Proceedings of Symposia in Pure Mathematics
Volume 60, 1997

Inference, Learning, and Recognition

Sanjoy K. Mitter

1. Introduction

I am deeply honored to give this talk on the occasion of the week long conference at M.I.T. to honor Norbert Wiener's many faceted contributions to mathematics and his penetrating use of mathematical ideas to illuminate and even create new fields in electrical engineering, neurobiology, and economics. The subject of my talk is inference, learning, and recognition, topics which have been touched upon in the last two chapters of Wiener's difficult and somewhat controversial book *Cybernetics* [**1**]. These topics are of great current interest today. But in many senses the fundamental problems including a sharp definition of the act of learning and recognition remains as elusive today as they were in Wiener's time. It is therefore not inappropriate to take a new look at Wiener's ideas on learning, self-organization and pattern recognition in light of current research. It is my purpose in this talk to do this by looking at several problems where Wiener's ideas have played an important role, although this may not have been recognized as such.

2. Theory of Filtering

An important problem of inference which Wiener [**2**] and independently Kolmogoroff solved was the problem of filtering a signal which can be observed only in the presence of measurement uncertainty. The highly original contribution of Wiener was to model the situation by representing the signal as a stationary stochastic process and the measurement noise as white Gaussian noise, an appropriate approximation of a wide band stationary process. In many physical problems the measurement noise may be assumed to be independent of the signal. It is worth mentioning that this model plays a key role in analog communication systems where the fundamental problem is the transmission of a signal reliably over a noisy channel.

I would like to interpret this problem in the somewhat more general context of a "recognition" problem where "learning" may have to take place through a process of "adaptation." There are four fundamental elements in all such recognition problems: an internal world not amenable to direct perception, an external world, a sensory apparatus through which the internal world interacts with the external

---

1991 *Mathematics Subject Classification.* Primary probability theory and stochastic processes (44), Systems theory; control (60), and Information and communication; circuits (61).

This research has been supported by the Army Research Office under Grant DAAL03-92-G-0115 (Center for Intelligent Control Systems).

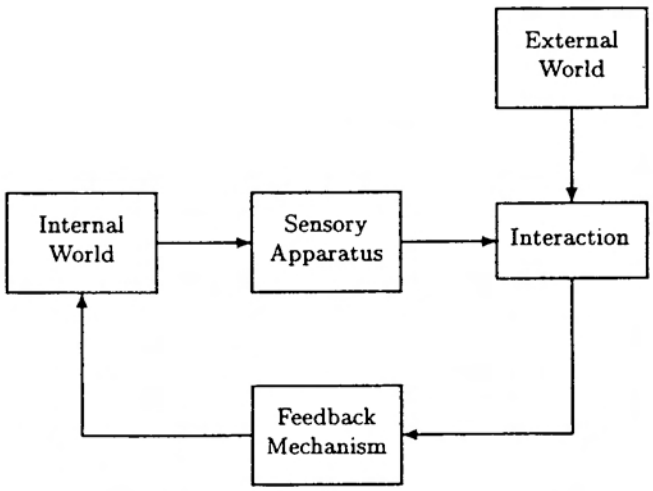

FIGURE 1. Feedback mechanism relationships.

world and a feedback mechanism which feeds back the result of the interaction to the internal world (see Figure 1.).

The process of recognizing the internal world through its interaction with the external world apparently requires [1] representation and coding of the *a priori* knowledge about the external world in the internal world, [2] modeling the sensory apparatus, [3] modeling the external world, [4] modeling the interaction between the internal and external world, [5] representing and coding of the signals resulting from the interaction, and [6] the feedback process. It is through the feedback process that adaptation and learning eventually takes place. There is one final point which requires attention; namely, the choice of criteria to determine the quality of the recognizing process. This is an essential part of recognition.

