Advances in Applied Mathematics **23,** 255–273 (1999)
Article ID aama.1999.0656, available online at http://www.idealibrary.com on IDEAL®

L^∞ Metric Criteria for Convergence in Bayesian Recursive Inference Systems*

Bruce M. Bennett and Rachel B. Cohen

Motivated by applications to probabilistic inference, we consider a sequence of probability measures, called "conclusion measures," on a fixed space X. The sequence is generated recursively via conditional probability, driven by a sequence of input measures (rather than by a sequence of punctual data, as in Bayesian statistical inference). The general problem is to give conditions on the input measures such that the sequence of conclusion measures converges weakly. We develop L^∞-metric criteria defined recursively on the input measures, which are sufficient (but not necessary) for the sequence of conclusion measures to converge at a given rate. We discuss the applications of this to the "directed convergence strategy" introduced in [1]. Finally, we show that if the input measures satisfy the criteria, then the input sequence also converges at a comparable rate. © 1999 Academic Press

INTRODUCTION

Probabilistic inference arises naturally in models of perception, rational deliberation, and other cognitive capacities. In this paper we are interested in the convergence of certain kinds of sequences of probabilistic inferences; a convergent sequence corresponds to a *stable state* of the inferencing system, e.g., a "stable percept" or "stable concept."

An *inference* is a function Ψ: $G \to F$ where F and G are sets of propositions; the elements of G are called the *premises* and the elements of F are called the *conclusions*. For example, in applications to perception, a conclusion proposition asserts that some state of affairs holds in an environment, and a premise proposition asserts that some state of affairs holds on a sensorium, or input receptor array. Given Ψ, we also use the

* This work has been supported in part by National Science Foundation grant No. SBR-9014278. to the University of California, Irvine.

0196-8858/99 $30.00

term ''inference'' to refer to the effect of Ψ on a particular element λ of G, i.e., $\lambda \mapsto \Psi(\lambda)$. A *probabilistic inference* is an inference for which there are measurable spaces Y and X, so that $G = \mathscr{P}(Y)$ and $F = \mathscr{P}(X)$, where $\mathscr{P}(Y)$ and $\mathscr{P}(X)$ denote the sets of probability measures on Y and X respectively. (We view a probability measure μ as the conjunction of propositions of the form ''the probability of A is $\mu(A)$,'' as A ranges over the measurable sets.) X is called the *conclusion configuration space*, and Y is called the *input configuration space*. Intuitively, X and Y are what the conclusions and premises ''would be'' if there were no noise and perfect resolution.

In Section 1 we introduce *recursively updated Bayesian probabilistic inference*, which is a dynamical system on a space of ''conclusion'' probability measures driven by ''input'' probability measures. This is a truly recursive system where the prior and the Bayes posterior are recalculated at every step. It should be distinguished from the usual Bayesian *statistical* inference where a fixed prior is successively conditioned on a sequence of punctual inputs. It is in this Bayesian recursive context that we will study convergence of inference sequences of probability measures: the main point of this article is to develop conditions on the sequence of input measures in $\mathscr{P}(Y)$ such that the corresponding sequence of conclusion measures in $\mathscr{P}(X)$ converges weakly.

Thus the natural spaces for our consideration are $F = \mathscr{P}(X)$ and $G = \mathscr{P}(Y)$ with their weak topologies. But in the present paper we take an indirect approach: we look at these spaces through an ''L^∞-window.'' This means that we choose some measure ν, say on X, and associate to each ν-essentially bounded measurable function f the measure $fd\nu$; this association is a continuous 1−1 map from $L^\infty(X, \nu)$ to $\mathscr{P}(X)$. In this way, we can transport the L^∞ metric to appropriate subsets of $\mathscr{P}(X)$ (to obtain a stronger topology than the weak topology). When we do this, we get effective recursively determined L^∞ metric criteria on the sequence of input measures, for the sequence of conclusion measures to be L^∞-convergent and hence to be weakly convergent (Theorems 13, 14).

Of course, the topology of weak convergence itself is metrizable in various ways, notably by means of the Prohorov metric [3]; in principle one can develop convergence criteria in terms of such metrics. In practice, however, it is dramatically less computationally expensive to use these L^∞ criteria for weak convergence than to use, for example, the Prohorov metric criteria. However, since the L^∞ topology on these measures is stronger than the weak topology, we cannot detect *all* weakly convergent sequences of measures by looking through the L^∞ window. In other words, if we use the L^∞ criteria we may miss some convergent inference sequences. Which ones will we miss? They are typified by the well-known example of a sequence of normal distributions with variances decreasing to

0, which converges weakly to Dirac measure at a point. Intuitively, say in the case of visual perception, for such sequences some feature of an input image is becoming *perfectly* resolved. But in practice, in any actual environment there is always a bound on the possible degree of resolution of inputs, so that sequences of this type would not be expected to occur. Thus, in applications to real-world perceptual inferences, for example, the theoretical limitations of the use of L^∞ may be of minimal significance.

One of the main applications of our convergence criteria is to the formulation of *directed convergence strategies* ([1]). A directed convergence strategy provides metric criteria to recursively select or reject inputs at a given stage of an inference sequence. To ''select'' an input means to use it to update the procedure to the next stage, and to ''reject'' an input means to ignore it for updating purposes, i.e., to wait for another input. In this manner a convergent inference sequence is generated, provided that inputs which satisfy the directed convergence criteria at each stage are forthcoming. Intuitively, we can think of directed convergence as a kind of ''tracking procedure'' whose goal is to ''lock on'' to some object in the environment; the ''forthcomingness'' of the criteria-satisfying inputs means that the rate of convergence of the sequence is consistent with locking on to an accessible object. But if the appropriate inputs are not sufficiently forthcoming to maintain a feasible rate of convergence, then the directed convergence strategy naturally leads to the casting of a wider net for accessible objects. Theorems 13 and 14 in Section 2 expresses the key technical point for directed convergence strategy via the L^∞ criteria.

