# Bayes Updating using Bounded Probability Densities for Model-Parameter Estimation in Dynamical Systems

**J. P. Norton**
**School of Electronic & Electrical Engineering, University of Birmingham**
**Edgbaston, Birmingham B15 2TT, UK**

**Abstract** Recursive parameter estimation is considered, updating a model of a system as new observations of the input and output are received. The most well established and widely used schemes, based on least squares or prediction-error minimisation, do not incorporate information on the distributions of the parameters or the observation errors. Bayes updating by contrast requires such information to be specified as probability density functions (pdf's). If the information about distribution is significant but imprecise, one possibility, widely canvassed, is to employ sets of pdf's. Two such approaches, Convex Bayes Estimation and Upper and Lower Probabilities (ULP's), are reviewed. The operations in applying ULP's to parameter estimation are examined and computational load is discussed. An example problem is presented, and a scalar measure of uncertainty for the posterior pdf is proposed.

## 1 INTRODUCTION

Algorithms for recursive estimation of the model parameters of a dynamical system, updating the estimates as new observations of the input and output are received, have matured into a small collection of widely accepted and well understood techniques. Among the most prominent are least-squares-based minimum-covariance and prediction-error-minimising algorithms (Ljung [1987], Norton [1986], Soderstrom and Stoica [1989]). They do not make restrictive assumptions about the distributions of the initial parameter error, observation error or modelling error and do not use such information (except incidentally to justify the algorithms in special cases). This is an advantage if little is known about the probability density functions (pdf's) of the random variables (parameter errors and model-output error, consisting of observation error plus modelling error, henceforth "observation error" for short), but it is inefficient if a significant amount *is* known.

A natural framework for parameter estimation taking distributional information fully into account is Bayes estimation, using each successive observation of the system output to update the pdf of the unknown parameters conditioned on the observations to date and on the initial information supplied as a prior pdf. Bayes estimation is attractive when a fuller characterisation of the estimation errors than their mean and covariance is desirable, as when they are markedly skewed or multimodal or have finite support. However, Bayes estimation may be rejected because, as well as often having heavy computational demand, it requires specification of two pdf's: the prior pdf of the parameters and the pdf of observation error, *i.e.* of the observations conditioned on the parameters. The need for a prior parameter pdf is not a serious practical difficulty so long as the cumulative information in the observations dominates the posterior pdf. Prior distributional information may, however, be hard to supply if the input:output records are short or the parameters are represented as time-varying, for instance as random walks or integrated random walks (Norton [1975, 1976], Jakeman and Young [1979]). For time-varying parameters, the pdf of the "process noise" forcing the parameter variation must be supplied and little may be known of it. An observation-error pdf is hard to supply when systematic modelling error is prominent.

For these reasons, it may be more attractive to specify a *set* of pdf's in place of one or more of the pdf's. Bayes updating using sets of pdf's, or more often of discrete probabilities, has been suggested by a number of authors. This paper reviews some of these suggestions in the context of parameter estimation and illustrates one of them, Upper and Lower Probabilities (ULP's), in an example. A scalar measure of uncertainty, geared to ULP updating, is also suggested.

## 2 VAGUE BELIEFS AND SETS OF PDF'S

The idea of employing sets of probabilities has a long history (Kyburg [1961], Good [1962], Levi [1974, 1980], Williams [1976], Fine [1988], Walley [1991], Lehner *et al.* [1996]). It has been closely associated with attempts to establish new calculi for quantifying beliefs and revising them on receiving new evidence; as a result it has been contentious. Although the meaning of probabilities or pdf's being specified must be discussed briefly here, this paper is more concerned with whether the estimation technique uses information likely to be available, is conceptually straightforward and yields easily interpreted results.

The raw material for updating estimates of the parameter vector $\theta$ of a selected model

$$y_t = h(\phi_t, \theta) + v_t \tag{1}$$

at time $t$ by Bayes' Rule consists of

- the new observation $y_t$
- the vector $\phi_t$ of known explanatory variables, which include the input
- the pdf $f(y_t|\theta)$, inferred from the pdf $f(v_t)$ of the observation error
- the prior pdf $f(\theta|Y_{t-1})$ conditioned on all information (initial and observations) up to and including time $t-1$.

