8

Perception and Evolution

BRUCE M. BENNETT,
Department of Mathematics, University of California, USA
DONALD D. HOFFMAN,
Department of Cognitie Science, University of California, USA
CHETAN PRAKASH
Department of Mathematics, California State University, USA

INTRODUCTION

The freshwater eel is notable not only for great sushi, but also for chromophore substitution (Beatty, 1984; Lythgoe, 1991). As young adults these eels spend most of their time in fresh water and, like many other freshwater creatures, have a long-wavelength visual pigment known as porphyropsin, with a maximal sensitivity at 522 nanometers. As the time draws near for the first leg of their breeding migration, which takes them into coastal waters, the porphyropsin changes to a rhodopsin with maximal sensitivity at 500 nanometers. And as the time draws near for the last leg of their breeding migration, which takes the eels into the deep sea, the rhodopsin changes its maximal sensitivity to 487 nanometers, a value typical for many deep-sea creatures. This is but one of many engaging examples of the adaptation, both phylogenetic and ontogenetic, of perceptual systems to environments.

The variety of adaptations in vision alone is remarkable. The optical systems used include: pigmented pits without lenses or mirrors, found in some platyhelminthes, protochordates, coelenterates, annelids, and molluscs (Salvini-Plawen & Maxx, 1977); multiple pigmented tubes, found in some tube-worms; spherical lens eyes, found in fishes, some molluscs, alciopid annelids, and copepod crustaceans (Pumphrey, 1961); corneal refraction, found in many terrestrial vertebrates and arachnids (Land, 1985); apposition compound eyes, found in many diurnal insects and crustacea; refracting superposition compound eyes, found in nocturnal insects and some crustacea (Exner, 1891; Kunze, 1979; Nilsson, 1989); simple mirror eyes,

Perception and the Physical World: Psychological and Philosophical Issue in Perception.
Edited by Dieter Heyer and Rainer Mausfeld.

found in the *Pecten* bivalve mollusc and the *Gigantocypris* crustacean; and reflecting superposition compound eyes, found in shrimps, prawns, crayfish, and lobsters (for an excellent review, see Land, 1991). Each solution works in its niche. Even something as simple as a pinhole eye has served the cephalopod *Nautilus* well for 400 million years (Lythgoe, 1991, p. 4).

Then there is the variety of visual pigments and colored oil drops used in vision. Humans are trichromats, as are some species in almost every animal class (Jacobs, 1981, p. 153). Squirrels, rabbits, tree shrews, some fishes and male New World monkeys are dichromats. Goldfish and turtles are tetrachromats (Crawford et al., 1990; Neumeyer, 1985, 1986). Pigeons may be pentachromats (Delius & Emmerton, 1979; Emmerton & Delius, 1980; Wright, 1979). The mantis shrimp has at least ten spectral types of photoreceptors (Cronin et al., 1994; Cronin & Marshall, 1989). Some visual pigments respond to ultraviolet (Neumeyer, 1985) and some to near infrared (Lythgoe, 1988).

The variety explodes when it comes to the neural processing and interpretation of visual images. On one extreme there is almost no processing, as in creatures with simple eye spots. On the other extreme there is human vision, with tens of billions of neurons devoted to image interpretation.

So if we are interested in perceptual evolution, and in particular how the perceptual interpretations of organisms are adapted to their environments, we must admit at once that the topic encompasses an incredible diversity of phenomena. Indeed the diversity is such that one might be tempted to conclude that there is little useful to be said in general about perceptual adaptation, but much to be said about individual cases.

That may yet turn out to be true. But recent developments in the formal study of perception, and developments with a longer history in the formal study of evolution, offer hope for a formal theory of perceptual adaptation that captures the unity behind the many cases while respecting the remarkable diversity. In this chapter we take initial steps directed toward making this hope a reality.

Currently the most rigorous and comprehensive formal theories of perception are Bayesian (see, e.g., Knill & Richards, 1996), and the most successful theories of evolution are neo-Darwinian. We will begin by developing the Bayesian approach to perception, and then place this approach in a neo-Darwinian context. This leads to concrete mathematical problems, regarding the convergence of certain probabilistic processes, whose resolution will provide key insights into the scope and limits of perceptual evolution.

BAYESIAN PERCEPTION

In the case of vision, the simplest motivation for the Bayesian approach is as follows. We are given an image or sequence of images, I, and we wish to reach conclusions about the visual scenes, S, responsible for I.

In general, countless different scenes could, in principle, be responsible for I. This ambiguity arises because the relationship between scenes and images is typically one of *projection,* e.g., from the three dimensions of the visual world to the two dimensions of retinal images. Since the mapping from scenes to images is many to one, the mapping from images to scenes is one to many.

So, given an image I, there are many scenes to be considered as possible interpretations since there are many scenes that could in principle be responsible for I. An observer would like to pick the "right" one, but might not have sufficient information to know with certainty which one to pick. If a guaranteed "right" choice is not possible, a probabilistic assessment over all possible choices is next best. Ideally, in this case, an observer would like to find a conditional probability,

$$\Pr(S \mid I) \tag{8.1}$$

which specifies the probabilities of various scenes given the image I.

As a simple example, suppose our image I_{line} has just one straight line segment in it. Consider the following two sets of scene interpretations. In the first, S_{line}, the interpretation is as some (any) straight line in 3-D which projects to the given line in the image. There are many such lines, at different orientations in space, all of which project to the same line in the image. Together these lines constitute the set S_{line}. In the second S_{curve}, the interpretation is as some (any) curve in 3-D which projects to the given line in the image. There are many such curves, including semicircles, various sinusoidal curves, and so on, all of which project to the same line in the image. Together these curves constitute the set S_{curve}. What we would like to compute are the conditional probabilities

$$\Pr(S_{\text{curve}} \mid I_{\text{line}}) \text{ and } \Pr(S_{\text{line}} \mid I_{\text{line}}) \tag{8.2}$$

Each set of interpretations, S_{line} and S_{curve} is uncountably large. So set size alone won't help us much in computing these conditional probabilities. What we need is the apparatus of Bayes' rule:

$$\Pr(S \mid I) = \frac{\Pr(I \mid S)\Pr(S)}{\Pr(I)} \tag{8.3}$$

It is conventional to call $\Pr(S \mid I)$ the posterior probability, $\Pr(I \mid S)$ the likelihood function, and $\Pr(S)$ and $\Pr(I)$ the prior probabilities. What we want to compute are the posterior probabilities $\Pr(S \mid I)$. They will give us the relative confidence we should place in the various possible interpretations. We can compute them, by Bayes' rule, if we know the likelihood function $\Pr(I \mid S)$ and the prior probabilities $\Pr(S)$ and $\Pr(I)$.

