IFAC

INTERNATIONAL FEDERATION
OF AUTOMATIC CONTROL

WARSZAWA 1969

Stochastic Process in Control and Information Systems

Fourth Congress of the International Federation of Automatic Control
Warszawa 16–21 June 1969

TECHNICAL SESSION
63

NOT

Organized by
Naczelna Organizacja Techniczna w Polsce

INTERNATIONAL FEDERATION OF AUTOMATIC CONTROL

Stochastic Process in Control and Information Systems

TECHNICAL SESSION No 63

FOURTH CONGRESS OF THE INTERNATIONAL
FEDERATION OF AUTOMATIC CONTROL
WARSZAWA 16 – 21 JUNE 1969

Organized by
Naczelna Organizacja Techniczna w Polsce

Contents

Wydawnictwa Czasopism Technicznych NOT
Warszawa, ul. Czackiego 3/5 — Polska

Zakład Poligraficzny WCT NOT. Zam. 17/69.

НАЧАЛА ИНФОРМАЦИОННОЙ ТЕОРИИ УПРАВЛЕНИЯ

Петров Б.Н. ,
Уланов Г.М.
Институт автоматики и телемеханики
Москва

Петров В.В.,
Агеев В.М.,
Запорожец А.В.,
Усков А.С.
Московский авиационный институт
Москва

Кочубиевский И.Л.
Сибирское отделение Академии Наук
Владивосток

С С С Р

I. Необходимость создания информационной теории управления

Кибернетика в широком смысле слова - это наука об информации и управлении в целенаправленно функционирующих динамических системах.

Современная теория управления хорошо описывает лишь относительно простые случаи. Поэтому системный подход к комплексной автоматизации производственных процессов, а также создание многомерных систем требуют развития единой информационной теории сложными динамическими системами.

Вместе с тем, если для систем связи на базе понятия информации создана достаточно общая теория, то в то же время для систем управления это понятие находит пока весьма ограниченное применение.

II. Особенности систем управления

Системы управления представляют собой динамические системы, оперирующие ограниченными ресурсами /энергия, количество вещества и т.п./. Эти свойства элементов и систем нашли свое отражение в отличной от единицы передаточной функции. Движение таких систем описывается переменными, ограниченными по модулю, рассматриваемыми в конечной полосе частот и ограниченном интервале времени. Следовательно, процессы в таких системах представляют собой последовательность взаимосвязанных состояний.

В целенаправленных динамических системах имеют место как детерминированные, так и случайные стационарные и нестационарные сигналы, а также их различные сочетания. В этих условиях возникает общая задача различимости состояний и динамической точности воспроизведения требуемых процессов, а также идентификации элементов и систем. Для решения указанных задач естественно привлечь аппарат теории информации.

Ш. Информационный подход к теории динамических систем

I. Различимость состояния объекта управления

Любой объект представляет собой сложную совокупность в общем случае разнородных элементов. Поэтому состояние такого объекта могут различаться лишь тогда, когда динамические переменные описывающие в целом этот объект отличаются на некоторую величину $\varepsilon > 0$, называемую порогом различимости [I,2] . Данный метод описания поведения динамической системы объединяет свойства непрерывного и дискретного представления. Введение различимости позволяет дать адэкватное описание объектов на различных иерархических уровнях организации и определять предельное количество информации, необходимое для функционирования системы управления.

Выбор порога различимости вводимого в математическую модель динамической системы может также базироваться на необходимой точности исследования отличной от естественных физических свойств объекта. Это позволяет оперировать с минимально необходимым количеством информации для решения поставленной задачи.

Введение порогов различимости и учет ограничений на фазовые переменные динамической системы определяют пороговые свойства всех параметров ее математической модели.

2. Теорема отсчетов при заданной динамической точности

В реальных системах при ограниченных ресурсах всегда имеет место динамическая ошибка, которая приводит к некоторой

потере информации и, как следствие, к снижению требований к точности воспроизведения. Поэтому для динамических систем представляется целесообразным ослабить требования в количестве минимального числа отсчетов по сравнению с известной теоремой отсчетов для динамических систем [3,4] достаточно пердавать значения процесса через интервалы времени:

$$\Delta t = \frac{1}{2(W-K)},$$

где K - уменьшение полосы частот, которое определяется из условия заданной динамической точности:

$$\int_{2\pi(W-2K)}^{2\pi W} S_x(\omega)\, d\omega \leq \pi\delta^2, \qquad (3.1)$$

где $S_x(\omega)$ - спектральная плотность сигнала

δ - предельно допустимая среднеквадратическая ошибка

Из условия $\delta = 0$ следует $K=0$ и мы приходим к условиям теоремы Котельникова.

3. Информационная теория управления базируется на энтропийном описании сложных и многомерных динамических систем [5]. Такой энтропийный подход позволяет дать оценку различных процессов управления.

4. Работа автоматической системы происходит благодаря специально организованной компенсации управлением случайных возмущений. Качество процесса управления зависит от степени этой компенсации.

Основные условия управления на информационном языке и представляют баланс энтропий, выражающий компенсирующее действие управления. В общем виде этот результат можно записать как [1]:

$$H_t(x) = H_t(v) - H_t(v/x) - H_t(z) + H_t(z/x,v). \qquad (3.2)$$

Здесь индексом H_t обозначена динамическая энтропия, характеризующая неопределенность некоторой величины за малый промежуток времени, соответствующий порогу различимости времени.

В случае полной компенсации достигается полная инвариантность системы управления. При этом уравнение баланса энтропий управляющего и возмущающего воздействий записывается в виде

$$H_t(v) - H_t(v/x) = H_t(z) + H_t(z/x, v). \qquad (3.8)$$

При небольших отклонениях от полного баланса (3.8) выполняется условия инвариантности до ε [2,6,7,8] .

Указанные информационные условия являются необходимыми для целенаправленного функционирования динамических систем.

IУ. Вопросы информационной теории управления и контроля

Рассмотрим результаты некоторых разработок в этой области, отражающие специфику систем управления, отмеченную выше.

I. Введение меры количества разнообразия динамической системы

Необходимость введния этой меры связана с ограничениями применения энтропии для оценки процессов управления.

Здесь лишь представляется возможным дать основную идею этой меры, показав это на примере непрерывной модели.

Пусть $X(t)$- произвольный /в общем случае нестационарный/процесс управления, заданный на произведении пространств $X \otimes T$. Представим этот процесс в виде:

$$X(t) = f(t) + \dot{X}(t), \qquad (4.1)$$

где $f(t)$- функция математического ожидания процесса $X(t)$, определяющая распределение значений $f \in F$ по области определения T и $\dot{X}(t)$- центрированный случайный процесс, определенный вероятностным распределением значений $\dot{x} \in \dot{X}$ для каждого $t \in T$.

Введем оценку распределения значений процесса $X(t)$ на $X \otimes T$ - меру разнообразия множества значений процесса, единую для детерминированных и случайных функций.

Разумно потребовать, чтобы эта мера разнообразия обладала свойством аддитивности и не зависела от конкретных значений процесса и масштабов, а учитывала только характер распределения значений процесса, т.е. обладала свойством, подобным свойствам энтропии случайных величин /процессов/ в теории информации.

Если задана плотность распределения вероятностей $P_t(\dot{x})$ значений $\dot{x} \in \dot{X}$ непрерывного процесса $\dot{x}(t)$ для каждого $t \in T$, то динамическая энтропия этого процесса /т.е. энтропия в момент t [I] записывается в виде

$$H_t(\dot{x}) = -\int_{\dot{X}} P_t(\dot{x}) \log [\varepsilon_x P_t(\dot{x})] dx, \tag{4.2}$$

где ε_x - порог различимости значений, определенной на множестве X .

Полагая, что наличие постоянного смещения не меняет вида распределений, имеем:

$$P_t(\dot{x}) = P_t(x),$$

$$H_t(\dot{X}) = -\int_X P_t(x) \log [\varepsilon_x P(x)] dx = H_t(x) \tag{4.3}$$

Следовательно, динамическая энтропия представляет собой меру разнообразия значений $x \in X$ процесса $x(t)$ только на X для каждого $t \in T$ и не учитывает функцию математического ожидания $f(t)$ т.е. важной динамической характеристики процесса $x(t)$.

Поэтому для полной меры разнообразия распределения значений процесса $x(t)$ на $X \otimes T$ необходимо ввести характеристику распределения значений процесса на T, определяемую $f(t)$.

Для возможности учета только характера распределения значений процесса на T вне зависимости от его масштаба и физического характера будем рассматривать нормированный процесс

$$\frac{x(t)}{X_{max}} = \frac{f(t)}{X_{max}} + \frac{\dot{x}(t)}{X_{max}} \tag{4.4}$$

Введем в качестве характеристики распределения значений непрерывный нормированной функции математического ожидания плотность распределения ее значений на T в виде

$$f^*(t) = \frac{|f'_t(t)|}{|X_{max}|} \tag{4.5}$$

$f^*(t)$ - характеризует интенсивность изменения значений процесса во времени.

Тогда плотность распределения значений процесса на $X \otimes T$ можно представить в виде

$$\mu(x,t) = f^*(t)\, P_t(x). \qquad (4.6)$$

Введем в качестве полной характеристики распределения значений процесса $X(t)$ в момент времени $t \in T$ динамическое разнообразие $R_t(X)$

$$R_t(X) = -\int \mu(x,t) \log[\varepsilon_x \varepsilon_t \mu(x,t)]\, dx = \\ = R_t(F) + f^*(t) H_t(X), \qquad (4.7)$$

где

$$R_t(F) = -f^*(t) \log[\varepsilon_t f^*(t)] \qquad (4.8)$$

есть динамическое разнообразие детерминированной составляющей /функции математического ожидания/ процесса $X(t)$ в момент времени $t \in T$.

$H_t(X)$ - динамическая энтропия определения в соответствии с (4.3)

$f^*(t) H_t(X)$ - динамическое разнообразие процесса $X(t)$ на $t \in T$ в момент

ε_t - порог различимости времени.

Введем в качестве полной меры разнообразия значений процесса $X(t)$ на $X \otimes T$ функционал $R(X,T)$ в виде

$$R(X,T) = -\int_T \int_X \mu(\bar{x},t) \log[\varepsilon_x \varepsilon_t \mu(x,t)]\, dx\, dt = \\ = R(F,T) + E_f H_t(X) \qquad (4.9)$$

где:

$$R(F,T) = -\int_T f^*(t) \log[\varepsilon_t f^*(t)]\, dt. \qquad (4.10)$$

Мера разнообразия детерминированной составляющей /функции математического ожидания/ процесса $X(t)$ на T

$$E_f H_t(X) = \int_T f^*(t) H_t(X)\, dt - \qquad (4.11)$$

- энтропия процесса $X(t)$, усредненная на T с учетом средней интенсивности изменения процесса $f^*(t)$.

Таким образом, предложенная мера разнообразия произвольного процесса управления представляет собой сумму меры разно-

образия детерминированной составляющей процесса и динамической энтропии этого процесса, усредненной с учетом интенсивности изменения детерминированной составляющей.

Предложенная оценка может служить основой для информационного анализа систем управления и контроля.

Детальное рассмотрение свойств разнообразия может быть предметом несостоятельного доклада.

2. Потенциальная характеристика элементов и систем управления

Одним из основных вопросов информационной теории является "что можно и чего нельзя достичь в автоматической системе, каковы ее потенциальные возможности".

Решение этого вопроса должно базироваться на фундаментальном понятии теории - потенциальной характеристике.

Величину характеризующие предельные динамические свойства будем называть потенциальной характеристикой элемента /системы/. В нашем случае

$$C(X) = \max R(X), \qquad (4.12)$$

где максимум рассматривается по всем возможным значениям воздействий $y(t)$, вызывающим процесс $X(t)$.

Само понятие потенциальной характеристики для автоматических систем в значительной степени относительно.

Прежде всего оно зависит от режима работы системы, критерия качества, и от того, что принято за "вход" и "выход".

Наиболее характерной для автоматических систем будет динамическая потенциальная характеристика, отражающая свойства системы в момент t или точнее на интервале, равном порогу различимости времени ε_t.

Введенное понятие потенциальной характеристики позволяет сформулировать и доказать для динамических систем основную теорему, соответствующую теореме Шеннона в теории информации [9].

Общая формулировка такой теоремы для динамических систем может быть представлена в следующей форме:

Если на входе объекта при различимости ε входное воздействие $y(t)$ имеет динамическое разнообразие $R_t(y)$, то существует возможность получить на выходе этого же объекта такое же разнообразие состояний $X(t)$, $[R_t(X)=R_t(Y)]$ индуцируемое этим же воздействием $y(t)$ если

$$R_t(Y) \leq C_t(X) \tag{4.13}$$

и это невозможно, если

$$R_t(Y) > C_t(X). \tag{4.14}$$

В предлагаемой модели эта теорема распространяется на произвольные процессы управления, в том числе и на детерминированные и нестационарные. Действительно, расписав выражение динамического разнообразия (4.7) по составляющим, получим вместо (4.13) выражение

$$R_t(F_y) + \int_y^x (t)\, H_t(Y) \leq C_t(X) \tag{4.15}$$

из которого следует, что возможности передачи обеих составляющих ограничены одной и той же потенциальной характеристикой,- поэтому увеличение разнообразия одной из составляющих возможно только за счет уменьшения разнообразия другой. Следовательно, в пределе как детерминированное воздействие, так и случайное может передаваться в границах одной и той же потенциальной характеристики.

Если в системе имеются помехи и искажения, то их разнообразие в динамике может быть учтено соответствующими аддитивными членами, не изменяя существо приведенной выше формулировки основной теоремы.

Для различных автоматических систем основная теорема может иметь различные формулировки такой теоремы для систем стабилизации приведена в [I].

Следует обратить внимание, что в формулировках основной теоремы применяются ε-оценки, связанные с порогом различимости или вообще с некоторой величиной ε , характеризующей динамическую точность. Разнообразия являются убывающей функцией от ε .

Таким образом, динамическая потенциальная характеристика является предельной характеристикой, ограничивающей выбор

между интенсивностью воздействия и точность его воспроизведения

Итак, всякое управление в своих возможностях ограничено потенциальной характеристикой. Именно эта величина отвечает на основной вопрос, что можно и чего нельзя достичь в системах управления.

3. Мера энтропийной устойчивости процессов управления

Условия энтропийной и информационной устойчивости можно рассматривать, как критерий определенности протекания процессов управления [10, 11, 12] .

Особенностью энтропийного представления большой совокупности событий или процессов является выделение из нее высоковероятной группы, что позволяет при анализе и расчете реализовать реальные режимы работы.

Именно это свойство групповой энтропии и информации должно позволить выделить характерные свойства динамических систем при их информационном описании.

В этих условиях представляется целесообразным ввести меру определенности протекания процесса..В случае неуправляемой системы наиболее неблагоприятным случаем является равномерное распределение и динамическая энтропия, при этом определяется как

$$H_t = \log \frac{|X|}{\varepsilon_x}$$

для $t \in T$.

Естественно, возникает вопрос, на сколько будет отличаться энтропийная оценка, если объект является управляемым и распределение становится отличным от равномерного.

Для решения этого вопроса, наряду с понятием энтропии как математического ожидания энтропийной плотности $[-\log \varepsilon_x P(x)]$ целесообразно рассмотреть ее дисперсию, т.е.

$$D_{H_t} = M\left[-\log \varepsilon_x P(x) - H_t(x)\right]^2.$$

Тогда

$$D_{H_t} = \int_{-\infty}^{\infty} P_t(x) \log^2 \left[P_t(x)\right] dx + H_0^2(x)\Big],$$

где $H_0(x)$- дифференциальная энтропия. Из последнего выражения следует, что величина D_H не зависит от шага квантования по

уровню, определяется только функцией распределения случайной величины и являются ограниченной величиной, в то время как абсолютная энтропия непрерывной случайной величины стремится к бесконечности при неограниченном уменьшении шага квантования по уровню.

Можно показать, что при равномерном распределении $D_H = 0$ величина D_H может характеризовать степень определенности протекания процессов управления.

Увеличение D_H отражает роль высоковероятных состояний управляемого процесса и может использоваться как мера информационной устойчивости.