For simplicity, $\theta$, $h$ and $f(v_t)$ will be taken as time-invariant. Initially, $f(\theta|Y_0)$ has to be supplied; commonly knowledge of $f(\theta|Y_0)$ is vague, and sometimes that of $f(v_t)$ is too, so either or both might be specified as sets rather than unique pdf's.

The observations are affected by unpredictable instrumentation errors, noise and unmonitored disturbances naturally characterised by their average behaviour in similar situations, so an ensemble-based interpretation of $f(v_t)$, and hence of $f(y_t|\theta)$, might seem appropriate. However, whether $f(y_t|\theta)$ can be specified objectively depends on circumstances. When instrumentation noise and errors dominate, their average behaviour in large samples may be measurable beforehand; if systematic but unknown modelling error dominates, or if the disturbance environment is non-stationary, guesswork is involved in specifying $f(y_t|\theta)$, so it is a subjective expression of belief. For $f(\theta|Y_0)$, there is no ensemble of realisations of $\theta$ to refer to. The unknown parameters are definite values which summarise the system's behaviour or predict its behaviour well. Thus $f(\theta|Y_0)$ is best viewed as expressing relative degrees of belief in possible values. [The acceptability of this view has been discussed at enormous length over many years; opting to employ Bayes' Rule implies acceptance].

If belief, perhaps vague, has to be quantified, it must be readily supplied, able to distinguish differences in confidence, and it must influence estimates in a reasonable way. Two simple ways to specify imprecise beliefs are as either (i) a parametric set of pdf's (*e.g.* a range of means and covariances defining a set of Gaussian pdf's), or (ii) a range of pdf's conforming to given constraints, *e.g.* upper and lower probabilities (ULP's) or probability densities at some or all values of the support, otherwise unconstrained except by the probability axioms. Neapolitan [1996] sees no compelling reason to resort to non-unique probabilities or pdf's, regarding imprecise beliefs as merely conditional on as yet unknown information. This may be true but is little help in this context. He also cites the objection that if, say, any probability in the range 0.3 to 0.5 is plausible, it isn't reasonable that a probability of 0.299 or 0.501, say, isn't. This does not recognise that probabilities at the extremes of the acceptable range are barely more credible than those just outside. Specifying a continuous function over the range might in principle be a better way to register varying confidence in belief, but would strain one's powers of introspection.

## 3 SET-BASED BAYES ESTIMATION

The closest set-based counterpart of classical estimation for dynamical systems is provided by Morrell and Stirling [1991], who consider estimation of state rather than parameters. State estimation adds a time-update (prediction) stage, also present in parameter estimation if the parameters are modelled as time-varying. They update a convex set of state means but a unique covariance, ensuring that the state set remains convex. The process noise and observation noise have unique pdf's.

Bayes updating of a convex set of prior probabilities of discrete values of state has been widely considered (Potter and Anderson [1980], White [1986], Snow [1991, 1996]).

The observations are often assumed also to be discrete-valued, which fits applications such as medical diagnosis from a succession of "yes/no" tests but can also approximate continuous-variable problems. Imprecise knowledge is embodied in a number of linear inequalities in the probabilities. An inequality might just say that one probability is greater than another, or a pair of inequalities might be used to confine a probability to an interval.

If the observation probabilities, conditioned on the parameters or state, are given unique values, the posterior state-probability set is convex. However, if they are uncertain and also defined by a set of linear inequalities, the posterior set is not generally convex. It does remain convex if the observation probability set is a box determined by interval bounds on individual (scalar) observations.

White [1986] considers estimation of a discrete state from discrete observations, with a system model supplying the probability of each possible observation value conditioned on every possible value of present and previous states. Usually it separates into a state-transition model and an observation model, both convex sets defined by linear inequalities. Exact propagation of the set of parameter or state values soon becomes too complicated; in both time and observation update, the features (such as vertices) of the prior set and the polytope of forcing or conditional observation probabilities show in the updated set, which is thus more complicated than either. White therefore gives an approximation algorithm to produce a set of inequalities containing the posterior set of state probabilities, and reports that (in a very simple example) it was heuristically possible to restrict the number of inequalities without undue loss of tightness.

Snow [1991, 1996] provides updating schemes for a prior defined by a set of linear inequalities and conditional observation probabilities confined to a box. Salo [1996] offers an alternative updating algorithm to give tighter posterior bounds. All these techniques require a number of linear programming solutions at each update.

