Communications in Information and Systems
Vol. 2, No. 2, pp. 167-182, December 2002

004

L^P METRIC CRITERIA FOR DIRECTED CONVERGENCE[*]

J. STRASSER MCINTOSH[†] AND BRUCE M. BENNETT[†]

Abstract. We consider a discrete-time probabilistic inference system where the conclusions and inputs are probability measures on measurable spaces $X, \mathfrak{X}$ and $Y, \mathcal{Y}$ respectively. We are given a kernel N from X to Y ($N : X \times \mathcal{Y} \to [0,1]$), which represents the way information about states of affairs (represented by X) is transmitted to a receptor (represented by Y). Suppose that at time $t = i$ the conclusion is given by a probability measure μ_i on X. Then given any input measure λ_i on Y we update μ_i, replacing it by $\mu_{i+1} = \lambda_i P_{\mu_i,N}$, where $P_{\mu_i,N}$ is the Bayes adjoint kernel for N and the 'prior' μ_i. In this way the conclusions evolve by conditional probability given a sequence of input measures. This is in contrast to classical Bayesian inference where the inputs are points of Y, and the conclusions are updated by conditioning a fixed Bayes posterior (defined with respect to a fixed prior) on the sequence of point inputs. In our case, as in classical Bayesian inference, the object is to obtain a (weakly) convergent sequence of conclusion measures μ_i. But in our case we have available a method called *directed convergence strategy* to facilitate convergence: metric criteria are employed to accept or reject input measures based on the degree of belief in the current conclusion. In this paper we develop sufficient conditions to execute this strategy using L^p metrics (after representing measures as their Radon-Nikodym derivatives). This work generalizes L^∞ criteria for directed convergence presented in [4].

Keywords. directed convergence, Bayesian recursive inference, weak convergence, L^p criteria
Primary. 60B10; **Secondary.** 62F15

1. Introduction.

Let $X, \mathfrak{X}$ and $Y, \mathcal{Y}$ be measurable spaces, N a kernel from X to Y ($N : X \times \mathcal{Y} \to [0,1]$), and μ_0 an initial probability measure on X. X is a configuration space for some class of states of some system of interest (probability measures on X are probabilistic models of those states). Y is a receptor for information about X, e.g., Y is a sensory receptor. N represents the channel which transmits information about the states of affairs to the receptor Y: If a state corresponding to a probability measure μ on X is transduced, there results the probability measure μN on Y. In the literature, N is sometimes called a 'likelihood kernel'.

In classical Bayesian inference, it is assumed that there is a state corresponding to μ which is unchanging over time, and that receptor information is acquired in the form of a sequence of points $\{y_n\}$ of Y whose distribution is μN. The object is then to infer μ from the y_n: With each successive y_n we obtain an updated measure

$$\mu_n = P_{(\mu_0,N)}(y_1, ..., y_n).$$

Here $P_{(\mu_0,N)}$ denotes the Bayes posterior kernel which depends only on μ_0 and N, so it is fixed throughout the procedure. Ideally the sequence μ_n will converge to

[*] Invited paper. Received on July 4, 2002; accepted for publication on August 19, 2002. Research supported in part by UNCF Special Programs and the Institute for Mathematical Behavioral Sciences at the University of California, Irvine.

[†] 103 Multipurpose Science and Technology Bldg., University of California Irvine, Department of Mathematics, Irvine, CA 92697, USA, E-mail: bbennett@math.uci.edu, strasser@math.uci.edu

μ in the weak topology; this is the problem of "Bayes consistency" which has been much discussed in the literature (Diaconis-Freedman 1986 [7]). It is important to note that in this classical Bayesian situation, the measure μ is supposed to represent *all* the available information about the state, its *overall* statistics. Moreover, *all* the available input data $\{y_n\}$ is to be used to infer μ; in particular there is no option to reject certain y_n's on the grounds that they are 'outliers,' since it is assumed that the whole sequence of y_n's instantiates μN.

In this study, however, we consider the case where *the inputs are probability measures on Y.* (For instance, our inference system may be part of a hierarchy in which the input measures are inferred at a lower level.) In this setting the updating of conclusions proceeds as follows. Suppose that at time $t = i$ the conclusion is given by a probability measure μ_i on X. Then, given any input measure λ_i on Y we update μ_i, replacing it by $\mu_{i+1} = \lambda_i P_{\mu_i,N}$, where $P_{\mu_i,N}$ is the Bayes adjoint kernel for N and the 'prior' μ_i. In this way the conclusions evolve by conditional probability given a sequence of input measures. This is called *structural probabilistic inference.*

Here we no longer assume that the inputs collectively instantiate some unique state of affairs to be inferred. Instead we adopt the more flexible viewpoint that any subsequence of input measures which gives rise to a (weakly) convergent sequence of conclusion measures is "valid," in the sense that it gives information about some stable *feature* in the environment. From this point of view the system, in its quest for convergent sequences of conclusions, should ideally have the option to select or reject inputs, i.e., to decide whether a given input measure should be used for updating the current conclusion, or should be ignored. In this paper we discuss how metric criteria can be effectively used for input selection, in order to achieve convergence of the conclusion sequence. The idea is that *at time $t = i$, λ should be close to $\mu_i N$ in order to accept it as the λ_i for purposes of updating μ_i to μ_{i+1}.* One can interpreted the minimum distance between λ and $\mu_i N$ in order for λ to be accepted, as indicative of the "degree of belief" in the current conclusion μ_i: the greater that degree of belief, the closer we will insist that λ is to $\mu_i N$ in order to accept it. This is based on the fact that if μ_i represented the actual state of affairs, then the input measure transduced at the receptor Y would be $\lambda = \mu_i N$. This is called *directed convergence strategy.* Of course, the environment must cooperate by making inputs available which satisfy the criteria. For this reason, if in a given observational situation the application of these criteria results in a sequence which converges to a μ, there is justification to conclude that there 'really' is a feature of the scene which is represented by μ.