It is instructive to examine Wiener's solution of the filtering problem in the above conceptual framework. Here the internal world is modeled as a stationary stochastic process $\{z_t \mid t \in \mathbb{R}\}$. Note that in the representation the invariance of the process under the bilateral shift is captured. The external world is modeled as white Gaussian noise $\{v_t \mid t \in \mathbb{R}\}$, interacting additively with Z_t leading to the observed signal

$$y_t = z_t + v_t, \qquad t \in \mathbb{R} \tag{2.1a}$$

and a further property of the interaction is captured through the independence of z_t and v_t as processes.

In this case there is no feedback to the internal world. Noting that white noise v_t is the distributional derivative of the Wiener process w_t, the interactive model represented by Eq. (2.1a) can be rewritten (by redefining symbols) as:

$$y_t = \int_{-\infty}^{t} z_s \, ds + w_t \tag{2.1b}$$

The problem of inference (recognition) in this case consists of estimating z_s from knowledge of the observation $\{y_s \mid s \in (-\infty, t]\}$. Wiener essentially made the further assumption that $\{z_s \mid s \in \mathbb{R}\}$ is a Gaussian process and chose as criteria for determining the quality of the estimate as:

$$E\left(z_t \mid F_t^y\right), \tag{2.2}$$

where F_t^y represents the σ-field generated by the process $\{y_s \mid s \in (-\infty, t]\}$ and $E(\cdot \mid \cdot)$ denotes conditional expectation. The solution to this problem can be obtained by the Projection Theorem and leads to the characterization of the estimate $\hat{z}_t := E(z_t \mid F_t^y)$ as:

$$\hat{z}_t = \int_{-\infty}^{t} h(t-s)\, dy_s. \tag{2.3}$$

The integral here is a stochastic integral and the linear filter is represented by the convolution kernel $h(\cdot)$ which is obtained through the solution of the celebrated Wiener-Hopf equation.

If we restrict ourselves to the category of filtering problems consisting of a stationary Gaussian process in additive white noise, then this linear filter is a "universal filter." The complexity of the filter in the sense of characterizing $h(\cdot)$ however may be large which is what one has to pay for universality. Another way of saying this is that if the process $\{y_s \mid s \in \mathbb{R}]\}$ has a rational spectral density, then the filter, conceived of as an independent entity has no obvious way of recognizing this and thereby reducing its complexity. It would appear that this fact of rational spectral density which is part of the *a priori* knowledge about the external world has to be coded or built-in in the estimator. The question of what information needs to be built-in and what information can be learned is one of the major open questions in the field.

One can prove that every stationary multivariate Gaussian process $(z_t \mid t \in \mathbb{R})$ with a rational spectral density can be prepared as the output of a linear stochastic dynamical system

$$\begin{cases} dx_t &= Ax_t\, dt + B\, dw_t \\ z_t &= Cx_t \end{cases} \tag{2.4}$$

where A, B, C are matrices of appropriate dimensions. w_t is standard Brownian motion with m components say. Stationarity requires that A be a stability matrix. In general there are many such representations and the problem of classifying all canonical representations is well understood [**3**]. The process $(x_t \mid t \in \mathbb{R})$ is now a Gauss-Markov process and locality has been built-in in the representation. The credit for introducing this representation belongs to Kalman [**4**]. We now have a universal linear filter in this more restrictive category, but more importantly the complexity of the filter is now manageable in the sense that the filter itself has a finite-dimensional stochastic dynamical representation

$$d\hat{x}_t = A\hat{x}_t\, dt + K\, dv_t \tag{2.5}$$

where $\hat{x}_t := E(x_t \mid F_t^y)$, K is an appropriate matrix which has a characterization (independent of the date y) and $dv_t = dy_t - C\hat{x}_t\, dt$, the so-called innovations process.

Wiener (and Kalman) filtering is a highly successful theory with wide-spread applications. The corresponding theory of non-linear filtering, a universal theory in some sense, is far less successful. Wiener's own approach to this problem is contained in his book *Nonlinear Problems in Random Theory* [**5**] and for a more

recent account see Arveson [**6**]. The approach via stochastic differential equations, which incidentally does not require stationarity assumptions, proceeds as follows.

