Some comments on cybernetics and control

Kickert, W.J.M.; Bertrand, J.W.M.; Praagman, J.

Published in:
IEEE Transactions on Systems, Man and Cybernetics

DOI:
10.1109/TSMC.1978.4309868

Published: 01/01/1978

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Some Comments on Cybernetics and Control
WALTER J. M. KICKERT, JAN-WILLEM M. BERTRAND, AND JAAP PRAAGMAN

Abstract—The theory of cybernetics as introduced by Ashby and developed by Ashby and Conant will be analyzed and commented upon. Ashby's law of requisite variety and, in particular, the underlying measure of optimality—the quantity of entropy—are examined. Next the cybernetic theorem of error control and cause control is observed, and finally the cybernetic theorem of the necessity of modeling for regulation is studied. In all three cases several practical conditions and restrictions for the applicability of the theorems to control engineering are pointed out.

I. INTRODUCTION
Since the first introduction of cybernetics by N. Wiener [1], as a "science of communication and control," numerous contributions to this field have been made. One of the most outstanding contributions, in our opinion, has been made by the late W. R. Ashby, who laid a basis for the link between information and control with his well-known law of requisite variety [2]. Both he and R. C. Conant further developed this theory. Some important and well-known elements of this development are their views on error-controlled regulators versus cause-controlled regulators [3], and their views on modeling as a necessary part of regulation [4]. These elements only represent part of their work, especially in view of their recent achievements in the field of complex and hierarchical systems [5]-[8]. The analysis and comments in this correspondence, however, will be restricted to the above-mentioned points, mainly because of their relevance for control.

There are three reasons which emphasize the necessity of analysis, comments, and criticism. First is the obvious importance of this branch of cybernetics for control theory. It will be clear that information and communication play an important role in control and that, more specifically, the above-mentioned issues are certainly very important for control theory. The second reason is the incomprehensible ignorance of this theory on the part of control engineers. It is astonishing how little attention the theory of communication, and in particular, the theory of cybernetics has received in control theory, though its importance is quite clear. The third reason is the unshakable popularity of this theory among system theorists. One might hope that some "cross talk between cyberneticians and control engineers" as Porter calls it [9] will result in a greater appreciation by control engineers of Ashby's link between information and control, and in a somewhat restricted but better founded popularity among system theorists.

II. THE LAW OF REQUISITE VARIETY [2], [3], [10]
The most famous law of cybernetics is undoubtedly Ashby's law of requisite variety. This very general law, which contrary to the usual control theory does not presuppose linearity, low-order structure, etc., gives an upper limit to the degree of controllability that a regulator can possibly achieve. In view of the generality of the law and its obvious importance for control, which can easily be shown in numerous examples, it is indeed astonishing that so little attention is paid to this law by control theorists.

The law of requisite variety states that the capacity of a device as a regulator cannot exceed its capacity as a channel of communication, or to put it in Ashby's words: "only variety in the regulator can force down the variety due to the disturbances; only variety can destroy variety." Imagine a system composed of a set $D$ of disturbances, a set $R$ of control actions, and a set $Z$ of outcomes, defined by a mapping $\phi: D \times R \rightarrow Z$. This obviously represents a control system. By taking finite discrete sets $D$, $R$, and $Z$ and by visualizing $\phi$ as a table, it can easily be shown that the goal of keeping the outcome $Z_k \in Z$ constant, that is, of decreasing the variety in the outcomes, can only be met by a corresponding increase in the variety of $R$.

The relation between information and control is essentially the following. As the criterion for the success of the regulator (the goal is constancy of the outcomes $Z$) Shannon's measure of selective information in a signal, the entropy $H(Z)$ is introduced:

$$H(Z) = - \sum_{z \in Z} p(z) \log_2 p(z).$$

Ashby states that optimality of a regulator $R$ is equivalent to the minimization of the entropy $H(Z)$ of the outcomes. One of the advantages of this measure of optimality is that it does not presume numerical variables; entropy also applies to variables that can only be classified (nominal).