Given that we are interested in weak convergence, it is natural that our criteria for acceptance or rejection of inputs should be expressed in terms of a metric which metrizes the weak topology of measures, such as the Prokhorov metric (Prokhorov, 1956, [2]). Unfortunately, the Prokhorov metric is, in practice, a fairly difficult computational tool. The idea is to replace Prokhorov metric computations with computations in L^p after identifying measures with their Radon-Nikodym derivatives. This involves first 'localizing' to the subset of measures which are absolutely continuous to a given one, and secondly restricting to those measures whose derivatives are in L^p. The technical problem is to see that this can be done effectively, in the sense that we can state effective directed convergence criteria in L^p. In Section 2 we give some mathematical background, and develop the formalism for this study. In Section 3 we obtain convergence criteria in L^1 (Theorem 3.17) , and in Section 4 we extend these results to L^p (the 'Main Theorem'). We also prove that if a directed convergence procedure with inputs $\{\lambda_n\}$ results in a convergent sequence of conclusions $\{\mu_n\}$, then the sequence $\{\lambda_n\}$ must also converge (Section 3, Theorem 3.19; and Section 4, (2) of the 'Main Theorem'). These results are generalizations of results for the special case L^∞, which were published in [4] (Bennett and Cohen-Lehman, 1999). It is a natural question whether the converse of (2) of the 'Main Theorem' holds: if $\{\lambda_n\}$ is a convergent sequence of input measures, then will the corresponding sequence $\{\mu_n\}$ of conclusion measures converge? At present, this is unknown.

2. Recursively Updated Bayesian Probabilistic Inference. Suppose that we are given measurable spaces $(X, \mathfrak{X})$ and $(Y, \mathcal{Y})$, where X models states of some system and Y models a receptor. We refer to the elements of Y, or probability measures on Y, as 'premises,' since they serve as premises for inferences about the state. Denote by $\mathcal{P}(X)$ and $\mathcal{P}(Y)$, the spaces of probability measures on X and Y, respectively. Suppose we are given a Markovian kernel $N : X \times \mathcal{Y} \to [0,1]$. 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 $\mathcal{Y}$, $N(x,B)$ is the probability that a premise y in B would be acquired if a state 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

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

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

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

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

With the data (μ, N), the apparatus of conditional probability canonically gives rise to a kernel $P_{(\mu,N)} : Y \times \mathfrak{X} \to [0,1]$, 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 our system at a given instant, and that N correctly describes the likelihood of y's given x's. For y in Y and $A \subset X$, $P_{(\mu,N)}(y, A)$ is the conditional probability that the state corresponds to a point in A, given the premise y.

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 $\mathcal{P}(Y) \to \mathcal{P}(X)$ given by $\lambda \mapsto \lambda P_{(\mu,N)}(A) \overset{\text{def}}{=} \int_Y P_{(\mu,N)}(y,A)\,\lambda(dy)$ for A in $\mathfrak{X}$. In this sense we can view, $P_{(\mu,N)}$ as defining an 'inference map'

$$\Psi : \mathcal{P}(Y) \to \mathcal{P}(X)$$

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

(given the premise measure λ the conclusion measure $\Psi(\lambda)$ is inferred).

Definition 2.1. *A Bayesian probabilistic inference is a map*

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

for a 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 $N : X \times \mathcal{Y} \to [0,1]$ gives rise to a measure on $X \times Y$, denoted $\mu \otimes N$, defined by

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

(for 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

$$(*) \qquad P_{(\mu,N)}(y,A) = \tfrac{d(\mu(\mathbf{1}_A N))}{d(\mu N)}(y) \qquad \mu N - a.e.\ y \in Y$$

If α and β are fixed probability measures on Y and X, respectively, then $(*)$ is equivalent to the familiar formulation of the Bayes posterior given in terms of probability densities:

$$P_{(\mu,N)}(y,A) = \int_A f(x,y)\,\beta(dx).$$

where

$$f(x,y) = \frac{\frac{d\mu}{d\beta}(x)\frac{dN(x,\cdot)}{d\alpha}(y)}{\int_X \frac{d\mu}{d\beta}(x)\frac{dN(x,\cdot)}{d\alpha}(y)\,\beta(dx)}.$$

Proposition 2.2. $\mu NP_{(\mu,N)} = \mu$

Proof. We have, for any μ-measurable subset $A \subset X$,

$$\mu NP_{(\mu,N)}(A) = \int_Y P_{(\mu,N)}(y,A)\,\mu N(dy).$$

Which, by definition of Bayes posterior,

$$\begin{aligned}
&= \int_Y \frac{d(\mu(\mathbf{1}_A N))}{d(\mu N)}(y)\,\mu N(dy)\\
&= \mu(\mathbf{1}_A N)(Y)\\
&= \int_X (\mathbf{1}_A(x)N(x,Y))\,\mu(dx)\\
&= \mu(A). \qquad\qquad (N(x,Y)=1)
\end{aligned}$$

□

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

The type of Bayesian inference described in Definition 2.1 can be updated recursively in discrete time 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 $\mathcal{P}(Y)$ to $\mathcal{P}(X)$. To simplify notation, let us denote this map by P_0. We use 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 $P_n = P_{(\mu_n,N)}$. We can 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 $(\mu_n, P_n) \longmapsto (\mu_{n+1}, P_{n+1})$. This procedure is more fully recursive than classical Bayesian inference, in which the posterior itself is not updated. Hence we may call it *Bayesian recursive updating* or *Bayesian structural updating.*

3. The L^1 Window.

NOTATION 3.1. *Let (U, d) be a metric space. Let $A \subset U$, and let $\varepsilon > 0$. We denote*

$$A^{\varepsilon} = \{u \in U : d(u, A) < \varepsilon\}.$$

With this we have:

DEFINITION 3.2. *Let (U, d) 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 $\rho_{prok}(\mu_1, \mu_2)$, is defined as follows:*