4. Вопросы фильтрации сигналов в ограниченной полосе частот

В некоторых задачах САУ и контроля представляется целесообразным точно воспроизводить сигнал лишь в части полосы частот W_1 .

/$W_1 = W - K$; W - полоса частот сигнала/

Для этой цели согласно обобщенной теореме отсчетов [4] достаточно производить измерения через интервал

$$\Delta t = \frac{1}{2(W-K)} \qquad /W, K\text{- в герцах/}$$

При этом функция отсчетов принимает вид

$$U(t) = \frac{\sin[2\pi(W-2K)t]}{2\pi(W-K)t} .$$

В реальных устройствах на полезный сигнал накладывается помеха.

При наличии высокочастотной помехи оптимальный прямоугольный фильтр найдем из условия минимума С.К.О.

Для сигналов, ограниченных по частоте, выражение С.К.О. имеет вид

$$\bar{\varepsilon}^2 = \frac{1}{\pi}\left[\int_{2\pi(W-K^*)}^{2\pi W} S_m(\omega)\,d\omega + \int_{0}^{+2\pi(W-K^*)} S_n(\omega)\,d\omega\right] \qquad (4.16)$$

Минимизируя выражение /4.16 по параметру K , имеем

$$S_m(W-K^*) = S_n(W-K^*). \qquad (4.17)$$

Соотношение (4.I7) может рассматриваться так же, как и обобщение теоремы Котельникова В.А. на случай, когда на полезный сигнал наложена помеха[x/].

Оценочный фильтр /фиг.I/ определяется из условия

$$\pi\delta^2_{max\,max} \geqslant \int_{W-K}^{W} S_m(\omega)\,d\omega + \int_{0}^{W-K} S_n(\omega)\,d\omega .$$

Для реализации оценочного фильтра с полосой $W-K$ может быть использован метод интегральной квадратичной аппроксимации прямоугольной характеристики выражением вида

$$\Phi(i\omega) = K\,\frac{\beta_m(i\omega)^m + \beta_{m-1}(i\omega)^{m-1} + \ldots + \beta_1 i\omega + 1}{\alpha_n(i\omega)^n + \alpha_{n-1}(i\omega)^{n-1} + \ldots + \alpha_1 i\omega + 1} .$$

Описанный метод не связан с проблемой регуляризации и позволяет синтезировать фильтры при заданной динамической точности с минимальной полосой пропускания.

При ограничении полосы пропускания оценочной величиной $W-K$ и точно вычисленных спектральных плотностях полезного сигнала и помехи задача синтеза может быть сведена к решению задачи Колмогорова-Винера:

$$\Phi(i\omega) = \frac{1}{2\pi\psi(i\omega)} \int_0^{\infty} e^{-i\omega t} \int_{-\infty}^{\infty} \frac{\Phi_0(i\omega)\,S_m(\omega)}{\psi^*(i\omega)}\, e^{i\omega t}\,d\omega , \qquad (4.18)$$

где $\psi(i\omega)$, $\psi^*(i\omega)$- комплексно-сопряженные множители, не имеющие ни нулей, ни полюсов соответственно в нижней и в верхней полуплоскостях комплексной плоскости ω .

$$\psi(i\omega)\psi^*(i\omega) = S_m(\omega) + S_n(\omega)$$

$\Phi_0(i\omega)$ - идеальный оператор и

$$|\Phi_0(i\omega)| = \begin{cases} e^{-\alpha[\omega - W + K]}, & \omega \geq (W-K)2\pi \\ 1, & |\omega| \leq (W-K)2\pi \\ e^{\alpha(\omega + W - K)}, & \omega \leq -2\pi(W-K) \end{cases} \qquad (4.19)$$

---

x/ Для сигналов с неограниченным спектром выражение [4.I7] имеет вид $S_m(W^*) = S_n(W^*)$, где W^* - полоса частот прямоугольного фильтра.

Полагая в выражении (4.I9) $\alpha > \alpha^*$ где α^* достаточно большое число, получим семейство операторов близких к оценочному /фиг.2/.

При $\alpha = 0$ формула /4.I8/ дает характеристику оптимального фильтра. Полагая в выражении (4.I9) $\alpha = 0$ и заменяя в первой интеграле (4.I8) 0 на ∞ получим физически нереализуемую оптимальную передаточную функцию

$$\Phi(i\omega) = \frac{S_m(\omega)}{S_m(\omega) + S_n(\omega)} . \qquad (4.20)$$

Соответствующая величина С.К.О. определяется формулой

$$\bar{\varepsilon}^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{S_m(\omega)\, S_n(\omega)}{S_m(\omega) + S_n(\omega)} d\omega . \qquad (4.2I)$$

Выражением (4.2I) может использоваться в качестве оценочной формулы при проектировании оптимальных фильтров. Формулы (4.20) , . , (4.2I) имеют принципиальное значение. Используя выражения (4.20) ,(4.2I) , нетрудно показать связь результатов Шеннона с результатами, полученными методом статистической оптимизации фильтров.

Рассматривая с этой целью ошибку в классе информационных сигналов, т.е. ограничивая ее приближенно полосой частот спектра W и протяженностью по времени T , получим энтропию ошибки в полосе Δf .

$$H_{\Delta f} = T \Delta f \log 2\pi e S_\varepsilon(f) \Delta f . \qquad (4.22)$$

Интегрируя (4.22) по полосе и используя выражение (4.2I) , получим энтропию, приходящуюся на одну степень сигнала ошибки

$$H_n(\varepsilon) = H_n(m) - \frac{1}{W} \int_W \log \frac{S_m(f) + S_n(f)}{S_n(f)} df \qquad (4.23)$$

Интеграл в правой части совпадает с Шенноновским выражением максимальной скорости передачи информации по каналу с полосой W .

Тот же результат может быть получен, используя формулу Шеннона для потери энтропии в линейном фильтре.

Таким образом получено соответствие между формулой Шеннона и оптимальным фильтром Колмогорова-Винера. Поскольку

при отыскании последнего условие физической осуществимости не учитывалось, максимальная скорость передачи информации по Шеннону не реализуема. То есть формула Шеннона дает завышенную оценку максимальной скорости передачи информации.

5. Количество информации и переходные процессы.

Оценим изменение количества информац ии при больших ошибках системы, имеющих место в переходных режимах.

Пусть задан ансамбль сигналов $x(t)$, принимающий значения на ограниченном множестве с метрикой $\rho(x_1, x_2)=(x_1-x_2)$, который в момент времени $t=0$ поступает на вход линейной динамической системы с импульсной переходной функцией $u(t)$. Начальные условия предполагаются нулевыми. На входе системы - ансамбль $y(t)$ с метрикой $\rho(y_1, y_2)=(y_1-y_2)$. Пусть сигналы $x(t)$ и $y(t)$ могут быть представлены следующим образом

$$x_{i+1}(t) = x_i(t) + \Delta x\, \xi(t),$$

$$y_{i+1}(t) = y_i(t) + \Delta y(t)$$

где $$\Delta y(t) = \Delta x \int_{-\infty}^{t} \xi(t-\tau)\, u(\tau)\, d\tau .$$

В любой момент времени

$$\lim_{\Delta y \to 0} \left(\frac{\Delta x}{\Delta y}\right)_t = \left(\frac{\partial x}{\partial y}\right)_t = J_t\left(\frac{x}{y}\right) = \frac{1}{\int_{-\infty}^{t} \xi(t-\tau)\, u(\tau) d\tau} = K(t)$$

где $J_t\left(\frac{x}{y}\right)$ - якобиан преобразования координат в момент времени t.

В теории информации, сведения о некоторой случайной величине X, получаемые в результате наблюдения случайной величины Y, изменяют ее неопределенность. Последнее характеризуется заменой безусловной энтропии величины X средней условной энтропией величины X относительно величины Y. С учетом линейности преобразования, количество информации в любой моемент в времени описывается следующим соотношением:

$$J_t(Y, X) = H_t(X) - \log K(t).$$

Необходимо отметить, что при изменении случайных величин функция $K(t)$ определяется просто:

$$K(t) = \frac{1}{\int^{t} u(\tau)\, d\tau} = \frac{1}{A(t)} .$$

Определив любым из известных способов переходную функцию системы $A(t)$, можно вычислить количественное изменение информации в переходном процессе.

Переходя к стационарным случайным сигналам с ограниченной полосой частот F, обладающих информационностью в любой момент времени $H(x)$, будем иметь:

$$J_t(x,t) = H(x) + \frac{1}{2F}\int_0^F \log |W(f)|^2 df - \log K(t) + \log \frac{\Delta x}{\Delta y}.$$

Получение исходных соотношений для полученного выражения в общем случае затруднительно. Однако в некоторых частных случаях /гауссовские процессы, фильтр описывается дифференциальным уравнением первого порядка/ эта задача решается относительно просто.

У. Перспективы и задачи информационной теории

Применение положений классической теории информации к задачам управления встречает значительные трудности. Для их устранения необходимо дальнейшее развитие основных идей теории информации в установлении количественных соотношений между потерями информации и точностью ее воспроизведения в динамических системах.

Основной вопрос, который должен быть решен для того, чтобы иметь информационную теорию управления и контроля - это решение задачи передачи разнообразия в системах с различными типами обратных связей и особенно в многосвязанных и многомерных системах.

Второй, не менее важной, проблемой является изучение систем с разветвленной иерархической структурой и приоритетом команд.

Третья неотложная задача, которая возникла сегодня в связи с необходимостью внедрения в практику результатов информационной теории управления, состоит в разработке вычислительных методов информационного анализа и синтеза сложных многомерных систем с многоярусной структурой.

Информационный подход позволяет с единой теоретической точки зрения рассматривать комплекс систем измерения управ-

ления и контроля в независимости от их назначения и способов реализации, что позволит уже на ранней стадии проектирования принимать научно обоснованные решения в тех случаях, где до сих пор господствуют опыт и интуиция.

Литература

I.. Петров Б.Н., Кочубиевский И.Д., Уланов Г.М., Информационные аспекты управления технологическими процессами. Изд. АН СССР."Техническая кибернетика", № 9,I967 г.

2. Петров Б.Н., Кочубиевский И.Д., Уланов Г.М.., Дудин Е.Б.. Различимость, инвариантность и информация в системах с жесткой и переменной структурой. В ст. "Многосязанные и инвариантные системы". Изд. "Наука", I968 г.

3. Петров В.В., Запорожец А.В. Информационная оценка динамической точности систем информации и управления.. Оптимальные системы. Статистические методы. Изд. "Наука". I967.

4. Петров В.В. Оценка динамической точности информационных систем управления. В кн "Современные методы проектирования САУ". М., Машгиз, I967 г.

5. Красовский А.А. Изменение энтропии непрерывных динамических систем. Изд. АН СССР "Техническая кибернетика", № 5, I964 г.

6. Петров Б.Н. "Принцип инвариантности и условия его применения при расчете линейных и нелинейных систем". Тр. I-го Международного конгресса ИФАК. М., I96I.

7. Петров В.В., Агеев В.М., Запорожец А.В. "Некоторые вопросы связи ξ- инвариантности и динамической точности систем автоматического управления". Труды III Всесоюзного совещания по теории инвариантности и ее применению в автоматических устройствах. Киев, I966 г.

8. Кочубиевский И.Д., Уланов Г.М. Информационные методы в теории инвариантности. Труды III Всесоюзного совещания по теории инвариантности и ее применению в автоматических устройствах. Киев, I966 г.

9. Шеннон К. Математическая теория связи. В кн. "Работы по теории информации в кибернетике". ИЛ.I963 г.

I0. Добрушин Р.А. Общая формулировка основной теоремы Шеннона в теории информации У.М.Н., 6/90/, I959 г.

II. Пинскер И.С. Информация и информационная устойчивость случайных величин и процессов.Изд.АН СССР, М. I960 г.

I2. Красовский А.А. Энтропийная устойчивость линейных непрерывных систем автоматического управления. Изв. АН СССР "Техническая кибернетика", № 5, I963 г.

MONTE CARLO TECHNIQUES FOR PREDICTION AND FILTERING OF NONLINEAR STOCHASTIC PROCESSES

J.E. Handschin (Switzerland)*
Centre for Computing and Automation
Imperial College of Science and Technology
University of London
London, S.W. 7. (U.K.)

1. Introduction

The object of this paper is to establish Monte Carlo techniques for the state estimation of nonlinear, discrete-time dynamical systems. In Section 2 we define the mathematical model of the problems considered. The prediction problem is discussed in Section 3. This section also serves to introduce the basic concepts of Monte Carlo work. Two variance reduction methods are derived in order to increase the efficiency of the Crude Monte Carlo estimator.

The importance of nonlinear filtering problems (Section 4) is reflected in the various contributions in this field during the last few years. Most of this work has been devoted to continuous-time systems. For a survey see Fisher[1] or Schwartz[2]. The main contributions for discrete-time systems are due to Cox[3] and Sorenson[4] in developing the approximate nonlinear filter equations.

The new feature of the approach presented here is that sampling methods are introduced which enable us to estimate parameters of probability density functions (p.d.f.), such as the density $p(\underline{x}_k)$ in the prediction problem and the posterior density $p(\underline{x}_k | \underline{y}^k)$ in the filtering problem.

Monte Carlo techniques provide stochastic solutions to the nonlinear filtering and prediction problems. There-

* This research was supported by SA Brown, Boveri & Co. CH-5401 Baden, Switzerland.

fore all results are estimates whose errors, in turn, can be estimated from their sampling variances. The sampling schemes can be improved by using variance reduction techniques. This permits the reduction of the sampling error to a value lower than that obtained by approximate nonlinear filters.

2. Problem Formulation

In this paper, k-stage, time-discrete systems are considered. The state $\underline{x}_k$ of the dynamical system evolves according to the following nonlinear stochastic difference equation:

$$\underline{x}_{k+1} = \underline{f}_k (\underline{x}_k, \underline{w}_k, k) \tag{2.1}$$

$\underline{x}_k$ is an n-dimensional state, and $\underline{w}_k$ is a $p \leq n$ dimensional disturbance vector. The explicit dependence of $\underline{f}_k(.,.,.)$ on the time parameter k accounts for any known input, e.g. control signals. The sequence $\underline{w}^k \triangleq \underline{w}_1, \underline{w}_2, \ldots, \underline{w}_k$ is assumed to be a white noise sequence with a known p.d.f.

$$p(\underline{w}^k) = \prod_{i=1}^{k} p(\underline{w}_i) \tag{2.2}$$

The initial condition of (2.1) is given as a p.d.f. $p(\underline{x}_1)$. The random variable $\underline{x}_1$ is uncorrelated with any other disturbance acting on the system.

The states $\underline{x}_k$ of eqn. (2.1) are observed through the m-dimensional observation vectors $\underline{y}_k$, which are functionally related to $\underline{x}_k$, and which contain random errors. The nonlinear transformation is assumed to be given by

$$\underline{y}_k = \underline{g}_k (\underline{x}_k, k) + \underline{v}_k \tag{2.3}$$

The m-dimensional noise vector $\underline{v}_k$ is supposed to be a member of a white noise sequence with known p.d.f. $p(\underline{v}_k)$ and uncorrelated with $\underline{w}_k$. The variance of $\underline{v}_k$ is denoted by Σ_v.

The problems considered in this paper are concerned with the determination of the state $\underline{x}_k$.

1) In the prediction problem, where we are given the initial density $p(\underline{x}_1)$ and the noise density $p(\underline{w}_k)$, it is desired to estimate the parameters of the p.d.f. $p(\underline{x}_k)$.

2) In the filtering problem, where we are given in addition to the information of 1) a sequence of observations $\underline{y}^k$, it is desired to estimate parameters of the posterior p.d.f. $p(\underline{x}_k \mid \underline{y}^k)$.

3. Nonlinear Prediction

Under the assumptions made in Section 2, the solution of the prediction problem is given by the Chapman-Kolmogorov equation[5]:

$$p(\underline{x}_k \mid \underline{x}_j) = \int_{-\infty}^{+\infty} p(\underline{x}_k \mid \underline{x}_{k-1}) \cdot p(\underline{x}_{k-1} \mid \underline{x}_j)\, d\,\underline{x}_{k-1} \quad (3.1)$$

where $j \leq k-2$. The integral appearing in eqn. (3.1), is a condensed form of the n-fold integral over all elements of the vector $d\underline{x}_{k-1}$.