The observed signal is modeled as

$$y_t = \int_0^t z_s \, ds + w_t \tag{2.6}$$

where $z_s = h(x_s)$ and x_s is a Markov process given as the unique solution of a stochastic differential equation

$$dx_t = f(x_t)\, dt + \sigma(x_t)\, db_t. \tag{2.7}$$

In contrast to the linear situation h, f and σ are non-linear smooth bounded functions of x. If we denote by $\pi_t(dy) := P(x_t \in dy \mid \sigma(y_s \mid 0 \leq s \leq t))$, the conditional distribution of x_t given the observations up to t, then a well known result of non-linear filtering theory (see e.g Davis-Marcus [**7**]) states that $\forall f$ which are bounded, continuous

$$\pi_t(f) = \frac{\rho_t(f)}{\rho_t(1)} \tag{2.8}$$

where $\rho_t(f) = \int f(x)\rho_t(dx)$ and ρ_t satisfies $\rho_t(f) = \rho_0(f) + \int_0^t \rho_s(\mathcal{L}f)\, ds + \int_0^t \rho_s(hf)\, dy_s$ where $\mathcal{L}$ is the generator of the Markov Diffusion Process (x_t). Under suitable assumptions ρ_t has a density with respect to Lebesgue measure and if we denote this density by $u(t,x)$ then $u(t,x)$ satisfies a stochastic partial differential equation (SPDE)

$$du(t,x) = \left(\mathcal{L}_0^* - \tfrac{1}{2}\mathcal{L}_1^2\right) u(t,x)\, dt + \mathcal{L}_1 u(t,x) \circ dy_t \tag{2.9}$$

where $\mathcal{L}_0^*$ is the formal adjoint of the generator of the Markov diffusion process x_t and $\mathcal{L}_1$ is the operator which is multiplication by $h(\cdot)$ where the SPDE is interpreted in the Stratanovich sense. This SPDE is a universal filter in the sense that any estimate can be computed from this through integration.

In the linear stationary situation, by restricting the covariance function of the external signal $h(\cdot)$, the complexity of the filter is effectively reduced in the sense that the filter has a finite-dimensional representation in the non-linear situation. It is an open problem as to how the externally observed signal should be represented so that the filter representation has manageable complexity. To sharpen this discussion, consider the Lie algebra generated by the operators $\mathcal{L}_0^* - \frac{1}{2}\mathcal{L}_1^2$ and $\mathcal{L}_1$. In the linear situation this Lie algebra is finite-dimensional and is related to the Oscillator algebra (for a discussion of these ideas see Mitter [**8**], [**9**] and the references cited there). This Lie algebra measures the complexity of the filter in the linear situation. Generalizing from the linear situation we define a finite-dimensional non-linear filter for a statistic η_t corresponding to $u(t,x)$ as a stochastic dynamical system.

$$\begin{cases} d\xi_t &= \hat{f}(\xi_t)\, dt + \hat{\sigma}(\xi_t) \circ dy_t \\ \eta_t &= \hat{h}(\xi_t) \end{cases} \tag{2.10}$$

where ξ_t resides in a smooth finite-dimensional manifold and $\eta_t = \int \eta(x) u(t,x)\, dx$. We may consider the representation Eq. (2.10) to be minimal in the sense of non-linear system theory. It is an Ansatz of Brockett [**10**] that for there to exist a finite-dimensional filter there must exist a homomorphism between the Lie algebra of operators generated by $\mathcal{L}_0^* - \frac{1}{2}\mathcal{L}_1^2$ and $\mathcal{L}_1$ and the Lie algebra of vector fields generated by the vector fields X_f and X_σ. This result has been rigorously proved

in certain situations. In the case that $x_t = b_t$, standard Brownian motion and $h(x) = x^3$ the Lie algebra of operators is the Weyl algebra and in this case the above Ansatz can be rigorously proved (this result was originally conjectured in [**8**]; for an elegant proof see Stafford [**11**]). It is believed but not rigorously proved that the class of non-linear problems with finite-dimensional filters forms a hypersurface (has zero measure with the complementary set having full measure) in the class of all non-linear filtering problems.