1. RECURSIVELY UPDATED BAYESIAN PROBABILISTIC INFERENCE

Having defined *inference* and *probabilistic inference* in the Introduction we now define *Bayesian probabilistic inference*. For this purpose, we will assume that we are given measurable spaces X and Y, which are to be the conclusion and premise configuration spaces. We will also assume that we are given a Markovian kernel N: $X \times \mathscr{Y} \to [0, 1]$, where $\mathscr{Y}$ denotes the σ-algebra of measurable sets of Y. (Recall that to say N is ''Markovian'' means that for x in X, $N(x, \cdot)$ is a probability measure on Y.) Intuitively, for x in X and B in $\mathscr{Y}$, $N(x, B)$ is the probability that a premise y in B would be acquired (assuming perfect resolution at the receptor array) given that an environmental state of affairs represented by x is transduced at the receptor array. N is called a *noise kernel*; in statistics N is sometimes called a ''likelihood function.'' N acts in a natural way as a

function

$$N: \mathscr{P}(X) \to \mathscr{P}(Y),$$

via $\mu \mapsto \mu N$, where μN is defined by

$$\mu N(x, B) = \int_X \mu(dx) N(x, B)$$

for μ in $\mathscr{P}(X)$ and B in $\mathscr{Y}$. Finally, we will assume that we are given a probability measure μ in $\mathscr{P}(X)$, called the *prior*. Intuitively, the prior μ represents the current preconception of the state of affairs in the world; the purpose of an inference now is to update that preconception, given a premise measure λ in $\mathscr{P}(Y)$.

With the data (N, μ), the apparatus of conditional probability canonically gives rise to a kernel $P_{(\mu, N)}: Y \times \mathscr{X} \to [0, 1]$, called the *Bayes adjoint* or *Bayes posterior of N with respect to the prior measure* μ. Let us assume that μ is a correct description of the probabilities of states of affairs in the world and that N correctly describes the likelihood of sets in Y given points of X. For y in Y and $A \subset X$, $P_{(\mu, N)}(y, A)$ is the conditional probability that the state of affairs in the world is ideally represented by a point in A, given that the punctual premise y is input. The probability measures $P_{(\mu, N)}(y, \cdot)$ are called the Bayesian *posterior probabilities* on X.

Now, via the usual operation of kernels on measures, $P_{(\mu, N)}$ defines the map $\mathscr{P}(Y) \to \mathscr{P}(X)$ given by $\lambda \mapsto \lambda P_{(\mu, N)}$, where $\lambda P_{(\mu, N)}(A) =_{\text{def}} \int_Y \lambda(dy) P_{(\mu, N)}(y, A)$ for A in $\mathscr{X}$. In this sense we can view $P_{(\mu, N)}$ as an inference map; i.e., we have

$$\Psi: \mathscr{P}(Y) \to \mathscr{P}(X)$$
$$\lambda \mapsto \lambda P_{(\mu, N)}.$$

DEFINITION 1. A *Bayesian probabilistic inference* is the probabilistic inference

$$\Psi: \lambda \mapsto \lambda P_{(\mu, N)}$$

for given X, Y, N, μ as above.

In other words, "Bayesian probabilistic inference" means that the inference is made exclusively on the basis of conditional probability in the form of the Bayesian posterior kernel. It will be useful to conceptualize this conditional probability mathematically as follows: Given spaces X and Y, a measure μ on X together with a kernel $N: X \times \mathscr{Y} \to [0, 1]$ gives rise

to a measure on $X \times Y$, denoted $\mu \otimes N$, defined by

$$\mu \otimes N(A \times B) =_{\text{def}} \int_A \mu(dx) N(x, B)$$

(for the sets $A \subset X$ and $B \subset Y$). Then $P_{(\mu, N)}(y, A)$ expresses the conditional probability of the set $A \subset X$ given the point y in Y with respect to this measure $\mu \otimes N$ on $X \times Y$. To make this completely precise, since the underlying measure $\mu \otimes N$ of the conditional probability is on $X \times Y$, we should express everything in terms of sets on $X \times Y$ and say that $P_{(\mu, N)}(y, A)$ is the conditional probability of the set $A \times Y \subset X \times Y$ given the set $X \times \{y\}$ in $X \times Y$. $P_{(\mu, N)}(y, A)$ may be expressed as the appropriate conditional expectation, or equivalently as a Radon–Nikodym derivative;

$$P_{(\mu, N)}(y, A) = \text{Prob}(A \mid y) = (\mu \otimes N)(A \times Y \mid X \times \{y\})$$

or

$$P_{(\mu, N)}(y, A) = \frac{d(\mu(1_A N))}{d(\mu N)}(y) \qquad \mu N\text{-a.e.}$$

PROPOSITION 2. $\mu N P_{(\mu, N)} = \mu$.

Proof. If σ is *any* measure on Y and K is *any* kernel from Y to X, it follows form the definitions of σP and $\sigma \otimes P$ that for a measurable set A in X,

$$\sigma P(A) = (\sigma \otimes P)(Y \times A).$$

In particular $(\mu N)P_{(\mu, N)}$ is the marginal measure of $(\mu N) \otimes P_{(\mu, N)}$ on X. But since N and $P_{(\mu, N)}$ are Bayes adjoints, the measures $\mu \otimes N$ and $\mu N \otimes P_{(\mu, N)}$ on $X \times Y$ are the same. Therefore $(\mu N)P_{(\mu, N)}$ is the marginal on X of $\mu \otimes N$, which is μ.

Note. For $\sigma \neq \mu$, the equation $\sigma N P_{(\mu, N)} = \sigma$ will not hold in general.