To avoid the evaluation of eqn. (3.1) using an approximation method, Monte Carlo procedures are employed to estimate parameters of the p.d.f. $p(\underline{x}_k)$, i.e. the p.d.f. of the state $\underline{x}_k$, k-1 steps ahead.

3.1. The Crude Monte Carlo Predictor

Most applications of Monte Carlo techniques are concerned with the evaluation of an integral. That is, Monte Carlo methods are better suited to finding some parameters of a function, rather than the entire function itself. The first order moment $\underline{m}_k$, i.e. the mean, of the p.d.f. $p(\underline{x}_k)$ at time k is defined as

$$\underline{m}_k = E[\underline{x}_k] \triangleq \int \underline{x}_k \cdot p(\underline{x}_k)\, d\,\underline{x}_k \quad (3.2)$$

By the strong law of large numbers[5], the random vector

$$\hat{\underline{m}}_k = N^{-1} \sum_{j=1}^{N} (\underline{x}_k)_j \quad (3.3)$$

converges with probability one to $\underline{m}_k$, if N samples $(\underline{x}_k)_j$, $j=1,2,\ldots,N$, are drawn from the p.d.f. $p(\underline{x}_k)$ and $N \rightarrow \infty$. By the central limit theorem[5], the sampling distribution of $\hat{\underline{m}}_k$ tends to a normal distribution as $N \rightarrow \infty$, and therefore approximate confidence limits can be given for the estimate $\hat{\underline{m}}_k$. The covariance matrix V associated with the random vector $\hat{\underline{m}}_k$ is:

$$V = E\left[(\hat{\underline{m}}_k - E[\underline{x}_k])(\hat{\underline{m}}_k - E[\underline{x}_k])^T\right] \tag{3.4}$$

Since the true moment $E[\underline{x}_k]$ is unknown we must use eqn. (3.5) as an estimate of V:

$$\hat{V} \doteq N^{-2} \sum_{j=1}^{N} [(\underline{x}_k)_j - \hat{\underline{m}}_k][(\underline{x}_k)_j - \hat{\underline{m}}_k]^T \tag{3.5}$$

The confidence limits are given in terms of the sampling error which is defined as $(V)^{\frac{1}{2}}$.

The final problem to be solved is drawing the random sample $(\underline{x}_k)_j$, $j=1,2,\ldots,N$, from the unknown p.d.f. $p(\underline{x}_k)$. The solution is given by direct simulation. $(\underline{x}_k)_j$ denotes the value of $\underline{x}_k$ obtained by simulation of the original nonlinear system (2.1). $(\underline{x}_1)_j$ is drawn from the given initial condition density $p(\underline{x}_1)$, $(\underline{w}^{k-1})_j$ are drawn from $p(\underline{w}^{k-1})$, and $(\underline{x}_k)_j$ is obtained as the solution of eqn. (2.1).

We will refer to a predictor based on eqn. (3.3) as a Crude Monte Carlo estimator. In the sequel, variance reduction methods are introduced to improve the Crude Monte Carlo estimator.

3.2. Variance Reduction Techniques

There are two basic methods for improving Monte Carlo estimates:

1) Change the sampling experiment, as described in Section 3.2.1.
2) Replace part of the sampling experiment by an analytical method, as described in Section 3.2.2.

Although there are many different ways for implementing

one or the other of these two alternatives, not all of them are applicable to the prediction problem.

3.2.1. Antithetic Variate Method

The principle of this method, introduced by Hammersley[6], requires the assumption that the p.d.f., from which samples are to be drawn, must be unimodal and symmetric. In addition to Section 2, we assume the initial condition is given by a normal distribution, denoted by:

$$p(\underline{x}_1) = n(\underline{x}_1 ; \underline{x}_m, \Sigma_x) \tag{3.6}$$

where $\underline{x}_m$ is the mean and Σ_x the variance. The noise p.d.f. $p(\underline{w}_k)$ is given by:

$$p(\underline{w}_k) = n(\underline{w}_k ; 0 , \Sigma_w) \tag{3.7}$$

In addition to the original sample $(\underline{x}_k)^+_j$, j=1,2,...N, a negatively correlated sample $(\underline{x}_k)^-_j$ is generated as the solution of eqn. (2.1), using the antithetic initial condition variate:

$$(\underline{x}_1)^-_j = 2\,\underline{x}_m - (\underline{x}_1)^+_j \tag{3.8}$$

The antithetic noise sequence $(\underline{w}^{k-1})^-_j$ is generated by:

$$(\underline{w}^{k-1})^-_j = - (\underline{w}^{k-1})^+_j \tag{3.9}$$

Finally, the first order moment $\underline{m}_k$ is estimated as:

$$\hat{\underline{m}}_k = (2N)^{-1} \sum_{j=1}^{N} (\underline{x}_k)^+_j + (\underline{x}_k)^-_j = \tfrac{1}{2} (\hat{\underline{m}}^+_k + \hat{\underline{m}}^-_k) \tag{3.10}$$

where

$$\hat{\underline{m}}^{\pm}_k = N^{-1} \sum_{j=1}^{N} (\underline{x}_k)^{\pm}_j \tag{3.11}$$

Since the two estimates $\hat{\underline{m}}^+_k$ and $\hat{\underline{m}}^-_k$ are correlated, the sampling covariance matrix

$$\mathrm{var}(\hat{\underline{m}}_k) = \tfrac{1}{4} \mathrm{var}(\hat{\underline{m}}^+_k) + \tfrac{1}{4} \mathrm{var}(\hat{\underline{m}}^-_k) + \tfrac{1}{2} \mathrm{cov}(\hat{\underline{m}}^+_k, \hat{\underline{m}}^-_k) \tag{3.12}$$

can be made smaller than $\mathrm{var}(\hat{\underline{m}}_k^+)$ of the crude estimator. This is because $(\underline{x}_k)_j^+$ and $(\underline{x}_k)_j^-$, being negatively correlated, yield $\mathrm{cov}(\hat{\underline{m}}_k^+, \hat{\underline{m}}_k^-) < 0$.

A useful property of an efficient Monte Carlo method is to have zero sampling variance, when applied to a linear system with additive Gaussian white noise $\underline{w}_k$. Indeed, for

$$\underline{x}_{k+1} = A_k \underline{x}_k + \underline{w}_k \tag{3.13}$$

an antithetic pair $(\underline{x}_1, \underline{w}^{k-1})_j^{\pm}$ yields:

$$(\underline{x}_k)_j^{\pm} = \prod_{i=1}^{k-1} A_i \, (\underline{x}_1)_j^{\pm} + \sum_{i=1}^{k-1} \prod_{r=i+1}^{k-1} A_r \, (\underline{w}_i)_j^{\pm} \tag{3.14}$$

Using eqn. (3.8) and (3.9), the first order moment $\underline{m}_k$ is estimated with eqn. (3.10) as:

$$\hat{\underline{m}}_k = \prod_{i=1}^{k-1} A_i \, \underline{x}_m \tag{3.15}$$

There is no randomness in this estimator (3.15), and thus the following result is established:

> The Antithetic Variate method yields an estimator for the mean $\underline{m}_k$, with zero sampling variance, when applied to a linear system with Gaussian noise (3.13).

This result holds for the first order moment only. The Antithetic Variate method (3.10) does not give zero sampling variance for estimates of higher order moments, even in the linear Gaussian case.

3.3.2. The Control Variate Method

In the Control Variate method part of the sampling procedure is replaced by an analytical method. A new two stage estimator is derived below for the nonlinear prediction problem. In order to keep the notation simple, the discussion is restricted to a scalar, nonlinear system of the form:

$$x_{k+1} = f_k (x_k, w_k, k) \tag{3.16}$$

The extension to the multidimensional case is discussed

in reference[7].

The first stage of the Control Variate estimator is concerned with the determination of the coefficient vector

$$\underline{\alpha}^T \triangleq [\alpha_1, \alpha_2, \ldots, \alpha_k] \tag{3.17}$$

in the linear model

$$x_k^* = \sum_{i=1}^{k-1} \alpha_i w_i + \alpha_k x_1 = \underline{\alpha}^T \cdot \underline{\omega} \tag{3.18}$$

in order to make eqn. (3.18) a close approximation to the nonlinear system (3.16).
The random vector $\underline{\omega}$ is defined as

$$\underline{\omega}^T \triangleq [w_1, w_2, \ldots, w_{k-1}, x_1] \tag{3.19}$$

The assumptions of Section 2 imply

$$p(\underline{\omega}) = \prod_{i=1}^{k-1} p(w_i) \cdot p(x_1) \tag{3.20}$$

and therefore, using eqn.(3.6) and (3.7) as a univariate p.d.f. (i.e. x_m, Σ_x, Σ_w are scalars), $p(\underline{\omega})$ is a k-variate normal p.d.f.

$$p(\underline{\omega}) = n(\underline{\omega}; \underline{a}, \Sigma) \tag{3.21}$$

where the mean $\underline{a}$ and the variance Σ are given by

$$\underline{a}^T = [0, 0, \ldots 0, x_m] \tag{3.21a}$$

$$\Sigma = \begin{bmatrix} \Sigma_w & & & 0 \\ & \ddots & & \\ & & \Sigma_w & \\ 0 & & & \Sigma_x \end{bmatrix} \tag{3.21b}$$

The parameter $\underline{\alpha}$ is said to be optimal, denoted as $\underline{\alpha}^o$, if the variance of the error e_k, defined as the difference

between system and model state x_k and x_k^* respectively, is minimal. Thus the functional $F(\underline{\alpha})$ to be minimized is:

$$F(\underline{\alpha})=\text{var}(e_k)=\text{var}(x_k-x_k^*)=\text{var}(x_k)+\text{var}(x_k^*)-2\ \text{cov}(x_k,x_k^*) \quad (3.22)$$

Since var($\underline{z}$) is defined as:

$$\text{var}(\underline{z}) \triangleq E[\underline{z}\ \underline{z}^T] - E[\underline{z}]\ .\ (E[\underline{z}])^T \quad (3.23)$$

then $F(\underline{\alpha})$ is a quadratic function in $\underline{\alpha}$ and thus the optimal $\underline{\alpha}^o$ is obtained as:

$$\underline{\alpha}^o = \underline{\alpha} - F_{\alpha\alpha}^{-1}\ \underline{F}_\alpha \quad (3.24)$$

where $\underline{F}_\alpha$ is the gradient and $F_{\alpha\alpha}$ the matrix of second order derivatives of $F(\underline{\alpha})$ with respect to $\underline{\alpha}$. Using eqn. (3.21), (3.22) and (3.23), the gradient $\underline{F}_\alpha$ is given by:

$$\underline{F}_\alpha = 2 \int \underline{\omega}\ \underline{\omega}^T\ \underline{\alpha}\ p(\underline{\omega})\ d\underline{\omega} - 2 \int \underline{\alpha}^T\ \underline{\omega}\ p(\underline{\omega})\ d\underline{\omega} \int \underline{\omega}\ p(\underline{\omega}) d\underline{\omega} \quad (3.25)$$

$$- 2 \iint x_k\ \underline{\omega}\ p(x_k,\underline{\omega}) d\,x_k\ d\underline{\omega} + 2 \int x_k\ p(x_k)\ dx_k \int \underline{\omega}\ p(\underline{\omega}) d\underline{\omega}$$

The matrix $F_{\alpha\alpha}$ is obtained by differentiating (3.25) with respect to $\underline{\alpha}$. This yields with eqn. (3.21):

$$F_{\alpha\alpha} = 2\ \Sigma \quad (3.26)$$

Since the p.d.f. $p(x_k)$ and the joint density $p(x_k,\underline{\omega})$ are unknown in eqn. (3.22), the gradient has to be estimated as:

$$\hat{\underline{F}}_\alpha = 2\ .\ N_1^{-1}\ \Big[\sum_{j=1}^{N_1} \underline{\omega}_j\ \underline{\omega}_j^T\ \underline{\alpha} - \sum_{j=1}^{N_1} \underline{\alpha}^T\ \underline{\omega}_j \sum_{j=1}^{N_1} \underline{\omega}_j$$

$$- \sum_{j=1}^{N_1} (x_k)_j\ \underline{\omega}_j + \sum_{j=1}^{N_1} (x_k)_j \sum_{j=1}^{N_1} \underline{\omega}_j \Big] \quad (3.27)$$

where $\underline{\omega}_j$, drawn from $p(\underline{\omega})$ of eqn. (3.21), denotes the

random sequence $(w_1, w_2, \ldots w_{k-1}, x_1)_j$, $j=1,2,\ldots,N_1$. The same random sequence $\underline{\omega}_j$ is used to simulate eqn.(3.16) to find $(x_k)_j$, which denotes the state x_k of the nonlinear system (3.16).

Since the gradient $\underline{F}_\alpha$ has to be replaced in eqn. (3.24) by its estimate $\hat{\underline{F}}_\alpha$, the optimal $\underline{\alpha}^o$ is obtained as an estimate $\hat{\underline{\alpha}}^o$, defined as

$$\hat{\underline{\alpha}}^o = \underline{\alpha} - F_{\alpha\alpha}^{-1} \hat{\underline{F}}_\alpha \tag{3.28}$$

The coefficient $\hat{\underline{\alpha}}^o$ could be estimated by ordinary regression analysis. The new feature of our method is the deterministically specified matrix $F_{\alpha\alpha}$ in eqn. (3.28).

The sample size N_1 affects the accuracy of the estimate $\hat{\underline{F}}_\alpha$ and therefore the accuracy of the estimate $\hat{\underline{\alpha}}^o$.

In the second stage of the sampling procedure, the linear model (3.18), with $\underline{\alpha}$ replaced by $\hat{\underline{\alpha}}^o$, is used to break eqn. (3.2) into two parts:

$$E[x_k] = \left[\int x_k \, p(x_k) \, d x_k - \int x_k^* \, p_m(x_k^*) \, d x_k^* \right] + \int x_k^* \, p_m(x_k^*) \, d x_k^* \tag{3.29}$$

The subscript 'm' indicates the p.d.f. $p_m(x_k^*)$ belongs to the linear model (3.18). We integrate the two parts of eqn. (3.29) separately, the first part by the Crude Monte Carlo method and the second analytically. Indeed, due to linearity, the last integral can be evaluated with eqn. (3.21) as:

$$E[x_k^*] = \int x_k^* \; p_m(x_k^*) \, d x_k^* = E[\underline{\alpha}^T \cdot \underline{\omega}] = \underline{\alpha}^T \cdot \underline{a} \tag{3.30}$$

and thus the new estimator takes the form:

$$\hat{m}_k = N^{-1} \sum_{j=1}^{N} \left[(x_k)_j - (x_k^*)_j \right] + \underline{\alpha}^T \underline{a} \tag{3.31}$$

Here $(x_k)_j$ denotes the state x_k of the system (3.16) using the random sequence $\underline{\omega}_j^T = (w^{k-1}, x_1)_j$, where $(x_1)_j$ is drawn from $p(x_1)$ and $(w^{k-1})_j$ are drawn from $p(wk-1)$.

The same random vector $\underline{\omega}_j$ is used to generate $(x_k^*)_j$ as the solution of the linear stochastic difference eqn.(3.18).

The middle term in eqn. (3.31) is known as the Control Variate of the first term. A reduced sampling variance $\text{var}(\hat{m}_k)$ is obtained provided that the linear model (3.18), giving rise to the control variate, is a close approximation to the original nonlinear problem, and absorbs most of the variations in the sampling procedure; i.e. $(x_k^*)_j$ must be a close approximation to the original state $(x_k)_j$ of eqn. (3.16).