It is worthwhile analyzing this situation further. The difficulty lies with starting from a representation of the signal z_t as a non-linear function of the Markov Diffusion process x_t. This is a *global* and *universal* representation (under appropriate conditions on f and σ) and has a built-in inherent complexity. It is not difficult to build a *universal* filter but the complexity of representation of the *a priori* knowledge is transmitted in a loss-less way to the universal representation of the filter. To make progress on this problem, the global information pertaining to a non-linear system has to be introduced *hierarchically* through *multiple representations* of the *a priori* knowledge about the system so that the inherent complexity of the problem becomes manageable. In the process, there may be a loss of optimality but this is inevitable if the complexity question is to be surmounted. For the non-linear filtering problem

$$\begin{cases} dx_t &= (2\tan^{-1} x_t - x_t)\,dt + db_t \\ dy_t &= x_t\,dt + dw_t \end{cases} \tag{2.11}$$

a hierarchical filtering approach is feasible [**12**]. This requires understanding the global qualitative behavior of the control system

$$\frac{dx}{dt} = \big(2\tan^{-1} x(t) - x(t)\big) + u(t) \tag{2.12}$$

and studying the quadratic variation properties of the process x_t.

It would appear that the solution of non-linear filtering problems would ultimately require a deep understanding of the geometry of implicit systems of non-linear differential equations.

3. From Filtering to Learning

In the previous section I have tried to cast the filtering problem in the larger context of a recognition problem. An examination of the elements constituting recognition shows that the internal representation of *a priori* knowledge about the external world (the x process in the filtering context), description of the sensory apparatus (the function $h(\cdot)$), the description of the interaction of the internal world with the external world through the sensory apparatus ($(dy_t = h(x_t)\,dt + dw_t)$) and the representation of the resulting external signal play an important role in the solution. Now for the filtering problem, the description of the *a priori* knowledge is probabilistically complete, that is, we have complete description of the evolution of the "state" $P(x_t \in dx)$ via the stochastic differential equation. The same remark applies to the other elements of recognition. Therefore no learning and adaptation is necessary for the filter to function effectively.

To examine the issue of learning, I shall look at the question of learning from data related to an external signal. I want to remark that this is an extreme situation where no *a priori* knowledge of the internal world producing the data is used. The

justification for doing this is to look at a limit situation in order to understand what can be "learned" purely from data about the external world.

Let me consider the problem of regression. Let $z = (x, y)$, where x is a $\mathbb{R}^n$-valued random variable and y is a real-random variable. We have a joint probability distribution $P(x, y)$ on z which however is unknown. We are given data $\mathcal{D}_N = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$ which is considered to be an independent i.i.d. sampling from $P(x, y)$. The problem is to estimate y from x. The estimator therefore is a function $f : \mathbb{R}^n \to \mathbb{R} : x \mapsto f(x)$ where the function f is required to belong to a suitably rich class $\mathcal{F}$. In practice, this class would be parametrized by some parameter α belonging to a set $\mathcal{A}$. In order to determine the quality of the estimate we postulate a loss function $\Phi(y, f(x, \alpha))$ which for the purposes of this discussion we take to be

$$\Phi\big(y, f(x, \alpha)\big) = \big(y - f(x, \alpha)\big)^2. \tag{3.1}$$

The regression problem is to minimize the functional

$$J(f) = \int \big(y - f(x, \alpha)\big)^2 \, dP(x, y). \tag{3.2}$$

If $P(x, y)$ were known then the optimal estimator is

$$f^*(x) = E(y \mid x). \tag{3.3}$$

When $P(x, y)$ is not known the natural strategy is to replace the function $J(f)$ by the corresponding empirical functional