The type of Bayesian inference described in Definition 1 can be updated recursively in a natural way. Given X, Y, N, and μ_0, we get the Bayes posterior $P_{(\mu_0, N)}$ which gives the inference map $\lambda \mapsto \lambda P_{(\mu_0, N)}$ from $\mathscr{P}(Y)$ to $\mathscr{P}(X)$. To simplify notation, let us denote this map by P_0. We use the following time index convention: we will view μ_0 and the associated inference map P_0 as arising at time $t = 0$, but the argument λ to which P_0 is applied as arising at $t = 1$. For this reason it is appropriate to denote the argument of the map P_0 by λ_1, and to view the new measure $\lambda_1 P_0$ on X as

a new prior μ_1 which arises at $t = 1$ together with its associated inference map $P_1 = P_{(\mu_1, N)}$. We remind the reader that N is time invariant.

In this way, given a sequence of premise measures $\{\lambda_n\}$, there is generated a sequence of priors $\{\mu_n\}$ and the associated sequence of Bayesian posteriors, i.e., of inference maps $\{P_n\}$, where $P_n = P_{(\mu_n, N)}$. It is appropriate to think of the inference map P_n as the "learning strategy" prepared at time n to be applied to a new premise λ_{n+1} which will be acquired at time $n + 1$. Thus at each time n there arises a *pair* consisting of a new prior μ_n and its associated learning strategy P_n. The acquisition of the premise λ_{n+1} triggers the transition $(\mu_n, P_n) \mapsto (\mu_{n+1}, P_{n+1})$. This procedure is truly recursive, and we will call it *Bayesian recursive updating* or *Bayesian structural updating*.

As we mentioned in the Introduction, this Bayesian recursive inference is not the same as the classical "Bayesian statistical inference" as that term is used in the statistics literature. However, it is not hard to show that the recursive inference is a *generalization* of the classical; we will not pursue this here.

2. THE L^∞ WINDOW

Notation 3. Let U be a metric space. Let $A \subset U$, and let $\epsilon > 0$. We denote

$$A^\epsilon = \{u \in U: \text{distance}(u, A) < \epsilon\}.$$

With this we have:

DEFINITION 4. Let U be a metric space with its associated Borel measurable structure. Let μ_1, μ_2 be measures on U. Then the *Prokhorov distance* between μ_1 and μ_2, denoted by $\rho_{\text{Prok}}(\mu_1, \mu_2)$, is defined as

$$\rho_{\text{Prok}}(\mu_1, \mu_2) = \max(\epsilon_{12}, \epsilon_{21}),$$

where

$$\epsilon_{12} = \inf\{\epsilon: \mu_1(A) < \mu_2(A^\epsilon) + \epsilon \text{ for all } A \in \mathscr{U}\}$$

$$\epsilon_{21} = \inf\{\epsilon: \mu_2(A) < \mu_1(A^\epsilon) + \epsilon \text{ for all } A \in \mathscr{U}\}$$

We will call convergence with respect to the Prokhorov metric *Prokhorov convergence* or ρ_{Prok}-*convergence*.

THEOREM 5 (Prokhorov [3]; see also [2]).

(i) *If U is a metric space then the Prokhorov metric topology is the weak topology on $\mathscr{P}(U)$.*

(ii) *If U is a complete separable metric space, then $\mathscr{P}(U)$, with the Prokhorov metric topology (i.e., the weak topology), is also a complete separable metric space.*

Let U be a complete separable metric space with a metric d; U is a measurable space with Borel algebra $\mathscr{U}$ associated to the metric topology. Let $\mathscr{P}(U)$ denote the set of probability measures on U. For a given measure ν on U, let

$$L^\infty(U, \nu) = \{\text{all measurable, } \nu\text{-essentially bounded functions on } U\},$$

where recall that a function f on U is ν-essentially bounded if there exists a real number b such that $\nu(\{u \in U\colon f(y) > b\}) = 0$. In that case, the essential sup norm $\|f\|_\nu$ is the infimum of the set of such b's. We denote by ρ_ν the metric on $L^\infty(U, \nu)$ associated to this norm. Let

$$\mathscr{B}_\nu(U) = \left\{\sigma \in \mathscr{P}(U)\colon \sigma \ll \nu \quad \text{and} \quad \frac{d\sigma}{d\nu} \in L^\infty(U, \nu)\right\},$$

where $\ll$ denotes absolute continuity. $\mathscr{B}_\nu(U)$ is a topological space for the weak topology induced from $\mathscr{P}(U)$.

Define the function

$$\Phi\colon L^\infty(U, \nu) \to \mathscr{B}_\nu(U)$$

by

$$\Phi(g) = g\nu.$$

With this we have the following theorem:

THEOREM 6.

(i) Φ *is a continuous bijective function onto its image* $\mathscr{B}_\nu(U)$.

(ii) Φ *is non-expansive*: $\rho_\nu(f, g) < \epsilon$ *implies* $\rho_{\text{Prok}}(\Phi(f), \Phi(g)) < \epsilon$.

Proof. (i) That Φ is a one-to-one correspondence follows immediately from the definitions. To show Φ is continuous, we will show that for any sequence $\{\sigma_n\}$ and any element σ in $\mathscr{B}_\nu(U)$,

$$\frac{d\sigma_n}{d\nu} \overset{\rho_\nu}{\to} \frac{d\sigma}{d\nu} \quad \Rightarrow \quad \sigma_n \overset{w}{\to} \sigma,$$

whre $\overset{\rho_\nu}{\to}$ and $\overset{w}{\to}$ denote conergence in the ρ_ν-metric and weak topologies, respectively. Suppose $d\sigma_n/d\nu \overset{\rho_\nu}{\to} d\sigma/d\nu$. Then in particular $d\sigma_n/d\nu(u) \to d\sigma/d\nu(u)$ as $n \to \infty$ for ν-a.e. $u \in U$. Let f be an arbitrary bounded

continuous function on U. Then as $n \to \infty$,

$$\int_U \sigma_n(du) f(u) = \int_U \nu(du) \frac{d\sigma_n}{d\nu}(u) f(u) \to$$

$$\int_U \nu(d\nu) \frac{d\sigma}{d\nu}(u) f(u) = \int_U \sigma(du) f(u).$$

Therefore, $\sigma_n \overset{w}{\to} \sigma$.