The computing routine for the control variate method can be summarized as follows:

1) Choose an arbitrary set of values for $\underline{\alpha}$ and estimate the gradient $\underline{F}_\alpha$ with eqn. (3.27), using a sample of size N_1.
2) Update the parameter $\underline{\alpha}$ using eqn. (3.26) and (3.28).
3) Compute the analytic result of eqn. (3.30).
4) Simulate the original system (3.16) and the linear model (3.18) to generate the random sample $(x_k)_j$ and $(x_k^*)_j$, $j=1,2,\ldots,N$.
5) Estimate the mean with eqn. (3.31). An analogous expression to eqn. (3.5) is used to find an estimate $\hat{V}$ for the reduced sampling covariance V.

3.3. Numerical Results

As an illustrative example let us consider the following scalar system:

$$x_{k+1} = x_k - 0.2\, x_k^3 + w_k \qquad (3.32)$$

with a deterministic initial condition $x_m=1.0$ and a noise variance $\Sigma_w = 0.0625$. Table 1 shows averages over ten ensembles with samples of size N=500. Thus $\bar{\hat{m}}_k$ is the ensemble average of $\hat{m}_k$ and $\bar{\hat{V}}_k$ is the ensemble average of $\hat{V}_k$, the estimate of $\text{var}(\hat{m}_k)$. The nonlinear prediction is nine intervals ahead. Further numerical results are given in reference[7].

	$\bar{\hat{m}}_{10}$	$\bar{\hat{V}}_{10} \cdot 10^{-4}$
Crude Estimator (3.3)	0.27020	4.879
Antithetic Variate Estimator (3.10)	0.26720	0.112
Control Variate estimator (3.31) $\alpha_1=0.4$	0.27774	1.202
Control Variate estimator (3.31) $\hat{\alpha}_1^{o}= 0.1814$	0.27322	0.262

Table 1: Nonlinear prediction nine intervals ahead.

The optimal Control Variate estimator shows a significant variance reduction compared with the crude estimator and the Control Variate estimator with an arbitrary chosen value for $\underline{\alpha}$. Although the results here seem to indicate that the Antithetic Variate estimator gives the smallest sampling variance for predicting the mean, the optimal Control Variate estimator is superior for estimating higher order moments.

Although there are approximate methods to solve the Chapman-Kolmogorov eqn. (3.1), judiciously designed Monte Carlo methods yield estimates, such as the mean $\hat{\underline{m}}_k$, whose sampling error $(\hat{V}_k)^{\frac{1}{2}}$ can be less than the approximation error of analytic methods. However, this is obtained at the expense of generating a random sample of appropriate size N.

4. Nonlinear, Multistage Filtering

Referring to the mathematical model introduced in Section 2, it is desired to estimate the state $\underline{x}_k$ conditioned on all past and present observations $\underline{y}^k \triangleq \underline{y}_1, \underline{y}_2, \cdots \underline{y}_k$. This requires the determination of the posterior p.d.f. $p(\underline{x}_k \mid \underline{y}^k)$ (see Doob[8]). Using Bayes' theorem, $p(\underline{x}_k \mid \underline{y}^k)$ can be replaced by the likelihood functions $p(\underline{y}_i \mid \underline{x}_i)$, $i=1,2,\ldots k$, and the prior p.d.f. $p(\underline{x}^k)$. The conditional mean $E[\underline{x}_k \mid \underline{y}^k]$, defined as

$$E[\underline{x}_k \mid \underline{y}^k] \triangleq \int \underline{x}_k \, p(\underline{x}_k \mid \underline{y}^k) \, d\underline{x}_k , \qquad (4.1)$$

can then be expressed as

$$\underline{x}_{k|k} = E[\underline{x}_k | \underline{y}^k] = \frac{\int\cdots\int \underline{x}_k \prod_{i=1}^{k} p(\underline{y}_i|\underline{x}_i)\, p(\underline{x}^k)\, d\underline{x}^k}{\int\cdots\int \prod_{i=1}^{k} p(\underline{y}_i|\underline{x}_i)\, p(\underline{x}^k)\, d\,\underline{x}^k} = \frac{\underline{\theta}_{n,k}}{\theta_{d,k}} \tag{4.2}$$

The numerator $\underline{\theta}_{n,k}$ is an n-dimensional vector which can be expressed as

$$\underline{\theta}_{n,k} = E[\underline{x}_k \prod_{i=1}^{k} p(\underline{y}_i \mid \underline{x}_i)] \tag{4.3}$$

The denominator $\theta_{d,k}$ is a scalar constant and can be expressed as:

$$\theta_{d,k} = E[\prod_{i=1}^{k} p(\underline{y}_i \mid \underline{x}_i)] \tag{4.4}$$

It is easy to verify that this scalar constant is related to the conditional p.d.f. $p(\underline{y}_k \mid \underline{y}^{k-1})$ by

$$\theta_{d,k} = \prod_{i=2}^{k} p(\underline{y}_i \mid \underline{y}^{i-1}) \cdot p(\underline{y}_1) \tag{4.5}$$

This result will be used again in Section 4.2.

4.1. A Crude Monte Carlo Estimator

Since the expectations in eqn. (4.3) and (4.4) are with respect to $p(\underline{x}^k)$ with $\underline{y}^k$ kept constant, the random variables

$$\hat{\underline{\theta}}_{n,k} = N^{-1} \sum_{j=1}^{N} (\underline{x}_k)_j \prod_{i=1}^{k} p(\underline{y}_i \mid (\underline{x}_i)_j) \tag{4.6}$$

and

$$\hat{\theta}_{d,k} = N^{-1} \sum_{j=1}^{N} \prod_{i=1}^{k} p(\underline{y}_i \mid (\underline{x}_i)_j) \tag{4.7}$$

converge again by the strong law of large numbers with probability one to $\underline{\theta}_{n,k}$ and $\theta_{d,k}$ respectively, provided that N samples $(\underline{x}^k)_j$, j=1,2,..,N, are drawn from $p(\underline{x}^k)$ and $N \to \infty$.

Eqn. (4.2) implies that the estimate $\hat{\underline{x}}_{k|k}$ of the

conditional mean $E[\underline{x}_k \mid \underline{y}^k]$ is given as the ratio of two random variables:

$$\hat{\underline{x}}_{k|k} = \frac{\hat{\underline{\theta}}_{n,k}}{\hat{\theta}_{d,k}} \qquad (4.8)$$

Although this estimator is a biased approximation it has been shown by Handschin and Mayne[9] that the error is negligible in comparison with the sampling error $(\mathrm{var}(\hat{\underline{x}}_{k|k}))^{\frac{1}{2}}$, provided that $N \to \infty$. An approximation V to the sampling covariance matrix V_o, defined as

$$V_o = \mathrm{var}(\hat{\underline{x}}_{k|k}) = E[(\hat{\underline{x}}_{k|k} - E[\underline{x}_k|\underline{y}^k])(\hat{\underline{x}}_{k|k} - E[\underline{x}_k|\underline{y}^k])^T], \qquad (4.9)$$

is obtained by expanding the RHS of eqn.(4.8) as a Taylor series around the respective means $\underline{\theta}_{n,k} = E[\hat{\underline{\theta}}_{n,k}]$ and $\theta_{d,k} = E[\hat{\theta}_{d,k}]$ and truncating terms higher than the second order. Denoting the r:th element of $\underline{\theta}_{n,k}$ as $\theta_{n,k}(r)$, the (r,p):th element of V is found to be:

$$V(r,p) = \frac{\theta_{n,k}(r).\theta_{n,k}(p)}{\theta_{d,k}^2}\left[\frac{\mathrm{var}(\hat{\theta}_{d,k})}{\theta_{d,k}^2} + \frac{\mathrm{cov}(\hat{\theta}_{n,k}(r),\hat{\theta}_{n,k}(p))}{\theta_{n,k}(r).\theta_{n,k}(p)}\right.$$

$$\left. - \frac{\mathrm{cov}(\hat{\theta}_{n,k}(p),\hat{\theta}_{d,k})}{\theta_{n,k}(p).\theta_{d,k}} - \frac{\mathrm{cov}(\hat{\theta}_{n,k}(r),\hat{\theta}_{d,k})}{\theta_{n,k}(r).\theta_{d,k}}\right] \qquad (4.10)$$

An estimate $\hat{V}$ of V is obtained by replacing all the terms in eqn. (4.10) by their estimates; e.g.

$$\mathrm{var}(\hat{\theta}_{d,k}) = N^{-2}\sum_{j=1}^{N}\left[\prod_{i=1}^{k} p(y_i|(x_i)_j) - \hat{\theta}_{d,k}^2\right]^2 \qquad (4.11)$$

In conclusion, eqn. (4.8) defines a Crude Monte Carlo estimator for the conditional mean $E[\underline{x}_k|\underline{y}^k]$ if the numerator $\underline{\theta}_{n,k}$ and denominator $\theta_{d,k}$ of eqn.(4.2) are estimated by eqn. (4.6) and (4.7). The sample $(\underline{x}^k)_j$, j=1,2,...,N, is again generated by the direct simulation principle: $(\underline{x}_k)_j$ denotes the state $\underline{x}_k$ of the nonlinear system (2.1) obtained by simulation of the system with the random sequence

$(\underline{x}_1, \underline{w}_1, \ldots, \underline{w}_{k-1})_j$. $(\underline{x}_1)_j$ is drawn from the p.d.f. $p(\underline{x}_1)$ and $(\underline{w}^{k-1})_j$ are drawn from $p(\underline{w}^{k-1})$.

The Crude Monte Carlo estimator (4.8) is not confined by any restrictive assumption. However, this generality is offset by the low accuracy of the estimator. In the following, Section 4.2, a Control Variate method is derived to give a reduced variance V.

4.2. Variance Reduction Techniques

A combination of the nonlinear filter equation and a Monte Carlo method is used to yield a Control Variate estimator for the conditional mean with reduced sampling variance. Up to the present time, several trials have been made on the physical realisation of optimal nonlinear filters in an approximate form of finite dimensional filter. Sunahara[10] uses a method of stochastic linearisation which has been applied by Handschin and Mayne[9] to derive a Monte Carlo estimator for the conditional mean. In this paper a different set of nonlinear filtering equations due to Sorenson[4] is used to specify an approximate solution. The nonlinear transformations f(.,.) and g(.,.) are subsequently linearized along this reference trajectory to yield a linear model required for the Control Variate method. This model is a statistical version of an exact nonlinear filter.

4.2.1. Approximate Nonlinear Filtering

A set of nonlinear filter equations is derived by Sorenson[4] for the following system

$$\underline{x}_{k+1} = \underline{f}_k (\underline{x}_k, k) + \underline{w}_k \qquad (4.12)$$

whose states are observed by eqn. (2.3). Under the assumption that $\underline{f}_k(.,.)$ has at least continuous first derivatives, and $\underline{g}_k(.,.)$ has first and second order derivatives, an approximate posterior p.d.f. is assumed to be Gaussian and defined as

$$p_a(\underline{x}_{k+1} | \underline{y}^{k+1}) = n(\underline{x}_{k+1}; \underline{\mu}_{k+1|k+1}, \underline{\Sigma}_{k+1|k+1}) \qquad (4.13)$$

where the mean $\underline{\mu}_{k+1|k+1}$ and the variance $\underline{\Sigma}_{k+1|k+1}$ are given by:

$$\underline{\mu}_{k+1|k+1} = \underline{z}_{k+1} + \Sigma_{k+1|k+1}\, G_{k+1} \Sigma_v^{-1} [\underline{y}_{k+1} - \underline{g}_{k+1}(\underline{z}_{k+1})]$$

$$\underline{z}_{k+1} = \underline{f}_k\,(\underline{\mu}_{k|k},\ k)$$

$$\Sigma_{k+1|k+1} = [(\Sigma_w + F_k \Sigma_{k|k} F_k^T)^{-1} + G_{k+1}^T \Sigma_v^{-1} G_{k+1} - \sum_{i=1}^{m} J_{k+1}(i) \cdot u_{k+1}(i)]^{-1} \qquad (4.14)$$

where $u_{k+1}(i)$ are the components of the vector $\underline{u}_{k+1} \triangleq \Sigma_v^{-1} [\underline{y}_{k+1} - \underline{g}_{k+1}(\underline{z}_{k+1})]$. In addition to the notation introduced in Section 2, we use locally the following abbreviations for the first partial derivatives:

$$F_k \triangleq \left.\frac{\partial \underline{f}_k}{\partial \underline{x}_k}\right|_{\underline{x}_k = \underline{\mu}_{k|k}} \qquad G_{k+1} \triangleq \left.\frac{\partial \underline{g}_{k+1}}{\partial \underline{x}_{k+1}}\right|_{\underline{x}_{k+1} = \underline{f}_k(\underline{\mu}_{k|k}, k)} \qquad (4.15)$$

The second partials of the i:th component of $\underline{g}_{k+1}$ are denoted as:

$$J_{k+1}(i) \triangleq \left.\frac{\partial^2 \underline{g}_{k+1}(i)}{\partial \underline{x}_{k+1}\, \partial \underline{x}_{k+1}}\right|_{\underline{x}_{k+1} = \underline{f}_k(\underline{\mu}_{k|k},\ k)} \qquad (4.16)$$

The recursive eqn. (4.14) starts at time k=1 with

$$\underline{\mu}_{1|1} = \underline{x}_m + \Sigma_{1|1}\, G_1 \Sigma_v^{-1} (\underline{y}_1 - \underline{g}_1(\underline{x}_m))$$

$$\Sigma_{1|1} = [\Sigma_x^{-1} + G_1^T \Sigma_v^{-1} G_1 - \sum_{i=1}^{m} J_1(i)\, u_1(i)]^{-1} \qquad (4.17)$$

where $\underline{u}_1 \triangleq \Sigma_v^{-1} [\underline{y}_1 - \underline{g}_1(\underline{x}_m)]$. The set of nonlinear eqn. (4.14) is obtained by expanding the exponents of the p.d.f. appearing in Bayes' theorem eqn. (4.2) into a Taylor series up to second order, around $\underline{\mu}_{k|k}$ for $\underline{f}_k(.,.)$, and around $f(\underline{\mu}_{k|k}, k)$ for $\underline{g}_{k+1}(.,.)$. The full derivation is given by Sorenson[4].

