THE BELL SYSTEM TECHNICAL JOURNAL

FEBRUARY 1982
VOL. 61, NO. 2

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL is published monthly, except for the May-June and July-August combined issues, by the American Telephone and Telegraph Company, C. L. Brown, Chairman and Chief Executive Officer; W. M. Ellinghaus, President; V. A. Dwyer, Vice President and Treasurer; F. A. Hutson, Jr., Secretary. Editorial inquiries should be addressed to the Editor, The Bell System Technical Journal, Bell Laboratories, Room WB 1L-336, Crawfords Corner Road, Holmdel, N.J. 07733. Checks for subscriptions should be made payable to The Bell System Technical Journal and should be addressed to Bell Laboratories, Circulation Group, 101 J. F. Kennedy Parkway, Short Hills, N.J. 07078. Subscriptions $20.00 per year; single copies $2.00 each. Foreign postage $1.00 per year; 15 cents per copy. Printed in U.S.A. Second-class postage paid at New Providence, New Jersey 07974 and additional mailing offices.

ISSN0005-8580

Comments on the technical content of any article or brief are welcome. These and other editorial inquiries should be addressed to the Editor, The Bell System Technical Journal, Bell Laboratories, Room WB 1L-336, Crawfords Corner Road, Holmdel, N.J. 07733. Comments and inquiries, whether or not published, shall not be regarded as confidential or otherwise restricted in use and will become the property of the American Telephone and Telegraph Company. Comments selected for publication may be edited for brevity, subject to author approval.

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

Volume 61 February 1982 Number 2

Effects of Day-to-Day Load Variation on Trunk Group Blocking

By A. KASHPER, S. M. ROCKLIN, and C. R. SZELAG

(Manuscript received September 4, 1981)

Modern trunking theory recognizes the need to account for day-to-day load variation when sizing a trunk group for an average blocking objective. This paper investigates the effects of high levels of load variation on average blocking, the measure of service used for sizing final trunk groups in the Public Switched Network. Specifically, we identify a curious phenomenon in which high day-to-day variation results in low average blocking and characterize the traffic theoretic models for which this occurs. By a similar analysis, we also investigate the behavior of an alternate measure of service, the probability of blocking, measured by the ratio of the number of unsuccessful attempts to the total number of attempts.

I. INTRODUCTION

1.1 Background

An important class of trunk groups in the Public Switched Network consists of those groups that provide the last-choice route for a call trying to reach its destination. The performance of such a "final trunk group" is defined to be the 20-day average blocking during the chosen busy hour of the busy season. In the Bell System, the final groups are sized for an objective of one percent average blocking ($\bar{B}.01$) in the busy season.

Figure 1 shows the history of the measured busy-hour loads offered to a trunk group over consecutive days. The degree of variability of the load measurements is not consistent with the hypothesis that the

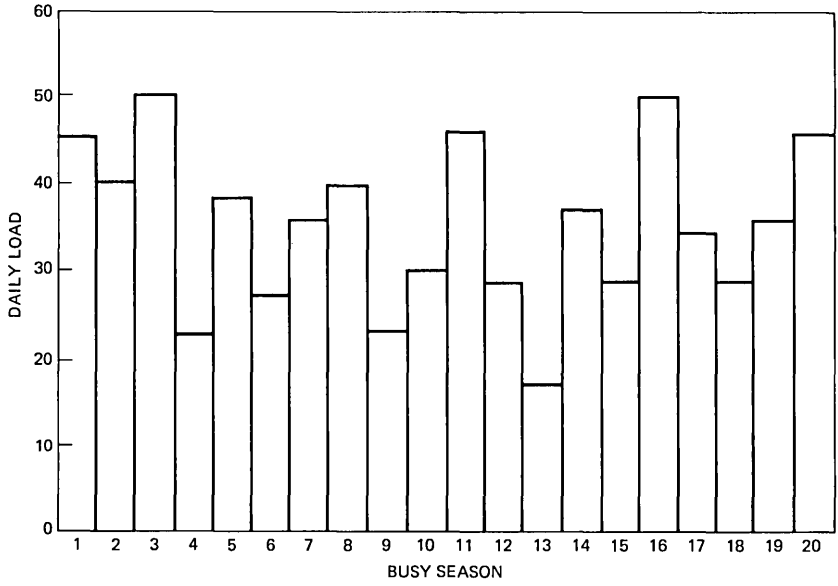


Fig. 1—Busy-season load variations.

daily busy-hour load is constant; rather, it varies from day-to-day in a somewhat random fashion. Modern teletraffic theory has recognized that this day-to-day variability in the offered load must be considered when determining the number of trunks required for an objective level of average blocking. It is generally believed that this variability tends to increase a group's average blocking, and therefore, its trunk requirement, over that observed during constant load conditions.

1.2 Motivation

Recent analyses of Bell Operating Company traffic data have revealed that levels of day-to-day load variation, much higher than those considered in current trunk engineering practices, occasionally appear in the network. To study the effects of such traffic on network service and trunk requirements, the currently deployed traffic models were extended into this region of high load variability. The results are illustrated in Fig. 2, which shows, for Poisson traffic with a fixed mean offered load, $\bar{a}$, and number of trunks, c, the average blocking $\bar{B}$ as a function of the variance, v, of the daily offered load. This quantity, called the day-to-day load variation, is usually parameterized by the variable ϕ, with $v = 0.13\bar{a}^{\phi}$ [1, 2]. In current engineering practices, ϕ is assumed to vary between 1.0 and 1.84. The graph shows that $\bar{B}$ increases with increasing load variation, as expected, but only up to a certain point. Beyond that point, $\bar{B}$ decreases monotonically to zero.

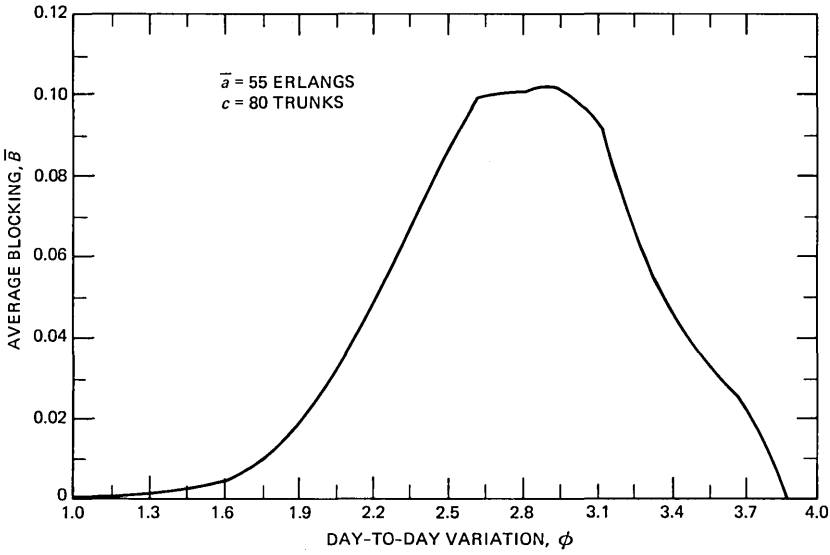


Fig. 2—Average blocking versus day-to-day variation.

Thus, the model predicts that a call arrival process with a highly variable daily load will have lower average blocking than one with a constant daily load.

The purpose of this paper is to analyze and explain this counter-intuitive phenomenon and to determine its implications on current service objectives and traffic models.

1.3 Overview

Section II reviews the model of day-to-day load variation currently deployed in Bell System trunk engineering practices. We then define the average blocking service criterion used for sizing final trunk groups and compare it to an alternate measure of service, the probability of blocking. In Section III, we analyze the behavior of the two measures of service under conditions of high day-to-day load variability and compare numerically their properties in different regions of engineering interest. Section IV summarizes our results and discusses the implications of our findings on trunk engineering practices and network service objectives.

II. TRAFFIC MODELS

This section reviews the model of day-to-day load variation that, in conjunction with models describing the within-hour call arrival process, is used to predict the relationship between trunk group size, average load, and average blocking.

2.1 Average blocking

The discovery of the effect of day-to-day load variation on average blocking is credited to a 1958 study by Wilkinson,[3] who proposed a mathematical model[4] that is now the basis for many of the Bell System's trunk engineering practices.

In Wilkinson's model, the load, a, offered to a group of c trunks within a time-consistent hour (e.g., 9 am to 10 am) varies from day to day in a random fashion. Specifically, the daily loads are modeled as independent random variables with a common gamma distribution $\Gamma(a \mid \bar{a}, v)$ with mean $\bar{a}$ and variance v. We assume further that within each hour, the call arrival process is Poisson and that blocked calls do not retry. The daily blocking probability for a trunk group with c trunks and offered load a is defined by the Erlang blocking function, $B(c, a)$; the average daily blocking probability, denoted $\bar{B}$, is given by

$$\bar{B} = \bar{B}(c, \bar{a}, v) = \int_0^\infty B(c, a) d\Gamma(a \mid \bar{a}, v). \tag{1}$$

2.2 Other considerations

Wilkinson's model (1) has been used within the Bell System to account for day-to-day load variation in the sizing of final trunk groups. Current trunk engineering practices also account for (*i*) the non-Poisson or "peaked" nature of overflow traffic,[5] and (*ii*) the finiteness of the one-hour measurement interval within which traffic measurements are collected.[1] However, the consideration of these additional factors affects only the choice of the daily blocking function in (1) but not the results of this paper, which hold for any single hour blocking function of practical interest.

2.3 Probability of blocking

In the next section, we will compare certain properties of the average blocking service measure, $\bar{B}$, with those of another standard service measure, the probability of blocking, defined below.

Let

$$o(c, a) = aB(c, a) \tag{2}$$

denote the overflow from a trunk group with c trunks and offered load a, and let

$$\bar{o} = \int_0^\infty aB(c, a) d\Gamma(a \mid \bar{a}, v) \tag{3}$$

denote the average overflow. We define the probability of blocking, P_B, as the ratio of the average overflow to the average offered load:

$$P_B = \frac{\bar{o}}{\bar{a}}$$

$$= \frac{1}{\bar{a}} \int_0^\infty aB(c, a)d\Gamma(a \mid \bar{a}, v). \tag{4}$$

Thus, P_B is simply the probability that an arriving call is blocked.

Comparing eqs. (1) and (4), we note that $\bar{B}$ is an unweighted average of the daily blockings, whereas P_B is a weighted average in which the daily blocking is weighted by the daily load.[4]

Next, we investigate the behavior of both $\bar{B}$ and P_B under conditions of high-load variability.

III. AVERAGE VERSUS PROBABILITY OF BLOCKING

In this section, we develop analytical results that validate and explain the high-variation, low-blocking phenomenon observed in Section I. We begin by relating the asymptotic behavior of $\bar{B}$ to the properties of the assumed daily load distribution.

3.1 Asymptotic behavior of $\bar{B}$

First, let us generalize the definition of average blocking given in (1) to the case in which the daily loads are independent, nonnegative random variables with a common distribution function $F(a)$. To simplify notation, let $B(a)$ denote the daily blocking probability on a particular trunk group expressed as a function of its offered load a. Then, the unweighted average blocking is the quantity

$$\bar{B} = \int_0^\infty B(a)dF(a). \tag{5}$$

Clearly, (5) coincides with Wilkinson's model (1) when $B(a)$ is the Erlang B blocking probability and the daily loads are gamma-distributed. By defining the average blocking in the more general form, we can investigate the role of both the blocking function $B(a)$ and the load distribution $F(a)$ in determining the behavior of $\bar{B}$.

Let $\bar{a}$ and v denote the mean and variance of the offered load distribution. Our first goal is to find general conditions on $F(a)$ under which $\bar{B} \to 0$ as $\sqrt{v}/\bar{a}$, the coefficient of variation of $F(a)$, increases. Our second goal is to show that, for a very general class of blocking functions $B(a)$, the commonly assumed gamma distribution satisfies the required conditions on the load distribution. Thus, by analysis we will both verify and explain the high-variation, low-blocking behavior described in Section I.

To give a precise answer to the first question posed above, we first define the general class of blocking functions to be considered. A real-

valued function $B(a)$ will be called a blocking function if it has the following four properties:

($B1$) $B(0) = 0$ and $B(a) > 0$ if $a > 0$;

($B2$) B is continuous;

($B3$) B is non-decreasing; and

($B4$) B is bounded.

Let $\{F_k\}$ be a sequence of probability distribution functions concentrated on $[0, \infty)$ and consider the sequence

$$\bar{B}_k = \int_0^\infty B(a) dF_k(a). \tag{6}$$

In Appendix A, we derive simple, necessary and sufficient conditions on the distribution sequence $\{F_k\}$ under which $\bar{B}_k \to 0$. Specifically, this will occur if and only if

$$\lim_{k \to \infty} F_k(a) = 1 \quad \text{for all} \quad a > 0. \tag{7}$$

The sufficiency of (7) follows from a standard convergence theorem (See Ref. 6, p. 249). A simple, direct proof of both the necessity and sufficiency of (7) is given in Appendix A.

Condition (7) says that all quantiles of the load distribution converge to zero. Equivalently, this means that all of the mass of the distribution converges toward the origin. Note, however, that (7) does not imply that the moments of the distributions (e.g., mean, variance) converge to zero.

Using this result, we can now analyze the effect of high day-to-day load variation on average blocking in the case of gamma-distributed daily loads.

Let $\{\Gamma_k(a)\}$ denote a sequence of gamma distribution functions with mean $\bar{a}_k$ and variance v_k. According to our result, the average blocking $\bar{B}_k = \int_0^\infty B(a) d\Gamma_k(a)$ converges to zero if and only if

$$\lim_{k \to \infty} \Gamma_k(a) = 1$$

for all $a > 0$. In Appendix B, we show that this occurs when the coefficient of variation of $\Gamma_k(a)$, $\sqrt{v_k}/\bar{a}_k$, increases without bound. In particular, if $\bar{a}_k = \bar{a}$ is fixed and $v_k \to \infty$, the average blocking $\bar{B}_k \to 0$ for arbitrary $c > 0$. Moreover, the number of trunks required to guarantee one percent blocking, $\bar{B}.01$, also tends to zero if $\bar{a}_k = \bar{a}$ and $v_k \to \infty$.

Thus, the results of this section give theoretical explanation for the phenomenon observed in Section I.

3.2 Asymptotic behavior of P_B

The results of the previous section can also be used to analyze the

asymptotic behavior of P_B, the probability of blocking, which we defined in Section II.

Recall that the probability of blocking can be expressed as

$$P_B = \frac{\bar{o}}{\bar{a}} = \frac{1}{\bar{a}} \int_0^\infty o(a) dF(a), \tag{8}$$

where $o(a)$ is the daily overflow load and $F(a)$ is the distribution of the daily offered load.

Let $u(a) = a - o(a)$ denote the daily carried load. Then

$$\bar{o} = \bar{a} - \bar{u}, \tag{9}$$

where $\bar{u} = \int_0^\infty u(a) dF(a)$ is the average carried load. Thus,

$$P_B = \frac{\bar{a} - \bar{u}}{\bar{a}}. \tag{10}$$

We can now analyze the effect of day-to-day load variation on the probability of blocking by studying the behavior of $\bar{u}$. To do this, we first note that the function $u(a)$ has the following properties:

($u1$) $u(0) = 0$ and $u(a) > 0$ for $a > 0$;

($u2$) $u(a)$ is continuous;

($u3$) $u(a)$ is non-decreasing; and

($u4$) $u(a)$ is bounded (by the number of trunks in the group).

Thus, the carried load $u(a)$ has the required properties of the "blocking function" and our result of Section 3.1 can be applied. That is, we know that $\bar{u} \to 0$ if and only if the load distributions converge as in (7). If $\bar{u} \to 0$ and $\bar{a}$ is fixed, then

$$P_B = \frac{\bar{a} - \bar{u}}{\bar{a}} \to 1. \tag{11}$$

In particular, if the daily loads are gamma-distributed, the probability of blocking converges to 1 as the coefficient of day-to-day variation increases with the mean held constant.

These results illustrate dramatically a fundamental difference between the average blocking ($\bar{B}$) and probability of blocking (P_B) service criteria under conditions of highly volatile network loads. For extremely high levels of day-to-day load variation, the average blocking measure takes on low values, indicating good service, even though most of the traffic is blocked ($P_B \approx 1$). Under such conditions, the unweighted average blocking is a poor indicator of the service experienced by the customer.

3.3 Comparison of $\bar{B}$ and P_B

In addition to the asymptotic results described above, we can also

analyze the general relationship between $\bar{B}$ and P_B. Specifically, we will show that $P_B \geq \bar{B}$ regardless of the load distribution.

To see this, note that our assumption that the blocking $B(a)$ is a nondecreasing function of the offered load implies that for any load distribution $F(a)$:

$$\int_0^{\bar{a}} (\bar{a} - a)B(a)dF(a) \leq \int_0^{\bar{a}} (\bar{a} - a)B(\bar{a})dF(a)$$

$$= \int_{\bar{a}}^{\infty} (a - \bar{a})B(\bar{a})dF(a) \leq \int_{\bar{a}}^{\infty} (a - \bar{a})B(a)dF(a). \quad (12)$$

Equivalently,

$$\int_0^{\infty} aB(a)dF(a) \geq \bar{a} \int_0^{\infty} B(a)dF(a),$$

or

$$P_B = \frac{\bar{\sigma}}{\bar{a}} \geq \bar{B}. \quad (13)$$

Next, we complement these results with a few numerical examples that illustrate the differences between $\bar{B}$ and P_B.

3.4 Numerical examples

Recall that final trunk groups in the public telephone network are sized for an average blocking objective of one percent. The behavior of $\bar{B}$ in this low blocking region is illustrated by the lower curve in Fig. 3, which shows, as a function of the day-to-day load variation, the actual average blocking experienced by a trunk group sized under the assumption of no day-to-day variation. As we observed in Section I, $\bar{B}$ initially increases to a maximum value much less than 1.0, and then decreases to zero. In contrast, the probability of a call's being blocked, P_B, monotonically increases to 1.0. However, within the range of day-to-day variation normally encountered ($1.0 \leq \phi \leq 1.84$), the numerical difference between $\bar{B}$ and P_B is not great.

Finally, let us examine the behavior of $\bar{B}$ on a trunk group sized for a high level of average blocking (e.g., 20 percent), assuming no day-to-day variation. Fig. 4 illustrates that day-to-day variation exerts an altogether different influence in this region. Namely, the average blocking decreases monotonically to zero, although the probability of a call's being blocked again increases monotonically to 1.0.

IV. SUMMARY AND CONCLUSIONS

This paper describes an investigation of certain properties of the

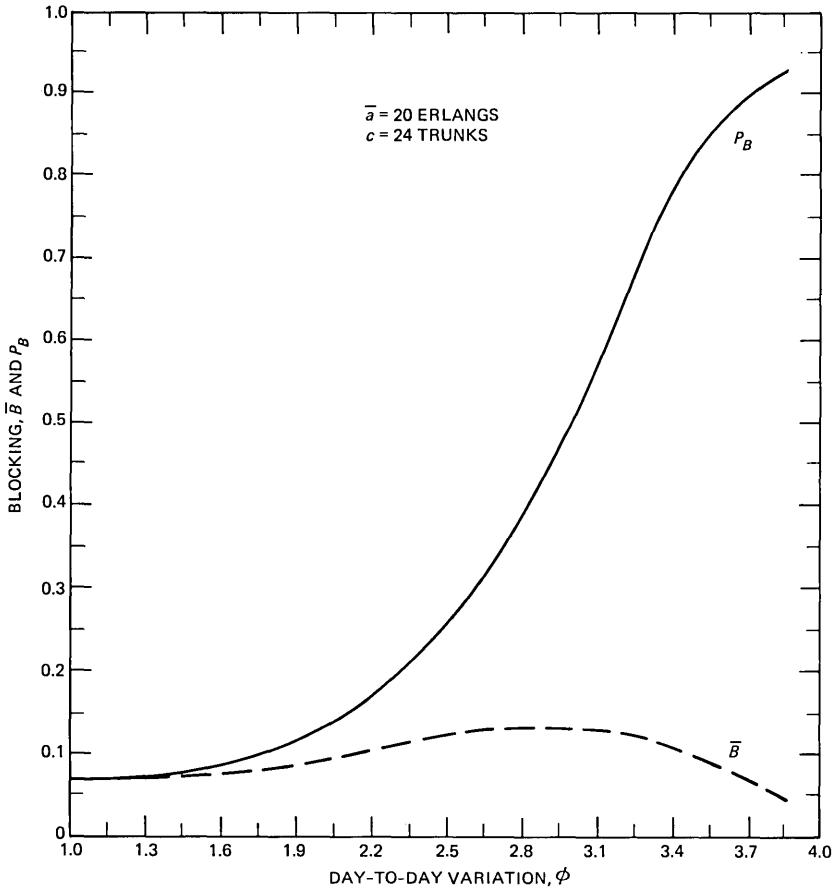


Fig. 3—Blocking versus day-to-day variation.

average blocking service measure ($\bar{B}$) used for sizing final trunk groups in the Public Switched Network. The work was motivated by the results of a recent data study that suggested that levels of day-to-day load variation much higher than those considered by Wilkinson[4] and current Bell System engineering practices occasionally appear in the network. The analytical development presented in Section III confirmed our numerical result of Section 1.2 that very high levels of load variation will result in low levels of average blocking. In addition, the properties of an alternate measure of blocking, P_B, were analyzed and compared to those of $\bar{B}$.

Our findings are summarized below:

(*i*) For the low, objective level of average blocking (one percent) and the traffic conditions normally encountered in the Public Switched

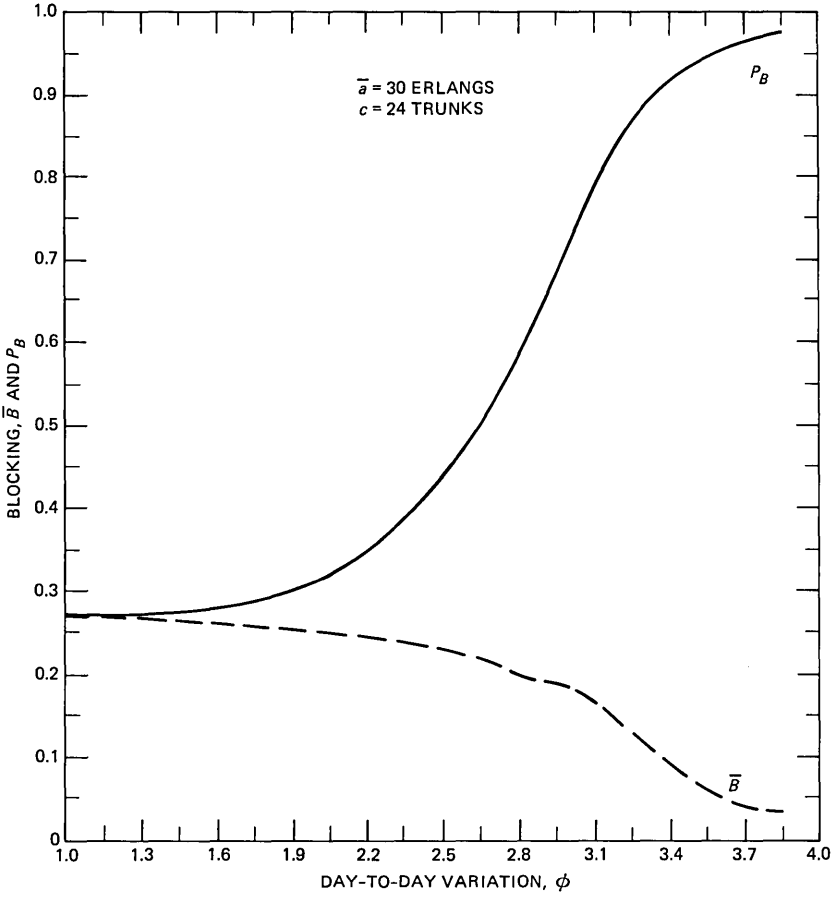


Fig. 4—Blocking versus day-to-day variation.

Network, the discrepancy between $\bar{B}$ and P_B is not great. That is, the current Bell System practice of sizing final trunk groups for a fixed, low level of average blocking also yields correspondingly low, though not uniform, levels of blocking probability.

(ii) In contrast with $\bar{B}$, engineering for a fixed level of P_B guarantees, on average, a fixed fraction of successful calls. However, in the presence of high day-to-day variation, the number of trunks required for an objective level of P_B is substantially greater than that required for the same level of $\bar{B}$. Because, the choice of an engineering objective should consider both customer satisfaction and cost, we cannot conclude that P_B is a better objective than $\bar{B}$.

(iii) In the range of high blocking, which is typical for private and/or special services networks, the discrepancy between $\bar{B}$ and P_B is sub-

stantial. In that case, our results suggest that consideration should be given to appropriateness of average blocking as a service measure.
(*iv*) We showed that, under volatile traffic conditions, the assumption of a particular daily load distribution is crucial in quantifying the relationship between load offered to a trunk group and the average blocking of the trunk group.

APPENDIX A

In Section III, we considered the general class of blocking functions $B(a)$ having the following four properties:

($B1$) $B(0) = 0$ and $B(a) > 0$ for $a > 0$;
($B2$) B is continuous;
($B3$) B is nondecreasing;
($B4$) B is bounded.

We now prove the following theorem:

Theorem 1: *Let* $\{F_k\}$ *be a sequence of probability distribution functions on* $[0, \infty)$ *and let* $B(a)$ *be any blocking function. Then*

$$\lim_{k\to\infty} \int_0^\infty B(a)dF_k(a) = 0 \qquad \textit{if and only if}$$

$$\lim_{k\to\infty} F_k(a) = 1 \qquad \textit{for} \qquad \textit{all } a > 0.$$

Proof: To show sufficiency, let $\epsilon > 0$. Since $B(0) = 0$ and B is continuous, $a < \delta$ implies $B(a) < \epsilon/2$ for sufficiently small δ. Assume that B is bounded by M. Since

$$\lim_{k\to\infty} F_k(a) = 1,$$

for sufficiently large k, $F_k(\delta) \geq 1 - \epsilon/2M$. For such k,

$$\int_0^\infty B(a)dF_k(a) = \int_0^\delta B(a)dF_k(a) + \int_\delta^\infty B(a)dF_k(a) \leq \epsilon/2 + \epsilon/2 = \epsilon.$$

Thus,

$$\lim_{k\to\infty} \int_0^\infty B(a)dF_k(a) = 0.$$

To show necessity, suppose that for some $a_0 > 0$,

$$\lim_{k\to\infty} F_k(a_0) \neq 1.$$

Then there exists a subsequence $\{n_k\}$ and a $\lambda_1 > 0$ such that $F_{n_k}(a_0) \leq 1 - \lambda_1$. Also, ($B1$) to ($B3$) imply that $B(a)$ is bounded away from zero on $[a_0, \infty)$, say by $\lambda_2 > 0$. Then,

$$\int_0^\infty B(a)dF_{n_k}(a) \geq \int_{a_0}^\infty B(a)dF_{n_k}(a) \geq \lambda_2 \int_{a_0}^\infty dF_{n_k}(a) \geq \lambda_2 \cdot \lambda_1 > 0,$$

which contradicts

$$\int_0^\infty B(a)dF_k(a) \to 0;$$

thus,

$$\lim_{k\to\infty} F_k(a) = 1 \quad \text{for all} \quad a > 0.$$

Q.E.D

APPENDIX B

Theorem 2: Let $\{\Gamma(a|\bar{a}, v)\}$ *be a family of gamma distributions with given mean* $\bar{a}$ *and variance* v *and suppose that* $\bar{a} \geq 1$ *and* $\sqrt{v}/\bar{a} \to \infty$. *Then,*

$$\Gamma(a|\bar{a}, v) \to 1 \quad \textit{for any} \quad a > 0.$$

Proof: By definition, the gamma distribution with fixed mean $\bar{a}$ and variance v is given by

$$\Gamma(a|\bar{a}, v) = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^a t^{\alpha-1} e^{-\beta t} dt, \tag{14}$$

where $\Gamma(\alpha)$ is the gamma function and the parameters α and β are determined by

$$\bar{a} = \alpha/\beta, \; v = \alpha/\beta^2. \tag{15}$$

Integrating (14) by parts we obtain

$$\Gamma(a|\bar{a}, v) = \frac{\beta^\alpha}{\Gamma(\alpha)\alpha} e^{-\beta t} t^\alpha \Big|_0^a + \frac{\beta^\alpha \beta}{\Gamma(\alpha)\alpha} \int_0^a e^{-\beta t} t^\alpha dt. \tag{16}$$

Let us start by showing that the first term in (16) tends to 1 as the coefficient of variation tends to infinity. From (15) $\alpha = \bar{a}^2/v$ and using the property of the gamma function we obtain

$$\Gamma(\alpha)\alpha = \Gamma(\alpha + 1) \to 1 \quad \text{as} \quad \alpha = \bar{a}^2/v \to 0. \tag{17}$$

From (15) we have

$$\beta^\alpha = \left|\frac{\bar{a}}{v}\right|^{\bar{a}^2/v} = \exp\left|\frac{\bar{a}^2}{v} \ln \frac{\bar{a}}{v}\right|.$$

Using the monotonicity of the exponential and logarithmic functions

we have

$$\exp\left|\frac{\bar{a}}{v}\ln\frac{\bar{a}}{v}\right| \le \exp\left|\frac{\bar{a}^2}{v}\ln\frac{\bar{a}}{v}\right| \le \exp\left|\frac{\bar{a}^2}{v}\ln\frac{\bar{a}^2}{v}\right|. \tag{18}$$

Since $\bar{a} \ge 1$ the ratio $\bar{a}/v \to 0$ as $\bar{a}^2/v \to 0$. Thus, passing to the limit in (18) and noting that $\lim_{x\to+0} x\ln(x) = 0$, we obtain that

$$\beta^\alpha \to 1 \quad \text{as} \quad \bar{a}^2/v \to 0. \tag{19}$$

Thus, from (17) and (19) the first term in (14) tends to 1 as $\alpha \to 0$ and $\beta \to 0$.

For the second term in (14) we get

$$\lim_{\substack{\alpha\to0\\ \beta\to0}} \frac{\beta^\alpha\beta}{\Gamma(\alpha)\alpha}\int_0^a e^{-\beta t}t^\alpha dt \le \lim_{\substack{\alpha\to0\\ \beta\to0}} \frac{\beta^\alpha}{\Gamma(\alpha)\alpha} \lim_{\substack{\alpha\to0\\ \beta\to0}} \beta\int_0^a a^\alpha dt = 0. \tag{20}$$

Thus, if $\bar{a} \ge 1$, for any $a > 0$,

$$\Gamma(a\,|\,\bar{a}, v) \to 1 \quad \text{as} \quad \sqrt{v}/a \to \infty.$$

The proof is complete.

We would like to remark that under the assumption $\bar{a} < 1$, Theorem 2 can be reformulated as follows:

$$\Gamma(a\,|\,\bar{a}, v) \to 1 \quad \text{as} \quad v/\bar{a} \to \infty$$

for arbitrary $a > 0$.

REFERENCES

1. D. W. Hill and S. R. Neal, "Traffic Capacity of a Probability Engineered Trunk Group," *55,* B.S.T.J. (September 1976), pp. 831–42.
2. R. I. Wilkinson, "Some Comparisons on Load and Loss Data with Current Teletraffic Theory," *50,* B.S.T.J. (October 1971), pp. 2807–34.
3. R. I. Wilkinson, "A Study of Load and Service Variation in Toll Alternate Route Systems," Proc. Second Int. Teletraffic Congress, The Hague, July 7–11, 1958, Document No. 29.
4. R. I. Wilkinson, *Nonrandom Traffic Curves and Tables for Engineering and Administrative Purposes,* Traffic Studies Center, Bell Telephone Laboratories, August 1970.
5. R. I. Wilkinson, "Theories for Toll Traffic Engineering in the USA," *35,* B.S.T.J. (March 1956), pp. 421–514.
6. W. Feller, An Introduction to Probability Theory and Its Applications, *II,* Second Edition, New York: John Wiley, 1971.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 2, February 1982
Printed in U.S.A.

The Continuing Evolution of the Military Standard 105D Sampling System

By B. S. LIEBESMAN

(Manuscript received June 18, 1981)

Military Standard 105D is the most widely used set of acceptance sampling plans in the world. This paper reviews the early development of the standard and points out the many contributions made by Bell System researchers, such as H. F. Dodge. The paper also reviews recent analyses and indicates areas where the special structure suggested by Dodge, and adopted in the standard, has been extremely valuable. Finally, the paper identifies many questions related to the standard that are still open for investigation.

I. INTRODUCTION

1.1 Genesis of military standard 105D

Lot-by-lot acceptance sampling began just prior to World War II and was given a large boost during the war because of the need to assure the quality of wartime material. Bell Laboratories personnel were heavily involved in the early development of sampling plans. The most prolific Bell Laboratories contributors were G. D. Edwards, H. F. Dodge, and H. G. Romig.

The initial system of acceptance sampling plans was developed to assure wartime material. This system evolved through a number of changes to the current system of plans, Military Standard 105D. This standard is described in Ref. 1. H. F. Dodge was one of the leading contributors to the final development of this system. W. R. Pabst, long-time editor of the Standards Section of the *Journal of Quality Technology*, discussed Dodge's contribution in a paper presented before the 17th annual convention of the American Society for Quality Control (ASQC):[2]

> "Much of the theoretical work underlying the new MIL-STD-105D is directly attributed to David Hill and indirectly to Harold Dodge."

Relatively few changes have occurred since 1963 when the MIL-STD-105D* system was published. What has happened instead has been an in-depth investigation of the properties of the system, with the result that changes have occurred in the way the system is used. Bell Laboratories Quality Assurance Center has been active in this investigation. The Center's effort has been in support of the Western Electric Company Purchased Product Inspection organization, which uses MIL-STD-105D almost exclusively to inspect products purchased by the Bell System.

Today, MIL-STD-105D is the most widely used system of acceptance sampling plans in the world as shown in a 1970 Japanese study.[3] It forms the basis for the American National Standards Institute system, ANSI Z1.4; the Japanese system, JIS 29015; the International Standards Organization system ISO 2859; and the British System DEF 131. Saniga and Shirland[4] estimated in 1977 that 76 percent of the quality control organizations in the United States use the system. The Japanese have made extensive use of MIL-STD-105D, a factor which may have contributed to the improved quality of Japanese products since World War II. Finally, it should be noted that the use of the system has spread to many types of items other than manufactured goods or raw materials. These include data records, maintenance operations, financial records, and administrative procedures.

II. DESCRIPTION OF MIL-STD-105D

2.1 Basic definitions

A number of terms are introduced in this section. These terms are important to the description of the system in Section 2.2 and of the issues, past and present, which are discussed in Sections III and IV. Definitions have been included in the glossary for handy reference.

First of all, we define a lot as a set of items under control of the inspection organization for which an acceptance or rejection must be made. The items forming the lot must be similar in nature. A random sample is a subset of the lot which is selected in a manner which makes it representative of the lot. In a truly random sample, each unit in the lot has the same probability of being included in the sample.

Inspection by attributes classifies individual samples as either "nondefective" or "defective" (good or bad, go or no go, etc.), depending upon whether they pass or fail certain tests of characteristics. The quality of a product is measured in terms of the percent defective or defects per hundred units. Inspection by attributes provides an estimate of the quality which is used for lot acceptance or rejection.

---

* Military Standard 105D is commonly abbreviated using MIL-STD-105D or just 105D. See the glossary for this and other abbreviations and definitions.

The operating characteristics (OC) is a curve of the probability of acceptance as a function of quality for a given sampling plan. Figure 1 is a typical OC curve. Note that there are two parameters marked on the curve, a good quality, AQL, and a poor quality, LTPD. Lots of quality equal to the AQL value have a high probability of acceptance, while lots of quality equal to the LTPD value have low probability of acceptance.

The acceptable quality level (AQL) is the index for the plans of MIL-STD-105D. It is defined in Ref. 1 as "the maximum percent defective (or the maximum number of defects per hundred units) that, for the purposes of sampling inspection, can be considered satisfactory as a process average." It is the AQL value at which the producer generally aims the quality of his process. If he produces at this level, he has a high probability that most of the lots will be accepted.

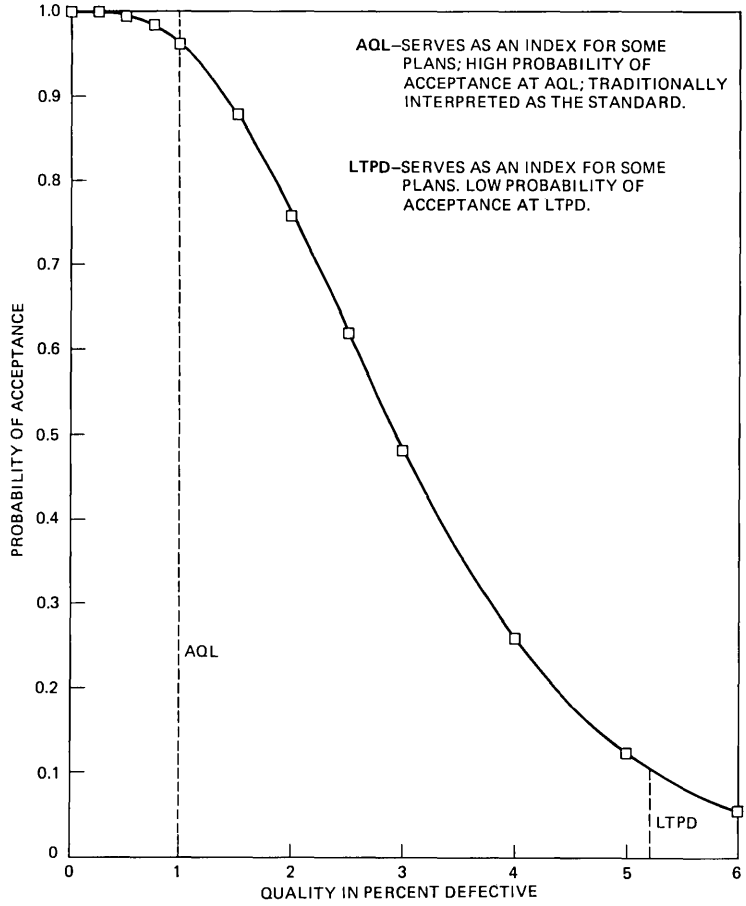


Fig. 1—Operating characteristic curve.

The lot tolerance percent defective (LTPD) is a quality level at which most lots will be rejected. Because of sampling error, some lots at this quality will be accepted.

The average outgoing quality (AOQ) is the average quality of all lots that are shipped. The average outgoing quality limit (AOQL) is the upper bound on AOQ when using a sampling plan that requires all units in all rejected lots be inspected and all defectives removed before shipment. The LTPD and AOQL values are not used directly in MIL-STD-105D. However, it is worth noting that there are plans based on these quantities as indices. The most commonly used set of LTPD and AOQL plans was developed by H. F. Dodge and H. G. Romig of Bell Laboratories.[5]

The final definitions introduced in this section are based on the fact that MIL-STD-105D is a system of plans with a feedback mechanism. This mechanism consists of four phases of operation and the switching rules for transferring between phases. The normal phase is used when there is no evidence that the quality being submitted is poorer or better than the specified quality level. The tightened phase is used when there is evidence of poorer quality, while the reduced phase is used when there is evidence of better quality. The discontinue phase is entered when the producer has not been successful in improving the poor quality of his product during the tightened phase.

2.2 Procedures of MIL-STD-105D

An outline of the basic MIL-STD-105D procedures is given here to further introduce important terminology and to give the reader an understanding of acceptance sampling plans. For more details, refer to the military standard,[1] and its accompanying handbook,[6] or to a quality control text, such as that by Duncan[7] or Grant and Leavenworth.[8]

There are five steps used in the selection of a sampling plan. First of all, the lot to be inspected must be formed and the lot size determined. The lot should be as homogeneous as possible. Next, the sample size code letter must be found in a table as a function of the lot size. The third step is to determine the AQL value to use based on the quality requirements of the product. The fourth step is to select the sampling plan from another table based on the sample size code letter and the AQL value. Finally, a random sample is selected, tested, and the lot accepted or rejected based on the allowable number of defectives for the plan.

These procedures are used to select an individual sampling plan for inspection of an individual lot. In reality, the system of plans operates over a sequence of lots using the phases and switching rules to provide consumer protection and control over the economy of inspection.

Figure 2 illustrates the basic switching rules. Generally, the inspection of a product starts in the normal phase (N), where quality is assumed to be at the desired level. When two out of five consecutive lots are rejected, this is taken to be evidence of poor quality. The next lot is inspected under the tightened phase (T). This results in using a plan that will reject a higher percentage of lots at a given quality than the corresponding plan under the normal phase.

The acceptance of five lots in a row during the tightened phase is evidence that the quality is back to the desired level. When this occurs, the next lots are inspected under the normal phase. However, if this does not occur before ten lots have been inspected under the tightened phase, sampling ceases and the discontinue phase (D) is entered.

When a product has been inspected under the normal phase for a number of consecutive lots, it is eligible for reduced inspection. The reduced phase (R) is entered after ten lots in a row are accepted and a defect limit number criterion is met. The product remains in the reduced phase until either a lot is rejected, or is accepted but the number of defectives exceeds a specified number (*Ac*).

Thus, we see that MIL-STD-105D is a sampling scheme with a feedback mechanism to reward good quality and improve poor quality. The manner in which this mechanism operates is not mathematically precise, but relies on the nature of human reaction to reward and punishment. This was recognized by the developers:

> "Whereas sampling plans are mathematically precise, the study of sampling schemes is not limited to pure mathematical considerations. In fact, many of the decisions related to the development of a sampling scheme are more a matter of art, opinion, esthetics, appeal, practical considerations and compromise."[9]

Often these words are lost on the practitioners who concentrate on the individual sampling plans rather than on the system. The author recently attended the 1981 ASQC Quality Congress where he had many discussions with users of MIL-STD-105D who were ignoring important aspects of the system. In essence, MIL-STD-105D was developed as a

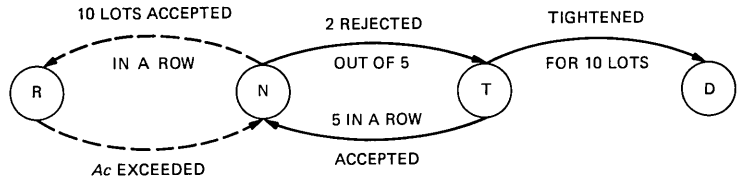


Fig. 2—Dynamics of the switching rules.

system of plans and any abrogation of its features destroys its effectiveness.

III. DEVELOPMENT OF MIL-STD-105D

We will now discuss the historical development of MIL-STD-105D. It is important to review this development because the current structure of the standard is a result of compromises that arose over issues relating to the standard. In fact, MIL-STD-105D was created because of the need to provide industry with a system of sampled plans indexed on AQL values.

3.1 Need for an AQL system

Two other types of sampling plans were developed just prior to MIL-STD-105D and served as competitors of 105D through much of its early history. These are plans indexed by the lot tolerance percent defective (LTPD) and the average outgoing quality limit (AOQL).[5]

I. D. Hill discusses the reasons why industry became disenchanted with LTPD and AOQL plans and pressured instead for AQL plans of the 105D type. He states that:

> "In normal situations, the process average corresponds to a high point on the OC curve; it has to, if the producer is going to make a profit. So it is really a high point rather than a low point which is the primary concern of both the producer and consumer."[10]

In addition, Hill states that an LTPD plan is not cost efficient:

> "This system [LTPD] leads then in general, to the producer making, and the consumer receiving, a quality considerably better than is really necessary, and the price must reflect this."[10]

To compare MIL-STD-105D to an AOQL system, we must first expand on the definitions of average outgoing quality (AOQ) and average outgoing quality limit (AOQL) given in Section 2.1. The AOQ is just the average quality of all lots that are shipped. In an AOQL system, a rejected lot must be inspected 100 percent and all defectives removed. Thus, a rejected lot is made perfect and shipped in sequence with other accepted lots. It can be shown[7,8] that the AOQ has an upper bound called the AOQL. Figure 3 illustrated this. The AOQL value is generally used as the "guaranteed" quality to be provided by a supplier.

The use of AOQL plans decreased when it was found that 100 percent inspection did not guarantee perfection. Hill states that:

> "Now it is well known in practice that 100 percent sorting is

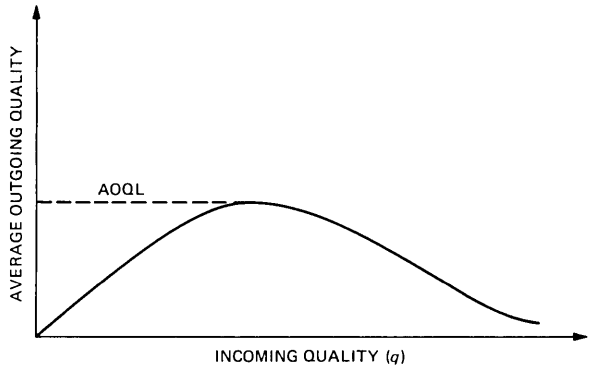


Fig. 3—Average outgoing quality limit (AOQL).

unlikely to be properly done ... In stressing, therefore, that the AOQL concept requires perfection in the inspection operation, I am not claiming that other methods do not require this. It is merely that the lack of such perfection seems to matter more in the case of AOQL."[10]

He then proceeds to show that because of inspection error, the AOQ curve tends to have the shape of the dashed curve in Fig. 4 rather than the solid curve shown in Fig. 3. Note that the maximum value of the AOQ curve in Fig. 4 depends on r, the probability that an inspector will call a bad unit good. The inspection error has an effect on the portion of the curve to the left of the dashed curve, but this effect is so small that it cannot be shown in Fig. 4.

The end result of an AOQL plan, according to Hill, is:

"This Table (Table 1 in Ref. 10) shows that the use of an AOQL

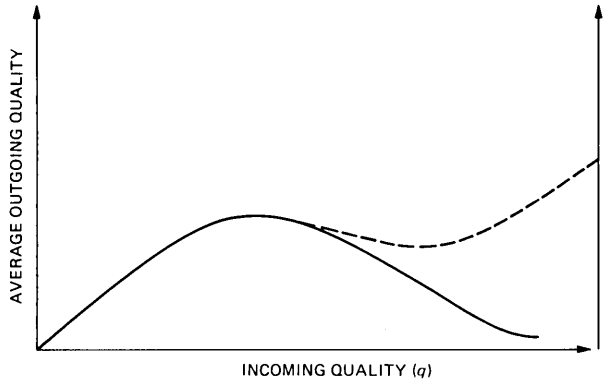


Fig. 4—Effect of inspection error on outgoing quality under an AOQL plan.

plan will, in general, force the producer to offer a quality a good deal better than the AOQL value, although not to such an extent as in the LTPD approach."[10]

And, to summarize the reason for taking the AQL approach over the LTPD and AOQL approaches, Hill states:

"Both the LTPD and AOQL concepts are capable of giving good protection against poor quality, provided they are used efficiently, but both do so at the cost of rejecting a good deal of satisfactory production."[10]

Thus, AQL sampling plans were developed to be indexed on quality values having a high probability of acceptance. This provided the producer with good protection against rejection of satisfactory lots. Protection for the consumer was provided by adoption of switching rules which recognize poor quality over a series of lots and take actions which put pressure on the supplier to improve.

3.2 Important issues during the development of MIL-STD-105D

A number of issues surfaced during the period leading to the 1963 publication of MIL-STD-105D. Most dealt with the practical application of the theoretical framework of the sampling system. Dodge was a prolific contributor during this period because he had close ties to both the theoretical academic world and the practical world of Bell Laboratories and Western Electric.

A change to the interpretation of the AQL is a prime example of Dodge's influence. The early tables were designed to have the probability of acceptance equal to 0.95 for quality equal to the AQL value. Dodge proposed a special structure which resulted in the probability of acceptance for quality equal to the AQL value varying between 0.88 and 0.99. This structure led to plans which were much easier to use by the general quality practitioner, because it consisted of a fixed set of sample sizes and a fixed set of AQL values for the entire range of lot sizes.