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

where

$$\begin{aligned} \epsilon_{12} &= inf\,\{\varepsilon : \mu_1(A) < \mu_2(A^{\varepsilon}) + \varepsilon \text{ for all } A \in \mathcal{U}\}, \\ \epsilon_{21} &= inf\,\{\varepsilon : \mu_2(A) < \mu_1(A^{\varepsilon}) + \varepsilon \text{ for all } A \in \mathcal{U}\}. \end{aligned}$$

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

THEOREM 3.3. *(Prokhorov 1956 [2])*

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

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

NOTATION 3.4. *Let U be a complete separable metric space with a metric d; U is then a measurable space with Borel σ-algebra, $\mathfrak{U}$, associated to the metric topology. Let $\mathcal{P}(U)$ denote the set of probability measures on U. For a given measure ν on U, and for an integrable function f defined on U, denote*

$$\|f\|_{\nu}^{1} = \int_U |f| d\nu.$$

Then, as usual, $L^1(U, \nu)$ denotes the set of ν - a.e. equivalence classes of measurable functions, f, on U such that $\|f\|_{\nu}^{1} < \infty$. We denote, by ρ_{ν}^{1}, the metric on $L^1(U, \nu)$ associated to the $\|\cdot\|_{\nu}^{1}$ norm.

Let S^+ denote the subset of the unit sphere in $L^1(U, \nu)$ consisting of non-negative functions such that $\|f\|_{\nu}^{1} = 1$.

NOTATION 3.5. *Let*

$$\mathcal{B}_{\nu}(U) = \{\sigma \in \mathcal{P}(U) : \sigma << \nu\}$$

where $<<$ *denotes absolute continuity of measures.* $B_\nu(U)$ *is a topological space for the weak topology induced from* $\mathcal{P}(U)$*.*

Define the function

$$\Phi : S^+ \subset L^1(U,\nu) \to B_\nu(U)$$

by

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

We now have the following proposition:

Proposition 3.6. Φ *is a continuous bijective function onto* $B_\nu(U)$*. Consequently, since* S^+ *is complete, (a closed subset of a complete space)* Φ *transforms Cauchy sequences in* S^+ *to Cauchy sequences in* $\mathcal{B}_\nu(U)$*.*

Proof. Consider $g_1, g_2 \in S^+$. If $g_1\nu = g_2\nu$, then

$$\int_B g_1 d\nu = \int_B g_2 d\nu$$

for every $B \in \mathfrak{U}$. And as g_1 and g_2 are integrable, we have

$$\int_B (g_1 - g_2) d\nu = 0$$

for every $B \in \mathfrak{U}$. Thus $g_1 - g_2 = 0$, ν-almost everywhere on U, proving Φ is *injective.*

For any probability measure $\sigma \in \mathcal{B}_\nu(U)$, since $\sigma << \nu$, we have that $\frac{d\sigma}{d\nu}$ exists, and (as σ is a probability measure) $\int_U \frac{d\sigma}{d\nu} d\nu = 1$. Thus, $\frac{d\sigma}{d\nu} \in S^+$ and $\Phi(\frac{d\sigma}{d\nu}) = \sigma$. That is, Φ is *surjective.*

To show that Φ is continuous, we show that for any sequence $\{\sigma_n\}_{n\in\mathbb{N}}$ and σ in $\mathcal{B}_\nu(U)$, if

$$\rho^1_\nu\left(\frac{d\sigma_n}{d\nu}, \frac{d\sigma}{d\nu}\right) = \left\|\frac{d\sigma_n}{d\nu} - \frac{d\sigma}{d\nu}\right\|^1_\nu \to 0$$

as $n \to \infty$, then $\{\sigma_n\}_{n\in\mathbb{N}}$ converges weakly to σ.

Let f be an arbitrary bounded continuous function on U. Then

$$\left|\int_U f\frac{d\sigma_n}{d\nu}\ d\nu - \int_U f\frac{d\sigma}{d\nu}\ d\nu\right| \le \int_U |f| \left|\frac{d\sigma_n}{d\nu} - \frac{d\sigma}{d\nu}\right|\ d\nu \ \longrightarrow 0 \tag{3.1}$$

as $n \to \infty$. Thus, by (3.1),

$$\lim_{n\to\infty} \int_U f\ \ d\sigma_n = \int_U f\ \ d\sigma$$

for any bounded continuous function f. That is, $\sigma_n \to \sigma$ weakly. □

Via the map Φ, we may now identify S^+ with $\mathcal{B}_\nu(U)$. In fact, we may even 'transport' the metric ρ^1_ν to the space $\mathcal{B}_\nu(U) \subset \mathcal{P}(U)$ in the following way:

Notation 3.7. *Let σ and τ be probability measures absolutely continuous with respect to a fixed measure ν.*

$$\rho^1_\nu(\sigma,\tau) \overset{def}{=} \rho^1_\nu(\frac{d\sigma}{d\nu},\frac{d\tau}{d\nu}). \tag{3.2}$$

Henceforth, we will use the notation, ρ^1_ν, both in its original sense as the L^1 metric on S^+ and as the metric on $\mathcal{B}_\nu(U)$ defined by equation (3.2). Since S^+ is complete with respect to ρ^1_ν, so is $\mathcal{B}_\nu(U)$. We can restate Proposition 3.6 in terms of a comparison of this metric with the Prokhorov metric, ρ_{prok}, on $\mathcal{B}_\nu(U)$.