$$J_N(f) = \tfrac{1}{N} \sum_{i=1}^{N} \big(y_i - f(x_i, \alpha)\big)^2$$

and minimizing $J_N(f)$ over the parametrized class $\mathcal{F}$. This formulation includes current feedforward neural network research where the parametrized class $\mathcal{F}$ corresponds to a neural network parametrization. A question one can ask about this problem is that of consistency in the sense that it is required that for $\forall_\epsilon > 0$,

$$\lim_{N \to \infty} P\left[\sup_{\alpha \in \mathcal{A}} \mid J(f) - J_N(f) \mid > \epsilon\right] = 0$$

Through the work of Vapnick-Cernovinkus [**13**], Pollard [**14**] and Dudley [**15**] and others, necessary and sufficient conditions for consistency are known. Basically, a uniform law of large numbers needs to be proved and a certain combinatorial dimension, the so-called VC-dimension and its finite-dimensionality plays a crucial role. For an analysis of this problem from a statistical point of view highlighting the trade-off between the bias and variance of the error see Geman-Bienenstock-Doursat [**16**] and for rates of convergence results see Stone [**17**]. The basic conclusion here is that for difficult problems of recognition such as recognition of speech and images, the amount of data needed in the training phase is prohibitive for reasonable generalization errors, that is, learning purely from data requires that the learning mechanism sees vast amounts of data before it can perform generalization with reasonable error probability. Besides the complexity of the learning mechanism also has to be suitably large as a function of the data. But then the computational problem of learning becomes insoluble.

It would appear that to make progress in these problems one needs to make appropriate use of prior knowledge of the internal mechanisms which produce the

data. Moreover representation of these internal mechanisms as well as representation of the data are the fundamental problems to be solved. In statistical terminology in the context of the learning problem I have discussed, the incorporation of systematic biases internally compensates for the need to have complex learning mechanisms (networks) and large amounts of data.

4. Approximation vs. Recognition

I have so far discussed the need for internal representations (possibly organized in a layered hierarchical manner) about external signals for the problem of recognition to be solvable with bounded complexity. There remains the question of evaluating the quality of the recognizer by postulating a performance measure (or possibly several performance measures from which a selection needs to take place) on the product space of data and the objects to be recognized. Now, the raw data will often be such that the information needed to perform the recognition is not easily extractable and may require coding or compression at several levels of abstraction. This coding may result in a loss of information in an information-theoretic sense but there may be a gain in "information" from the recognition point of view. In any case this coding may be necessary so that the complexity of the recognition process is manageable and also that the process is robust. To examine this question let me consider after Chernoff [**18**] a classification problem where the classes are described by normal distributions $N(\mu_i, \Sigma_i), i = 1, 2$ where μ_i denotes the mean and Σ_i denotes the covariance of the distribution. Let $\mu_1 = 0$ and $\mu_2^T = (\cdot 2, \cdot 2, \cdot 2)$ and let

$$\Sigma_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & .9 & 0 \\ 0 & 0 & .001 \end{pmatrix} \quad \text{and} \quad \Sigma_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & .001 & 0 \\ 0 & 0 & .91 \end{pmatrix}.$$

Suppose that we want to reduce the dimension of the data space using Principal Component Analysis. In the terminology of principal-component analysis (see e.g. Rao [**19**]) the first principal direction is the x_1 direction. But this direction has little effect if our task is to discriminate between the two classes. The x_1 direction gives us information as to where the action is in terms of the mixture of the two distributions but for classification purposes the space spanned by (x_2, x_3) is more informative, since the $\text{cov}(x_2, x_3)$ is very different for the two populations. The point is that approximation or data compression using the Euclidean metric which plays a fundamental role in Principal Component Analysis is totally inappropriate for classification in this example. The performance measure is to be "induced" by the task at hand which is classification.