Dodge suggested that the values of AQL and sample size both follow the same geometric progression based on multiples of $\sqrt[5]{10} = 1.585$.[11] This resulted in the sequence of AQL values currently used: $\cdots$, 1, 1.5, 2.5, 4, 6.5, $\cdots$; and the sequence of sample sizes (n) currently used: 2, 3, 5, 8, 13, 20, $\cdots$. Furthermore, the structure in the table of sampling plans was enhanced, because along any diagonal, the product of AQL and n is essentially constant and the acceptance number is a constant. (See Fig. 5.) Recently, this structure has been quite useful in a number of analyses which will be described in Section IV.

SINGLE SAMPLING PLANS FOR NORMAL INSPECTION

SAMPLE SIZE	ACCEPTABLE QUALITY LEVELS (NORMAL INSPECTION)																	
	0.25		0.40		0.65		1.0		1.5		2.5		4.0		6.5		10	
	Ac	*Re*	*Ac*	*Re*	*Ac*	*Re*	*Ac*	*Re*	*Ac*	*Re*	*Ac*	*Re*	*Ac*	*Re*	*Ac*	*Re*	*Ac*	*Re*
2	⇩		⇩		⇩		⇩		⇩		⇩		⇩		0	1	⇩	
3													0	1	⇧			
5											0	1	⇧		⇩		1	2
8									0	1	⇧		⇩		1	2	2	3
13							0	1	⇧		⇩		1	2	2	3	3	4
20					0	1	⇧		⇩		1	2	2	3	3	4	5	6
32			0	1	⇧		⇩		1	2	2	3	3	4	5	6	7	8
50	0	1	⇧		⇩		1	2	2	3	3	4	5	6	7	8	10	11
80	⇧		⇩		1	2	2	3	3	4	5	6	7	8	10	11	14	15
125	⇩		1	2	2	3	3	4	5	6	7	8	10	11	14	15	21	22
200	1	2	2	3	3	4	5	6	7	8	10	11	14	15	21	22	⇧	
315	2	3	3	4	5	6	7	8	10	11	14	15	21	22	⇧			
500	3	4	5	6	7	8	10	11	14	15	21	22	⇧					
800	5	6	7	8	10	11	14	15	21	22	⇧							
1250	7	8	10	11	14	15	21	22	⇧									
2000	10	11	14	15	21	22	⇧											

⇩ USE FIRST SAMPLING PLAN BELOW ARROW

⇧ USE FIRST SAMPLING PLAN ABOVE ARROW

Ac ACCEPTANCE NUMBER

Re REJECTION NUMBER

Fig. 5—Part of a table of sampling plans from MIL-STD-105D.

The structure devised by Dodge had very strong practical implications. The fixed set of AQL values and sample sizes were easy to understand, interpret, and use by the general quality practitioner.

Dodge also provided inputs concerning tightened inspection and the switching rules. He suggested[11] that a switch to tightened inspection be done when two out of five consecutive lots fail. This was a change from the process average criteria used in early versions of MIL-STD-105D and was based on Dodge's experience with users.[12] For tightened plans, he suggested the use of the same sample size as for normal plans

(adopted), so that inspection costs would not increase, and the use of an acceptance number one less than the normal acceptance number (partially adopted). The tightened plans adopted as part of 105D do have AOQL values that are close to AQL values of the corresponding normal plans, as Dodge had suggested.[11]

There were also questions concerning the sample size during the reduced phase, the sample sizes during double and multiple sequential sampling, and the relationships between lot size and sample size. Compromises were reached which resulted in the following:

(*i*) The reduced sample size was set at 40 percent of the normal sample size:

(*ii*) Sequential samples were set to be the same size as the first sample;

(*iii*) Empirical relationships for relating lot size to sample size were used; and

(*iv*) Multiple sampling levels were added to give the user a variety of sampling options.

Finally, there was controversy over analytical tools published to help the practitioner analyze his plans. The major tool that eventually was added to MIL-STD-105D was a set of charts and tables defining the OC curve for each plan. Discussions arose over the distributions to be used to calculate the probabilities, and a compromise was made:[13]

(*i*) The Poisson distribution would be used for AQL > 10;

(*ii*) The Poisson approximation to the binomial distribution would be used for AQL ≤ 10 and sample size ≥ 50; and

(*iii*) The binomial distribution would be used for all other plans.

Some analysts felt that different measures of plan capability would be more useful. For one, Professor Barnard of the Royal Statistical Society, in his comments on Hill's paper,[10] suggested the use of the average run length (ARL) needed to detect shifts in quality as a better measure of plan capability.

The current version of MIL-STD-105D published in 1963 includes other analytical tools. These are tables to determine (*i*) the AOQL and LTPD values corresponding to each plan, and (*ii*) charts showing the average sample size.

We are now ready to examine the most recent period in the life of MIL-STD-105D. During this period, only minor changes have been made to the procedures. This stability has allowed researchers time to observe the use of the standard and to analyze the impact of the total system. Prior to this time, the research had concentrated on the properties of individual plans.

An interest developed to provide tools to aid users in selecting sets of plans from 105D and to analyze the effectiveness of the combination of these plans under specified conditions. Some of this work will be discussed in the next four sections.

IV. ANALYSIS OF MIL-STD-105D

4.1 System of OC curves

The development of OC curves for the system of plans represented a breakthrough in the use of MIL-STD-105D because it recognized the effect of the total system and not just individual plans. This section describes the evolution of these curves, which took approximately thirteen years, and illustrates the difficulty of changing aspects of widely used procedures such as MIL-STD-105D.

The MIL-STD-105D was developed under the assumption that the normal, tightened, and discontinue phases would be strictly adhered to.[1] Use of the reduced phase is considered optional. The need to follow all procedures was recognized by Dodge,[14] Hald and Thyregod,[15] and Stephens and Larson.[16]

Stephens and Larson of Western Electric Company were the first to explore the system OC concept. They constructed a Markov model of MIL-STD-105D considering two cases: (*i*) tightened, normal, and reduced phases combined, and (*ii*) tightened and normal phases combined. The discontinue phase was not included in either case.

The end result of their model was a composite operating characteristic curve for the system:

$$P_{ac}(q) = r_T P_{aT}(q) + r_N P_{aN}(q) + r_R P_{aR}(q), \tag{1}$$

where

P_{ac} = system OC curve
r_T, r_N, r_R = expected proportion of lots inspected during tightened, normal, and reduced inspection
P_{aT}, P_{aN}, P_{aR} = tightened, normal, and reduced OC curves
q = quality.

In addition, they used the Markov model to find the expected number of units sampled per lot or the average sample number (ASN):

$$\text{ASN} = r_T n_T + r_N n_N + r_R n_R, \tag{2}$$

where

n_T, n_N and n_R = sample sizes under tightened, normal, and reduced inspection.

More recently, Schilling and Sheesley[17,18] used the Markov Model of Stephens and Larson to develop tables of system values for the AOQL, the limiting quality, the operating characteristics curve, the average sample number, the average outgoing quality, and the average total inspection. They have suggested incorporation of these curves in future revisions of MIL-STD-105D.

Schilling and Sheesley[17] made a number of observations. Three of these will now be discussed. First of all they state that:

"Unfortunately, the standard [MIL-STD-105D] is frequently misused, particularly in nonmilitary applications, through the selection and use of normal plans only—disregarding the tightened and reduced plans and the switching rules."[17]

Perhaps the major objective of their papers was to show the value of using all of the rules, which results in "enhanced protection for both the producer and consumer."[17]

They also note that the limiting quality, which is synonymous with the LTPD value, "is for use with isolated lots and does not reflect the limiting quality afforded by the MIL-STD-105D sampling system."[17]

Finally, they state that "reduced inspection provides an obvious reward to the producer of a good quality product in terms of lower sample size and slightly higher probability of acceptance."[17] But, they indicate that part of the switching procedures (the use of reduced limit numbers) have a minimal effect on the system OC curves.

Although the work cited in this section represented a breakthrough in the use of MIL-STD-105D, Hald and Thyregod,[15] Stephens and Larson,[16] and others recognized a shortcoming in their own approaches. The main criticism is that they considered a static situation with a fixed level of quality as input to the inspection system. As noted by W. R. Pabst in the discussion of Ref. 15:

"What would also be interesting, and perhaps more difficult to explore is the dynamic effect of these switching rules on the production process."[15]

Stephens and Larson agreed and stated that:

"Hence, the actual behavior of the process under the influence of the sampling procedure may thus be very dynamic."[16]

These comments influenced some of the work described in the next section and led the author to investigate the dynamic effects of the switching rules.

4.2 Controlling average outgoing quality

The average outgoing quality (AOQ) was chosen as the parameter of interest in the analysis of the dynamic effects of the switching rules. This is because it is a measure of the quality of the product supplied to the customer. Hence, AOQ is the "bottom line" for any sampling plan. The investigation in Ref. 19 centered on measurement of the effects of switching rule feedback to the supplier and the resultant quality provided to the customer. It was reasoned that the switching rules were included in 105D to cause the supplier to take actions resulting in improved quality when the tightened phase is entered:

"When the quality of a product degrades, the tightened inspection and discontinue features of MIL-STD-105D can be used to motivate the producer to improve."[19]

This is because fewer lots will be accepted under tightened inspection if quality does not improve. This may be observed in Fig. 6, which is a typical comparison of OC curves for a normal and tightened pair of plans. The difference between the probabilities of acceptance is shown in the curve at the bottom. For example, when quality equals 2.5 percent, this difference is about 25 percent. Hence, if quality remains at 2.5 percent defective during the tightened phase, about 25 percent more lots will be rejected than during the normal phase. This should act as a strong stimulus to the supplier to improve his quality. In addition, the discontinue phase and its consequences will be the end results of no improvement during the tightened phase.

The approach described in Ref. 19 was based on three assumptions which differed from those used in Refs. 16 to 18. First, the discontinue phase was included in the analysis because the threat of discontinue (when enforced) strengthens the effect of the tightened phase on the supplier. The premise was that the switching rules were meant to provide feedback to the supplier causing him to improve his quality. Secondly, it was assumed that a reasonable supplier does respond and that the average outgoing quality will be based on the level of improvement. Thus, two levels of quality were considered: (*i*) q_N, the quality during the normal phase, and (*ii*) q_T, the quality during the tightened phase. This is in marked contrast to the assumption of a single quality level in Refs. 16 to 18. Finally, only accepted lots were included in the analysis, whereas the previous work had been based on all inspected lots. Note that only accepted lots affect the quality received by the customer.

Under these assumptions, the equation for the average outgoing quality (AOQ) for the system of plans is

$$\text{AOQ} = \frac{q_N E_{AN}}{E_{AN} + E_{AT}} + \frac{q_T E_{AT}}{E_{AN} + E_{AT}}, \tag{3}$$

where

q_N, q_T = quality during normal, tightened phase

E_{AN}, E_{AT} = expected number of lots accepted during normal, and tightened phases.

If the level of response (q_T) is fixed and q_N is varied, the result is a curve that is very similar to the AOQL curve (Fig. 3). Then, if different levels of response are analyzed, we obtain a set of curves similar to those shown in Fig. 7. This figure gives an example of the AOQ for the system of 105D plans. The curves are for an AQL of 0.65 percent, a

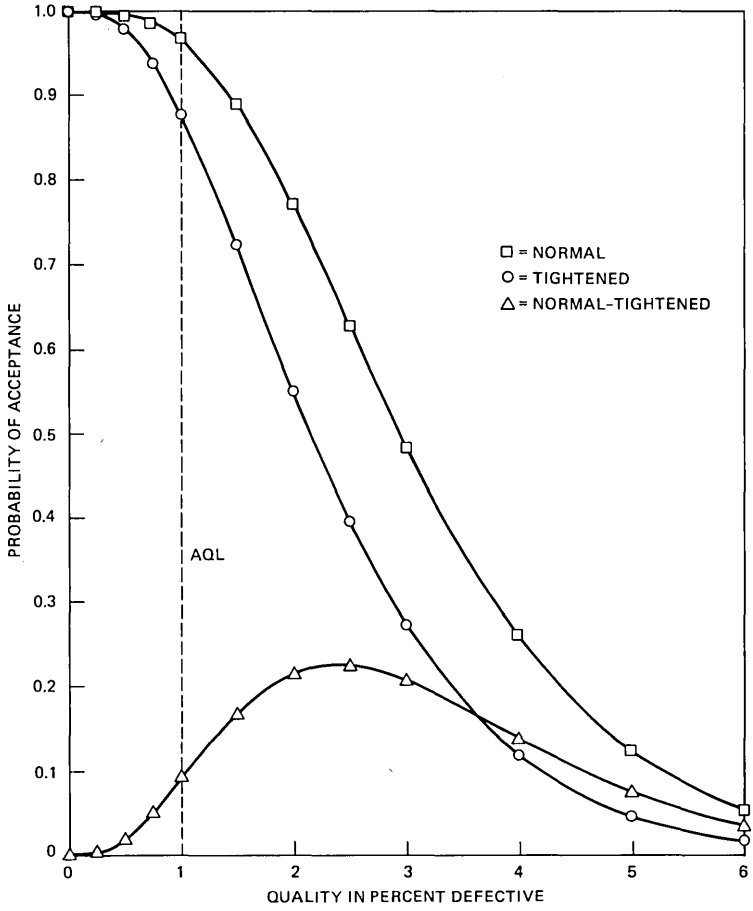


Fig. 6—Effects of tightened inspection.

sample size of 125, and an Ac of 2. There is one curve for each set of values of quality during the tightened phase (q_T). The maximum value on each curve is called the maximum average outgoing quality (AOQM) value.

Figure 8 is a sample plot of AOQM as a function of the level of response (q_T). The AOQM value is analogous to the AOQL value and if q_T can be estimated, AOQM represents a limit on the quality provided to the customer. The result of this analysis is a demonstration that "MIL-STD-105D has feedback properties which tend to limit the worst average outgoing quality without relying on the mechanism of screening rejected lots."[19]

4.3 Selection of specific plans from MIL-STD-105D

One of the main results of the AOQM analysis was a more comprehensive look at the properties of the 105D sampling plans. An under-

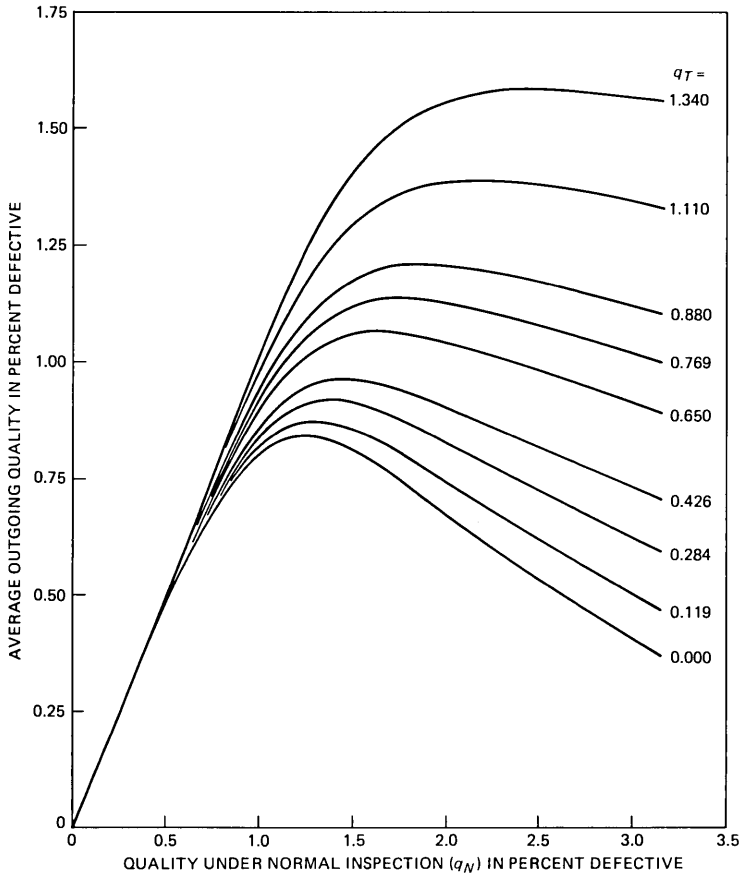


Fig. 7—Example of average outgoing quality for the system of MIL-STD-105D plans.

standing of these properties generated a desire to use them in the selection of specific plans from the 105D system.

The main result in Ref. 19 was a measure of the average outgoing quality based on the responsiveness of the supplier. Other measures of the plan capability were based on the ability of the 105D system to detect changes in quality. These were developed from the standpoint of control engineering. Professor Barnard's comments on Hill's paper point out that some of the early developers recognized the need for this type of analysis:

> "One particular thing which would have come out of this [view the scheme in terms of control engineering] would have been a description of the scheme, not in terms of the OC curve ··· but in terms of the average amount of production passed after quality has deteriorated before the deterioration is picked up by the inspection."[10]

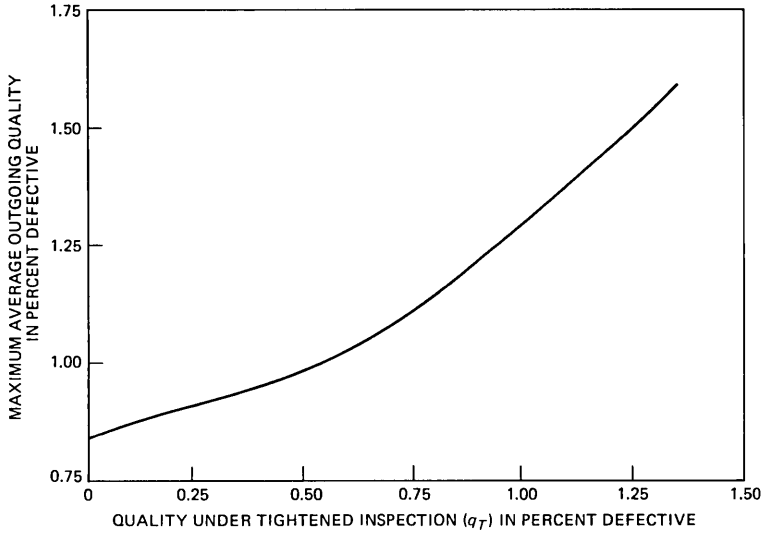


Fig. 8—Maximum average outgoing quality (AOQM).

The results in Ref. 19 led to the development of a preliminary set of criteria for evaluating MIL-STD-105D plans. Three quantities were considered in Ref. 20. First, the expected number of lots accepted during the normal phase prior to switching to tightened phase (E_{AN}) estimates both the speed of detecting a change to poor quality and the false alarm jeopardy when quality remains good. Secondly, the expected number of lots accepted during the reduced phase prior to switching back to the normal phase (E_{RN}) estimates the same quantities under reduced inspection. Both E_{AN} and E_{RN} are functions of quality during their respective phases. Finally, AOQM estimates the limiting quality leaving the inspection system. The value of AOQM chosen for this estimate assumes that the supplier will respond to a tightened quality level, q_T^*, which will have only a small chance of causing a switch to the discontinue phase.

Table I is a summary of the decision parameters developed to select sets of 105D plans. The value q_G in the table represents a good quality level while q_B represents a poor quality level. The value of q_B is typically three times that of q_G.

The computation of the parameters in Table I for the plans of MIL-STD-105D is a large task. However, the structure proposed by Dodge[11] and adopted in MIL-STD-105D (see Section 3.2) simplified this problem a great deal.

A set of normalized tables and curves were presented in Ref. 20 that reduces the required information by a factor of 14. When quality is normalized by the AQL value, a single curve is needed for each acceptance number rather than for each sampling plan. For example, the 152

Table I—Decision parameters for selection of 105D plans

Parameter	Phase	Purpose
1. $E_{AN}(q_G)$	Normal	Estimate false alarm rate during the normal phase.
2. $E_{AN}(q_B)$	Normal	Estimate response of a plan to a shift to bad quality during the normal phase.
3. AOQM(q_T^*)	Tightened	Estimate maximum average outgoing quality for a supplier's change to a safe quality q_T^*.
4. $E_{RN}(q_G)$	Reduced	Estimate false alarm rate during the reduced phase.
5. $E_{RN}(q_B)$	Reduced	Estimate response to a shift to bad quality during the reduced phase.

OC curves under normal inspection can be reduced to the eleven normalized curves in Fig. 9. These curves give the probability of acceptance as a function of the normalized quality. The units of normalized quality are multiples of the AQL value. Other curves and tables are provided in Ref. 20 which may be used to facilitate analysis of 105D sampling plans and help in the selection of appropriate AQL values.

4.4 Other results based on the structure of MIL-STD-105D

Analysts in the Bell Laboratories Quality Assurance Center recently investigated three questions which arose during the normal use of MIL-STD-105D by Western Electric inspectors. The first of these was concerned with the distribution of proportion defective in outgoing lots, assuming a distribution of quality in the incoming lots. Brush et al.* assume a beta distribution for incoming lot quality, and using the set of sampling plans from 105D, they find the mean, variance, and equivalent 0.90 beta quantile for the outgoing distribution.[21] Rejected lots are assumed to be scrapped. The special structure of 105D leads to a small set of normalized curves for the three outputs. These results provide the analyst with a powerful tool for determining the effect of a single sampling plan on the outgoing quality.

The second question was also resolved with the aid of the special structure of 105D. This is the multiple group situation.[22] Questions arose over the use of 105D when the set of inspection characteristics is divided into groups each with its own sampling plan. The MIL-STD-105D encourages this situation.[1,6] A conflict may arise because the supplier views quality in terms of each individual group, whereas the customer views quality in terms of collections of groups called categories. The ratio, k, of category AQL to the individual group AQL was determined in Ref. 22, assuming each group uses the same sampling plan. The special structure of 105D led to the development of tables of k as a function of the acceptance number and the number of groups in the category.

The third problem area was that of developing limiting quality or LTPD plans which are compatible with 105D plans. Again, the special

* Members of Bell Laboratories Quality Assurance Center.

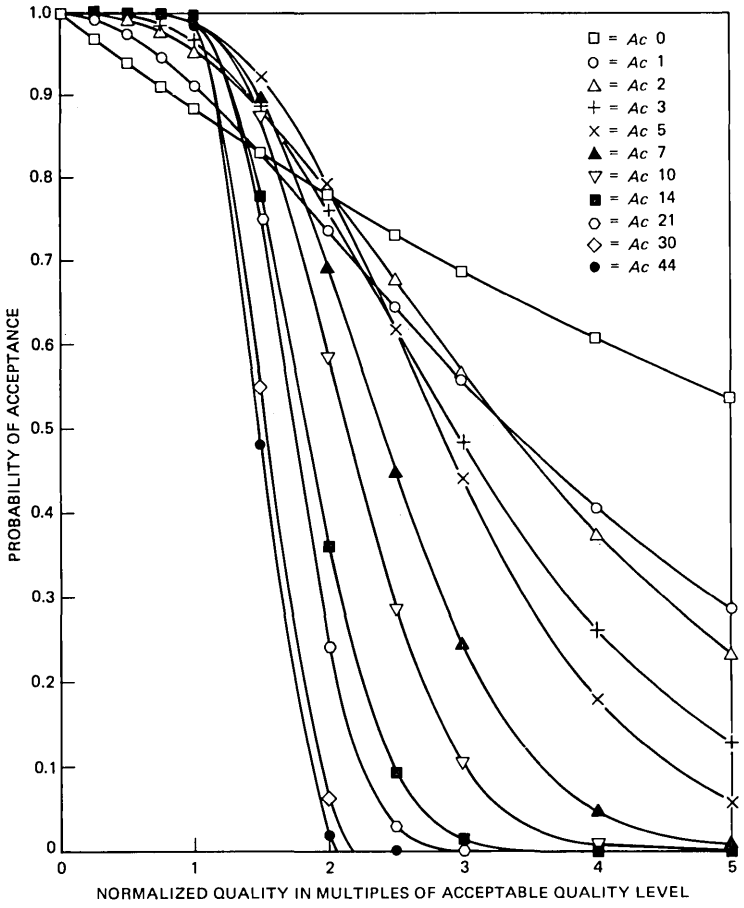


Fig. 9—Normalized OC curves.

structure of 105D aided in the resolution of this problem. The plans developed by Duncan, Mundel, Godfrey,* and Partridge* use the same lot size-sample size relationship as 105D.[23] In addition, the sequence of limiting quality levels that indexes the plans and the sample sizes follow the same geometric progression as 105D. The authors have proposed a table of these plans for inclusion in the next revision of ISO 2859.[24]

V. FUTURE RESEARCH

5.1 Open questions

The selection of the "best" sampling plans is certainly an open question. The analysis described in Section 3.3 gives the user a meth-

* Members of Bell Laboratories Quality Assurance Center.

odology for selection of an AQL value when costs are unknown. When costs are known, or partially known, the problem becomes one of including the effects of MIL-STD-105D on total cost. The most recent work in this area was given in a paper presented at the 1981 ASQC Quality Congress.[25] In that paper, a cost model and simplified procedures were developed for selecting AQL values. The analysis should be extended to include sample size variation and its effect on total cost.

A second open question is the amount of switching desirable during plan operation. Most of the interest in this question has centered in Japan, where researchers have developed a modification to 105D that includes different switching rules.[3] Little attention has been given to this modification in the United States.

A third open question is a comparison of the procedures of MIL-STD-105D with the procedures used by the Bell Laboratories Quality Assurance Center to audit Western Electric Company's manufactured product. A new reporting system for the audit, QMP, has recently been developed.[26] Future work should compare the cost basis and the feedback properties of the two systems.

Finally, the three results discussed in Section 4.4 should each be extended. First, the study of outgoing distributions should be extended to include the effects of the system of plans. The original study encompassed only individual plans. Secondly, the multiple group situation should be extended to groups using a mixture of sampling plans. And, finally, LQL plans should be incorporated in a system similar to 105D, or these plans should be combined with 105D plans into a complete attribute acceptance system.

Much of the early 105D development work is still open to review because of the nature of the system and its basis in compromise. In a continually changing environment, the 105D system may have to change. Good data are needed to evaluate many of the effects of 105D, while good analysts are needed to understand and extend the features of 105D.

GLOSSARY

Ac	Acceptance number; if this number is exceeded, either a lot is rejected or a switching rule is applied.
AOQ	The average outgoing quality; the average quality of all lots shipped.
AOQL	Upper bound of AOQ when all rejected lots are inspected 100 percent and all defectives are removed.
AOQM	Maximum average outgoing quality.
AQL	Acceptable quality level (good quality).
ARL	Average run length.
ASN	Average sample number.

Attribute	A characteristic of a product which is classified as nondefective or defective, good or bad, yes or no, etc.
Discontinue	The phase entered from the tightened phase when the supplier does not respond and improve his quality.
E_{AN}, E_{AT}, E_{RN}	Expected number of lots accepted during the normal, tightened, and reduced phases.
IQ	Incoming quality.
Lot	A set of items under control of the inspection organization for which an acceptance or rejection decision must be made.
Lot quality (q)	Percent defective or defects per hundred units in the lot.
LTPD	Lot tolerance percent defective (poor quality level). Also called the limit quality level (LQL) or limit quality (LQ).
MIL-STD-105D	Military Standard 105D.
n	Sample Size.
Normal phase	Entered when there is no evidence of better or poorer quality than desired.
Operating characteristic (OC) curve	Probability of acceptance as a function of quality.
OQ	Outgoing quality.
P_{ac}	System OC curve.
P_{aT}, P_{aN}, P_{aR}	Tightened, normal and reduced OC curves.
q_N, q_T	Quality level during normal, tightened phase.
r	The probability that an inspector will call a bad unit good.
r_N, r_T, r_R	The expected portion of all lots inspected during normal, tightened and reduced phases.
Random sample	A truly representative subset of a lot; each unit in the lot has the same probability of being included in the sample.
Reduced phase	Entered when there is evidence of good quality.
Screening	100 percent inspection of a rejected lot.
Sequential sampling	A sequence of samples are used during which the decisions are to accept the lot, reject the lot or take the next sample; the average total sample size is smaller than for comparable single sampling plans.
Switching rules	Rules for switching between phases.

Tightened phase 105D	Entered when there is evidence of poor quality. Military Standard 105D.

REFERENCES

1. "Sampling Procedures and Tables for Inspection by Attributes," MIL-STD-105D, April, 1963, Department of Defense, Washington, D.C.
2. W. R. Pabst, Jr., "Some Background Consideration of MIL-STD-105D," 17th Annual ASQC Convention and Exhibit, May 20, 1963, Chicago, Ill.
3. T. Koyama et al., "MIL-STD-105D and the Japanese Modified Standard," *J. Quality Tech.*, , No. 2 (April 1970), pp. 99–108.
4. E. M. Saniga and L. E. Shirland, "Quality Control in Practice—A Survey," Quality Progress, *10*, No. 4 (May 1977), pp. 30–3.
5. H. F. Dodge and H. G. Romig, *Sampling Inspection Tables, 2nd Edition*, Wiley, New York, 1959.
6. "Guide for Sampling Inspection," Handbook H-53, June 30, 1965, Office of the Assistant Secretary of Defense (Installations and Logistics), Washington, D.C. 20301.
7. A. J. Duncan, *Quality Control and Industrial Statistics*, 4th ed., Homewood, Illinois: Richard D. Irwin, Inc., 1974.
8. E. L. Grant and R. W. Leavenworth, *Statistical Quality Control*, 4th ed., New York: McGraw Hill Book Co., 1972.
9. O. A. Cocca, "MIL-STD-105D, An International Standard for Attribute Sampling," Ind. Quality Control, *21*, No. 5 (November 1964), pp. 249–53.
10. I. D. Hill, "Sampling Inspection and Defense Specification DEF-131," J. Royal Stat. Soc., Series A, *125*, Part 1, pp. 31–73, 1962.
11. H. F. Dodge, "A General Procedure for Sampling Inspection by Attributes—Based on the AQL Concept," Rutgers Statistics Center Technical Report No. 10, December 1959.
12. H. F. Dodge, "Notes on the Evolution of Acceptance Sampling Plans, Part III," J. Quality Tech. *1*, No. 4 (October, 1969), pp. 225–32.
13. H. F. Dodge, "Notes on the Evolution of Acceptance Sampling Plans, Part II," J. Quality Tech., *1*, No. 3 (July 1969), pp. 155–62.
14. H. F. Dodge, "Evaluation of a Sampling Inspection Having Rules for Switching Between Normal and Tightened Inspection," Technical Report No. 14, Rutgers University, August 1965.
15. A. Hald and P. Thyregod, "The Composite Operating Characteristic Under Normal and Tightened Sampling Inspection by Attributes," 35th Session Int. Stat. Inst., Beograd, 1965.
16. K. S. Stephens and K. E. Larson, "An Evaluation of the MIL-STD-105D System of Sampling Plans," Ind. Quality Control, *23*, No. 7 (January 1967), pp. 310–9.
17. E. G. Schilling and J. H. Sheesley, "The Performance of MIL-STD-105D Under Switching Rules, Part 1: Evaluation," J. Quality Tech., *10*, No. 2 (April 1978), pp. 76–83.
18. E. G. Schilling and J. H. Sheesley, "The Performance of MIL-STD-105D Under Switching Rules, Part 2: Tables," J. Quality Tech., *10*, No. 3 (April 1978), pp. 104–24.
19. B. S. Liebesman, "The Use of MIL-STD-105D to Control Average Outgoing Quality," J. Quality Tech., *11*, No. 1 (January 1979), pp. 36–43.
20. B. S. Liebesman, "The Characterization of MIL-STD-105D Sampling Plans Using Normalized OC Curves," Workshop on Statistical Quality Control, West Berlin, June 11–13, 1980.
21. G. G. Brush, H. Cautin, and B. R. Lewin. "Outgoing Quality Distributions for MIL-STD-105D Sampling Plans," J. Quality Tech., *13*, No. 4 (October 1981), pp. 254–63.
22. B. S. Liebesman, "The Quality Standard for Multiple Group Categories," 1980 ASQC Technical Conf. Trans., pp. 468–74.
23. A. J. Duncan et al., "LQL Indexed Plans that are Compatible with the Structure of MIL-STD-105D," J. Quality Tech., *12*, No. 2 (January 1980), pp. 40–6.
24. International Organization for Standardization, "Draft Collection of Sampling Plans and Tables to Complement ISO 2859," ISO/TC69/SC5, N69, London, September 1973.
25. B. S. Liebesman, "Selection of MIL-STD-105D Plans Based on Costs," 1981 ASQC Quality Congress Trans., San Francisco, California, May 1981, pp. 475–84.
26. A. B. Hoadley, "The Quality Measurement Plan (QMP)," B.S.T.J.., *60*, No. 2 (February 1981), pp. 215–73.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 2, February 1982
Printed in U.S.A.

Expansions for Nonlinear Systems*

By I. W. SANDBERG

(Manuscript received April 22, 1981)

In this paper we study operator-type models of dynamic nonlinear physical systems, such as communication channels and control systems. Attention is focused on the problem of determining conditions under which there exists a power-series-like expansion, or a polynomial-type approximation, for a system's outputs in terms of its inputs. Related problems concerning properties of the expansions are also considered and nonlocal, as well as local, results are given. In particular, we show for the first time the existence of a locally convergent Volterra-series representation for the input-output relation of an important large class of nonlinear systems containing an arbitrary finite number of nonlinear elements.

I. INTRODUCTION

In this paper we study operator-type models of dynamic nonlinear physical systems, such as communication channels and control systems. Attention is focused on the problem of determining conditions under which there exists a power-series-like expansion, or a polynomial-type approximation, for a system's outputs in terms of its inputs. Related problems concerning properties of the expansions are also considered and nonlocal, as well as local, results are presented. In particular, we show for the first time the existence of a locally convergent Volterra-series representation for the input-output relation of an important large class of nonlinear systems containing an arbitrary finite number of nonlinear elements.

With regard to background material, functional power series of the form

* The material given in this paper was described in the writer's "Conference Course" at the 1981 European Conference on Circuit Theory and Design, The Hague, August 1981.

$$k_0 + \sum_{m=1}^{\infty} \int_a^b \cdots \int_a^b k_m(t, \tau_1, \cdots, \tau_m) u(\tau_1) \cdots u(\tau_m) d\tau_1 \cdots d\tau_m, \quad (0)$$

in which k_0 is a constant, t is a parameter, and u and the k_m for $m \geq 1$ are continuous functions, were considered in 1887 by Vito Volterra[1,2] in connection with his studies of functions of functions. (These studies provided much of the initial motivation to develop the field now known as functional analysis.) About twenty years later, Fréchet[3] proved that a continuous real functional (i.e., a continuous real scalar-valued map) defined on a compact set of real continuous functions on $[a, b]$ could be approximated by a sum of a finite number of terms in Volterra's series (0), but with (in analogy with the well-known Weierstrass approximation theorem) the number of terms, as well as the k_m, dependent on the degree of approximation.

It was Norbert Wiener[4] who first used a Volterra-series representation in the analysis of a nonlinear system.* The form of Volterra's expansion provided also the basis for Wiener's later work (see, for example, Refs. 5 and 6) on nonlinear analysis and synthesis. His studies, which were concerned mainly with the modeling of systems when only input-output data (rather than the system's equations) are available, stimulated considerable interest concerning Volterra and other† functional expansions for nonlinear systems. It was appreciated from the outset that such expansions, when they exist, could provide important insight of a qualitative nature concerning the input-output behavior of a system, and that they could be useful in connection with, for example, the estimation and/or equalization of distortion caused by nonlinearities.

There is a fairly large literature related directly or indirectly to the material of the present paper (see, for example, Refs. 1 through 29 and the references cited there). In most cases the functional expressions considered are Volterra series (or truncated Volterra series). With regard to systems for which the governing equations are known, with relatively few exceptions, questions concerning the existence of an expansion, its convergence, and/or the nature of the approximation provided by a truncated series are either not addressed or are left unanswered. (See, for instance, the remarks in Ref. 6, p. 137, on the lack of understanding concerning convergence.)

On the other hand, some material regarding the range of validity and specific properties of functional expansions has appeared, both for systems governed by ordinary differential equations defined on a finite time interval $[0, \tau]$ (Refs. 15, 16, 20, and 23 are representative refer-

* In Ref. 4 Wiener considers the problem of evaluating the output moments of a specific type of detector circuit driven by a random input.

† See, for instance, Ref. 7.

ences), and for polynomic systems, which are modeled by operator generalizations of ordinary functions of polynomial form.[26-29] The work on differential equations is concerned mainly with the particular case of bilinear systems and with "linear-analytic" systems.[16,23] The main result obtained asserts that under certain conditions (see, for example, Ref. 23) there does exist a locally convergent Volterra series for the solution (but with the size of the region of convergence dependent on τ).

The studies of polynomic systems, which are operator theoretic in nature and which draw on the theory of multilinear forms, are more closely related to the results reported in this paper. The most pertinent earlier proposition[29] is one to the effect that if a certain contraction mapping condition is met, then it is possible to construct in a particular way a local inverse of a certain generalized power series. While we do not use previous results in the polynomic systems area, there are some points of contact with the earlier material, and this is discussed at appropriate places in Section II.

We now briefly outline the remainder of the paper. Section II begins with some mathematical preliminaries. In Section 2.1, we introduce the general setting of concern throughout the rest of Section II. This involves two maps f and g related by $f[g(u)] = u$ for u in a certain set that can be thought of as a set of system inputs such that each u contained produces an output $g(u)$. The remaining portion of Section II describes, proves, and discusses results concerning expansions and approximations of $g(u)$. The material in Section II is somewhat abstract. Examples which illustrate how the material can be used to obtain more specific results of general interest are given in Section III.

II. APPROXIMATIONS AND EXPANSIONS

Throughout the paper, $\mathscr{B}$ and $\mathscr{B}_0$ denote two Banach spaces, each with real or complex scalars, and X denotes a nonempty open subset of $\mathscr{B}_0$. We use the symbol $\|\cdot\|$ for the norm associated with $\mathscr{B}$, as well as for the norm associated with $\mathscr{B}_0$, and θ is used to denote the zero element of $\mathscr{B}$ and of $\mathscr{B}_0$.

It will become clear shortly that a central role in our development is played by first and higher order Fréchet derivatives (see, for instance, Ref. 30). Before proceeding, we recall a few pertinent facts and definitions.

Let F map X into $\mathscr{B}$, and let x_0 be a point in X. If there is a bounded linear map L_{x_0} from $\mathscr{B}_0$ to $\mathscr{B}$ such that $\|F(x_0 + h) - F(x_0) - L_{x_0}h\| = o(\|h\|)$ as $\|h\| \to 0$, then F is said to be *Fréchet differentiable* at x_0 with Fréchet derivative L_{x_0}, which we denote by $dF(x_0)$. If F is Fréchet differentiable at every point in X, then we say that F is differentiable on X. Similarly, if F is Fréchet differentiable on X and $dF(\cdot)$ is

continuous, then F is said to be *continuously differentiable* on X. Higher order Fréchet derivatives of F are defined in the usual inductive manner.

Note that the second-order Fréchet derivative $d^2F(x_0)$, when it exists, is a bounded linear map from $\mathscr{B}_0$ into the space $L(\mathscr{B}_0, \mathscr{B})$ of bounded linear maps from $\mathscr{B}_0$ into $\mathscr{B}$. For h_1 and h_2 in $\mathscr{B}_0$, $d^2F(x_0)h_1h_2$, by which we mean $[d^2F(x_0)h_1]h_2$, is an element of $\mathscr{B}$, and, therefore, $d^2F(x_0)$ can be regarded as a bilinear map from $\mathscr{B}_0 \times \mathscr{B}_0$ into $\mathscr{B}$, i.e., it can be *identified* in the obvious way with such a map. This bilinear map satisfies the symmetry condition that $[d^2F(x_0)h_1]h_2 = [d^2F(x_0)h_2]h_1$. In general,[30] the mth order Fréchet derivative $d^mF(x_0)$ for $m > 1$ is a bounded linear map from $\mathscr{B}_0$ into a Banach space of bounded linear maps with $\| d^mF(x_0) \|$ defined in the usual way in terms of induced norms, and $d^mF(x_0)$ can be regarded alternatively as a symmetric m-linear map from $\mathscr{B}_0^m$ into $\mathscr{B}$. Moreover, for $1 \leq l < m$ and $h_1, h_2, \cdots, h_l$ elements of $\mathscr{B}_0$, $d^mF(x_0)h_1h_2 \cdots h_l$ is a bounded linear map from $\mathscr{B}_0$ into a Banach space of bounded linear operators.

A result that we shall use is the following essentially standard inverse function proposition (Ref. 30, p. 273).*

Lemma 1: Let $F:X \to \mathscr{B}$ be continuously Fréchet differentiable on X, and let $x_0 \in X$. If $dF(x_0)$ is an invertible map of $\mathscr{B}_0$ onto $\mathscr{B}$, there is an open neighborhood $V \subset X$ of x_0 such that F restricted to V is a homeomorphism of V onto an open neighborhood of $F(x_0)$ in $\mathscr{B}$. In addition, if F is r times continuously differentiable on V, the inverse mapping G of $F(V)$ onto V is r times continuously differentiable on $F(V)$.

Also of importance in our work are derivatives of Banach-space-valued maps defined on an open subset S of the real or complex numbers. If F maps S into $\mathscr{B}$ and $s_0 \in S$, then F is said to have a *derivative* $dF(s_0)/ds$ at s_0 if $dF(s_0)/ds$ is an element of $\mathscr{B}$ and we have

$$\lim_{|\gamma| \to 0} \| \gamma^{-1}[F(s_0 + \gamma) - F(s_0)] - dF(s_0)/ds \| = 0.$$

Again, higher order derivatives are defined in the usual inductive way. Here derivatives are elements of $\mathscr{B}$.

2.1 f and g

Throughout the paper, $f:X \to \mathscr{B}$ denotes a map with the property that there is a nonempty open convex subset U of $\mathscr{B}$ such that for each $u \in U$, there is in X a unique x_u such that $f(x_u) = u$. (Recall that X is a nonempty open subset of $\mathscr{B}_0$).

---

* With regard to a difference in a hypothesis of Lemma 1 and the cited proposition in Ref. 30, we note that if A is a bounded linear invertible map of $\mathscr{B}_0$ onto $\mathscr{B}$ with inverse A^{-1}, then, by a result of Banach, this inverse is a *bounded* linear map of $\mathscr{B}$ onto $\mathscr{B}_0$.

We shall be concerned primarily with the map $g: U \to X$ defined by $f[g(u)] = u$ for every $u \in U$.

2.2 A representation theorem for $g(\cdot)$

We shall refer to the following two hypotheses:

H.1: For some positive integer p, the mth order F-derivative (i.e., Fréchet derivative) $d^m f$ exists and is continuous on X for $m = 1, 2, \cdots, (p+1)$.

H.2: $[df(x)]^{-1}$ exists [i.e., $df(x)$ is an invertible map of $\mathscr{B}_0$ onto $\mathscr{B}$] for $x \in X$.

Theorem 1: Suppose that H.1 and H.2 are met, and let u and u_0 be points in U. For $m = 1, 2, \cdots, p+1$ and each $\beta \in [0, 1]$, let $g_m(u_0, u - u_0, \beta)$ be defined as follows:

1.(a) $g_1(u_0, u - u_0, \beta) = df\{g[u_0 + \beta(u - u_0)]\}^{-1}(u - u_0)$

1.(b) $g_m(u_0, u - u_0, \beta) =$

$$-df\{g[u_0 + \beta(u - u_0)]\}^{-1} \sum_{l=2}^{m} (l!)^{-1} \sum_{\substack{k_1+k_2+\cdots+k_l=m \\ k_j>0}} \cdot d^l f\{g[u_0 + \beta(u - u_0)]\} g_{k_1}(u_0, u - u_0, \beta) g_{k_2}(u_0, u - u_0, \beta) \cdots g_{k_l}(u_0, u - u_0, \beta)^* \quad (1)$$

for $m = 2, \cdots, p+1$.

Then $g_{p+1}(u_0, u - u_0, \beta)$ depends continuously on β for $\beta \in [0, 1]$, and we have

$$g(u) = g(u_0) + g_1(u_0, u - u_0) + \cdots + g_p(u_0, u - u_0) + (p+1) \int_0^1 (1 - \beta)^p g_{p+1}(u_0, u - u_0, \beta) d\beta,$$

in which $g_m(u_0, u - u_0) = g_m(u_0, u - u_0, 0)$ for $m = 1, \cdots, p$.

Moreover,

2.(a) *there are positive constants ρ and σ, which do not depend on u, such that*

$$\left\| g(u) - g(u_0) - \sum_{m=1}^{p} g_m(u_0, u - u_0) \right\| \leq \rho \|u - u_0\|^{p+1} \quad \text{for} \quad \|u - u_0\| \leq \sigma,$$

2.(b) *there are positive constants $\rho_1, \rho_2, \cdots, \rho_p$, which do not depend on u, such that $\|g_m(u_0, u - u_0)\| \leq \rho_m \|u - u_0\|^m$ for $m = 1, 2, \cdots, p$, and*

---

* In (1), $\sum_{\substack{k_1+k_2+\cdots+k_l=m \\ k_j>0}}$ denotes a sum over all positive $k_1, \cdots, k_l$ that add to m.

2.(c) *for* $m = 1, 2, \cdots, p$ *we have* $g_m(u_0, u_a - u_0) = r^m g_m(u_0, u_b - u_0)$ *for* u_a *and* u_b *in* U *such that* $(u_a - u_0) = r(u_b - u_0)$ *for some real number* r.

Proof: By H.1, H.2, and Lemma 1, the mth order F-derivative $d^m g(w)$ exists and is continuous for $m = 1, 2, \cdots, (p + 1)$ and $w \in U$.

By a version of Taylor's theorem for Fréchet differentiable maps (Ref. 30, pp. 190–1),

$$g(u) = g(u_0) + dg(u_0)(u - u_0) + \frac{1}{2!} d^2 g(u_0)(u - u_0)^2 + \cdots + \frac{1}{p!} d^p g(u_0)(u - u_0)^p + \int_0^1 \frac{(1 - \beta)^p}{p!} d^{(p+1)} g[u_0 + \beta(u - u_0)](u - u_0)^{p+1} d\beta, \quad (2)$$

where $(u - u_0)^m$, for $m = 1, 2, \cdots, p + 1$, denotes $[(u - u_0), (u - u_0), \cdots, (u - u_0)]$ with m terms.

Since $d^m g(w)$ exists for $m = 1, 2, \cdots, p + 1$ for w belonging to the convex set U, $d^m g[u_0 + \beta(u - u_0)]/d\beta^m$ exists and is equal to $d^m g[u_0 + \beta(u - u_0)](u - u_0)^m$ for $\beta \in [0, 1]$ and $m = 1, 2, \cdots, p + 1$ [31, p. 198].*

Let $q(\beta)$ denote $g[u_0 + \beta(u - u_0)]$ for all real β such that $u_0 + \beta(u - u_0) \in U$, and let $q^{(m)}(\beta)$ stand for the mth derivative of $q(\beta)$ with respect to β at the arbitrary point $\beta \in [0, 1]$.

We have $f[q(\beta)] = u_0 + \beta(u - u_0)$ when $u_0 + \beta(u - u_0) \in U$. By a version of the chain rule (Ref. 31, p. 173) for the derivative of a composition function,

$$df[q(\beta)]q^{(1)}(\beta) = u - u_0, \qquad \beta \in [0, 1] \quad (3)$$

since $df[q(\beta)]$ and $q^{(1)}(\beta)$ exist for $\beta \in [0, 1]$. Thus,

$$dg[u_0 + \beta(u - u_0)](u - u_0) = q^{(1)}(\beta) = df[g(u_0 + \beta(u - u_0))]^{-1}(u - u_0), \qquad \beta \in [0, 1] \quad (4)$$

Now let $2 \le m \le (p + 1)$, and let $f_0: \mathscr{B}_0 \to \mathscr{B}$ and $q_0: (-\infty, \infty) \to \mathscr{B}_0$ be defined by

$$f_0(y) = \sum_{l=1}^{m} (l!)^{-1} d^l f[q(\beta)][y - q(\beta)]^l \quad (5)$$

---

* More specifically, since with $h = (u - u_0)$, $g[u_0 + (\beta + \sigma)h] = g(u_0 + \beta h) + dg(u_0 + \beta h)\sigma h + o(\|\sigma h\|)$ as $\sigma \to 0$ (for $\beta \in [0, 1]$), it is clear that $dg(u_0 + \beta h)/d\beta$ exists and is equal to $dg(u_0 + \beta h)h$ for $\beta \in [0, 1]$. Similarly, using $d^{m-1}g[u_0 + (\beta + \sigma)h] = d^{m-1}g(u_0 + \beta h) + d^m g(u_0 + \beta h)\sigma h + o(\|\sigma h\|)$ as $\sigma \to 0$ for $2 \le m \le (p + 1)$, we see that $d^m g(u_0 + \beta h)/d\beta^m = d^m g(u_0 + \beta h)h^m$ for $1 \le m \le (p + 1)$.

$$q_0(r) = q(\beta) + \sum_{l=1}^{m} (l!)^{-1}(r-\beta)^l q^{(l)}(\beta) \tag{6}$$

for $y \in \mathscr{B}_0$ and $r \in (-\infty, \infty)$, where $\beta \in [0, 1]$. We will use the following generalization of the classical rule for differentiating a product of two differentiable functions.