Do we know the likelihood function $\Pr(I \mid S)$? We do, if we know how scenes get rendered as images. In fact $\Pr(I \mid S)$ is sometimes called the rendering function for this reason. If we know, for instance, that the projection is orthographic and noise free, then we know that for each possible scene S, the likelihood function is a dirac delta function on that image I which is the orthographic projection of S. If instead we know that the projection is orthographic and noisy, and that the noise is distributed as a gaussian $N(0, \sigma)$, then we know that for each possible scene S, the likelihood function is a gaussian of variance σ^2 centered on that image I which is the orthographic projection of S. So if we know how images are rendered as scenes, and we often do have a reasonable idea about this, then we know the likelihood function $\Pr(I \mid S)$.

Do we know the prior probability $\Pr(I)$? Not really. But this might not be a problem in practice. As long as $\Pr(I)$ is not zero, it really doesn't much matter what it is when computing the posterior by Bayes' rule. We can simply view it as a normalizing factor, to make all the conditional probabilities sum to one. Or we can ignore it and just look at the ratios of various posterior probabilities to find the relative likelihoods of various scene interpretations. Of course if $\Pr(I)$ is zero, as may be the case if the space of possible images is nondiscrete, then there is a serious technical issue to deal with. One must reformulate Bayes' rule in a more sophisticated setting, using Radon–Nikodym derivatives and kernels, to get a rigorous result (Bennett et al., 1996). This has been done, but is outside the purview of this chapter.

Do we know the prior probability $\Pr(S)$? To know this would be to know, under a frequentist interpretation of probability, in fact how frequently different scenes in the world actually occur. This is surely a lot to know. It is hard to imagine how, in practice, one could empirically obtain such information. One cannot, in practice, measure all scenes. There simply is not enough time, phylogenetically or ontogenetically, to do so. And one cannot, in principle, arrange for an appropriate statistical sample of scenes from which one can infer the proper population statistics. It's not clear even how one would try to do this. When pollsters sample voters prior to an election, they try to obtain a stratified random sample. They stratify their sampling based on their prior knowledge about properties of the population of voters. Such knowledge about the population of scenes is not at hand.

So a frequentist interpretation of $\Pr(S)$ seems to lead us into trouble. But if we adopt a subjectivist interpretation, then $\Pr(S)$ is the prior probability that the observer assumes to hold for purposes of computing interpretations. $\Pr(S)$ codifies the assumptions of the observer. These assumptions heavily influence, via Bayes' rule, the posterior probabilities assigned by the observer to various scenes, and thus heavily influence what the observer sees. In the example at hand, however, it is hard to imagine a principled assumption to make about the prior probabilities of lines versus curves. Intuitively one might expect that there are more curves than lines, but beyond this it's hard to assign specific probabilities. Let's assume, for now, that both are just given some positive probability.

Assumptions can also affect the likelihood function. For instance, one assumption that the observer might hold in the example at hand is the assumption of a generic view. That is, the observer might assume that all possible viewing directions are equally likely. If one thinks of the set of viewing directions as being isomorphic to the unit sphere, where each point on the sphere represents a different view directed toward the center of the sphere, then the probability of a given set of views is proportional to the area of that set on the unit sphere. Under this assumption it is easy to prove that the set of viewing directions for which curves in space project to the given line in the image is a set which has no area. Indeed this set is a great circle on the unit sphere.

Since this set has no area, it has probability zero. Thus the probability

$$\Pr(I_{\text{line}} \mid S_{\text{curve}}) = 0 \tag{8.4}$$

On the other hand, the set of viewing directions for which straight lines in space project to the image I_{line} is the entire unit sphere. Thus the probability

$$\Pr(I_{\text{line}} \mid S_{\text{line}}) = 1 \tag{8.5}$$

If we assume that $\Pr(S_{\text{curve}}) > 0$, $\Pr(S_{\text{line}}) > 0$, and $\Pr(I_{\text{line}}) > 0$, then by putting (4) and (5) respectively into Bayes' rule (3), we find that

$$\Pr(S_{\text{curve}} \mid I_{\text{line}}) = 0 \tag{8.6}$$

and that

$$\Pr(S_{\text{line}} \mid I_{\text{line}}) = 1 \tag{8.7}$$

Thus an observer who makes the plausible assumption of a generic viewpoint, and who also assumes that curves and lines occur with nonzero probability, is led inexorably to the conclusion that a straight line in an image must be interpreted as a straight line in space.

This example is simple, in that the critical probabilities involved in the computation, namely the likelihoods, are either zero or one so that the resulting posteriors are also zero or one. This simplicity is intentional, to illustrate how Bayesian approaches to vision work, but without getting bogged down in detailed mathematical computations. However the Bayesian approach is, of course, not restricted to these simple cases. The priors and likelihood functions, and therefore the posteriors, can be as nasty as you like. In this manner many interesting problems of vision have been successfully addressed within a Bayesian framework, including visual motion (Bennett et al., 1996; Jepson et al., 1996; Knill et al., 1996), texture (Blake et al., 1996; Witkin, 1981), stereovision (Belhumeur, 1996), shading and lighting (Adelson & Pentland, 1996; Freeman, 1996), and color constancy (Brainard & Freeman, 1994). In many of the cases just cited, the Bayesian analysis leads to an effective computational procedure, thus allowing one to build a process model that can be implemented by computer. So in addition to providing interesting theoretical insights, the Bayesian approach also aids in the construction of working computer vision systems.

TWO APPROACHES TO ONTOGENETIC ADAPTATION

Perceptual adaptation is a refining of an observer's perceptual conclusions in consequence of its interactions with its environment.

In the simple example of the previous section, we saw that (usually unconscious) assumptions by the observer play a key role in the perceptual conclusions it reaches.