The probability-set updating methods all rely on the sets being convex. Kyburg and Pittarelli [1996] point out that adding a constraint of mutual independence among variables with a set specification of their joint probabilities may render that set non-convex.

The next section discusses estimation with prior and posterior probability sets defined by upper and lower bounds, *i.e.* in a box.

## 4 ULP PARAMETER UPDATING WITH UNIQUE OBSERVATION-ERROR PDF

### 4.1 Updating procedure

The literature on ULP's also has considered discrete probabilities rather than pdf's, and for ease of explanation this section does so, even though the aim is estimation of continuous-valued parameters from continuous-valued observations. Updating bounds on a pdf is discussed briefly in Section 4.2. The probabilities here may be regarded as defining piecewise constant approximated pdf's.

Setting aside for now the question of how to segment parameter space, consider events

$$\varepsilon_i \equiv \{\theta \in \Theta_i \subset \Re^n\}, \quad i=1, 2, \ldots, N \qquad (2)$$

where $\theta$ is the parameter vector and the $\Theta_i$ are disjoint regions of parameter space (histogram bins, if you like). At time $t$, the prior probabilities

$$p(\varepsilon_i | Y_{t-1}) \equiv prob(\theta \in \Theta_i | Y_{t-1}) \qquad (3)$$

are to be updated to posterior probabilities $p(\varepsilon_i | Y_t)$ by processing observation vector $y_t$, given

$$p(y_t | \theta_i) = \int_{\Theta_i} f(v_t = y_t - h(\phi_t, \theta_i)) dv_t \qquad (4)$$

for $i=1,2, \ldots, N$. Bayes' Rule gives

$$p(\varepsilon_i | Y_t) = \frac{p(y_t|\varepsilon_i)p(\varepsilon_i)}{p(y_t|\varepsilon_i)p(\varepsilon_i) + \sum_{\substack{j=1 \\ j\neq i}}^{N} p(y_t|\varepsilon_j)p(\varepsilon_j)}$$

$$\equiv \frac{p(y_t|\varepsilon_i)p(\varepsilon_i)}{p(y_t|\varepsilon_i)p(\varepsilon_i) + \sigma_t(\{p(\varepsilon_j), j \neq i\})} \qquad (5)$$

where for conciseness dependence on $Y_{t-1}$ of probabilities on the right-hand side is not made explicit. The problem is first to distribute the $p(\varepsilon_j)$, subject to

$$\sum_{j=1}^{N} p(\varepsilon_j) = 1, \qquad (6)$$

to minimise $p(\varepsilon_i | Y_t)$, giving the lower probability $\check{p}(\varepsilon_i | Y_t)$, then as a separate exercise maximise it to give the upper probability $\hat{p}(\varepsilon_i | Y_t)$. Since

$$\frac{\partial p(\varepsilon_j | Y_t)}{\partial p(\varepsilon_i)} = -\frac{p(y_t|\varepsilon_i)p(\varepsilon)\partial\sigma_t \,/\, \partial p(\varepsilon_i)}{d^2} \qquad (7)$$

where $d$ is the denominator in (5), and

$$\frac{\partial \sigma_t(\{p(\varepsilon_j, j \neq i\})}{\partial p(\varepsilon_i)} < 0 \quad (8)$$

because increasing $p(\varepsilon_i)$ reduces at least one of the $p(\varepsilon_j), j \neq i$ through (6), $p(\varepsilon_i | Y_t)$ is minimised by setting $p(\varepsilon_i)$ to its lower bound $\breve{p}(\varepsilon_i)$ then distributing $1 - \breve{p}(\varepsilon_i)$ among the $\{p(\varepsilon_j, j \neq i\}$ so as to maximise $\sigma_t(\{p(\varepsilon_j, j \neq i\})$, subject to their lower and upper bounds.