Strictly speaking, the use of this entropy measure implies that the variables involved (outcomes) are stochastic. (Notice that the requirement of stochastic system variables does not imply that the system relations be stochastic; on the contrary, most of this cybernetic theory leads to the necessity of deterministic systems relations.) Most frequently, however, the entropy measure is applied as a measure of variety without strict probability density functions. It then serves as a measure of the number of possible alternatives. The assumption behind this use of entropy is that all alternatives have the same probability:

$$H(Z) = - \sum_{z \in Z} p(z) \log_2 p(z) = - \sum_{z \in Z} \frac{1}{n} \log_2 \frac{1}{n} = \log n.$$

This probability assumption is often omitted. Hence, in fact, the entropy measure is not only used in case of stochastic variables but also with (varying) deterministic variables.

It should be remarked that although the use of entropy has an advantage over classical control theory in that it incorporates stochastic variables, it does not solve the kind of problems that are solved by classical control theory, simply because entropy does not deal with them. Although stability analysis, transfer function theory, etc., do not exclude the existence of stochastic signals, those theories just do not consider it; they deal with analytical functions in time (pulse, step, ramp, sinusoid, transient response, steady state, etc.). In contrast, entropy only considers aggregates of variables, such as density functions and varieties. Hence, a great deal of control theory is not covered by this theory of cybernetics.

Secondly, it seems questionable to equate optimality with minimal entropy. A well-known fact of information theory is that the
entropy $H(Z)$ is minimal when there is one $i$ such that $p(z_i) = 1$ and $p(z_j) = 0$, $\forall j \neq i$, and that $H(Z)$ is maximal when all $p(z_i)$, $z_i \in Z$ are equal. Minimizing $H(Z)$ means compressing the probability distribution of $Z$. Hence, whatever the position of $z_i$, $H(Z)$ will decrease if the probability distribution is compressed around $z_i$ (see Fig. 1). This means that entropy only measures variety and does not take into account the absolute position of the optimal variable. This omission has serious implications in questions of steady-state errors, bias, etc.

Thirdly, some attention should be paid to the concept of variety, for which entropy is a measure. A good regulator should minimize the variability of the output. Apart from the possibly additional requirement of a particular output value—which entropy does not guarantee—this requirement usually means that an output variable should fluctuate as little as possible. The usual measure for this is variance. However, it should be emphasized that between these two there is a difference of practical relevance, and not just of mathematics. Let us try to illustrate this difference through an example.

**Example:** Imagine a control system with a disturbance set $D = \{a, b, c\}$, a control action set $R = \{p, q, r\}$, and an output set $Z = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. The transition table of $\Phi$: $D \times R \rightarrow Z$ is

\[
\begin{array}{ccc}
  & a & b & c \\
p & 1 & 9 & 4 \\
q & 5 & 2 & 8 \\
r & 7 & 6 & 3 \\
\end{array}
\]

Let the three values of $D$ be equally likely. Then the measure of variety, the entropy, of $D$ is

\[H(D) = \log 3.\]

Suppose we take the following two control strategies: 1) maintain a fixed control action $R^* = p$, and 2) relate the control action to the disturbance as follows:

\[
a \rightarrow q \\
b \rightarrow r \\
c \rightarrow p.
\]

The resulting outputs will be, respectively,

1) $Z = \{1, 4, 9\}$,
2) $Z = \{4, 5, 6\}$.

The entropy of both outputs is $H(Z) = \log 3$. Hence, according to this measure, neither of the control strategies suppresses the variety in the disturbances at all, and no difference exists between the static 1) and dynamic 2) control strategy. But supposing the output values admit a numerical interpretation and taking the variance as a measure of variability, it will be clear that control strategy 2) is far better than strategy 1). The variance of $(1, 4, 9)$ is much larger than that of $(4, 5, 6)$, irrespective of any desired mean value.

Thus entropy is neither a measure for the mean value of the fluctuating variable, nor a measure for the variance around this mean value. Entropy is a measure for variety, that is, the number of possibilities weighted according to their probabilities of occurrence.