Proposition 1: With S a Banach space and A an open interval in $(-\infty, \infty)$ containing a point r_0, let $L(r)$ denote a bounded linear map from $\mathscr{B}_0$ into S for each $r \in A$, and let $e(\cdot)$ be a map from A to $\mathscr{B}_0$. If $dL(r_0)/dr$ and $de(r_0)/dr$ exist, then $d[L(r)e(r)]/dr$ exists at $r = r_0$, and $d[L(r)e(r)]/dr = L(r)de(r)/dr + [dL(r)/dr]e(r)$ at $r = r_0$.

A proof of Proposition 1 is given in Appendix A. Using Proposition 1 and the observation that $d^m f[q(\beta)]/d\beta^m = \theta$ for $\beta \in [0, 1]$ when $2 \leq m \leq (p+1)$, we show in Appendix B that $d^m\{f_0[q_0(r)]\}/dr^m|_{r=\beta} = \theta$ for $\beta \in [0, 1]$.

Since we have

$$f_0[q_0(r)] = \sum_{l=1}^{m} (l!)^{-1} d^l f[q(\beta)] \left(\sum_{k=1}^{m} (k!)^{-1}(r-\beta)^k q^{(k)}(\beta) \right)^l$$

for every r, and each $d^l f[q(\beta)]$ can be regarded as a l-linear operator on $\mathscr{B}_0^l$, it follows that for $\beta \in [0, 1]$:

$$\begin{aligned} \theta = d^m f_0[q_0(r)]/dr^m|_{r=\beta} \\ = m! \sum_{l=1}^{m} (l!)^{-1} \sum_{\substack{k_1+k_2+\cdots+k_l=m \\ k_j>0}} (k_1!k_2! \cdots k_l!)^{-1} \\ \cdot d^l f[q(\beta)]q^{(k_1)}(\beta)q^{(k_2)}(\beta) \cdots q^{(k_l)}(\beta). \end{aligned} \tag{7}$$

Referring now to the $g_m(u_0, u - u_0, \beta)$ in the statement of the theorem, notice that $g_1(u_0, u - u_0, \beta) = q^{(1)}(\beta)$ for $\beta \in [0, 1]$, and that, by (7), $g_m(u_0, u - u_0, \beta) = (m!)^{-1}q^{(m)}(\beta)$ for $2 \leq m \leq (p+1)$ and $\beta \in [0, 1]$. Therefore, using (2) and $q^{(m)}(\beta) = d^m g[u_0 + \beta(u - u_0)] \cdot (u - u_0)^m$ for $\beta \in [0, 1]$ and each m, we obtain the formula for $g(u)$ given in the theorem, in which $g_{p+1}(u_0, u - u_0, \beta)$, which is equal to

$$[(p+1)!]^{-1} d^{(p+1)} g[u_0 + \beta(u - u_0)](u - u_0)^{(p+1)},$$

depends continuously on β for $0 \leq \beta \leq 1$.

Since

$$g_m(u_0, u - u_0) = (m!)^{-1} d^m g(u_0)(u - u_0)^m$$

for $m = 1, 2, \cdots, p$ in which $d^m g(u_0)$ is an mth order Fréchet derivative,* it follows at once that properties 2.(b) and 2.(c) hold.

---

* In particular, $d^m g(u_0)$ and $d^m g(u_0)v_1 \cdots v_{m-l}$ for $1 \leq l \leq m-1$ and $v_j \in \mathscr{B}$ for $1 \leq j \leq m-l$ are bounded linear maps on $\mathscr{B}$.

Finally, and referring to (2), since

$$\| d^{(p+1)}g(u_0)(u-u_0)^{(p+1)} \| \leq \| d^{(p+1)}g(u_0) \| \cdot \| u-u_0 \|^{(p+1)},$$

property 2.(a) is a consequence of the result (Ref. 30, pp. 190–1) that for every $\sigma_0 > 0$ there is a $\sigma > 0$ such that

$$\left\| g(u) - dg(u_0)(u-u_0) - \frac{1}{2!} d^2g(u_0)(u-u_0)^2 - \cdots - \frac{1}{(p+1)!} d^{(p+1)}g(u_0)(u-u_0)^{(p+1)} \right\| \leq \sigma_0 \| u-u_0 \|^{(p+1)}$$

for $\| u-u_0 \| \leq \sigma$.

2.3 Corollary to Theorem 1

Corollary 1: Suppose that H.1 is met, and that $u_0 \in U$ is such that $df[g(u_0)]$ is invertible (i.e., is an invertible map of $\mathscr{B}_0$ onto $\mathscr{B}$). Then there are positive constants ρ and σ such that for $u \in U$ with $\| u-u_0 \| \leq \sigma$,

$$\| g(u) - g(u_0) - g_1(u_0, u-u_0) - \cdots - g_p(u_0, u-u_0) \| \leq \rho \| u-u_0 \|^{(p+1)},$$

in which $g_1(u_0, u-u_0), \cdots, g_p(u_0, u-u_0)$ are defined in Theorem 1. In addition 2.(b) and 2.(c) of Theorem 1 hold.

Proof: Since $df(\cdot)$ is continuous on X, and, by a theorem of Banach,[32] $df[g(u_0)]^{-1}$ is bounded, by a standard type of argument (see, for example, Ref. 30, pp. 154–5) $df(\cdot)^{-1}$ exists and is continuous* in some open neighborhood Γ of $g(u_0)$ in X. Also, since $df[g(u_0)]^{-1}$ exists, Γ contains an open neighborhood $N_{g(u_0)}$ and U contains an open neighborhood N_{u_0} of u_0 such that f restricted to $N_{g(u_0)}$ is a homeomorphism of $N_{g(u_0)}$ onto N_{u_0} (see Lemma 1). Let Ξ be a nonempty open convex subset of N_{u_0}. At this point the corollary follows from Theorem 1 with $X = N_{g(u_0)}$ and $U = \Xi$.

2.4 Comments

For a fixed u_0, the expansion given in Theorem 1 has the properties that the homogeneity condition 2.(c) is met and the remainder, the integral, is bounded above by $\rho \| u-u_0 \|^{(p+1)}$ for $\| u-u_0 \| \leq \sigma$ for some positive constants ρ and σ. A proof given in Ref. 33, p. 174 can easily be modified to show that the expansion is *unique* in the sense that there is no other similar [i.e., $g(u_0)$ plus p terms plus remainder] expansion of $g(u)$ valid for all $u \in U$ with these homogeneity and

* The continuity is not used in the present proof. It is used in Section 2.7 and in Appendix C, where reference is made to this proof.

remainder properties. Of course, a corresponding uniqueness proposition holds in the case of the truncated expansion in Corollary 1.

In some cases, $d^l f[g(u_0)]$ of Theorem 1 is the zero map whenever l is even and $2 \leq l \leq p$. Then $g_m(u_0, u - u_0) = \theta$ for m even with $2 \leq m \leq p$. This follows from a simple inductive argument using 1.(b) and the observation that $k_1 + k_2 + \cdots + k_l$ is an odd number if l is odd and each k_j is positive and odd.

Referring to Corollary 1, we can establish the existence of an expansion in a more general setting. Specifically, let $\mathscr{B}_1$ be a third Banach space and let $h(\cdot, \cdot)$ be a $(p + 1)$-times Fréchet continuously differentiable map of $S_0 \times S$ into $\mathscr{B}_1$, where S_0 and S are nonempty open subsets of $\mathscr{B}_0$ and $\mathscr{B}$, respectively. Let x_0 and u_0 be elements of S_0 and S, respectively, such that $h(x_0, u_0) = \theta$, in which here θ is used to denote the zero element of $\mathscr{B}_1$. Finally, assume that $D_1 h(x_0, u_0)$ the Fréchet partial derivative of $h(x, u)$ with respect to x, at the point (x_0, u_0), is an invertible map of $\mathscr{B}_0$ onto $\mathscr{B}_1$.

By the implicit function theorem in Ref. 30, p. 270 and a related proposition in Ref. 30, Result (10.2.3), it follows that there is an open convex neighborhood N of u_0 in S, and a $(p + 1)$-times Fréchet continuously differentiable map w of N into S_0 such that $w(u_0) = x_0$ and $h[w(u), u] = \theta$ for $u \in N$.* Therefore (2), with g replaced with w, is a representation about u_0 of w valid for $u \in N$ (Ref. 30, pp. 190–1). It can be shown that the terms in the representation can be determined by successively differentiating $h\{w[u_0 + \beta(u - u_0)], u_0 + \beta(u - u_0)\}$ with respect to β and setting the result equal to θ. [Recall that $D_1 h(x_0, u_0)$ is assumed to be invertible, and see the proof of Theorem 1.]†

The proof of Theorem 1 shows that the recursive relation (1) arises in a natural way. Such formulas concerning the inversion of ordinary power series and/or the derivatives of composite ordinary functions are probably well known in some circles. In Ref. 29, similar relations are given in an abstract setting for the different problem of constructing a local inverse of a mapping that has a power series expansion.

The expansion of $g(u)$ in Theorem 1, and its associated truncation in Corollary 1, each contains a constant term $g(u_0)$, a term $g_1(u_0, u - u_0)$ that can be written as $L_{u_0}(u - u_0)$ in which L_{u_0} is a bounded linear map, and a sum $R_{u_0}(u - u_0)$ of higher order terms such

---

* Also, N can be chosen so that w is the *only* continuous map of N into S_0 such that $w(u_0) = x_0$ and $h[w(u), u] = \theta$ for $u \in N$.

† Similar remarks apply also in the case of Theorem 4, below. By applying either Theorem 2 or Theorem 5, below, to the map $H: S_0 \times S \to B_1 \times B$ defined by $H(x, v) = [h(x, v), v]$ for $(x, v) \in S_0 \times S$, the writer has obtained an *explicit* expansion, involving partial derivatives of $h(\cdot, \cdot)$, for the solution x of $h(x, u) = w$ in terms of u and w, under certain reasonable assumptions concerning $h(\cdot, \cdot)$ and the sets from which u and w are drawn. The details will be given in another paper.

that $R_{u_0}(h) = o(\|h\|)$ as $\|h\| \to 0$. The following result shows that the hypothesis of Theorem 1 that H.2 is met, and of Corollary 1 that $df[g(u_0)]$ is invertible, are not merely ones of convenience which allow an explicit expression to be given for the terms.

Proposition 2: Let u_0 be an element of U such that $df[g(u_0)]$ exists [respectively, such that f is continuously F-differentiable on a neighborhood of $g(u_0)$]. Suppose that there is a constant $\sigma > 0$ such that

$$g(u) = g(u_0) + L(u - u_0) + R(u - u_0) \tag{8}$$

for $u \in U$ with $\|u - u_0\| < \sigma$, in which L is a bounded linear map from $\mathscr{B}$ into $\mathscr{B}_0$, and $R(\cdot)$ is a map of $\mathscr{B}$ into $\mathscr{B}_0$ with the property that $R(h) = o(\|h\|)$ as $\|h\| \to 0$. Then $df[g(u_0)]^{-1}$ exists [respectively, $df(x)^{-1}$ exists and is continuous in x for x in some neighborhood of $g(u_0)$].

The proposition is proved in Appendix C. It shows, for example, that H.2 is a consequence of the hypotheses that $g(U)$ is an open set, $X = g(U)$, f is differentiable on X, and (8) holds for each $u_0 \in U$. In this connection, notice that because U is open, $g(U)$ is open under merely the condition that f is *continuous* on X.

2.5 Results for complex Banach spaces

Theorem 2: Suppose that $\mathscr{B}$ and $\mathscr{B}_0$ are over the complex field, that the F-derivative $d^m f(x)$ exists at each $x \in X$ for all m, and that H.2 is met. Let $u_0 \in U$, and let ρ be a positive constant with the property that $u_0 + v \in U$ for $\|v\| < \rho$. Then for $u \in U$ such that $\|u - u_0\| < \rho$, we have

$$g(u) = g(u_0) + \sum_{m=1}^{\infty} g_m(u_0, u - u_0), \tag{9}$$

in which

$$g_1(u_0, u - u_0) = df[g(u_0)]^{-1}(u - u_0), \tag{10}$$

and

$$g_m(u_0, u - u_0) = -df[g(u_0)]^{-1} \sum_{l=2}^{m} (l!)^{-1} \sum_{\substack{k_1+k_2+\cdots+k_l=m \\ k_j>0}} d^l f[g(u_0)] g_{k_1} \cdot (u_0, u - u_0) g_{k_2}(u_0, u - u_0) \cdots g_{k_l}(u_0, u - u_0), \qquad m \geq 2. \tag{11}$$

Proof: Here $d^m g$ exists on U for each m (see the proof of Theorem 1). In particular, dg exists throughout U, and therefore we see that

$$\lim_{z \to 0} z^{-1}[g(x + zh) - g(x)]$$

exists in $\mathscr{B}_0$ (and equals $dg(x)h$) for each $x \in U$ and $h \in \mathscr{B}$, where z is

a *complex* scalar variable. Thus, by Theorem 3.16.2 of Ref. 34, p. 111 and the development preceding it,

$$d^m g[u_0 + z(u - u_0)]/dz^m|_{z=0}$$

exists for each $m \geq 1$, and we have

$$g(u) = g(u_0) + \sum_{m=1}^{\infty} (m!)^{-1} d^m g[u_0 + z(u - u_0)]/dz^m|_{z=0}$$

for $\|u - u_0\| < \rho$.

Notice that

$$d^m g[u_0 + z(u - u_0)]/dz^m|_{z=0} = d^m g[u_0 + r(u - u_0)]/dr^m|_{r=0}$$

in which r is a real variable. Since (see the proof of Theorem 1)

$$d^m g[u_0 + r(u - u_0)]/dr^m|_{r=0} = m! g_m(u_0, u - u_0),$$

the proof is complete.

Theorem 3: Let the hypotheses of Theorem 2 *hold. Then for each* u_0 *there is a* $\sigma > 0$ *such that the series on the right side of* (9) *converges* uniformly *for* $\|u - u_0\| < \sigma$.

Proof: The proof of Theorem 2 shows, using the openness of U, that for each u_0 there is a $\rho > 0$ such that the series converges for $\|u - u_0\| < \rho$. Since dg exists on U, g is continuous on U. The map g is, therefore, *locally bounded* on U in the sense of Ref. 34, Definition 3.17.1, and thus the proof of Theorem 3.17.1 of Ref. 34, p. 112, shows that, given u_0, there is a $\sigma > 0$ such that the convergence is uniform for $\|u - u_0\| < \sigma$.

The following result is obtained from Theorems 2 and 3 in the same way that Corollary 1 is proved.

Theorem 4: Assume that $\mathcal{B}$ *and* $\mathcal{B}_0$ *are over the complex field, and that* $d^m f$ *exists on* X *for each* m. *Let* $u_0 \in U$, *and suppose that* $df[g(u_0)]$ *is an invertible map of* $\mathcal{B}_0$ *onto* $\mathcal{B}$. *Then there is a* $\sigma > 0$ *such that the expansion*

$$g(u) = g(u_0) + \sum_{m=1}^{\infty} g_m(u_0, u - u_0)$$

is valid and uniformly convergent for $u \in U$ *with* $\|u - u_0\| < \sigma$, *where* $g_1(u_0, u - u_0)$, $g_2(u_0, u - u_0)$, $\cdots$ *are defined by* (10) *and* (11).

2.6 Comments

Under the conditions of Theorem 2 (respectively, Theorem 4) the infinite sum $R(u - u_0)$of the terms $g_2(u_0, u - u_0)$, $g_3(u_0, u - u_0)$, $\cdots$ has the property that $R(h) = o(\|h\|)$ as $\|h\| \to 0$. (This follows from the fact that $dg(u_0) = df[g(u_0)]^{-1}$.) Therefore, Proposition 2, as well as remarks similar to those of Section 2.4, apply here too with regard

to the necessity of the hypothesis that H.2 is met (respectively, $df[g(u_0)]$ is invertible).

Following is an interesting corollary of Theorems 2 and 3.

Corollary 2: Suppose that $\mathscr{B}$ and $\mathscr{B}_0$ are complex Banach spaces, and that f is a C^∞-diffeomorphism of X onto U (i.e., that f is a homeomorphism of X onto U such that f and its inverse g have F-derivatives of all orders on X and U, respectively). Let $u_0 \in U$, and let ρ be a positive constant with the property that $u_0 + v \in U$ for $\|v\| < \rho$. Then the series representation (9), *in which the $g_m(u_0, u - u_0)$ are given by* (10) *and* (11), *is valid for $\|u - u_0\| < \rho$, and there is a $\sigma > 0$ such that the series on the right side of* (9) *converges uniformly for $\|u - u_0\| < \sigma$.*

Proof: Since $(df)^{-1}$ exists on X under the conditions of Corollary 2*, Corollary 2 follows from Theorems 2 and 3.

2.7 Discussion

Uniqueness propositions similar to the one described in Section 2.4 apply in the cases of Theorems 2 and 4, as well as Corollary 2. Consider, for example, Theorem 2 and assume that its hypotheses are met. From (10) and (11) (or from the proofs of Theorems 1 and 2), we see that the expansion on the right side of (9) has the homogeneity property that for each m, $g_m(u_0, u_a - u_0) = r^m g_m(u_0, u_b - u_0)$ for $u_a, u_b \in U$ such that $\|u_a - u_0\| < \rho$, $\|u_b - u_0\| < \rho$, and $(u_a - u_0) = r(u_b - u_0)$ for some real r. Suppose that $g(u_0) + \sum_{m=1}^{\infty} h_m(u_0, u - u_0)$ is also an expansion of $g(u)$ about u_0 valid for $\|u - u_0\| < \rho$, and that it has the corresponding homogeneity property. Assuming, for the purpose of induction that $h_m(u_0, u - u_0) = g_m(u_0, u - u_0)$ for $\|u - u_0\| < \rho$ and $1 \le m \le n$ for some nonnegative integer n, we see† that for any fixed u such that $\|u - u_0\| < \rho$,

$$h_{n+1}(u_0, u - u_0) - g_{n+1}(u_0, u - u_0) = \sum_{m=n+2}^{\infty} [g_m(u_0, u - u_0) - h_m(u_0, u - u_0)] r^{(m-n-1)} \quad (12)$$

for $0 < |r| < 1$. Since

$$\sup_m \|g_m(u_0, u - u_0) - h_m(u_0, u - u_0)\|$$

---

* We have $f[g(u)] = u$ and $g[f(x)] = x$ for each $u \in U$ and each $x \in X$. Thus, by a version of the chain rule for differentiating a composite function (Ref. 31, pp. 171–2), $df[g(u)]dg(u) = I$ and $dg[f(x)]df(x) = I_0$ for each u and x, where I and I_0 are the identity maps on $\mathscr{B}$ and $\mathscr{B}_0$, respectively. This shows that $df(x)$ has both a right inverse and a left inverse for each $x \in X$, and, therefore, that $(df)^{-1}$ exists on X.

† This type of observation is used in Ref. 33, p. 174, to prove the uniqueness result given there.

is finite, the right side of (12) approaches zero as $r \to 0$. Thus $h_{n+1} \cdot (u_0, u - u_0) = g_{n+1}(u_0, u - u_0)$, and, therefore, $h_m(u_0, u - u_0) = g_m(u_0, u - u_0)$ for all m and all u such that $\| u - u_0 \| < \rho$.

The proof of Theorem 2 is based in part on basically well-known results concerning Banach space valued functions of a complex variable. Such results are used also in, for example, Refs. 20 and 24 for related but different purposes.

A comparison of Theorems 1 and 4 leads us to ask whether Theorem 1 can be used to prove a result along the lines of Theorem 4 for cases in which $\mathscr{B}$ and $\mathscr{B}_0$ are not necessarily over the complex field, but the $\| d^m f(x) \|$ are sufficiently small in some not too restrictive sense for x near $g(u_0)$. In this connection, we have the following.

Theorem 5: Let $d^m f$ exist on X for each m, and let u_0 be a point in U. Suppose that there are positive constants δ and γ, and a neighborhood N_0 of $g(u_0)$ in X, such that $\| d^m f(x) \| \leq m! \delta \gamma^m$ for $x \in N_0$ and every $m \geq 2$. Assume that $df[g(u_0)]$ is an invertible map from $\mathscr{B}_0$ onto $\mathscr{B}$. Then the conclusion of Theorem 4 holds.

Proof: By the proof of Corollary 1, it suffices to show that there is a $\sigma > 0$ such that

$$(p + 1) \int_0^1 (1 - \beta)^p g_{p+1}(u_0, u - u_0, \beta) d\beta,$$

the remainder in the expansion for $g(u)$ of Theorem 1, approaches θ as $p \to \infty$ uniformly for $\| u - u_0 \| < \sigma$. Since $\int_0^1 (1 - \beta)^p d\beta = (p + 1)^{-1}$, it is enough to prove that there are constants $\sigma > 0$, $d > 0$, and $c > 0$ such that for all p, all $\beta \in [0, 1]$, and all $u \in U$ with $\| u - u_0 \| < \sigma$, we have $\| g_p(u_0, u - u_0, \beta) \| \leq c e^{-dp}$. That we do as follows.

By the continuity of g at u_0, and the continuity of $(df)^{-1}$ at $g(u_0)$ (see the first part of the proof of Corollary 1), choose $\sigma > 0$ so that U contains the open ball of radius σ centered at u_0, and $\| df\{g[u_0 + \beta(u - u)]\}^{-1} \| \leq c_1$ and $\| d^l f\{g[u_0 + \beta(u - u_0)]\} \| \leq l! \delta \gamma^l$ for some constant c_1 whenever $\beta \in [0, 1]$ and $\| u - u_0 \| < \sigma$.

Choose any positive number d. For each m, let h_m denote $\sup\{\| e^{dm} g_m(u_0, u - u_0, \beta) \| : \beta \in [0, 1], \| u - u_0 \| < \sigma\}$. From Parts 1.(a) and 1.(b) of Theorem 1, and our hypotheses, the h_m are finite, and we have

$$\sum_{m=1}^{p} h_m \leq e^d c_1 \sigma + c_1 \delta \sum_{m=2}^{p} \sum_{l=2}^{m} \sum_{\substack{k_1 + k_2 + \cdots + k_l = m \\ k_j > 0}} \gamma^l h_{k_1} h_{k_2} \cdots h_{k_l}$$

for each $p > 1$. Since

$$\left(\sum_{m=1}^{p-1} h_m \right)^l = \sum h_{k_1} h_{k_2} \cdots h_{k_l}$$

over $(k_1, k_2, \cdots, k_l) \in \{1, 2, \cdots, (p-1)\}^l$, it easily follows that

$$\sum_{m=2}^{p} \sum_{l=2}^{m} \sum_{\substack{k_1+k_2+\cdots+k_l=m \\ k_j>0}} \gamma^l h_{k_1} h_{k_2} \cdots h_{k_l} \leq \sum_{l=2}^{p} \gamma^l \left(\sum_{m=1}^{p-1} h_m \right)^l.$$

Thus, with

$$s_p = \sum_{m=1}^{p} h_m,$$

we have

$$s_p \leq e^d c_1 \sigma + c_1 \delta \sum_{l=2}^{p} (\gamma s_{p-1})^l$$

for $p > 1$.

Now let $c > 0$ be chosen such that

$$c_1 \delta \sum_{l=2}^{\infty} (\gamma c)^l \leq \frac{1}{2} c,$$

and, if necessary, reduce σ so that $e^d c_1 \sigma \leq \frac{1}{2}c$. Since $s_1 \leq \frac{1}{2}c$, and $s_{(p-1)} \leq c$ implies that $s_p \leq c$ for $p > 1$, it is clear that $s_p \leq c$ (and hence $h_p \leq c$) for all p, which completes the proof.

2.8 Comments

Since the hypotheses of Theorem 5 can be shown to ensure the local existence of a Fréchet "power series" expansion (see Ref. 30, p. 190) of f about $g(u_0)$, another way to prove Theorem 5 is to use the result stated in Ref. 29. The observation concerning the representation of $(s_{p-1})^l$ as a sum of products used in our proof to obtain the inequality involving s_p and s_{p-1} (but not the exponential weighting approach) is used also in Ref. 27 for a case that corresponds here to the one in which only a finite number of the $\| d^l f \|$ do not vanish.

Theorem 1 can also be used to prove *nonlocal* convergence results when $\mathscr{B}$ and $\mathscr{B}_0$ are not necessarily complex spaces. For example, let ρ be the radius of any finite open ball contained in U and centered at u_0. Suppose that for every $x \in X$ the following is true: $d^m f(x)$ exists for each m, $df(x)^{-1}$ exists, and $\| df(x)^{-1} \| \leq \rho_0$ for some constant ρ_0. Then, using Theorem 1, it can be shown that if the $\| d^m f(\cdot) \|$ satisfy certain smallness conditions on X for $m \geq 2$, the expansion described in the conclusion of Theorem 2 converges uniformly to $g(u)$ for $\| u - u_0 \| < \rho$. The "smallness conditions" are met if, for example,

$$\sup_{x \in X} \| d^m f(x) \| = 0 \quad \text{for} \quad m \geq M \quad \text{for some} \quad M \geq 2,$$

and each nonzero

$$\sup_{x \in X} \| d^m f(x) \|$$

with $m > 1$ is sufficiently small. The details and a proof will be given in a later paper.

2.9 Properties 1 and 2

We conclude this section with a proof of a proposition used in Section III, where attention is directed to cases in which the elements of $\mathscr{B}$ and $\mathscr{B}_0$ are functions of a time variable t. The proposition is used to show that under certain very reasonable conditions, causality and time invariance (or periodicity of variation)* are properties which, when possessed by g, are inherited by the terms $g_1(u_0, u - u_0), \cdots, g_p(u_0, u - u_0)$ in Theorem 1, and by the terms $g_1(u_0, u - u_0), g_2(u_0, u - u_0), \cdots$ in Theorem 2. We first introduce some preliminaries.

Let Ω denote a nonempty set of real numbers. For each $\omega \in \Omega$, let T_ω and $T_{0\omega}$ denote linear transformations of $\mathscr{B}$ and $\mathscr{B}_0$, respectively, such that $\| T_{0\omega} w \| \leq \| w \|$ for $\omega \in \Omega$ and $w \in \mathscr{B}_0$. Let S be a subset of $\mathscr{B}$ such that $T_\omega S \subseteq S$ for $\omega \in \Omega$. Let J denote an open interval in the set R^1 of real numbers.

We say that a map $F: J \times S \to \mathscr{B}_0$ has *Property 1* on S at a point $r \in J$ if

$$T_{0\omega} F(r, v) = T_{0\omega} F(r, T_\omega v)$$

for all $v \in S$ and $\omega \in \Omega$. Finally, we say that $F: J \times S \to \mathscr{B}_0$ has *Property 2* on S at a point r in J if

$$T_{0\omega} F(r, v) = F(r, T_\omega v)$$

for $v \in S$ and $\omega \in \Omega$.

Proposition 3: Suppose that $F: J \times S \to \mathscr{B}_0$ has Property 1 (respectively, Property 2) on S for each $r \in J$, and that for an arbitrary $v \in S$ the derivative $dF(r, v)/dr$ exists at each $r \in J$. Then the map $H: J \times S \to \mathscr{B}_0$, defined by $H(r, v) = dF(r, v)/dr$ for each r and each v, has Property 1 (respectively, Property 2) on S for each $r \in J$.

Proof: Assume initially that F has Property 1. Let arbitrary $r \in J$ and $v \in S$ be given and let β be a real variable. Using

$$\lim_{\beta \to 0} \| \beta^{-1} [F(r + \beta, v) - F(r, v)] - H(r, v) \| = 0,$$

we have

$$\lim_{\beta \to 0} \| \beta^{-1} [T_{0\omega} F(r + \beta, v) - T_{0\omega} F(r, v)] - T_{0\omega} H(r, v) \| = 0 \quad (13)$$

---

* See Section 3.1 for the pertinent definitions.

for any $\omega \in \Omega$. Since (13) holds also with v replaced with $T_\omega v$, and using the hypotheses that F has Property 1 on S at r and at $(r + \beta)$ for sufficiently small β, we find that

$$\lim_{\beta \to 0} \| \Delta(\beta) + T_{0\omega} H(r, v) - T_{0\omega} H(r, T_\omega v) \| = 0,$$

in which $\Delta(\beta) = \beta^{-1}[T_{0\omega} F(r + \beta, v) - T_{0\omega} F(r, v)] - T_{0\omega} H(r, v)$. By (13), $\| \Delta(\beta) \| \to 0$ as $\beta \to 0$. Therefore, we have $T_{0\omega} H(r, v) = T_{0\omega} H(r, T_\omega v)$ for arbitrary $\omega \in \Omega$, as claimed. The Property 2 part of the proposition can be proved in essentially the same way.

III. APPLICATIONS AND EXAMPLES

Throughout this section, we consider cases where each element of $\mathscr{B}$, and also of $\mathscr{B}_0$, is a function of a time variable t. Specifically, we now assume that each element of $\mathscr{B}$ is a map from a set T of real numbers into a linear space V with zero element θ_V, and, similarly, that the elements of $\mathscr{B}_0$ are maps from T into a linear space V_0 with zero element θ_V.

We shall be concerned mainly with the cases where either $\mathscr{B} = \mathscr{B}_0 = L_\infty(\mathrm{R})$ or $\mathscr{B} = \mathscr{B}_0 = L_\infty(\mathrm{C})$, in which by $L_\infty(\mathrm{R})$ [respectively, $L_\infty(\mathrm{C})$] we mean the real (respectively, complex) Banach space of (Lebesgue) measurable* real (respectively complex) column n-vector valued functions v defined on the interval $[0, \infty)$ such that the jth component v_j of v satisfies

$$\sup_{t \geq 0} |v_j(t)| < \infty \quad \text{for} \quad j = 1, 2, \cdots, n,$$

and where the norm $\| \cdot \|$ on $L_\infty(\mathrm{R})$ or $L_\infty(\mathrm{C})$ is given by

$$\| v \| = \max_j \sup_t |v_j(t)|.$$

(As usual, n denotes an arbitrary positive integer.) If, for example, $\mathscr{B} = \mathscr{B}_0 = L_\infty(\mathrm{R})$, then $T = [0, \infty)$ and we can take V and V_0 to be R^n.

3.1 Causality and time-invariance

Referring to Proposition 3 and the associated definitions, let $\Omega = T$, and initially let T_ω (respectively, $T_{0\omega}$) be the "time-truncation" operator defined on $\mathscr{B}$ (respectively, $\mathscr{B}_0$) by $(T_\omega v)(t) = v(t)$ for $t \leq \omega$, and $(T_\omega v)(t) = \theta_V$ for $t > \omega$ (respectively, by $(T_{0\omega} v)(t) = v(t)$ for $t \leq \omega$, and $(T_{0\omega} v)(t) = \theta_{V_0}$ for $t > \omega$) for each v and each ω. Assume that for $\omega \in T$, T_ω and $T_{0\omega}$ map $\mathscr{B}$ and $\mathscr{B}_0$ into themselves, and that $\| T_{0\omega} v \| \leq \| v \|$

* See, for example, Ref. 35.

for $\omega \in T$ and all v. [Notice that these assumptions are satisfied if, for example, $\mathscr{B} = \mathscr{B}_0 = L_\infty(\mathrm{R})$ or $\mathscr{B} = \mathscr{B}_0 = L_\infty(\mathrm{C})$.]*

Suppose that the hypotheses of Theorem 1 are met, that U is an open ball centered at $u_0 = \theta$, that $T_\omega U \subseteq U$ for $\omega \in T$ [which is clearly satisfied if $\mathscr{B} = L_\infty(\mathrm{R})$ or $L_\infty(\mathrm{C})$] and initially assume that g is causal on U in the sense that $T_{0\omega} g(v) = T_{0\omega} g(T_\omega v)$ for any $v \in U$ and $\omega \in T$.

Let $G_m(u)$ denote $g_m(\theta, u)$ of Theorem 1 for $m = 1, 2, \cdots, p$ and $u \in U$. We observe that $g(\beta u)$ is an element of $\mathscr{B}_0$ for each $\beta \in (-1, 1)$ and each $u \in U$. By the proof of Theorem 1, $d^m g(\cdot)$ exists on U for $m = 1, 2, \cdots, p$, from which it follows that $d^m g(\beta u)/d\beta^m$ exists for $\beta \in (-1, 1)$, $u \in U$, and $m = 1, 2, \cdots, p$ [31, p. 198]. By the proof of Theorem 1, $m! G_m(u) = d^m g(\beta u)/d\beta^m$ at $\beta = 0$ for each m and u. Therefore, by the Property 1 part of Proposition 3 [with $S = U$ and $J = (-1, 1)$] and an obvious inductive argument, it follows that each $G_m: U \to \mathscr{B}_0$ is causal in the same sense that g is causal.†

Now suppose that T is one of the four sets $[0, \infty)$, $(-\infty, \infty)$, $\{0, 1, 2, \cdots\}$, or $\{0, \pm 1, \pm 2, \cdots\}$. Again take $\Omega = T$. Let T_ω (respectively, $T_{0\omega}$) denote the "time delay operator" defined by $(T_\omega v)(t) = \theta_V$ for $t < \omega$ and $(T_\omega v)(t) = v(t - \omega)$ for $t \geq \omega$ when either $T = [0, \infty)$ or $T = \{0, 1, 2, \cdots\}$, and by $(T_\omega v)(t) = v(t - \omega)$ when $T = (-\infty, \infty)$ or $T = \{0, \pm 1, \pm 2, \cdots\}$ (respectively, $(T_{0\omega} v)(t) = \theta_{V_0}$ for $t < \omega$ and $(T_{0\omega} v)(t) = v(t - \omega)$ for $t \geq \omega$ if T is either $[0, \infty)$ or $\{0, 1, 2, \cdots\}$, and $(T_{0\omega} v)(t) = v(t - \omega)$ when T is either $(-\infty, \infty)$ or $\{0, \pm 1, \pm 2, \cdots\}$). Assume here, as above, that T_ω and $T_{0\omega}$ map $\mathscr{B}$ and $\mathscr{B}_0$, respectively, into themselves, that $\| T_{0\omega} v \| \leq \| v \|$ for each v and ω, that the hypotheses of Theorem 1 are met, that U is an open ball centered at θ, and that $u_0 = \theta$. Consider the case in which g is causal on U, and g maps the zero element of $\mathscr{B}$ into the zero element of $\mathscr{B}_0$. Assume that g is *time invariant* on U in the sense that $T_{0\omega} g(u) = g(T_\omega u)$ for $u \in U$ and $\omega \in T$. Let G_m be as defined in the preceding paragraph. By the Property 2 part of Proposition 3, and the observations concerning $d^m g(\beta u)/d\beta^m$ in the preceding paragraph, we see that each G_m ($m = 1, 2, \cdots, p$) is time invariant on U.

The material just described can be modified to address the case in which g is periodically varying with a given period τ. Specifically, suppose that T is $[0, \infty)$, $(-\infty, \infty)$, $\{0, 1, 2, \cdots\}$, or $\{0, \pm 1, \pm 2, \cdots\}$, and that τ is a positive element of T. Let T_ω and $T_{0\omega}$ be as defined in the preceding paragraph, but with Ω taken to be the single-element set $\{\tau\}$ rather than T. Then, in the setting described in the preceding

---

* The assumptions are not met for $\mathscr{B}$ the set of bounded *continuous* functions from $[0, \infty)$ to R^1.

† Our definition is consistent with the one introduced in Ref. 36, p. 888, concerning causality for operators between abstract spaces. Also, a related result is given in Ref. 37, p. 40, for polynomial operators.

paragraph, $T_{0\omega}g(u) = g(T_\omega u)$ for $u \in U$ and $\omega \in \Omega$ means that g is *periodically varying with period* τ on U, and we see that if g has this property, it is inherited by the G_m.

In the case of Theorem 2, each $d^m g(\cdot)$ exists on U (because H.2 holds), and $g_m(\theta, u) = (m!)^{-1} d^m g(\beta u)/d\beta^m$ at $\beta = 0$ for each $m = 1, 2, \cdots$ and $\|u\| < \rho$. Therefore, results essentially the same as those developed in the preceding four paragraphs hold also with regard to the terms in (9).*

3.2 An application of theorem 1

Our first example, as well as the example in Section 3.3, concerns a nonlinear integral equation that plays an important role in the theory of feedback systems. To introduce the equation, we need the following definitions.

Let α_0 and α_1 be positive numbers with $\alpha_0 \leq \alpha_1$, and let $\psi_1, \psi_2, \cdots, \psi_n$ be a collection of $(p+1)$-times continuously differentiable functions from R^1 onto R^1 such that $\psi_i(0) = 0$ and $\alpha_0 \leq d\psi_i(\lambda)/d\lambda \leq \alpha_1$ for all i and all λ. Below, for convenience, we shall use $\psi_i^{(m)}$ to denote the mth derivative of ψ_i. Let ψ denote the map from R^n into R^n defined by $[\psi(s)]_i = \psi_i(s_i)$ for $i = 1, 2, \cdots, n$ and all $s \in R^n$, in which $[\psi(s)]_i$ and s_i are the ith components of $\psi(s)$ and s, respectively.

Let k denote an $n \times n$ matrix-valued function defined on $[0, \infty)$ such that each k_{ij} is measurable, bounded, and satisfies

$$\int_0^\infty |k_{ij}(\tau)|\, d\tau < \infty.$$

In this and the following section, each k_{ij} is assumed to be real valued.

Consider the equation

$$x(t) + \int_0^t k(t-\tau)\psi[x(\tau)]d\tau = u(t), \qquad t \geq 0, \tag{14}$$

as well as the related equation

$$y(t) + \int_0^t k(t-\tau)D(\tau)y(\tau)d\tau = v(t), \qquad t \geq 0, \tag{15}$$

in which u and v are elements of $L_\infty(R)$, and D is a real $n \times n$ diagonal-matrix-valued function defined on $[0, \infty)$ such that each D_{ii} is measurable on $[0, \infty)$ and satisfies $\alpha_0 \leq D_{ii}(\tau) \leq \alpha_1$ for each τ. Since ψ satisfies

* Specifically, the development remains valid if U is taken to be $\{u \in \mathscr{B}: \|u\| < \rho\}$, and if "Theorem 1" and "$m = 1, 2, \cdots, p$", respectively, are replaced with "Theorem 2" and "1, 2, $\cdots$."

a global Lipschitz condition on R^n, and k is as indicated, a standard successive-approximations approach (see, in particular, Ref. 38, Section 1.13) can be used to show that for each u and v in $L_\infty(R)$, (14) and (15) have solutions x and y, respectively, in the space E of functions w from $[0, \infty)$ into R^n such that

$$\int_0^\lambda |w_i(\tau)|^2 d\tau < \infty$$

for each $\lambda \in (0, \infty)$ and each i, and that x and y are unique in E.* To fix ideas, assume (this is often very easy to justify) that the only solutions of (14) of interest to us are those that are contained in E.

Let A.1 denote the hypothesis that for each u and v in $L_\infty(R)$, $L_\infty(R)$ contains any solution x of (14) in E, as well as any solution y of (15) in E. For explicit conditions on k, α_0, and α_1 under which A.1 holds, see Ref. 36, Theorem 3.

Assume initially that A.1 is met, and notice that then for each pair of elements $u, v \in L_\infty(R)$, the space $L_\infty(R)$ contains exactly one element x such that (14) is satisfied and exactly one element y such that (15) is met. In particular, we see that we can take f in Section 2.1 to be the map of $L_\infty(R)$ into itself defined by

$$f(w)(t) = w(t) + \int_0^t k(t - \tau)\psi[w(\tau)]d\tau, \qquad t \geq 0$$

for each $w \in L_\infty(R)$, and can take $\mathscr{B}$, $\mathscr{B}_0$, X, and U to be $L_\infty(R)$. In this example, g is defined on all of $\mathscr{B}$.

To discuss the example in more detail, it is convenient to let K and Ψ denote the maps of $L_\infty(R)$ into $L_\infty(R)$ defined by

$$(Kw)(t) = \int_0^t k(t - \tau)w(\tau)d\tau, \qquad t \geq 0$$

$$(\Psi w)(t) = \psi[w(t)], \qquad t \geq 0$$

for each $w \in L_\infty(R)$. Thus, here $f = I + K\Psi$ in which I is the identity map on $L_\infty(R)$. Consider Ψ.

Proposition 4: *$d\Psi$ exists on $L_\infty(R)$, and for w and h in $L_\infty(R)$, we have $[d\Psi(w)h](t) = D_0(t)h(t)$ for $t \geq 0$ in which $D_0(t)$ is the diagonal matrix* $\text{diag}[\psi_1^{(1)}[w_1(t)], \psi_2^{(1)}[w_2(t)], \cdots, \psi_n^{(1)}[w_n(t)]]$.

---

* The integral on the left side of (14) can easily be shown to be an element of R^n for each t whenever $x \in E$. Since the value of the integral is unchanged if x is replaced by any element of E that agrees with x almost everywhere, (14) has a solution if there is an element of E that satisfies the equation almost everywhere, and, moreover, any solution $x \in E$ is unique and not merely *essentially* unique. Similar remarks apply in the case of (15).

Proof: For w and h in $L_\infty(\mathrm{R})$,

$$\Psi(w+h)(t) - \Psi(w)(t)$$

$$= D_0(t)h(t) + \int_0^1 [D_{0\beta}(t) - D_0(t)]d\beta \cdot h(t), \qquad t \geq 0, \quad (16)$$

in which $D_{0\beta}(t)$ is the diagonal matrix of order n whose ith diagonal element is $\psi_i^{(1)}[\beta[w_i(t) + h_i(t)] + (1-\beta)w_i(t)]$. Since each $\psi_i^{(1)}$ is uniformly continuous on compact subsets of R^1, we see that for any fixed w, the difference $(D_{0\beta i} - D_{0i})$ of ith diagonal elements satisfies

$$\sup_{t\geq 0} |D_{0\beta i}(t) - D_{0i}(t)| \to 0 \quad \text{as} \quad \|h\| \to 0.$$

Thus, with $\delta \in L_\infty(\mathrm{R})$ defined by $\delta(t) = \int_0^1 [D_{0\beta}(t) - D_0(t)]d\beta \cdot h(t)$ for $t \geq 0$, we see that $\|\delta\| = o(\|h\|)$, which proves the proposition.

Proposition 4, together with the discussion above concerning (14) and (15), show that df exists, that (using the linearity of K) $df = I + Kd\Psi(w)$ for each w, and that for any w the map $df(w)$ is an invertible map of $L_\infty(\mathrm{R})$ onto itself.

To make further progress, we need the following result which is more general than Proposition 4.

Proposition 5: For each $m = 1, 2, \cdots, (p+1)$, $d^m\Psi$ exists and is continuous on $L_\infty(\mathrm{R})$, and, with $h_1, h_2, \cdots, h_m$ any m elements of $L_\infty(\mathrm{R})$, we have

$$[d^m\Psi(w)h_1 \cdots h_m(t)]_i = \psi_i^{(m)}[w_i(t)] \prod_{j=1}^{m} h_{ji}(t), \qquad t \geq 0$$

for each $i = 1, 2, \cdots, n$ and any $w \in L_\infty(\mathrm{R})$.

Proof: Let w be given. By Proposition 4 and, (with regard to continuity) the observation that $\|d\Psi(w+v)h - d\Psi(w)h\| \leq \max_i \sup_{t\geq 0} |\psi_i^{(1)}[w_i(t) + v_i(t)] - \psi_i^{(1)}[w_i(t)]|$ for $\|h\| = 1$, the assertion for $m = 1$ is true. Suppose that the assertion is true for m such that $1 \leq m \leq l$ with $l < (p+1)$.

Let $\tilde{Q}_l(w)$ denote the continuous multilinear mapping of $L_\infty(\mathrm{R})^{(l+1)}$ into $L_\infty(\mathrm{R})$ defined by

$$[\tilde{Q}_l(w)(p_1, p_2, \cdots, p_{l+1})(t)]_i = \psi_i^{(l+1)}[w_i(t)] \prod_{j=1}^{l+1} p_{ji}(t), \qquad t \geq 0$$

for each i and for $p_j \in L_\infty(\mathrm{R})$ for all j. We shall use $Q_l(w)$ to denote the usual associate (Ref. 31, p. 318) of $\tilde{Q}_l$ that belongs to $L(L_\infty(\mathrm{R}), L(L_\infty(\mathrm{R}), \cdots, L(L_\infty(\mathrm{R}), L_\infty(\mathrm{R}))\cdots))$ with $(l+1)$ L's, in which $L(A_1, A_2)$ stands for the set of continuous linear operators from the Banach space A_1 into the Banach space A_2.*

---

* For example, if $l = 2$, $L(L_\infty(\mathrm{R}), L(L_\infty(\mathrm{R}), \cdots, L(L_\infty(\mathrm{R}), L_\infty(\mathrm{R}))\cdots)) = L(L_\infty(\mathrm{R}), L(L_\infty(\mathrm{R}), L(L_\infty(\mathrm{R}), L_\infty(\mathrm{R}))))$.

By our induction hypothesis, and with h, as well as h_j for $j = 1, 2, \cdots, l$, elements of $L_\infty(\mathrm{R})$,

$$\| d^l\Psi(w + h)h_1 \cdots h_l - d^l\Psi(w)h_1 \cdots h_l - Q_l(w)hh_1 \cdots h_l \|$$

$$= \max_i \sup_{t\geq 0} \left| \psi_i^{(l)}[w_i(t) + h_i(t)] \prod_{j=1}^{l} h_{ji}(t) - \psi_i^{(l)}[w_i(t)] \prod_{j=1}^{l} h_{ji}(t) \right.$$

$$\left. - \psi_i^{(l+1)}[w_i(t)]h_i(t) \prod_{j=1}^{l} h_{ji}(t) \right| \leq c(h) \prod_{j=1}^{l} \|h_j\|,$$

where

$$c(h) = \max_i \sup_{t\geq 0} |\psi_i^{(l)}[w_i(t) + h_i(t)] - \psi_i^{(l)}[w_i(t)] - \psi_i^{(l+1)}[w_i(t)]h_i(t)|.$$

Thus, $\| d^l\Psi(w + h) - d^l\Psi(w) - Q_l(w)h\| \leq c(h)$. By the uniform continuity on compact sets of the $\psi_i^{(l+1)}$ (see the proof of Proposition 4), we have $c(h) = o(\|h\|)$ as $\|h\| \to 0$, which shows that $d^{(l+1)}\Psi(w)$ exists and that $d^{(l+1)}\Psi(w) = Q_l(w)$. Since $\| Q_l(w + h)p_1 \cdots p_{l+1} - Q_l(w)p_1 \cdots p_{l+1}\| \leq$

$$\max_i \sup_{t\geq 0} |\psi_i^{(l+1)}[w_i(t) + h_i(t)] - \psi_i^{(l+1)}[w_i(t)]| \cdot \prod_{j=1}^{l+1} \|p_j\|,$$

and the $\psi_i^{(l+1)}$ are continuous, we see that

$$\| Q_l(w + h) - Q_l(w) \| \to 0 \quad \text{as} \quad \|h\| \to 0,$$

showing that $d^{(l+1)}\Psi(w)$ depends continuously on w. This completes the proof.

Returning now to our example, by Proposition 5 and the linearity and boundedness of K, we see that $d^m f(w)$ exists and is continuous for $w \in L_\infty(\mathrm{R})$ and $m = 1, 2, \cdots, (p + 1)$. (Of course, $d^m f = Kd^m\Psi$ for $1 < m \leq (p + 1)$.) Therefore, the hypotheses of Theorem 1 are satisfied. Now choose $u_0 = \theta$ in Theorem 1 and notice that $g(\theta) = \theta$.