4.2.2. A Linear Model

The coefficients $\underline{a}_k$, B_k, $\underline{c}_k$ and D_k of the linear model

$$\underline{x}^*_{k+1} = \underline{a}_k + B_k \, (\underline{x}^*_k - \underline{\mu}_{k|k}) \tag{4.18}$$

$$\underline{y}_k = \underline{c}_k + D_k \, (\underline{x}^*_k - \underline{\mu}_{k|k}) \tag{4.19}$$

are obtained by expanding $\underline{f}_k(\underline{x}_k,k)$ and $\underline{g}_k(\underline{x}_k,k)$ into a Taylor series around $\underline{\mu}_{k|k}$ such that the mean squared norm of the remainders of the expansions are minimal with respect to $p_a(\underline{x}_k|\underline{y}^k)$. This yields:

$$\underline{a}_k = E_{p_a}\left[\underline{f}_k (\underline{x}_k,k) | \underline{y}^k \right] \tag{4.20}$$

$$B_k = E_{p_a}\left[\underline{f}_k (\underline{x}_k,k) - \underline{a}_k)(\underline{x}_k - \underline{\mu}_{k|k})^T | \underline{y}^k \right] \Sigma_{k|k}^{-1} \tag{4.21}$$

$$\underline{c}_k = E_{p_a}\left[\underline{g}_k (\underline{x}_k,k) | \underline{y}^k \right] \tag{4.22}$$

$$D_k = E_{p_a}\left[(\underline{g}_k (\underline{x}_k,k) - \underline{c}_k)(\underline{x}_k - \underline{\mu}_{k|k})^T | \underline{y}^k \right] \Sigma_{k|k}^{-1} \tag{4.23}$$

Using Bayes' theorem in the form of eqn. (4.2) for the linear model (4.18) and (4.19), the combination of nonlinear system and linear models allows to rewrite the numerator $\underline{\theta}_{n,k}$ and denominator $\theta_{d,k}$ as follows:

$$\underline{\theta}_{n,k} = \int\cdots\int_{\underline{x}_k} \prod_{i=1}^{k} p(\underline{y}_i|\underline{x}_i) p(\underline{x}^k) d\underline{x}^k - \int\cdots\int_{\underline{x}^*_k} \prod_{i=1}^{k} p_m(\underline{y}_i|\underline{x}^*_i) p_m(\underline{x}^{*k}) d\underline{x}^{*k}$$

$$+ \underline{\theta}_{mn,k} \tag{4.24}$$

$$\theta_{d,k} = \int\cdots\int \prod_{i=1}^{k} p(\underline{y}_i|\underline{x}_i) p(\underline{x}^k) d\underline{x}^k - \int\cdots\int \prod_{i=1}^{k} p_m(\underline{y}_i|\underline{x}^*_i) p_m(\underline{x}^{*k}) d\underline{x}^{*k}$$

$$+ \theta_{md,k} \tag{4.25}$$

The index m refers to the linear model and due to linearity the correction terms $\underline{\theta}_{mn,k}$ and $\theta_{md,k}$ have an analytic result defined as:

$$\underline{\theta}_{mn,k} = \int\cdots\int \underline{x}^*_k \cdot \prod_{i=1}^{k} p_m(\underline{y}_i|\underline{x}^*_i) \; p_m(\underline{x}^{*k}) \; d\underline{x}^{*k} \tag{4.26}$$

$$\theta_{md,k} = \int\cdots\int \prod_{i=1}^{k} p_m(\underline{y}_i|\underline{x}^*_i) \; p_m(\underline{x}^{*k}) \; d\underline{x}^{*k} \tag{4.27}$$

Because of the linear structure of the model (4.18) and (4.19), the p.d.f. $p_m(\underline{y}_{k+1}|\underline{y}^k)$ is an m-variate Gaussian density with mean $\underline{m}_{k+1|k}$ and variance $V_{k+1|k}$ given by

$$\underline{m}_{k+1|k} = \underline{c}_{k+1} + D_{k+1}[\underline{a}_k + B_k(\underline{m}_{k|k} - \underline{\mu}_{k|k}) - \underline{\mu}_{k+1|k+1}]$$

$$V_{k+1|k+1} = D_{k+1} B_k V_{k|k} B_k^T D_{k+1}^T + D_{k+1} \Sigma_w D_{k+1}^T + \Sigma_v \quad (4.28)$$

These recursive relations (4.29) start with

$$\underline{m}_{1|0} = \underline{c}_1 + D_1 (\underline{x}_m - \underline{\mu}_{1|1})$$

$$V_{1|0} = D_1 \Sigma_x D_1^T + \Sigma_v \quad (4.29)$$

Clearly $\underline{\mu}_{k|k}$ is given by eqn. (4.14), and the mean $\underline{m}_{k|k}$ and variance $V_{k|k}$ of the posterior p.d.f. $p_m(\underline{x}_{k+1}|\underline{y}^{k+1})$ of the linear model (4.18) and (4.19) are obtained by linear filter theory[11]:

$$\underline{m}_{k+1|k+1} = \underline{a}_k + B_k(\underline{m}_{k|k} - \underline{\mu}_{k|k}) + V_{k|k} D_k \Sigma_v^{-1}(\underline{y}_k - \underline{m}_{k|k-1})$$

$$V_{k+1|k+1} = [(\Sigma_w + B_k V_{k|k} B_k^T)^{-1} + D_k^T \Sigma_v^{-1} D_k]^{-1} \quad (4.30)$$

These recursive eqn. (4.30) start with

$$\underline{m}_{1|1} = \underline{x}_m + V_{1|1} D_1 \Sigma_v^{-1}(\underline{y}_1 - \underline{m}_{1|0}) \quad (4.31)$$

$$V_{1|1} = [\Sigma_x^{-1} + D_1^T \Sigma_v^{-1} D_1]^{-1}$$

Using eqn. (4.5) the solution of eqn. (4.26) and (4.27) is given by eqn. (4.28) as

$$\theta_{md,k} = \prod_{i=1}^{k-1} p_m(\underline{y}_{i+1}|\underline{y}^i) \cdot p(\underline{y}_1) \quad (4.32)$$

and

$$\underline{\theta}_{mn,k} = \underline{m}_{k|k} \cdot \theta_{md,k} \quad (4.33)$$

4.2.3. The Control Variate Estimator

Based on the foregoing discussion, the computational

procedure for the Control Variate method can be summarized as follows:

1) Compute the approximate nonlinear filter eqn.(4.14).
2) Establish the linear model (4.18) and (4.19) and the linear filtering solution (4.30), (4.31).
3) Compute the correction terms (4.32) and (4.33) using eqn. (4.28).
4) Estimate the numerator $\underline{\theta}_{n,k}$, using eqn. (4.24), as

$$\hat{\underline{\theta}}_{n,k} = N^{-1} \sum_{j=1}^{N} (\underline{x}_k)_j \prod_{i=1}^{k} p(\underline{y}_i | (\underline{x}_i)_j) - (\underline{x}_k^*)_j \prod_{i=1}^{k} p_m(\underline{y}_i | (\underline{x}_i^*)_j) + \underline{\theta}_{mn,k} \qquad (4.34)$$

and the denominator $\theta_{d,k}$, using (4.25), as:

$$\hat{\theta}_{d,k} = N^{-1} \sum_{j=1}^{N} \prod_{i=1}^{k} p(y_i | (x_i)_j) - \prod_{i=1}^{k} p_m(y_i | (x_i^*)_j) + \theta_{md,k} \qquad (4.35)$$

5) The estimate $\hat{\underline{x}}_{k|k}$ of the conditional mean $E[\underline{x}_k | \underline{y}^k]$ is given by eqn. (4.8). Its sampling variance matrix V is given by eqn. (4.10).

Applying this Monte Carlo procedure to a linear system with Gaussian noise $\underline{w}_k$ the following result holds.

The Control Variate method based on eqn. (4.34) and (4.35) yields an estimate $\hat{\underline{x}}_{k|k}$ of the conditional mean $E[\underline{x}_k | \underline{y}^k]$ with zero sampling variance.

The proof of this result follows from the property that the nonlinear filter eqn. (4.14) reduces to the ordinary linear filter eqn. (4.30) when applied to a linear system. Thus, a linear model is obtained which is identical with the original system. That implies the differences between the original variates and the control variates in eqn.(4.34) and (4.35) are identically equal to zero, thus removing all randomness from the estimator. This proof indicates that the Control Variate method yields zero sampling variance estimates for any order moments in linear Gaussian systems.

4.3. Numerical Example

The following scalar example is used to illustrate the multistage nonlinear filtering method using Monte Carlo techniques:

$$x_{k+1} = x_k - 0.2\, x_k^3 + w_k \tag{4.36}$$

$$y_k = \text{Tanh}\,(x_k) + v_k \tag{4.37}$$

where

$$p(x_1) = n(x_1;\, 1,\, 0.01)\quad p(w_k;\, 0,\, 0.01)\quad p(v_k) = n(v_k;\, 0,\, 0.1) \tag{4.38}$$

In the following table the methods of Section 4.1 and 4.2 are compared. These results are averages over ten ensembles with samples of size N=500. The following results are shown in this order: the time parameter k, the given sequence of observations y_k, the ensemble average $\bar{\hat{x}}_{k|k}^{(1)}$ of the conditional mean $E[x_k|y^k]$ using a crude Monte Carlo estimator, the sampling error $(\text{var}(\hat{x}_{k|k}^{(1)}))^{\frac{1}{2}}$, the ensemble average $\bar{\hat{x}}_{k|k}^{(2)}$ of the conditional mean using a Control Variate estimator, the sampling error $(\text{var}(\hat{\bar{x}}_{k|k}^{(2)}))^{\frac{1}{2}}$, the approximate nonlinear result $\mu_{k|k}$ and finally, the ensemble average of the error $\bar{\hat{e}}_{k|k}$ defined as

$$\bar{\hat{e}}_{k|k} = |\,\bar{\hat{x}}_{k|k} - \mu_{k|k}\,| \tag{4.39}$$

k	y_k	$\bar{\hat{x}}_{k\|k}^{(1)}$	$[\text{var}(\hat{x}_{k\|k}^{(1)})]^{\frac{1}{2}}$ 10^{-3}	$\bar{\hat{x}}_{k\|k}^{(2)}$	$[\text{var}(\hat{x}_{k\|k}^{(2)})]^{\frac{1}{2}}$ 10^{-4}	$\mu_{k\|k}$	$\bar{\hat{e}}_k$ 10^{-3}
1	1.1	1.01126	4.397	1.01418	1.407	1.01367	0.502
2	.79	.80662	4.632	.80832	3.912	.81281	4.490
3	.68	.70567	5.077	.70533	4.332	.71140	6.067
4	.58	.63420	5.370	.63416	4.544	.64100	6.836
5	.5	.57918	5.620	.57804	5.009	.58480	6.755
6	.44	.53390	5.762	.53060	5.589	.53681	6.213
7	.4	.49461	6.028	.49043	6.483	.49533	4.889
8	.36	.45797	6.208	.45417	7.554	.45728	3.119
9	.33	.42600	6.487	.42136	8.495	.42271	1.346
10	.28	.39110	6.787	.38737	9.867	.38613	1.244

Table 2: Nonlinear, multistage filtering using Monte Carlo techniques.

These results indicate that the approximation error is larger than the sampling error of the Control Variate estimator, and thus the following two objects have been achieved:

1) A judiciously designed sampling experiment improves the efficiency of the estimator compared with Crude Monte Carlo techniques.
2) The combination of the nonlinear filtering equation and Monte Carlo techniques yields an estimate whose sampling error is less than the approximation error of the nonlinear filter.

5. Summary and Conclusions

The application of Monte Carlo techniques has been found to be of great use for filtering and predicting the states of nonlinear, dynamic systems. The prediction problem has been solved by three different estimators. The efficiency of the Crude Monte Carlo estimator was improved by either using the Antithetic Variate method or the more general Control Variate method. A two stage procedure for the latter case offers an elegant alternative to ordinary regression analysis to establish a linear model.

The Bayesian approach is adopted for the filtering problem, which requires the estimation of parameters for the posterior p.d.f. The Crude Monte Carlo estimator is applicable under very general conditions, but its efficiency is significantly improved by the Control Variate Method.

For a specific example it has been shown that the approximation error of the nonlinear filter is larger than the sampling error of the Control Variate estimator, and thus the latter is a considerable improvement over existing nonlinear filters. Although throughout the paper we assumed to know the system dynamics, the parameter identification problem of nonlinear systems can readily be solved with the presented methods, by increasing the dimensionality of the state space.

Acknowledgement

I would like to thank Dr. D.Q. Mayne for his encouragement, advice and constructive criticism relating to this

research work. I would also like to acknowledge the financial support of SA Brown, Boveri & Co. Baden, Switzerland.

References

1. Fisher, J.R.; "Optimal Nonlinear Filtering", Chapter of Vol. 5. Advances in Control Systems, C.J. Leondes, Editor, Academic Press, New York, 1967.
2. Schwartz, L.; "Approximate Continuous Nonlinear Minimal Variance Filtering", Hughes Aerospace Tech. Res. Report, SSD60472 R, Dec. 1966.
3. Cox, H.; "On the Estimation of State Variables and Parameters for Noisy Systems", IEEE Trans. AC-9 p.5-12, 1964.
4. Sorenson, H.W.; "A Nonlinear Perturbation Theory for Estimation and Control of Time-Discrete Stochastic Systems", Department of Engineering, University of California, Report No. 68-2, Los Angeles, 1968.
5. Papoulis, A.; "Probability, Random Variables and Stochastic Processes", McGraw Hill, New York, 1965.
6. Hammersley, J.M. and Handscomb, D.C.; "Monte Carlo Methods", Methuen's Monographs, London, 1965.
7. Handschin, J.E.; "Monte Carlo Techniques for Filtering and Prediction of Nonlinear Stochastic Processes", Imperial College, Internal Report, London, 1968.
8. Doob, J.L.; "Stochastic Processes", Wiley, New York, 1953.
9. Handschin, J.E. and Mayne, D.Q.; "Monte Carlo Techniques to Estimate the Conditional Expectation in Multistage Nonlinear Filtering", Int. J. Control, (to appear).
10. Sunahara, Y.; "An Approximate Method of State Estimation for Nonlinear Dynamical Systems, Brown University, Tech. Report 1967-68, 1967.
11. Kalman, R.E. and Bucy, R.S.; "New Results in Linear Filtering and Prediction Theory", Trans. ASME, J. Basic Eng. 83 D, p.95, 1961.

INTRODUCTION TO MULTICHANNEL STOCHASTIC COMPUTATION AND CONTROL

G.A. Ferraté, L. Puigjaner and J. Agulló
High Technical School of Engineering
Barcelona (Spain)

1. INTRODUCTION

As a departure from conventional digital or analog computing techniques, the stochastic (random-pulse) computer utilizes logical elements (gates) to process the analog magnitude that has been chosen to represent the variables. The analog magnitude referred to is the probability of pulse-occurrence in a train of random pulses.

It is easy to see, for instance, that given two statistically independent stationary random-pulse trains driving the two inputs of an AND gate the output pulse train, once eventually reshaped, will have a probability of occurrence equal to the product of the probabilities of the incoming inputs. A crude and simple form of multiplier will thus have been obtained.

The use of random-pulse sequences with measurable mean-rates to drive logical operators was first introduced by von Newmann with the aim to show that reliably accurate results could be obtained, through redundancy, from a basically inaccurate representation of variables and unreliable components. Very recently, Poppelbaum [1], Ribeiro [2], and Gaines [3], have extended those ideas to the development of practical computing systems.

As will be shown later, addition, multiplication, delay, integration (and even differentiation), function generation, etc., can be performed with a good accuracy by relatively simple logic arrays. The inherently parallel structure of analog computation together with the increasing availability of complex logic functions in integrated or large scale integrated form suggest a wealth of applications and a very fast development of stochastic computing techniques in the field of process control. This specialized application is further enhanced by the ease with which algebraic and non linear operations can be performed. However, the evaluation of the expected accuracy and computing speed of the stochastic methods should be a prerequisite before each special purpose real time on line control application is envisaged.

With the control application in view, the research that has been carried out at the Automatic Control Dept. of the High Technical School of Engineering in Barcelona is mainly concerned with the generation of stochastic pulse sequences from analog or digital outputs of sensors, together with the feasibility of special multichannel processing techniques to increase the accuracy to bandwidth ratio. Among these

is of particular interest the development of a floating-point stochastic information processing method. Attention has also been payed to novel ways of stochastic function-generators through the manipulation of stochastic digital or analog noise of special probability distribution functions.

2. STOCHASTIC REPRESENTATION OF VARIABLES

In the stochastic computing techniques the physical magnitudes are internally represented by the probability associated with the corresponding random-pulse-rate. Several coding schemes are possible or have been proposed. Initially, non-clocked random-pulse sequences were used,[1] but present trends favour clocked (or syncronous) random-pulse sequences which have many advantages as far as the ease with which some operators can be realized is concerned, if not because they lead to a somewhat simpler mathematical analysis.[2]

In this paper, we will be exclusively concerned with clocked random-pulse sequences (CRPS) to stochastically represent a variable. Several such sequences may be used in multichannel operation, either to deal with the sign transmission problem, to improve the accuracy/speed ratio (the ergodicity of the different pulse trains being assumed), to allow for an easy way of differentiation or to increase the dynamic range of the variables through the use of a floating point stochastic technique.

At this point it is worth to note that the use of clocked random pulse sequences to code an information in analog form introduces from the beginning a kind of sampling of the variables, the probability of pulse occurrence in each clock interval being related to the value of the sampled variable.

2.1. Stochastic Codification

Given a function of time $x = f(t)$, normalized in the interval $0 \leq x \leq 1$ for any value of t, and a sequence $\underline{\delta}^*(t)$ of clock pulses of period θ, we can consider the values $x(t^*)$ of $x(t)$ at the sampling instants coded in analog, digital or stochastic form:

$$^{A}x(t^*),\ ^{D}x(t^*),\ ^{S}x(t^*)$$

or, in simplified notation:

$$^{A}x,\ ^{D}x,\ ^{S}x$$

^{S}x representing the instantaneous probability of pulse occurrence of the corresponding associated random-pulse train $^{S}\underline{x}$ at the sampling instants t^*, this probability being equal to $x(t^*)$.