This remark represents quite a serious objection against Ashby’s measure of entropy in case of a numerical output variable, so let us examine this somewhat more formally. Entropy is defined as

\[H(Z) = -\sum_{z_j \in Z} p(z_j) \log p(z_j).\]

while variance is defined as

\[\text{var}(Z) = \sum_{z_j \in Z} p(z_j)(z_j - \bar{z})^2.\]

These definitions show the above-mentioned difference: entropy is a nominal measure, whereas variance is an interval (metric) measure. These reflections lead us to the conclusion that because entropy is indifferent to distances between $z_i$, it is generally not a measure of variance.

There still remains the interesting question of the conditions under which the entropy is a measure of variance. In other words, what properties should the probability density function $p(Z)$ have, so as to permit conclusions from entropies about variances. Let us start by requiring order preservation:

\[H(z_1) > H(z_2), \quad \text{iff var}(z_1) > \text{var}(z_2).\]

A sufficient condition for order preservation is that a strictly monotonic mapping exists between entropy and variance. Now it is clear from proposition 20.5 of Shannon [11] that when order preservation is required we have to restrict ourselves to certain classes of distribution functions, as the proposition implies that for any arbitrary non-Gaussian distribution, there exists a Gaussian distribution with the same variance but higher entropy. Classes of distributions for which order preservation does hold are the Gaussian, uniform, and exponential distribution. For these classes the entropy equals, respectively,

\[\log \sigma_1 \sqrt{2\pi e}, \quad \log \sigma_2 \sqrt{12}, \quad \log \sigma_3 \cdot e\]

where $\sigma_z$ is the variance. From these results it also follows that there is no class of distribution functions for which anything stronger than order preservation holds.

In summary, it can be stated that minimal entropy may sometimes be a necessary, but surely will not often be a sufficient condition for optimality.

III. ERROR-CONTROLLED REGULATORS [3], [12]

A diagram such as Fig. 2 should be quite familiar to control engineers, for it is the representation of the well-known form of regulation by error control widely used by control engineers, e.g., in servomechanisms. It is surprising to note that a huge body of theories and techniques has evolved around this error-controlled regulator—in fact most of classical control theory deals with it whereas in cybernetics it can easily be shown that it might well be better not to control by the error but by what gives rise to the error. It can be proven, namely in cybernetics, that such an error-controlled regulator can never regulate perfectly, that is, can never keep $Z$ constant. The fundamental property of the error-controlled regulator, that it can never be perfect, is hardly
CORRESPONDENCE

Fig. 2. Error-controlled regulator.

Fig. 3. Cause-controlled regulator.

considered by control engineers. Simply stated, the argument is that because R counters D only by learning of it from Z, which has already been affected by D, complete success is impossible. Of course the proof of this common-sense observation is not so trivial [12]. The deduction assumes that the transformation of Z to R is represented as a finite state machine so that if the entropy of Z is zero, it will be passed on to R, that is, the entropy of R also becomes zero. Because \( \phi: D \times R \rightarrow Z \) is the mapping \( Z = D - R \), it can be proven that the case of complete control, \( H(Z) = 0 \), can only occur when the input \( D \) is constant too, that is, when \( H(D) = 0 \).

In contrast to an error-controlled regulator which can never attain complete success (as has been proved), a cause-controlled regulator, as visualized in Fig. 3, is able to perfectly regulate \( Z \) because the regulator reacts directly to the disturbances which affect the system. This is the type of regulation in which the regulator anticipates \( S \), so that the regulatory action is simultaneous with that of \( S \). In cybernetic terms, the job of the regulator is to block the flow of information from \( D \) to \( Z \).

Although an error-controlled regulator may be inferior in principle to a cause-controlled regulator, a well-known practical fact is that the success of the feedforward prediction in the latter scheme completely depends on the adequacy and validity of the system’s model in the regulator. Slight differences between model and reality or changes in time of the real process will result in an accumulation of errors, which might result in instability (in the Lyapunov sense of the term). Moreover, in the next section it will be shown that the statement that a cause-controlled regulator can be optimal should be somewhat more specified.