This constrained maximisation is readily performed by setting every $p(\varepsilon_j)$ to its lower bound then increasing first whichever $p(\varepsilon_j)$ has the largest influence on $\sigma_t$, *i.e.* that for which $p(y_t | \varepsilon_j)$ is largest, until it reaches its upper bound, then increasing the $p(\varepsilon_j)$ with the second largest $p(y_t | \varepsilon_j)$, and so on until (6) is met. The result is

$$\left.\begin{array}{l} p(\varepsilon_j) = \hat{p}(\varepsilon_j), \ \ 1 \leq j \leq r-1 \\ \hat{p}(\varepsilon_r) \geq p(\varepsilon_r) \geq \breve{p}(\varepsilon_r) \\ p(\varepsilon_j) = \breve{p}(\varepsilon_j), \ \ r+1 \leq j \leq n-1 \end{array}\right\} \quad (9)$$

where the $p(y_t | \varepsilon_j), \ j = 1,2,\ldots,N, \ j \neq i$ have been sorted into decreasing order, then

$$\breve{p}(\varepsilon_i) = \frac{p(y_t | \varepsilon_i)\breve{p}(\varepsilon_i)}{p(y_t | \varepsilon_i)\breve{p}(\varepsilon_i) + \hat{\sigma}_t(\breve{p}(\varepsilon_i))} \quad (10)$$

$$\hat{\sigma}_t(\breve{p}(\varepsilon_i)) = \sum_{j=1}^{r-1} p(y_t | \varepsilon_j)\hat{p}(\varepsilon_j) + p(y_t | \varepsilon_r)p(\varepsilon_r) + \sum_{j=r+1}^{N-1} p(y_t | \varepsilon_j)\breve{p}(\varepsilon_j) \quad (11)$$

and

$$p(\varepsilon_r) = 1 - \breve{p}(\varepsilon_i) - \sum_{\substack{j=1 \\ j \neq i}}^{r-1} \hat{p}(\varepsilon_j) - \sum_{\substack{j=r+1 \\ j \neq i}}^{N-1} \breve{p}(\varepsilon_j) \quad (12)$$

Maximisation of each posterior probability is exactly similar; it is important to carry it out rather than inferring each upper bound as *1-(sum of lower bounds of all other probabilities),* as those bounds are generally mutually incompatible, being derived by independent minimisations. If they are treated as if compatible, they are jointly too low and make the inferred "upper bound" too high. Failure to carry out both minimisation and maximisation accounts for the dilation of ULP's encountered by some authors.

### 4.2 Application to continuous-valued variables; computing load

Finding the upper or lower bounds on the posterior pdf for continuous-valued parameter vector $\theta$ and observations is a variational problem with point constraints (due to the bounds on the prior pdf) and an integral constraint (integration of the posterior pdf to unity). The constraints make solution conceptually easy. The counterpart of the maximisation of $\sigma_t$ described above is to find the set of parameter values over which the prior probability density must be at its upper bound, by lowering $f_0$ determining a level set

$$U(f_0) = \{\theta | f(y_t | \theta) \geq f_0\} \quad (13)$$

(with $f$ a function of $\theta$ only, as $y_t$ *is* known) until

$$\int_{U(f_0)} \hat{f}(\theta | Y_{t-1}) d\theta + \int_{\bar{U}(f_0)} \breve{f}(\theta | Y_{t-1}) d\theta = 1 \quad (14)$$

where $\bar{U}(f_0)$ is the complement of $U(f_0)$. This is not computationally easy unless $f(y_t | \theta)$ has a simple parameterisation (and perhaps not even then, *e.g.* with an orthogonal expansion). Segmentation of $\theta$ space to make the problem discrete looks inevitable, pending more research on parameterising $f(y_t | \theta)$.

The main computing load in the discrete-$\theta$ updating process outlined above is ranking the $p(y_t | \varepsilon_j), \ j = 1, 2,\ldots, n, \ \ j \neq i$ and running through the corresponding $\hat{p}(\varepsilon_j)$ until (6) is met. The load is much smaller than this might seem to indicate, as only a single sort of all $N$ $p(y_t | \varepsilon_j)$ need be done per time instant. Moreover, a single calculation of the running totals of the ranked $p(y_t | \varepsilon_j)$ allows cheap determination of $r$ for each $i$.

For acceptable computing load, $N$ must not be too large. A prime motive for Bayes updating is its ability to deal with pdf's not restricted to any particular parametric family. The price is the difficult task of quantifying a nonparametric pdf at adequate resolution over as many dimensions as elements in $\theta$. In off-line parameter estimation, the difficulty may be eased by iteration, initially quantising $\theta$ space coarsely, using the results to select a much smaller region for finer coverage, and so on. On line, segments of $\theta$ space might be merged or split according to the size and variation of the evolving probabilities. Efficiency may also be improved by segmenting $\theta$ space according to the sensitivity of continuous-time items such as time constants and gains to the discrete-time parameters $\theta$. Economy in parameterisation, *e.g.* employing a rational transfer-function model, is essential.