(ii) Suppose f, g are in $L^\infty(U, \nu)$, with $\rho_\nu(f, g) < \epsilon$. Let $\sigma = f\nu$ and $\tau = g\nu$, so that we may write

$$\rho_\nu\left(\frac{d\sigma}{d\nu}, \frac{d\tau}{d\nu}\right) < \epsilon,$$

in other words

$$\left|\frac{d\sigma}{d\nu}(u) - \frac{d\tau}{d\nu}(u)\right| < \epsilon \qquad \text{for } \nu\text{-a.e. } u \in U.$$

Choose $B \in \mathscr{U}$. We have

$$\begin{aligned}
\sigma(B) &= \int_B \sigma(du) \\
&= \int_B \nu(du) \frac{d\sigma}{d\nu}(u) \\
&< \int_B \nu(du) \left[\frac{d\tau}{d\nu}(u) + \epsilon\right] \\
&= \int_B \nu(du) \frac{d\tau}{d\nu}(u) + \int_B \nu(du)\epsilon \\
&\leq \tau(B) + \epsilon.
\end{aligned}$$

A fortiori,

$$\sigma(B) < \tau(B^\epsilon) + \epsilon.$$

Reversing the roles of σ and τ, we obtain similarly

$$\tau(B) < \sigma_n(B^\epsilon) + \epsilon.$$

These last two inequalities imply that

$$\rho_{\text{Prok}}(\sigma, \tau) < \epsilon,$$

and the proof is concluded.

THEOREM 7. *Let $(U, \mathscr{U})$ and $(V, \mathscr{V})$ be two measurable spaces. Let K be any Markovian kernel from U to V, and let μ and ν be probability measures on U; then*

$$\mu \ll \nu \quad \Rightarrow \quad \mu K \ll \nu K.$$

Proof. Let B be an arbitrary subset of V, such that $\nu K(B) = 0$; we show that $\mu K(B) = 0$. We have

$$\nu K(B) = \int_U \nu(du) K(u, B) = 0,$$

and therefore, since $K(u, B) \geq 0$, $K(u, B) = 0$ for ν-a.e. $u \in U$.

So, letting $A = \{u \in U\colon K(u, B) > 0\}$, we have that $\nu(A) = 0$, and hence $\mu(A) = 0$ by hypothesis. Therefore, since $K(u, B) \leq 1$ and $A^c = \{u \in U\colon K(u, B) = 0\}$,

$$\begin{aligned} \mu K(B) &= \int_A \mu(du) K(u, B) + \int_{A^c} \mu(du) K(u, B) \\ \mu K(B) &\leq \mu(A) + 0 \\ &= 0. \end{aligned}$$

We now apply Theorem 7 to Bayesian recursive updating using the notation and terminology introduced at the end of Section 1.

COROLLARY 8. *Let μ_0 be a fixed prior probability measure on X, let N be a time invariant noise kernel from X to Y, let $\{\lambda_n\}$ be a sequence of premise measures on Y, and let $\{\mu_n\}$ be the sequence of percepts generated thereby. Then for each $n = 1, 2, \ldots,$*

$$\lambda_n \ll \mu_{n-1} N \quad \Rightarrow \quad \mu_n \ll \mu_{n-1}.$$

Proof. By definition $\lambda_n P_{n-1} = \mu_n$, and by Proposition 2 $\mu_{n-1} N P_{n-1} = \mu_{n-1}$. Therefore, applying Theorem 7 with $\nu = \lambda_n$, $\mu = \mu_{n-1} N$, and $K = P_{n-1}$ (which is a Markovian kernel for each $n \geq 0$), we obtain the desired result.

COROLLARY 9. *For each $n \geq 0$,*

$$\mu_n \ll \mu_{n-1} \quad \Rightarrow \quad \mu_n N \ll \mu_{n-1} N.$$

Proof. Since N is a Markovian kernel, the result follows immediately from Theorem 7.

Notation. Let λ and σ be probability measures on a measurable space U. For convenience, put

$$\rho_\nu(\lambda, \sigma) = \rho_\nu\left(\frac{d\lambda}{d\nu}, \frac{d\sigma}{d\nu}\right) = \left\| \frac{d\lambda}{d\nu} - \frac{d\sigma}{d\nu} \right\|_\nu.$$

Thus, the notation "$\rho_\nu(\lambda, \sigma)$" means that we transport the metric ρ_ν to $\mathscr{B}_\nu(U)$ via Φ as in Theorem 6.

Let μ and ν be probability measures on a measurable space X. If $\mu \ll \nu$, let us adopt the convention that $d\mu/d\nu = 0$ outside the support "supp(ν)" of ν; in other words, we assume chosen a version of the Radon–Nikodym derivative with this property. Denote

$$1_\nu = 1_{\text{supp}(\nu)}.$$

Then, by our convention, we have

$$\frac{d\mu}{d\nu} = \frac{d\mu}{d\nu} 1_\nu.$$

THEOREM 10. *Let $(U, \mathscr{U})$ and $(V, \mathscr{V})$ be two measurable spaces. Let K be any Markovian kernel from U to V, and let μ and ν be probability measures on U such that $\mu \ll \nu$. Then for any $\epsilon > 0$,*

$$\rho_\nu(\mu, \nu) < \epsilon \Rightarrow \rho_{\nu K}(\mu K, \nu K) < \epsilon.$$

Proof. Let $\epsilon > 0$ be fixed, and $\rho_\nu(\mu, \nu) < \epsilon$; then this implies that

$$\left\| \frac{d\mu}{d\nu} - 1_\nu \right\|_\nu < \epsilon \qquad \text{for } \nu\text{-a.e. } u \in U. \tag{1}$$