At this point we must emphasize that unlike the analog or digital form of representation, an instantaneous probability cannot be measured directly. This implies that in order to recover the value of an

stochastically coded variable we must use statistical methods, either averaging over a period of time if stationarity is assumed or in the limit, averaging the instantaneous outputs of a multichannel system of stochastic codification, making use of the ergodicity of those channels.

Averaging a random-pulse train (ratio of actual pulses to number of clock pulses in the averaging period T) we obtain a value for the instantaneous probability every clock pulse. As it is well known this value, for a stationary random-pulse train will fluctuate according to a binomial distribution which, for a sample of sufficient size, can be approximated by the normal distribution. The standard deviation will be:

$$\sigma = \sqrt{\frac{p(1-p)}{n}} \qquad \begin{array}{l} p = \text{probability of R.P.T.} \\ n = \frac{T}{\theta} = \text{sample size} \end{array} \qquad (1)$$

Equation (1) relates the sample size, the precision and hence its confidence level α. We can observe that as the precision is inversely proportional to σ it will be proportional to $\sqrt{n}$. This is also true when the probability is averaged over N ergodic channels if $n = N.T/\theta$.

In Fig. 1 is shown the sample-size/confidence-level relationship for a precision of 0,1 % and a signal probability (worst case) of 0.5.

2.2. The Dynamics of the Averaging Process

To determine the dynamic behaviour of the stochastic codification, intimately related to the statistical problem of value retrieval through the measurement of pulse rate average, it is useful to define the averaging process as:

$$AV({}^{s}x) = \int_0^t {}^{s}\underline{x}(\tau^*) q(t - \tau^*) d\tau \qquad (2)$$

where ${}^{s}\underline{x}(\tau^*)$ is the clocked sequence of binary values of the random pulse train ${}^{s}\underline{x}$ and q is the weighting function which defines the process.

The conventional averaging process (Fig. 2.a) can be found taking

$$q = \frac{1}{T} \qquad \tau + T \geq t$$

$$q = 0 \qquad \tau + T < t$$

and Eq. (2) reduces to

$$AVa({}^{s}x) = \frac{1}{T}\int_0^t {}^{s}\underline{x}(\tau^*) . u(T - t + \tau^*) d\tau$$

the corresponding Laplace transform of the averager itself being:

$$\mathcal{L}[AVa] = \frac{1}{T}\frac{1 - e^{-Ts}}{s}$$

from which the module and the argument of its frequency response

$$\text{Mod} = \frac{\text{sen } T\omega/2}{T\omega/2} \qquad \text{Arg} = \frac{-T\,\omega}{2}$$

As can be seen in Fig. 3 and Fig. 4 the crossover frequency for an attenuation of 3 db is $\omega_o \approx \pi/T$ ($f_o \approx 1/2T$ cps). If now it is assumed that a static accuracy of 0.001 is required, Fig. 1 gives a sample size of 10.000.000 for a confidence level of 0'16%.

If the clock frequency were 10 MHz the crossover frequency, in cps, would be 0.5, corresponding to a sampling period T = 1 sec.

At the assumed 3db crossover frequency the error would be about 30%, however, should the sample size be reduced to n = 100.000 the new 3db crossover frequency would increase to 50 cps and the attenuation at the former 0.5 cps point could be neglected (0.004%), the new static accuracy being now 0.01. Fig. 5, shows the total error (static and dynamic) versus sample-size for different frequencies.

From the foregoing follows that the dynamic range of the SC can be increased at the expense of the static accuracy, decreasing the sample size, the clock frequency remaining constant. This fact, peculiar to the stochastic computation, has no direct counterpart in the conventional analog or digital computers and may have interesting implications in the on line control field.

Wide band-pass together with a high accuracy easily involve the use of extremely high clock frequencies which may reach a few hundred megacycles. Besides multichannel operation, which will be discussed later, other methods can be considered in order to optimize the precision-bandwidth ratio. It is proposed here the use of more sophisticated weighting functions for the averaging process.

The ideal averager would be a low-pass filter (numerical or analog) such that the attenuation would remain constant over the entire bandwidth.

The weighting function of the ideal averager will be described by the inverse Laplace transform

$$q = \frac{1}{j\,2\pi} \int_{\gamma - j\infty}^{\gamma + j\infty} Q(s) e^{ts} ds$$

where

$$|Q(j\omega)| = \begin{cases} 1 & \text{for } |\omega| < \Omega \\ 0 & \text{otherwise} \end{cases} \qquad \underline{Q(j\omega)} = 0 \qquad \Omega = \frac{2\pi}{T}$$

resulting

$$q = \frac{2}{T} \cdot \frac{\text{sen}\Omega t}{\Omega t}$$

The function of Eq. (3) which involves the use of negative values of t (Fig. 2.b), presuposes that the system must respond prior to the input excitation. This filter is physically unrealizable but an approximation can be made if the computer time is allowed to lag the problem time by an amount equal to the maximum value of negative time used for the approximation. To avoid the pure delay thus introduced, the positive time going portion of Eq. 3 can be used at the expense of the armonic response (Fig. 2.c), which in this latter case, after normalization, is given by

$$\mathcal{L}\left[\frac{4}{T} \cdot \frac{\text{sen}\Omega t}{\Omega t}\right] = \frac{2}{\pi} \text{Arctg} \frac{\Omega}{s} \qquad (4)$$

The module of Eq. 4 for the range $\omega \leq \Omega$ is

$$\sqrt{\pi^2 + \left[\frac{1}{2} \ln \left| \frac{\Omega/\omega - 1}{\Omega/\omega + 1} \right| \right]^2}$$

which has an infinite resonance-peak at $\omega = \Omega$ (see Fig. 3).

Again, the averager of Eq. (4) can be approximated by stochastic computing operators according to Fig. 2.d and 2.f.

The approximated averager of Fig. 2.d. is shown for two different values of the constants K_1 and K_2 of the weighting function

$$K_1 u(t) - (K_1 + K_2) u(t - T/2) + K_2 u(t - T).$$

The first set of values has been chosen according to:

$$\rho = \frac{K_1 T/2}{K_2 T/2} = 5.55 \qquad \text{and} \qquad K_1 T/2 - K_2 T/2 = 1$$

ρ being the ratio of the areas of the first two lobes of Fig. 2.c. The second set is for $\rho = 10$. The averager of Fig.3.g is of the same kind, with $\rho = 10$, but for double sampling time. Special attention has been payed to the rectangular or multirectangular approximation because the implementation can be easily performed either by the the use of pulse delay-lines or shifted channels in a multichannel technique.

The response of the single time constant averager, which has been proposed elsewhere, has been included in Fig.3 and Fig. 4 as a reference.

The above considerations would be strictly valid for deterministic pulse-rate codification systems. When the input to the averager is a random pulse-train $^{5}\underline{x}$ the variance of the fluctuation of the statistical

measurement of $^{s}\underline{x}$, supposed stationary, will depend on the weighting function of the averager. When the averager is not of the type shown in Fig.2.a. the approach of Paragraph 2.1. must be revised. Furthermore the power density spectrum of the superimposed noise due to the statistical fluctuations will be affected by the frequency response of the averager. These questions are beyond the scope of this paper and have been the subject of a publication by the authors[4].

2.3. Generation of Stochastic Series with Specified Pulse-Rate Probability.

Let be an analog variable ^{A}x, normalized in the positive interval (0, +1), a random analog noise $\hat{\mathcal{N}}$, and let be a random pulse-train $^{s}\underline{z}$ defined in the following way:

$$^{s}\underline{z}(t^*) = \begin{cases} 1 & \text{if } \hat{\mathcal{N}}(t^*) \leq x(t^*) \\ 0 & \text{otherwise} \end{cases}$$

The instantaneous probability of $^{s}\underline{z}$ will be

$$^{s}z = \text{Prob}\ (^{s}\underline{z}) = \text{Prob}\ (\hat{\mathcal{N}} \leq {}^{A}x) = P_{\mathcal{N}}\ (^{A}x) \qquad (5)$$

where $P_{\mathcal{N}}$ is the cumulative probability function of the noise $\hat{\mathcal{N}}$. Equation (5), implies the instantaneous nature of the probability density function of $\mathcal{N}$ at the sampling instants t^*, that is to say, the value of $\mathcal{N}$ at one sampling instant does not affect at all the value at the next.

If,
$$P_{\mathcal{N}}(v) = v \qquad (6)$$

then $^{s}z = {}^{A}x$, and hence, the random-pulse train is a true stochastic representation of the variable. Ec. (6) requires a uniform probability density function,

$$p_{\mathcal{N}}(v) = 1.$$

Gaussian noise, in spite of its apparently easier obtention is not to be recommended for this purpose owing to the non-linearity of its cpf. An acceptable linearity would be only between 0.25 to 0.75 and an "extended" linearity could be put into consideration for values ranging from 0.16 to 0.84.

An schematic diagram explaining the above process is shown in Fig. 6. A similar method can be used for digital to stochastic conversion in which the converter compares both digital values of variable and noise at the clock intervals.

2.4. Sampled Digital or Analog Noise with Rectangular Instantaneous Probability Density Function.

In the preceding paragraph the need of digital or analog random noise

of rectangular instantaneous probability density function has been established:

$$p_{\mathcal{N}^*}(v) = 1$$

A qualified random noise of the above characteristics in digital form can be obtained from several (one per binary digit) statistically independent clocked random binary-pulse-trains [6] (BPT). The analog random noise, in sampled form, can be obtained by a digital to analog conversion. See Fig. 7. The several statistically independent random BPTs can be approximately generated from a single random BPT, using a multiple output multiplexing network to decorrelate them.

Pseudo random BPTs can be used instead of random binary noise. The advantage is that it can be easily generated by the well known technique of maximum length sequences through the use of shift registers with modulo-two addition feedback paths[5]. The periodic nature of the pseudo random sequences can be taken to advantage in some computing process to reduce the dispersion of the results.

2.5 The Sign Transmission in the Stochastic Codification

Several schemes have been proposed to deal with the problem of sign transmission. Among them:

a) Coding the module of the variable in a normal way and transmiting the sign on a separate binary channel.

b) Coding the sign as negative or positive pulses in the stochastic train.

c) Positive values are coded in the 0.5 to 1 probability range and negative values in the 0 to 0.5 range, so that ${}^{s}x = \text{Prob}({}^{s}\underline{x}) - 0.5$.

d) Coding the variable as the difference between the probability of two stochastic channels, minuhend and sustrahend:

$${}^{s}x = {}_{m}^{s}x - {}_{s}^{s}x = \text{Prob}({}_{m}^{s}\underline{x}) - \text{Prob}({}_{s}^{s}\underline{x})$$

If the two channels are non coincident i.e. no pulses occur simultaneously in both channels, they will be called "exclusive" and represented as follows:

$${}_{e,m}^{s}x \qquad \text{and} \qquad {}_{e,s}^{s}x$$

Coding schemes a) and b) usually result in unnecessary complexity for some operators. The method c) is very simple but reduces the dynamic range of the coded variables. The coding method under d) besides the sign transmission, has several advantages that will be discussed later. Also, according to this scheme and in order to have the uniqueness in the encoding process it is convenient to set down the following extra requirements.

$$_{e,s}{}^{s}\underline{x} = 0 \text{ for positive variable}$$

$$_{e,m}{}^{s}\underline{x} = 0 \text{ for negative variable}$$

2.6. Random Floating Point Stochastic Codification

The classical stochastic codification has the disadvantage that the range of the computer variable is limited to values smaller than one. The implications of this fact may be extremely troublesome when the scaling of nonlinear problems is required, because the stochastic trains tend to vanish as they proceed through some logical operators. To circunvent this problem and in order to increase the dynamic range of the variables, the authors propose a generalized stochastic codification whereby the concept of mathematical expectation is introduced through the use of weighted probabilites.

In Paragraph 2.1 the stochastic codification was defined as:

$$^{s}y = \text{Prob}\left[{}^{s}\underline{y}\right], \quad |y| \leq 1$$

the proposed generalized codification establishes:

$$X = {}^{w,s}X = \text{Math Exp}\left[{}^{w,s}\underline{X}\right] \tag{7}$$

in which X is an unscaled variable and ${}^{w,s}\underline{X}$ is the bichannel signal:

$$\left[\underline{{}^{w,s}x} \middle| \underline{w_x}\right] = {}^{w,s}\underline{X}$$

where ${}^{w,s}\underline{x}(t^*)$ is a binary random pulse train associated to a sampled multilevel (digital or analog) information channel $\underline{w_x}(t^*)$ which, at each sampling instant t^*, weights the probability ${}^{w,s}x$ of the binary valued train ${}^{w,s}\underline{x}$.

Taking at the encoding stage,

$$\underline{w_x}(t^*) = \log_b\left[\frac{X(t^*)}{x(t^*)}\right]$$

with

$$\frac{1}{b} < \left|x(t^*)\right| = \left|{}^{w,s}x(t^*)\right| \leq 1$$

a "floating-point" stochastic codification, of base b, is obtained.

The mathematical expectation of Eq. (7) cannot be found directly. Again, an statistical estimation of that value is required:

$$X \approx \text{Estim}\left\{\text{MExp}\left[{}^{w,s}\underline{X}\right]\right\} = \text{AV}\left[b^{\underline{w}_x} \cdot {}^{w,s}\underline{x}\right].$$

This coding scheme, compatible with the substractive method of sign transmission, does not impair the inherent simplicity of the stochastic logic operators, and dramatically simplifies the scaling.

3. STOCHASTIC OPERATORS

A brief survey of the main types of operators will follow including, in particular, the implementation of the generalized (random floating point) stochastic codification with logical elements. It will be noticed that a wide use is made of linear forms with exclusive (non inclusive) stochastic coeficients, due to the fact that the additions are performed by a single OR gate.

3.1. Addition and Substraction

In normal stochastic codification addition can be performed with the methods shown in the self-explanatory logic diagrams of Figs. 8 and 9. The first one does not require the use of auxiliary random-pulse noise while the last needs two exclusive 0.5 stochastic constants. Both operators are shown for bipolar variables.

In generalized stochastic codification the basic scheme of Fig. 9 suggests the random floating point adder shown in Fig. 10. The operation can be easily explained. The output of the adder is:

$${}^{w,s}\underline{z} = {}_{.5}^{s}\underline{k} \cdot {}^{w,s}\underline{X} \;\text{v}\; {}_{.5}^{s}\overline{\underline{k}} \cdot {}^{w,s}\underline{Y}$$

which means

$${}_{e,m}^{w,s}\underline{z} = {}_{.5}^{s}\underline{k} \cdot {}_{e,m}^{w,s}\underline{x} + {}_{.5}^{s}\overline{\underline{k}} \cdot {}_{e,m}^{w,s}\underline{y}$$

$${}_{e,s}^{w,s}\underline{z} = {}_{.5}^{s}\underline{k} \cdot {}_{e,s}^{w,s}\underline{x} + {}_{.5}^{s}\overline{\underline{k}} \cdot {}_{e,s}^{w,s}\underline{y}$$

$$\underline{w_z} = {}_{.5}^{s}\underline{k} \cdot \underline{w_x} + {}_{.5}^{s}\overline{\underline{k}} \cdot \underline{w_y}$$

where $\overline{\underline{k}} = 1 - \underline{k}$ (complementary random pulse trains). The estimation of this output is:

$$\text{Est}\left[\underline{z}\right] = \text{AV}\left[{}_{e,m}^{w,s}\underline{z}\right] - \text{AV}\left[{}_{e,s}^{w,s}\underline{z}\right] =$$

$$= \text{AV}\left[b^{\underline{w}_z} \cdot {}_{e,m}^{w,s}\underline{z}\right] - \text{AV}\left[b^{\underline{w}_z} \cdot {}_{e,s}^{w,s}\underline{z}\right]$$

and substituting from Eq. (8),

$$Est\left[Z\right] = AV\left[b^{\left({}^{s}_{5}\underline{k}\ \underline{w_x} + {}^{s}_{5}\underline{\bar{k}}\ \underline{w_y}\right)} \cdot \left({}^{s}_{5}k \cdot {}^{w,s}_{e,m}\underline{x} + {}^{s}_{5}\underline{\bar{k}} \cdot {}^{w,s}_{e,m}\underline{y} - {}^{s}_{5}\underline{k} \cdot {}^{w,s}_{e,s}\underline{x} - {}^{s}_{5}\underline{\bar{k}}\ {}^{w,s}_{e,s}\underline{y}\right)\right]$$

rearranging and accounting for the exclusivity of the complementary constants,

$$Est\left[Z\right] = AV\left[{}^{s}_{5}\underline{k} \cdot b^{\underline{w_x}} \cdot {}^{w,s}_{e,m}\underline{x}\right] - AV\left[{}^{s}_{5}\underline{k}^{w} \cdot b^{\underline{w_x}} \cdot {}^{w,s}_{e,s}\underline{x}\right] +$$

$$+ AV\left[{}^{s}_{5}\underline{\bar{k}} \cdot b^{\underline{w_y}} \cdot {}^{w,s}_{e,m}\underline{y}\right] - AV\left[{}^{s}_{5}\underline{\bar{k}} \cdot b^{\underline{w_y}} \cdot {}^{w,s}_{e,s}\underline{y}\right] =$$

$$= \frac{1}{2} Est\left[{}^{w,s}_{e,m}X\right] - \frac{1}{2} Est\left[{}^{w,s}_{e,s}X\right] + \frac{1}{2} Est\left[{}^{w,s}_{e,m}Y\right] - \frac{1}{2} Est\left[{}^{w,s}_{e,s}Y\right] =$$

$$= \frac{1}{2}\left\{Est\left[{}^{w,s}X\right] + Est\left[{}^{w,s}Y\right]\right\}$$

so that the operation performed is:

$$Z = \frac{1}{2}\left(X + Y\right).$$

The substraction is obtained with this adder by simply interchanging the minuhend and substrahend binary channels.