### 4.3 A scalar measure of uncetainty

The final phase of Bayes estimation is to extract estimates from the posterior pdf or discrete probabilities by minimising a risk, the mean of a cost function (Norton [1986]). ULP's confer the extra choice of which pdf or discrete probability set to use in minimising the risk. For instance, a conservative policy is to use that pdf which maximises the minimum risk; a less pessimistic course would be to minimise the mean risk over the whole set of pdf's admitted by the ULP's. As such options are computationally expensive, one might prefer to wait until the ULP's are close enough for a nominal pdf or discrete probabilities to be calculated from them. If the ULP's vary in a complicated way, a scalar measure of remaining uncertainty at each point, or over each segment, may help in deciding when to extract the estimates. It may also serve to highlight deficiencies in the data set.

A measure may be designed by requiring that it ranges from 0 to 1, increases monotonically with $\overset{\frown}{p}-\breve{p}$, decreases monotonically with increasing $\overset{\frown}{p}$ for fixed $\overset{\frown}{p}-\breve{p}$, is zero when $\overset{\frown}{p}=\breve{p}$, and equals $\overset{\frown}{p}$ when $\breve{p}$ is zero. The simplest function found to do so is

$$u(\overset{\frown}{p},\breve{p}) = \overset{\frown}{p}\frac{\overset{\frown}{p}-\breve{p}}{\overset{\frown}{p}+\breve{p}} \qquad (15)$$

## 5 EXAMPLE OF ULP UPDATING

To allow easy display of the evolving ULP's, a single parameter $\theta$ in

$$y_t = \theta + v_t \qquad (16)$$

is estimated. The range of $\theta$ is divided into 15 bins. Uninformative initial ULP's [0, 0.25] for the whole range are updated by the observations shown in Fig. 1, which have Gaussian errors $\{v_t\}$ with zero mean and variance 1.

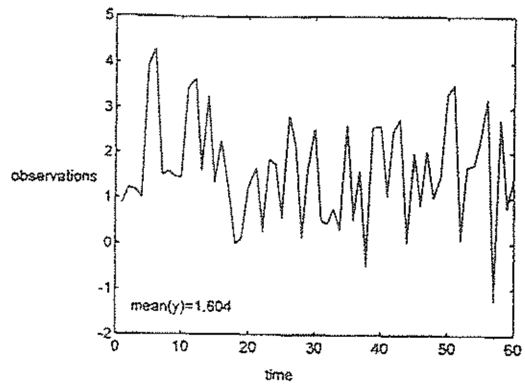


**Figure 1** Observations for example

Fig. 2 shows how the ULP's evolve, and Fig. 3 the uncertainty $u(\overset{\frown}{p},\breve{p})$.

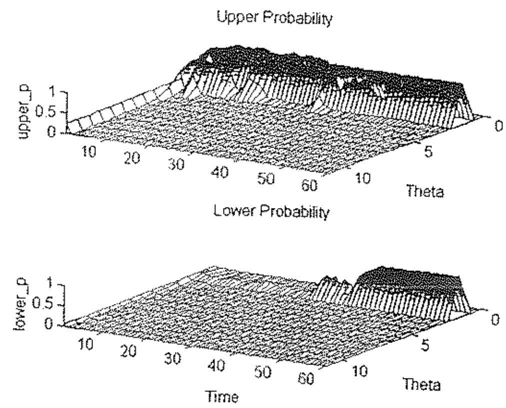


**Figure 2** Evolution of $\overset{\frown}{p}$ and $\breve{p}$

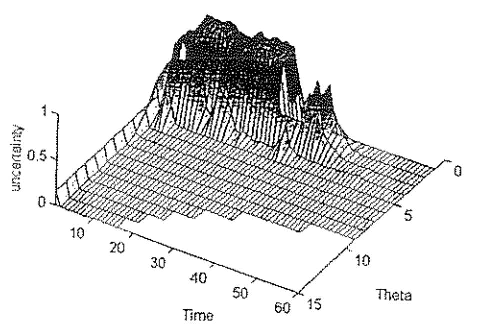


**Figure 3** Evolution of uncertainty $u(\overset{\frown}{p},\breve{p})$

Convergence of the UPL's and reduction of $u$ is rapid between about times 25 and 35, when the observation errors happen to be relatively small.