**IV. THE NECESSITY OF MODELING IN REGULATION [4]**

Although most control engineers will start their design of a control system for a complex dynamic process by making a model of the process, this modeling is often regarded as optional. Making a model has the intuitive appeal of being helpful, but it is quite possible that other methods of design without any model might do as well or even better, in which case the making of a model would be a waste of time. In other words, as long as it is not proven that modeling is a necessary part of regulation, the usefulness of this often difficult and time-consuming activity remains doubtful. It is the great merit of Conant and Ashby that they have proved that modeling is not only a helpful but a strictly necessary part of regulation. They show that any regulator which is both optimal and simple must be isomorphic with the system being regulated (see their paper for a precise definition of the

terms used in this statement). Take Fig. 3 as the general configuration of a control system, with

\[
\rho: D \rightarrow R \\
\sigma: D \rightarrow S \\
\phi: S \times R \rightarrow Z.
\]

Given the disturbances \( D \), the control actions \( R \), the system variables \( S \), and the outcome variables \( Z \), their modeling theorem is the simplest optimal regulator produces control actions \( R \), related to \( S \) by a mapping \( h: S \rightarrow R \) so that for all \( d \in D: \rho(d) = [\rho(d)] \).

There is an aspect of their proof that is somewhat artificial: they first assume a probability distribution \( p(S) \) and a conditional distribution \( p(R|S) \) specifying the behavior of the regulator. Although a conditional distribution \( p(R|S) \) does not necessarily imply any causal relationship between \( R \) and \( S \), this notation is the usual one for a stochastic system relation. Their proof might therefore be misinterpreted as suggesting that \( S \) affects \( R \) (contradicting their configuration of a cause-controlled regulator) or that the regulator itself is a stochastic system. In view of their conclusion about the deterministic relation \( h: S \rightarrow R \), this seems artificial (compare with the proof given further on).

A more relevant remark to be made here is that formally they proved only that a simple optimal regulator must be a model. They did not prove that a model automatically is the simplest optimal regulator, and that is the way in which the control engineer would like to work. So let us examine the cause-controlled regulator somewhat more closely.

Assume the control system configuration of Fig. 3; then the following holds:

\[
\]

Because there is a mapping \( \phi: S \times R \rightarrow Z, H(Z|S, R) = 0 \), so that

\[
\]

By definition \( H(R|S) \geq 0 \), and thus \( H(Z, S, R) \geq H(S) \). The equality \( H(Z, S, R) = H(S) \) occurs if \( H(R|S) = 0 \), that is, if there is a mapping \( h: S \rightarrow R \). On the other hand, the following is also an identity:

\[
\]

Because entropy is nonnegative, \( H(Z) \leq H(Z, S, R) \) always holds. Define the multiplicity \( k \) of the \( (\text{many-to-one}) \) mapping \( \Phi: R \times S \rightarrow Z \) as the largest number of pairs \( (r_i, s_j) \in R \times S \) which map to a same \( z \in Z \). Call \( K \) the binary log of \( k \). Then \( H(R, S|Z) \leq K \). Once \( \Phi \) is defined, \( K \) is a fixed value. Thus

\[
H(Z) \geq H(Z, R, S) - K \geq H(S) - K,
\]

while also, generally, \( H(Z) \leq H(Z, S, R) \) with \( H(Z) \leq H(S) \) in case \( h: S \rightarrow R \) exists. This results in a dual inequality

\[
H(S) \geq H(Z) \geq H(S) - K,
\]

where the left-hand inequality assumes the existence of a mapping \( h: S \rightarrow R \) and the right-hand inequality is general.

A few remarkable conclusions can be drawn from this inequality. If the mapping \( \Phi: R \times S \rightarrow Z \) is one-to-one, then \( K = 0 \), so that \( H(Z) \geq H(S) \). Two cases can be distinguished in this inequality:

\[
H(Z) \geq H(S) \text{ if no mapping } h: S \rightarrow R \text{ exists, that is, if the control action } R \text{ varies independently of } S \text{ at least to some degree,}
\]
produce an optimal output. The conclusion is that the set of optimal regulators is a subset of the set of model regulators.