3.2. Product and Quotient

The logic diagram for performing the product is shown in Fig. 11. The output is

$${}^{w,s}\underline{Z} = {}^{w,s}\underline{X} \circledast {}^{w,s}\underline{Y}$$

The symbol ⊛ meaning:

$${}^{w,s}_{e,m}\underline{z} = {}^{w,s}_{e,m}\underline{x} \cdot {}^{w,s}_{e,m}\underline{y} + {}^{w,s}_{e,s}\underline{x} \cdot {}^{w,s}_{e,s}\underline{y}$$

$${}^{w,s}_{e,s}\underline{z} = {}^{w,s}_{e,m}\underline{x} \cdot {}^{w,s}_{e,s}\underline{y} + {}^{w,s}_{e,s}\underline{x} \cdot {}^{w,s}_{e,m}\underline{y}$$

$$\underline{w_z} = \underline{w_x} + \underline{w_y}$$

from which it can be deduced, in a similar way as for the addition, that the performed operation is

$$Z = X \cdot Y$$

The quotient needs the previous invertion of the divisor obtained i.e. by means of an inverse function generator.

3.3. Integration and Derivation

The operation

$$ {}^{s}z = \int {}^{s}x\,dt \qquad \text{or} \qquad {}^{w,s}z = \int {}^{w,s}x\,dt $$

can be easily performed by means of a bidirectional counter (with weighted inputs in the case of the random floating point codification) followed by a digital to stochastic converter if necessary. The operation

$$ {}^{w,s}z = \int_{0}^{{}^{w,s}x} {}^{w,s}y\, d\left[{}^{w,s}x\right] = \int_{0}^{t} {}^{w,s}y \frac{d\left[{}^{w,s}x\right]}{dt} dt $$

depends on the obtention of the derivative.

First Derivative

The substraction between two channels, the variable ${}^{w,s}x$ and the delayed variable ${}^{w,s}x_{\Delta}$, gives:

$$ \frac{1}{2} d\left[{}^{w,s}x\right] = \frac{1}{2}\left[{}^{w,s}x - {}^{w,s}x_{\Delta}\right] \approx \frac{1}{2} \frac{d\left[{}^{w,s}x\right]}{dt} \Delta \qquad (9) $$

Δ being the known time delay between the two channels, Eq. (9) generates a generalized stochastic signal proportional to the derivative.

Higher Order Derivatives

The use of a set of delayed sequences of the variable

$$ {}^{w,s}x_{\Delta}\,,\ {}^{w,s}x_{2\Delta}\,,\ \ldots\ {}^{w,s}x_{n\Delta} $$

can produce, by iterative substractions, the set of approximations:

$$ \frac{\Delta}{2} \frac{d\left[{}^{w,s}x\right]}{dt}\,,\ \frac{\Delta^2}{4} \frac{d^2\left[{}^{w,s}x\right]}{dt^2}\,,\ \ldots \frac{\Delta^n}{2^n} \frac{d^n\left[{}^{w,s}x\right]}{dt^n} $$

3.4. Function Generation

For function generation in stochastic computation a very stimulating method, peculiar to this technology, is envisaged. If at the stochastic converting stage the cumulative probability function of the sampled random noise is not linear, as indicated in Eq. (6) but is an arbitrary monotone increasing function

$$ P_{y} (v) = f (v) $$

then, the encoded variable in Eq. (5) would be the stochastic representation of the function $f(x)$,

$$^{s}z = P_{\mathcal{F}}(x) = f(x)$$

This kind of functional noises can be used in the generalized stochastic encoding process if provision is made of auxiliary logic manipulation of the sign and of the weighting train $\underline{w_x}$.

3.5. Readout

The readout is the value retrieval of stochastically coded variables by means of the averaging process. The implementation of some of the averager transfer functions described in Paragraph 2.2. can be made by three basically distinct methods: a) simulating the transfer function by means of stochastic operators in feedback loops, b) rectangular approximations implementation by delayed weighted sequences, and counters or operational amplifiers, or c) with delay lines or shift registers driving the weighted inputs of analog or digital adders. The method b), using counters, is not to be recommended when the functional noises are not periodic and pseudorandom, as it can lead to integration errors. Further research is being carried on concerning these subjects.

4. FUNCTIONAL RANDOM NOISES

The logic or analog manipulation of multilevel noises $\mathcal{N}$ can produce functional noises $\mathcal{F}$. The logic manipulation method described in the Appendix generates a set of functional noises with the following cumulative probability functions:

$$P_{\mathcal{F}}(v) = v^2,\ v^3,\ v^4\ \ldots\ldots\ v^r.$$

Linear forms of these noises, with exclusive stochastic coefficients that verify,

$$\sum_{i=1}^{i=r} {}^{s}k_i = 1$$

constitute the "generalized functional module" $^{w,s}\left[\mathcal{F}\right]$ of any Taylor series-expandable function , which together with the binary stochastic trains that take account of the sign and weight of the output variable implement the generalized functional noises in digital or analog form.

Analog multilevel functional noises can also be obtained implementing

$$^{A}\mathcal{F} = f^{-1}\left[^{A}\mathcal{N}\right], \qquad f(v) = \text{increasing monotone}$$

with conventional analog function-generators. Again the sign and weight in the general case, must be taken care of with auxiliary binary signals.

The great advantage of the functional noise method, is that given $\mathcal{F}$, any number of functions of independent variables can be simultaneously obtained from it.

5. CONCLUDING REMARKS

The stochastic pulse rate technology is particularly attractive for on-line applications. Besides its use as differential analyzer, the fact that interconnection parameters may be very easily stored and changed through pulse gating, and that complex parallel computing arrays be feasible at moderate prices (for medium accuracies), greatly extend its field of applications. Matrix operation, algebraic equations, finite differences, adaptive structures and even pattern recognition, are just a few of the subjects where either its success has already been proved or for which its potentialities are presently being investigated.

APPENDIX

Let be $\mathcal{N}_1, \mathcal{N}_2 \ldots \mathcal{N}_r$, r statistically independent sampled multilevel noises with $p_{\mathcal{N}}(v) = 1$. These noises are inputs to a multilevel gate that, at each sampling interval t^*, selects the highestly valued. The pdf of the output noise will be:

$$
\begin{aligned}
p_{\mathcal{F}}(v) &= \text{Prob}\ (v < \mathcal{F} \leq v+dv) = \\
&= \text{Prob}\ (v < \mathcal{N}_1 \leq v+dv, \mathcal{N}_2 \leq v+dv, \ldots \mathcal{N}_r \leq v+dv) + \ldots \\
&\ldots + \text{Prob}\ (\mathcal{N}_1 \leq v+dv, \mathcal{N}_2 \leq v+dv, \ \ldots\ v < \mathcal{N}_r \leq v+dv) = \\
&= \text{Prob}\ (v < \mathcal{N}_1 \leq v+dv)\ \text{Prob}\ (\mathcal{N}_2 \leq v+dv)\ \ldots\ \text{Prob}\ (\mathcal{N}_r \leq v+dv) + \ldots \\
&\ldots + \text{Prob}\ (\mathcal{N}_1 \leq v+dv)\ \text{Prob}\ (\mathcal{N}_2 \leq v+dv)\ \ldots\ \text{Prob}\ (v < \mathcal{N}_r \leq v+dv) = \\
&= 1.dv.v.v.\ \ldots\ v + \ldots\ v.v.\ \ldots\ v.1.dv = r.v^{r-1}dv
\end{aligned}
$$

and the cumulative probability function:

$$P_{\mathcal{F}}(v) = \int_0^v p_{\mathcal{F}}(u)du = \int_0^v r.u^{r-1}.du = [u^r]_0^v = v^r$$

REFERENCES

1 Poppelbaum W.J., Afuso C., and Esch J.W.: "Stochastic Computing Elements and Systems". Fall Joint Comp. Conf. 1967.

2 Ribeiro S.T.: "Random-Pulse Machines". IEEE Transactions on Electronic computers, June, 1967.

3 Gaines B.R.: "Stochastic Computer thrives on Noise", Electronics, July, 1967.

4 Ferraté G.A., and Puigjaner L.: "Static and Dynamic Precision in Generalized Stochastic Codification" (in Spanish), Pub. de la C. de Automática, E.T.S. de Ingenieros, Barcelona, 1968.

5 Roberts T.A.: "Analysis and Synthesis of Linear and Nonlinear Shift-Register Generators", Proceedings International Telemetering Conf. London, September, 1963.

6 Tausworthe R.C.: "Random Numbers Generated by Linear Reccurrence Modulo Two", Math of Computation, April, 1965.

7 Agulló J. and Ferraté G.A.: "The Generation of Stochastically Coded Functions with Functional Noise" (in Spanish), Pub. de la C. de Automática, E.T.S. de Ingenieros, Barcelona, 1969.

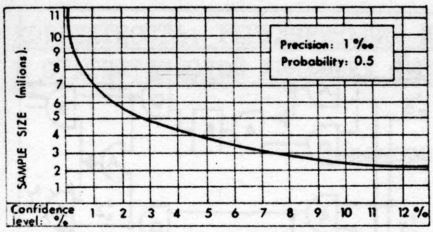

FIG. 1.— Sample size against confidence level for a given precision (0.001) and probability. (0.5)

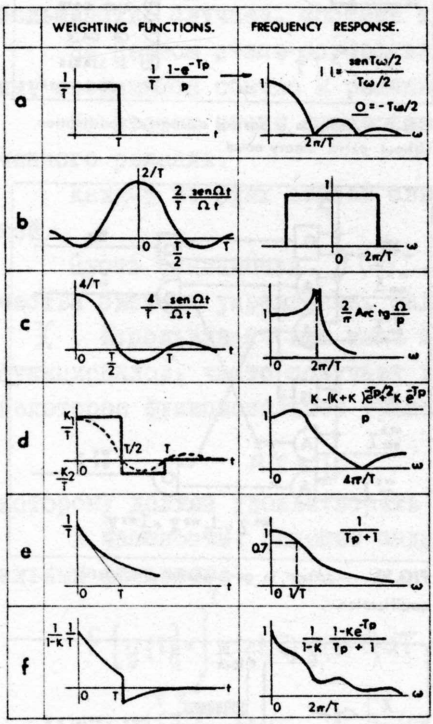

FIG. 2. — AVERAGERS: Weighting functions andthe corresponding frequency responses in schematic form.

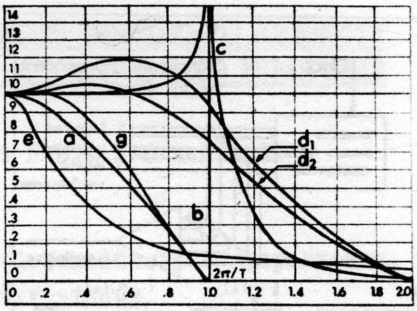

FIG. 3. — Frequency responses of several averagers.

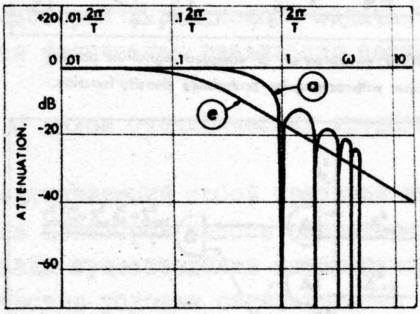

FIG. 4. — Frequency responses of averagers (a), (e) in the Bode's diagram.

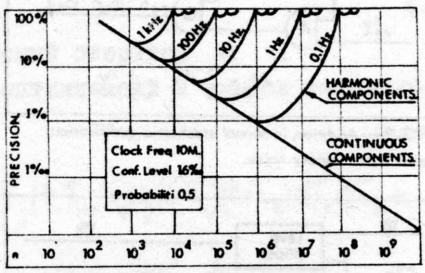

FIG. 5. — Precision against sample size for a rectangular averagers.

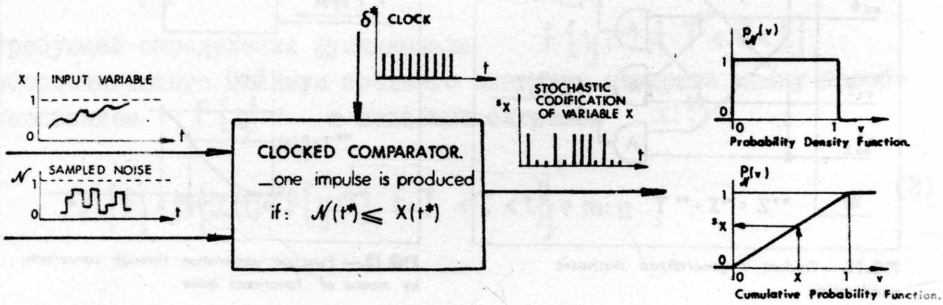

FIG. 6. — Generation of stochastic series with specified Pulse-rate probability.

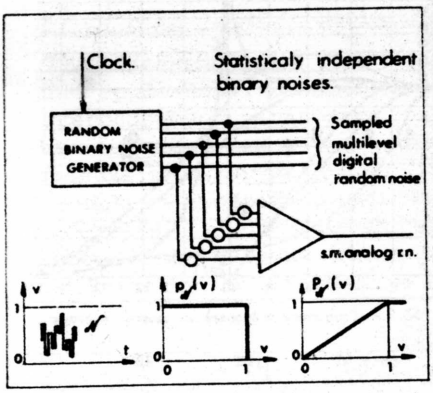

FIG. 7.— Obtention of sampled digital or analog noise with rectangular probability density function.

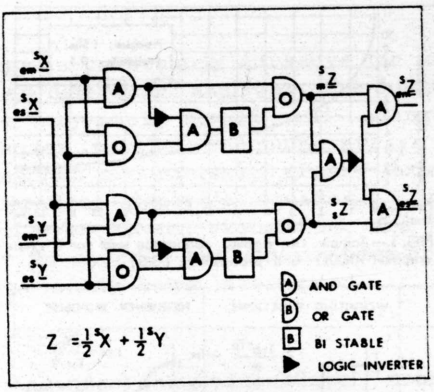

FIG. 8.—Addition, in normal stochastic codification, without extra binary noise.