Proposition 3.8. *The ρ^1_ν metric topology is weaker than the ρ_{prok} metric topology, i.e., if a sequence of measures in $\mathcal{B}_\nu(U)$ converges in the ρ^1_ν metric, then it converges in the ρ_{prok} metric. Consequently, since $\mathcal{B}_\nu(U)$ is ρ^1_ν-complete, a ρ^1_ν-Cauchy sequence of measures in $\mathcal{B}_\nu(U)$ is a ρ_{prok}-Cauchy sequence.*

Remark. If U is a complete metric space, then $\mathcal{P}(U)$ is complete with respect to ρ_{prok}, but $\mathcal{B}_\nu(U)$ is not. (Cauchy sequences that converge to *Dirac* measure, for example, have limit outside of $\mathcal{B}_\nu(U)$.) However, $\mathcal{B}_\nu(U)$ *is* complete with respect to ρ^1_ν.

We now need the following theorems to use $L^1(U,\nu)$ as a 'window'. We obtain recursively determined metric criteria on the sequence of input measures for the sequence of conclusion measures to be L^1-convergent (via Φ), and, hence, weakly convergent.

Proposition 3.9. *Let $(U,\mathfrak{U})$ and $(V,\mathfrak{V})$ be two measurable spaces. Let K be a Markovian kernel from U to V, and let μ and ν be probability measures on U. Then*

$$\mu << \nu \qquad \Rightarrow \qquad \mu K << \nu K. \tag{3.3}$$

Proof. Let $B \in \mathfrak{V}$ be given, and suppose that $\nu K(B) = 0$. Then

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

Since K is non-negative, this means

$$K(u,B) = 0 \qquad for\ \nu - a.e.\ u \in U.$$

Let $A = \{u \in U : K(u,B) > 0\}$. By above, $\nu(A) = 0$, so $\mu(A) = 0$, since $\mu << \nu$. We have that, as $K(u,B) = 0$ on A^c,

$$\begin{aligned}
\mu K(B) &= \int_U K(u,B)\,d\mu \\
&= \int_A K(u,B)\,d\mu + \int_{A^c} K(u,B)\,d\mu \\
&\le \int_{A^c} 1\,d\mu + 0 && \text{(K is Markovian)} \\
&= \mu(A) + 0 \\
&= 0.
\end{aligned}$$

That is, $\mu K(B) = 0$ and, therefore, $\mu K << \nu K$. □

As a corollary, we apply the result of Proposition 3.9 to Bayesian recursive inference terminology.

COROLLARY 3.10. *Let X and Y be measurable spaces. Let μ_0 be a fixed prior probability measure on X, and suppose N is a time invariant noise kernel from X to Y. If $\{\lambda_n\}_{n\in\mathbb{N}}$ is a sequence of premise measures on Y, and $\{\mu_n\}_{n\in\mathbb{N}}$ is the generated sequence of conclusions, then, for each $n = 1, 2, 3, \ldots$,*

$$\lambda_n << \mu_{n-1}N \qquad \Rightarrow \qquad \mu_n << \mu_{n-1} \tag{3.4}$$

Proof. Use $\nu = \lambda_n$, $\mu = \mu_{n-1}N$, and $K = P_{n-1}$ in Proposition 3.9. Then, as $\lambda_n P_{n-1} = \mu_n$ and $\mu_{n-1}NP_{n-1} = \mu_{n-1}$, the result is immediate. □

COROLLARY 3.11. *For each $n \geq 1$,*

$$\mu_n << \mu_{n-1} \qquad \Rightarrow \qquad \mu_n N << \mu_{n-1}N.$$

NOTATION 3.12. *Let μ and ν be probability measures on a measurable space U. If $\mu << \nu$, then $\frac{d\mu}{d\nu}$ exists. Let us adopt the convention to use a representative of the equivalence class $\frac{d\mu}{d\nu}$ such that $\frac{d\mu}{d\nu} = 0$ outside of the* support *of ν (supp(ν)).*
We will write $\mathbf{1}_\nu$ to mean the indicator function for supp(ν), i.e., $\mathbf{1}_{supp(\nu)}$. As a result, $\frac{d\mu}{d\nu} = \frac{d\mu}{d\nu}\mathbf{1}_\nu$.

THEOREM 3.13. *Let $(U, \mathfrak{U})$ and $(V, \mathfrak{V})$ be two measurable spaces. Let K be any Markovian kernel from U to V, and let σ and ν be probability measures on U such that $\sigma << \nu$. Then, for any $\varepsilon > 0$,*

$$\rho^1_\nu(\sigma, \nu) < \varepsilon \qquad \Rightarrow \qquad \rho^1_{\nu K}(\sigma K, \nu K) < \varepsilon. \tag{3.5}$$

Proof. Let $\varepsilon > 0$ be given, and suppose $\rho^1_\nu(\sigma, \nu) < \varepsilon$. That is,

$$\left\| \frac{d\sigma}{d\nu} - \mathbf{1}_\nu \right\|^1_\nu = \int_U |\frac{d\sigma}{d\nu} - \mathbf{1}_\nu| \, d\nu < \varepsilon.$$

Let $A^+ = \{u \in U : \frac{d\sigma}{d\nu}(u) > \mathbf{1}_\nu(u)\}$ and $A^- = \{u \in U : \frac{d\sigma}{d\nu}(u) < \mathbf{1}_\nu(u)\}$. Then we have,

$$\int_{A^+} (\frac{d\sigma}{d\nu}(u) - \mathbf{1}_\nu(u)) \, d\nu + \int_{A^-} (\mathbf{1}_\nu(u) - \frac{d\sigma}{d\nu}(u)) \, d\nu < \varepsilon. \tag{3.6}$$

We make use of the following facts, noting first that, by Proposition 3.9, $\frac{d\sigma K}{d\nu K}$ makes sense. By definition, for every $B \in \mathfrak{V}$,

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

And, by the Radon-Nikodym theorem,

$$\sigma K(B) = \int_B \frac{d\sigma K}{d\nu K}\, d\nu K.$$

So,

$$\int_U K(u,B)\frac{d\sigma}{d\nu}\, d\nu = \int_B \frac{d\sigma K}{d\nu K}\, d\nu K. \tag{3.7}$$