For a different example let us consider the following detection problem. Let

$$\begin{cases} h_1(t) = 0.125, & 0.1 \le t \le 0.91 \\ h_1(t) = 0 & \text{otherwise} \end{cases}$$

$$\begin{cases} h_2(t) = 1, & 0.5 \le t \le 0.51 \\ h_2(t) = 0 & \text{otherwise} \end{cases}$$

and

$$\begin{cases} g(t) = 1, & 0.52 \le t \le 0.53 \\ g(t) = 0 & \text{otherwise} \end{cases}$$

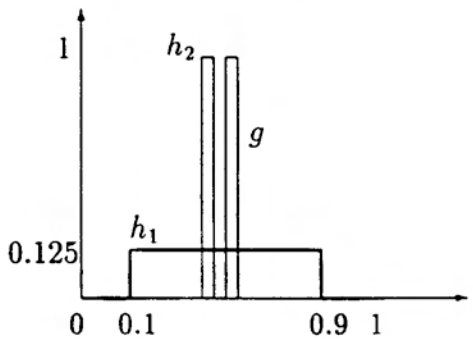

FIGURE 2. Waveforms in detection example.

We think of $g(t)$ as an observed signal and we are required to detect whether $g(t)$ is an uncertain version of $h_1(t)$ or $h_2(t)$ (see Figure 2).

If we postulate that $g(t)$ is an additively perturbed version of $h_1(t)$ or $h_2(t)$ and we use an $\mathcal{L}^2$-criterion to determine whether h_1 or h_2 is the true signal, then h_1 is to be chosen as the detected signal. On the other if we postulate that $g(t) = h_i(\sigma(t)) + n(t)$ where σ is a diffeomorphism, then the following invariant metric

$$d(f_1, f_2) = \inf_{\sigma} \left(\sup_t |\sigma(t) - t| + \sup_t |f_2(t) - f_1(\sigma(t))| \right)$$

will detect h_2 as the signal. The above metric is the Skorokhod metric on the space $D(0,1)$ and is well known in the theory of stochastic processes. But in order to arrive at this metric we need to have prior knowledge about the signal generating mechanism. For a detailed discussion of the role of invariance in the choice of metric see Akra and Mitter [**20**].

The question of a choice of performance measure for recognition arises in the context of character recognition in the form of asking how we measure the similarity of two characters. The mathematical notation of similarity is that of an equivalence relation. In the space of characters this equivalence relation can be imposed by defining a metric on the space of characters.

However a similarity in the mathematical sense need not correspond to perceptual similarity. To discuss this further define a character to be a binary image having a value of 1 (foreground) along the trace of a character and 0 (background) outside. For convenience, define the domain of the character to be the foreground only. Hence a character is a constant function 1 but defined over different domains. Therefore a character can be specified by defining its domain and hence is a set of complex numbers $(z_i)_{i \in I}$. A natural measure of similarity would be that induced by the Hausdorff distance between two sets A and B in $\mathbb{R}^2$:

$$H(A,B) = \max \left[\sup_{a \in A} \inf_{b \in B} \rho(a,b), \sup_{b \in B} \inf_{a \in A} \rho(a,b) \right]$$

where $\rho(\cdot,\cdot)$ is a metric on $\mathbb{R}^2$. However, it is well known that the Hausdorff metric is sensitive to outliers and therefore the distance between an M (ideal character) and an M with a blob above it (see Figure 3) would be large in the Hausdorff

FIGURE 3. Ideal and Noisy Character.

metric. If we think of the M with the blob over it as a noisy version of an ideal M then the Hausdorff metric gives equal weight to both and as we shall see in the next section this may be an incorrect view of the recognition process. On the other hand if we use the one-sided distance between the ideal M and the noisy M, then this distance will be small and the two M's will be considered equivalent.

5. Toward a Theory of Character Recognition

In this section I suggest an outline of a theory of character recognition which addresses many of the ideas raised in the previous sections. Character recognition which may be a simpler problem than image or speech recognition remains largely unsolved. For the details of this theory see Akra and Mitter [**21**]. In this discussion it is assumed that additive noise is removed in a pre-processing stage.