By adopting a different definition of optimality, namely, that of attaining a prescribed goal, a very simple alternative proof of the hypothesis that the simplest optimal regulator must be an isomorphism \( h: S \rightarrow R \) can be given [13]. Given the system mapping \( \Phi: R \times S \rightarrow Z \), the goal \( G \) can be viewed as a subset of the total set of outcomes: \( G \subset Z \). The measure of optimality here is whether or not \( z \in G \). In fact, this is a measure of effectiveness. To ensure that \( z \in G \), it is necessary that the pair \( (r, z) \in \phi^{-1}(G) \). Here \( \phi^{-1}(G) \) is defined to be \( (r, s) \mid (r, s, z) \in \phi \land z \in G \). Evidently \( \phi^{-1}(G) \subset R \times S \), which exists everywhere under the assumption that the system is controllable. Thus defined, this is a relation which, however, is not necessarily single valued. The regulator can choose from various possible control actions. The simplest regulator will be the mapping \( h: S \rightarrow R \) with \( h \in \phi^{-1}(G) \). This deduction of the simplest optimal controller is much more simple than that of Conant and Ashby; actually it is trivial.

We will end this section with some general remarks about the simplest controller. The theorem which states that when a regulator is optimal there must exist a mapping from process variables to control variables appears to be of little practical relevance for classical control engineers. For instead of working with instantaneous values of time variables as in cybernetics, where mappings such as \( \sigma \) are generally not one-to-one and consequently \( p \sigma^{-1} \) is generally not a mapping, classical control theory works with variables in the Laplace domain. In this domain the mappings—the transfer functions—are all deterministic and one-to-one, so that the systems mapping \( \sigma \) and the regulators mapping \( \rho \) can always be transformed into a “model mapping” \( h: \rho \cdot \sigma^{-1}: S \rightarrow R \). Every controller would then be a model. Because the use of Laplace transforms presupposes that all initial conditions of the differential equations are equal to zero, one can even state that in the time domain the time function mappings of classical control theory are one-to-one. Note that in that time domain one works with mappings between time series and not with instantaneous values of variables like in cybernetics.

Secondly, it should be noted that the regulator \( \rho \) is a model of the subsystem \( \sigma \), and not of the whole system from disturbance up to the output. It seems obvious that a sensible control engineer wants to predict, and hence to model, the whole mechanism from disturbance to output before designing a controller. The system part consisting of the transformation \( \phi: R \times S \rightarrow Z \) should be incorporated in the model. Strongly related to this point is the question of the meaning which should be attached to both subsystems \( \sigma \) and \( \phi \). Much depends on the interpretation of these abstract blocks.

Thirdly, the exact meaning of the statement that “every good regulator must be a model” should once more be emphasized. It does mean that optimality implies modeling, but it does not mean that modeling implies optimality. In other words, the class of optimal regulators is only a subset of the class of model regulators. The condition of modeling for optimal regulation is necessary but not sufficient. In practice this means that designing a regulator indeed must start by modeling, but that not every model leads to optimal regulation.

V. Conclusion

It has been shown that the basis of the cybernetic theory of information and control, the concept of entropy, has some practical limitations as a measure of optimality. Furthermore, the theorem that a cause-controlled regulator can be completely successful as opposed to an error-controlled regulator, fails to deal
with the problem of stability. The theorem that every regulator must necessarily consist of a model of the system to be regulated has been shown to be of less practical use than it seems.

These comments definitely do not imply that those theorems are useless. On the contrary, they represent the first valuable steps towards a fruitful and practical link between the two well-developed theories of information and control. It may be that this theory of cybernetics will shed a new and useful light on the development of control systems.

REFERENCES


WLODZIMIERZ GREBLICKI

Abstract—Modified class conditional density estimates for pattern classification obtained by replacing the sample sizes for particular classes by the overall sample size in expressions for the original estimates are presented, and their consistency is proved. Pattern recognition procedures derived from original and modified Rosenblatt–Parzen, Loftsgaarden–Quesenberry, and orthogonal series estimators are given, and Bayes risk consistency is established.