Similarly,

$$\nu K(B) = \int_U K(u,B)\, d\nu = \int_U K(u,B)\mathbf{1}_\nu\, d\nu.$$

and

$$\nu K(B) = \int_B \mathbf{1}_{\nu K}(v)\, d\nu K.$$

So that,

$$\int_U K(u,B)\mathbf{1}_\nu\, d\nu = \int_B \mathbf{1}_{\nu K}(v)\, d\nu K. \tag{3.8}$$

Consider the case:

$$B = \{v \in V : \frac{d\sigma K}{d\nu K}(v) > \mathbf{1}_{\nu K}(v)\}.$$

For this B, by (3.7) and (3.8) we have,

$$\begin{aligned}\int_B (\frac{d\sigma K}{d\nu K} - \mathbf{1}_{\nu K})\, d\nu K &= \int_U K(u,B)[\frac{d\sigma}{d\nu} - \mathbf{1}_\nu]\, d\nu \\ &= \int_{A^+} K(u,B)[\frac{d\sigma}{d\nu} - \mathbf{1}_\nu]\, d\nu + \int_{A^-} K(u,B)[\frac{d\sigma}{d\nu} - \mathbf{1}_\nu]\, d\nu \\ &\leq \int_{A^+} [\frac{d\sigma}{d\nu} - \mathbf{1}_\nu]\, d\nu + 0.\end{aligned}$$

In a similar fashion, we may define

$$C = \{v \in V : \frac{d\sigma K}{d\nu K}(v) < \mathbf{1}_{\nu K}(v)\}$$

so that (with C replacing B above)

$$\begin{aligned}\int_C (\mathbf{1}_{\nu K} - \frac{d\sigma K}{d\nu K})\, d\nu K &= \int_U K(u,C)[\mathbf{1}_\nu - \frac{d\sigma}{d\nu}]\, d\nu \\ &= \int_{A^+} K(u,C)[\mathbf{1}_\nu - \frac{d\sigma}{d\nu}]\, d\nu + \int_{A^-} K(u,C)[\mathbf{1}_\nu\, d\nu - \frac{d\sigma}{d\nu}]\, d\nu \\ &\leq 0 + \int_{A^-} [\mathbf{1}_\nu - \frac{d\sigma}{d\nu}]\, d\nu.\end{aligned}$$

Thus,

$$\begin{aligned}\int_U \mid \frac{d\sigma K}{d\nu K} - \mathbf{1}_{\nu K} \mid d\nu K &= \int_B (\frac{d\sigma K}{d\nu K} - \mathbf{1}_{\nu K})\, d\nu K + \int_C (\mathbf{1}_{\nu K} - \frac{d\sigma K}{d\nu K})\, d\nu K \\ &\leq \int_{A^+} [\frac{d\sigma}{d\nu} - \mathbf{1}_\nu]\, d\nu + \int_{A^-} [\mathbf{1}_\nu - \frac{d\sigma}{d\nu}]\, d\nu. \\ &< \varepsilon. \qquad \text{(by (3.6))}\end{aligned}$$

That is, $\rho^1_{\nu K}(\sigma K, \nu K) < \varepsilon$. □

We apply the result of Theorem 3.13 to Bayesian recursive updating to obtain the following corollaries; as always, the notation is as introduced in Section 2.

Corollary 3.14. *For any $n \geq 0$ and $\varepsilon > 0$, if $\lambda_{n+1} << \mu_n N$, then*

$$\rho^1_{\mu_n N}(\lambda_{n+1}, \mu_n N) < \varepsilon \qquad \Rightarrow \qquad \rho^1_{\mu_n}(\mu_{n+1}, \mu_n) < \varepsilon.$$

Proof. Use $K = P_n$ in Theorem 3.13 and the fact that $\mu_n N P_n = \mu_n$. □

Corollary 3.15. *For any $n \geq 0$ and $\varepsilon > 0$, if $\mu_{n+1} << \mu_n$, then*

$$\rho^1_{\mu_n}(\mu_{n+1}, \mu_n) < \varepsilon \qquad \Rightarrow \qquad \rho^1_{\mu_n N}(\mu_{n+1} N, \mu_n N) < \varepsilon.$$

The next lemma provides sufficient criteria for the recursive selection of a sequence of input measures which yields a convergent sequence of conclusion measures (convergent with respect to a fixed prior measure).

Lemma 3.16. *Let μ_0 be given. For any $n \geq 0$ and $\kappa > 0$, if $\lambda_{n+1} << \mu_n N$ and $\rho^1_{\mu_n N}(\lambda_{n+1}, \mu_n N) < \kappa$, then*

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

Proof. Note that, $\lambda_{n+1} << \mu_n N$ for all n, implies that $\mu_{n+1} << \mu_n$ for all n by Proposition 3.9. If $\rho^1_{\mu_n N}(\lambda_{n+1}, \mu_n N) < \kappa$, then, by Corollary 3.14, we have

$$\rho^1_{\mu_n}(\mu_{n+1}, \mu_n) < \kappa.$$

We have,

$$\begin{aligned}\rho^1_{\mu_0}(\mu_{n+1}, \mu_n) &= \left\| \frac{d\mu_{n+1}}{d\mu_0} - \frac{d\mu_n}{d\mu_0} \right\|^1_{\mu_0} \\ &= \int_U \mid \frac{d\mu_{n+1}}{d\mu_n} - \mathbf{1}_{\mu_n} \mid\mid \frac{d\mu_n}{d\mu_0} \mid d\mu_0 \quad \left(\text{since } \frac{d\mu_n}{d\mu_0} = \frac{d\mu_n}{d\mu_0}\mathbf{1}_{\mu_0}, \text{ by convention}\right) \\ &= \int_U \mid \frac{d\mu_{n+1}}{d\mu_n} - \mathbf{1}_{\mu_n} \mid d\mu_n \\ &= \left\| \frac{d\mu_{n+1}}{d\mu_n} - \mathbf{1}_{\mu_n} \right\|^1_{\mu_n} \\ &= \rho^1_{\mu_n}(\mu_{n+1}, \mu_n).\end{aligned}$$