These assumptions are modeled in a Bayesian framework by priors and likelihood functions. Changes in the observer's assumptions must, on the Bayesian formulation, be modeled by changes in the observer's priors and likelihoods. These changes will, in general, also lead to changes in the resulting posteriors, i.e., in the observer's perceptual conclusions.

Therefore one way to model perceptual adaptation within a Bayesian framework is to model changes in the likelihoods and priors an observer uses as it interacts with its environment. These changes will systematically affect the perceptual conclusions of the observer, resulting in systematic perceptual adaptation. This approach seems to us most natural, and is the one we will pursue in some detail here. We call it the *structural adaptation* approach.

But there is another approach that one might take. If one never alters the priors or likelihoods, one can still get changes in the posteriors, and therefore perceptual adaptation, by simply conditioning on more and more data. As the observer has more and more commerce with its environment, the observer obtains a larger pool of data on which to do its Bayesian computations and arrive at a posterior. This accumulated data need not affect the *structure* of the likelihood function or priors at all. It can simply change the *argument* given to the likelihood function, and thus lead to a different posterior. This approach seems to us less natural as a model either of ontogenetic or phylogenetic adaptation. It seems unlikely that adaptation is simply a matter of conditioning on more data, without concomitant structural changes in the priors and likelihoods effectively used by the organism. But it is a logical possibility, one that we call the *nonstructural adaptation* approach. In a later section we formally compare the structural and nonstructural approaches to adaptation. This provides clearer insight into both.

BAYES MEETS DARWIN: BASIC IDEAS

Whether one choses a structural or nonstructural approach to ontogenetic adaptation, there remains the problem of placing this ontogenetic adaptation in a phylogenetic context in such a way that satisfies, at least in broad outline, neo-Darwinian accounts of evolution by natural selection. What will not work, of course, are Lamarckian-style theories of either the structural or nonstructural types, in which a parent passes on to its offspring the results of its ontogenetic adaptation. The parent cannot, on a structural approach, pass on to its offspring the structural changes in its likelihoods and priors that took place in the course of its perceptual interactions with its environment. Moreover the parent cannot, on a nonstructural approach, pass on to its offspring the body of data it has accumulated in the course of its perceptual interactions with its environment.

What it can pass on, and all it can pass on, is its genome. On the Bayesian hypothesis this genome encodes, inter alia, the various Bayesian structures and processes (including priors and likelihoods) involved in perception and ontogenetic perceptual adaptation. This genome is subject to random mutations, so that the offspring can, in principle, have slightly different Bayesian structures and processes than the parents.

The key idea (and an old one) is this. If an organism's genetic endowment allows its perceptual system to adapt so well and so quickly to its environment that it can, with greater probability than its competitors, reproduce and pass on its genes, then in succeeding generations the offspring will inherit this advantaged perceptual ability, with minor random mutations. Some of these mutations will be beneficial to perceptual adaptation, more will be harmful. But natural selection will, in this manner, tend to grant higher frequency of offspring to those whose mutations are beneficial.

So there are two evolutionary processes intertwined. There is the ontogenetic evolution of the perceptual system within the life of an individual. And there is the phylogenetic evolution of the perceptual system across generations. Those genomes which grant more effective ontogenetic evolution tend to persist in the phylogenetic evolution.

This story sounds promising in broad outline. To see if it really works it is necessary, of course, to formalize structural and nonstructural Bayesian approaches to perceptual adaptation and to prove that they have the ability to properly converge to or track with those aspects of a changing environment that are crucial to an organism's survival to reproductive age. We now turn to these formal issues.

STRUCTURAL AND CLASSICAL BAYES UPDATING

To recapitulate: The *classical* model operates by sequentially updating the posterior as data arrive, while holding the prior and likelihood function fixed. The sequencing here is based upon successively richer conditioning on present and past data; the strategy of instantaneous inference remains unaffected by such information. In our notation, given a data sequence $(I_1, \ldots, I_k)$, the posterior is updated to

$$\Pr(S \mid (I_1, \ldots, I_k)) = \frac{N((I_1, \ldots, I_k) \mid S) \Pr(S)}{\Pr((I_1, \ldots, I_k))} \tag{8.8}$$

where N is a markovian kernel known as the "likelihood function". Classical updating is also known as Bayesian statistical inference and has a long history in the statistical literature (see, e.g., Diaconis & Freedman, 1986 and references cited therein). A different scheme has been proposed by Bennett and Cohen (1997) as part of their "directed convergence" scheme for acquiring stable percepts. In this scheme, their *structural* procedure involves updating not only of posteriors but also of priors consequent to an observation. When the image I_1 arrives, Bayes' rule produces a posterior $\Pr_1(S) = \Pr(S \mid I_1)$. This posterior is then taken as the new candidate prior to be used for a Bayesian inference at the next arrival of a premise. In contrast to the *iterative* character of classical updating, the structural variety is *recursive*: the result at any given stage changes the very strategy of computation at the next stage.

In order to clarify the differences between these two kinds of procedure, we will develop our notation somewhat. The space of punctual premises will be denoted Y. We wish to include the possibility, in a performance model, of probabilistic premises: those that arrive as a *probability measure*. We will denote such premises (e.g., the data I) by $\lambda(\mathrm{d}y)$. Thus a punctual premise y_0 reappears as the "Dirac measure", or "point mass" $\delta_{y_0}(\mathrm{d}y)$ at y_0: the measure that assigns to any set $B \subset Y$ the value 1 if $y_0 \in B$ and 0 otherwise. The space of punctual percepts, on the other hand, will be denoted X. A probabilistic premise will, in general, lead to a probabilistic percept: we are therefore interested in percepts which are themselves measures on X. Conversely, in the presence of noise, even a punctual state of affairs in the world—say $x \in X$—will lead to the appearance of a probabilistic premise. In this sense, the "noise kernel" or likelihood function N plays the role of an image-rendering function which assigns, to each punctual percept x, the corresponding *probabilistic* premise $N(x, \mathrm{d}y)$. That is, for x in X, $N(x, dy)$ is the probability distribution on Y which expresses the likelihood that premises will be acquired assuming that the system is subject to an ambient state of affairs represented by x. As a final motivation for probabilistic premises and percepts, note that Bayes' rule gives a procedure whereby premises (probabilistic *or* punctual) are transformed to (probabilistic) conclusions: we will call this updating procedure the *updating law* and denote it by P. Thus, given a premise λ, the updating law produces a posterior probability $\mu(\mathrm{d}x)$ on X by integrating over the premise distribution:

$$\mu(\mathrm{d}x) \underset{\mathrm{def}}{=} \lambda P \underset{\mathrm{def}}{=} \int \lambda(\mathrm{d}y) P(y, \mathrm{d}x) \tag{8.9}$$