I. INTRODUCTION

Pattern recognition algorithms with nonparametric density estimates have been studied by several authors. The Rosenblatt–Parzen estimate has been used by Van Ryzin [13], [14], while Devroye and Wagner [3] have employed both the Rosenblatt–Parzen and the Loftsgaarden–Quesenberry estimates. Van Ryzin [14] has also examined procedures with orthogonal series estimates.

In this correspondence we present pattern recognition procedures with the Rosenblatt–Parzen, the Loftsgaarden–Quesenberry, and the orthogonal series class density estimates of their original forms, and we introduce modified class density estimates and derive appropriate procedures. The modified estimates are obtained by replacing the sample sizes for particular classes by the overall sample size in the expressions for the original estimates. We prove Theorems 4 and 7 on consistency of the modified class density estimates, and then using general the Greblicki [7] and the Wolpert–Wagner [16] theorems on Bayes risk consistency, we establish Theorems 5 and 8 on asymptotical optimality of so obtained procedures. We also show that procedures introduced by Van Ryzin [13], [14] and Devroye and Wagner [3] can be derived from either original or modified class density estimates.

II. PRELIMINARIES

Let \( \Omega = \left\{ 1, \cdots, M \right\} \); elements of \( \Omega \) will be called classes. Let \( (\Omega, X) \) be a pair of random variables. \( \Omega \) takes values in \( \Omega \), and \( p_i = P(\Omega = i) \). \( X \) takes values in \( R^p \), and \( f_i \) is the class conditional density, i.e., the conditional density of \( X \) given the class \( i \). \( L_i(x) \) is the loss we incur in taking action \( i \in \Omega \) when the class is \( j \).

We assume the 0–1 loss function. For a decision function \( \psi \), i.e., for a function mapping \( R^p \) into \( \Omega \), the expected loss is

\[
R(\psi) = \sum_{j=1}^{M} p_j \int L(\psi(x), j) f_j(x) \, d\mu(x)
\]

where \( \mu \) is the Lebesgue measure on \( R^p \). A decision function \( \psi_0 \) which classifies every \( x \) as coming from any class \( i \) for which

\[
p_i f_i(x) = \max_j p_j f_j(x)
\]

is a Bayes decision function. All the class densities as well as the class prior probabilities are assumed to be unknown and will be estimated from the learning sequence

\[(\Omega_1, X_1), \cdots, (\Omega_n, X_n),\]

i.e., a sequence of \( n \) independent observations of the pair \((\Omega, X)\). Let \( p_i = N_i / n \), where \( N_i \) is the number of observations from the class \( i \), be an estimate of \( p_i \), and let \( f_i(x) \) be an estimate of \( f_i(x) \).

The \( \Omega \)-valued function \( \psi_n \), defined for all \( x \in R^p \) and all realizations of the learning sequence is called an empirical decision function. Throughout this correspondence we are concerned with pattern recognition procedures that are sequences \( \psi_n \) of empirical decision functions classifying every \( x \) among any class \( i \) for which

\[
\tilde{f}_i(x) = \max_j f_j(x), \quad n = 1, 2, \cdots.
\]

We say that the procedure is Bayes risk consistent if

\[
\lim_{n \to \infty} ER(\psi_n) = R(\psi_0)
\]

where \( \psi_0 \) is any Bayes decision function. It is clear that the asymptotical optimality, i.e., Bayes risk consistency depends on properties of class density estimates. The next two theorems on the asymptotical optimality are due to Greblicki [7] and Wolpert and Wagner [16], respectively.

Theorem 1: If

\[
\tilde{f}_i(x) - \frac{1}{n} f_i(x)
\]

as \( n \to \infty \), at almost all \( (\mu) \) in \( R^p \), for \( i = 1, \cdots, M \), then the procedure is Bayes risk consistent.

It should be mentioned that under some additional assumption,