The starting point of this theory is that one is given a set of templates $\mathcal{T} = (T_i)_{i\in I}$ corresponding to representatives of the characters. This set may have to be *learned* from data. As discussed in the previous section this corresponds to a set of complex numbers $T_i = \left(z^j_{T_i}\right)_{j\in J} \subset \mathbb{R}^2$. We also attach a triplet consisting of label, size, and position to each character. Now, to capture the variability and richness of the characters we are likely to see, a set of allowable deformations acting on the templates are introduced. This set of deformations $\mathcal{D}$ is a closed, bounded subset of the set of affine maps $\mathcal{A} : \mathbb{R}^2 \to \mathbb{R}^2$. This subset corresponds to arbitrary compositions of translations, partial rotations, scaling, and slanting. The particular deformation may be character dependent. This is an instance of using *a priori* knowledge to model the external signal generation process. The boundary of the set $\mathcal{D}$ is estimated from data.

The application of a deformation $D \in \mathcal{D}$ on a template $T \in \mathcal{T}$ produces a deformed template $D(T)$. Now consider the problem of minimizing

$$\operatorname*{Min}_{D\in\mathcal{D}} \operatorname*{Min}_{T\in\mathcal{T}} h\big(D(T), O\big) \tag{5.1}$$

Here $h(\cdot,\cdot)$ denotes one-sided Hausdorff distance and O is the observed character. The minimizing T^* is the template which is a candidate for the character which is similar to the observed character. Note that one-sided Hausdorff distance essentially consists of checking set containment and therefore if we see an observed character such as a distorted E, both an I and F will be contained in it and these two ideal characters will also be found as candidate templates when the minimization problem Eq. (5.1) is solved. The main computational problem is a minimization of one-sided Hausdorff distance. This is a well-known problem in computational geometry [**23**].

In the use of the one-sided Hausdorff distance as a measure of similarity, we *break the symmetry* by weighting the templates more than the data. This has the effect of reducing the amount of training data needed for recognition. We are in effect postulating that the *recognition process is a search for what we already know is what we observe.* This is an *internalist* view. There remains the question of resolving the non-uniqueness inherent in the solution of the minimization problem Eq. (5.1). At this level of representation this is solved using the Minimum Description Length Principle of Rissanen [**24**]. That is, we choose the template which requires the minimum number of bits to encode it. It can be shown that in this instance this would result in the choice of E as the template which is similar to the distorted $\mathcal{E}$. There may be situations where this ambiguity may have to be resolved at the word or concept level.

In this theory the recognition process is a layered hierarchical process representing different levels of abstraction. Sets of points are recognized as characters, sets of characters as words, and sets of words as concepts. I have described some of the principles of recognition at the character level. The same principles apply at the word and concept level and suitable feedback mechanisms which regulate the bi-directional information flow between the levels are inserted. Details of these ideas can be found in the previously cited paper of Akra and Mitter and in the patent application [**25**].

6. Conclusions

The main thesis of my talk is that the problem of recognition should be viewed as a layered hierarchical process where prior knowledge about the signal generating process is represented hierarchically at different levels of abstraction. The process of recognition is itself a hierarchical process where recognition is done at each level in a non-unique way using criteria appropriate to that level. If possible, ambiguities are resolved at a particular level using the Minimum Description Length Principle. The recognition at a particular level may be carried out under the supervision of higher levels. The information about the results of recognition at a particular level is then transmitted to higher levels possibly in coded form and the recognition at the higher levels is done conditional on the recognition at lower levels. The process is therefore one where multiple feedback loops are an integral part of the recognition process.

We adopt an internalist perspective where greater weight is placed on the templates and their space of deformations than on the data. Representation of the data is important but we restrict processing to essentially noise removal. This has the effect of reducing the size of the data needed for learning.