## 6 CONCLUSIONS

ULP's offer a way to incorporate vague prior knowledge into parameter estimation. They can be updated by Bayes' Rule fairly easily when the observation error pdf is known, the main extra operation being a sort once per update. A rapprochement with propagation of sets of linear inequalities, an alternative which has been well explored, seems overdue. Both techniques are complicated greatly if the observation error is also vaguely specified. Any such technique is limited by the need to divide the support into a large number of segments, with correspondingly large computing load. A scalar uncertainty measure is useful in assessing progress of the estimation.

### 6.1 Acknowledgements

Part of this paper is based on work supported by the Defence Evaluation and Research Agency, Malvern, UK. The example was programmed by Dr. Andrew Downing. Facilities provided by the Centre for Resource and Environmental Studies, Australian National University, during preparation of the paper are gratefully acknowledged.

### 6.2 References

Fine, T.L., Lower probability models for uncertainty and nondeterministic processes, *J. of Stat. Planning & Inference*, 20, 389-411, 1988.

Good, I.J., Subjective probability as the measure of a nonmeasurable set, in *Logic, Methodology and Philosophy of Science*, E. Nagel, P. Suppers and A. Tarski, Eds., Stanford Univ. Press, Palo Alto, CA, 319-329, 1962.

Jakeman, A.J. and P.C. Young, Joint parameter-state estimation, *Electronic Lett.*, 15, 582-583, 1979.

Kyburg, H.E., *Probability and the Logic of Rational Belief*, Wesleyan Univ. Press, Middletown, CT, 1961.

Kyburg, H.E. and M. Pittarelli, Set-based Bayesianism, *IEEE Trans. on Systems, Man & Cybernetics*, SMC-26, Pt. A, 324-339, 1996.

Lehner, P.E., K.B. Laskey and D. Dubois, An introduction to issues in higher order uncertainty, *IEEE Trans. on Systems, Man & Cybernetics*, SMC-26, Pt. A, 289-293, 1996 [Introduction to special issue on Higher Order Uncertainty] .

Levi, I., On indeterminate probabilities, *J. Philosphy*, 71, 391-418, 1974.

Levi, I., *The Enterprise of Knowledge*, MIT Press, London, 1980.

Ljung, L., *System Identification: Theory for the User*, Prentice-Hall, Englewood Cliffs, NJ, 1987.

Morrell, D.R. and W.C. Stirling, Set-valued filtering and smoothing, *IEEE Trans. on Systems, Man & Cybernetics*, SMC-21, 184-193, 1991.

Neapolitan, R.E., Is higher-order uncertainty needed?, *IEEE Trans. on Systems, Man & Cybernetics*, SMC26, Pt. A, 294-302, 1996.

Norton, J.P., Optimal smoothing in the identification of linear time-varying systems, *Proc. IEE*, 122, 663-668, 1975.

Norton, J.P., Identification by optimal smoothing using integrated random walks, *Proc. IEE*, 1232, 451-452, 1976.

Norton, J.P., *An Introduction to Identification*, Academic Press, London, 1986.

Potter, J.M. and B.D.O. Anderson, Partial prior information and decisionmaking, *IEEE Trans. on Systems, Man & Cybernetics*, SMC-10, 125-133, 1980.

Snow, P., Improved posterior probability estimates from prior and conditional linear constraint systems, *IEEE Trans. on Systems, Man & Cybernetics*, SMC-21, 464-469, 1991.

Snow, P., The posterior probabilities of linearly constrained priors and interval-bounded conditionals, *IEEE Trans. on Systems, Man & Cybernetics*, SMC-26, 655-659, 1996.

Soderstrom, T. and P. Stoica, *System Identification*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

Walley, P., *Statistical Reasoning with Imprecise Knowledge*, Chapman & Hall, London, 1991.

White, C.C., *A posteriori* representations based on linear inequality descriptions of *a priori* and conditional probabilities, *IEEE Trans. on Systems, Man & Cybernetics*, SMC-16, 570-573, 1986.

Williams, P.M., Intermediate probabilities, in *Formal Methods in the Methodology of Empirical Sciences*, M. Przelecki, K. Szaniawski and R. E. Wojciki, Eds., D. Reidel, Boston, MA, 229-246,1976.