Of course, this updating law[1] P has further structure that depends on the image rendering function N and the current prior μ_0. Bayes' rule in the discrete case expresses this dependence as

$$P_{(\mu_0, N)}(y, A) = \frac{\sum_{x \in A} \mu_0(x) N(x, y)}{\sum_{x \in X} \mu_0(x) N(x, y)} \tag{8.10}$$

where we have explicitly displayed the dependence on the current prior and on the likelihood function as subscripts. In the continuous case we have

$$P_{(\mu_0, N)}(y, A) = \frac{\mathrm{d}\big(\int_{x \in A} \mu_0(\mathrm{d}x) N(x, \mathrm{d}y)\big)}{\mathrm{d}\big(\int_{x \in A} \mu_0(\mathrm{d}x) N(x, \mathrm{d}y)\big)}(y) \tag{8.11}$$

Here the right-hand side is given as a Radon–Nikodym derivative (see Bennett et al., 1996, equations (5.21) and (5.28)). The likelihood, or image rendering function N takes percepts and yields premises, according to physical laws of projection, refraction, noise, etc. P, in turn, accepts prior probabilities and, employing the offices of N, gives a procedure for transforming those priors into posteriors upon the arrival of a premise.[2]

In classical updating, P remains the same regardless of the sequence of percepts. In structural updating, by contrast, at stage n, where the prior is μ_n, the law is $P_{(\mu_n, N)}$. Then, in the $(n+1)$th stage, a premise λ_{n+1} is acquired; this results in the new conclusion μ_{n+1} by

$$\mu_n = \lambda_n P_{n-1} \tag{8.12}$$

This μ_{n+1} then becomes the prior at the next stage. It is a fact, though we shall not prove it here, that the two procedures are identical for the sequences of punctual premises, i.e., Dirac measures, that the classical case considers. However, the structural procedure is vastly more general, in that it allows the use of nonpoint masses and therefore of much more general kinds of premise. Moreover, the structural procedure allows us to improve on the efficiency of convergence: Even with a sequence $(y_1, \ldots, y_k)$ of punctual premises, the structural method allows us to use, at stage n, the premise

$$\lambda_n(\mathrm{d}y) = \frac{1}{n} \sum_{k=1}^{n} \delta_{y_k}(\mathrm{d}y) \tag{8.13}$$

We expect the structural posteriors using this sequence of premises to converge more rapidly than the classical posteriors (which, perforce, use punctual premises) to the true state of affairs.

Let us use a simple example to illustrate updating. Suppose we are to infer the relative probabilities of heads and tails for a biased coin. The probability of a head, to be inferred, is some number $x \in X = [0, 1]$. Successive premises y_k consist of tossing the coin and observing the outcome. So $y_k = 1$ or 0, depending on whether we observe a head or a tail. Now if the probability of a head were actually x, the distribution of premises y would be governed by $N(x, \mathrm{d}y) = x^{|y|}$, where $|y| = \sum_{k=1}^{n} y_k$ is the number of heads in the n observations. Then the random vector $\{y_k\}_{k=1}^{n} = (y_1, \ldots, y_n)$ is distributed binomially, i.e., the probability of any n-long sequence of heads and tails equals a product of x's (one for each $y_k = 1$) and of $(1 - x)$'s (one for each $y_k = 0$). This is just $x^{|y|}(1 - x)^{n-|y|}$. Finally, the random variable x is itself taken to be distributed according to the prior μ_0, so in terms of the posterior probability on [0, 1], the previous expression is proportional to a *density* with respect to μ_0. Hence classical Bayesian statistical inference says that, after n observations have been obtained, and given a prior probability μ_0 (on the x's), the posterior probability is

$$\mu_n\left(\mathrm{d}x \mid \{y_k\}_{k=1}^{n}\right) = \frac{x^{|y|}(1 - x)^{n-|y|}\mu_0(\mathrm{d}x)}{\int x'^{|y|}(1 - x')^{n-|y|}\mu_0(\mathrm{d}x')} \tag{8.14}$$

the denominator in (8.14) is the normalization which makes the left-hand side a probability.

We would like these posteriors in Equation (8.14) to converge to the point mass (or Dirac delta) at the true value $\bar{x}$, given that the data $\{y_k\}$ are independent and identically distributed with a probability of $\bar{x}$ for heads. Such a convergence of the posteriors to the true coin probability is called *consistency* in the literature. It is well known that whenever the prior μ_0 assigns positive measure to every open interval around $\bar{x}$, we have consistency: The measures μ_n converge *weakly* to Dirac measure at $\bar{x}$. That is, if A is any (measurable) subset of the interval [0, 1],

$$\mathrm{Lim}_{n\to\infty}\, \mu_n\left(A|\{y_k\}_{k=1}^{n}\right) = \begin{cases} 1 & \text{if } \bar{x} \in A \\ 0 & \text{otherwise} \end{cases} \tag{8.15}$$

for almost all[3] input sequences $\{y_k\}_{k=1}^{\infty}$. In other words, the limit is the point mass at $\bar{x}$.

This is satisfying; there are, however, more complicated situations where consistency does not obtain, in a generic sense (Diaconis & Freedman, 1986).