That is,

$$\rho^1_{\mu_0}(\mu_{n+1},\mu_n) = \rho^1_{\mu_n}(\mu_{n+1},\mu_n) \quad (3.9)$$
$$< \kappa.$$

□

THEOREM 3.17. *Let X and Y be complete, separable metric spaces. Suppose that $\{\lambda_n\}_{n\in\mathbb{N}}$ is a sequence of probability measures on Y, such that $\lambda_{n+1} << \mu_n N$, which recursively generates the sequence $\{\mu_n\}_{n\in\mathbb{N}}$ on X via $\mu_{n+1} = \lambda_{n+1}P_{(\mu_n,N)}$. Suppose we are given a sequence $\{\kappa_n\}_{n\in\mathbb{N}}$ of non-negative numbers such that $\sum_{i=0}^{\infty}\kappa_i < \infty$. Suppose that, for each $n = 0, 1, 2, \ldots$,*

$$\rho^1_{\mu_n N}(\lambda_{n+1},\mu_n N) < \kappa_n.$$

Then, the sequence $\{\mu_n\}_{n\in\mathbb{N}}$ converges weakly in $\mathcal{P}(X)$.

Proof. By Lemma 3.16, since $\rho^1_{\mu_n N}(\lambda_{n+1},\mu_n N) < \kappa_n$ for each n, we have,

$$\rho^1_{\mu_0}(\mu_{n+1},\mu_n) < \kappa_n \qquad \textit{for each } n \in \mathbb{N}. \quad (3.10)$$

Then, since $\sum_{i=0}^{\infty}\kappa_i$ converges, equation (3.10) means that $\{\frac{d\mu_n}{d\mu_0}\}_{n\in\mathbb{N}}$ is a Cauchy sequence in $S^+ \subset L^1(X,\mu_0)$. Thus, by Proposition 3.6, via Φ, $\{\mu_n\}_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathcal{B}_{\mu_0}(X)$. Hence, $\{\mu_n\}_{n\in\mathbb{N}}$ is a ρ_{prok}-Cauchy sequence in $\mathcal{B}_{\mu_0}(X)$ and converges weakly in $\mathcal{P}(X)$. □

Theorem 3.17 describes concretely how a system can formulate a directed convergence strategy. The goal of the system is to acquire a weakly convergent sequence of conclusion measures, a stable percept. Suppose the system has the ability to accept or reject input measures λ. To *accept* λ at the $(n+1)$st stage means to take $\lambda = \lambda_{n+1}$ and obtain updated conclusion measure $\mu_{n+1} = \lambda_{n+1}P_{(\mu_n,N)}$. To *reject* λ means not to use λ for the purpose of updating the conclusion. Suppose, moreover, that the system has the capability to determine, for each n, the $L^1(Y,\mu_n N)$ - metric distance of an input λ to the measure $\mu_n N$, i.e., to determine $\rho^1_{\mu_n N}(\lambda,\mu_n N)$. Now, if a sequence of numbers, $\{\kappa_n\}$, such that $\sum_{i=0}^{\infty}\kappa_i < \infty$, is given, then at stage $n+1$, the system can wait for an input measure λ such that $\rho^1_{\mu_n N}(\lambda,\mu_n N) < \kappa_n$. When such a λ is acquired, it will be accepted as λ_{n+1}. According to the preceding theorem, the sequence of conclusions, $\{\mu_n\}$, corresponding to the sequence of inputs $\{\lambda_n\}$ selected in this manner, will converge weakly in $\mathcal{P}(X)$.

In practice, the choice of the sequence of numbers $\{\kappa_n\}$ 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 - a correct conclusion, by definition, is a weak limit in $\mathcal{P}(X)$ of a recursively generated sequence of conclusions, $\{\mu_n\}$. The greater the confidence, the smaller κ_n is. That is, incoming inputs λ must fall within a more restrictive neighborhood to be accepted as λ_{n+1}. If such a λ is forthcoming, we say that the system's *belief* (about the closeness of conclusion μ_n to the correct conclusion μ) is *confirmed.* If a belief-confirming premise is *not* acquired in a reasonable length of time at the *nth* stage, then the system's confidence in conclusion μ_n may decrease. In this case, it is reasonable that the degree of belief, κ_n, may be replaced by a larger number, to improve the possibility of acquiring an acceptable λ (one such that $\rho^1_{\mu_n N}(\lambda, \mu_n N) < \kappa_n$) to be used as λ_{n+1}. In this manner, the 'direction' of the search for stable percepts is responsive to the actual environmental conditions.

We now show in Theorem 3.19 the conditions on the λ_n's which ensure the convergence of the μ_n's (Theorem 3.17) also ensure the convergence of the sequence $\{\lambda_n\}_{n\in\mathbb{N}}$, and at a rate comparable to that of $\{\mu_n\}_{n\in\mathbb{N}}$ in $\mathcal{P}(X)$.

LEMMA 3.18. *For every n and every $\varepsilon > 0$, if $\mu_n << \mu_0$, then*

$$\rho^1_{\mu_0}(\mu_n, \mu_{n-1}) < \varepsilon \qquad \Rightarrow \qquad \rho^1_{\mu_0 N}(\mu_n N, \mu_{n-1} N) < \varepsilon.$$

Proof. We note that

$$\rho^1_{\mu_0 N}(\mu_n N, \mu_{n-1} N) = \rho^1_{\mu_{n-1} N}(\mu_n N, \mu_{n-1} N).$$