Referring to the standard successive approximations technique (Ref. 38, Section 1.13) that can be used to construct a unique solution x in E of (14) for each $u \in L_\infty(\mathrm{R})$, by the rule by which the iterates are generated it follows that g is causal and time invariant on $L_\infty(\mathrm{R})$ in the sense of Section 3.1. Therefore, the material in Section 3.1 shows that $g_m(\theta, \cdot)$ of Theorem 1 is both causal and time invariant on $L_\infty(\mathrm{R})$ for each $m = 1, 2, \cdots, p$.*

The terms $g_1(\theta, u), g_2(\theta, u), \cdots, g_p(\theta, u)$ in the expansion in Theorem 1 can be determined using 1.(a) and 1.(b).† For example, with H

* The same conclusion can be reached by considering the specific form of the $g_m(\theta, u)$, and using the fact that the operator H introduced below can be shown to be causal and time invariant. (In this connection, see Lemma 2.)

† The recursive process is straightforward in principle, but the complexity mounts rapidly with increasing order.

denoting $[I + Kd\Psi(\theta)]^{-1}$, we have $g_1(\theta, u) = Hu$, and, using $g_1(\theta, u) = Hu$, we find that when $p \geq 2$,

$$g_2(\theta, u) = -\tfrac{1}{2}HKd^2\Psi(\theta)(Hu)^2 \tag{17}$$

since $k_1 + k_2 = 2$ with k_1 and k_2 positive integers requires that $k_1 = k_2 = 1$, and

$$g_3(\theta, u) = -\tfrac{1}{6}HKd^3\Psi(\theta)(Hu)^3 + \tfrac{1}{2}HKd^2\Psi(\theta)(Hu)[HKd^2\Psi(\theta)(Hu)^2]. \tag{18}$$

[The derivation of (18) uses $g_1(\theta, u) = Hu$, (17), and the observation that $k_1 + k_2 = 3$ is met if either $k_1 = 1$ and $k_2 = 2$ or $k_1 = 2$ and $k_2 = 1$.]

Proposition 5 provides an important interpretation of the terms in the expressions for the $g_m(\theta, u)$. For example, by Proposition 5, we see that $d^3\Psi(\theta)(Hu)^3$, which appears in the first term on the right side of (18), is the element s of $L_\infty(\mathrm{R})$ given by $s_i(t) = \psi_i^{(3)}(0)[(Hu)(t)_i]^3$ for $t \geq 0$ and each i. Similarly, the ith component of $d^2\Psi(\theta)(Hu) \cdot [HKd^2\psi(\theta)(Hu)^2]$ in (18) has values $\psi_i^{(2)}(0)[(Hu)(t)_i]q_i(t)$, where $q = HKd^2\Psi(\theta)(Hu)^2$. Of course, q also has an immediate interpretation.

3.3 An application of corollary 1

Corollary 1 is in many respects a local version of Theorem 1. Here we give an example of an application of the corollary. As in Section 3.2, let $\mathscr{B} = \mathscr{B}_0 = L_\infty(\mathrm{R})$, and let K, Ψ, I, and u_0 be as defined there, but here it is not required that A.1 be met.

Let $F: L_\infty(\mathrm{R}) \to L_\infty(\mathrm{R})$ be given by $F = I + K\Psi$. As in Section 3.2, d^mF exists and is continuous on $L_\infty(\mathrm{R})$ for $m = 1, 2, \cdots, (p + 1)$. Of course, $dF(\theta) = I + Kd\Psi(\theta)$.

Let z be a complex scalar variable, and let $\tilde{K}$, the Laplace transform of k, be given by

$$\tilde{K}(z) = \int_0^\infty k(t)e^{-zt}dt, \qquad \mathrm{Re}(z) \geq 0.$$

Assume that

$$\det[1_n + \tilde{K}(z)\mathrm{diag}\{\psi_1^{(1)}(0), \cdots, \psi_n^{(1)}(0)\}] \neq 0$$

for $\mathrm{Re}(z) \geq 0$, in which 1_n is the identity matrix of order n. As a consequence, it can be shown (see Lemma 2 in Section 3.5) that $dF(\theta): L_\infty(\mathrm{R}) \to L_\infty(\mathrm{R})$ is invertible. Thus, by Lemma 1, there are open subsets S_1 and S_2 of $L_\infty(\mathrm{R})$, each containing θ, such that F restricted to S_1 is a homeomorphism of S_1 onto S_2.

Therefore, the hypotheses of Corollary 1 are met if X is chosen to be S_1, f is taken to be the restriction of F to X, and U is any open convex

set contained in S_2 and containing θ. This establishes the existence of, and shows how to obtain, a series approximation with error $o(\|u\|^p)$ as $\|u\| \to 0$ of the locally unique solution x of $Fx = u$, for u of sufficiently small norm.

3.4 Physical systems, and an application of theorem 4

In the following, $H(\mathrm{C})$ denotes the linear space of complex column n-vector-valued functions h defined on $[0, \infty)$ such that, with T_ω the "time truncation" operator of Section 3.1, we have $T_\omega h \in L_\infty(\mathrm{C})$ for $\omega \in [0, \infty)$ (i.e., such that any truncation of h is bounded and measurable). Clearly, unlike $L_\infty(\mathrm{C})$, $H(\mathrm{C})$ can contain unbounded functions.

Consider a physical system with an input v drawn from $L_\infty(\mathrm{C})$ and an output w contained in $H(\mathrm{C})$. Let the system be composed of linear elements, as well as nonlinear elements. Suppose that the nonlinear elements can be viewed as collectively introducing a constraint that can be written as $y = Nx$, in which N is a map from one subset of $H(\mathrm{C})$ into another, and where x and y, respectively, are the $H(\mathrm{C})$ input and output of the nonlinear part of the system.

With regard to the remainder of the system, which is linear, assume that there are linear maps A, B, C, and D of $H(\mathrm{C})$ into itself such that

$$x = Av + Cy \tag{19}$$

$$w = Dv + By. \tag{20}$$

Concerning the degree of generality of the model, and the assumption that the values of v, w, x, and y have the *same dimension* n, notice that we have not ruled out the possibility that some components of v, x, and/or y have no effect on the system, and, similarly, that certain of the components of w can be ignored. Nonzero initial conditions, if any, are assumed to be able to be taken into account either in N or as inputs to the system. A signal-flow-graph representation of the relations under consideration is given in Fig. 1.*

In this section, we use Theorem 4 to obtain a result concerning the response w of the system to inputs v having sufficiently small norm. To state the result, we introduce the following hypotheses and definition.

B.1: The restrictions of A, B, C, and D to $L_\infty(\mathrm{C})$ are bounded linear maps of $L_\infty(\mathrm{C})$ into itself.

B.2: There are open neighborhoods S_1 and S_2 of θ in $L_\infty(\mathrm{C})$ such that N restricted to S_1 is an invertible map of S_1 onto S_2. The map N also satisfies $N(\theta) = \theta$.

---

* This type of representation of a system has been used in different but related settings in Refs. 39, 40, and 41. The maps A, B, C, and D exist for a very large class of systems, but it is not difficult to give examples in which one or more map does not exist (see Ref. 40, pp. 244–5, for a very simple linear example along these lines).

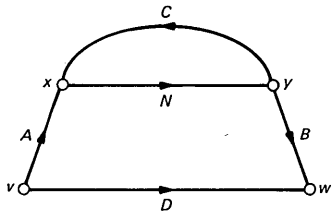


Fig. 1—Signal flow graph.

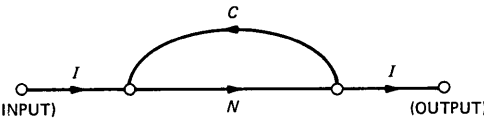


Fig. 2—Feedback part of the flow graph of Fig. 1.

Definition: When B.2 holds, the inverse of the restriction of N to S_1 is denoted by Φ. (Of course, Φ maps S_2 onto S_1, and we have $\Phi(\theta) = \theta$.)

B.3: There is an open neighborhood S_0 of θ in $L_\infty(C)$ such that $d^m\Phi$ exists on S_0 for $m = 1, 2, \cdots$.

B.4: $[d\Phi(\theta) - C_\infty]$ is an invertible map of $L_\infty(C)$ onto $L_\infty(C)$, where C_∞ is the restriction of C to $L_\infty(C)$.

Theorem 6: When B.1, B.2, B.3, and B.4 are met, there is a positive number δ and a neighborhood S of θ in $L_\infty(C)$ with the following properties:

(i) For each $v \in L_\infty(C)$ with $\|v\| < \delta$, there exist unique y, x, and w of S, S_1, and $L_\infty(C)$, respectively, such that (19), (20), *and $y = Nx$ hold.*

(ii) The function w described in (i) is given by

$$w = Dv + \sum_{m=1}^{\infty} Bg_m(Av) \tag{21}$$

for $\|v\| < \delta$, in which $g_1(Av) = [d\Phi(\theta) - C_\infty]^{-1}Av$, and

$$g_m(Av) = -[d\Phi(\theta) - C_\infty]^{-1} \cdot \sum_{l=2}^{m} (l!)^{-1} \sum_{\substack{k_1+k_2+\cdots+k_l=m \\ k_j>0}} d^l\Phi(\theta)g_{k_1}(Av)g_{k_2}(Av) \cdots g_{k_l}(Av)$$

for $m \geq 2$, and the series on the right side of (21) *converges uniformly for $\|v\| < \delta$.*

Proof: Using B.4, there are open neighborhoods S and S_3 of θ in $L_\infty(C)$, with $S \subseteq S_2$, such that for each $p \in S_3$ there is in S a unique y with the property that $\Phi(y) - C_\infty y = p$ (see Lemma 1). By the boundedness of the restriction of A to $L_\infty(C)$, $\delta_1 > 0$ can be chosen so that $Av \in S_3$ when $v \in L_\infty(C)$ and $\|v\| < \delta_1$, and thus so that for each such v there is a unique $y \in S$ such that $\Phi(y) - C_\infty y = Av$. For each such v and its associate y, let $x = \Phi(y)$ and $w = Dv + By$. Observe that for $\|v\| < \delta_1$ the corresponding triple (y, x, w) has the properties indicated in (i).

Now assume, as we may without loss of generality, that S_3 in the

preceding paragraph is convex, and that S_0 of B.3 satisfies $S \subseteq S_0$. Choose $\mathscr{B}$, $\mathscr{B}_0$, X, U, and f, respectively, of Section 2.1 to be $L_\infty(\mathrm{C})$, $L_\infty(\mathrm{C})$, S, S_3, and the map defined by $f(s) = \Phi(s) - C_\infty s$ for $s \in S$. We have $df = d\Phi - C_\infty$ on X, $d^m f = d^m \Phi$ for $m = 2, 3, \cdots$ on X, and, by B.4, $df(\theta)$ is an invertible map. Thus, by Theorem 4, there is a $\sigma > 0$ such that S_3 contains an open ball centered at θ of radius σ, and the solution $s \in S$ of $\Phi(s) - C_\infty s = p$ for $p \in L_\infty(\mathrm{C})$ with $\|p\| < \sigma$ is given by the uniformly convergent series

$$\sum_{m=1}^{\infty} g_m(p),$$

in which $g_1(p) = [d\Phi(\theta) - C_\infty]^{-1} p$, and where

$$g_m(p) = -[d\Phi(\theta) - C_\infty]^{-1} \sum_{l=2}^{m} (l!)^{-1} \sum_{\substack{k_1+k_2+\cdots+k_l=m \\ k_j>0}} d^l\Phi(\theta) g_{k_1}(p) \cdots g_{k_l}(p)$$

for $m \geq 2$. Therefore, (*ii*) of Theorem 6 holds if for some $\delta \in (0, \delta_1)$ we have $\|Av\| < \sigma$ whenever $\|v\| < \delta$, and this condition is met because by B.1 the restriction of A to $L_\infty(\mathrm{C})$ is bounded. [Notice that here, and with regard to the *uniform* convergence of the series, we also use the boundedness of the restriction of B to $L_\infty(\mathrm{C})$.]

3.5 Volterra-series representations

In this, our final section, we first consider the single-input case in which v in Theorem 6 satisfies $[v(t)]_i = 0$ for $t \geq 0$ and $i > 1$. We give a theorem which provides explicit conditions on A, B, C, D, and N under which the hypotheses of Theorem 6 are met and the series for w can be interpreted as a *Volterra series* in v_1 the function on $[0, \infty)$ whose values are $[v(t)]_1$. Towards the end of the section, an n-input extension of our result is given.

We will use the following definitions and hypotheses.

Definition: For any positive integers l and q, let $S_q^{(l)}$ denote the set of all functions h from the l-dimensional interval $[0, \infty)^l$ into the set of complex $n \times q$ matrices such that each h_{ij} is measurable and bounded on $[0, \infty)^l$, and satisfies

$$\int_{[0,\infty)^l} |h_{ij}(\tau_1, \tau_2, \cdots, \tau_l)| \, d(\tau_1, \tau_2, \cdots, \tau_l) < \infty. \tag{22}$$

Definition: If r and s are two complex column n-vectors, then rs denotes the column n-vector defined by $(rs)_i = r_i s_i$ for $i = 1, 2, \cdots, n$.

Definition: ${}^{(1)}L_\infty(\mathrm{C})$ denotes $L_\infty(\mathrm{C})$ with $n = 1$.

C.1: Referring to A, B, C, and D of (19) and (20), there are elements a, b, c, and d of $S_n^{(1)}$ such that for each $p \in L_\infty(\mathrm{C})$,

$$(Ap)(t) = \int_0^t a(t-\tau)p(\tau)d\tau$$

$$(Bp)(t) = \int_0^t b(t-\tau)p(\tau)d\tau$$

$$(Cp)(t) = \int_0^t c(t-\tau)p(\tau)d\tau$$

$$(Dp)(t) = \int_0^t d(t-\tau)p(\tau)d\tau$$

for $t \geq 0$.

C.2: For some $\gamma > 0$, N of Section 3.4 is defined on $\Gamma = \{s \in L_\infty(\mathrm{C}): \|s\| < \gamma\}$ by

$$[(Ns)(t)]_j = n_j\{[s(t)]_j\}, \qquad t \geq 0 \tag{23}$$

for $j = 1, 2, \cdots, n$, in which each n_j maps complex numbers z with $|z| < \gamma$ into complex numbers such that $n_j(0) = 0$ and such that $dn_j(z)/dz$ exists for $|z| < \gamma$ and is nonzero at $z = 0$. (Clearly, when C.2 is met, N restricted to Γ can be represented by n single-input single-output memoryless nonlinear operators.)

C.3: $\det\{1_n - \mathrm{diag}[dn_1(0)/dz, \cdots, dn_n(0)/dz] \int_0^\infty c(\tau)e^{-z\tau}d\tau\} \neq 0$ for $\mathrm{Re}(z) \geq 0$.*

A result (see Lemma 3, below) concerning elements of $S_1^{(l)}$ that we shall use is the following:

Proposition 6: If $k_l \in S_1^{(l)}$ for some l, then the iterated integral

$$\int_0^t \cdots \int_0^t k_l(t-\tau_1, t-\tau_2, \cdots, t-\tau_l)\mu(\tau_1)\mu(\tau_2) \cdots \mu(\tau_l)d\tau_1 d\tau_2 \cdots d\tau_l$$

exists for $t \geq 0$ and $\mu \in {}^{(1)}L_\infty(\mathrm{C})$, and $V_{k_l}(\mu)$ defined on $[0, \infty)$ by

$$V_{k_l}(\mu)(t) = \int_0^t \cdots \int_0^t k_l(t-\tau_1, t-\tau_2, \cdots, t-\tau_l)\,\mu(\tau_1)\,\mu(\tau_2) \cdots \mu(\tau_l)d\tau_1 d\tau_2 \cdots d\tau_l$$

for an arbitrary $\mu \in {}^{(1)}L_\infty(\mathrm{C})$, is an element of $L_\infty(\mathrm{C})$.

Theorem 7: Suppose that C.1, C.2, and C.3 are met. Then

(i) The hypotheses of Theorem 6 are satisfied.

(ii) For each $l = 1, 2, \cdots$ there is a $k_l \in S_1^{(l)}$ such that, under the condition that $[v(t)]_i = 0$ for $t \geq 0$ and $i > 1$, we have

* As in Section 3.3, 1_n denotes the identity matrix of order n.

$$w = \sum_{l=1}^{\infty} V_{k_l}(v_1) \quad \textit{for} \quad \|v\| < \delta,$$

with the series uniformly convergent for $\|v\| < \delta$, *where* v, w, *and* δ *are described in Theorem* 6, *and* $V_{k_l}(\cdot)$ *is as indicated in Proposition* 6.

(iii) Each k_l *can be taken to be continuous* on* $[0, \infty)^l$ *when* a *and* d *are continuous on* $[0, \infty)$.

Proof of Theorem 7: Hypothesis B.1 is clearly met. Also, with N as described in C.2, an easy modification of Proposition 4 shows† that $dN{:}L_\infty(\mathrm{C}) \to L_\infty(\mathrm{C})$ exists on Γ, and that for $p \in \Gamma$ and $h \in L_\infty(\mathrm{C})$,

$$[dN(p)h](t) = D(t)h(t), \qquad t \geq 0, \tag{24}$$

where $D(t) = \mathrm{diag}\{dn_1[p_1(t)]/dz, \cdots, dn_n[p_n(t)]/dz\}$.

Similarly, an easy modification of Proposition 5 establishes that $d^m N(p)$ exists for $p \in \Gamma$ and all m. (Observe that, because z is complex, the existence of $dn_j(z)/dz$ for each j and $|z| < \gamma$ means that the derivatives of each n_j of *all orders* exist for $|z| < \gamma$.) Since each $dn_j(z)/dz$ is not zero at $z = 0$, and hence each is nonzero throughout a neighborhood of $z = 0$, it is clear that the $|dn_j(z)/dz|$ are bounded away from zero on some neighborhood of $z = 0$. It follows from (24) that $dN(p)$ is invertible for p in a neighborhood of θ in $L_\infty(\mathrm{C})$. Thus, there are open neighborhoods S_1 and S_2 of θ such that N restricted to S_1, which we denote by N_0, is an invertible map of S_1 onto S_2 and $d^m(N_0^{-1})$ exists throughout S_2 for each $m = 1, 2, \cdots$ (see Lemma 1). Therefore, B.2 and B.3 are satisfied.

Let $\Psi{:}S_2 \to S_1$ denote N_0^{-1}, and notice that Ψ satisfies

$$[(\Psi s)(t)]_j = n_j^{-1}\{[s(t)]_j\}, \qquad t \geq 0$$

for each j and all s in some neighborhood of θ, where n_j^{-1}, defined in a neighborhood of the origin of the complex plane, is a local inverse of n_j. Since each $dn_j(z)/dz$ exists throughout a neighborhood of the origin, and does not vanish at the origin, we see that for each j and m, $d^m n_j^{-1}(z)/dz^m$ exists in a neighborhood of the origin. Therefore, by a direct modification of Proposition 5,

$$[d^m\Psi(\theta)h_1 \cdots h_m(t)]_i = d^m n_j^{-1}(0)/dz^m \prod_{j=1}^{m} h_{ji}(t) \tag{25}$$

for $t \geq 0$, each m and i, and any $h_1, h_2, \cdots, h_m$ in $L_\infty(\mathrm{C})$.

We shall use (25) subsequently. At the moment, concerning Ψ,

---

* Of course, by k_l "continuous" we mean that each component of k_l is continuous.

† Notice that if $dn_j[z(p_j(t) + h_j(t)) + (1 - z)p_j(t)]/dz$ exists at a point $(\alpha, 0)$, then $dn_j[\beta(p_j(t) + h_j(t)) + (1 - \beta)p_j(t)]$, with β a real variable, exists at $\beta = \alpha$ and the values of the two derivatives are the same.

note merely that $[d\Psi(\theta)]^{-1} = dN(\theta)$. Since $[d\Psi(\theta) - C_\infty] = d\Psi(\theta)\{I - [d\Psi(\theta)]^{-1}C_\infty\}$, where I is the identity transformation in $L_\infty(\mathrm{C})$ and C_∞ is defined in B.4, by C.3 and Lemma 2 below, we see that the hypotheses of Theorem 6 are satisfied.

In Lemma 2 we refer to the following.

D.1: $\lambda \in S_n^{(1)}$, Λ denotes the map of $L_\infty(\mathrm{C})$ into itself defined by

$$(\Lambda p)(t) = \int_0^t \lambda(t-\tau)p(\tau)d\tau, \qquad t \geq 0$$

for $p \in L_\infty(\mathrm{C})$, and, with z a scalar complex variable, $\tilde{\Lambda}(z)$ denotes

$$\int_0^\infty \lambda(t)e^{-zt}dt, \qquad \mathrm{Re}(z) \geq 0.$$

Lemma 2: Let D.1 hold, and suppose that $\det\{1_n - \tilde{\Lambda}(z)\} \neq 0$ *for* $\mathrm{Re}(z) \geq 0$. *Then*

(i) $(I - \Lambda)$ *is an invertible map of* $L_\infty(\mathrm{C})$ *onto itself,*

(ii) there is a $\kappa \in S_n^{(1)}$ *such that*

$$(I-\Lambda)^{-1}p(t) = p(t) + \int_0^t \kappa(t-\tau)p(\tau)d\tau, \qquad t \geq 0$$

for $p \in L_\infty(\mathrm{C})$, *and*

(iii) if λ *is continuous for* $t \geq 0$, *then* κ *can be taken to be continuous on* $[0, \infty)$.*

Lemma 2 is proved in Appendix D.

We also need the following two lemmas which are proved in Appendices E and F.

Lemma 3: Suppose that $h \in S_1^{(l)}$ *for some* $l \geq 1$, *that* $s \in S_n^{(1)}$, *and that* u *is a bounded measurable function from* $[0, \infty)^l$ *into the complex numbers. Then*

(i) With $\tilde{h}$ *defined on* $(-\infty, \infty)^l$ *by* $\tilde{h} = h$ *on* $[0, \infty)^l$ *and* $\tilde{h} = 0_{n1}$ *(the zero* $n \times 1$ *matrix) otherwise, the function* k, *defined by*

$$k(\alpha_1, \alpha_2, \cdots, \alpha_l) = \int_0^\infty s(\tau)\tilde{h}(\alpha_1 - \tau, \cdots, \alpha_l - \tau)d\tau$$

for $(\alpha_1, \alpha_2, \cdots, \alpha_l) \in [0, \infty)^l$, *belongs to* $S_1^{(l)}$.

(ii) If h *is continuous on* $[0, \infty)^l$, *then so is* k.

(iii) The iterated integrals

$$\int_0^t \cdots \int_0^t h(t-\tau_1, \cdots, t-\tau_l)u(\tau_1, \cdots, \tau_l)d\tau_1 \cdots d\tau_l$$

* Part *(iii)* is not used in the proof of Theorem 7, and is included because it is useful for other purposes.

and

$$\int_0^t \cdots \int_0^t k(t-\tau_1, \cdots, t-\tau_l)u(\tau_1, \cdots, \tau_l)d\tau_1 \cdots d\tau_l$$

exist, and are invariant with respect to interchanges of orders of integration, for $t \geq 0$; *and* p *defined by*

$$p(t) = \int_0^t \cdots \int_0^t h(t-\tau_1, \cdots, t-\tau_l)u(\tau_1, \cdots, \tau_l)d\tau_1 \cdots d\tau_l, \; t \geq 0$$

is an element of $L_\infty(C)$.

(*iv*) *We have*

$$\int_0^t s(t-\tau) \int_0^\tau \cdots \int_0^\tau h(\tau-\tau_1, \cdots, \tau-\tau_l)u(\tau_1, \cdots, \tau_l)d\tau_1 \cdots d\tau_l d\tau$$

$$= \int_0^t \cdots \int_0^t k(t-\tau_1, \cdots, t-\tau_l)u(\tau_1, \cdots, \tau_l)d\tau_1 \cdots d\tau_l, \qquad t \geq 0.$$

Lemma 4: If $h \in S_1^{(p)}$ *and* $k \in S_1^{(q)}$, *then the function* s, *defined* on* $[0, \infty)^{p+q}$ *by*

$$s(\tau_1, \cdots, \tau_{p+q}) = h(\tau_1, \cdots, \tau_p)k(\tau_{p+1}, \cdots, \tau_{p+q})$$

for $(\tau_1, \cdots, \tau_{p+q}) \in [0, \infty)^{p+q}$, *belongs to* $S_1^{(p+q)}$.

We now return to the proof of Theorem 7.

With v, w, and δ as in Part (*ii*) of Theorem 6, we have

$$w = Dv + \sum_{m=1}^{\infty} Bg_m(Av), \qquad \|v\| < \delta$$

in which the series converges uniformly, $g_1(Av) = [d\Psi(\theta) - C_\infty]^{-1}Av$, and

$$g_m(Av) = -[d\Psi(\theta) - C_\infty]^{-1} \cdot \sum_{l=2}^{m} (l!)^{-1} \sum_{\substack{k_1+k_2+\cdots+k_l=m \\ k_j>0}} d^l\Psi(\theta)g_{k_1}(Av) \cdots g_{k_l}(Av), \qquad m \geq 2.$$

Assume now that $[v(t)]_i = 0$ for $t \geq 0$ and $i > 1$. By Lemma 3, it suffices to show that for each m there is a $q_m \in S_1^{(m)}$ such that

$$g_m(Av)(t) = \int_0^t \cdots \int_0^t q_m(t-\tau_1, \cdots, t-\tau_m)v_1(\tau_1) \cdots v_1(\tau_m)d\tau_1 \cdots d\tau_m \quad (26)$$

---

* See the second definition at the beginning of Section 3.5.

for $t \geq 0$ and $\|v\| < \delta$, and that q_m is continuous on $[0, \infty)^m$ when a is continuous on $[0, \infty)$. Thus, suppose for the purpose of induction that (26) holds with q_m as indicated for $m = 1, 2, \cdots, p$ for some $p \geq 1$. Observe that by Lemma 3 and Part (*ii*) of Lemma 2 the induction hypothesis is met for $p = 1$.

We have

$$g_{p+1}(Av) = -[d\Psi(\theta) - C_\infty]^{-1} \cdot \sum_{l=2}^{(p+1)} (l!)^{-1} \sum_{\substack{k_1+k_2+\cdots+k_l=p+1 \\ k_j>0}} d^l\Psi(\theta) g_{k_1}(Av) \cdots g_{k_l}(Av). \quad (27)$$

Now let $l \geq 2$ be fixed, and let $k_1, \cdots, k_l$ be any l positive integers such that $k_1 + k_2 + \cdots + k_l = p + 1$. By Lemma 4, (25), and Lemma 3, r defined by

$$r(\tau_1, \cdots, \tau_{p+1}) = q_{k_1}(\tau_1, \cdots, \tau_{k_1}) q_{k_2}(\tau_{k_1+1}, \cdots, \tau_{k_1+k_2}) \cdots q_{k_l}(\tau_{k_1+\cdots+k_{l-1}+1}, \cdots, \tau_{p+1})$$

for $(\tau_1, \cdots, \tau_{p+1}) \in [0, \infty)^{p+1}$ belongs to $S_1^{(p+1)}$, and

$$d^l\Psi(\theta) g_{k_1}(Av) \cdots g_{k_l}(Av)(t) = D \int_0^t \cdots \int_0^t r(t - \tau_1, \cdots, t - \tau_{p+1}) v_1(\tau_1) \cdots v_1(\tau_{p+1}) d\tau_1 \cdots d\tau_{p+1}$$

for $t \geq 0$, in which D is the diagonal matrix of order n whose jth diagonal element is $d^l n_j^{-1}(0)/dz^l$. Observe that r is continuous on $[0, \infty)^{p+1}$ when each q_{k_j} is continuous on $[0, \infty)^{k_j}$.

Therefore, using (27), and by Lemmas 2 and 3, there is a $q_{p+1} \in S_1^{(p+1)}$, which is continuous on $[0, \infty)^{p+1}$ when a is continuous on $[0, \infty)$, such that (26) holds with m replaced with $(p + 1)$. This completes the proof of the theorem.

Comments: By Lemma 2 and Proposition 7 (in Appendix G), an interpretation of C.3 is simply that the "feedback part" of the graph of Fig. 1, shown in Fig. 2, is bounded-input bounded-output stable* when C.1 and C.2 are met, N is replaced with its linearization $dN(\theta)$ extended in the natural way to all of $H(\mathrm{C})$, and $C{:}H(\mathrm{C}) \to H(\mathrm{C})$ is defined by $(Cp)(t) = \int_0^t c(t - \tau) p(\tau) d\tau$, $t \geq 0$ for each $p \in H(\mathrm{C})$.

The k_l in Theorem 7 can be taken to be real valued (i.e., to have

* By this we mean that for each $L_\infty(\mathrm{C})$ input to the graph of Fig. 2, there is in $H(\mathrm{C})$ a unique output, and the output belongs to $L_\infty(\mathrm{C})$. In this case, existence and uniqueness of an output in $H(\mathrm{C})$ follows from the hypotheses concerning c via the usual successive approximations approach (Ref. 38, Section 1.13). In Fig. 2, I denotes the identity transformation in $H(\mathrm{C})$.

zero imaginary part) if a, b, c, and d are real valued, and $d^m n_j^{-1}(0)/dz^m$ is real for all m and j. (It is not difficult to show that the $d^m n_j^{-1}(0)/dz^m$ are real when $d^m n_j(0)/dz^m$ is real for every m and j.) In particular, we see that Theorem 7 establishes the existence of a Volterra-series expansion for the important corresponding case in which v_1, w, x, and y in Fig. 1 are restricted to be real valued, C.1 is met, and N (which then would be a map between real-valued function spaces) can be analytically extended so that C.2 and C.3 are satisfied.*† In this connection, Theorem 5 can be used to prove results along the same lines as Theorems 6 and 7, but with $L_\infty(\mathrm{C})$ replaced throughout with $L_\infty(\mathrm{R})$. Similarly, Corollary 1 can be used to obtain corresponding pth order *approximation* results under weaker differentiability hypotheses concerning N.

Theorem 6 provides explicit expressions for the $g_m(Av)$. For example,

$$g_1(Av) = [d\Psi(\theta) - C_\infty]^{-1}Av$$

$$g_2(Av) = -\tfrac{1}{2}[d\Psi(\theta) - C_\infty]^{-1}d^2\Psi(\theta)\{[d\Psi(\theta) - C_\infty]^{-1}Av\}^2$$

$$g_3(Av) = -[d\Psi(\theta) - C_\infty]^{-1}d^2\Psi(\theta)g_1(Av)g_2(Av) - \tfrac{1}{6}[d\Psi(\theta) - C_\infty]^{-1}d^3\Psi(\theta)[g_1(Av)]^3,$$

and so on. Therefore, assuming merely that we can write down the representation of $[d\Psi(\theta) - C_\infty]$ along the lines of Part (*ii*) of Lemma 2, notice that it is not difficult in principle to give an explicit expression for any of the k_l of Theorem 7.

Theorem 7 (and Proposition 6) can be extended to cover the case in which the restriction that $[v(t)]_i = 0$ for $t \geq 0$ and $i > 1$ is not met.‡ Specifically, using the results and techniques described in connection with the proof of Theorem 7, it is not difficult to prove the following extension in which for each l, $\chi[v(\tau_1), \cdots, v(\tau_l)]$ denotes the column vector of order n^l whose elements are the n^l distinct products $v_{\omega_1}(\tau_1)v_{\omega_2}(\tau_2) \cdots v_{\omega_l}(\tau_l)$, corresponding to distinct sequences $\omega_1, \omega_2, \cdots, \omega_l$ with each ω_j drawn from $\{1, 2, \cdots, n\}$, arranged in an arbitrary predetermined order.

---

* Observe that this extendability condition is often met. (In particular, polynomial nonlinearities frequently arise in locally-valid models, and polynomials in z are entire functions.)

† Theorem 7 bears on problems concerning the transmission of digital signals over analog communication channels. Discussions with this writer's colleague, J. Salz, concerning such problems provided part of the motivation to formulate the system model in Section 3.4 and to develop a theorem along the lines of Theorem 7.

‡ The case of more than one input is of importance, for example, in connection with studies of the effects of initial conditions. Also, straightforward modifications suffice to establish corresponding results for the *time-discrete* case in which t takes values in $\{0, 1, 2, \cdots\}$.

Theorem 8: Under the hypotheses of Theorem 7, for every $l = 1, 2, \cdots$ *there is a* $k_l \in S_{n^l}^{(l)}$, *with* k_l *continuous on* $[0, \infty)^l$ *when* a *and* d *are continuous on* $[0, \infty)$, *such that*

(*i*) *The iterated integral*

$$\int_0^t \cdots \int_0^t k_l(t - \tau_1, \cdots, t - \tau_l)\chi[v(\tau_1), \cdots, v(\tau_l)]d\tau_1 \cdots d\tau_l$$

exists for $t \geq 0$ *and* $v \in L_\infty(\mathrm{C})$, *and* $V_{k_l}(v)$ *defined on* $[0, \infty)$ *by*

$$V_{k_l}(v)(t) = \int_0^t \cdots \int_0^t k_l(t - \tau_1, \cdots, t - \tau_l)\chi[v(\tau_1), \cdots, v(\tau_l)]d\tau_1 \cdots d\tau_l$$

for any $v \in L_\infty(\mathrm{C})$ *is an element of* $L_\infty(\mathrm{C})$.

(*ii*) *With* $V_{k_l}(\cdot)$ *as indicated in* (*i*) *above, and with* v, w, *and* δ *as described in Theorem* 6, *the expansion*

$$w = \sum_{l=1}^{\infty} V_{k_l}(v)$$

converges uniformly for $\|v\| < \delta$.

It is not difficult to verify that Theorems 7 and 8 remain valid if C.1 is modified to allow a constant multiple (or, more generally, a constant $n \times n$-matrix multiple) of the identity operator on $H(\mathrm{C})$ to be added to B.

We do not give here the details of other extensions of Theorem 7,* but it is worthwhile to appreciate that in some extensions in which C.1 is weakened, series representations can arise in which, unless generalized-function kernels are admitted, the terms do not have the form normally associated with a Volterra series. Consider, for example, that if $n = 1$, if A, as well as B, is the identity transformation on $H(\mathrm{C})$, if C and D have the representations on $L_\infty(\mathrm{C})$ given in C.1, and if C.2 and C.3 are satisfied with $d^2n_1^{-1}(0)/dz^2 = 2$, then the hypotheses of Theorem 6 are met and the second term in the sum in (21) is a function whose values are

* We leave for another paper results concerning cases in which N can be of a more general form, and the restrictions to $L_\infty(\mathrm{C})$ of A, B, C, and D are not necessarily time invariant, and generalized functions may be involved in their representations. Also left for a later paper are applications to differential equations. Assuming merely that A, B, C, and D are defined on *all* of $H(\mathrm{C})$ as convolution operators with kernels a, b, c, and d, it is a simple corollary of Theorem 8 that there exists a Volterra-series expansion also for the case in which the conditions of Theorem 8 are met, with the exception that the $dn_j(z)/dz$ are not necessarily *nonzero* at $z = 0$. More specifically, N can be replaced with N plus a multiple βI of the identity operator I on $H(\mathrm{C})$, with $|\beta|$ sufficiently small that the four linear parts of the system can be modified accordingly and remain $S_n^{(1)}$ convolutions.

$$- \rho^3 v(t)^2 - 2\rho^2 v(t) \int_0^t k(t-\tau)v(\tau)d\tau + \rho^2 \int_0^t k(t-\tau)v(\tau)^2 d\tau$$

$$+ 2\rho \int_0^t k(t-\tau)v(\tau) \int_0^\tau k(\tau-\tau_1)v(\tau_1)d\tau_1 d\tau$$

$$+ \int_0^t \int_0^t k_a(t-\tau_1, t-\tau_2)v(\tau_1)v(\tau_2)d\tau_1 d\tau_2,$$

in which ρ is a nonzero constant, $k \in S_1^{(1)}$ and $k_a \in S_1^{(2)}$.

APPENDIX A

Proof of Proposition 1

Let σ be real, nonzero, and such that $(r_0 + \sigma) \in A$. Then, using the linearity of $L(r_0 + \sigma)$,

$$\sigma^{-1}[L(r_0+\sigma)e(r_0+\sigma) - L(r_0)e(r_0)]$$

$$= L(r_0+\sigma)\sigma^{-1}[e(r_0+\sigma) - e(r_0)] + \sigma^{-1}[L(r_0+\sigma) - L(r_0)]e(r_0)$$

$$= L(r_0)de(r_0)/dr + [dL(r_0)/dr]e(r_0) + \Delta_1(\sigma) + \Delta_2(\sigma) + \Delta_3(\sigma) + \Delta_4(\sigma),$$

where

$$\Delta_1(\sigma) = L(r_0)\{\sigma^{-1}[e(r_0+\sigma) - e(r_0)] - de(r_0)/dr]\}$$

$$\Delta_2(\sigma) = [L(r_0+\sigma) - L(r_0)]\{\sigma^{-1}[e(r_0+\sigma) - e(r_0)] - de(r_0)/dr\}$$

$$\Delta_3(\sigma) = [L(r_0+\sigma) - L(r_0)]de(r_0)/dr$$

$$\Delta_4(\sigma) = \{\sigma^{-1}[L(r_0+\sigma) - L(r_0)] - dL(r_0)/dr\}e(r_0).$$

Since $L(r_0)$ is a *bounded* operator and $\|\sigma^{-1}[e(r_0+\sigma) - e(r_0)] - de(r_0)/dr\| \to 0$ as $\sigma \to 0$, $\Delta_1(\sigma) \to 0$ in S as $\sigma \to 0$. It is clear that $\Delta_2(\sigma)$, $\Delta_3(\sigma)$, and $\Delta_4(\sigma)$ approach zero in S as $\sigma \to 0$. This completes the proof.

APPENDIX B

Part of the Proof of Theorem 1

Let k be any integer such that $1 \le k \le p$, and let $v_1(\cdot)$, $v_2(\cdot)$, $\cdots$, $v_k(\cdot)$ denote k maps from an open subset of $(-\infty, \infty)$ containing $[0, 1]$ into $\mathscr{B}_0$ such that each $v_j(\cdot)$ is differentiable on $[0, 1]$. With l an integer such that $0 \le l \le k - 1$, consider $d^k f[q(\beta)]v_1(\beta) \cdots v_{k-l-1}(\beta)v_{k-l}(\beta)$. Since $d^k f[q(\beta)]$ for $0 \le \beta \le 1$ is a Fréchet derivative, $d^k f[q(\beta)]v_1(\beta) \cdots v_{k-l-1}(\beta)$ is a bounded linear map from $\mathscr{B}_0$ into a Banach space S for each $\beta \in [0, 1]$. By the version of the chain rule in Ref. 31, p. 173, $d\{d^k f[q(\beta)]\}/d\beta$ exists for $\beta \in [0, 1]$, and

$$d\{d^k f[q(\beta)]\}/d\beta = d^{k+1} f[q(\beta)] q^{(1)}(\beta), \qquad \beta \in [0, 1]. \tag{28}$$

By Proposition 1, (28), and an obvious inductive argument, $d\{d^k f[q(\beta)] v_1(\beta) \cdots v_{k-l-1}(\beta)\}/d\beta$ exists for $\beta \in [0, 1]$. Thus, by Proposition 1, $d\{d^k f[q(\beta)] v_1(\beta) \cdots v_{k-l}(\beta)\}/d\beta$ exists and satisfies

$$\begin{aligned} d\{d^k f[q(\beta)] v_1(\beta) \cdots v_{k-l}(\beta)\}/d\beta \\ = d^k f[q(\beta)] v_1(\beta) \cdots v_{k-l-1}(\beta)[dv_{k-l}(\beta)/d\beta] \\ + d\{d^k f[q(\beta)] v_1(\beta) \cdots v_{k-l-1}(\beta)\}/d\beta \, v_{k-l}(\beta) \end{aligned} \tag{29}$$

for $\beta \in [0, 1]$.

By relations (28) and (29), and the fact that $df[q(\beta)]/d\beta = df[q(\beta)] q^{(1)}(\beta)$ for $0 \le \beta \le 1$, we see that an expression $E_m(\beta)$ can be given for $d^m f[q(\beta)]/d\beta^m$ for $\beta \in [0, 1]$ which depends only on $d^l f[q(\beta)]$ and $q^{(l)}(\beta)$ for $l = 1, 2, \cdots, m$. For example, $E_2(\beta) = df[q(\beta)] q^{(2)}(\beta) + d^2 f[q(\beta)] q^{(1)}(\beta) q^{(1)}(\beta)$. Since $df[q(\beta)]/d\beta = (u - u_0)$ for $\beta \in [0, 1]$, we see that $E_m(\beta) = \theta$ for $\beta \in [0, 1]$.

Now consider (5) and (6). Since $d^l f_0[q_0(r)]$ and $d^l q_0(r)/dr^l$ exist at $r = \beta$, with $d^l f_0[q_0(\beta)] = d^l f[q(\beta)]$ and $d^l q_0(r)/dr^l = q^{(l)}(\beta)$ for $r = \beta$ and $l = 1, 2, \cdots, m$, by Proposition 1 and observations similar to those of the preceding three paragraphs, we see that, as claimed in Section 2.2, $d^m f_0[q(r)]/dr^m|_{r=\beta}$ exists and that it equals $E_m(\beta)$.

APPENDIX C

Proof of Proposition 2

Under the conditions indicated, $dg(u_0)$ exists (and equals L). Thus, using $f[g(u)] = u$ for $u \in U$, and with I the identity operator on $\mathscr{B}$, we have $df[g(u_0)] dg(u_0) = I$. This shows that $df[g(u_0)]$ has a *right* inverse.

On the other hand, for $u \in U$ with $\|u - u_0\| < \sigma$, $g(u) - g(u_0) = Lf[g(u)] - Lf[g(u_0)] + R\{f[g(u)] - f[g(u_0)]\}$ and, thus,

$$\begin{aligned} \{I_0 - Ldf[g(u_0)]\}[g(u) - g(u_0)] = R\{df[g(u_0)][g(u) - g(u_0)] \\ + R_1[g(u) - g(u_0)]\} + R_2[g(u) - g(u_0)] \end{aligned} \tag{30}$$

in which I_0 is the identity operator on $\mathscr{B}_0$, $R_1(h) = o(\|h\|)$ as $\|h\| \to 0$, and, using the boundedness of L, $R_2(h) = o(\|h\|)$ as $\|h\| \to 0$.

Now let $v \in \mathscr{B}$ be arbitrary. By the continuity of f at $g(u_0)$, and the openness of X and U, choose $\beta > 0$ so that $g(u_0) + \alpha v \in X$, $\|f[g(u_0) + \alpha v] - f[g(u_0)]\| < \sigma$, and $f[g(u_0) + \alpha v] \in U$ for $|\alpha| < \beta$. Notice that for each α with $|\alpha| < \beta$, and with $u = f[g(u_0) + \alpha v]$,

we have $\|u - u_0\| < \sigma$, as well as $f[g(u)] = f[g(u_0) + \alpha v]$, and hence* $g(u) = g(u_0) + \alpha v$.

Therefore, by (30),

$$\{I_0 - Ldf[g(u_0)]\}v = \alpha^{-1}R\{df[g(u_0)]\alpha v + R_1(\alpha v)\} + \alpha^{-1}R_1(\alpha v) \quad (31)$$

for $0 < |\alpha| < \beta$. Since the right side of (31) approaches θ as $\alpha \to 0$, and since v is arbitrary, it follows that $df[g(u_0)]$ has a *left* inverse. Since $df[g(u_0)]$ has both a left inverse and a right inverse, $df[g(u_0)]^{-1}$ exists. Finally, a standard type of argument shows that $(df)^{-1}$ exists and is continuous throughout some neighborhood of $g(u_0)$ when f is continuously differentiable on a neighborhood of $g(u_0)$ (see the proof of Corollary 1 and the references given there).

APPENDIX D

Proof of Lemma 2

In the following, we use L_1 to denote the set of complex-valued functions summable over $[0, \infty)$.

For $\mathrm{Re}(z) \geq 0$, we have $[1_n - \tilde{\Lambda}(z)]^{-1} = M(z)\{\det[1_n - \tilde{\Lambda}(z)]\}^{-1}$, in which M is the matrix of transposed cofactors of $[1_n - \tilde{\Lambda}(z)]$ if $n > 1$ and $M = 1$ if $n = 1$. Since $\lambda_{ij} \in L_1$ for each i and j, and the convolution of any two bounded L_1 functions belongs to L_1 and is bounded, it follows that there is a $q \in L_1$, and an $r \in S_n^{(1)}$, such that

$$\det[1_n - \tilde{\Lambda}(z)] = 1 - \int_0^\infty q(t)e^{-zt}dt$$

$$M(z)\tilde{\Lambda}(z) = \int_0^\infty r(t)e^{-zt}dt$$

for $\mathrm{Re}(z) \geq 0$.

Since $\det[1_n - \tilde{\Lambda}(z)] \neq 0$ for $\mathrm{Re}(z) \geq 0$, it follows (see Ref. 42, pp. 60–63) that there is an element s of L_1 such that

$$\{\det[1_n - \tilde{\Lambda}(z)]\}^{-1} = 1 + \int_0^\infty s(\tau)e^{-z\tau}d\tau, \qquad \mathrm{Re}(z) \geq 0.^\dagger$$

Thus, κ defined by

$$\kappa(t) = r(t) + \int_0^t r(t - \tau)s(\tau)d\tau, \qquad t \geq 0$$

---

* Here we use the hypothesis that for each $u \in U$, there is exactly one $x \in X$ such that $f(x) = u$.

† Notice that when $\tilde{\Lambda}(z)$ is *rational* in z, it is a simple matter to show the existence of such an s.

belongs to $S_n^{(1)}$, and $H(z)$ given by

$$H(z) = \int_0^\infty \kappa(\tau)e^{-z\tau}d\tau, \qquad \mathrm{Re}(z) \geq 0$$

satisfies $H(z) = [1_n - \tilde{\Lambda}(z)]^{-1}\tilde{\Lambda}(z)$ for all $\mathrm{Re}(z) \geq 0$.

Since $[1_n - \tilde{\Lambda}(z)]^{-1}\tilde{\Lambda}(z) = \tilde{\Lambda}(z) + \tilde{\Lambda}(z)[1_n - \tilde{\Lambda}(z)]^{-1}\tilde{\Lambda}(z)$ for $\mathrm{Re}(z) \geq 0$, we have

$$\kappa(t) = \lambda(t) + \int_0^t \lambda(\tau)\kappa(t-\tau)d\tau \tag{32}$$

for almost every $t \geq 0$.*

For an arbitrary $p \in L_\infty(\mathrm{C})$, let $q \in L_\infty(\mathrm{C})$ be defined by

$$q(t) = p(t) + \int_0^t \kappa(t-\tau)p(\tau)d\tau, \qquad t \geq 0. \tag{33}$$

We have for $t \geq 0$,

$$q(t) - (\Lambda q)(t) = p(t) + \int_0^t \kappa(t-\tau)p(\tau)d\tau - \int_0^t \lambda(t-\tau)p(\tau)d\tau$$
$$- \int_0^t \lambda(t-\tau_1)\int_0^{\tau_1} \kappa(\tau_1-\tau_2)p(\tau_2)d\tau_2 d\tau_1.$$

Since

$$\int_0^t \lambda(t-\tau_1)\int_0^{\tau_1} \kappa(\tau_1-\tau_2)p(\tau_2)d\tau_2 d\tau_1$$

can be expressed as

$$\int_0^t \int_0^{t-\beta} \lambda(\alpha)\kappa(t-\beta-\alpha)p(\beta)d\alpha d\beta$$

for $t \geq 0$ [the justification of the interchange of order of integration being provided by Theorems of Fubini and Tonelli (Ref. 43, pp. 137–45)], we have, using (32), $(I - \Lambda)q = p$. Thus, $(I - \Lambda)$ maps $L_\infty(\mathrm{C})$ onto itself. Similarly, (32) holds with λ and κ interchanged in the integral, and (32) so modified can be used to show (see the proof of Theorem I of Ref. 39) that whenever there is a solution $q \in L_\infty(\mathrm{C})$ of $(I - \Lambda)q = p$ with $p \in L_\infty(\mathrm{C})$, then (33) holds. This establishes Parts (*i*) and (*ii*) of the lemma.