Let B be an arbitrary subset of V such that $\nu K(B) \neq 0$. Then

$$\mu K(B) = \int_U \mu(du) K(u, B)$$

$$= \int_U \nu(du) \frac{d\mu}{d\nu}(u) K(u, B).$$

Therefore, in view of (1),

$$\mu K(B) < \int_U \nu(du) K(u, B)(1 + \epsilon) = \nu K(B)(1 + \epsilon)$$

and

$$\mu K(B) > \int_U \nu(du) K(u, B)(1 - \epsilon) = \nu K(B)(1 - \epsilon),$$

i.e.,

$$\nu K(B)(1-\epsilon) < \mu K(B) < \nu K(B)(1+\epsilon).$$

Therefore, for any subset B of U, such that $\nu K(B) \neq 0$, we have

$$(1-\epsilon) < \frac{\mu K(B)}{\nu K(B)} < (1+\epsilon). \tag{2}$$

Since $\nu \ll \mu$, by Theorem 7 we know that $\nu K \ll \mu K$; hence $d\mu/d\nu$ makes sense and (2) implies

$$\left\| \frac{d_{\mu K}}{d_{\nu K}} - 1_{\nu K} \right\|_{\nu K} < \epsilon,$$

in other words

$$\rho_{\nu K}(\mu K, \nu K) < \epsilon.$$

We now apply Theorem 10 to Bayesian recursive updating; as always the notation is as introduced at the end of Section 1.

COROLLARY 11. *For any $n \geq 0$ and any $\epsilon > 0$, if $\lambda_{n+1} \ll \mu_n N$, then*

$$\rho_{\mu_n N}(\lambda_{n+1}, \mu_n N) < \epsilon \quad \Rightarrow \quad \rho_{\mu_n}(\mu_{n+1}, \mu_n) < \epsilon.$$

Proof. Since for any n P_n is a Markovian kernel from Y to X, and since $\lambda_{n+1} \ll \mu_n N$, we get $\lambda_{n+1} P_n \ll \mu_n N P_n$ by Theorem 7. However, $\lambda_{n+1} P_n = \mu_{n+1}$ and $\mu_n N P_n = \mu_n$, so the result follows immediately from Theorem 10.

COROLLARY 12. *For any $n \geq 0$ and any $\epsilon > 0$, if $\mu_{n+1} \ll \mu_n$, then*

$$\rho_{\mu_n}(\mu_{n+1}, \mu_n) < \epsilon \quad \Rightarrow \quad \rho_{\mu_n N}(\mu_{n+1} N, \mu_n N) < \epsilon.$$

Proof. Since N is a Markovian kernel from X to Y, and from Corollary 9, since $\mu_{n+1} \ll \mu_n$ implies that $\mu_{n+1} N \ll \mu_n N$, the result follows immediately from Theorem 10.

THEOREM 13. *Let μ_0 be given. Suppose that $\lambda_k \ll \mu_{k-1} N$ for $k = 1, 2, \ldots, n$ and suppose that $\rho_{\mu_{k-1} N}(\lambda_k, \mu_{k-1} N) \leq \epsilon_{k-1}$ for $\epsilon_{k-1} > 0$, $k = 1, 2, \ldots, n$. Then, given $\kappa > 0$, if $\lambda_{n+1} \ll \mu_n N$ and if*

$$\rho_{\mu_n N}(\lambda_{n+1}, \mu_n N) < \frac{\kappa}{(1+\epsilon_0)(1+\epsilon_2)\cdots(1+\epsilon_{n-1})},$$

then

$$\rho_{\mu_0}(\mu_{n+1}, \mu_n) < \kappa.$$

Proof. Note first that since $\lambda_k \ll \mu_{k-1}N$ for $k = 1, 2, \ldots, n$, by successively applying Corollary 8 and Corollary 9 for $m = 0, 1, \ldots, n$, we have that $\mu_n \ll \mu_m$ and $\mu_n N \ll \mu_m N$. We will use these facts freely below

$$\begin{aligned}\left\| \frac{d\mu_{n+1}}{d\mu_0} - \frac{d\mu_n}{d\mu_0} \right\|_{\mu_0} &= \left\| \frac{d\mu_{n+1}}{d\mu_n}\frac{d\mu_n}{d\mu_0} - 1_{\mu_n}\frac{d\mu_n}{d\mu_0} \right\|_{\mu_0} \\ &\le \left\| \frac{d\mu_{n+1}}{d\mu_n} - 1_{\mu_n} \right\|_{\mu_0} \left\| \frac{d\mu_n}{d\mu_0} \right\|_{\mu_0} \\ &\le \left\| \frac{d\mu_{n+1}}{d\mu_n} - 1_{\mu_n} \right\|_{\mu_0} \left\| \frac{d\mu_n}{d\mu_{n-1}} \right\|_{\mu_0} \left\| \frac{d\mu_{n-1}}{d\mu_{n-2}} \right\|_{\mu_0} \cdots \left\| \frac{d\mu_1}{d\mu_0} \right\|_{\mu_0}.\end{aligned}$$

So we have that

$$\begin{aligned}\rho_{\mu_0}(\mu_{n+1}, \mu_n) &= \left\| \frac{d\mu_{n+1}}{d\mu_0} - \frac{d\mu_n}{d\mu_0} \right\|_{\mu_0} \\ &\le \left\| \frac{d\mu_{n+1}}{d\mu_n} - 1_{\mu_n} \right\|_{\mu_0} \left\| \frac{d\mu_n}{d\mu_{n-1}} \right\|_{\mu_0} \left\| \frac{d\mu_{n-1}}{d\mu_{n-2}} \right\|_{\mu_0} \cdots \left\| \frac{d\mu_1}{d\mu_0} \right\|_{\mu_0}.\end{aligned} \tag{1}$$

Now for $\delta > 0$, if

$$\rho_{\mu_n N}(\lambda_{n+1}, \mu_n N) = \left\| \frac{d\lambda_{n+1}}{d\mu_n N} - 1_{\mu_n N} \right\|_{\mu_n N} < \delta,$$

then, by Corollary 11, we have that

$$\rho_{\mu_n}(\mu_{n+1}, \mu_n) = \left\| \frac{d\mu_{n+1}}{d\mu_n} - 1_{\mu_n} \right\|_{\mu_n} < \delta.$$