Namely, by definition,

$$\rho^1_{\mu_0 N}(\mu_n N, \mu_{n-1} N) = \left\| \frac{d\mu_n N}{d\mu_0} - \frac{d\mu_{n-1} N}{d\mu_0} \right\|^1_{\mu_0 N},$$

where,

$$\begin{aligned} \left\| \frac{d\mu_n N}{d\mu_0} - \frac{d\mu_{n-1} N}{d\mu_0} \right\|^1_{\mu_0 N} &= \int_U \left| \frac{d\mu_n N}{d\mu_{n-1}} - \mathbf{1}_{\mu_{n-1} N} \right| \left| \frac{d\mu_{n-1} N}{d\mu_0} \right| \, d\mu_0 N \\ &= \int_U \left| \frac{d\mu_n N}{d\mu_{n-1}} - \mathbf{1}_{\mu_{n-1} N} \right| \, d\mu_{n-1} N \\ &= \rho^1_{\mu_{n-1} N}(\mu_n N, \mu_{n-1} N). \end{aligned}$$

Thus, for a given $\varepsilon > 0$, if $\mu_n << \mu_0$ for all n, then

$$\begin{aligned} \rho^1_{\mu_0}(\mu_n, \mu_{n-1}) < \varepsilon &\Rightarrow \rho^1_{\mu_{n-1}}(\mu_n, \mu_{n-1}) < \varepsilon && \text{(by equation (3.9))} \\ &\Rightarrow \rho^1_{\mu_{n-1} N}(\mu_n N, \mu_{n-1} N) < \varepsilon && \text{(by Theorem 3.13)} \\ &\Rightarrow \rho^1_{\mu_0 N}(\mu_n N, \mu_{n-1} N) < \varepsilon. \end{aligned}$$

□

THEOREM 3.19. *Suppose we have a sequence of probability measures, $\{\mu_n\}_{n\in\mathbb{N}}$ on X, which is obtained recursively from a sequence of premises $\{\lambda_n\}_{n\in\mathbb{N}}$ which satisfy $\lambda_{n+1} << \mu_n N$, for all n. Let $\{\kappa_n\}$ be a sequence of non-negative numbers such that $\sum_{i=0}^{\infty} \kappa_i < \infty$. Assume that*

$$\rho^1_{\mu_n N}(\lambda_{n+1}, \mu_n N) < \kappa_n.$$

Then,

$$\rho^1_{\mu_0 N}(\lambda_{n+1}, \lambda_n) < \kappa_n + 2\kappa_{n-1}. \tag{3.11}$$

Consequently, $\{\lambda_n\}$ *converges weakly in* $\mathcal{P}(Y)$.

Proof. We have

$$\begin{aligned} \rho^1_{\mu_n N}(\lambda_{n+1}, \mu_n N) < \kappa_n &\Rightarrow \rho^1_{\mu_0}(\mu_{n+1}, \mu_n) < \kappa_n && \text{(by Lemma 3.16)} \\ &\Rightarrow \rho^1_{\mu_0 N}(\mu_{n+1}N, \mu_n N) < \kappa_n && \text{(by Lemma 3.18).} \end{aligned}$$

Thus,

$$\rho^1_{\mu_n N}(\lambda_{n+1}, \mu_n N) < \kappa_n \Rightarrow \rho^1_{\mu_0 N}(\mu_{n+1}N, \mu_n N) < \kappa_n. \tag{3.12}$$

Moreover, for any $n = 1, 2, 3, \dots,$

$$\begin{aligned} \rho^1_{\mu_0 N}(\lambda_n, \mu_{n-1}N) &= \int_U \left| \frac{d\lambda_n}{d\mu_{n-1}N} - \mathbf{1}_{\mu_{n-1}N} \right| \left| \frac{d\mu_{n-1}N}{d\mu_0 N} \right| \, d\mu_0 N \\ &= \int_U \left| \frac{d\lambda_n}{d\mu_{n-1}N} - \mathbf{1}_{\mu_{n-1}N} \right| d\mu_{n-1}N. \\ &= \rho^1_{\mu_{n-1}N}(\lambda_n, \mu_{n-1}N). \end{aligned}$$

So that,

$$\rho^1_{\mu_0 N}(\lambda_n, \mu_{n-1}N) = \rho^1_{\mu_{n-1}N}(\lambda_n, \mu_{n-1}N). \tag{3.13}$$

By the triangle inequality,

$$\rho^1_{\mu_0 N}(\lambda_{n+1}, \lambda_n) \le \rho^1_{\mu_0 N}(\lambda_{n+1}, \mu_n N) + \rho^1_{\mu_0 N}(\mu_n N, \mu_{n-1}N) + \rho^1_{\mu_0 N}(\mu_{n-1}N, \lambda_n). \tag{3.14}$$

Using (3.12) and (3.13) in (3.14), we have

$$\begin{aligned} \rho^1_{\mu_0 N}(\lambda_{n+1}, \lambda_n) &< \kappa_n + \kappa_{n-1} + \kappa_{n-1} \\ &= \kappa_n + 2\kappa_{n-1}. \end{aligned}$$

Now, since $\sum_{i=0}^{\infty} \kappa_i < \infty$, we have that

$$\sum_{i=1}^{\infty} (\kappa_i + 2\kappa_{i-1}) = \sum_{i=1}^{\infty} \kappa_i + 2\sum_{i=0}^{\infty} \kappa_i < \infty.$$

Thus, $\{\lambda_n\}$ is a $\rho^1_{\mu_0 N}$-Cauchy sequence. By Proposition 3.8, $\{\lambda_n\}$ is then a ρ_{prok} -Cauchy sequence in $\mathcal{B}_{\mu_0 N}(Y) \subset \mathcal{P}(Y)$ and, therefore, weakly convergent in $\mathcal{P}(Y)$. □

Remark. In practice, as the numbers κ_n may correspond to degree of belief, there is no reason to expect that $\{\kappa_n\}$ is a strictly *decreasing* sequence of non-negative numbers. (Even though the κ_n's must go to 0.)

