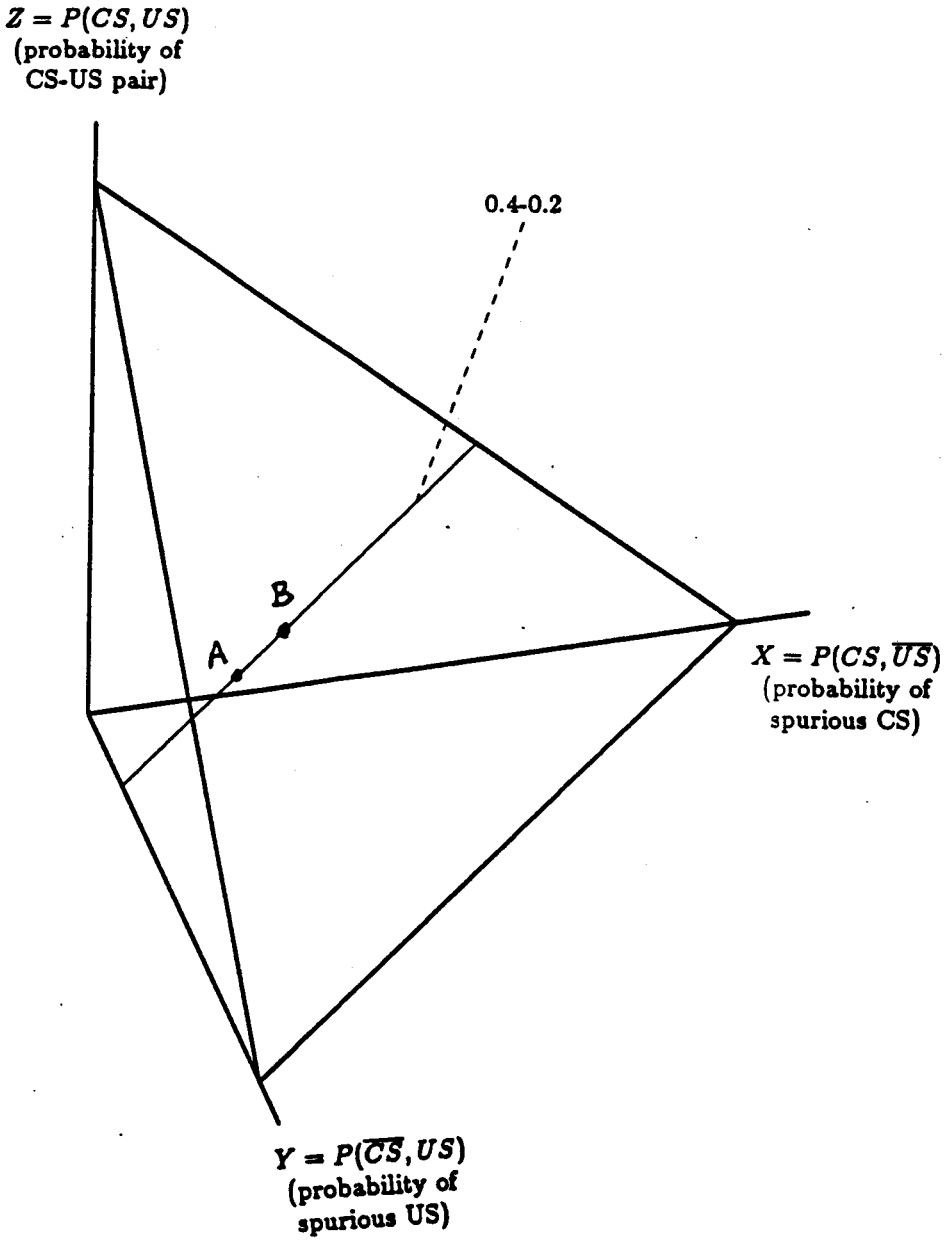
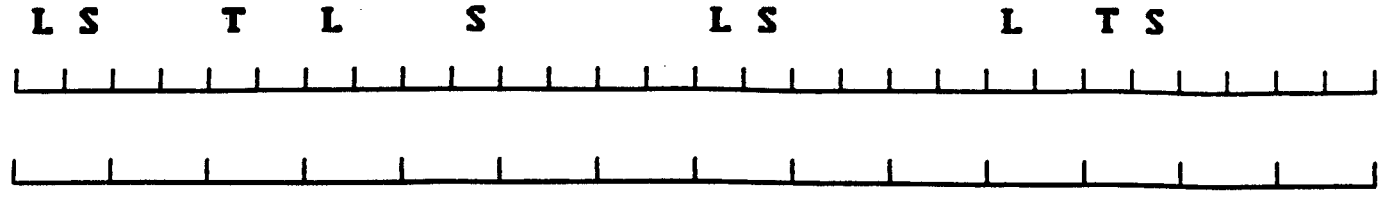


Figure 9: The 0.4–0.2 presentation condition



For Tone: 1 Pairing, 1 Spurious CS, 3 Spurious USs, and 9 Empty trials
$P(US \mid T) = 1/(1+1) = 0.5$; $P(US \mid \bar{T}) = 3/(3+9) = 0.25$
For Light: 2 Pairings, 2 Spurious CSs, 2 Spurious USs, and 2 Empty trials
$P(US \mid L) = 2/(2+2) = 0.5$; $P(US \mid \bar{L}) = 2/(2+8) = 0.2$

For Tone: 1 Pairing, 1 Spurious CS, 3 Spurious USs, and 4 Empty trials
$P(US \mid T) = 1/(1+1) = 0.5$; $P(US \mid \bar{T}) = 3/(3+4) = 0.43$
For Light: 1 Pairing, 3 Spurious CSs, 3 Spurious USs, and 2 Empty trials
$P(US \mid L) = 1/(1+3) = 0.25$; $P(US \mid \bar{L}) = 3/(3+2) = 0.6$

For Tone: 1 Pairing, 1 Spurious CS, 3 Spurious USs, and 2 Empty trials
$P(US \mid T) = 1/(1+1) = 0.5$; $P(US \mid \bar{T}) = 3/(3+2) = 0.6$
For Light: 3 Pairings, 1 Spurious CS, 1 Spurious US, and 2 Empty trials
$P(US \mid L) = 3/(3+1) = 0.75$; $P(US \mid \bar{L}) = 1/(1+2) = 0.33$

Figure 10: Three different trial-window duration assumptions yield different contingencies for the same set of trials