References

[1] Wiener, N: *Cybernetics*, Technology Press and Wiley, New York, 1948.

[2] ____: *Extrapolation, Interpolation and Smoothing of Stationary Time Series*, Technology Press and Wiley, New York, 1949.

[3] Lindquist, A. and Picci, G.: *Realization Theory for Multivariate Stationary Gaussian Processes*, tech. report, TRITA-MAT-1984-7, Dept. of Mathematics, Royal Institute of Technology, Stockholm.

[4] Kalman, R.E.: A New Approach to Linear Filtering and Prediction Problems, *J. of Basic Engineering* (Trans. ASME Ser. D) Vol. 83, pp. 35-45, 1960.

[5] Wiener, N.: *Nonlinear Problems in Random Theory*, The Technology Press of M.I.T. and Wiley, 1958.

[6] Arveson, W.: *Nonlinear Spectral Theory with Applications to Nonlinear Noise*, unpublished manuscript.
[7] Davis, M.H.A. and Marcus, S.J.: An Introduction to Nonlinear Filtering, in *Stochastic Systems: The Mathematics of Filtering Identification and Applications*, Reidel, 1981.
[8] Mitter, S.K.: On the Analogy Between Mathematical Problems of Nonlinear Filtering and Quantum Physics, *Ricerche di automatica*, Vol. x, No. 2, pp. 163-206, 1979.
[9] ____: Geometric Theory of Nonlinear Filtering, *Outils et Modèles Mathématiques pour L'Automatique, L'Analyse de Systèms et le Traitement du Signal*, Editions du CNRS, 1983.
[10] Brockett, R.W.: Remarks on Finite Dimensional Nonlinear Estimation, in *Analyse de Systèms Asterisque*, Vol. 75-76, pp. 199-205, 1980.
[11] Stafford, J.T.: The Weyl Algebra and Finite Dimensional Filtering, *Stochastics*, vol. 14, pp. 29-31, 1984.
[12] Mitter, S.K.: A Hierarchical View of Nonlinear Filtering, to appear.
[13] Vapnik, V.M.: *Estimation of Dependencies Based on Empirical Data*, Springer-Verlag, New York, 1982.
[14] Pollard, D.: *Convergence of Stochastic Processes*, Springer-Verlag, Berlin, 1984.
[15] Dudley, R.M.: *Uniform Central Limit Theorems*, M.I.T. Course Notes of 18.177, 1992.
[16] Geman, S., Bienenstock, E., and Doursat, R.: Neural Networks and the Bias/Variance Dilemma, *Neural Computation*, **4**, pp.. 1-58, 1992.
[17] Stone, C.J.: Optimal Global Rates for Convergence of Nonparametric Regression, *Annals of Statistics*, **10**, pp. 1040-153, 1982.
[18] Chernoff, H.: unpublished manuscript, 1977.
[19] Rao, C.R.: *Linear Statistical Inference and its Applications*, Wiley, New York, 1973.
[20] Akra, A. and Mitter, S.K.: Waveform Recognition in the Presence of Domain and Amplitude Noise, accepted for publication *IEEE Trans. on Information Theory.*
[21] ____: A New View of Optical Character Recognition, to appear.
[22] Preparata, F.P. and Shamos M.I.: *Computational Geometry: An Introduction*, Springer-Verlag, New York, 1988.
[23] Grenander, U.: *Pattern Analysis, Vol. 2 of Lectures in Pattern Theory*, Springer-Verlag, New York, 1978.
[24] Rissanen, J.: *Stochastic Complexity in Statistical Enquiry*, World Scientific, River Edge, New Jersey, 1989.
[25] ____: U.S. Patent Application No: 08/254,938, Filed June 7, 1994.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY, CAMBRIDGE, MA 02139
Current address: Department of Electrical Engineering and Computer Science, Laboratory for Information and Decision Systems, and Center for Intelligent Control Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139
E-mail address: `mitter@lids.mit.edu`