4. Extension to L^p. In this section, we use fundamental relationships between the L^p spaces to prove the following main result:

MAIN THEOREM. *Suppose that $\{\lambda_n\}_{n\in\mathbb{N}}$ is a sequence of probability measures on Y which recursively generate the sequence $\{\mu_n\}_{n\in\mathbb{N}}$ on X via $\mu_{n+1} = \lambda_{n+1}P_{(\mu_n,N)}$. Suppose that, for each n, $\lambda_{n+1} << \mu_n N$. Let $\{\kappa_n\}_{n\in\mathbb{N}}$ be a sequence of non-negative numbers such that $\sum_{i=0}^{\infty}\kappa_i < \infty$. For each n, assume that (see Notation 4.1)*

$$\rho^p_{\mu_n N}(\lambda_{n+1}, \mu_n N) \leq \kappa_n.$$

Then,
(1) The sequence $\{\mu_n\}_{n\in\mathbb{N}}$ converges weakly in $\mathcal{P}(X)$.
and
(2) The sequence $\{\lambda_n\}_{n\in\mathbb{N}}$ converges weakly in $\mathcal{P}(Y)$.

NOTATION 4.1. *Let $(U, \mathfrak{U})$ be as in Notation 3.2. Let ν be a given probability measure on the measurable space $(U, \mathfrak{U})$. For $p \in [1, \infty)$, let us denote*

$$\|f\|^p_\nu = \left\{\int_U |f|^p \, d\nu\right\}^{1/p}.$$

Then, as usual, $L^p(U, \nu)$ denotes the set of ν - a.e. equivalence classes of measurable functions, f, on U such that $\|f\|^p_\nu < \infty$. We denote by ρ^p_ν the metric on $L^p(U, \nu)$ associated to the p-norm.

For $p = \infty$, $L^\infty(U, \nu)$ denotes the set of ν-a.e. equal equivalence classes of measurable, ν-essentially bounded functions on U, where recall that a function f on U is ν-essentially bounded if there exists a real number M such that $\nu(\{u \in U : f(u) > M\}) = 0$. In that case, the essential sup norm, $\|f\|_\nu$, is the infimum of the set of such M's. We denote by ρ_ν the metric on $L^\infty(U, \nu)$ associated to this norm.

PROPOSITION 4.2. *Let $(U, \mathfrak{U}, \nu)$ be a probability space. For any p and q such that $1 \leq q < p \leq \infty$, $L^p(U, \nu) \subset L^q(U, \nu)$. In fact, for any measurable function g on U,*

$$\|g\|^q_\nu \leq \|g\|^p_\nu.$$

Proof. Proposition 4.2 is a well-known result. (Yeh 2000 [1], for instance). □

It follows from Proposition 4.2 that, for any open ball, $B_q(f, \varepsilon)$, in $L^q(U, \nu)$ there exists an open ball with respect to the p-metric - namely $B_p(f, \varepsilon)$ - contained inside. That is, $B_p(f, \varepsilon) \subset B_q(f, \varepsilon)$. This implies that $B_q(f, \varepsilon)$ is an *open set* with respect to the p-metric, i.e., ρ^p_ν is a finer metric on $L^p(U, \nu)$ than ρ^q_ν *restricted* to $L^p(U, \nu)$. We apply these results to a special case below:

PROPOSITION 4.3. *Let $p > 1$.*
(1) $L^p(U, \nu) \subset L^1(U, \nu)$.
(2) $\|\cdot\|^1_\nu \leq \|\cdot\|^p_\nu$.

(3) $\rho^p_\nu \preceq \rho^1_\nu \downharpoonright_{L^p(U,\nu)}$, *where* $\preceq$ *denotes finer topology.*

Proof. Let $q = 1$ in Proposition 4.2. □

Proof of Main Theorem. Suppose we have $\{\lambda_n\}_{n\in\mathbb{N}}$, a sequence of probability measures on Y, such that $\lambda_{n+1} << \mu_n N$ which recursively generate the sequence $\{\mu_n\}_{n\in\mathbb{N}}$ on X via $\mu_{n+1} = \lambda_{n+1}P_{(\mu_n,N)}$. Suppose we are given a sequence, $\{\kappa_n\}_{n\in\mathbb{N}}$ of non-negative numbers such that $\sum_{i=0}^{\infty} \kappa_i < \infty$. Let $p \in [1,\infty]$ be given and suppose that, for each $n \in \mathbb{N}$, we have that

$$\rho^p_{\mu_n N}(\lambda_{n+1}, \mu_n N) < \kappa_n.$$

Then, by (2) of Proposition 4.3,

$$\rho^1_{\mu_n N}(\lambda_{n+1}, \mu_n N) < \kappa_n.$$

We then apply Theorem 3.17 to obtain that the sequence $\{\mu_n\}_{n\in\mathbb{N}}$ converges weakly in $\mathcal{P}(X)$.

Moreover, by Theorem 3.19, $\{\lambda_n\}_{n\in\mathbb{N}}$ converges weakly in $\mathcal{P}(Y)$.

REFERENCES

[1] J. Yeh, *Lectures On Real Analysis*, World Scientific, Singapore-New Jersey-London-Hong Kong, 2000.

[2] Y. Prokhorov, *Convergence of Random Processes and Limit Theorems in Probability Theory*, Theory of Prob. and its App., 1(1956), pp. 157–214.

[3] B. Bennett and R. Cohen-Lehman, *Directed Convergence in Stable Percept Acquisition*, J. Math. Psych., (1997).

[4] ———, *L^∞ Metric Criteria for Convergence in Bayesian Recursive Inference Systems*, Adv. in App. Math., 23:(1999), pp. 255-273.

[5] B. Bennett, D. Hoffman, and C. Prakash, *Perception and Evolution*, to appear in *Perception and the Physical World*, D. Weyer and R. Mausfeld eds., Wiley (Chichester).

[6] P. Billingsley, *Convergence of Probability Measures,* Wiley, New York, 1968.

[7] P. Diaconis and D. Freedman, *On the Consistency of Bayes Estimates*, Ann. of Stat., 14(1986), pp. 1–26.