But since $|d\mu_{n+1}/d\mu_n - 1_{\mu_n}| = 0$ outside $\operatorname{supp}(\mu_n)$, we may write

$$\left\| \frac{d\mu_{n+1}}{d\mu_n} - 1_{\mu_n} \right\|_{\mu_n} = \left\| \frac{d\mu_{n+1}}{d\mu_n} - 1_{\mu_n} \right\|_{\mu_0}.$$

Hence

$$\rho_{\mu_n N}(\lambda_{n+1}, \mu_n N) < \delta \quad \Rightarrow \quad \left\| \frac{d\mu_{n+1}}{d\mu_n} - 1_{\mu_n} \right\|_{\mu_0} < \delta.$$

With this, and in view of (1), if $\rho_{\mu_n N}(\lambda_{n+1}, \mu_n N) < \delta$, then

$$\rho_{\mu_0}(\mu_{n+1}, \mu_n) < \delta \left\| \frac{d\mu_n}{d\mu_{n-1}} \right\|_{\mu_0} \left\| \frac{d\mu_{n-1}}{d\mu_{n-2}} \right\|_{\mu_0} \cdots \left\| \frac{d\mu_1}{d\mu_0} \right\|_{\mu_0}. \tag{2}$$

But by assumption, for $k = 1, 2, \ldots, n$,

$$\rho_{\mu_{k-1} N}(\lambda_k, \mu_{k-1} N) \leq \epsilon_{k-1},$$

so by Corollary 8, $\| d\mu_k / d\mu_{k-1} - 1_{\mu_{k-1}} \|_{\mu_{k-1}} \leq \epsilon_{k-1}$. Hence, since we can write

$$\left\| \frac{d\mu_k}{d\mu_{k-1}} - 1_{\mu_{k-1}} \right\|_{\mu_{k-1}} = \left\| \frac{d\mu_k}{d\mu_{k-1}} - 1_{\mu_{k-1}} \right\|_{\mu_0},$$

we have

$$\left\| \frac{d\mu_k}{d\mu_{k-1}} - 1_{\mu_{k-1}} \right\|_{\mu_0} \leq \epsilon_{k-1},$$

whence

$$\left\| \frac{d\mu_k}{d\mu_{k-1}} \right\|_{\mu_0} \leq 1 + \epsilon_{k-1}$$

for $k = 1, 2, \ldots, n$. So in view of (2), if $\rho_{\mu_n N}(\lambda_{n+1}, \mu_n N) < \delta$, then

$$\rho_{\mu_0}(\mu_{n+1}, \mu_n) < \delta(1 + \epsilon_{n-1}) \cdots (1 + \epsilon_1)(1 + \epsilon_0).$$

Therefore, if

$$\rho_{\mu_n N}(\lambda_{n+1}, \mu_n N) < \frac{\kappa}{(1 + \epsilon_{n-1}) \cdots (1 + \epsilon_1)(1 + \epsilon_0)},$$

then

$$\rho_{\mu_0}(\mu_{n+1}, \mu_n) < \kappa.$$

THEOREM 14. *With the notation of Theorem* 13, *suppose that* $\{\lambda_n\}_{n \geq 1}$ *is a sequence of probability measures on Y which recursively generate the sequence* $\{\mu_n\}_{n \geq 1}$ *on X via* $\mu_{n+1} = \lambda_{n+1} P_{(\mu_n, N)}$. *Suppose we are given a decreasing sequence* $\kappa_1 > \kappa_2 > \ldots$ *of positive numbers such that* $\sum_{i=1}^{\infty} \kappa_i$ *converges.*

Let $\epsilon_k = \rho_{\mu_k N}(\lambda_{k+1}, \mu_k N)$ *and suppose for each* n

$$\epsilon_n \leq \frac{\kappa_n}{(1+\epsilon_0)(1+\epsilon_1)\cdots(1+\epsilon_{n-1})}; \tag{1}$$

then the sequence $\{\mu_n\}$ *converges weakly in* $\mathscr{P}(X)$, *with* $\rho_{\text{Prok}}(\mu_{n+1}, \mu_n) < \epsilon$.

Proof. By Theorem 13, (1) above implies that for all n,

$$\rho_{\mu_0}(\mu_{n+1}, \mu_n) = \left\| \frac{d\mu_{n+1}}{d\mu_0} - \frac{d\mu_n}{d\mu_0} \right\|_{\mu_0} < \kappa_n.$$

Then since $\sum_{i=1}^{\infty} \kappa_i$ converges, this means that $\{d\mu_n/d\mu_0\}_n$ is a Cauchy sequence in $L^\infty(X, \mu_0)$ so it converges since L^∞ is complete. Then by Theorem 6 (i), $\{\mu_n\}$ converges weakly in $\mathscr{B}_{\mu_0}(X) \subset \mathscr{P}(X)$. By Theorem 6 (ii), since $\rho_{\mu_0}(\mu_{n+1}, \mu_n) < \kappa_n$, we also have $\rho_{\text{Prok}}(\mu_{n+1}, \mu_n) < \kappa_n$.

Note. The condition (1) on the λ's may be verified recursively.

Theorem 14 describes concretely how a system can formulate a *directed convergence strategy*. The goal of the system is to acquire a "stable percept," i.e., a weakly convergent sequence of conclusion measures μ_n. Suppose that the system has the capability to accept or reject input measures λ. To accept a λ at the $(n+1)$st stage of the process means to take $\lambda = \lambda_{n+1}$, and use it to acquire μ_{n+1} in the form $\mu_{n+1} = \lambda_{n+1} P_{(\mu_n, N)}$. To reject λ means to not use λ for purposes of updating the conclusion. Suppose moreover that the system has the capability to measure, for each n, the $L^\infty(Y, \mu_n N)$-metric distance of input measures λ to $\mu_n N$, i.e., to measure $\rho_{\mu_n N}(\lambda, \mu_n N)$. Suppose a sequence of numbers κ_n such that $\sum_{i=1}^{\infty} \kappa_i$ converges is given. Suppose $\lambda_1, \ldots, \lambda_n$ have already been chosen so that, if ϵ_k denotes $\rho_{\mu_k N}(\lambda_{k+1}, \mu_k N)$, then

$$\epsilon_k \leq \frac{\kappa_k}{(1+\epsilon_0)(1+\epsilon_1)\cdots(1+\epsilon_{k-1})},$$

for $k = 1, \ldots, n-1$. Then the system can wait for an input measure λ such that

$$\rho_{\mu_n N}(\lambda, \mu_n N) \leq \frac{\kappa_n}{(1+\epsilon_0)(1+\epsilon_1)\cdots(1+\epsilon_{n-1})}.$$

When such a λ is acquired, it will be accepted as λ_{n+1}. According to Theorem 14, the sequence of conclusions μ_n corresponding to the sequence of inputs λ_n selected in this manner will converge weakly in $\mathscr{P}(X)$.