In general, consistency for Bayes updating is defined as follows: Given that the data arrive under the law $N(\bar{x}, \mathrm{d}y)$ for a (punctually) true state of affairs $\bar{x}$, then we say that the pair $(\mu_0, \bar{x})$ consisting of the initial prior and the actual situation $\bar{x}$ is *consistent* if , the posteriors μ_n defined in (8.9) and (8.10) (or (8.11)) above *converge weakly* to Dirac measure at $\bar{x}$. This means that for every bounded, continuous, real-valued function f on X, if we define

$$\mu_n(f) \underset{\text{def}}{=} \int_X \mu_n(\mathrm{d}x) f(x) \tag{8.16}$$

then

$$\mu_n(f) \longrightarrow \delta_{\bar{x}}(f) = f(\bar{x}) \tag{8.17}$$

as $n \to \infty$. Weak convergence is a natural notion, in that the collection of bounded continuous functions on X can be thought of as *observables* for the states of the world: in state x, the observable f has the value $f(x)$. Thus consistency is the requirement that in the Bayes updating scheme, the limiting values of all observables are their true values – surely an operationally sound notion. Moreover, weak convergence has the technical advantage that it is the easiest of the various kinds of convergence criteria to satisfy: if a given sequence of measures is weakly convergent, then it is so in other ways too.

We will refer to the above definition of consistency as *classical adaptability*: the pair (μ_0, N) is (classically) adaptable to the state of affairs $\bar{x}$ if $(\mu_0, \bar{x})$ is consistent as in (8.16) above. That is, for almost all input sequences $y_1, y_2, y_3, \ldots$, the sequence of distributions $\mu_1, \mu_2, \mu_3, \ldots$ on X obtained as $n \to \infty$, by successively conditioning on the first n terms of $y_1, y_2, y_3, \ldots$, converges weakly to Dirac measure at $\bar{x}$. Classical adaptability corresponds, in evolutionary situations, to the capability of an organism to attain a *stable* perceptual representation of the persistent environmental feature represented by $\bar{x}$, beginning with the initial perceptual representation μ_0. Note, however, that this definition of adaptibility is of practical value only if punctual inputs are received according to the law $N(\bar{x}, \cdot)$ for some *fixed* $\bar{x}$ in X.

We now state the natural generalization of classical adaptability to the structural situation:

> The pair (μ_0, N) is *adaptable* to the probability measure μ on X if the structural process with initial prior μ_0 and repeated premise $\lambda = \mu N$ converges weakly to a probability measure μ_∞. We call this weak limit μ_∞ the *adaptation* of (μ_0, N) to μ.

Structural adaptability corresponds, in evolutionary situations, to the capability of an organism to attain a stable perceptual representation of the stable environmental feature represented by μ, beginning with the initial perceptual representation μ_0. Clearly the faster the rate of adaptation, the more likely is it that this particular adaptability will persist through generations.

BAYES AND DARWIN MEET: DIRECTED CONVERGENCE

Consider an inferencing system (X, Y, N, μ_0), where X is the space of conclusions for the inference, Y is the space of premises, and N is the 'image rendering kernel'. μ_0 is the 'prior' measure on X which encodes the system's initial subjective probabilities for conclusions in X.

The classical process can reasonably be used to acquire stable inferences only in situations where (i) the points of Y represent premises which are acquired with perfect discrimination by the system, and where (ii) the points in X irredundantly parameterize, via N, *all* possible distributions of images which might occur in practice. In fact, with regard to (ii), given N and the prior μ_0, the question of whether or not classical adaptibility holds is meaningful only when the law governing the punctual inputs at each stage is $N(\bar{x}, \cdot)$ for some fixed $\bar{x}$ in X. If for some $\bar{x}$ in X, (μ_0, N) is not adaptable to $N(\bar{x}, dy)$, then a state of affairs corresponding to $\bar{x}$ (i.e., a state of affairs which generates premises according to the law $N(\bar{x}, dy)$) will never be stably inferred. On the other hand, if there is more than one $\bar{x}$ in X with the same $N(\bar{x}, \cdot)$ then classical adaptability becomes almost meaningless. This suggests that we should weaken the definition to be something like: (μ_0, N) is adaptable to $\bar{x}$ if the posteriors in the classical process converge to a measure supported on the set $V_{\bar{x}} = \{x' \in X : N(x', \cdot) = N(\bar{x}, \cdot)\}$. Putatively this measure will be the restriction of μ_0 to V_x, normalized to a probability measure. Without additional information there is no way to make a stronger inference than this, so this weaker type of adaptability at least provides a "platform" for a more specialized inference that might permit discrimination within V_x. Note that if $V_{\bar{x}}$ has more than one element in it, such a measure on $V_{\bar{x}}$ represents a *multistable* percept.

We will assume that we have a standard way to measure the rate of weak convergence. Then if (μ_0, N) is adaptable to $\bar{x}$, we will denote by $f(\mu_0, N; \bar{x})$ the *reciprocal* of the rate of convergence to Dirac measure at $\bar{x}$ of the classical process starting with (μ_0, N). We might then interpret $f(\mu_0, N; x)$ as the 'length of time required for convergence'. For example, if (μ_0, N) is not adaptable to x then $f(\mu_0, N; x)$ will be infinite.

For our purposes we can define an *environment* to mean the collection of possible ambient states of affairs, together with their probabilities of occurrence. Suppose an organism's perceptual inferencing system utilizes (X, Y, N, μ_0). In the case of systems which employ the classical process to acquire stable percepts, we will assume that these "possible ambient states of affairs" are represented by points of X. Thus, a *(classical) environment E is specified by a probability measure ρ_E on X*. This measure, called the *underlying environmental measure*, is meant to give a much more global and long-term description of the environment than does μ_0, or do the measures on X which embody perceptual inferences. In fact ρ_E is intended to provide comprehensive information about the relative frequency of occurrence, *over an extended period of time*, of the various states of affairs represented by points of X. By contrast, an inference is intended to describe the state of affairs encountered by the organism more instantaneously, i.e., an inference intends to describe a much more specific and transient state than does the underlying environmental measure ρ_E.

In addition to frequency of occurence of environmental states of affairs, we will also assume that the underlying environmental measure ρ_E contains information about the degree to which the states of affairs are adaptively critical, i.e., information about the *survival value* of the inference. Consider, for example, a creature which must drink water every few days for survival, and which is also the prey of an extremely deadly predator which comes into the vicinity, say, once every few months. For purposes of survival the abilities to correctly infer the presence of the predator, or to correctly infer the presence of water, are of equal importance. So, in spite of the fact that the corresponding states of affairs occur with very unequal frequency, they may be given equal weight by ρ_E. We conclude that if $\rho_E(S)$ is large for some $S \subset X$, then the environmental states of affairs represented by S are collectively significant for survival, perhaps because they are only moderately critical but occur frequently, or perhaps because they are enormously critical but occur only rarely.