Suppose now that λ is continuous on $[0, \infty)$. Since $\kappa \in S_n^{(1)}$, by Part

* Equation (32) is a matrix-valued-function version of the usual equation for the resolvent kernel.

(*ii*) of Lemma 3, the integral in (32) depends continuously on t for $t \geq 0$. Thus, by (32), the values of κ agree almost everywhere on $[0, \infty)$ with those of a continuous function. This completes the proof of the lemma.

APPENDIX E

Proof of Lemma 3

Consider k. That each k_{i1} is measurable on $[0, \infty)^l$ and satisfies

$$\int_{[0,\infty)^l} |k_{i1}(\tau_1, \cdots, \tau_l)| \, d(\tau_1, \cdots, \tau_l) < \infty$$

follows from a direct application of Theorems of Fubini and Tonelli (Ref. 43, pp. 137–45). (See the proof in Ref. 44, pp. 99–100, for the $l = 1$ case.) Since every s_{ij} is summable over $[0, \infty)$, and h is bounded, we see that k is bounded. Thus, (*i*) holds.

Suppose now that h is continuous on $[0, \infty)^l$. Let $\alpha_m \geq 0$ be given for $m = 1, 2, \cdots, l$, let δ_m for $m = 1, 2, \cdots, l$ be real variables such that each $(\alpha_m + \delta_m)$ is nonnegative, and let Δ be defined by

$$\Delta(\alpha + \delta) = \int_0^\infty s_{ij}(\tau) \tilde{h}_{j1}(\alpha_1 + \delta_1 - \tau, \cdots, \alpha_l + \delta_l - \tau) d\tau$$

for any i and j. Let $\gamma > 0$ be given. With b_1 and b_2 such that $|\tilde{h}_{j1}(\tau_1, \cdots, \tau_l)| \leq b_1$ and $|s_{ij}(\tau)| \leq b_2$ for $(\tau_1, \cdots, \tau_l) \in [0, \infty)^l$ and $\tau \in [0, \infty)$, choose $\tau_0 \in (0, \infty)$ so that

$$2b_1 \int_{\tau_0}^\infty |s_{ij}(\tau)| \, d\tau \leq \frac{1}{2}\gamma,$$

and observe that

$$|\Delta(\alpha + \delta) - \Delta(\alpha)|$$

$$\leq \frac{1}{2}\gamma + b_2 \int_0^{\tau_0} |\tilde{h}_{j1}(\alpha_1 + \delta_1 - \tau, \cdots, \alpha_l + \delta_l - \tau) - \tilde{h}_{j1}(\alpha_1 - \tau, \cdots, \alpha_l - \tau)| \, d\tau. \quad (34)$$

Since γ is arbitrary, and, by the boundedness and uniform continuity of h on compact subsets of $[0, \infty)^l$, the value of the integral in (34) can be made arbitrarily small by choosing

$$\sum_{m=1}^{l} |\delta_m|$$

to be sufficiently small, we see that k is continuous on $[0, \infty)^l$, which proves (*ii*).

Since $h(t-\tau_1, \cdots, t-\tau_l)u(\tau_1, \cdots, \tau_l)$ and $k(t-\tau_1, \cdots, t-\tau_l)(\tau_1, \cdots, \tau_l)$ are bounded and measurable on $(\tau_1, \cdots, \tau_l) \in [0, t]^l$, the multiple integrals

$$\int_{[0,t]^l} h(t-\tau_1, \cdots, t-\tau_l)u(\tau_1, \cdots, \tau_l)d(\tau_1, \cdots, \tau_l)$$

and

$$\int_{[0,t]^l} k(t-\tau_1, \cdots, t-\tau_l)u(\tau_1, \cdots, \tau_l)d(\tau_1, \cdots, \tau_l)$$

exist. Therefore, two repeated [i.e., two $(l-1)$-fold] applications of Fubini's theorem (Ref. 43, p. 137) show that the iterated integrals in (*iii*) exist, that each equals the corresponding multiple integral, and that each is invariant under changes in the order of integration. Notice that the existence of

$$\int_0^t \left[\int_0^t \cdots \int_0^t h(t-\tau_1, \cdots, t-\tau_l)u(\tau_1, \cdots, \tau_l)d\tau_1 \cdots d\tau_l \right] d\tau \tag{35}$$

for any $t > 0$ can be established in essentially the same way.

Now let p be defined on $[0, \infty)$ by

$$p(t) = \int_0^t \cdots \int_0^t h(t-\tau_1, \cdots, t-\tau_l) \cdot u(\tau_1, \cdots, \tau_l)d\tau_1 \cdots d\tau_l, \qquad t \geq 0. \tag{36}$$

Since $h \in S_1^{(l)}$, and u is bounded on $[0, \infty)^l$, it is clear from the relationship between the iterated integral in (36) and the corresponding multiple integral that p is bounded on $[0, \infty)$. That p is *measurable* on $[0, \infty)$, is a consequence of the existence of (35) for all $t > 0$.

Similarly, again using Fubini's theorem and the fact that a bounded measurable function defined on a set E of finite measure is summable over E, we have, for any $t \geq 0$,

$$\begin{aligned}
&\int_0^t s(t-\tau) \int_0^\tau \cdots \int_0^\tau h(t-\tau_1, \cdots, t-\tau_l)u(\tau_1, \cdots, \tau_l)d\tau_1 \cdots d\tau_l d\tau \\
&\quad = \int_0^t s(\tau) \int_0^{t-\tau} \cdots \int_0^{t-\tau} h(t-\tau-\tau_1, \cdots, t-\tau-\tau_l) \\
&\qquad \cdot u(\tau_1, \cdots, \tau_l)d\tau_1 \cdots d\tau_l d\tau \\
&\quad = \int_0^t s(\tau) \int_{[0,t]^l} \tilde{h}(t-\tau-\tau_1, \cdots, t-\tau-\tau_l)
\end{aligned}$$

$$\cdot u(\tau_1, \cdots, \tau_l)d\tau_1 \cdots d\tau_l d\tau$$

$$= \int_{[0,t]^l} \int_0^t s(\tau)\tilde{h}(t-\tau-\tau_1, \cdots, t-\tau-\tau_l)d\tau$$

$$\cdot u(\tau_1, \cdots, \tau_l)d(\tau_1, \cdots, \tau_l)$$

$$= \int_{[0,t]^l} k(t-\tau_1, \cdots, t-\tau_l)u(\tau_1, \cdots, \tau_l)d(\tau_1, \cdots, \tau_l)$$

$$= \int_0^t \cdots \int_0^t k(t-\tau_1, \cdots, t-\tau_l)u(\tau_1, \cdots, \tau_l)d\tau_1 \cdots d\tau_l,$$

showing that (iv) is met. This completes the proof of the lemma.

APPENDIX F

Proof of Lemma 4

By Fubini's theorem (Ref. 43, p. 137), and the proposition that bounded measurable functions on a set E of finite measure are summable on E,

$$\int_{[0,T]^{(p+q)}} |h_{i1}(\tau_1, \cdots, \tau_p)k_{i1}(\tau_{p+1}, \cdots, \tau_{p+q})|\, d(\tau_1, \cdots, \tau_{p+q})$$

$$\leq \int_{[0,\infty)^p} |h_{i1}(\tau_1, \cdots, \tau_p)|\, d(\tau_1, \cdots, \tau_p)$$

$$\cdot \int_{[0,\infty)^q} |k_{i1}(\tau_{p+1}, \cdots, \tau_{p+q})|\, d(\tau_{p+1}, \cdots, \tau_{p+q})$$

for each i and any finite $T > 0$, from which it is clear that the lemma holds.

APPENDIX G

On the Necessity of the Condition That $\mathrm{Det}[1_n - \tilde{\Lambda}(z)] \neq 0$ *for* $\mathrm{Re}(z) \geq 0$

Proposition 7: Let D.1 (which appears just before Lemma 2) hold. If for each $q \in L_\infty(\mathrm{C})$ there is a $p \in L_\infty(\mathrm{C})$ such that $(I - \Lambda)p = q$, then $\det[1_n - \tilde{\Lambda}(z)] \neq 0$ for $\mathrm{Re}(z) \geq 0$.

Proof:

Since $\lambda \in S_n^{(1)}$, by a standard successive approximations approach (Ref. 38, Section 1.13), it can be shown that there is an $n \times n$ matrix-valued function κ defined on $[0, \infty)$ such that each κ_{ij} is square summable on any finite interval $[0, \beta]$, and

$$\kappa(t) = \lambda(t) + \int_0^t \lambda(\tau)\kappa(t-\tau)d\tau, \qquad t \geq 0 \tag{37}$$

(i.e., and such that (32) is met for $t \geq 0$). From (37) and the Schwarz inequality, the κ_{ij} are bounded on finite intervals. Using Fubini's theorem (see the proof of Theorem I of Ref. 39), if $(I - \Lambda)p = q$ with q and p in $L_\infty(\mathrm{C})$, then

$$p(t) = q(t) + \int_0^t \kappa(t - \tau)q(\tau)d\tau, \qquad t \geq 0.$$

Thus, by the hypothesis of the proposition, each κ_{ij} is summable on $[0, \infty)$. In particular, the Laplace transform $\tilde{H}(z)$ of κ, given by

$$\tilde{H}(z) = \int_0^\infty \kappa(t)e^{-zt}dt$$

exists for all $\mathrm{Re}(z) \geq 0$.

We have $(I - \Lambda)(I + H)q = q$ for every $q \in L_\infty(\mathrm{C})$, in which H is the convolution transformation defined in $L_\infty(\mathrm{C})$ in the usual way in terms of κ. Now let q be given by $q(t) = e^{-t}c_i$ for $t \geq 0$, in which c_i is the column n-vector whose ith component is unity and all other components are zero. Upon taking the Laplace transform of both sides of $(I - \Lambda)(I + H)q = q$, we find that $[1_n - \tilde{\Lambda}(z)][1_n + \tilde{H}(z)]c_i = c_i$ for $\mathrm{Re}(z) \geq 0$ and each i. Therefore, $[1_n - \tilde{\Lambda}(z)][1_n + \tilde{H}(z)] = 1_n$ for $\mathrm{Re}(z) \geq 0$, which shows that $\det[1_n - \tilde{\Lambda}(z)] \neq 0$ for $\mathrm{Re}(z) \geq 0$.

REFERENCES

1. V. Volterra, "Sopra le Funzioni che Dipendono da Altre Funzioni, Nota 1," Rend. Lincei Ser. 4, 3 (1887), pp. 97–105.
2. V. Volterra, *The Theory of Functionals and of Integral and Integro-differential Equations*, New York: Dover, 1959.
3. M. Fréchet, "Sur les Fonctionnelles Continues," Ann. de L'Ecole Normale sup., 3rd Series, *27* (1910).
4. N. Wiener, "Response of a Non-linear Device to Noise," Massachusetts Institute of Technology Radiation Laboratory Report 129 (April 1942).
5. N. Wiener, *Nonlinear Problems in Random Theory*, Cambridge, Mass.: M.I.T. Press, 1958.
6. M. Schetzen, *The Volterra and Wiener Theories of Nonlinear Systems*, New York: John Wiley, 1980.
7. L. A. Zadeh, "On the Representation of Nonlinear Operators," IRE Wescon Convention Record, Part 2 (1957), pp. 105–13.
8. E. Bedrosian and S. O. Rice, "The Output Properties of Volterra Systems Driven by Harmonic and Gaussian Inputs," Proc. IEEE, *59* (December 1971), pp. 1688–707.
9. M. B. Brilliant, "Theory of the Analysis of Nonlinear Systems," Massachusetts Institute of Technology Research Laboratory of Electronics, Report No. 345 (1958).
10. P. d'Allessandro, A. Isidori, and A. Ruberti, "Realizations and Structure Theory of Bilinear Dynamical Systems," SIAM J. Control, *12*, No. 3 (August 1974), pp. 517–35.
11. S. B. Karmaker, "Approximate Analysis of Non-Linear Systems by Laplace Transform," J. Sound and Vibration, *69*, No. 4 (April 1980), pp. 597–602.
12. J. Katzenelson and L. A. Gould, "The Design of Nonlinear Filters and Control Systems, Part II," Information and Control, 7 (June 1964), pp. 117–45.
13. E. Weber, "Complex Convolution Applied to Non-linear Problems," Proc. Symp. Non-linear Circuit Analysis, Polytechnic Inst. of Brooklyn, *VI* (1956), pp. 409–27.
14. B. J. Leon and D. J. Schaefer, "Volterra Series and Picard Iteration for Nonlinear

Circuits and Systems," IEEE Trans. Circuits and Systems, *25* No. 9 (September 1978), pp. 789–93.
15. C. Bruni, G. DiPillo, and G. Koch, "On the Mathematical Models of Bilinear Systems," Ricerche Di Automatica, *2(i)*, (1971).
16. R. W. Brockett, "Volterra Series and Geometric Control Theory," Automatica *12* (1976), pp. 167–76. See also R. W. Brockett and E. G. Gilbert, "An Addendum to Volterra Series and Geometric Control Theory," Automatica, *12* (1976), p. 635.
17. R. B. Parente, "Nonlinear Differential Equations and Analytic Systems Theory," SIAM J. Appl. Math., *18* (January 1970), pp. 41–66.
18. H. L. Van Trees, "Functional Techniques for the Analysis of the Nonlinear Behavior of Phase-locked Loops," Proc. IEEE, *52* (August 1964), pp. 894–911.
19. S. Narayanan, "Application of Volterra Series to Intermodulation Distortion Analysis of a Transistor Feedback Amplifier," IEEE Trans. Circuit Theory, *17* (November 1970), pp. 518–27.
20. E. G. Gilbert, "Functional Expansions for the Response of Nonlinear Differential Systems," IEEE Trans. Automatic Control, *22* (December 1977), pp. 909–21.
21. L. O. Chua and C. Y. Ng, "Frequency Domain Analysis of Nonlinear Systems: Formulation of Transfer Functions," Elect. Circuits and Systems, *3* No. 6 (November 1979), pp. 257–69.
22. S. Narayanan, "Transistor Distortion Analysis using Volterra Series Representations," B.S.T.J., *46* No. 5 (May–June 1967), pp. 991–1023.
23. C. Lesiak and A. J. Krener, "The Existence and Uniqueness of Volterra Series for Nonlinear Systems," IEEE Trans. Automatic Control, *23* No. 6 (December 1978), pp. 1090–5.
24. A. V. Balakrishnan, *Applied Functional Analysis*, New York: Springer Verlag, 1976, pp. 158–61.
25. M. Fliess, "A Note on Volterra Series for Nonlinear Differential Systems," IEEE Trans. Automatic Control, *25* (February 1980), pp. 116–7.
26. W. A. Porter, "An Overview of Polynomic Systems Theory," Proc. IEEE, *64* (January 1976), pp. 18–23.
27. A. Halme, J. Orava, and H. Blomberg, "Polynomial Operators in Nonlinear System Theory," Int. J. Syst. Sci., *2* No. 1 (July 1971), pp. 25–47.
28. P. A. Prenter, "On Polynomial Operators and Equations," *Nonlinear Functional Analysis and Applications*, L. B. Rall, ed., New York: Academic Press (1971).
29. J. Orava and A. Halme, "Inversion of Generalized Power Series Representation," J. of Mathematical Analysis and Applications, *45* No. 1 (January 1974), pp. 136–41.
30. J. Dieudonné, *Foundations of Modern Analysis*, New York: Academic Press, 1969.
31. T. M. Flett, *Differential Analysis*, London: Cambridge University Press, 1980.
32. I. J. Maddox, *Elements of Functional Analysis*, London: Cambridge University Press, 1977, pp. 118–9.
33. L. M. Graves, "Reimann Integration and Taylor's Theorem in General Analysis," Trans. Amer. Mathematical Soc., *29* (January 1927), pp. 163–77.
34. E. Hille and R. S. Phillips, *Functional Analysis and Semi-groups*, Providence: Amer. Mathematical Soc., 1957.
35. F. Riesz and B. Sz.-Nagy, *Functional Analysis*, New York: Frederick Ungar Publishing Co., 1955.
36. I. W. Sandberg, "Some Results on the Theory of Physical Systems Governed by Nonlinear Functional Equations," B.S.T.J., *44* No. 5 (May 1965), pp. 871–98.
37. A. Halme, "Polynomial Operators for Nonlinear Systems Analysis," Acta Polytech. Scand. Ma, *24* (1972), pp. 1–63. (Dissertation).
38. F. G. Tricomi, *Integral Equations*, New York: Interscience, 1957.
39. I. W. Sandberg, "Signal Distortion in Nonlinear Feedback Systems," B.S.T.J., *42* No. 6 (November 1963), pp. 2533–49.
40. I. W. Sandberg, "On the Theory of Linear Multi-loop Feedback Systems," *Feedback Systems*, J. B. Cruz, Jr., ed., New York: McGraw-Hill, 1972.
41. J. G. Truxal, *Automatic Feedback Control System Synthesis*, New York: McGraw-Hill, 1955, p. 114.
42. R. E. Paley and N. Wiener, *Fourier Transforms in the Complex Domain*, Providence: Amer. Mathematical Soc., 1934.
43. S. McShane, *Integration*, Princeton: Princeton Univ. Press, 1944.
44. S. Bochner and K. Chandrasekharan, *Fourier Transforms*, Princeton: Princeton Univ. Press, 1949.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 2, February 1982
Printed in U.S.A.

Volterra Expansions for Time-Varying Nonlinear Systems

By I. W. SANDBERG

(Manuscript received August 13, 1981)

Recent results show for the first time the existence of a locally convergent Volterra-series representation for the input-output relation of a certain important large class of time-invariant *nonlinear systems containing an arbitrary finite number of nonlinear elements. (Systems of the type considered arise, for example, in the area of communication channel modeling.) Here corresponding results are given for time-varying systems, which arise frequently. A key hypothesis of our main theorem, which asserts that a convergent Volterra expansion exists under certain specified conditions, has the useful property that it is met if a certain "linearized subgraph" of the system is bounded-input bounded-output stable.*

I. INTRODUCTION

This paper is a continuation of the study initiated in Ref. 1 concerning operator-type models of dynamic nonlinear physical systems, such as communication channels and control systems. Reference 1 addresses the problem of determining conditions under which there exists a power-series-like expansion, or a polynomial-type approximation, for a system's outputs in terms of its inputs. Related problems concerning properties of the expansions are also considered, and nonlocal as well as local results are presented. In particular, the results in Ref. 1 show for the first time the existence of a locally convergent Volterra-series representation for the input-output relation of a certain important large class of *time-invariant* systems containing an arbitrary finite number of nonlinear elements.

The main purpose of this paper is to give corresponding results applicable to time-varying systems, which arise frequently. Also, the results obtained here by specializing to the time-invariant case involve weaker hypotheses concerning the nonlinear elements (here mutual

coupling is not ruled out, and, at the expense of somewhat more complicated proofs, we show how to proceed without the local invertibility of a certain mapping associated with the nonlinear elements*).

With regard to background material, functional power series of the form

$$k_0 + \sum_{m=1}^{\infty} \int_a^b \cdots \int_a^b k_m(t, \tau_1, \cdots, \tau_m) u(\tau_1) \cdots u(\tau_m) d\tau_1 \cdots d\tau_m, \quad (1)$$

in which k_0 is a constant, t is a parameter, and u and the k_m for $m \geq 1$ are continuous functions, were considered in 1887 by Vito Volterra[2, 3] in connection with his studies of functions of functions (which provided much of the initial motivation to develop the field now known as functional analysis). About twenty years later, Fréchet[4] proved that a continuous real functional (i.e., a continuous real scalar-valued map) defined on a compact set of real continuous functions on $[a, b]$ could be approximated by a sum of a finite number of terms in Volterra's series (1), but with (in analogy with the well-known Weierstrass approximation theorem) the number of terms as well as the k_m dependent on the degree of approximation. Further background material (concerning, in particular, bilinear and polynomic systems) omitted here to avoid unnecessary repetition, can be found in Ref. 1.

Our results are given in the next section, which begins with some mathematical preliminaries followed by a description of the general class of systems to be addressed. Of interest with regard to our main result, Theorem 2 below, is that a key hypothesis has the useful property that it is met if a certain "linearized subgraph" of the system is bounded-input bounded-output stable.

II. SYSTEMS AND EXPANSIONS

2.1 Preliminaries

Throughout Section II we use $L_\infty(\mathrm{C})$ to denote the complex Banach space of Lebesque measurable complex column n-vector-valued functions v defined on the interval $[0, \infty)$ such that the jth component v_j of v satisfies $\sup_{t \geq 0} |v_j(t)| < \infty$ for $j = 1, 2, \cdots, n$, and where the norm $\|\cdot\|$ on $L_\infty(\mathrm{C})$ is given by $\|v\| = \max_j \sup_t |v_j(t)|$. (As usual, n denotes an arbitrary positive integer.) The symbol θ stands for the zero element of $L_\infty(\mathrm{C})$. We use $H(\mathrm{C})$ to denote the linear space of complex column n-vector-valued functions h defined on $[0, \infty)$ such that truncations of h belong to $L_\infty(\mathrm{C})$ (i.e., such that $h_{(\omega)} \in L_\infty(\mathrm{C})$ for $\omega \in (0, \infty)$, where

* A simple way to circumvent the need for invertibility mentioned in (Ref. 1, Comments of Section 3.5) is often much less satisfactory for the purpose of obtaining explicit expressions for the Volterra kernels.

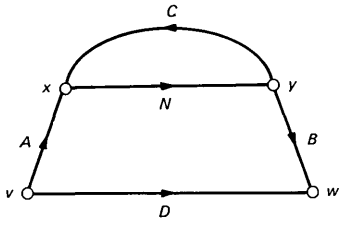


Fig. 1—Signal flow graph.

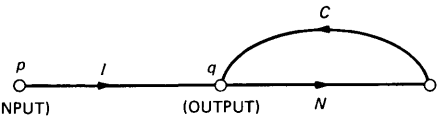


Fig. 2—Feedback part of the flow graph of Fig. 1.

$h_{(\omega)}(t) = h(t)$ for $t \leq \omega$ and $h_{(\omega)}(t) = 0$ otherwise). Clearly unlike $L_\infty(\mathrm{C})$, $\mathrm{H}(\mathrm{C})$ can contain unbounded functions.

2.2 The class of systems

Consider a physical system with an input v drawn from $L_\infty(\mathrm{C})$ and an output w contained in $H(\mathrm{C})$. Let the system be composed of linear elements as well as nonlinear elements. Suppose that the nonlinear elements can be viewed as collectively introducing a constraint that can be written as $y = Nx$, in which N is a map from one subset of $H(\mathrm{C})$ into another, where x and y, respectively, are the $H(\mathrm{C})$ input and output of the nonlinear part of the system.

With regard to the remainder of the system, which is linear, assume that there are linear maps A, B, C, and D of $H(\mathrm{C})$ into itself such that

$$x = Av + Cy \tag{2}$$

$$w = Dv + By. \tag{3}$$

A signal-flow-graph representation of the relations under consideration is given in Fig. 1.* Concerning the degree of generality of the model, and the assumption that the values of v, w, x, and y have the *same dimension* n, notice that we have not ruled out the possibility that some components of v, x, and/or y have no effect on the system, and, similarly, that certain of the components of w can be ignored. Nonzero initial conditions, if any, are assumed to be able to be taken into account either in N or as inputs to the system.

2.3 General expansions

Consider three hypotheses:

A.1: The restrictions of A, B, C, and D to $L_\infty(\mathrm{C})$ are bounded linear maps of $L_\infty(\mathrm{C})$ into itself.

A.2: There is an open neighborhood S_0 of θ in $L_\infty(\mathrm{C})$ such that N maps

* This is the same class of systems introduced in Ref. 1. Such representations of systems have been used in different but related settings in Refs. 5, 6, and 7. The maps A, B, C, and D exist for a very large class of systems.

S_0 into $L_\infty(C)$ with $N(\theta) = \theta$, and d^mN (the mth order Fréchet derivative of N) exists on S_0 for every $m = 1, 2, \cdots$.

A.3: $[I - C_\infty dN(\theta)]$ is an invertible map of $L_\infty(C)$ onto $L_\infty(C)$, in which I is the identity transformation on $L_\infty(C)$, and C_∞ is the restriction of C to $L_\infty(C)$.*

We shall prove the following general result.

Theorem 1: *When A.1, A.2, and A.3 are met, there is a positive number δ and an open subset S of S_0 with the following properties:*

(i) $\theta \in S$, and for each $v \in L_\infty(C)$ with $\|v\| < \delta$ there exist unique x, y, and w of S, $L_\infty(C)$, and $L_\infty(C)$, respectively, such that (2), (3), and $y = Nx$ hold.

(ii) The function w described in (i) is given by

$$w = Dv + \sum_{m=1}^{\infty} B[g_m(Av)]_2 \tag{4}$$

for $\|v\| < \delta$, in which the $[g_m(Av)]_2$ are defined recursively by the relations

$$[g_1(Av)]_1 = [I - C_\infty dN(\theta)]^{-1} Av \tag{5}$$

$$[g_1(Av)]_2 = dN(\theta)[g_1(Av)]_1 \tag{6}$$

and

$$h_m = \sum_{l=2}^{m} (l!)^{-1} \sum_{\substack{k_1+k_2+\cdots+k_l=m \\ k_j>0}} d^lN(\theta)[g_{k_1}(Av)]_1 \cdot [g_{k_2}(Av)]_1 \cdots [g_{k_l}(Av)]_1^{\dagger} \tag{7}$$

$$[g_m(Av)]_1 = [I - C_\infty dN(\theta)]^{-1} C_\infty h_m \tag{8}$$

$$[g_m(Av)]_2 = dN(\theta)[g_m(Av)]_1 + h_m \tag{9}$$

for $m \geq 2$. In addition, the series on the right side of (4) converges uniformly with respect to $\|v\| < \delta$.

2.3.1 Proof of Theorem 1

Notice that (2) and $y = Nx$ can be written as $x - CNx = Av$ and $Nx - y = \theta$, for y and Nx belonging $L_\infty(C)$, and that an expansion for w in terms of v can be obtained at once from (3) and an expansion for y in terms of v. These observations motivate us to proceed as follows.‡

---

* Of course, $dN(\theta)$ denotes the Fréchet derivative of N at the point θ.

† In (7), $\sum_{\substack{k_1+k_2+\cdots+k_l=m \\ k_j>0}}$ denotes a sum over all positive integers $k_1, \cdots, k_l$ that add to m.

‡ An alternative way not pursued here to prove a result along the lines of Theorem 1 involves obtaining an expansion for x in terms of v, one for y in terms of x, and substituting the former into the latter.

Let $\mathscr{B}$ denote the Banach space $L_\infty(C) \times L_\infty(C)$, whose elements we take to be two-component *column* vectors, normed by $\|u\| = \max(\|u_1\|, \|u_2\|)$ for $(u_1, u_2)' \in L_\infty(C) \times L_\infty(C)$, in which "′" denotes transpose. [We use the same symbol for the norms associated with $\mathscr{B}$ and $L_\infty(C)$. The meaning of the symbol will be clear from the context in which it is used.] Let S_1 be any open ball in $L_\infty(C)$ of positive radius centered at θ.

Define $F: S_0 \times S_1 \to \mathscr{B}$ to be the map given by

$$F_1(p_1, p_2) = p_1 - C_\infty N p_1$$

$$F_2(p_1, p_2) = N p_1 - p_2$$

for $p \in S_0 \times S_1$.

The set $S_0 \times S_1$ is open in $\mathscr{B}$. By A.2 it easily follows that the derivative $dF(p): \mathscr{B} \to \mathscr{B}$ exists and is continuous at each point p of $S_0 \times S_1$, and that it is given by

$$dF(p) = \begin{pmatrix} [I - C_\infty dN(p_1)] & 0 \\ dN(p_1) & -I \end{pmatrix}, \tag{10}$$

in which here "0" denotes the transformation of $L_\infty(C)$ into itself that replaces each element by θ. By A.3, it follows that $dF(p)$ is an invertible map of $\mathscr{B}$ onto $\mathscr{B}$, with inverse given by

$$dF(p)^{-1} = \begin{pmatrix} [I - C_\infty dN(p_1)]^{-1} & 0 \\ dN(p_1)[I - C_\infty dN(p_1)]^{-1} & -I \end{pmatrix} \tag{11}$$

for $p \in S_0 \times S_1$. Since $dF(p)$ is invertible at $p = (\theta, \theta)'$, by a standard inverse function theorem (Ref. 8, page 273; see also the comment in Ref. 1 concerning Lemma 1 of Ref. 1), there are open neighborhoods S_2 and S_3 of $(\theta, \theta)'$ in $\mathscr{B}$, with $S_2 \subset S_0 \times S_1$, such that for each $q \in S_3$ there is in S_2 a unique p such that $F(p) = q$. Using the boundedness of the restriction of A to $L_\infty(C)$, $\delta_1 > 0$ can be chosen so that $(Av, \theta)' \in S_3$ *for* $v \in L_\infty(C)$ with $\|v\| < \delta_1$, and thus so that for each such v, there is in S_2 a unique $(x, y)'$ with the property that $F(x, y) = (Av, \theta)'$ [i.e., such that (2) and $y = Nx$ are met].

Observe that the set $S = \{u: (u, Nu)' \in S_2\}$ is an open subset of S_0, and that for any $\delta \in (0, \delta_1)$ and for each $v \in L_\infty(C)$ with $\|v\| < \delta$, there is a unique (x, y, w) in $S \times L_\infty(C) \times L_\infty(C)$ such that (2), (3), and $y = Nx$ hold, as claimed in (*i*).

With regard to part (*ii*), we shall use the following Lemma in which f denotes any map from an open subset X of the Banach space $\mathscr{B}$ into $\mathscr{B}$ with the property that there is a nonempty open convex subset U of $\mathscr{B}$ such that for each $u \in U$ there is in X a unique x_u such that $f(x_u) = u$, and in which g stands for the map of U into X defined by $f[g(u)] = u$ for $u \in U$.

Lemma 1: *Assume that the Fréchet derivative $d^m f$ exists on X for each m. Let $u_0 \in U$, and suppose that $df[g(u_0)]$ is an invertible map of $\mathscr{B}$ onto itself. Then there is a $\sigma > 0$ such that the expansion*

$$g(u) = g(u_0) + \sum_{m=1}^{\infty} g_m(u_0, u - u_0)$$

is valid and uniformly convergent for $u \in U$ with $\|u - u_0\| < \sigma$, where

$$g_1(u_0, u - u_0) = df[g(u_0)]^{-1}(u - u_0),$$

and

$$g_m(u_0, u - u_0) = -df[g(u_0)]^{-1} \sum_{l=2}^{m} (l!)^{-1} \sum_{\substack{k_1+k_2+\cdots+k_l=m \\ k_j>0}} d^l f[g(u_0)]$$

$$\cdot g_{k_1}(u_0, u - u_0) g_{k_2}(u_0, u - u_0) \cdots g_{k_l}(u_0, u - u_0), \qquad m \geq 2.$$

Lemma 1 is a special case of Theorem 4 of Ref. 1 (see also Ref. 9).

With S_2 and S_3 as indicated above before Lemma 1, choose $X = S_2$, assume without loss of generality that S_3 is convex, and take $U = S_3$. Throughout the remainder of this section, let f denote the restriction of F to X. The following lemma is proved in Appendix A.

Lemma 2: *Under the conditions of Theorem* 1, *for each $p \in X$ and every $l = 2, 3, \cdots$ the l*th *order Fréchet derivative $d^l f(p)$ exists, and we have*

$$d^l f(p) h_1 h_2 \cdots h_l = \begin{pmatrix} -C_\infty d^l N(p_1) h_{11} h_{21} \cdots h_{l1} \\ d^l N(p_1) h_{11} h_{21} \cdots h_{l1} \end{pmatrix} \tag{12}$$

for any elements $h_1, h_2, \cdots, h_l$ of $\mathscr{B}$ (where h_{j1} denotes the first component of h_j for each j).

By Lemmas 1 and 2, there is a $\sigma > 0$ such that S_3 contains an open ball in $\mathscr{B}$ centered at $(\theta, \theta)'$ of radius σ, and the solution $r \in X$ of $f(r) = s$ for $s \in \mathscr{B}$ with $\|s\| < \sigma$ is given by the uniformly convergent series $\sum_{m=1}^{\infty} g_m(s)$, in which $g_1(s) = df[(\theta, \theta)']^{-1} s$, and

$$g_m(s) = -dF[(\theta, \theta)']^{-1} \sum_{l=2}^{m} (l!)^{-1} \sum_{\substack{k_1+k_2+\cdots+k_l=m \\ k_j>0}}$$

$$\cdot d^{\mathscr{L}} f[(\theta, \theta)'] g_{k_1}(s) \cdots g_{k_l}(s)$$

for $m \geq 2$. In particular, by (11) and Lemma 2, when $s \in B$ with $\|s\| < \sigma$ and s has the form $(s_1, \theta)'$, we have

$$g_1(s) = \begin{pmatrix} [I - C_\infty dN(\theta)]^{-1} s_1 \\ dN(\theta)[I - C_\infty dN(\theta)]^{-1} s_1 \end{pmatrix} \tag{13}$$

and

$$g_m(s) = -\begin{pmatrix} [I - C_\infty dN(\theta)]^{-1} & 0 \\ dN(\theta)[I - C_\infty dN(\theta)]^{-1} & -I \end{pmatrix} \sum_{l=2}^{m} (l!)^{-1}$$

$$\cdot \sum_{\substack{k_1+k_2+\cdots+k_l=m \\ k_j>0}} \begin{pmatrix} -C_\infty d^l N(\theta)[g_{k_1}(s)]_1 \cdots [g_{k_l}(s)]_1 \\ d^l N(\theta)[g_{k_1}(s)]_1 \cdots [g_{k_l}(s)]_1 \end{pmatrix} \quad (14)$$

for $m \geq 2$.

Now choose $\delta \in (0, \delta_1)$ so that $\|v\| < \delta$ implies that $\|Av\| < \sigma$, and, referring to the g_m of (13) and (14), observe that for $\|v\| < \delta$, y of part (i) of the theorem is given by $y = \sum_{m=1}^{\infty} \{g_m[(Av, \theta)']\}_2$. From (14),

$$[g_m(s)]_1 = [I - C_\infty dN(\theta)]^{-1}$$

$$\cdot \sum_{l=2}^{m} (l!)^{-1} \sum_{\substack{k_1+k_2+\cdots+k_l=m \\ k_j>0}} C_\infty d^l N(\theta)[g_{k_1}(s)]_1 \cdots [g_{k_l}(s)]_1,$$

and

$$[g_m(s)]_2 = dN(\theta)[I - C_\infty dN(\theta)]^{-1}$$

$$\cdot \sum_{l=2}^{m} (l!)^{-1} \sum_{\substack{k_1+k_2+\cdots+k_l=m \\ k_j>0}} C_\infty d^l N(\theta)[g_{k_1}(s)]_1 \cdots [g_{k_l}(s)]_1$$

$$+ \sum_{l=2}^{m} (l!)^{-1} \sum_{\substack{k_1+k_2+\cdots+k_l=m \\ k_j>0}} d^l N(\theta)[g_{k_1}(s)]_1 \cdots [g_{k_l}(s)]_1$$

for $m \geq 2$ and $\|s_1\| < \sigma$ which, together with (13), completes the proof of the theorem. (The $[g_m(Av)]_i$ in Theorem 1 correspond to the $\{g_m[(Av, \theta)']\}_i$ here.)

2.3.2 Comments

In principle, it is straightforward to give an explicit expression for any term in the series in (4). For example, it is a simple exercise to verify that the third-order term $B[g_3(Av)]_2$ is

$$\tfrac{1}{6}B\{dN(\theta)[I - C_\infty dN(\theta)]^{-1}C_\infty + I\}\, d^3N(\theta)$$

$$\cdot \{[I - C_\infty dN(\theta)]^{-1}Av\}^3, \quad (15)$$

when $d^2N(\theta)$ is the zero operator (i.e., is the zero operator in the space to which $d^2N(\theta)$ belongs). If $dN(\theta)$ also is the zero operator, then of course (15) is simply

$$\tfrac{1}{6}Bd^3N(\theta)(Av)^3.$$

The interpretation of (15), and more general expressions, under certain assumptions concerning the forms of N, A, B, and C is addressed in the following section.

Theorem 1 (and Lemma 1) hold if $L_\infty(\mathrm{C})$ and $H(\mathrm{C})$, respectively, are replaced with any complex Banach space and any linear space containing the elements of the complex Banach space.

2.4 Volterra expansions

In this section, we introduce, discuss, and prove our main result. We begin by considering the following definitions and two hypotheses, B.1 and B.2.

For each $l = 1, 2, \cdots$, let $R_0(l)$ denote the subset of $\mathrm{R}^{(l+1)}$ given by $R_0(l) = \{(v_0, v_1, \cdots, v_l) \in \mathrm{R}^{(l+1)}: v_0 \geq v_i \geq 0 \text{ for } i = 1, 2, \cdots, l\}$.

For any positive integers q and l, let $S_q^{(l)}$ denote the set of complex $n \times q$ matrix-valued functions h defined on $R_0(l)$ such that each h_{ij} is Lebesgue measurable and bounded on $R_0(l)$, and satisfies

$$\sup_{t \geq 0} \int_{[0,t]^l} |h_{ij}(t, \tau_1, \cdots, \tau_l)| \, d(\tau_1, \cdots, \tau_l) < \infty. \tag{16}$$

B.1: There are elements a, b, c, and d of $S_n^{(1)}$ such that for each $p \in H(\mathrm{C})$,

$$(Ap)(t) = \int_0^t a(t, \tau)p(\tau)d\tau$$

$$(Bp)(t) = \int_0^t b(t, \tau)p(\tau)d\tau$$

$$(Cp)(t) = \int_0^t c(t, \tau)p(\tau)d\tau$$

$$(Dp)(t) = \int_0^t d(t, \tau)p(\tau)d\tau$$

for $t \geq 0$.

In hypothesis B.2 below, Γ_0 denotes the set $\{z \in \mathrm{C}^n: |z_i| < \gamma \text{ for } i = 1, 2, \cdots, n\}$, in which γ is a positive constant and C^n is the normed linear space of complex column n-vectors with zero element θ_C and norm $|\cdot|$ given by $|z| = \max_i |z_i|$ for $z \in \mathrm{C}^n$.

B.2: N is defined on $\Gamma = \{s \in L_\infty(\mathrm{C}): \|s\| < \gamma\}$ by

$$(Ns)(t) = \eta[s(t), t], \qquad t \geq 0$$

where η is a map from $\Gamma_0 \times [0, \infty)$ into C^n with the following properties:

(*i*) $\eta(\theta_\mathrm{C}, t) = \theta_\mathrm{C}$ for $t \geq 0$.

(*ii*) The function ξ given by $\xi(t) = \eta[s(t), t]$, $t \geq 0$ is Lebesgue measurable on $[0, \infty)$ for each $s \in \Gamma$.

(*iii*) For each $t \in [0, \infty)$, $\eta(\cdot, t)$ is a continuous map of Γ_0 into C^n, and for each $t \in [0, \infty)$, for $1 \leq i, j \leq n$, and for any point $\alpha \in \Gamma_0$, the function $z_j \mapsto \eta_i(\alpha_1, \cdots, \alpha_{j-1}, z_j, \alpha_{j+1}, \cdots, \alpha_n, t)$ is differentiable with respect to the complex variable z_j for $|z_j| < \gamma$. [This implies (see Ref. 8, pages 204, 205, 226, 227, 230) the existence throughout Γ_0 of every mth order partial derivative

$$\frac{\partial^m \eta_i}{\partial z_{j_m} \partial z_{j_{m-1}} \cdots \partial z_{j_1}} \tag{17}$$

for each t and all m.]

(*iv*) For any $m, j_1, \cdots, j_m$, and i, the partial derivative (17), which we denote by $p(z_1, \cdots, z_n, t)$, satisfies the conditions that the function $t \mapsto p(0, \cdots, 0, t)$ is bounded on $[0, \infty)$, and that p is uniformly continuous on closed subsets of Γ_0 uniformly in t, in the sense that given a closed $\Gamma_{00} \subset \Gamma_0$ and a $\delta_1 > 0$ there is a $\delta_2 > 0$ such that

$$|p(z_{a1}, \cdots, z_{an}, t) - p(z_{b1}, \cdots, z_{bn}, t)| < \delta_1$$

for $t \geq 0$ whenever z_a and z_b are elements of Γ_{00} such that $|z_a - z_b| < \delta_2$.

Following are comments and an example.

If $\eta(\cdot, t)$ is independent of t and (*iii*) is met, then (*ii*) and (*iv*) are met.

The conditions on η of B.2 are met if, for example,

$$\eta_i(z, t) = \sum_{j=1}^{\rho} \beta_{ij}(t) \lambda_{ij}(z_i), \qquad t \geq 0$$

for each i and $z \in \Gamma_0$, in which ρ is a positive integer, the β_{ij} are C^1-valued bounded measurable functions, and each λ_{ij} is an analytic function from the disk $|z_i| < \gamma$ in C^1 into C^1 such that $\lambda_{ij}(0) = 0$. In this important case, N restricted to Γ can of course be represented by n single-input single-output memoryless, possibly time-varying, nonlinear operators.

In order to introduce another needed hypothesis, consider the following proposition.

Proposition 1: When $c \in S_n^{(1)}$ and η satisfies the conditions of B.2, for each $p \in H(C)$ there exists a unique $q \in H(C)$ such that

$$p(t) = q(t) - \int_0^t c(t, \tau) L(\tau) q(\tau) d\tau, \qquad t \geq 0, \tag{18}$$

where L is the $n \times n$ matrix-valued function defined on $[0, \infty)$ by $L_{ij}(t) = \partial \eta_i(z_1, \cdots, z_n, t)/\partial z_j$ at $z_1 = z_2 = \cdots = z_n = 0$ for each i, j, and t.

Since L is measurable on $[0, \infty)$ (see the proof of Lemma 3 in

Appendix B), Proposition 1 follows at once from Lemma 6 in Section 2.4.2.

B.3: Under the hypotheses of Proposition 1, (18) has the further property that $p \in L_\infty(C)$ implies that the solution q also belongs to $L_\infty(C)$.

The "further property" of B.3 has the interpretation that the feedback part of the graph of Fig. 1, shown in Fig. 2, is *bounded-input bounded-output stable* in the indicated sense when N is replaced with its linearization at the origin [by which we mean its linearization (see Lemma 3 below) at θ extended in the natural way to all of $H(C)$]. The node labeled "output" in Fig. 2 is an intermediate output node. For our purposes here, the output label can be moved to the node to the right of N when the matrix $L(t)^{-1}$ exists for each $t \geq 0$ and has uniformly bounded elements.

We shall use also the following definition and proposition:

Definition: Throughout the remainder of this section, for each l, $\chi[v(\tau_1), \cdots, v(\tau_l)]$ denotes the column vector of order n^l whose elements are the n^l distinct products $v_{\omega_1}(\tau_1)v_{\omega_2}(\tau_2) \cdots v_{\omega_l}(\tau_l)$, corresponding to distinct sequences $\omega_1, \omega_2, \cdots, \omega_l$ with each ω_j drawn from $\{1, 2, \cdots, n\}$, arranged in an arbitrary predetermined order.

Proposition 2: If $k_l \in S_n^{(l)}$ for some l, then the iterated integral

$$\int_0^t \cdots \int_0^t k_l(t, \tau_1, \cdots, \tau_l)\chi[v(\tau_1), \cdots, v(\tau_l)]d\tau_1 \cdots d\tau_l$$

exists and is invariant with respect to interchanges in the order of integration for each $t \geq 0$ and $v \in L_\infty(C)$, and $V_{k_l}(v)$, defined on $[0, \infty)$ by

$$V_{k_l}(v)(t) = \int_0^t \cdots \int_0^t k_l(t, \tau_1, \cdots, \tau_l)\chi[v(\tau_1), \cdots, v(\tau_l)]d\tau_1 \cdots d\tau_l$$

for an arbitrary $v \in L_\infty(C)$, is an element of $L_\infty(C)$.

Proposition 2 is a special case of Lemma 4 of Section 2.4.2.

The following is our main result.

Theorem 2: Suppose that B.1, B.2, and B.3 are met. Then

(i) The hypotheses of Theorem 1 are satisfied.

(ii) For each $l = 1, 2, \cdots$ there is a $k_l \in S_n^{(l)}$ such that

$$w = \sum_{l=1}^{\infty} V_{k_l}(v) \quad \text{for} \quad \|v\| < \delta, \tag{19}$$

with the series uniformly convergent with respect to $\|v\| < \delta$, where v, w, and δ are described in Theorem 1, and $V_{k_l}(\cdot)$ is as indicated in Proposition 2.

(iii) Each k_l can be taken to be continuous on $R_0(l)$ when a and d

are continuous on $R_0(1)$, and b and c meet the condition imposed on s in part (ii) of Lemma 4 below [the condition is met if b and c are continuous on $R_0(1)$].

2.4.1 Comments

Theorem 2 is proved in the next section. Using the proof given there, it can be shown that, as one would expect, the Volterra kernels k_l can be taken to depend on only $(t - \tau_1), \cdots, (t - \tau_l)$ when a, b, c, and d depend only on $(t - \tau)$, and $\eta(\cdot, t)$ is independent of t.*

Similarly, each k_l can be taken to be real valued (i.e., to have zero imaginary part) if a, b, c, d, and the partial derivatives of $\eta(\cdot, t)$ at the origin are real valued. This shows that Theorem 2 establishes the existence of a Volterra-series expansion for the important corresponding case in which v, w, x, and y in Fig. 1 are restricted to be real valued and N (which then would be a map between real-valued function spaces) can be analytically extended so that the hypotheses of the theorem are met.†

For the single-input case in which either $n = 1$ or $v_i(t) = 0$ for all t and $i = 2, 3, \cdots, n$, (19) takes the more familiar form

$$w(t) = \sum_{l=1}^{\infty} \int_0^t \cdots \int_0^t h_l(t, \tau_1, \cdots, \tau_l) \cdot v_1(\tau_1)v_1(\tau_2) \cdots v_1(\tau_l)d\tau_1 d\tau_2 \cdots d\tau_l, \qquad t \geq 0$$

for $\sup_{t\geq 0}|v_1(t)| < \delta$, with the h_l belonging to $S_1^{(l)}$.

By modifying the proof given in Section 2.4.2, results similar to Theorem 2 can be obtained for cases in which the basic underlying function space $L_\infty(C)$ is replaced with another complex Banach space, and/or A, B, C, and D have a more general‡ (or different) form. Of some importance is the case in which $L_\infty(C)$ is replaced with the corresponding set $L_\infty(C)(T)$ of bounded functions defined on a finite interval $[0, T]$, and a theorem along the lines of Theorem 2 for this case is given in Appendix F.

2.4.2 Proof of Theorem 2

Our proof uses five lemmas, which are proved in the appendix, and an inductive argument using Theorem 1. We begin with a description of the lemmas and some associated definitions.

---

* See Proposition 7 and Lemma 2 of Ref. 1.

† In this connection, Theorem 5 of Ref. 1 can be used in place of Lemma 1 to prove results along the same lines as Theorems 1 and 2, but with $L_\infty(C)$ replaced with the corresponding function space over the *real* field, and Corollary 1 of Ref. 1 can be used to obtain corresponding pth order *approximation* results under weaker differentiability hypotheses.

‡ Detailed results for cases in which a, b, c, and d are replaced with certain generalized functions, and N is not necessarily memoryless, will be given in another paper.

Lemma 3: *Suppose that B.2 is met. Then N maps* Γ *into* $L_\infty(C)$, *each*

$$\frac{\partial^m \eta_i[s_1(\cdot), \cdots, s_n(\cdot), \cdot]}{\partial z_{j_m} \partial z_{j_{m-1}} \cdots \partial z_{j_1}}$$

is bounded and measurable on $[0, \infty)$ *for each* $s \in \Gamma$, $d^m N(s)$ *exists for each* $s \in \Gamma$ *and all* $m = 1, 2, \cdots$, *and, for any m, we have* $[d^m N(s) h_1 \cdots h_m(t)]_i =$

$$\sum_{j_1=1}^{n} \sum_{j_2=1}^{n} \cdots \sum_{j_m=1}^{n} \frac{\partial^m \eta_i[s_1(t), \cdots, s_n(t), t]}{\partial z_{j_m} \partial z_{j_{m-1}} \cdots \partial z_{j_1}} h_{1j_1}(t) h_{2j_2}(t) \cdots h_{mj_m}(t), \quad t \geq 0$$

for each $s \in \Gamma$, *each i, and any m elements* $h_1, h_2, \cdots, h_m$ *of* $L_\infty(C)$.