Remark. The choice of the sequence of κ_n's corresponds to the system's degree of confidence at each stage about how close the current conclusion

μ_n is to a "correct" conclusion μ; by definition, a correct conclusion is a weak limit in $\mathscr{P}(X)$ of a recursively generated sequence of μ_n's. The greater the confidence, the smaller κ_n is, i.e., the more restrictive is the condition on an incoming input measure λ for it to be accepted as λ_{n+1}. If such a λ is forthcoming, we say that the system's *belief* (about the closeness of the current conclusion μ_n to a correct conclusion μ) is *confirmed*. For this purpose the environment must cooperate with the system to provide enough belief-confirming inputs λ so that one of them will be acquired at the nth stage after not too long a wait. The idea is that the successive acquisition of belief-confirming inputs at each stage n (so that a convergent sequence μ_n is actually generated) occurs with probability 0, unless there is an object in the environment which transduces the successive inputs. Note that if a belief-confirming premise is *not* acquired in a reasonable length of time at the nth stage, then the system's confidence in μ_n may justifiably decrease; in this case it will be reasonable for the system to modify κ_n, replacing it with a larger number, to enhance the possibility of acquiring an acceptable λ to be used for λ_{n+1}. This introduces flexibility into the system, so that the "direction" of the quest for stable percepts (i.e., for convergent sequences of conclusions) is responsive to the actual environmental conditions. For more details on directed convergence, the reader is referred to [1].

Consider a directed convergence procedure for a Cauchy sequence κ_n as in Theorem 14, i.e., consider a sequence of input measures λ_n for which the hypotheses of Theorem 14 are satisfied. Then, by that theorem, the corresponding sequence of recursively generated conclusions μ_n converges weakly in $\mathscr{P}(X)$. This paper concludes with several results which enable us to state that *in directed convergence, when the sequence* $\{\mu_n\}$ *converges, the sequence of inputs* λ_n *must also converge weakly (in* $\mathscr{P}(Y)$*), and at a rate comparable to that of* μ_n *in* $\mathscr{P}(X)$. Recall that directed convergence proceeds by comparing λ_{n+1} to $\mu_n N$, and *not* to λ_n. Hence that the $\{\lambda_n\}$ is a Cauchy sequence is not transparent from the fact that $\{\mu_n\}$ is a Cauchy sequence—naive triangle inequality arguments adduced for this purpose seem to blow up as $n \to \infty$.

THEOREM 15. *For every n and for every* $\epsilon > 0$, *if* $\mu_n \ll \mu_0$, *then*

$$\rho_{\mu_0}(\mu_n, \mu_{n-1}) < \epsilon \quad \Rightarrow \quad \rho_{\mu_0 N}(\mu_n N, \mu_{n-1} N) < \epsilon.$$

Proof. Let $\epsilon > 0$ be fixed. Let B be an arbitrary subset of Y such that $\mu_0 N(B) > 0$. Then

$$\mu_n N(B) = \int_X \mu_n(dx) N(x, B) = \int_X \mu_0(dx) \frac{d\mu_n}{d\mu_0}(x) N(x, B).$$

Now, since $\rho_{\mu_0}(\mu_n, \mu_{n-1}) < \epsilon$, this means that

$$\left\| \frac{d\mu_n}{d\mu_0} - \frac{d\mu_{n-1}}{d\mu_0} \right\|_{\mu_0} < \epsilon,$$

so that we have for μ_0-a.e. $x \in X$

$$\frac{d\mu_n}{d\mu_0}(x) < \frac{d\mu_{n-1}}{d\mu_0}(x) + \epsilon$$

and

$$\frac{d\mu_n}{d\mu_0}(x) > \frac{d\mu_{n-1}}{d\mu_0}(x) - \epsilon$$

Therefore, we have that

$$\begin{aligned} \mu_n N(B) &< \int_X \mu_0(dx) \left(\frac{d\mu_{n-1}}{d\mu_0}(x) + \epsilon \right) N(x, B) \\ &= \mu_{n-1} N(B) + \epsilon \mu_0 N(B), \end{aligned}$$

and similarly

$$\begin{aligned} \mu_n N(B) &> \int_X \mu_0(dx) \left(\frac{d\mu_{n-1}}{d\mu_0}(x) - \epsilon \right) N(x, B) \\ &= \mu_{n-1} N(B) - \epsilon \mu_0 N(B), \end{aligned}$$

so that

$$\mu_{n-1} N(B) - \epsilon \mu_0 N(B) < \mu_n N(B) < \mu_{n-1} N(B) + \epsilon \mu_0 N(B).$$

Therefore, for any subset B of Y such that $\mu_0 N(B) \neq 0$,

$$\frac{\mu_{n-1} N(B)}{\mu_0 N(B)} - \epsilon < \frac{\mu_n N(B)}{\mu_0 N(B)} < \frac{\mu_{n-1} N(B)}{\mu_0 N(B)} + \epsilon. \tag{1}$$