We can now define the *classical adaptivity of the organism to an environment E* for the case where stable inferences are acquired via the classical process. We assume X and Y are fixed, so that in effect we are defining the adaptivity $A_{\mathrm{cl}}(\mu_0, N; E)$ of (μ_0, N) to E as

$$A_{\mathrm{cl}}(\mu_0, N; E) = \frac{1}{\int_X \rho_E(\mathrm{d}x) f(\mu_0, N; x)} \tag{8.18}$$

The adaptivity is an indicator of the time required for the organism to arrive at stable percepts which stably represent environmental states of affairs as they are encountered. This indicator takes into account, via ρ_E, the relative likelihood of encountering the various states of affairs, as well as the survival value of a correct inference in the context of the encounter. The smaller the average value of f, i.e., the less the average time required for adaptation, the larger will be the value of the adaptivity. Hence the larger the value of the adaptivity, the more rapidly does the organism's perceptual system make correct inferences about environmental states of affairs. To say that the adaptivity is infinite means that the function $f(\mu_0, N; x)$ on X has value 0 except possibly on some subset of X which has ρ_E-measure 0. This means that stable percepts are instantaneously inferred in almost all environmental conditions, i.e., in all conditions except those represented by the ρ_E-measure zero subset of X. On the other hand, to say that the adaptivity is 0 means that there is a class of environmental states of affairs which are significant for survival (i.e., they are represented by a set S in X for which $\rho_E(S) > 0$), but the perceptual adaptation to these environmental states is very slow (i.e., f takes large values on S). In particular, if (μ_0, N) is *not adaptable* to x in X, then we make take $f(\mu_0, N; x)$ to be ∞, so if nonadaptability holds on any set S in X with $\rho_E(S) > 0$, then the adaptivity will be 0.

We now consider the case of an organism which acquires stable percepts using the structural process; we will indicate the appropriate definitions of adaptability and adaptivity. In this case, as in the classical case, we assume that the basic data for the organism's perceptual inferencing system is (X, Y, N, μ_0), where these symbols have the same meanings as above. But in this case the priors, beginning with μ_0, are updated recursively, based on premises which are probabilistic, i.e., the premises are probability measures λ on the scene space Y. To review the updating procedure, suppose that at time n the updated prior is μ_n. We then have the Bayesian posterior kernel $P_{(\mu_n,N)}$ for this prior μ_n and the likelihood kernel N.

(Recall that $P_{(\mu_n,N)}$ is the kernel from Y to X which has the following interpretation: Assume that the distribution of states of affairs is given by the probability measure μ_n on X, and assume that $N(x, \cdot)$ describes the probabilities of scenes in Y being acquired as premises given that the actual state of affairs is x. Then, for $y \in Y$, $P_{(\mu_n,N)}(y, S)$ is the probability that an environmental state of affairs which is represented by some point of S was transduced, given that the scene y was acquired as a premise.)

Suppose that at the next $(n + 1)$th instant the premise λ_{n+1} is acquired. Recall that the system then infers the measure $\lambda_{n+1} P_{(\mu_n,N)}$ on X, which becomes the next prior μ_{n+1}, i.e., the priors are updated recursively according to the law $\mu_{n+1} = \lambda_{n+1} P_{(\mu_n,N)}$, where μ_n denotes the prior at time n, $P_{(\mu_n,N)}$ is the Bayes' posterior kernel for this prior and for the likelihood kernel N, and λ_{n+1} is the premise at time $n + 1$. Thus, for $A \subset X$,

$$\mu_{n+1}(A) = \int_Y \lambda_{n+1}(\mathrm{d}y) P_{(\mu_n,N)}(y, A) \tag{8.19}$$

In this situation, a stable percept is a weakly convergent sequence of measures μ_n on X which arises from this updating procedure for some sequence of premises λ_n.

Now suppose that an environmental state of affairs is described as a probability measure μ on X. Here, the meaning of the likelihood kernel N is that the premise scene transduced from μ is described by the probability measure μN on Y, defined by $\mu N(B) = \int_X \mu(\mathrm{d}x) N(x, B)$. We will say that (μ_0, N) *is structurally adaptable to* μ if the sequence of measures μ_n on X, defined recursively by (8.19) (beginning with μ_0), converges weakly *for the constant sequence of premise measures* $\lambda_n = \mu N$ *on* Y. In other words, (μ_0, N) is adaptable to μ if, beginning with μ_0, the system acquires a stable percept in the presence of that persistent environmental state of affairs which corresponds to μ. Let μ_∞ denote the weak limit of the sequence μ_n; this μ_∞ exists by definition in case of adaptability as above. For the definition here of structural adaptability of (μ_0, N) to μ it is too much to require that $\mu_\infty = \mu$, just as for classical adaptability of (μ_0, N) to x in X it is too much to require that the classical process converges to x. All that we can ask is that a stable percept be acquired in the presence of the given persisting environmental state of affairs. In the classical case, if an environmental state corresponds to the punctual x in X, then for the system to be in the "presence of a persistent environmental state" means that the system obtains sequences of punctual premises y_n in Y which are independent and identically distributed with the law $N(x, \cdot)$. In the structural case, if an environmental state corresponds to the probability measure μ on X, then for the system to be in the "presence of a persistent environmental state" means that the system obtains a sequence of probabilistic premises λ_n which are identically equal to μN.

We now consider the meaning of adaptivity in the structural case. As in the classical case, we assume we have a measure of the rate of weak convergence, and we denote its reciprocal by f. If (μ_0, N) is adaptable to μ, then $f(\mu_0, N; \mu)$ may be interpreted as the length of time required for the convergence of the structural process $\{\mu_n\}$, beginning with μ_0 generated by the constant premise sequence $\lambda_n = \mu N$. If (μ_0, N) is not adaptable to μ, then we may set $f(\mu_0, N; \mu) = \infty$. We must also specify a precise definition of "environment" as we did for the classical case; the idea again is that an environment E is represented by an underlying environmental measure ρ_E, which is a measure on the space of environmental states of affairs. In the structural case, while X is a "configuration space" for these states, the actual states are identified with probability measures μ on X. Thus, denoting the set of probability measures on X by $\mathcal{P}(X)$, we will define a *(structural) environment* to be a probability measure ρ_E on $\mathcal{P}(X)$.[4] We can now define the *structural adaptivity of an organism* (μ_0, N) *to an environment* E as

$$A_{\text{st}}(\mu_0, N; E) = \frac{1}{\int_{\mathcal{P}(X)} \rho_E(d\mu) f(\mu_0, N; \mu)} \tag{8.20}$$

Note that μ is now the variable of integration on $\mathcal{P}(X)$. As in the classical case, the number $A_{\text{st}}(\mu_0, N; E)$ may be interpreted as the expected rate of perceptual adaptation to the environment, a rate which is adjusted for survival value of the various environmental conditions.