Definition: For each $h \in S_q^{(l)}$, $\tilde{h}$ denotes the function defined on $[0, \infty)^{(l+1)}$ by $\tilde{h} = h$ on $R_0(l)$ and $h = 0_{nq}$ (the zero $n \times q$ matrix) otherwise.

Lemma 4: *Suppose that* $h \in S_1^{(l)}$ *for some* $l \geq 1$, *that* $s \in S_n^{(1)}$, *and that u is a bounded measurable function from* $[0, \infty)^l$ *into the complex numbers. Then*

(*i*) *The function k defined by*

$$k(t, \tau_1, \cdots, \tau_l) = \int_0^t s(t, \tau) \tilde{h}(\tau, \tau_1, \cdots, \tau_l) d\tau$$

for $(t, \tau_1, \cdots, \tau_l) \in R_0(l)$, *belongs to* $S_1^{(l)}$.

(*ii*) *If h is continuous on* $R_0(l)$, *and* $\tilde{s}$ *meets the condition that each* δ_{ij}, *given by*

$$\delta_{ij}(\alpha, t) = \int_0^\infty |\tilde{s}_{ij}(t + \alpha, \tau) - \tilde{s}_{ij}(t, \tau)| d\tau$$

for $t \geq 0$ *and* $(t + \alpha) \geq 0$, *satisfies* $\delta_{ij}(\alpha, t) \to 0$ *as* $\alpha \to 0$ *for each t, then k is continuous on* $R_0(l)$.

(*iii*) *The iterated integrals*

$$\int_0^t \cdots \int_0^t h(t, \tau_1, \cdots, \tau_l) u(\tau_1, \cdots, \tau_l) d\tau_1 \cdots d\tau_l$$

and

$$\int_0^t \cdots \int_0^t k(t, \tau_1, \cdots \tau_l) u(\tau_1, \cdots, \tau_l) d\tau_1 \cdots d\tau_l$$

exist, and are invariant with respect to interchanges of orders of integration, for $t \geq 0$, *and p defined by*

$$p(t) = \int_0^t \cdots \int_0^t h(t, \tau_1, \cdots, \tau_l) u(\tau_1, \cdots, \tau_l) d\tau_1 \cdots d\tau_l, \qquad t \geq 0$$

is an element of $L_\infty(\mathrm{C})$.

(*iv*) *We have*

$$\int_0^t s(t,\tau)\int_0^\tau \cdots \int_0^\tau h(\tau,\tau_1,\cdots,\tau_l)u(\tau_1,\cdots,\tau_l)d\tau_1\cdots d\tau_l d\tau$$
$$= \int_0^t \cdots \int_0^t k(t,\tau_1,\cdots,\tau_l)u(\tau_1,\cdots,\tau_l)d\tau_1\cdots d\tau_l, \qquad t \geq 0.$$

Comment

The condition on s of part (*ii*) of Lemma 4 is met if s is continuous on $R_0(1)$, or if $s(t,\tau)$ depends only on the difference $(t-\tau)$ (see Ref. 11, page 12).

Definition: If r and s are two complex column n-vectors, then rs denotes the column n-vector defined by $(rs)_i = r_i s_i$ for $i = 1, 2, \cdots, n$.

Lemma 5: *If* $h \in S_1^{(p)}$ *and* $k \in S_1^{(q)}$, *then the function* s, *defined on* $R_0(p+q)$ *by*

$$s(t,\tau_1,\cdots,\tau_{p+q}) = h(t,\tau_1,\cdots,\tau_p)k(t,\tau_{p+1},\cdots,\tau_{p+q})$$

for $(t,\tau_1,\cdots,\tau_{p+q}) \in R_0(p+q)$, *belongs to* $S_1^{(p+q)}$.

Lemma 6: *If* $\lambda \in S_n^{(1)}$, *then for each* $p \in H(\mathrm{C})$ *there is a unique* $q \in H(\mathrm{C})$ *such that*

$$p(t) = q(t) - \int_0^t \lambda(t,\tau)q(\tau)d\tau, \qquad t \geq 0. \tag{20}$$

In Lemma 7, below, we refer to the following two hypotheses.

C.1: $\lambda \in S_n^{(1)}$, and Λ denotes the map of $L_\infty(\mathrm{C})$ into itself defined by

$$(\Lambda p)(t) = \int_0^t \lambda(t,\tau)p(\tau)d\tau, \qquad t \geq 0$$

for $p \in L_\infty(\mathrm{C})$.

C.2: $\lambda \in S_n^{(1)}$, and, for each $p \in L_\infty(\mathrm{C})$, the unique element q of $H(\mathrm{C})$ such that

$$p(t) = q(t) - \int_0^t \lambda(t,\tau)q(\tau)d\tau, \qquad t \geq 0$$

satisfies the condition that $q \in L_\infty(\mathrm{C})$.

Lemma 7: *Suppose that C.*1 *and C.*2 *hold. Then* $(I-\Lambda)$ *is an invertible map* of* $L_\infty(\mathrm{C})$ *onto itself, and there is a* $\kappa \in S_n^{(1)}$ *such that*

* Here, as in Section 2.3, I denotes the identity transformation on $L_\infty(\mathrm{C})$.

$$(I - \Lambda)^{-1}p(t) = p(t) - \int_0^t \kappa(t, \tau)p(\tau)d\tau, \qquad t \geq 0$$

for every $p \in L_\infty(\mathrm{C})$, *and such that if* λ *meets the conditions imposed on* s *of part* (*ii*) *of Lemma* 4, *then so does* κ.

This concludes our statement of the lemmas that we shall use. As mentioned at the beginning of this section, Lemmas 3 through 7 are proved in the appendix.

It is clear that (under the hypotheses of Theorem 2) A.1 is met, Lemma 3 shows that A.2 is satisfied, and, by Lemmas 3 and 6, as well as the observation that C.1 together with C.2 imply that $(I - \Lambda)^{-1}$ exists, we see that A.3 also is satisfied. Therefore, the hypotheses of Theorem 1 are met.

With v, w, and δ as in part (*ii*) of Theorem 1,

$$w = Dv + \sum_{m=1}^{\infty} B[g_m(Av)]_2$$

for $\|v\| < \delta$, where the $[g_m(Av)]_2$ are defined by (5) through (9), which involve associated functions $[g_m(Av)]_1$.

For each positive integer p, let H_p denote the hypothesis that we have

$$[g_m(Av)]_1(t) = \int_0^t \cdots \int_0^t q_m(t, \tau_1, \cdots \tau_m) \cdot \chi[v(\tau_1), \cdots, v(\tau_m)]d\tau_1 \cdots d\tau_m,$$

and

$$[g_m(Av)]_2(t) = L(t)[g_m(Av)]_1(t) + \sum_{l=2}^{m} (l!)^{-1} \sum_{\substack{k_1+k_2+\cdots+k_l=m \\ k_j>0}} B_l(t) \int_0^t \cdots \int_0^t r_{k_1,\ldots,k_l}(t, \tau_1, \cdots, \tau_m) \cdot \chi[v(\tau_1), \cdots, v(\tau_m)]d\tau_1 \cdots d\tau_m$$

for $t \geq 0$, $\|v\| < \delta$, and $m = 1, 2, \cdots, p$, in which

(*i*) by the sum over l when $m = 1$ is meant the zero n-vector,

(*ii*) each q_m belongs to $S^{(m)}_{n^{(m)}}$,

(*iii*) L is the $n \times n$ matrix-valued function described in Proposition 1,

(*iv*) for $l \geq 2$, the B_l are bounded* measurable $n \times n^l$ matrix-valued functions over $[0, \infty)$,

(*v*) for $m \geq 2$, the $r_{k_1,\ldots,k_l}$ are $n^l \times n^m$ matrix-valued functions

* By B_l bounded is meant that its elements are bounded.

defined on $R_0(m)$ such that each $(r_{k_1,\dots,k_l})_{ij} \in S_1^{(m)}$ with $n = 1$, and

(*vi*) the q_m, and the $r_{k_1,\dots,k_l}$ for $m \geq 2$, are continuous on $R_0(m)$ when a is continuous on $R_0(1)$ and c meets the conditions on s of part (*ii*) of Lemma 4.

By Lemma 3, L is bounded and measurable. Using (5) and (6) as well as Lemmas 3, 4, and 7, we see that H_1 is met. [Notice that s given by $s(t, \tau) = u(t, \tau)v(\tau)$ meets the condition of part (*ii*) of Lemma 4 when $u \in S_n^{(1)}$, u meets the condition, and v is a bounded measurable $n \times n$ matrix-valued function on $[0, \infty)$.] Thus, by Lemma 4, there is a $k_1 \in S_n^{(1)}$ such that

$$Dv(t) + B[g_1(Av)]_2(t) = \int_0^t k_1(t, \tau)v(\tau)d\tau, \qquad t \geq 0$$

for $\|v\| < \delta$, and k_1 is continuous under the conditions on a, b, c, and d of part (*iii*) of Theorem 2.

By Lemma 4 (which holds for any n), it easily follows that if H_p is met for some $p \geq 2$ then there is a $k_p \in S_{n^p}^{(p)}$ such that

$$B[g_p(Av)]_2(t) = \int_0^t \cdots \int_0^t k_p(t, \tau_1, \cdots, \tau_p) \cdot \chi[v(\tau_1), \cdots, v(\tau_p)]d\tau_1 \cdots d\tau_p, \qquad t \geq 0$$

for $\|v\| < \delta$, and such that k_p meets the continuity requirement of part (*iii*) of the theorem. Therefore, to complete the proof of the theorem it suffices to show that H_p is met for every p. For this purpose, suppose that H_p is satisfied for some p. Using (7), we have

$$h_{(p+1)} = \sum_{l=2}^{(p+1)} (l!)^{-1} \sum_{\substack{k_1+k_2+\cdots+k_l=(p+1) \\ k_j>0}} d^l N(\theta) \cdot [g_{k_1}(Av)]_1[g_{k_2}(Av)]_1 \cdots [g_{k_l}(Av)]_1, \qquad \|v\| < \delta.$$

Now let l be a fixed integer such that $2 \leq l \leq (p + 1)$, and let $k_1, \cdots, k_l$ be positive integers such that $k_1 + k_2 + \cdots + k_l = p + 1$. Using Lemma 3,

$$\{d^l N(\theta)[g_{k_1}(Av)]_1[g_{k_2}(Av)]_1 \cdots [g_{k_l}(Av)]_1(t)\}_i$$

$$= \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} \cdots \sum_{j_l=1}^{n} b_i(t, j_1, \cdots, j_l) \left[\int_0^t \cdots \int_0^t q_{k_1}(t, \tau_1, \cdots, \tau_{k_1})\chi \cdot [v(\tau_1), \cdots, v(\tau_{k_1})]d\tau_1 \cdots d\tau_{k_1} \right]_{j_1} \cdots \left[\int_0^t \cdots \int_0^t q_{k_l}(t, \tau_1, \cdots \tau_{k_l}) \cdot \chi[v(\tau_1), \cdots, v(\tau_{k_l})]d\tau_1 \cdots d\tau_{k_l} \right]_{j_l}$$

for $t \geq 0$, $\|v\| < \delta$, and each i, in which the $b_i(\cdot, j_1, \cdots, j_l)$ are bounded and measurable. By Lemmas 4 and 5, we see that

$$d^l N(\theta)[g_{k_1}(Av)]_1 \cdots [g_{k_l}(Av)]_1(t)$$

$$= B_l(t) \int_0^t \cdots \int_0^t r_{k_1,\ldots,k_l}(t, \tau_1, \cdots, \tau_{(p+1)})$$

$$\cdot \chi[v(\tau_1), \cdots, v(\tau_{(p+1)})] d\tau_1 \cdots \tau_{(p+1)}, \qquad t \geq 0$$

for $\|v\| < \delta$ and for some B_l and $r_{k_1,\ldots,k_l}$ of the type required. [Here we have used the observations that a product of integrals

$$\int_0^t \cdots \int_0^t q_{k_1}(t, \tau_1, \cdots, \tau_{k_1})_{j_1 l_1} \{\chi[v(\tau_1), \cdots, v(\tau_{k_1})]\}_{l_1}$$

$$\cdot d\tau_1 \cdots d\tau_{k_1} \cdots \int_0^t \cdots \int_0^t q_{k_l}(t, \tau_1, \cdots, \tau_{k_l})_{j_l l_l}$$

$$\cdot \{\chi[v(\tau_1), \cdots, v(\tau_{k_l})]\}_{l_l} d\tau_1 \cdots d\tau_{k_l},$$

in which l_j is drawn from $\{1, 2, \cdots, n^{k_j}\}$ for each j, can be written as the iterated integral

$$\int_0^t \cdots \int_0^t q_{k_1}(t, \tau_1, \cdots, \tau_{k_1})_{j_1 l_1} \cdots q_{k_l}(t, \tau_{(k_1+\ldots+k_{l-1}+1)}, \cdots,$$

$$\cdot \tau_{(p+1)})_{j_l l_l} \{\chi[v(\tau_1), \cdots, v(\tau_{k_1})]\}_{l_1} \cdots$$

$$\cdot \{\chi[v(\tau_{(k_1+\ldots+k_{l-1}+1)}), \cdots, v(\tau_{(p+1)})]\}_{l_l} d\tau_1 \cdots d\tau_{(p+1)},$$

and that r, given by $r(t, \tau_1, \cdots, \tau_{p+1}) = q_{k_1}(t, \tau_1, \cdots, \tau_{k_1}) \cdot {}_{j_1 l_1} \cdots q_{k_l}(t, \tau_{(k_1+\ldots+k_{l-1}+1)}, \cdots, \tau_{p+1})_{j_l l_l}$ on $R_0(p+1)$, is continuous when each q_{k_j} is continuous on $R_0(k_j)$.]

Finally, using (8) and (9), and Lemmas 4 and 7, we observe that $H_{(p+1)}$ is satisfied, showing that H_p is met for all p. This completes the proof.*

APPENDIX A

Proof of Lemma 2

Assume that $p \in X$ is given.

Let Q denote the linear map from $\mathscr{B}$ into the space $L(\mathscr{B}, \mathscr{B})$ of bounded linear operators from $\mathscr{B}$ into $\mathscr{B}$, given by

---

* Our proof shows also that the theorem holds if B.1 and B.2 are modified to the extent that an arbitrary constant (scalar or $n \times n$-matrix) multiple of the identity map in $H(C)$ is added to B, and $\eta(\cdot, t)$ is required to be independent of t.

$$Qr = \begin{bmatrix} -C_\infty d^2N(p_1)r_1 & 0 \\ d^2N(p_1)r_1 & 0 \end{bmatrix} \tag{21}$$

for any $r \in \mathscr{B}$. Since $\sup\{\|Qh_1h_2\|: h_1, h_2 \in \mathscr{B} \text{ with } \|h_1\| = \|h_2\| = 1\}$ is finite, it follows that Q is bounded.

Let $h \in \mathscr{B}$ be such that $(p + h) \in X$. Observe that, using (10) and (21),

$$\|df(p + h) - df(p) - Qh\| = \sup\{\|df(p + h)h_1 - df(p)h_1 - Qhh_1\|: h_1 \in B, \|h_1\| = 1\} = o(\|h\|),$$

which shows that $d^2f(p)$ exists, that $d^2f(p) = Q$, and hence that the expression for $d^2f(p)h_1h_2$ given in the lemma is valid.

Now suppose that for some $l \geq 2$, $d^lf(p)$ exists and that it satisfies (12). Met $\tilde{M}$ denote the continuous multilinear mapping of $\mathscr{B}^{(l+1)}$ into $\mathscr{B}$ given by

$$\tilde{M}(q_1, q_2, \cdots, q_{(l+1)}) = \begin{bmatrix} -C_\infty d^{(l+1)}N(p_1)q_{11}q_{21} \cdots q_{(l+1)1} \\ d^{(l+1)}N(p_1)q_{11}q_{21} \cdots q_{(l+1)1} \end{bmatrix}$$

for $q_1, q_2, \cdots, q_{(l+1)}$ belonging to $\mathscr{B}$. We shall use M to denote the usual associate (Ref. 10, page 318) of $\tilde{M}$ that belongs to $L(\mathscr{B}, L(\mathscr{B}, \cdots, L(\mathscr{B}, \mathscr{B})\cdots))$ with $(l + 1)$ L's, in which $L(A_1, A_2)$ stands for the set of continuous linear operators from the Banach space A_1 into the Banach space A_2.*

Using the fact that

$$\|d^lf(p + h) - d^lf(p) - Mh\| = \sup\{\|d^lf(p + h) \cdot h_1h_2 \cdots h_l - d^lf(p)h_1h_2 \cdots h_l - Mhh_1h_2 \cdots h_l\|: \|h_1\| = \|h_2\| = \cdots = \|h_l\| = 1\}$$

for $(p + h) \in X$, as well as the boundedness of C_∞, we find that $\|d^lf(p + h) - d^lf(p) - Mh\| = o(\|h\|)$ as $\|h\| \to 0$, which shows that $d^{(l+1)}f(p)$ exists and equals M. This proves the lemma.

APPENDIX B

Proof of Lemma 3

For each t, (*iii*) implies (Ref. 8, pages 204, 205, 226, 227, 230) the existence throughout Γ_0 of the F-derivatives of all orders of the map $\eta(\cdot, t)$: $\Gamma_0 \subset C^n \to C^n$. In particular, each partial derivative (17) exists in Γ_0 for any $t \geq 0$.†

---

* For example, if $l = 2$, $L(\mathscr{B}, L(\mathscr{B}, \cdots, L(\mathscr{B}, \mathscr{B})\cdots)) = L(\mathscr{B}, L(\mathscr{B}, L(\mathscr{B}, \mathscr{B})))$.

† See Section 8.9 of Ref. 8.

Given any $s \in \Gamma$,

$$\eta[s(t), t] = \int_0^1 d\eta[\beta s(t), t] d\beta \cdot s(t), \qquad t \geq 0$$

in which $d\eta[\beta s(t), t]$ is the F-derivative of $\eta(\cdot, t)$ at the point $\beta s(t)$ (i.e., $d\eta[\beta s(t), t]$ is the $n \times n$ matrix whose ijth element is $\partial \eta_i(z, t)/\partial z_j$ evaluated at $z = \beta s(t)$). By (iv), the elements of $d\eta[\beta s(t), t]$ are bounded on $(\beta, t) \in [0, 1] \times [0, \infty)$. Thus, using (ii), N maps Γ into $L_\infty(\mathrm{C})$.

Similarly, for any $s \in \Gamma$ and any $h \in L_\infty(\mathrm{C})$ such that $(s + h) \in \Gamma$,

$$\eta[s(t) + h(t), t] - \eta[s(t), t] - d\eta[s(t), t]h(t) = \int_0^1 \{d\eta[\beta(s(t) + h(t)) + (1 - \beta)s(t), t] - d\eta[s(t), t]\} d\beta \cdot h(t), \qquad t \geq 0.$$

This, together with the continuity described in (iv), yields

$$\sup_{t \geq 0} |\eta[s(t) + h(t), t] - \eta[s(t), t] - d\eta[s(t), t]h(t)| = o(\|h\|) \tag{22}$$

as $\|h\| \to 0$. Since the pointwise limit function of a sequence of (Lebesgue) measurable functions is measurable, and, for each $i = 1, 2, \cdots, n$, (22) holds with $h(t) = \sigma u(i)$ for $t \geq 0$, in which σ is a scalar and $u(i)$ is the element of C^n with $u(i)_i = 1$ and $u(i)_j = 0$ for $i \neq j$, it easily follows that the elements of $d\eta[s(\cdot), \cdot]$ are measurable on $[0, \infty)$. By (iv) these elements are bounded. Thus, using (22), $dN(s)$ exists and

$$[dN(s)h(t)]_i = \sum_{j_1=1}^{n} \frac{\partial \eta_i[s_1(t), \cdots, s_n(t), t]}{\partial z_{j_1}} h_{j_1}(t), \qquad t \geq 0$$

for each i. This shows that the $m = 1$ part of the lemma is true.

Now assume that the assertions of the lemma are true for $1 \leq m \leq l$, and again let $s \in \Gamma$ be given, and let $h \in L_\infty(\mathrm{C})$ satisfy $(s + h) \in \Gamma$.

By (iv), each

$$\frac{\partial^{(l+1)} \eta_i[s_1(\cdot), \cdots, s_n(\cdot), \cdot]}{\partial z_{j_{(l+1)}} \cdots \partial z_{j_1}} \tag{23}$$

is bounded on $[0, \infty)$. To see that each is measurable, observe that for $h_1, h_2, \cdots, h_l$ belonging to $L_\infty(\mathrm{C})$,

$$\sup_{t \geq 0} \max_i \left| \sum_{j_1=1}^{n} \cdots \sum_{j_l=1}^{n} \frac{\partial^l \eta_i[s_1(t) + h_1(t), \cdots, s_n(t) + h_n(t), t]}{\partial z_{j_l} \cdots \partial z_{j_1}} h_{1j_1}(t) \cdots h_{lj_l}(t) - \sum_{j_1=1}^{n} \cdots \sum_{j_l=1}^{n} \frac{\partial^l \eta_i[s_1(t), \cdots, s_n(t), t]}{\partial z_{j_l} \cdots \partial z_{j_1}} h_{1j_1}(t) \right.$$

$$\cdots h_{lj_l}(t) - \sum_{j_1=1}^{n} \cdots \sum_{j_{(l+1)}=1}^{n} \frac{\partial^{(l+1)}\eta_i[s_1(t), \cdots, s_n(t), t]}{\partial z_{j_{(l+1)}} \cdots \partial z_{j_1}} h_{1j_1}(t)$$

$$\cdots h_{lj_l}(t) h_{j_{(l+1)}}(t) \| \leq o(\|h\|) \prod_{j=1}^{l} \|h_j\|, \tag{24}$$

which is a consequence of (*iv*) and the relation

$$\frac{\partial^l \eta_i[s_1(t) + h_1(t), \cdots, s_n(t) + h_n(t), t]}{\partial z_{j_l} \cdots \partial z_{j_1}} - \frac{\partial^l \eta_i[s_1(t), \cdots, s_n(t), t]}{\partial z_{j_l} \cdots \partial z_{j_1}}$$

$$= \sum_{j_{(l+1)}=1}^{n} \frac{\partial^{(l+1)}\eta_i[s_1(t), \cdots, s_n(t), t]}{\partial z_{j_{(l+1)}} \partial z_{j_l} \cdots \partial z_{j_1}} h_{j_{(l+1)}}(t)$$

$$+ \int_0^1 \left\{ \sum_{j_{(l+1)}=1}^{n} \frac{\partial^{(l+1)}\eta_i[s_1(t) + \beta h_1(t), \cdots, s_n(t) + \beta h_n(t), t]}{\partial z_{j_{(l+1)}} \partial z_{j_l} \cdots \partial z_{j_1}} h_{j_{(l+1)}}(t) \right.$$

$$\left. - \sum_{j_{(l+1)}=1}^{n} \frac{\partial^{(l+1)}\eta_i[s_1(t), \cdots, s_n(t), t]}{\partial z_{j_{(l+1)}} \partial z_{j_l} \cdots \partial z_{j_1}} h_{j_{(l+1)}}(t) \right\} d\beta, \qquad t \geq 0.$$

It easily follows from (24) that each function (23) is the pointwise limit of a sequence of measurable functions, and is therefore measurable.

In particular, $\tilde{Q}_l(s)$ defined by

$$[\tilde{Q}_l(s)(p_1, \cdots, p_{(l+1)})(t)]_i$$

$$= \sum_{j_{(l+1)}=1}^{n} \sum_{j_l=1}^{n} \cdots \sum_{j_1=1}^{n} \frac{\partial^{(l+1)}\eta_i[s_1(t), \cdots, s_n(t), t]}{\partial z_{j_{(l+1)}} \partial z_{j_{(l)}} \cdots \partial z_{j_1}}$$

$$\cdot p_{1j_1}(t) p_{2j_2}(t) \cdots p_{(l+1)j_{(l+1)}}(t), \qquad t \geq 0$$

for $p_1, p_2, \cdots, p_{(l+1)}$ in $L_\infty(\mathrm{C})$ and $i = 1, 2, \cdots, n$, is a continuous multilinear mapping of $L_\infty(\mathrm{C})^{(l+1)}$ into $L_\infty(\mathrm{C})$.

Proceeding as in the proof of Lemma 2, let $Q_l(s)$ denote the usual associate of $\tilde{Q}_l(s)$ that belongs to $L(L_\infty(\mathrm{C}), L(L_\infty(\mathrm{C}), \cdots, L(L_\infty(\mathrm{C}), L_\infty(\mathrm{C})) \cdots))$ with $(l+1)$ L's, in which $L(A_1, A_2)$ stands for the set of continuous linear operators from the Banach space A_1 into the Banach space A_2. Using $\|d^l N(s+h) - d^l N(s) - Q_l(s)h\| = \sup\{\|d^l N(s+h)h_1 \cdots h_l - d^l N(s)h_1 \cdots h_l - Q_l(s)hh_1 \cdots h_l\|: \|h_1\| = \|h_2\| = \cdots = \|h_l\| = 1\}$, as well as our induction hypothesis and (24), we see that $\|d^l N(s+h) - d^l N(s) - Q_l(s)h\| = o(\|h\|)$ as $\|h\| \to 0$. Therefore $d^{(l+1)}N(s)$ exists and is equal to $Q_l(s)$. This completes the proof.

APPENDIX C

Proof of Lemmas 4 and 5

It suffices to prove the lemmas for $n = 1$ and $u(\tau_1, \cdots, \tau_l) = 1$ for $(\tau_1, \cdots, \tau_l) \in [0, \infty)^l$, and attention is now restricted to that case.

For $t > 0$, one has

$$\int_{[0,t]^l} \left| \int_0^t s(t, \tau)\tilde{h}(\tau, \tau_1, \cdots, \tau_l)d\tau \right| d(\tau_1, \cdots, d\tau_l)$$

$$\leq \int_{[0,t]^l} \int_0^t |s(t, \tau)| \cdot |\tilde{h}(\tau, \tau_1, \cdots, \tau_l)| d\tau d(\tau_1, \cdots, \tau_l)$$

$$\leq \int_0^t \int_{[0,\ t]^l} |\tilde{h}(\tau, \tau_1, \cdots, \tau_l)| d(\tau_1, \cdots, \tau_l)| s(t, \tau)| d\tau$$

$$\leq \sup_{\tau \geq 0} \int_{[0,\tau]^l} |h(\tau, \tau_1, \cdots, \tau_l)| d(\tau_1, \cdots, \tau_l) \cdot \sup_{t \geq 0} \int_0^t |s(t, \tau)| d\tau,$$

in which the measurability of

$$\int_0^t s(t, \tau)\tilde{h}(\tau, \tau_1, \cdots, \tau_l)d\tau \tag{25}$$

in $(\tau, \cdots, \tau_l)$, and the justification for the interchange of the order of integration, follow from theorems of Fubini and Tonelli (Ref. 12, pages 137–145). The measurability of (25) in $(t, \tau_1, \cdots, \tau_l)$ is also a consequence of these theorems.* Thus, since it is clear that k is bounded, *(i)* holds.

Now let h and s satisfy the conditions of part *(ii)*. Let $(t, \tau_1, \cdots, \tau_l) \in R_0(l)$ be given, let $\alpha, \alpha_1, \cdots, \alpha_l$ be real variables such that $(t + \alpha, \tau_1 + \alpha_1, \cdots, \tau_l + \alpha_l) \in R_0(l)$, and notice that

$$k(t + \alpha, \tau_1 + \alpha_1, \cdots, \tau_l + \alpha_l) - k(t, \tau_1, \cdots, \tau_l)$$

$$= \int_0^\infty [\tilde{s}(t + \alpha, \tau) - \tilde{s}(t, \tau)]\tilde{h}(\tau, \tau_1 + \alpha_1, \cdots, \tau_l + \alpha_l)d\tau$$

$$+ \int_0^\infty \tilde{s}(t, \tau)[\tilde{h}(\tau, \tau_1 + \alpha_1, \cdots, \tau_l + \alpha_l) - \tilde{h}(\tau, \tau_1, \cdots, \tau_l)]d\tau. \tag{26}$$

Using the hypothesis of part *(ii)* concerning s, the boundedness of h and s, and the uniform continuity of h on compact subsets of $R_0(l)$, we see that each integral in (26) approaches zero as $(t + \alpha, \tau_1 + \alpha_1, \cdots, \tau_l + \alpha_l) \to (t, \tau_1, \cdots, \tau_l)$, showing that *(ii)* is true.

Straightforward modifications of the proof of part *(iii)* of Lemma 3 in Ref. 1 establish that *(iii)* here holds.

With regard to part *(iv)*, using the theorems of Fubini and Tonelli cited above, and the proposition that a bounded measurable function

---

* Consider, for arbitrary finite $T > 0$, the existence and iterated-integral representations of the multiple integral $\int_{[0,\ T]^{(l+2)}} \tilde{s}(t, \tau)\tilde{h}(\tau, \tau_1, \cdots, \tau_l)d(t, \tau_1, \cdots, \tau_l, \tau)$.

on a set E of finite measure is summable over E, we have

$$\int_0^t s(t,\tau)\int_0^\tau \cdots \int_0^\tau h(\tau,\tau_1,\cdots,\tau_l)d\tau_1\cdots d\tau_l d\tau$$

$$= \int_0^t s(t,\tau)\int_{[0,\tau]^l} h(\tau,\tau_1,\cdots,\tau_l)d(\tau_1,\cdots,\tau_l)d\tau$$

$$= \int_0^t s(t,\tau)\int_{[0,t]^l} \tilde{h}(\tau,\tau_1,\cdots,\tau_l)d(\tau_1,\cdots,\tau_l)d\tau$$

$$= \int_{[0,t]^l}\int_0^t s(t,\tau)\tilde{h}(\tau,\tau_1,\cdots,\tau_l)d\tau d(\tau_1,\cdots,\tau_l)$$

$$= \int_0^t \cdots \int_0^t k(t,\tau_1,\ldots,\tau_l)d\tau_1\cdots d\tau_l$$

for $t \geq 0$, which establishes (*iv*) and completes the proof of Lemma 4.

Under the hypothesis of Lemma 5,

$$\int_{[0,t]^{(p+q)}} |h(t,\tau_1,\cdots,\tau_p)k(t,\tau_{(p+1)},\cdots,\tau_{(p+q)})|d(\tau_1,\cdots,\tau_{(p+q)})$$

$$\leq \sup_{t\geq 0}\int_{[0,t]^p} |h(t,\tau_1,\cdots,\tau_p)|d(\tau_1,\cdots,\tau_p)$$

$$\times \sup_{t\geq 0}\int_{[0,t]^q} |k(t,\tau_{(p+1)},\cdots,\tau_{(p+q)})|d(\tau_{(p+1)},\cdots,\tau_{(p+q)})$$

for every $t \geq 0$, which proves the lemma.

APPENDIX D

Proof of Lemma 6

By the proof of Theorems 2.3 and 2.5 of Ref. 13, there exists a measurable function κ from $R_0(1)$ into the set of complex $n \times n$ matrices such that the elements of κ are bounded on bounded subsets of $R_0(1)$, and κ satisfies the resolvent equations

$$\kappa(t,\tau) + \lambda(t,\tau) = \int_\tau^t \lambda(t,u)\kappa(u,\tau)du \tag{27}$$

$$\kappa(t,\tau) + \lambda(t,\tau) = \int_\tau^t \kappa(t,u)\lambda(u,\tau)du \tag{28}$$

for $t \geq \tau \geq 0$.

For each $p \in H(C)$, the function q defined on $[0,\infty)$ by

$$q(t) = p(t) - \int_0^t \kappa(t, \tau)p(\tau)d\tau, \qquad t \geq 0 \tag{29}$$

belongs to $H(\mathrm{C})$, and, using (27) as well as theorems of Fubini and Tonelli (Ref. 12, pages 137–145) to justify an interchange of order of integration, it is simple matter to show that q given by (29) satisfies (20) for each $p \in H(\mathrm{C})$. Similarly, it is essentially well known that (28) can be used to show that if there is a $q \in H(\mathrm{C})$ that satisfies (20) for a given $p \in H(\mathrm{C})$, then q satisfies (29), which completes the proof.

APPENDIX E

Proof of Lemma 7

By Lemma 6 and its proof, $(I - \Lambda)$ is an invertible map of $L_\infty(\mathrm{C})$ onto $L_\infty(\mathrm{C})$, and there is a measurable matrix-valued κ, defined on $R_0(1)$ such that the elements of κ are bounded on bounded subsets of $R_0(1)$, with the property that (28) is satisfied and

$$(I - \Lambda)^{-1}p(t) = p(t) - \int_0^t \kappa(t, \tau)p(\tau)d\tau, \qquad t \geq 0$$

for each $p \in L_\infty(\mathrm{C})$. Since $(I - \Lambda)^{-1}$ maps $L_\infty(\mathrm{C})$ into itself, it follows (Refs. 14 and 15) that each κ_{ij} satisfies

$$\sup_{t \geq 0} \int_0^t |\kappa_{ij}(t, \tau)|d\tau < \infty. \tag{30}$$

Using (28), (30), and the boundedness of λ, we see that κ is bounded on $R_0(l)$. Therefore, $\kappa \in S_n^{(1)}$.

Assume now that λ satisfies the condition on s of part (ii) of Lemma 4, recall that κ satisfies (27), and let r be defined by

$$r(t, \tau) = \int_0^t \lambda(t, u)\tilde{\kappa}(u, \tau)du$$

for $t \geq \tau \geq 0$.*

Let $t \geq 0$ be given. For arbitrary i and j, let

$$\Delta_{ij}(\alpha, t) = \int_0^\infty |\tilde{r}_{ij}(t + \alpha, \tau) - \tilde{r}_{ij}(t, \tau)|d\tau$$

for $(t + \alpha) \geq 0$ (see the definition preceding Lemma 4 for the meaning of $\tilde{r}$; r belongs to $S_n^{(1)}$ because κ and λ do and (27) is met). Notice that to complete the proof of our lemma, it suffices to show that $\Delta_{ij}(\alpha, t) \to 0$ as $\alpha \to 0$.

* With regard to the meaning of $\tilde{\kappa}$, see the definition immediately preceding Lemma 4.

It is clear that $\Delta_{ij}(\alpha, 0) \to 0$ as $\alpha \to 0$. Assume now that $t > 0$, and let α be such that $t - |\alpha| > 0$. We have

$$\Delta_{ij}(\alpha, t) = \int_0^{t-|\alpha|} |r_{ij}(t + \alpha, \tau) - r_{ij}(t, \tau)| d\tau + \int_{t-|\alpha|}^{t+|\alpha|} |\tilde{r}_{ij}(t + \alpha, \tau) - \tilde{r}_{ij}(t, \tau)| d\tau, \quad (31)$$

in which (by the boundedness of r) the second integral goes to zero as $\alpha \to 0$. Further,

$$\int_0^{t-|\alpha|} |r_{ij}(t + \alpha, \tau) - r_{ij}(t, \tau)| d\tau$$

$$= \int_0^{t-|\alpha|} \left| \sum_{k=1}^{n} \int_0^{t+\alpha} \lambda_{ik}(t + \alpha, u)\tilde{\kappa}_{kj}(u, \tau) du - \sum_{k=1}^{n} \int_0^{t} \lambda_{ik}(t, u)\, \tilde{\kappa}_{kj}(u,\tau) du \right| d\tau$$

$$\leq \sum_{k=1}^{n} \int_0^{t-|\alpha|} \int_0^{\infty} |\tilde{\lambda}_{ik}(t + \alpha, u) - \tilde{\lambda}_{ik}(t, u)| \cdot |\tilde{\kappa}_{kj}(u, \tau)| du d\tau$$

which, using the boundedness of the κ_{kj}, shows that the first integral on the right side of (31) also approaches zero as $\alpha \to 0$.

APPENDIX F

Volterra Expansions on a Finite Time Interval

In this appendix, T denotes an arbitrary positive constant, L_∞ (C) (T) stands for the complex Banach space of measurable complex column n-vector-valued functions v defined on $[0, T]$ such that the jth component v_j of v satisfies $\sup_{t\in[0,T]} |v_j(t)| < \infty$ for $j = 1, 2, \cdots, n$, and where the norm $\|\cdot\|_T$ on $L_\infty(\mathrm{C})(T)$ is given by $\|v\|_T = \max_j \sup_t |v_j(t)|$, and for each $l = 1, 2, \cdots, R_0(l)(T)$ denotes the subset of $\mathrm{R}^{(l+1)}$ given by $R_0(l)(T) = \{(v_0, v_1, \cdots, v_l) \in \mathrm{R}^{(l+1)}: T \geq v_0 \geq v_i \geq 0$ for $i = 1, 2, \cdots, l\}$.

Similarly, for any positive integers q and $l, S_q^{(l)}(T)$ denotes the set of complex $n \times q$ matrix-valued functions h defined on $R_0(l)(T)$ such that each h_{ij} is Lebesgue measurable and bounded on $R_0(l)(T)$.

We shall refer to the following two hypotheses.

D.1: There are elements a, b, c, and d of $S_n^{(1)}(T)$ such that for each $p \in L_\infty(\mathrm{C})(T)$,

$$(Ap)(t) = \int_0^t a(t, \tau)p(\tau)d\tau$$

$$(Bp)(t) = \int_0^t b(t, \tau)p(\tau)d\tau$$

$$(Cp)(t) = \int_0^t c(t, \tau)p(\tau)d\tau$$

$$(Dp)(t) = \int_0^t d(t, \tau)p(\tau)d\tau$$

for $t \in [0, T]$.

D.2: With γ, Γ_0, C^n, and θ_c as indicated in the paragraph preceding B.2 of Section 2.4, N is defined on $\Gamma = \{s \in L_\infty(C)(T): \|s\|_T < \gamma\}$ by

$$(Ns)(t) = \eta[s(t), t], \qquad t \in [0, T],$$

where η is a map from $\Gamma_0 \times [0, T]$ into C^n with the following properties:

(*i*) $\eta(\theta_c, t) = \theta_c$ for $t \in [0, T]$.

(*ii*) The function ξ given by $\xi(t) = \eta[s(t), t]$, $0 \le t \le T$, is Lebesgue measurable on $[0, T]$ for each $s \in \Gamma$.

(*iii*) For each $t \in [0, T]$, $\eta(\cdot, t)$ is a continuous map of Γ_0 into C^n, and for each $t \in [0, T]$, for $1 \le i, j \le n$, and for any point $\alpha \in \Gamma_0$, the function $z_j \mapsto \eta_i(\alpha_1, \cdots, \alpha_{j-1}, z_j, \alpha_{j+1}, \cdots, \alpha_n, t)$ is differentiable with respect to the complex variable z_j for $|z_j| < \gamma$. [This implies (Ref. 8, pages 204, 205, 226, 227, 230) the existence throughout Γ_0 of every mth order partial derivative

$$\frac{\partial^m \eta_i}{\partial z_{j_m} \partial z_{j_{m-1}} \cdots \partial z_{j1}} \tag{32}$$

for each t and all m.]

(*iv*) For any $m, j_1, \cdots, j_m$, and i, the partial derivative (32), which we denote by $p(z_1, \cdots, z_n, t)$, satisfies the conditions that the function $t \mapsto p(0, \cdots, 0, t)$ is bounded on $[0, T]$, and that p is uniformly continuous on closed subsets of Γ_0 uniformly in t, in the sense that given a closed $\Gamma_{00} \subset \Gamma_0$ and a $\delta_1 > 0$ there is a $\delta_2 > 0$ such that

$$|p(z_{a1}, \cdots, z_{an}, t) - p(z_{b1}, \cdots, z_{bn}, t)| \le \delta_1$$

for $t \in [0, T]$ whenever z_a and z_b are elements of Γ_{00} such that $|z_a - z_b| < \delta_2$.

Direct modifications of the proof in Section 2.4.2 suffice to establish the following result, in which by Proposition 2′ we mean the corollary of Proposition 2 obtained from Proposition 2 by replacing $S_{n'}^{(l)}$, $t \ge 0$,

$L_\infty(\mathrm{C})$, and $[0, \infty)$ by $S_{n^l}^{(l)}(T)$, $t \in [0, T]$, $L_\infty(\mathrm{C})(T)$, and $[0, T]$, respectively.*

Theorem 3: When D.1 and D.2 are met, there is a positive number δ and an open subset S of Γ of D.2 with the following properties.

(*i*) S contains the origin in $L_\infty(\mathrm{C})(T)$, and for each $v \in L_\infty(\mathrm{C})(T)$ with $\|v\|_T < \delta$, there exist unique x, y, and w of S, $L_\infty(\mathrm{C})(T)$, and $L_\infty(\mathrm{C})(T)$, respectively, such that (2), (3), and $y = Nx$ hold.

(*ii*) For each $l = 1, 2, \cdots$ there is a $k_l \in S_{n^l}^{(l)}(T)$ such that

$$w = \sum_{l=1}^{\infty} V_{k_l}(v) \quad \text{for} \quad \|v\|_T < \delta,$$

with the series uniformly convergent with respect to $\|v\|_T < \delta$, where $V_{k_l}(\cdot)$ is as indicated in Proposition 2′ (which is described just before Theorem 3).

(*iii*) Each k_l can be taken to be continuous on $R_0(l)(T)$ when a, b, c, and d are continuous on $R_0(1)(T)$.

REFERENCES

1. I. W. Sandberg, "Expansions for Nonlinear Systems," B.S.T.J., this issue. See also I. W. Sandberg, "Functional Expansions for Nonlinear Systems," Proceedings of the 1981 Int. Conf. on Digital Signal Processing, Florence, September 2-5, 1981.
2. V. Volterra, "Sopra le Funzioni che Dipendono da Altre Funzioni, Nota 1," Rend. Lincei Ser. 4, 3 (1887), pp. 97–105.
3. V. Volterra, *The Theory of Functionals and of Integral and Integro-differential Equations*, New York: Dover, 1959.
4. M. Fréchet, "Sur les Fonctionnelles Continues," Ann. de L'Ecole Normale sup., 3rd Series, *27* (1910).
5. I. W. Sandberg, "On the Properties of Some Systems that Distort Signals — II," B.S.T.J., *43*, No. 1 (January 1964), pp. 91–112.
6. I. W. Sandberg, "On the Theory of Linear Multi-loop feedback Systems," in *Feedback Systems*, ed. J. B. Cruz, Jr., New York: McGraw-Hill, 1972.
7. J. G. Truxal, *Automatic Feedback Control System Synthesis*, New York: McGraw-Hill, 1955, p. 114.
8. J. Dieudonné, *Foundations of Modern Analysis*, New York: Academic Press, 1969.
9. A. Halme, J. Orava, and H. Blomberg, "Polynomial Operators in Nonlinear System Theory," Int. J. Syst. Sci. *2*, No. 1 (1971), pp. 25–47.
10. T. M. Flett, *Differential Analysis*, London: Cambridge University Press, 1980.
11. N. Wiener, *The Fourier Integral and Certain of its Applications*, New York: Dover Publications, 1933.
12. S. McShane, *Integration*, Princeton: Princeton Univ. Press, 1944.
13. R. K. Miller, *Nonlinear Volterra Integral Equations*, Menlo Park, California: Benjamin Inc., 1971.
14. D. C. Youla, "On the Stability of Linear Systems," IEEE Trans. Circ. Th., *CT-10*, No. 2 (1963), pp. 276–9.
15. C. A. Desoer and A. J. Thomasian, "A Note on Zero-State Stability of Linear Systems," Proc. First Annual Allerton Conf. Circ. and Syst. Th. (1963), pp. 50–2.

* In this connection, see the last paragraph of Section 2.3.2. Also, a hypothesis corresponding to B.3 is not needed because when D.1 and D.2 hold, and $L(t)$ is as indicated in Proposition 1 for $t \in [0, T]$, for each $p \in L_\infty(\mathrm{C})(T)$ there exists a unique $q \in L_\infty(\mathrm{C})(T)$ such that $p(t) = q(t) - \int_0^t c(t, \tau)L(\tau)q(\tau)d\tau$, $0 \le t \le T$. On the other hand, δ and S of Theorem 3 depend on T.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 2, February 1982
Printed in U.S.A.

An Approximate Thermal Model For Outdoor Electronics Cabinets

By J. C. COYNE

(Manuscript received November 18, 1980)

Electronic systems are installed in outdoor loop plant in metal cabinets which are essentially unventilated. The current trend toward greater circuit miniaturization and higher power density increases the difficulty in the thermal design and points up the need for comprehensive thermal design guidelines. As a step toward that end, an approximate lumped thermal-conductance model is presented for calculating steady-state temperatures at three points of the cabinet (the hot sunny wall, the cool shaded wall, and the top row of circuit boards) as functions of input parameters of cabinet geometry, wind speed, solar radiation, and internal heat dissipation (assumed to be uniformly distributed). Calculated results are found to be in agreement with tests within about 10 percent.

I. INTRODUCTION

Electronic systems are typically installed in outdoor loop plant[1] in metal cabinets which are essentially unventilated. The cabinets are subject to ambient temperature excursions from −40 to 120°F and subject to solar heating which can raise the cabinet interior temperatures 30°F above ambient. With a thermal design limit of 185°F at circuit boards, the allowable temperature rise because of circuit dissipation is, thus, limited to about 35°F. The current trend toward greater circuit miniaturization and higher power density further aggravates an already difficult thermal design problem and points up the need for a comprehensive thermal analysis of the outdoor electronic cabinet. This paper is a step in that direction.

The circuit boards are typically arranged in outdoor cabinets (see Fig. 1) in much the same way as in conventional central office equipment bays and, to this extent, the heat transfer mechanisms are similar. In both cases, heat is removed from the circuit boards by

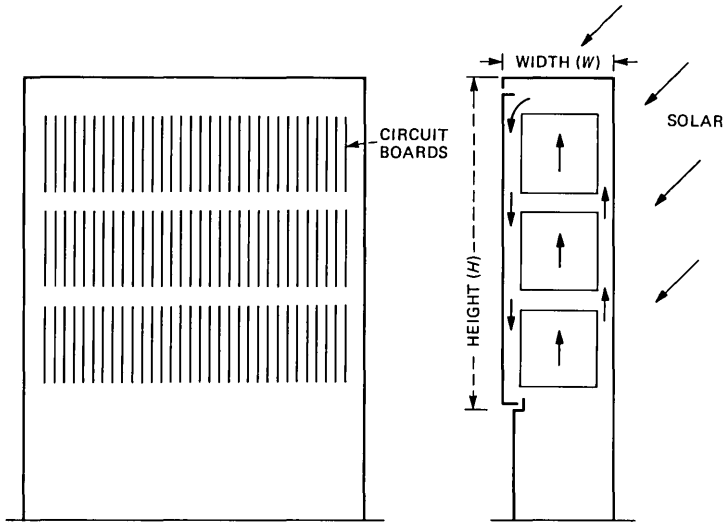


Fig. 1—Convective heat flow in an unventilated electronic cabinet, isolated on one wall.

natural convection; air is warmed by the circuitry and it expands and rises between the columns of boards. But unlike the central office bay situation, the hot air is confined inside the cabinet and must recirculate to transfer its heat to the cabinet walls. In addition, solar heat is absorbed on the sunny exterior cabinet surfaces, some of which is transferred to the cabinet interior and convected to the opposing shaded walls along with the internally dissipated heat. At the exterior surfaces of the cabinet, both solar and dissipated heat are convected and radiated to the ambient.

Whereas, the central office bay problem has been analyzed extensively, very little has been done to date on this more difficult cabinet problem. However, a problem closely related to it, which has received a great deal of study by Elder,[2] Eckert, and Carlson[3] and others, is that of a fluid-filled two-dimensional enclosure whose opposing walls are held at a uniform temperature difference as shown in Fig. 2. This enclosure problem is approximately the same as that of an empty (air-filled) unpowered cabinet in the sun. The main difference is that the walls of the electronic cabinet are not isothermal but, instead, increase in temperature vertically because of natural convection.