However, if we iteratively apply Corollary 9 we have that $\mu_k \ll \mu_0 \Rightarrow \mu_k N \ll \mu_0 N$. Hence $d\mu_k N / d\mu_0 N$ makes sense and (1) shows that

$$\left\| \frac{d\mu_n N}{d\mu_0 N} - \frac{d\mu_{n-1} N}{d\mu_0 N} \right\|_{\mu_0 N} < \epsilon,$$

so that

$$\rho_{\mu_0 N}(\mu_n N, \mu_{n-1} N) < \epsilon.$$

THEOREM 16. *Suppose we have a sequence* $\mu_0, \mu_1, \ldots, \mu_n, \ldots$ *which is obtained recursively from a sequence of premises* $\lambda_0, \lambda_1, \ldots, \lambda_n, \ldots$ *Let* $\{\kappa_n\}$ *be a decreasing sequence of positive numbers such that* $\sum \kappa_n$ *converges. Let* $\epsilon_n = \rho_{\mu_n N}(\lambda_{n+1}, \mu_n N)$, *and assume that*

$$\epsilon_n \leq \frac{\kappa_n}{(1+\epsilon_0)(1+\epsilon_1)\cdots(1+\epsilon_{n-1})}. \tag{1}$$

Then

$$\rho_{\mu_0 N}(\lambda_{n+1}, \lambda_n) < \kappa_{n-1}(3+2K),$$

where

$$K = \sum_{n=0}^{\infty} \kappa_n.$$

Proof. In view of (1) and the fact that $\epsilon_n = \rho_{\mu_n N}(\lambda_{n+1}, \mu_n N)$, Theorem 13 guarantees that

$$\kappa_n > \rho_{\mu_0}(\mu_{n+1}, \mu_n),$$

which then guarantees (by Theorem 15) that

$$\kappa_n > \rho_{\mu_0 N}(\mu_{n+1} N, \mu_n N). \tag{2}$$

This means

$$\left\| \frac{d\mu_i N}{d\mu_0 N} - \frac{d\mu_{i-1} N}{d\mu_0 N} \right\|_{\mu_0 N} < \kappa_i$$

for $i = 1, \ldots, n$, so that

$$\left\| \frac{d\mu_i N}{d\mu_0 N} \right\|_{\mu_0 N} < \kappa_i + \left\| \frac{d\mu_{i-1} N}{d\mu_0 N} \right\|_{\mu_0 N},$$

for $i = 1, \ldots, n$. Therefore, applying this iteratively as i decreases from n to 1, we obtain

$$\left\| \frac{d\mu_n N}{d\mu_0 N} \right\|_{\mu_0 N} < \kappa_n + \kappa_{n-1} + \cdots + \kappa_0 + 1. \tag{3}$$

From the triangle inequality we have

$$\rho_{\mu_0 N}(\lambda_{n+1}, \lambda_n) < \rho_{\mu_0 N}(\lambda_{n+1}, \mu_n N) + \rho_{\mu_0 N}(\mu_n N, \mu_{n-1} N) + \rho_{\mu_0 N}(\mu_{n-1} N, \lambda_n). \tag{4}$$

However,

$$\rho_{\mu_0 N}(\lambda_i, \mu_{i-1}N) = \rho_{\mu_{i-1}N}(\lambda_i, \mu_{i-1}N)\left\|\frac{d\mu_{i-1}N}{d\mu_0 N}\right\|_{\mu_0 N}, \tag{5}$$

for every i. Hence (2)–(5) yield

$$\begin{aligned}
\rho_{\mu_0 N}(\lambda_{n+1}, \lambda_n) &< \rho_{\mu_n N}(\lambda_{n+1}, \mu_n N)\left\|\frac{d\mu_n N}{d\mu_0 N}\right\|_{\mu_0 N} + \kappa_{n-1} \\
&\quad + \rho_{\mu_{n-1}N}(\lambda_n, \mu_{n-1}N)\left\|\frac{d\mu_{n-1}N}{d\mu_0 N}\right\|_{\mu_0 N} \\
&< \epsilon_n\left(\sum_{i=0}^{n}\kappa_i + 1\right) + \kappa_{n-1} + \epsilon_{n-1}\left(\sum_{i=0}^{n-1}\kappa_i + 1\right) \\
&< \epsilon_n K + \epsilon_n + \kappa_{n-1} + \epsilon_{n-1}K + \epsilon_{n-1} \\
&< \kappa_n K + \kappa_n + \kappa_{n-1} + \kappa_{n-1}K + \kappa_{n-1} \\
&< \kappa_{n-1}K + \kappa_{n-1} + \kappa_{n-1} + \kappa_{n-1}K + \kappa_{n-1} \\
&= \kappa_{n-1}(3 + 2K).
\end{aligned}$$

COROLLARY 17. *With the hypotheses of Theorem* 16 $\{\lambda_n\}$ *is Cauchy for* ρ_{Prok}, *with* $\rho_{\text{Prok}}(\lambda_{n+1}, \lambda_n) < \kappa_{n-1}(3 + 2K)$.

Proof. Since the same conditions hold as in Theorem 16, this means that

$$\rho_{\mu_0 N}(\lambda_{n+1}, \lambda_n) < \kappa_{n-1}(3 + 2K).$$

By Theorem 6(ii), this implies $\rho_{\text{Prok}}(\lambda_{n+1}, \lambda_n) < \kappa_{n-1}(3 + 2K)$.

Remark. In this case, the *rate* of convergence of the λ's is "comparable" to that of the μ's, in that the Cauchy distances differ by the constant factor $(3 + 2K)$ (and they are offset by one index).

ACKNOWLEDGMENTS

We thank Don Hoffman, Chetan Prakash, Scott Richman, Brian Skyrms, and Weian Zheng for helpful discussions. We are grateful to the Institute for Mathematical Behavioral Sciences and the National Science Foundation for their support.

REFERENCES

1. B. Bennett and R. Cohen, Directed convergence in stable percept acquisition, *J. Math. Psychol.*, to appear.
2. P. Billingsley, "Convergence of Probability Measures," Wiley, New York, 1968.
3. Y. Prokhorov, Convergence of random processes and limit theorems in probability theory, *Theory Probab. Appl.* **1** (1956), 157–214.