CONCLUSIONS

One significant distinction between structural and classical Bayesian updating of perceptual inferences is that a *directed convergence* strategy is available in the structural case (Bennett & Cohen, 1997). Directed convergence is a strategy for acquiring stable percepts, i.e., convergent sequences of instantaneous percepts, even in a noisy environment with ubiquitous distractors. The idea is to decide whether or not to incorporate a premise λ into the updating procedure based on (i) how close λ appears to be to the current percept, and (ii) how strong is the belief that the current percept is close to a "correct percept". To make this precise, suppose at time n an organism (μ_0, N) which uses structural updating has percept μ_n, so that its updating law is the Bayesian posterior kernel $P_{(\mu_n, N)}$. Suppose that there is a strong belief that μ_n is close to a correct percept "μ". In fact the degree of that belief may be expressed as the distance within which it is believed that μ_n lies from μ. Then $\mu_n N$ should be correspondingly close to μN, since N is continuous as a function from $\mathcal{P}(x)$ to $\mathcal{P}(Y)$ ($\mathcal{P}(X)$ denotes the probability measures on X, etc.). Now when we say that μ is a "correct percept" we mean that it represents a stable environmental feature *of interest* to the organism. Because it represents a stable environmental feature, we expect that it will be transduced, i.e., there is a nontrivial probability that any given premise transduces that feature. And if a premise λ does transduce the feature represented by μ, then λ will be close to μN and consequently close to $\mu_n N$. Thus, to the extent to which the degree of belief in μ_n is justified, within a reasonable time interval the organism will receive a premise λ that *confirms the degree of belief in* μ_n, in the sense that it lies sufficiently close to $\mu_n N$. Thus, suppose that at the next $(n+1)$th instant a premise λ is received. If λ confirms the degree of belief in μ_n, since the organism is *interested* in the feature μ, λ will be accepted as λ_{n+1}, and hence incorporated in the updating procedure. Indeed we will then have $\mu_{n+1} = \lambda P_{(\mu_n, N)}$. On the other hand, if a λ which confirms the degree of belief in μ_n is not received within a reasonable time, then there will be reason to modify the degree of belief to suppose that μ_n lies at a greater distance from a correct percept than was originally thought, and hence justifying acceptance of premises λ which lie further from $\mu_n N$ than was acceptable previously.

The use of such a strategy introduces a "flexibility of direction" into the updating procedure, that maximizes the possibility of convergence to a conclusion which represents one of a possible multitude of environmental phenomena, each of which may be responsible for a share of the raw premises obtained by the organism. By using the directed convergence strategy, the conclusions are updated only in response to the premises in a recursively selected subsequence (of the sequence of *all* premises); the actual selection occurs as the result of an ongoing balancing dance of belief and confirmation.

It seems clear that the organism (μ_0, N) must use something similar to a directed convergence strategy to selectively incorporate premises into the updating procedure. Otherwise, given the complexity of the environment, with numerous features being transduced on the sensorium in the presence of noise and other perturbations, it is hard to imagine that the raw sequence of incoming premises would correspond to just one of those features, and would do so in a manner which would produce, via updating, a convergent percept sequence. But in any case it is unreasonable to expect that, in practice, the premise sequences that yield the stable percepts (i.e., that yield convergent percept sequences), are *constant*. In other words, suppose the organism obtains a sequence of premises λ_n which give rise to the convergent percept sequence μ_n, whose limit μ represents a stable environmental feature. In fact, suppose for simplicity that this premise sequence arose from transduction of that very feature. In practice, this does *not* mean that $\lambda_n = \mu N$ for all n. For, because of various perturbations whose effect is not completely subsumed in the "noise" kernel N, the transduction will result in premises which are close to μN but not equal to it. Therefore the most we can reasonably expect in this situation is that the premises λ_n converge to $\lambda = \mu N$ (as the percepts μ_n converge to μ). Indeed if we imagine that a raw sequence of premises which are random perturbations of μN is obtained, then the directed convergence procedure will lead to the recursive selection of a subsequence which optimizes the possibility of convergence of the μ_n to μ.

But recall that the *adaptivity* of (μ_0, N) to μ and the associated function $f(\mu_0, N; \mu)$ were previously defined in terms of the sequence of percepts μ_n which is generated in response to the *constant* premise sequence $\lambda_n = \lambda = \mu_n N$ for all n. How is this definition relevant to the "real world" situation where the adaptivity of the organism to μ depends on its response to a non-constant premise sequence λ_n which converges to λ? The idea is that the adaptivity defined in terms of the constant sequence λ is an idealized version of the response of the organism to a random premise sequence which converges to λ. And in this spirit, if μ and μ' are measures on X, the relative values of $f(\mu_0, n; \mu)$ and $f(\mu_0, N; \mu')$ represent the relative expected rates of convergence of the percepts, in response to random premise sequences which converge on the one hand to μN and on the other hand to $\mu' N$.

We expect that ongoing mathematical investigations will clarify these intuitions, and lead ultimately to computable models.

NOTES

1. Note that we often write the argument *before* a function, rather than after it: this leads to a certain notational convenience, as we shall see.
2. So far, these notational changes are purely formal and seek to express the essential functional relationships in Bayes' rule. We will see, however, that there is much more than formality to the content of this notation.
3. By "almost all" input sequences we mean "with respect to the usual measure extended from that defined on cylinder sets of the set of sequences, i.e., infinite product set $Y \times Y \times \cdots$, given by $N(\bar{x}, dy)$ in each factor.
4. For this purpose we should assume that X *is a complete separable metric space*, whose σ-algebra of measurable sets is generated by the open sets of its metric topology. Then, by theorems of Prohorov [Billingsley, 1968], $\mathcal{P}(X)$ is also a complete separable metric space, with its corresponding σ-algebra of measurable sets; ρ_E is a measure for this σ-algebra.