The core region of the cabinet of Fig. 1 and the enclosure of Fig. 2 both exhibit a large vertical temperature gradient. According to Elder,[2] the centerline temperature (midway between walls of the enclosure) approaches the hot wall temperature at the top and approaches the cold wall temperature at the bottom, with a gradient at the midway height given by half the difference in wall temperatures divided by the

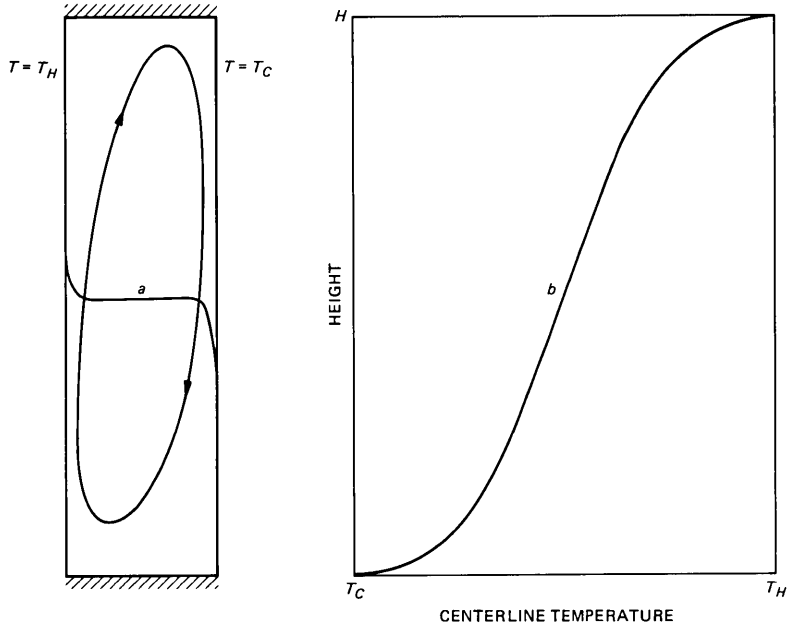


Fig. 2—Two-dimensional enclosure with isothermal side walls, heated at one side, showing flow path and temperature profiles: *a* is horizontal, *b* is vertical.

enclosure height. Laterally, the temperature is nearly constant except near the walls. The vertical gradient violates the isothermal assumption implicit in text-book, heat-transfer relationships for natural convection on vertical walls. As shown later, the error is appreciable.

The enclosure problem shown in Fig. 2 also approximates the convective heat transfer between heat-dissipating circuit boards and the walls of the powered electronic cabinet. The boards can be thought of as the "hot wall" of the enclosure analyzed in the literature. The main assumption here is that the boundary layer flow down the shaded cabinet wall and the heat transfer at this wall is the same regardless of whether the heat originates at an opposing hot wall or at the surface of circuit boards. Using this assumption, heat transfer coefficients from the literature will be incorporated into a simple lumped thermal conductance model from which approximate temperatures can be calculated for a powered electronic cabinet in the sun.

Several empirical relationships for the heat transfer of an enclosure depicted in Fig. 2 have been proposed in the recent literature. For conditions applicable to loop electronic cabinets (Rayleigh number based on height about 10^9, height-to-width ratio about 3, and Prandtl number equal to 0.7) they all are in essential agreement. For instance, Seki et al.[4] in a recent paper propose

$$\mathrm{Nu}_H = 0.36\ \mathrm{Pr}^{0.051}\left(\frac{H}{W}\right)^{-0.11}\mathrm{Ra}_H^{0.25}, \tag{1}$$

where Nu_H = Nusselt number based on enclosure height, Pr = Prandtl number, Ra_H = Rayleigh number based on enclosure height, H = enclosure height and W = enclosure width. After substitution of air properties at 120°F into the dimensionless numbers, Seki's relationship becomes

$$Q = 0.05 A_W W^{0.11} H^{-0.36} (T_H - T_C)^{1.25}, \tag{2}$$

where Q is heat transfer (watts), T_H and T_C are the hot and cold wall temperatures (°F), and A_W is the area (ft^2) of a wall.

Catton[5] recommends the Berkovsky-Polevikov[6] relationship which at 120°F air temperature becomes

$$Q = 0.043 A_W W^{0.09} H^{-0.25} (T_H - T_C)^{1.28}. \tag{3}$$

For typical values of H = 3 ft, W = 1 ft, and $T_H - T_C$ = 20°F, eqs. (2) and (3) agree within 10 percent.

By comparison, the recommended[7] heat transfer relationship for a vertical isothermal plate (assumed to exist in an infinite constant-temperature air space) predicts 23 percent less heat transfer for the same conditions. Thus, the core region in an enclosure, in particular the existence of a vertical temperature gradient in the core, has a significant effect on the heat transfer at the enclosure wall.

For the powered electronic cabinet problem modeled in the next section, eq. (2) will be used as an approximation for the convective heat transfer both between the opposing cabinet walls and between the circuit boards and cabinet walls.

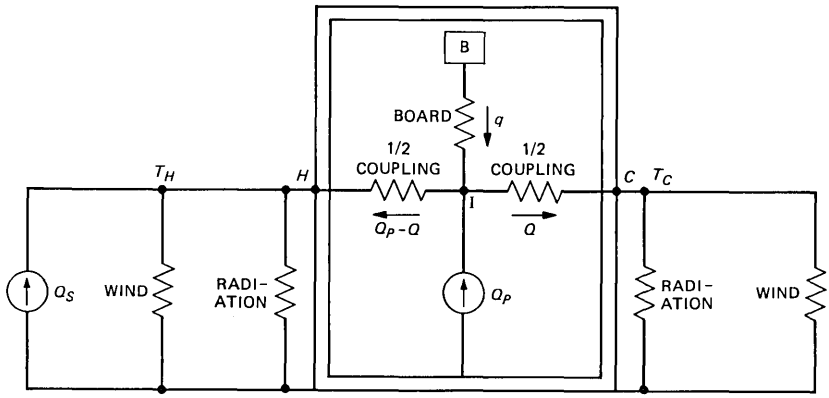
II. A LUMPED THERMAL CONDUCTANCE MODEL

A simple electrical analog for the calculation of approximate steady-state temperatures in a powered electronic cabinet subject to sun and wind is shown on Fig. 3. The absorbed solar radiation, assumed incident on half the cabinet surface, is represented by the heat source Q_s (watts) and the internal heat dissipation by Q_P (watts). Heat is transferred to ambient from the exterior cabinet surfaces by combined long-wave radiation and wind-dependent convection. For radiation, the linearized Stefan-Boltzman equation is used

$$Q_{\mathrm{rad}} = 4\sigma e A_W T_A^3 (T_W - T_A), \tag{4}$$

where σ = Stefan-Boltzman constant (5×10^{-10} watt/ft^2 − °R^4), e = emissivity ($0 < e < 1$), A_W = surface area of one wall (ft^2), T_W = wall temperature (°R), and T_A = ambient temperature (°R).

For the wind-induced heat transfer at the cabinet's exterior surfaces,



1/2 COUPLING : $Q = 0.119\, W^{0.11} H^{-0.36} A_W (T_I - T_C)^{1.25}$

RADIATION : $Q = 4\sigma A_W T_A^3 (T_C - T_A)$

WIND : $Q = 0.2 (u/D)^{0.5} A_W (T_C - T_A)$

BOARD : $q = hS (T_B - T_I)$

Fig. 3—Lumped conductance model for a powered cabinet in the sun.

consider first the cabinet top. Assume horizontal laminar air flow. The heat transfer (Nusselt number) expressed in terms of Re (Reynolds number) and Pr (Prandtl number) for this situation is given by[8]

$$\mathrm{Nu} = 0.664\ \mathrm{Pr}^{0.33}\mathrm{Re}^{0.5}. \tag{5}$$

Substitution of air properties at 100°F gives

$$h_{\mathrm{wind}} = 0.21(u/D)^{0.5}, \tag{6}$$

where D is the distance across the top measured parallel to the wind direction and u is wind speed (ft/s). For random wind direction, D is approximately given by the average side dimension of the cabinet.

The wind-induced heat transfer at the cabinet's vertical walls at any instant of time depends on the incident wind direction and speed at that time. However, as will be shown, if random wind direction is assumed, then the average heat transfer coefficient (h) at each wall can also be approximated by the simple relationship given in eq. (6). The object here is to preserve the model's simplicity.

Consider the cabinet to be a long square cylinder having side D whose axis is normal to the wind. As before, assume D to be the average side dimension of the cabinet's rectangular cross section. For this situation, the convective heat transfer is given by[9]

$$\mathrm{Nu} = C\ \mathrm{Pr}^{0.35}\mathrm{Re}^{m}, \tag{7}$$

where C and m are constants which depend on wind direction. For a range of Re from 5×10^3 to 10^5, and for the wind normal to a cabinet

wall, $C = 0.1$ and $m = 0.675$. Over the same range of Re, and for the wind direction along a diagonal (45 degrees to a wall), $C = 0.25$ and $m = 0.588$. Assuming random wind direction, the average heat transfer coefficient ($\bar{h}$) is approximately given by the average of these two cases. For a typical case $D = 2$ ft, $u = 5$ ft/s and for air properties at 100°F, $\bar{h}$ is evaluated to be 0.337. The same substitutions into eq. (6) gives $\bar{h} = 0.335$. So, although each wall of the square cylindrical cabinet has a different instantaneous heat-transfer coefficient depending on wind direction, the overall cabinet coefficient is approximately the same as that of a hypothetical cabinet which has wind parallel to all its walls simultaneously. Consequently, with the assumption of random wind direction, the average heat transfer at each cabinet wall is approximately given by

$$Q_{\text{wind}} = 0.21 A_W (u/D)^{0.5} (T_W - T_A). \tag{8}$$

For the internal convective heat transfer from the hot (sunny) cabinet wall to the cool (shaded) cabinet wall, and also from the circuit boards to the cabinet walls, Seki's relationship given in eq. (2) is used. In Fig. 3, this convective coupling is represented by a nonlinear conductance between walls, center-tapped to the heat source Q_P. In adopting Seki's results to the cabinet problem, several differences between the ideal enclosure and the real electronic cabinet are being ignored. The effect of these differences is discussed in Appendix B.

In the no-solar case ($Q_s = 0$ in Fig. 3), Q_P divides equally, half being transferred to each wall. Thus, the temperature at the center tap represents the hottest internal air temperature, which occurs at the top row of electronics if the heat sources are uniformly distributed. Its magnitude relative to the wall temperature is given in the model by half the dissipated power flowing through half the coupling conductance. Thus,

$$T_I - T_W = \frac{R_{\text{coupl}}}{2} \left(\frac{Q_P}{2} \right)^{0.8}. \tag{9}$$

An isothermal board (B) shown in Fig. 3, having an area (two sides) of S ft^2, located in the top row of electronics and dissipating q watts will experience an additional temperature rise above the internal cabinet ambient (center tap) given by

$$T_B - T_I = q/(\text{Sh}), \tag{10}$$

where h is the heat transfer coefficient at the board and S is the board surface area (both sides).

With no internal dissipation ($Q_P = 0$ in Fig. 3), solar heat in the model flows from the hot sunny wall through the coupling conductance to the cool shaded wall. In this case, the temperature at the center-tap

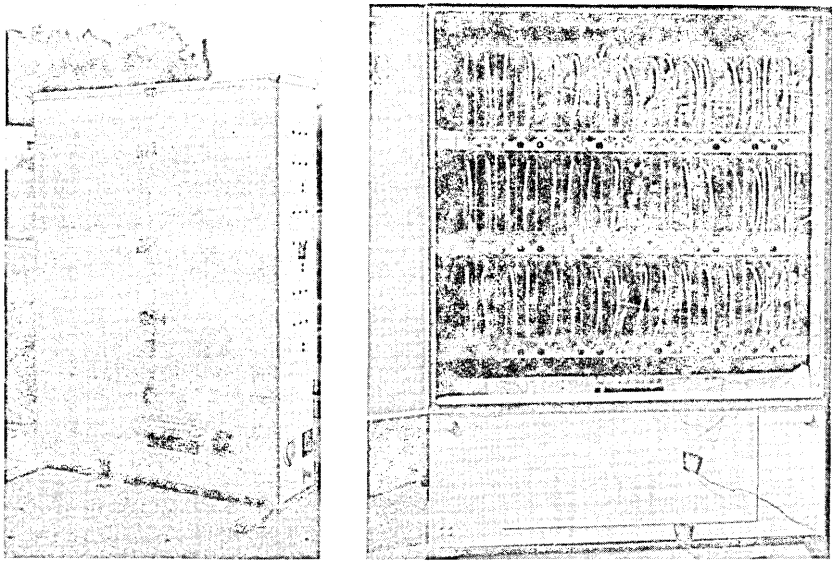


Fig. 4—40E Feeder distribution interface cabinet at test site.

represents the internal temperature at an elevation corresponding to the mean between the hot and cold wall temperatures, which occurs near the midway height of the enclosure in Fig. 2, but somewhat higher in the cabinet of Fig. 1 because of the cabinet wall's vertical temperature gradient.

Thus, in both limiting cases of $Q_P = 0$ or $Q_S = 0$, the proposed thermal conductance model accounts for the flow of heat in roughly the correct way. It remains to be seen how well it agrees with experimental data. For this purpose, the model predictions will be compared with test data taken by the author during the summer of 1979. In these tests, 96 circuit boards, each with eight resistors, were mounted in three rows of *BELLPAC** housings inside a Western Electric 40E FDI cabinet. Figure 4 shows photos of the front and back views of the cabinet at the test site. With all boards powered, the heat dissipation was fairly uniformly distributed in the cabinet.

In evaluating the thermal conductances of the model for this test, assume that the 35 ft^2 surface area of the FDI cabinet is equally divided into 17.5 ft^2 of hot sunny surface and 17.5 ft^2 of cool shaded surface, each at a uniform temperature. In eqs. (2), (4), and (8) substitute $W = 1$ ft, $H = 3.3$ ft, $D = 2$ ft, $u = 10$ ft/s, $T_A = 540°\text{R}$, $A_W = 17.5$ ft^2, and $e = 1$ to obtain $Q_{\text{RAD}} = 5.51(T_W - T_A)$, $Q_{\text{wind}} = 7.86(T_W$

* Trademark of Western Electric Company.

$- T_A$), and $Q_{coupl} = 0.58(T_H - T_C)^{1.25}$. The inverse relationships, shown on Fig. 5, are

$$T_W - T_A = 0.075(Q_{rad} + Q_{wind}) \tag{11}$$

$$T_H - T_C = 1.54Q_{coupl}^{0.8}. \tag{12}$$

The relationship given by eq. (12) for the coupling conductance has been divided into two series elements in Fig. 5 each having a coefficient of 0.77.

The value of h in eq. (10) was determined experimentally. The measured temperature rise at the center of the board in the middle of the top row relative to the air directly above (about one inch below the cabinet top) gives an effective heat transfer coefficient (h) of about 0.3 watt/ft^2−°F. Substitution of this value of h and $S = 1$ ft^2 into eq. (12) gives for a board in the top row

$$T_B - T_I = 3.3q. \tag{13}$$

Solutions to the thermal model of Fig. 5 for $Q_P = 150$ watts as a function of solar input power are plotted in Fig. 6. To show the sensitivity to wind speed, solutions for both 5 and 10 ft/s winds are presented. Shown on the figure are calculated curves for the cool wall, the hot wall and the top row of boards (points C, H, and B, respectively of Fig. 5). Recognize that the model predicts only a single temperature at each of these locations, whereas in actuality, temperature distributions exist in each case. Experiments were then performed to verify the analytical results at points where the model applies.

III. THERMAL MODEL PREDICTIONS AND COMPARISONS WITH TEST DATA

For purposes of comparing the analytic results with experimental

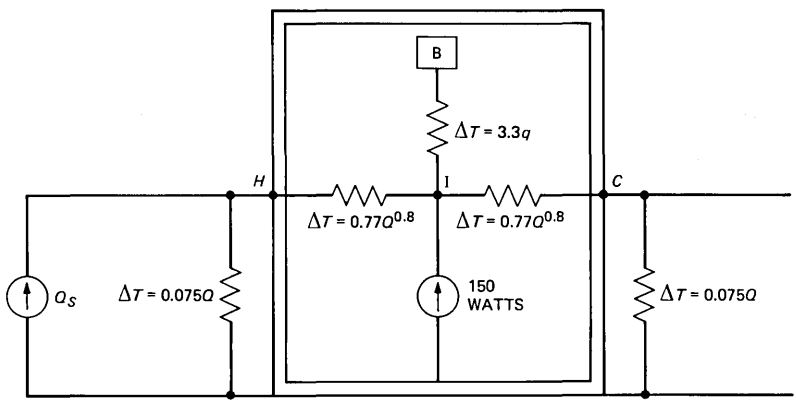


Fig. 5—Lumped conductance model for a powered cabinet in the sun for specific conditions of analysis and tests.

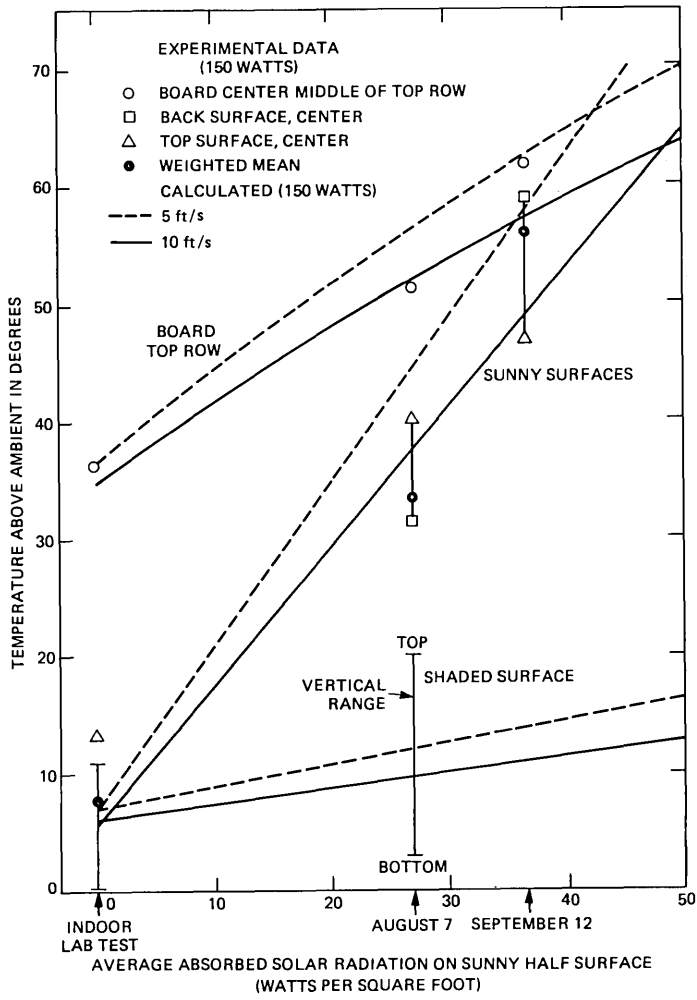


Fig. 6—Measured cabinet temperatures compared with model predictions for Q_p = 150 watts and uniform breeze on all surfaces.

data obtained by the author, two test dates have been selected because of their exceptionally clear skies, August 7 and September 12, 1979. The absence of clouds on these dates simplifies the task of correlating solar intensity with measured cabinet temperatures. Transient effects are small since the cabinet's time constant (measured in the laboratory to be about 0.75 hours) is small compared with the time scale of the solar input waveform. Assuming the solar input to be approximated by a half sine-pulse of 10-hours' duration, the peak daily temperature in the cabinet would be 98 percent of an ideal cabinet having no thermal lag.

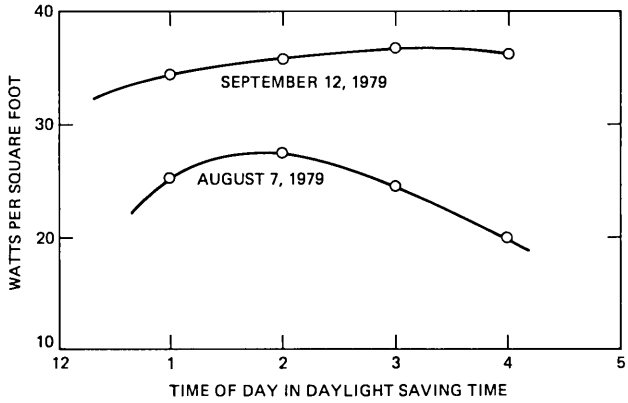


Fig. 7—Average absorbed solar flux calculated from pyranometer traces for two test dates.

Figures 13 and 14 in Appendix A show the incident solar radiation on a horizontal plane for these two dates using a Weatherman R413 star pyranometer. The solar radiation absorbed by the test cabinet for each test date can be calculated from this data. Also needed is a specification of the absorptivity and area of each cabinet surface, as well as the orientation of each surface relative to the direct rays of the sun. This calculation is shown in Appendix A and the results given in Fig. 7.

As shown in Fig. 7, the peak solar power absorbed by the cabinet on September 12 was about 35 percent greater than that on August 7. About half of this large difference was due to the sun's lower position in the sky in September, which resulted in a larger component of direct solar radiation on the cabinet's large back surface. The remaining half was due to a greater solar intensity on September 12, attributable to reduced atmospheric attenuation compared with August 7. As shown in Appendix A, the September 12 intensity equaled the values published by ASHRAE[10] for this date, while the August 7 intensity fell short of ASHRAE values.

Figures 8 and 9 show measured cabinet temperatures for the same two dates. The cabinet top and back surface temperatures, which constitute the sunny hot walls, are plotted versus time of day. Observe that the back surface was hotter than the top surface on September 12, but that the reverse was true on August 7, indicative of the sun's lower altitude in September. The cabinet internal temperature was measured at the center of the board located at the middle of the top row and is believed to be fairly representative of the cabinet's maximum temperature (exclusive of hot spot effects near the resistor heat sources). It is close to the temperature that would occur if the cabinet's

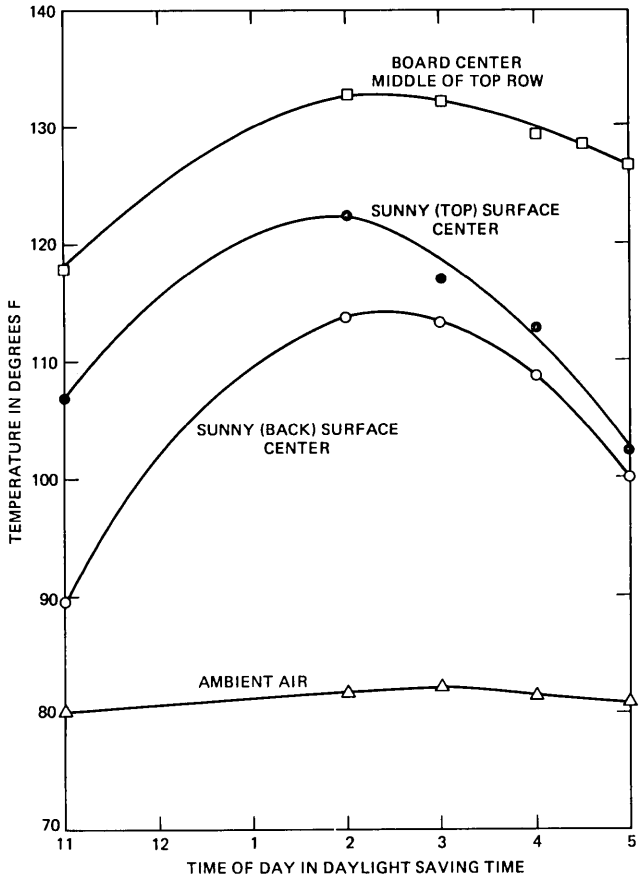


Fig. 8—Cabinet temperatures on August 7, 1979, 150-watt dissipation.

internal dissipation were truly uniform. On September 12 (from Figs. 7 and 9), the cabinet's peak internal temperature rise above ambient was about 62°F; the absorbed solar radiation was 36.5 watts/ft^2. On August 7 (Figs. 7 and 8), the cabinet's peak internal temperature rise was about 51°F; the absorbed solar radiation was 27 watts/ft^2. These results, along with measured surface temperatures, are plotted alongside the analytic results in Fig. 6. Agreement is within 15 percent of the 10 ft/s (6.8 mph) wind curve. The average wind speed during the tests is estimated to be in the range of 5 to 10 mph.

Also given in Fig. 6 is the cool-shaded wall for August 7, shown as a range of temperatures (no measurement made on September 12). It is typical for the shaded wall to have such a vertical temperature gradient since it is the principal heat transfer surface for internal dissipation. By contrast, the sunny hot walls have a fairly uniform temperature.

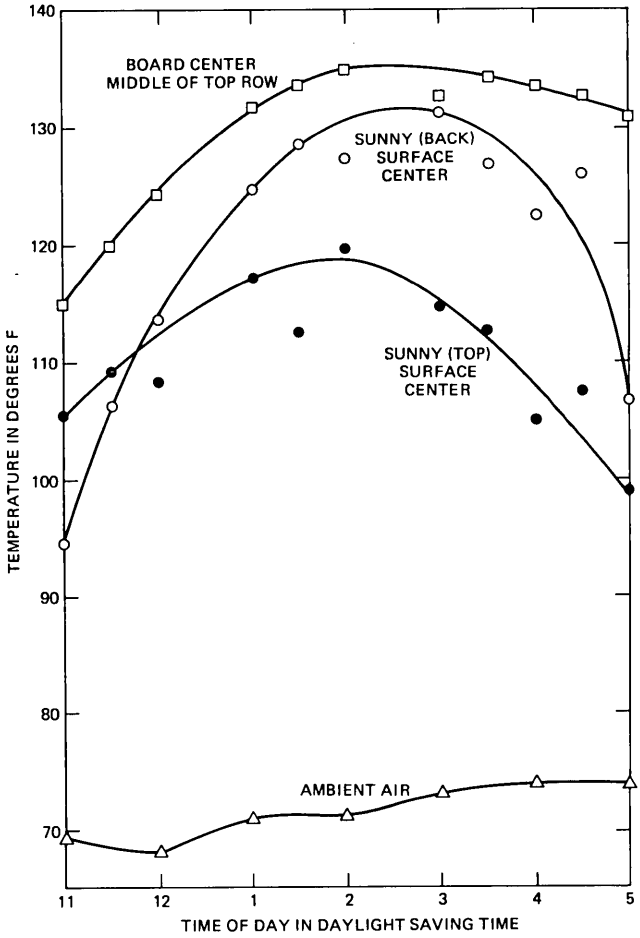


Fig. 9—Cabinet temperatures on September 12, 1979, 150-watt dissipation.

Evidence of this is provided by Fig. 10 which shows vertical temperature profiles measured on the sunny (cabinet back) surface, the shaded (cabinet front) surface and the cabinet centerline (center of boards in the middle of rows). Observe the large vertical temperature gradients on the cabinet center line (1°F/in.) and the shaded surface (0.5°F/in.). By comparison, the sunny surface gradient of 0.1°F/in. is small.

For the case of no solar radiation, Fig. 6 shows experimental temperatures measured in the lab at the same cabinet locations as before. Agreement is almost exact with respect to the 5 ft/s wind curve. The same measurements taken outdoors in the evening or early morning (not shown in the figure) were slightly less because of the action of

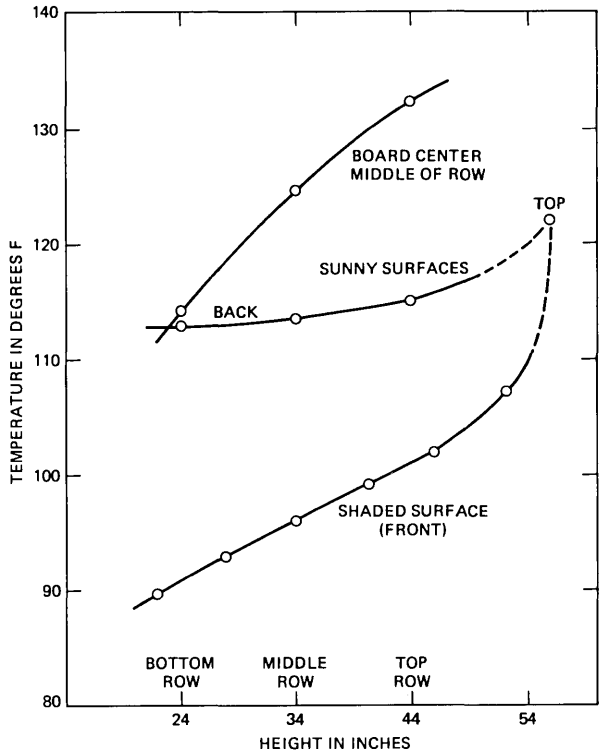


Fig. 10—Vertical temperature profiles on 2 p.m. daylight saving time, August 7, 1979, 150-watt dissipation.

light outdoor breezes. For instance, the board temperature averaged 2 to 3°F cooler, giving closer agreement with the 10 ft/s wind curve.

More generally, for the case of no solar radiation, the model's prediction can be expressed algebraically. From inspection of Fig. 5, the board temperature (T_B) for $Q_S = 0$ and $u = 10$ ft/s is seen to be

$$T_B - T_A = 0.77(Q_P/2)^{0.8} + 0.075(Q_P/2) + 3.3q, \qquad (14)$$

where $q = Q_P/96$. For a 5-ft/s breeze, the 0.075 coefficient becomes 0.106. A comparison of eq. (14) with laboratory experimental temperatures is shown in Fig. 11. The agreement is within 6 percent for both wind speeds over a 250 watt range of Q_P. The calculated curve for wind speed of 5 ft/s is close to lab conditions of natural convection.

Also plotted on Fig. 11 is the measured air temperature (corresponding to T_I in the model) at a point on the cabinet center line about one inch below the cabinet top. The measured temperature rise at the board (T_B) relative to this (T_I) gives the effective heat transfer coefficient of 0.3 watt/ft^2 − °F, which was used to obtain eq. (13).

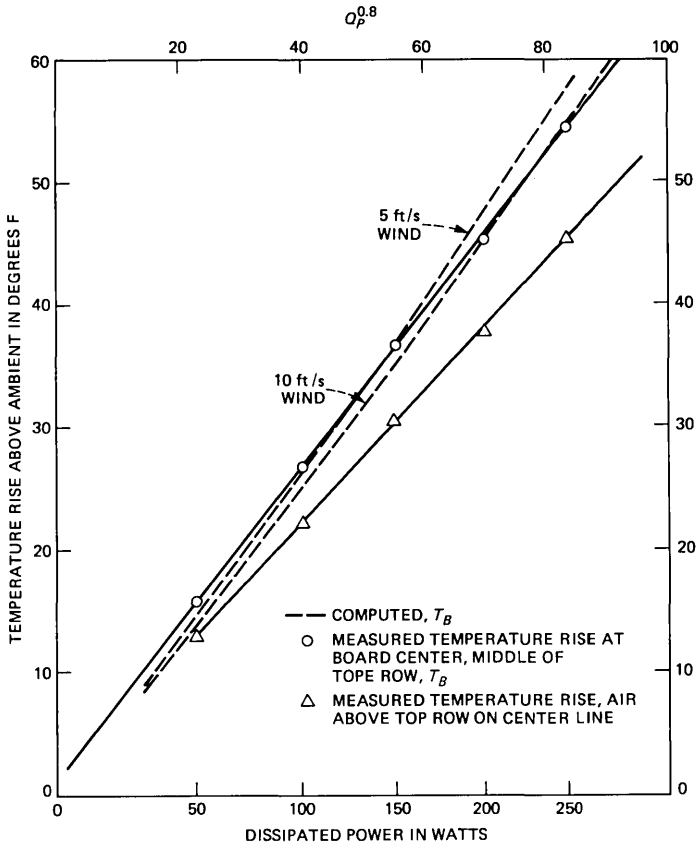


Fig. 11—Comparison of analysis with indoor experimental results.

Returning to Fig. 6, one sees that the measured board temperature agrees fairly well with the analysis for assumed breezes of 5 to 10 ft/s on all cabinet surfaces. Also, the mean hot wall temperature for September 12 (weighted mean of top and back wall) agrees. However, the mean hot wall temperature for August 7 is less than the model prediction. It is believed that the reason for this disagreement is different magnitude breezes on different cabinet surfaces. On August 7, the prevailing breeze direction at the test site was from the south, tending to produce more convective heat transfer at the sunny surfaces than at the shaded surfaces. Figure 12 shows the results of recalculating the cabinet temperatures, increasing the assumed hot wall breeze to 15 ft/s (10.2 mph) and decreasing the assumed shaded wall breeze to 5 ft/s (3.4 mph). The agreement with August 7 data is improved, showing that a reasonable adjustment to the assumed average wind speed at each wall brings the analytic and experimental wall temper-

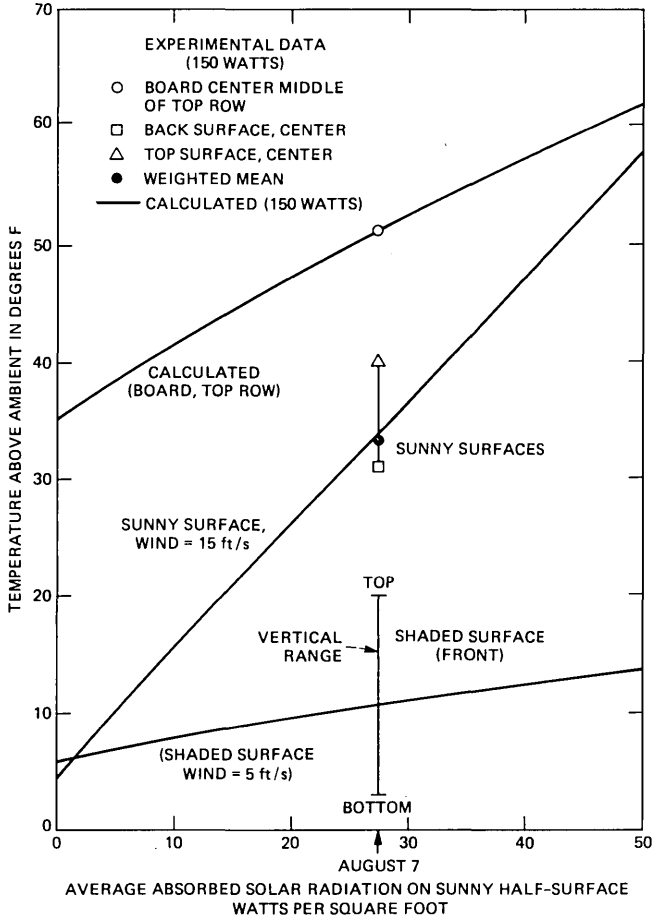


Fig. 12—Measured cabinet temperatures compared with model predictions for Q_p = 150 watts and different wind conditions on opposite cabinet walls.

ature into near perfect agreement. The calculated board temperature is only slightly effected since it depends primarily on average wall conditions.

IV. SUMMARY

An approximate lumped thermal conductance model is presented for calculating maximum steady-state board temperatures in unventilated electronic cabinets, subject to both solar and uniformly-distributed, internally-dissipated heat. The model has general applicability, with input parameters consisting of cabinet geometry, solar radiation, wind speed, and internal heat dissipation.

The main feature of the model is the use of a convective coupling

conductance based on an empirical relationship found in the recent literature for the convective heat transfer between the hot and cold wall of an unpowered enclosure, and the attachment of a heat source (representing internally dissipated heat) to the center tap of this coupling conductance.

Several phenomena are ignored in the analysis. The sensitivity to these are shown in Appendix B to be in the order of 10 percent. Consistent with this, the agreement between the analysis and experimental results is within 10 percent.

The results of this work provides a useful design tool for determining the minimum cabinet size needed to safely dissipate the total heat of a system. The temperature rises of individual devices and circuit boards, which can be determined from individual laboratory tests, can be added to the predicted maximum cabinet ambient (T_I of the model) to give the maximum device or circuit board temperature operating in the full-system cabinet ambient. Also, the analysis tells the designer how much temperature rise to allow for solar radiation and how much cooling can be expected from breezes.

Although useful for predicting the maximum cabinet ambient at the top row, the model does not predict the cooler temperatures at lower rows. This is not a serious limitation for two reasons. First, the thermal design is limited by the maximum temperature which normally occurs at the top. Second, one can determine the vertical temperature profile up between the boards by scaling the results from central office frame analyses. The board temperatures in the cabinet are hotter than in the central office frame because of the reduced convective air flow in the cabinet, but the profiles are similar with respect to height. In both cases, one can assume ambient temperature below the bottom shelf of boards.

A more severe limitation is the model's inability to account for nonuniformly distributed heat sources. A similar difficulty exists in analyzing central office frames where it is common practice to compute temperatures based on average heat dissipation. The cabinet model, by considering just total dissipation, is doing a similar thing. For modest departures from uniformity (e.g., alternate boards powered) the model's effectiveness is unimpaired. Also, certain extremely nonuniform dissipations can be handled. For instance, if all the heat dissipation is at one shelf level, good experimental agreement is obtained if an effective area is substituted into the model which excludes the wall below that shelf level.

These and other topics are under study to refine the approximate model presented here and ultimately to be incorporated into general engineering design guidelines for reliable packaging of loop electronics in outdoor environments.

V. ACKNOWLEDGMENT

The author gratefully acknowledges the encouragement and helpful suggestions given by P. J. Lauriello in the preparation of this paper.

APPENDIX A

Computation of Absorbed Solar Power

Figures 13 and 14 give the vertical component of solar intensity on August 7, 1979, and September 12, 1979, measured by P. E. Fiechter using a Weatherman Star pyranometer. Also shown on each figure are calculated curves of this vertical component using data and relationships given by ASHRAE.[10] Observe that the computed curve of total radiation (direct + diffuse) agrees very well with the September 12 pyranometer trace. By comparison, the pyranometer trace for August 7 falls short of the computed curve for this date. The discrepancy on August 7 is attributed to a light haze on this date.

The total solar power absorbed by the cabinet can be computed from Figs. 13 and 14 and a specification of the orientation of each cabinet wall with respect to the direct rays of sunlight. For this purpose, Fig. 15 shows traces of the sun's altitude angle (L) and azimuth angle (Z) (with respect to south) taken from ASHRAE.[10] The solar power absorbed by each vertical wall is given by

$$Q_V = \alpha A_W I \cos(L)\cos(Z - N)$$

and that absorbed by the top (horizontal) surface

$$Q_H = \alpha A_W I \sin(L),$$

where α is the absorptivity (0.78), N is the angle made by the wall's

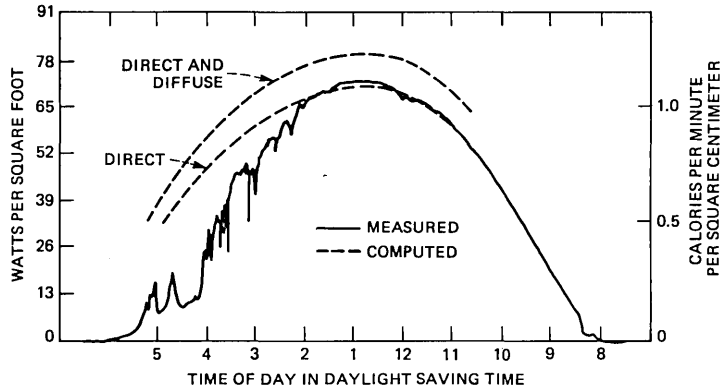


Fig. 13—Pyranometer trace compared with solar intensity computed from ASHRAE data for a horizontal surface on August 7, 1979.

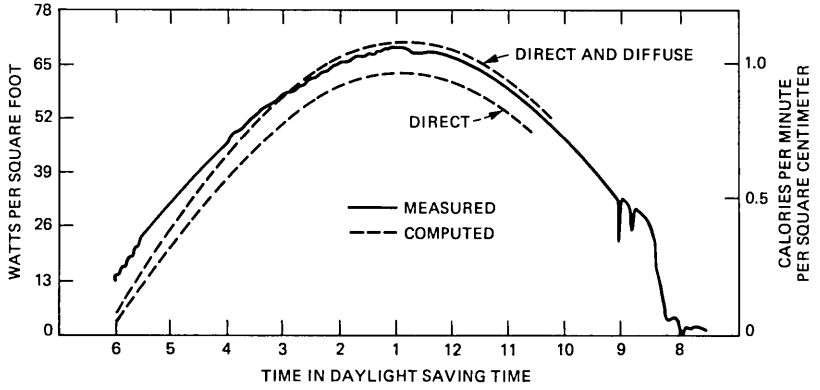


Fig. 14—Pyranometer trace compared with solar intensity computed from ASHRAE data for a horizontal surface on September 12, 1979.

normal with south, A is the wall area (ft^2), and I is the intensity of the direct radiation (watts/ft^2) given by dividing the pyranometer trace by $\sin(L)$. The cabinet surfaces exposed to sun were: top – 3.3 ft^2, side – 3.3 ft^2, back – 11.1 ft^2. The back surface faced 30° west of south (N = 30°).

The total absorbed solar power for each date, shown in Fig. 7 of the text, is obtained by adding the contribution of each wall.

APPENDIX B

The Effect of Factors Ignored in the Analysis

Several phenomena which exist in a real electronic cabinet have been ignored in the analysis. This appendix discusses five of these phenomena and estimates their effect on temperature.

In using Seki's results [eq. (1)], the restriction to air flow caused by equipment in the cabinet is ignored. Most of the convected heat is carried in a thin (typically less than one inch) boundary layer at the cabinet walls. The air drag of the equipment, which is mostly in the core of the cabinet, is small provided adequate clearance exists at the walls. The effect on temperatures of this air drag is estimated to be the same order of magnitude as that caused by reducing Seki's enclosure width (W) to a value equal to the sum of the clearances at opposing cabinet walls, say two inches. From eq. (2), one can see that the increase in temperature for an enclosure whose original dimensions are H = 40 in. and W = 10 in., which is reduced in width to W = 2 in., is about 15 percent. Thus, the effect of equipment in the cabinet is estimated to be of the same order of magnitude.

Note that eq. (1) is applicable only within the boundary layer flow-regime which, according to Eckert and Carlson,[3] occurs for height-

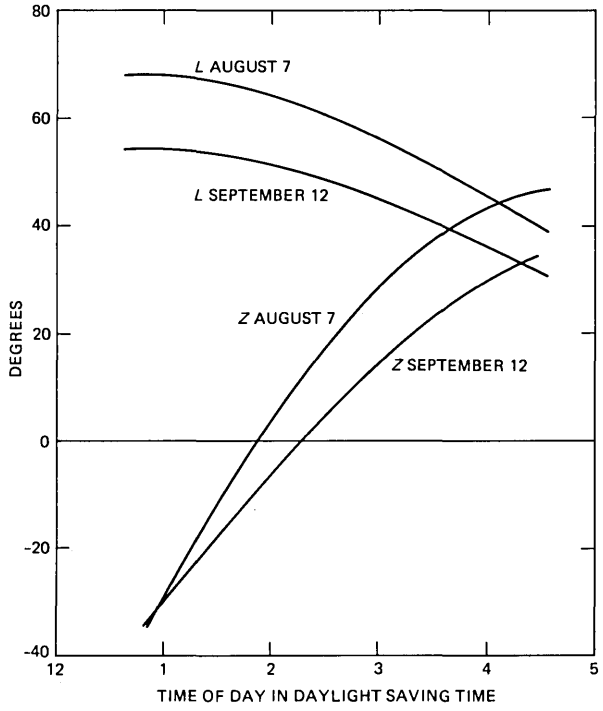


Fig. 15—Solar altitude and azimuth (from south) for test dates.

width ratio less than 2.32×10^{-4} Gr_W, where Gr_W is the Grashof number based on enclosure width. The limiting clearance per wall based on this constraint, for $T_H - T_C = 25°F$ and $H = 40$ in., is 0.9 in. Thus, eq. (2) should be valid for the 1-inch clearance.

A second difference between Seki's enclosure and the real cabinet is the fact that the cabinet walls are not forced to be isothermal. The vertical temperature gradient that naturally occurs up the cabinet walls alters the overall heat transfer from that measured by Seki and others. The effect of this on the internal temperatures is difficult to estimate, but a rough idea can be obtained by comparing the maximum temperature of a constant heat flux wall with that of an isothermal wall, both standing vertically in a uniform ambient. For either wall, $\Delta T = NH^{0.2}Q^{0.8}$, where $N = 8.7$ for constant heat flux and $N = 7.2$ for constant wall temperature. The maximum temperature of the constant heat flux wall is 19 percent greater. Assuming the cabinet walls to lie somewhere between these extremes, then the analysis underestimates the internal cabinet temperatures by about 10 percent.

Conductive and radiative heat transfer have been ignored in the analysis. Conduction from the circuit boards to the cabinet walls is negligible. However, conduction from the hot sunny wall to the cool

shaded wall through the cabinet metal is estimated to be about 0.1 watt/°F for the test cabinet. This conductance amounts to about 10 percent of the convective heat flow given by eq. (2). The average wall temperature and internal cabinet temperature are affected to a much lesser extent.

The heat transfer by radiation from the heat dissipating boards to the cabinet walls depends largely on geometry. For the test conditions consisting of 8-inch deep boards on 1-inch centers, the view factor at the board center is 0.03. Because of this small view factor and because the emissivity of the boards approaches zero in the plane of the board, the heat transfer by radiation is small, calculated to amount to about 10 percent of the boards' dissipation. Consequently, board temperatures are about 10 percent cooler than that predicted by the model.

A fifth factor is the cabinet roof which, in the real cabinet, is a principal heat transfer surface but, in the ideal enclosure, is adiabatic. In the analyses, the roof area was lumped in with that of the vertical walls, thereby assigning to the roof the same average heat transfer as that of the walls. Some error results from the approximation. For instance, in the no-solar case, the heat transfer per unit area through the roof is approximately twice that of the walls (on average), since its temperature equals that at the top (maximum) of walls. Thus, for the test cabinet whose roof comprises 10 percent of the total cabinet surface, the analysis underestimates the total heat transfer and overestimates the cabinet temperature by about 10 percent.

Five factors (more could be added), which are ignored in the analysis, are discussed in this Appendix. Some tend to increase temperature and some decrease temperature. All, broadly speaking, are estimated to have effects in the range of 10 percent. This is consistent with the agreement found between analysis and experiment.

REFERENCES

1. R. W. Henn and D. H. Williamson, "The Loop Plant—Physical Design Considerations," B.S.T.J., *57,* No. 4 (April 1978), pp. 1185–223.
2. J. W. Elder, "Laminar Free Convection in a Vertical Slot," J. Fluid Mech., *23,* Part 1 (1965), pp. 77–98.
3. E. R. G. Eckert and W. O. Carlson, "Natural Convections in an Air Layer Enclosed Between Two Vertical Plates With Different Temperatures," Int. J. Heat Mass Transfer, *2,* (1961), p. 106.
4. N. Seki, S. Fukusako, and H. Inoba, "Visual Observations of Natural Convective Flow in a Narrow Vertical Cavity," J. Fluid Mech., *84,* Part 4, (1978), pp. 695–704.
5. I. Catton, "Natural Convection in Enclosures," Keynote papers of Sixth Int. Heat Transfer Conf., Toronto, *6,* (August 1978) pp. 13–31.
6. D. B. Spalding and H. Afgan, "Heat Transfer and Turbulent Buoyant Convection," *I* and *II,* Washington, DC: Hemisphere Publishing (1977), pp. 443–55.
7. W. H. McAdams, "Heat Transmission," New York: McGraw-Hill 1954, pp. 172–3.
8. H. Grober and S. Erk, "Fundamentals of Heat Transfer," New York: McGraw-Hill, 1961, p. 216.
9. S. S. Kutateladze and V. M. Borishanskii, "A Concise Encyclopedia of Heat Transfer," Elmsford, N. Y.: Pergamon Press, 1966, pp. 140–3.
10. ASHRAE Handbook, American Society of Heating Refrigerating and Air Conditioning Engineers, New York, 1977, pp. 26.1–9.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 2, February 1982
Printed in U.S.A.

Fail-Safe Nodes for Lightguide Digital Networks

By A. ALBANESE

(Manuscript received October 6, 1981)

Lightguide digital networks that use fail-safe nodes made of an optical regenerator and optical couplers are described and analyzed. Every node in the network can regenerate or overwrite information traveling in a ring or bus network, and in the case of a power failure at one of the nodes, the network continues to function because the coupler keeps the continuity at the failing node. Fail-safe nodes that operate at 16 Mb/s were built to implement a digital network of ring architecture. A description of the components is presented, together with an analysis of the design constraints of the different parts of the fail-safe nodes.