REFERENCES

Adelson, E.H. & Pentland, A.P. (1996). The perception of shading and reflectance. In D.C. Knill & W.A. Richards (Eds.), *Perception as Bayesian inference* (pp. 409–423). Cambridge: Cambridge University Press.

Beatty, D.D. (1984). Visual pigments and the labile scotopic visual system of fish. *Vision Research*, **24**, 1563–1573.

Belhumeur, P.N. (1996). A computational theory for binocular stereopsis. In D.C. Knill & W.A. Richards (Eds.), *Perception as Bayesian inference* (pp. 323–364). Cambridge: Cambridge University Press.

Bennett, B.M. & Cohen, R.B. (1997). Directed convergence and stable percept acquisition. (In preparation). xxx

Bennett, B.M., Hoffman, D.D., Prakash, C. & Richman, S.N. (1996). Observer theory, Bayes' theory, and psychophysics. In D.C. Knill & W.A. Richards (Eds.), *Perception as Bayesian inference* (pp. 163–212). Cambridge: Cambridge University Press.

Billingsley, P. (1968). *Convergence of probability measures*. Wiley, New York.

Blake, A., Bülthoff, H.H. & Sheinberg, D. (1996). Shape from texture: Ideal observers and human psychophysics. In D.C. Knill & W.A. Richards (Eds.), *Perception as Bayesian inference* (pp. 287–321). Cambridge: Cambridge University Press.

Brainard, D.H. & Freeman, W.T. (1994). Bayesian method for recovering surface and illuminant properties from photosensor responses. *Proceedings of SPIE, 2179*. San Jose, California, February 1994.

Crawford, M.L.J., Anderson, R.A., Blake, R., Jacobs, G.H. & Neumeyer, C. (1990). Interspecies comparisons in the understanding of human visual perception. In L. Spillman & J.S. Werner (Eds.) *Visual perception: The neurophysiological foundations.* San Diego: Academic Press.

Cronin, T.W. & Marshall, N.J. (1989). A retina with at least ten spectral types of photoreceptors in a stomatopod crustacean. *Nature*, **339**, 137–140.

Cronin, T.W., Marshall, N.J. & Land, M.F. (1994). The unique visual sysem of the mantis shrimp. *American Scientist*, **82** (4), 356–365.

Delius, J.D. & Emmerton, J. (1979). Visual performance in pigeons. In A.M. Granda & J.H. Maxwell (Eds.) *Neural mechanisms of behavior in the pigeon.* New York: Plenum Press.

Diaconis, P. & Freedman, D. (1986). On the consistency of Bayes estimates. *The Annals of Statistics*, **14** (1), 1–26.

Emmerton, J. & Delius, J.D. (1980). Wavelength discrimination in the "visible" and ultraviolet spectrum by pigeons. *Journal of Comparative Physiology A*, **141**, 47–52.

Exner, S. (1891). *The physiology of the compound eyes of insects and crustaceans.* Translated by R.C. Hardie (1989). Berlin: Springer.

Freeman, W.T. (1996). The generic viewpoint assumption in a Bayesian framework. In D.C. Knill & W.A. Richards (Eds.), *Perception as Bayesian inference* (pp. 365–389). Cambridge: Cambridge University Press.

Jacobs, G.H. (1981). *Comparative color vision.* New York: Academic Press.

Jepson, ., Richards, . & Knill, D. (1996). xxx

Knill, D.C. & Richards, W.A. (Eds.) (1996). *Perception as Bayesian inference.* Cambridge: Cambridge University Press.

Knill, D.C., Kersten, . & Mamassian, . (1996). xxx

Kunze, P. (1979). Apposition and superposition eyes. In H.-J. Autrum (Ed.) *Handbook of sensory physiology* (*vol. VII/6A*, pp. 441–502). Berlin: Springer.

Land, M.F. (1985). The morphology and optics of spider eyes. In F.G. Barth (Ed.) *Neurobiology of arachnids* (pp. 53–78). Berlin: Springer.

Land, M.F. (1991). Optics of the eyes of the animal kingdom. In J.R. Cronly-Dillon & R.L. Gregory (Eds.) *Evolution of the eye and visual system* (pp. 118–135). Boca Raton, FL: CRC Press.

Lythgoe, J.N. (1988). Light and vision in the aquatic environment. In J. Atema, R.R. Fary, A.N. Popper & W.N. Tavolga (Eds.) *Sensory biology of aquatic animals* (pp. 57–82). New York: Springer.

Lythgoe, J.N. (1991). Evolution of visual behavior. In J.R. Cronly-Dillon & R.L. Gregory (Eds.) *Evolution of the eye and visual system* (pp. 3–14). Boca Raton, FL: CRC Press.

Neumeyer, C. (1985). An ultraviolet receptor as a fourth receptor type in goldfish color vision. *Naturwissenschaften*, **72**, 162–163.

Neumeyer, C. (1986). Wavelength discrimination in the goldfish. *Journal of Comparative Physiology*, **158**, 203–213.

Nilsson, D.-E. (1989). Optics and evolution of the compound eye. In D.G. Ṡtavenga & R.C. Hardie (Eds.) *Facets of vision* (pp. 30–73). Berlin: Springer.

Pumphrey, R.J. (1961). Concerning vision. In J.A. Ramsay & V.B. Wigglesworth (Eds.) *The cell and the organism* (pp. 193–208). Cambridge: Cambridge University Press.

Salvini-Plawen, L.V. & Maxx, R. (1977). On the evolution of photoreceptors and eyes. *Evolutionary Biology*, **10**, 207–263.

Witkin, A. (1981). Recovering surface shape and orientation from texture. *Artificial Intelligence*, **17**, 17–45.

Wright, A. (1979). Color-vision psychophysics: a comparison of pigeon and human. In A.M. Granda & J.H. Maxwell (Eds.) *Neural mechanisms of behavior in the pigeon.* New York: Plenum Press.

AQ1: See Msp page 3 for Year?

AQ2: See Msp page 24 for References?