I. INTRODUCTION

The use of regenerators at the nodes of a lightwave network introduces a reliability problem when the power at one node fails. Optical passive couplers solve this problem, but the number of passive couplers in a network is limited by the maximum insertion loss that can be tolerated between a transmitter and the receiver farthest away from it.[1]

This paper describes a new arrangement for a lightguide digital network built with fail-safe nodes and with the characteristics that the number of stations is independent of the coupler insertion loss, and that the network keeps functioning when the power at one or more nodes fails. A fail-safe node consists of a lightguide receiver and a lightguide transmitter electrically connected by a regenerator and optically connected by a directional coupler. Figure 1 shows a configuration for a fail-safe node consisting of a lightwave receiver and transmitter pair connected by a regenerator and a directional coupler that provides optical continuity when the power at the node fails. The feasibility of the network was tested using lightwave transmitters

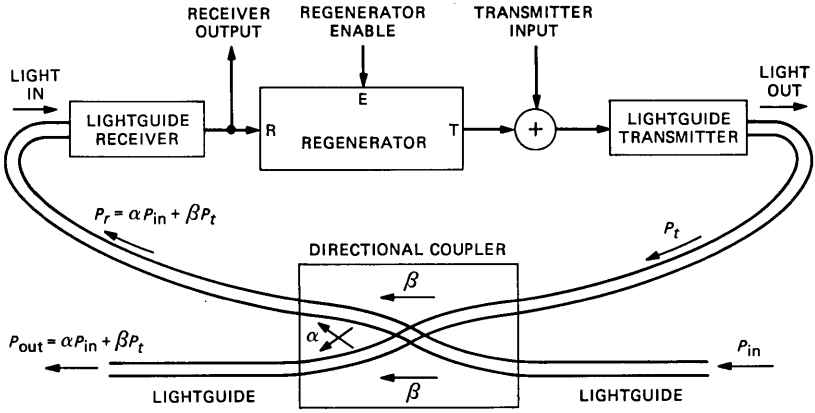


Fig. 1—Fail-safe node.

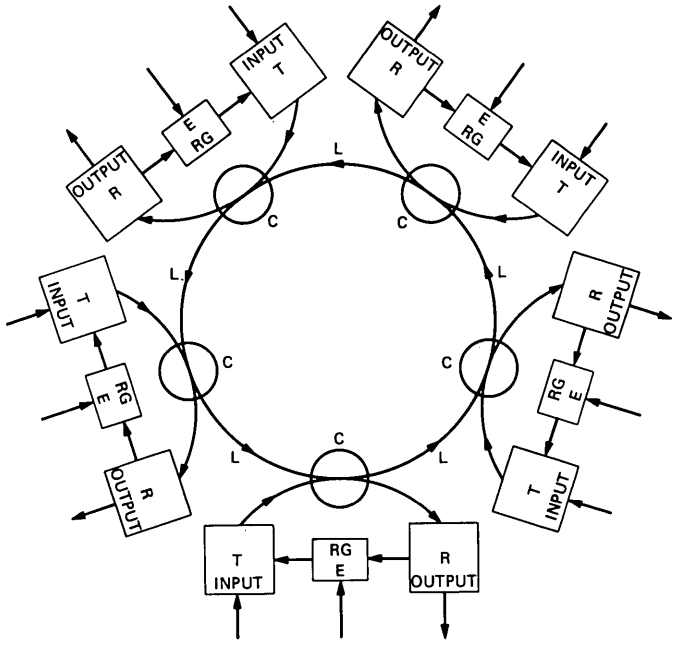


Fig. 2—Ring-type lightguide network.

(GaAlAs LED, $\lambda = 0.8$ μm) and lightwave, avalanche photodiode (APD) receivers made by Western Electric. The repeaters were built using transistor-transistor logic (TTL) integrated circuits, and the network was operated with 16 Mb/s digital signals.

A node can regenerate, overwrite, or be off, depending on whether

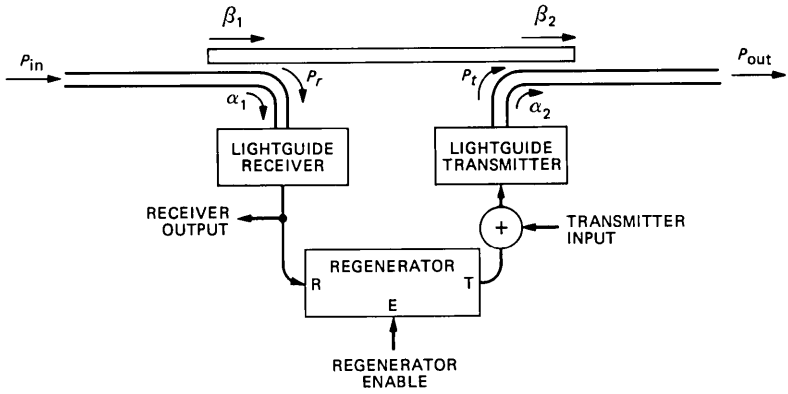


Fig. 3—A two-coupler node.

the regenerator is operating. In the regenerating configuration, the transmitter is controlled by the receiver, while in the overwriting state, the transmitter is independent of the receiver.

II. FAIL-SAFE NETWORKS

Fail-safe nodes can be connected together by a lightguide to form a ring-type network as shown in Fig. 2. Each node consists of a coupler, C, a receiver, R, a regenerator, RG, and a transmitter, T. The nodes are connected by lightguides, L. The E input disables the regenerator. The ring architecture was selected as an example; fail-safe nodes may also be used in other optical bus-type networks.[1] The nodes in the network are normally regenerating; that is, each node listens and regenerates the information flowing in the network. When a node wants to transmit, it turns its regenerator off and the information is inserted in the network by the lightguide transmitter. If the power at one node fails, or if the electronic components are removed for maintenance, the node is in the off state, and the optical coupler provides the continuity needed for the operation of the network. An additional advantage is that all the signals from different stations arrive with the same intensity at every receiver.

For proper operation of the fail-safe network, every node in the network must meet the three following constraints:

(*i*) *Sensitivity constraint*—The receiver of any node must be sensitive enough to receive the signal from a preceding transmitter when several nodes between the transmitter and the receiver are off.

(*ii*) *Interference constraint*—When two nodes are transmitting simultaneously, a node down the line should receive the signal from the closer node. This discrimination between the two transmitters is

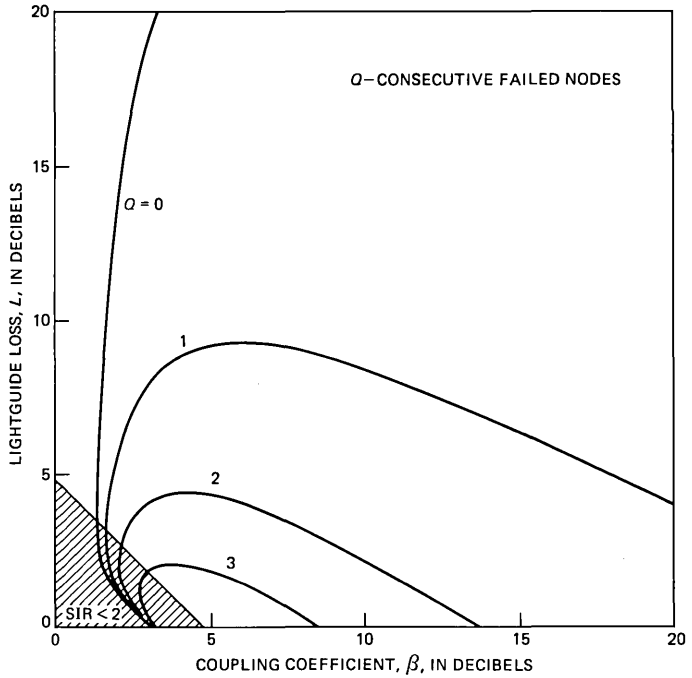


Fig. 4—Feasibility graph for one-coupler nodes.

achieved by the automatic gain control (AGC) of the receiver that adjusts the gain so the comparator circuit can detect only the stronger signal, while the less intense signal is below the threshold level.

(*iii*) *Automatic gain control response time*—In the case of a ring-type network, the sending node should be prevented from regenerating its own pulses to avoid having pulses traveling around the ring forever. This constraint is satisfied when the response time of the AGC in the lightguide receiver is longer than the time it takes a pulse to go around the ring once.

III. ANALYSIS

Each fail-safe node in the network may have one or two couplers, depending on whether the network uses return-to-zero or nonreturn-to-zero formats. Figure 1 shows the case where one coupler is used. The receiver has to be off during the time the transmitter is on to avoid saturation. This is achieved using a signal with a duty ratio less than 50 percent, and having the transmitter operating out of phase from the receiver. Figure 3 shows the coupler arrangement when two couplers are used.[2] In this case, the receiver can always be on because it does not receive light from the transmitter.

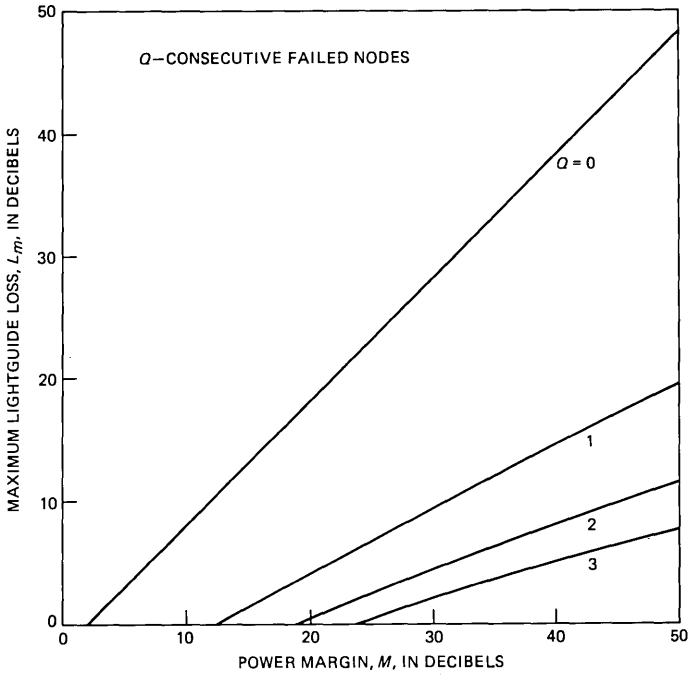


Fig. 5—Maximum lightguide loss for one-coupler nodes.

The coupler used in both cases can be characterized by a four-port device with a transmission coefficient α, a coupling coefficient β, and an excess loss coefficient $\gamma = \alpha + \beta$. Couplers with γ better than -1 dB have been reported in the literature.[3]

In the case of Fig. 1, P_{in} is the light entering the coupler from the ring; P_t is the light entering the coupler from the transmitter; $\alpha P_{in} + \beta P_t$ is the amount of light entering the receiver, P_r; $\alpha P_t + \beta P_{in}$ is the amount of light leaving the coupler and going into the ring, P_{out}. And in the case of Fig. 3, $P_r = \alpha_1 P_{in}$, and $P_{out} = \alpha_2 P_t + \beta_1 \beta_2 P_{in}$. These two node configurations will be analyzed next.

3.1 One-coupler nodes

Let us consider first the case of one-coupler node and analyze a hypothetical network where Q adjacent nodes have failed and they are off. The power received after Q failed nodes is $P_s + P_i$; P_s is the power received from the closest active transmitter:

$$P_s = P_t L^{Q+1} \beta^Q \alpha^2, \tag{1}$$

and P_i the sum of all the powers received from all the other previous active transmitters that may cause interference:

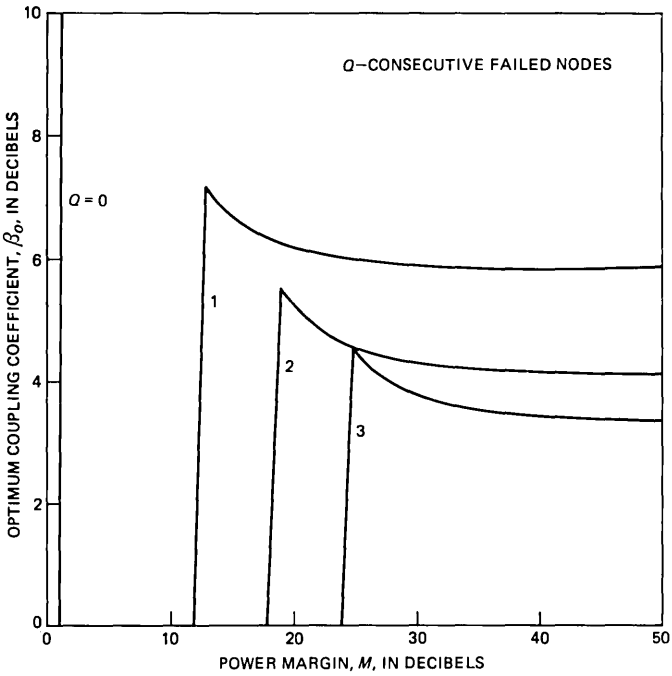


Fig. 6—Optimum coupling coefficient for one-coupler nodes.

$$P_i = \sum_{i=Q}^{i=\infty} P_t \alpha^2 L^{i+2} \beta^{i+1} = \frac{P_t L^{Q+2} \beta^{Q+1} \alpha^2}{1 - \beta L}. \tag{2}$$

In eqs. (1) and (2), L is the average lightguide attenuation between two adjacent nodes, and the infinite summation accounts for the worst case of interference.

The sensitivity and the interference constraints are satisfied when the effective received power, $P_{re}(Q)$, is larger than the sensitivity of the receiver, S,

$$P_{re}(Q) = P_s - P_i \geq S; \tag{3}$$

the minus sign accounts for the reduction in the opening of the eye diagram caused by the interference.

Equations 1 and 2, and $\alpha = \gamma - \beta$ are used to rewrite eq. 3 as

$$L^{Q+1} \beta^Q (\gamma - \beta)^2 \left| \frac{1 - 2\beta L}{1 - \beta L} \right| \geq \frac{S}{P_t} = M^{-1}, \tag{4}$$

which limits the values of β and L that satisfy the network constraints. In eq. 4, M is the optical power margin between the transmitter and the receiver.

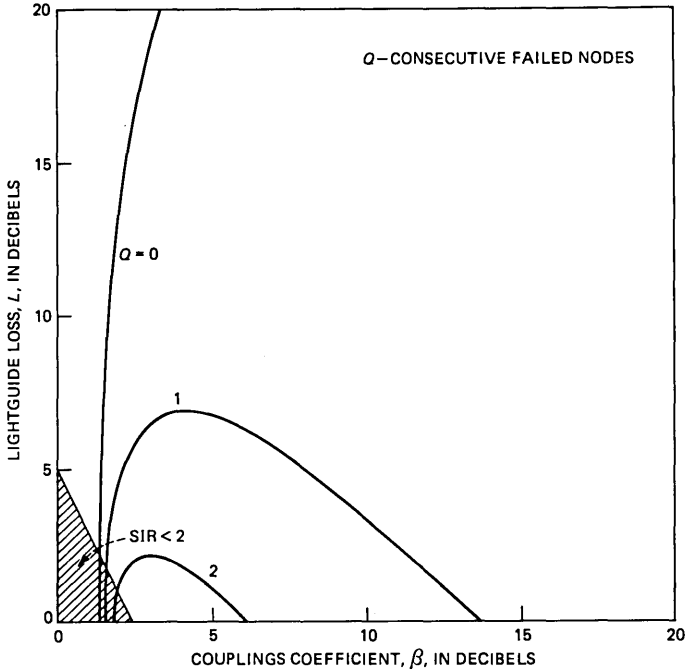


Fig. 7—Feasibility graph for two-coupler nodes.

The interference is deterministic, and it will not be seen by the comparator in the receiver because the AGC circuit sets the threshold automatically to one half of the peak amplitude which is precisely in the middle of the eye pattern. In addition to satisfying the sensitivity constraint expressed by eq. 3, one should have a signal-to-interference ratio (SIR) greater than 2 to eliminate any possible error caused by variations in the pulse amplitude,

$$\text{SIR} = \frac{P_s}{P_i} = \frac{1 - \beta L}{\beta L} \geq 2. \tag{5}$$

The SIR value of 2 was selected experimentally as the value where the error rate doubles.

Equations 4 and 5 are used to find the minimum value of L for a given β. This is done by defining a variable $V = \beta L$ that allows us to express L and β as a function of V:

$$L = \frac{V}{\gamma} + F(V) + [F^2(V) + 2VF(V)/\gamma]^{1/2}$$

$$\beta = \frac{V}{L}, \tag{6}$$

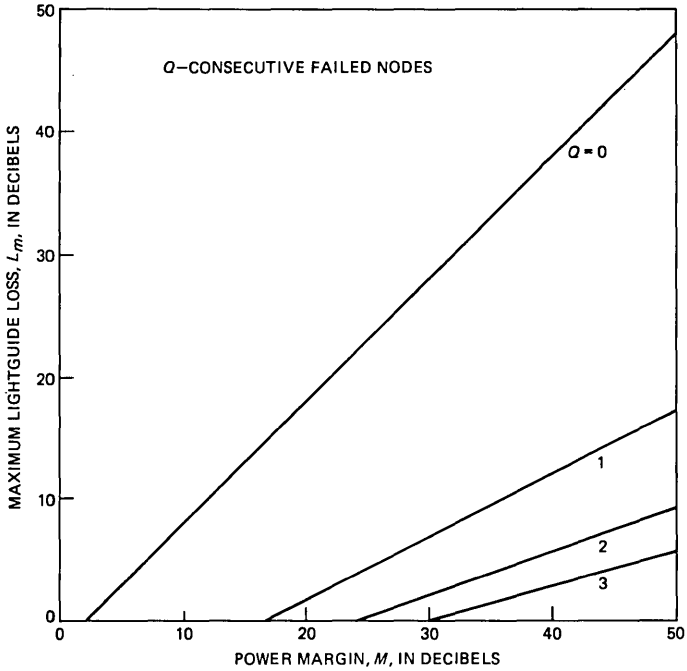


Fig. 8—Maximum lightguide loss for two-coupler nodes.

where

$$F(V) = \frac{1 - V}{2M\gamma^2(1 - 2V)V^Q}.$$

Equation 5 restricts the possible values of V to be within the range from 0 to 0.33.

Figure 4 shows a relation between L and β as given by eq. 6 for values of $Q = 0$, 1, 2, and 3, $\gamma = -1$ dB, and $M = 30$ dB. Values L and β are generally expressed in decibels, and L is commonly called the lightguide loss.

We can define an optimum coupling, β_o, as the value of β that allows the maximum lightguide loss, L_m, for a given power margin M. Figures 5 and 6 show the values of maximum L_m, and β_o as a function of M, and for a $\gamma = -1$ dB.

Figures 4 and 5 also show that the use of the fail-safe nodes in the network reduces the maximum lightguide loss between repeaters. This fact cannot be tolerated in transmission systems, but it may be possible in local area networks where the lightguide loss may not be a limiting factor.

Figure 4 shows that the maximum lightguide loss L_m is not sensitive

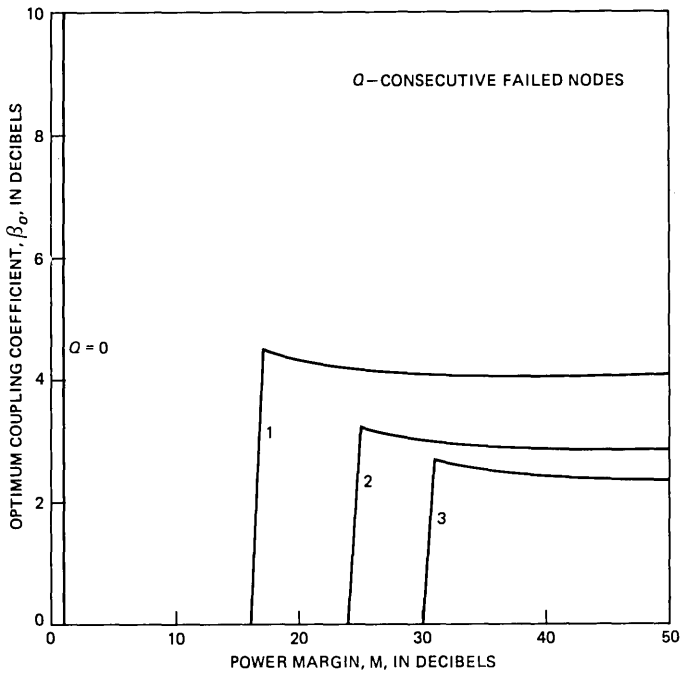


Fig. 9—Optimum coupler coefficient for two-coupling nodes.

to variations of β within ±1 dB from the optimum value of β, β_o. From Fig. 6, one could establish that a coupler with β between 4 and 6 dB is adequate for different values of M and Q.

3.2 Two-coupler nodes

The analysis of a network with two-coupler nodes is similar to the case of one-coupler nodes substituting β by β^2 in eqs. 1, 2, and 5. Note that the expression of $\alpha = \gamma - \beta$ remains the same for both cases. In this case, β and L are related by the expression

$$L^{Q+1}\beta^{2Q}(\gamma - \beta)^2 \left|\frac{1 - 2\beta^2 L}{1 - \beta^2 L}\right| \geq \frac{S}{P_t} \tag{7}$$

for the sensitivity constraint, and

$$\text{SIR} = \frac{1 - \beta^2 L}{\beta^2 L} = 2 \tag{8}$$

for the interference constraint. Figure 7 shows a relation between L and β as given by eqs. (7) and (8) for values of $Q = 0$, 1, and 2, $\gamma = -1$ dB, and $M = 30$ dB. Figures 8 and 9 show the values of L_m and β_o as a function of M when $\gamma = -1$ dB.

A comparison of Figs. 7, 8, and 9 against Figs. 4, 5, and 6 indicates that there is a 3-dB penalty when using nodes made of two couplers instead of one.

IV. CONCLUSIONS

Two configurations of optical couplers were analyzed to provide continuity in an optical network in the case of a power failure at several consecutive nodes. The analysis determines the optimum coupling coefficient, and the maximum lightguide loss that a network can have. Optical networks with fail-safe nodes are of interest in local area networks where the lightguide transmission loss is substantially less than the optical power margin between the transmitter and the receiver.

REFERENCES

1. H. W. Giertz, V. Vucins, and L. Ingre, "Experimental Fiber Optic Databus," Proc. Fourth European Conf. Opt. Commun., Genoa, Italy, September 12–15, 1978, pp. 641–5.
2. M. Chown and J. G. Farrington, "Data Transmission System," U. S. Patent No. 4,166,946, September 4, 1979.
3. B. S. Kawasaki and K. O. Hill, "Low-loss Access Coupler for Multimode Optical Fiber Distribution Networks," Appl. Opt., *16*, No. 7, (July 1977), pp. 1794–5.

CONTRIBUTORS TO THIS ISSUE

Andres Albanese, Ingeniero Electrico, 1970, University of Central Venezuela; M.Sc., 1972, University of Texas at Austin; Ph.D., 1976, Stanford University; Instituto Venezolano de Investigaciones Cientificas, 1969–1970; Bell Laboratories, 1975—. Mr. Albanese's current research interests are systems and components for optical communications and computer networks.

James C. Coyne, B.S.M.E., 1954; M.S., 1958; Eng. Sc.D., 1966, Columbia University.—Mr. Coyne has worked on wire-terminal connection reliability, techniques and equipment for burying telephone cable, and systems for automating the main distributing frame. Since 1976, he has been involved in thermal and moisture studies of outdoor loop electronics equipment.

Arik Kashper, M.Sc. (Mathematics), 1969, Leningrad University, USSR; Ph.D. (Systems Engineering), 1979, University of Arizona, Tucson; Bell Laboratories 1979—. At the University of Arizona, Mr. Kashper worked in the area of system identification and parameter estimation. At Bell Laboratories, he is concerned with problems in trunk network engineering. Member, SIAM.

Burton S. Liebesman, B.S. (General Engineering), 1957, United States Naval Academy; M.S. (Electrical Engineering), 1962, New York University; Ph.D. (Operations Research), 1971, New York University; Bell Laboratories, 1960—. In addition to his employment at Bell Laboratories, Mr. Liebesman is also on the staff of Rutgers University Graduate School of Management. His experience at Bell Laboratories has been in all areas of the quality and reliability disciplines. He is currently a member of the Quality Assurance Center, where he is working with a team responsible for evaluating general trade switching systems. Mr. Liebesman has published numerous papers on Military Standard 105D. Member, Operations Research Society of America, Reliability and Statistics Divisions of the American Society for Quality Control.

Sol M. Rocklin, B.S. (Aeronautical Engineering), 1964, Auburn University; M.S. (Aeronautical Engineering), 1967, Princeton University; Ph.D. (Engineering Science), 1973, University of California, Berkeley; National Institute of Health, 1976–1978; Bell Laboratories, 1978—. At National Institute of Health, Mr. Rocklin was a Postdoc-

toral Fellow. When he joined Bell Laboratories he worked on problems in trunk network administration. Since 1978, Mr. Rocklin has been at Lincoln Laboratory, M.I.T., where he works in the area of estimation theory. He is also interested in the application of optimal control theory to problems in population dynamics and evolutionary ecology.

Irwin W. Sandberg, B.E.E., 1955, M.E.E., 1956, and D.E.E., 1958, Polytechnic Institute of Brooklyn; Bell Laboratories, 1958—. Mr. Sandberg has been concerned with analysis of radar systems for military defense, synthesis and analysis of active and time-varying networks, several fundamental studies of properties of nonlinear systems, and with some problems in communication theory and numerical analysis. His more recent interests include compartmental models, the theory of digital filtering, global implicit-function theorems, and functional expansions for nonlinear systems. Former Vice Chairman IEEE Group on Circuit Theory, and Former Guest Editor IEEE Transactions on Circuit Theory Special Issue on Active and Digital Networks. Fellow and member, IEEE; member, American Association for the Advancement of Science, Eta Kappa Nu, Sigma Xi, Tau Beta Pi, National Academy of Engineering.

C. Russell Szelag, B.S. (Mathematics), 1972. Stevens Institute of Technology; M.S., 1974, M.S. (Computer Science), 1977, University of Wisconsin-Milwaukee; Bell Laboratories, 1977—. As a member of the Traffic Network Planning Department, Mr. Szelag has worked on various aspects of trunk network administration for the public telephone network. Member, IEEE.

PAPERS BY BELL LABORATORIES AUTHORS

COMPUTING/MATHEMATICS

Bounded Straight-Line Approximation of Digitized Curves and Lines. C. M. Williams, Computer Graphics and Image Processing, *15* (August 1981), pp 370–81.

ENGINEERING

The Evolution of the Electronic Loop Network in the 1980's. R. W. Wyndrum, Jr., Conf Proc, *2* (September 1981), pp 431–4.
Expediting Insulation Evaluation Through Standardized Test Procedures. R. A. Frantz and J. A. Klatte, Proc 15th Electrical/Electronics Insulation Conf (October 19–22, 1981), pp 336–42.
Experimental Wideband Switching System Capability. R. E. Cardwell and H. R. Lehman, Proc Int Switching Symp 1981, *2* (September 21–25, 1981), pp 1–7.
Forecasting Reliable Systems Into the Future. R. C. Hansen, Proc Natl Elec Conf, XXXV (October 1, 1981), pp 41–3.
The FT3 Lightguide Transmission Medium—Initial Performance Results. M. I. Schwartz, P. F. Gagen, and N. E. Hardwick, III, Third Int Conf Integrated Optics and Optical Fiber Commun—Tech Digest (April 27–29, 1981), pp 12–13.
Lateral Oxidation and Redistribution of Dopants. B. R. Penumalli, *Numerical Analysis of Semiconductor Devices and Integrated Circuits*, Eds. B. T. Browne and J. J. H. Miller, Dublin, Ireland: Boole Press 1981, pp 264–9.
Maintenance, Control, and Protection of Remote Electronics—An Overview. M. A. Schwartz, IEEE Trans Commun, *COM-29*, No. 10 (October 1981), pp 1415–8.
Mechanical Tempering of Optical Fibers. L. Rongved, C. R. Kurkjian, and F. T. Geyling, J Non-Crystalline Solids, *42* (1980), pp 579–84.
Moisture Dependence of the Leakage Current of Tantalum Film Capacitors with Negative Bias. D. O. Melroy, Int J Hybrid Microelect, *4*, No. 2 (October 1981), pp 263–8.
No. 4 ESS—Design and Performance of Reliable Switching Software. P. K. Giloth and J. R. Witsken, Proc 10th Int Switching Symp, *3*, No. 33A-2 (September 1981), pp 1–9.
No. 5 ESS—Hardware Design. H. L. Bosco, R. K. Eisenhart, F. A. Saal, and W. G. Scheerer, Proc Int Switching Symp, *3* (September 1981), pp 31-A-3-1—31-A-3-7.
No. 5 ESS—Software Design. S. M. Bauman, R. J. Carline, J. S. Nowak, and H. Oehring, Proc Intl Switching Symp, *3* (September 21–25, 1981), pp 1–6.
Packet Networks: Layered Protocols, Routing, and Flow Control. J. W. Palmer, Proc NEC 1981, *35* (October 1, 1981), pp 454.
Photodetectors for Optical Fibre Communication. T. P. Pearsall, J Optical Commun, *2*, No. 2 (May 31, 1981), pp 42–8.
Prediction of the Visibility of Asynchronous Gratings by a Single-Channel Model. J. O. Limb, Vision Research, *21* (1981), pp 1409–12.
Sampling Techniques for Determining Fault Coverage in LSI Circuits. V. D. Agrawal, J Digital Systems, *V*, No. 3 (Fall 1981), pp 189–202.
Test Generation for Highly Sequential Scan-Testable Circuits Through Logic Transformation. M. R. Mercer, V. D. Agrawal, and C. M. Roman, 1981 Int Test Conf Digest of Papers (October 1981), pp 561–5.
A Transport Protocol Providing Efficient Datagram and Virtual Circuit Service. B. H. Bharucha, D. E. Butler, and S. Chakrabarti, Proc Int Switching Symp, Session 34B (September 1981) Paper No. 3.

PHYSICAL SCIENCES

Atmospheric Corrosion of Clad Palladium and Palladium-Silver Alloys Part I. Film Growth and Contamination Effects. C. A. Haque and M. Antler, Proc 27th Annual Holm Conf Elec Contacts (September 21, 1981), pp 183–90.
Characteristic Differences in the Surface Space Charge Layer of Red and Green Forms of Lutetium Diphthalocyanine. S. C. Dahlberg, C. B. Reinganum, C. Lundgren, and C. E. Rice, J Electrochem Soc *128*, No. 10 (1981), pp 2150–3.

Charge Distribution in Potassium Graphite. S. B. DiCenzo, G. K. Wertheim, and S. Basu, Phys Rev B, *24,* No. 4 (1981), pp 2270–3.

Conformational Defects and Associated Molecular Motions in Crystalline Poly(Vinylidene Fluoride). A. J. Lovinger, J Appl Phys, *52,* No. 10 (1981), pp 5934–38.

Core-Electron Spectroscopy of Intermediate Valence Systems. G. K. Wertheim, *Valence Fluctuations in Solids,* Eds. L. M. Falicov and W. Hanke, Amsterdam: North Holland, 1981, pp 67–71.

Core Level Photoelectron Spectroscopy. G. K. Wertheim, *Emission and Scattering Techniques,* Ed. P. Day, Dordrecht, Holland: D. Reidel, 1981, pp 61–74.

Development of Stereopsis and Cortical Binocularity in Human Infants: Electrophysiological Evidence. B. Petrig, B. Julesz, W. Kropfl, M. Anliker, and G. Baumgartner, Science, *213* (September 18, 1981), pp 1402–5.

Electron Impact Ionization Cross Sections of F_2 and Cl_2. F. A. Stevie and M. J. Vasile, J Chem Phys, *74,* No. 9 (May 1, 1981), pp 5106–10.

Inhibition of Copper Sulphidation. G. W. Kammlott, C. M. Preece, T. E. Graedel, J. P. Franey, E. N. Kaufmann, and A. Staudinger, Corrosion Sci, *21* (1981), pp 541–5.

In-plane Ordering in Stage Two Lithium Graphite. S. B. DiCenzo, S. Basu, and G. K. Wertheim, Synthetic Metals, *3* (1981), pp 139–45.

Island Formation in Metal Halide Intercalation Compounds. G. K. Wertheim, Solid State Commun, *38* (1981), pp 633–5.

Laser-Induced Melt Dynamics of Si on Silica. M. A. Bosch and R. A Lemons, Phys Rev Lett, *47,* No. 16 (October 19, 1981), pp 1151–5.

The Microspectroscopy of Silicon. D. C. Joy and O. M. Maher, Proc 2nd Oxford Mtg on Microscopy of Semiconductors, Conf Series 60 (November 1981), pp 229–36.

The Morphology and Corrosion Resistance of a Conductive Silvery-Epoxy Paste. J. P. Franey, T. E. Graedel, and G. W. Kammlott, J Matls Sci, *16* (1981), pp 2360–8.

Morphology of Lightly Plasticized PVC. H. E. Bair and P. C. Warren, J Macromol. Sci-Phys., *B20,* No. 3 (1981), pp 381–902.

Non-Boussinesq and Penetrative Convention in a Cylindrical Cell. R. W. Walden and G. Ahlers, J Fluid Mechanics, *109* (1981), pp 89–114.

Nondestructive Concentration Profiling of Fiber Optic Performs by Analysis of Raman Spectra. W. A. Sproson, K. B. Lyons, and J. W. Fleming, J Non-Crystalline Solids, *45,* No. 1 (1981), pp 69–81.

Plasma Silicon Oxide Films on Garnet Substrates: Measurement of Their Thickness and Refractive Index by the Prism Coupling Technique. T. W. Hou and C. J. Mogab, Appl Optics, *20,* No. 8 (September 15, 1981), pp 3184–8.

Reduction of Threading Dislocation in 180-Epitaxial Layers Grown on (001) InP Substrate by Misfit Stresses. S. N. G. Chu, S. Mahajan, K. E. Strege, and W. D. Johnston, Jr., Appl Phys Lett, *38,* No. 10 (May 15, 1981), pp 766–8.

X-ray Photoelectron and Auger Analysis of Thin Fibers. C. C. Chang, J Vac Sci Tech, *18,* No. 2 (March 1981), pp 276–81.

X-ray Photoemission Studies of Superficially Oxidized Cesium Antimonide Photoemitters. C. W. Bates, Jr., P. M. Th. M. van Attekum, G. K. Wertheim, D. N. E. Buchanan, and K. E. Clements, Appl Phys Letts, *38,* No. 5 (March 1, 1981), pp 387–9.

Orthorhombic Phase of Nickel Bromine Boracite $Ni_3B_7O_{13}Br$: Room Temperature Ferroelectric-Ferroelastic Crystal Structure. S. C. Abrahams, J. L. Bernstein, and C. Svensson, J Chem Phys, *74,* No. 4 (August 15, 1981), pp 1912–8.

CONTENTS, MARCH 1982

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 2, February 1982
Printed in U.S.A.

B.S.T.J. BRIEF

Fabrication and Properties of Single-Mode Optical Fiber Exhibiting Low Dispersion, Low Loss, and Tight Mode Confinement Simultaneously

By A. D. PEARSON, P. D. LAZAY, and W. A. REED

(Manuscript received October 5, 1981)

Single-mode fiber with a new index profile design has been fabricated. The design allows the decoupling of bandwidth related factors from considerations which affect curvature-induced losses.

I. INTRODUCTION

In a previous publication, Lazay et al.[1] reported a single-mode fiber design with a new refractive index profile that allows the wavelength of zero total dispersion to be positioned in the vicinity of 1.31 μm, while simultaneously providing low loss and tight mode confinement. This report will describe the fabrication and properties of fibers of this new design.

The design uses a heavily fluorine-doped phosphosilicate cladding and a germania-doped silica core. The core and cladding compositions were chosen to provide a reduced material dispersion that cancels the waveguide dispersion near 1.3 μm. At the same time, the total core-to-clad index difference was chosen to give a delta of 0.5 percent, sufficient to produce a small modal size and excellent resistance to curvature-induced losses. We anticipate that the losses ultimately achievable with this new design will be lower than those of step-index GeO_2–SiO_2 core fibers of the same core diameter, largely because of the reduced contribution from Ge scattering. Ainslie et al.[2] have pointed out the

loss advantage to be gained by the use of a low refractive index cladding and moderately doped core.

Previous single-mode fiber designs which we have studied were based on either a low Δ (~0.20 percent) and large core diameter (~10 μm) or a high Δ (~0.5 percent) and a small core diameter (~7.5 μm). The first class of fibers is capable of having zero dispersion wavelengths near 1.3 μm but has exhibited sensitivity to cabling induced loss because of mode confinement problems. The second class of fiber has exhibited no cabling induced loss but has a zero dispersion wavelength near 1.35 μm and, consequently, the dispersion near 1.3 μm is too large for high data rate (274 mb/s and above) undersea systems having repeater spacings of 30 to 50 km.[3,4]

II. DESIGN

The basic goal is to make a fiber with a lightly GeO_2 doped SiO_2 core such that the material dispersion contribution will cancel the waveguide dispersion at or very near to the system operating wavelength, which is expected to be around 1.3 μm.[4] At the same time, it is imperative to retain the relatively small core (e.g., 7.5 to 8 μm) and large Δ in order to have low bending-induced loss. This is accomplished by providing a phosphosilicate cladding heavily doped with fluorine. The negative Δ^- of the cladding, combined with the positive Δ^+ of the core, both relative to silica, gives the required total Δ value of about 0.5 percent. With a core diameter of 7.5 to 8 μm, this design provides excellent resistance to bending-induced loss, and cutoff wavelengths of around 1.2 μm. Most importantly, since the core and cladding doping levels can be varied independently, the core composition can be adjusted to make the wavelength of zero total dispersion fall at or very close to the projected system operating wavelength.

Phosphorous doping of the cladding was necessary to provide reasonable processing temperatures. Although the P_2O_5 raises the refractive index slightly, the F doping depresses it much more, without substantially affecting the processing temperature. The reduction in refractive index of silicate glasses when fluorine replaces oxygen was discovered by Schott and Abbe in the late nineteenth century,[5] and the use of fluorine doping to reduce the refractive index of silicate compositions in optical fibers has been reported previously.[6,7,8,9]

III. FIBER FABRICATION

The preform was made by Modified Chemical Vapor Deposition (MCVD) in a 19- by 25-mm Heraeus T08-WG silica tube. The delivery rates of the reactants for the cladding ($SiCl_4$, $POCl_3$, and CF_2Cl_2) were chosen to give a glass deposition rate of about 0.35 g/min and a Δ^- of

−0.185 percent. The cladding was deposited in 16 passes without pressurization. The core was deposited in two passes using $SiCl_4$ and $GeCl_4$ with a Δ^+ of +0.315 percent. A compensated collapse procedure was used, but no other chemical drying agents were used. After drawing, the fiber dimensions were OD 114 μm, core diameter 7.5 μm, deposited cladding diameter 44 μm, and length 1 km.

IV. CHARACTERIZATION OF FIBER

Cutoff was determined, using the method described by Lazay,[10] as the location of the rapid drop in power transmitted through a 3-meter length of fiber as the wavelength of the incident light was increased. A well-defined cutoff was located at $\lambda_c = 1.192 \pm .005$ μm.

The loss spectrum was measured from 1.0 to 1.7 μm using the far-end/near-end technique with a 3-meter near-end length. The loss was measured with and without a single 40-mm radius loop in the near-end length. Figure 1 shows the loss curve measured with the loop. Quite surprisingly, the loss curve without the loop was essentially identical, even in the vicinity of cutoff at 1.19 μm. It has been our experience that this is a signature of very good mode confinement. The loss has a local minimum at 1.30 μm of 0.57 ± .03 dB/km, and a minimum loss of 0.40 dB/km at 1.50 μm. Beyond 1.5 μm, the loss rises rapidly and all evidence indicates that the loss is unbounded. This loss "edge" occurs when the effective mode index (the propogation constant divided by $2\pi/\lambda$) falls below the index of the substrate tube. When this happens, the mode becomes cut off because the power can leak through the cladding and be lost by the process of radiation.[11] The wavelength at which the loss rises can be moved to longer wavelengths by increasing the thickness of the deposited cladding. In any case, the loss at 1.3 μm is not adversely affected.

The total chromatic dispersion in the single-mode regime was calculated from the derivative of group delay versus wavelength data.[12] These data were obtained using narrow pulses generated by stimulated Raman scattering in a single-mode fiber pumped with 1.06-μm pulses from a mode-locked Q-switched Nd:YAG laser. The wavelength of the pulses emerging from the Raman fiber was selected with a grating monochromator. Figure 2 shows the spectral dependence of the dispersion. The zero dispersion wavelength, λ_0, is located at 1.312 μm.

V. CONCLUSIONS

We have made a fiber that implements a new design. This design allows dispersion optimization through the manipulation of core and cladding glass compositions, while providing low curvature-induced losses through proper choice of core diameter and Δ. The ability to

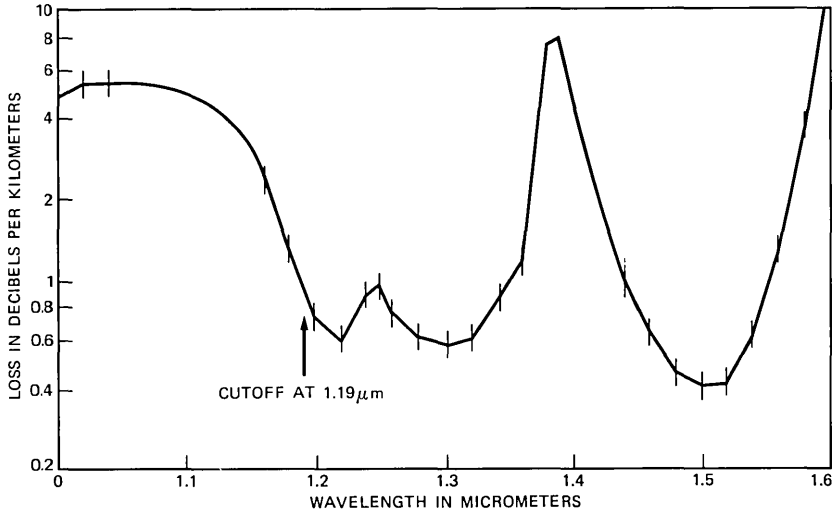


Fig. 1—Loss versus wavelength for depressed index single-mode fiber. The error bars represent 3σ.

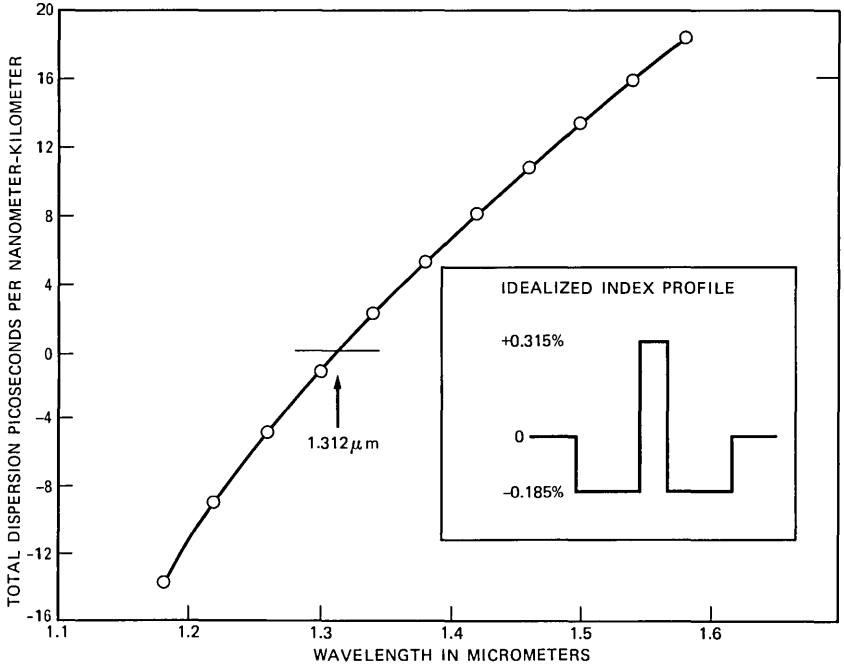


Fig. 2—Total dispersion versus wavelength for depressed index single-mode fiber. The inset shows the idealized refractive index profile.

largely decouple bandwidth design from loss related requirements is a new and unique feature of this design.

It is anticipated that further reduction in loss will be possible. A more complete study of the curvature-induced loss sensitivity of this fiber is underway.

VI. ACKNOWLEDGMENTS

We wish to acknowledge the contribution of F. V. DiMarcello for drawing the fiber and E. A. Sigety for loss measurements. W. T. Anderson provided useful assistance in the bandwidth calculations. Many useful discussions with P. J. Lemaire, and P. A. Christian regarding fluorine doping, contributed to the success of this work.

REFERENCES

1. P. D. Lazay et al., "An Improved Single Mode Fiber Design, Exhibiting Low Loss, High Bandwidth and Tight Mode Confinement Simultaneously," Conf. on Lasers and Electro-optics, 1981, Technical Digest, 1981, Washington, D.C. Postdeadline Paper WG6.
2. B. J. Ainslie et al., "Design and Fabrication of Monomode Fibers for Long-Wavelength Operation," Conf. Integrated Optics and Optical Commun. 1981, Oral Presentation TUC3, 1981, San Francisco.
3. L. G. Cohen et al., "Transmission Studies of a Long Single-Mode Fiber," B.S.T.J., *60* (October 1981), pp. 1713–25.
4. C. D. Anderson et al., "An Undersea Communication System Using Fiberguide Cables," Proc. IEEE, 1980, *68,* pp. 1299–303.
5. F. V. Tooley, editor, "Handbook of Glass Manufacture Vol. II," New York: Ogden Publishing Co., 1960, p. 17.
6. A. Mühlich et al., "A New Doped Synthetic Fused Silica as Bulk Material for Low-Loss Optical Fibers," First Eur. Conf. on Opt. Fiber Comm., 1975, London, Postdeadline paper.
7. W. A. Gambling et al., "Optical Fibers Based on Phosphosilicate Glass," Proc. IEEE, 1976, *123,* pp. 570–6.
8. B. J. Ainslie et al., "Optimized Structure for Preparing Long Ultra-Low-Loss Single-Mode Fibers," Elec. Lett., 1980, *16,* pp. 692–3.
9. J. Irven and A. P. Harrison, "Single-Mode and Multimode Fibers Co-doped with Fluorine," Int. Optical Fiber Commun. Conf. 1981, Technical Digest, 1981, San Francisco, pp. 50–2.
10. P. D. Lazay, "Effect of Curvature on the Cutoff Wavelength of Single Mode Fibers," Symp. on Opt. Fiber Measurements, Technical Digest, 1980, Boulder, Colorado, pp. 93–5.
11. L. G. Cohen, D. Marcuse, and W. L. Mammel, unpublished work.
12. L. G. Cohen and C. Lin, "A Universal Fiber Optic Measurement System Based on a Near-IR Fiber Raman Laser," IEEE J. Quant. Elec., 1978, *QE-14,* pp. 855–9.

THE BELL SYSTEM TECHNICAL JOURNAL is abstracted or indexed by *Abstract Journal in Earthquake Engineering, Applied Mechanics Review, Applied Science & Technology Index, Chemical Abstracts, Computer Abstracts, Current Contents/ Engineering, Technology & Applied Sciences, Current Index to Statistics, Current Papers in Electrical & Electronic Engineering, Current Papers on Computers & Control, Electronics & Communications Abstracts Journal, The Engineering Index, International Aerospace Abstracts, Journal of Current Laser Abstracts, Language and Language Behavioral Abstracts, Mathematical Reviews, Science Abstracts (Series A, Physics Abstracts; Series B, Electrical and Electronic Abstracts; and Series C, Computer & Control Abstracts), Science Citation Index, Sociological Abstracts, Social Welfare, Social Planning and Social Development,* and *Solid State Abstracts Journal.* Reproductions of the Journal by years are available in microform from University Microfilms, 300 N. Zeeb Road, Ann Arbor, Michigan 48106.


Bell System