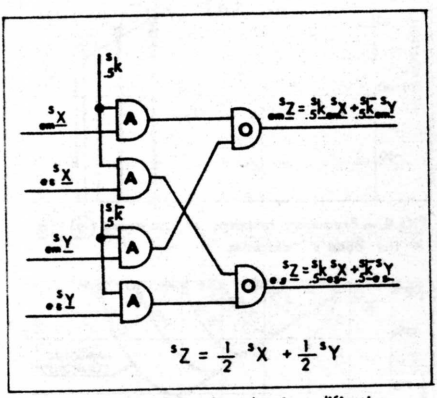

FIG. 9.— Addition, in normal stochastic codification, with extra binary noise.

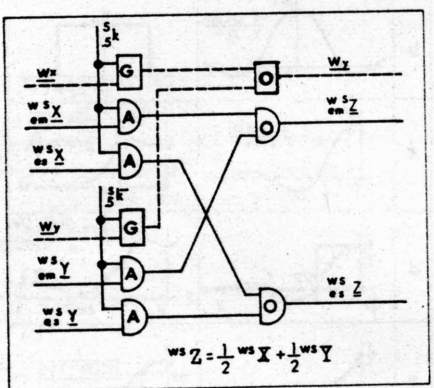

FIG. 10.— Addition in generalized stochastic codification.

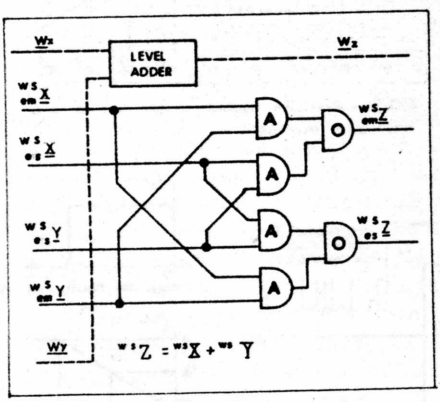

FIG. 11.— Product in generalized stochastic codification.

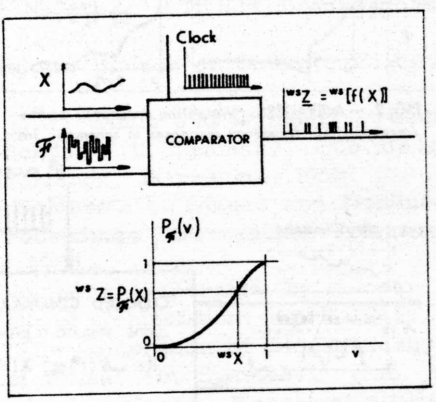

FIG. 12.— Function generation through conversion, by means of functional noise.

КОРРЕКТНОСТЬ, РЕГУЛЯРИЗАЦИЯ И ПРИНЦИП МИНИМАЛЬНОЙ СЛОЖНОСТИ В СТАТИСТИЧЕСКОЙ ДИНАМИКЕ СИСТЕМ АВТОМАТИЧЕСКОГО УПРАВЛЕНИЯ

Солодовников В.В., Ленский В.Л.

Создание оптимальной системы автоматического управления, в большинстве случаев, состоит из двух этапов.

На первом этапе осуществляется математический синтез системы, сводящийся обычно к решению некоторой вариационной задачи.

На втором этапе осуществляется физическая реализация полученного решения.

Каждый из этих этапов связан с рядом специфических трудностей.

Пусть функционал $J(x)$, представляющий собой критерий качества системы управления, задан на некотором классе операторов X. Используя те или иные признаки существования экстремумов функционалов, часто получают в качестве условия оптимальности некоторое функциональное уравнение вида

$$Ax = y, \tag{1}$$

которому должен удовлетворять искомый оператор x.

В частности, решение задачи оптимизации в классе полиномиальных фильтров:

$$F[y(\tau)] = \int_0^T k_1(t,\tau)\,y(t-\tau)\,d\tau + \int_0^T\int_0^T k_2(t,\tau_1,\tau_2)\,y(t-\tau_1)(t-\tau_2)\,d\tau_1 d\tau_2 +$$

$$+ \dots + \int_0^T\int_0^T k_n(t,\tau_1,\dots\tau_n)\,y(t-\tau_1)\dots y(t-\tau_n)\,d\tau_1\dots d\tau_n, \tag{2}$$

требующей определения функционала $F[y(\tau), t-T \le \tau < t]$, обеспечивающего минимум среднего значения квадрата между преобразованием $F[y]$ и желаемым сигналом $x(t)$:

$$J\{F\} = M\left\{x(t) - F[y(\tau), t-T \le \tau < t]\right\}^2 = \min, \tag{3}$$

приводит к системе интегральных уравнений

$$\sum_{i=1}^{n} \int_0^T \dots \int_0^T k_i(t,\tau_1,\dots\tau_i)\,\Gamma_y(t,\tau_1,\dots\tau_i;\theta_1,\dots\theta_j)\,d\tau_1\dots d\tau_j = \tag{4}$$

$$= \Gamma_{xy}(t,\theta_1,\dots\theta_j), \quad j=1,2\dots n,$$

где $$\Gamma_y = M\{y(t-\tau_1)\dots y(t-\tau_i)\,y(t-\theta_1)\dots y(t-\theta_j)\},$$

$$\Gamma_{xy} = M\{x(t)\,y(t-\theta_1)\dots y(t-\theta_j)\} \tag{5}$$

и через k_i обозначены многомерные импульсные переходные функции.

Эту систему уравнений кратко можно записать в виде

$$\Gamma_y\, k = \Gamma_{xy} \tag{6}$$

где k и Γ_{xy} элементы гильбертова пространства.

Во многих случаях точное решение уравнения (I) невозможно и поэтому приходится использовать численные процедуры отыскания приближенного решения.

Кроме того часто исходные данные для решения уравнения (I) задаются с некоторой ошибкой.

Все это приводит к необходимости решения проблемы устойчивости решения уравнения (I) относительно численных процедур и ошибок исходных данных.

Степень устойчивости уравнения (I) относительно вариаций правой части определяется модулем непрерывности обратного отображения:

$$\omega(\delta,X) = \sup \rho(x,x_1) \text{ при } x,x_1 \in X,\ \rho(Ax,Ax_1) \le \delta \tag{7}$$

где функция $\rho(x,x_1)$ определяет метрику в классе X.

Очевидно, что возможная ошибка определения оптимального оператора растет с расширением класса, на котором ищется оптимальный оператор, т.е. имеет место следующее неравенство:

$$\omega(\delta,X_1) \le \omega(\delta,X_2), \qquad X_1 \subset X_2 \tag{8}$$

В случае некорректности уравнения (I), решение становится неус-

тойчивым относительно ошибок задания исходных данных и поэтому синтез оптимальной системы становится принципиально неразрешимой задачей. Но даже в случае корректных задач, мы часто сталкиваемся с чрезвычайной сложностью численных процедур. Это обстоятельство приводит при использовании для расчетов ЦВМ к увеличению машинного времени и памяти, необходимых для получения решения с заданной точностью.

Полученное решение, чтобы его можно было физически реализовать, обычно приходится тем или иным способом аппроксимировать. Желание уменьшить потери в качестве управления из-за отклонения от оптимального решения заставляет увеличивать точность его аппроксимации и расширять класс операторов, на котором ищется экстремум функционала $J(x)$, что приводит к ухудшению технологических и эксплуатационных свойств системы управления (рост стоимости, уменьшение надежности и т.п.).

Действительно, пусть для каждого оператора $x \in X$ определена функция $C_\varepsilon(x)$ - минимально необходимая стоимость практической реализации системы управления, оператор которой аппроксимирует x с точностью ε. Нетрудно видеть, что если $X_1 \subset X_2$, то

$$\max_{x \in X_1} C_\varepsilon(x) \leq \max_{x \in X_2} C_\varepsilon(x) \tag{9}$$

Аналогичное неравенство справедливо для любой функции $V_\varepsilon(x)$, являющейся мерой объема вычислительной работы, необходимой для определения оператора x с точностью ε:

$$\max_{x \in X_1} V_\varepsilon(x) \leq \max_{x \in X_2} V_\varepsilon(x) \tag{10}$$

Пусть для каждого $x \in X$ определена функция $\tau_\varepsilon(x)$ - максимально достижимая вероятность безотказной работы в течение некоторого фиксированного промежутка времени системы управления, оператор которой аппроксимирует x с точностью ε.

Тогда также очевидно, что, если $X_1 \subset X_2$, то

$$\min_{x \in X_1} \tau_\varepsilon(x) \geq \min_{x \in X_2} \tau_\varepsilon(x) \tag{11}$$

Назовем оператором аннулирования Θ оператор, соответствующий случаю отсутствия системы управления. Для оператора аннулирования Θ естественно положить $C_\varepsilon(\Theta) = 0$, $V_\varepsilon(\Theta) = 0$ и

$\tau_\varepsilon(x)=1$. Пусть имеется семейство классов $\mathfrak{M}$, вполне упорядоченное по включению и такое, что

$$\bigcap X \cdot \theta$$

$$X \in \mathfrak{M} .$$

С учетом неравенств очевидно, что желательно искать оператор системы управления, принадлежащий наиболее узкому классу семейства. Но сужение класса приводит к ухудшению критерия качества. Получающееся противоречие может быть разрешено, если сформулировать постановку задачи синтеза систем управления следующим образом:

Пусть задан допустимый уровень качества системы управления q . Требуется среди всех операторов, обладающих заданным уровнем качества, найти оператор, принадлежащий минимальному классу, относительно некоторого семейства, при котором заданный уровень качества достижим.

Предложенной постановке задачи синтеза можно придать более компактную формулировку.

Действительно, рассматривая два оператора x_1, x_2 таких что $x_1 \in X_1, x_2 \in X_2$ и $X_1 \subset X_2$, видим, что определение и реализация оператора x_2 с заданной точностью будет более сложной задачей, чем определение и реализация оператора x_1 с той же точностью.

В соответствии с этим будем называть оператор x_2 более сложным, чем оператор x_1 , если нет никакой информации, кроме той, что

$$x_1 \in X_1 , \; x_2 \in X_2 \quad \text{и} \quad X_1 \subset X_2 .$$

Семейство классов $\mathfrak{M}$ в этом случае играет роль шкалы сложности в множестве, являющемся объединением всех классов семейства.

Теперь задача синтеза систем управления может быть сформулирована в виде следующего принципа, который можно назвать принципом минимальной сложности:

среди всех операторов, обладающих заданным уровнем качества, необходимо выбрать оператор минимальной сложности относительно заданной шкалы.

На ряду с принципом минимальной сложности можно пользоваться также принципом ограниченной сложности, который формулируется

следующим образом:

для фиксированного класса из заданной шкалы сложности найти оператор, обеспечивающий экстремальное значение уровня качества.

Для того, чтобы принципом минимальной сложности можно было пользоваться, необходимо дать методы конструирования шкалы сложности.

Пусть задан некоторый непрерывный функционал $G(x)$, имеющий абсолютный минимум на операторе аннулирования; тогда одно-параметрическое семейство классов $X_t = \{x \mid G(x) \le t\}$ будет обладать свойствами выше определенного семейства классов $\mathfrak{M}$.

В этом случае сужению класса соответствует минимизация функционала $G(x)$.

Применение принципа минимальной сложности приводит к задаче на условный экстремум: найти минимум $G(x)$ при условии $J(x) = q$. Решение этой задачи, как известно, сводится к минимизации функционала вида

$$\lambda G(x) + J(x),$$

где λ - множитель Лагранта.

Другой метод построения шкалы сложности заключается в следующем.

Пусть имеется возрастающая система конечно-мерных классов

$$\Theta \subset X_1 \subset X_2 \subset \dots \subset X_n \subset \dots \subset X,$$

где индекс обозначает размерность класса.

Эта система классов может служить шкалой сложности и применение принципов сложности в этом случае сводится к экстремальным задачам для функций многих переменных. Построение такой шкалы сложности возможно, например, для класса X, имеющего базис $x_1, x_2, \dots x_n, \dots$. В этом случае конечномерным классом X_n будет множество всевозможных линейных комбинаций из базисных элементов.

Возможны также другие методы построения шкалы сложности.

Поскольку, вообще говоря, система классов, образующая шкалу сложности, неединственна, то для каждой конкретной постанов-

ки задачи синтеза необходимо строить шкалу сложности, учитывающую специфику данной задачи и возможности физической реализации системы.

Применение принципов сложности целесообразно не только потому, что получаемые системы обладают лучшими технологическими и эксплуатационными свойствами, но также и потому, что получаемые необходимые признаки существования экстремумов вида (I), при соответствующем выборе шкалы сложности, являются корректными задачами в смысле А.Н.Тихонова. Для корректности признаков существования экстремумов вида (I) достаточно, чтобы классы $X_t = \{x \mid G(x) \leqslant t\}$ были компактны и оператор Эйлера для функционала $G(x)$ был вполне непрерывным[1]. В этом случае функционал $G(x)$ будет регуляризирующим функционалом для уравнения вида (I). Некорректность уравнений вида (I) не является теоретической возможностью с которой можно практически не считаться. Большинство задач статистической динамики, сводящихся к линейным уравнениям первого рода являются некорректными. Действительно, синтез системы оптимальной по минимуму средквадратической ошибки сводится к определению минимума квадратичного функционала вида

$$J(x) = (Ax, x) - 2(x, y), \qquad (12)$$

где A - положительный самонапряженный линейный оператор;
(x, y) - символ скалярного произведения элементов x и y.
Как известно, чтобы элемент x обеспечивал минимум функционала $J(x)$, необходимо и достаточно, чтобы x удовлетворял линейному уравнению вида (I).

Решение уравнения (I) можно записать в виде интеграла Стильтьеса, используя спектральное разложение самосопряженного оператора A:

$$x = \int_0^\infty \frac{1}{\lambda} d E_\lambda y, \qquad (13)$$

где E_λ - разложение единицы, соответствующей самосопряженному оператору.

Решение существует в том и только в том случае, если

$$\int_0^\infty \frac{1}{\lambda^2} d\,(E_\lambda y, y) < \infty . \qquad (14)$$

В случае, если нуль есть предельная точка спектра оператора A, то интеграл (14) может быть расходящимся, в зависимости от распределения спектральной меры $(E_\lambda y, y)$. Это означает, что уравнение (I) может не иметь решения с конечной нормой.

Такая ситуация возникает, например, во всех случаях синтеза линейных фильтров, когда оптимальная импульсная переходная функция содержит в своем составе δ - функции и ее производные, не интегрируемые в квадрате. Как известно такие фильтры физически переализуемы. Пусть элемент y задан с некоторой ошибкой h, тогда, в силу аддитивности обратного оператора, квадрат нормы ошибки решения выразится следующим образом:

$$\|\delta x\|^2 = \int_0^\infty \frac{1}{\lambda^2} d\,(E_\lambda h, h) . \qquad (15)$$

Отсюда видно, что квадрат нормы ошибки может быть каким угодно в зависимости от распределения спектральной меры $(E_\lambda h, h)$.

Как известно, математическая задача называется корректной, если решение этой задачи существует, единственно и непрерывно зависит от вариации исходных данных. В этом смысле задачи статистической динамики, сводящиеся и уравнению (I) - некорректны.

Для применения принципов сложности функционал сложности можно задавать, например, в виде

$$G(x) = (Bx, Bx), \qquad (16)$$

где B - некоторый положительный и непрерывный оператор.

В частном случае, если B - единичный оператор $G(x) = (x,x) = \|x\|^2$ - квадрат нормы элемента x.

Применение принципа минимальной сложности в этом случае приводит к следующей вариационной задаче.

Найти минимальное значение функционала $G(x) = (x,x)$ при условии $I(x) = (Ax,x) - 2(x,y) = q$. Элемент, дающий ее решение, удовлетворяет следующему уравнению второго рода:

$$\lambda x + Ax = y \qquad (17)$$

В этом случае применение принципа минимальной сложности эквивалентно слабой регуляризации в смысле А.Н.Тихонова. Если B -

дифференциальный оператор, то применение принципов сложности эквивалентно сильной регуляризации.

Используя спектральное разложение оператора A, можно построить шкалу сложности следующим образом

$$X_t = \left\{ x \mid x = \int_t^{\infty} \frac{1}{\lambda} dE_\lambda u \right\},$$

где $\|u\| < \infty$

Тогда $X_{t_1} \subset X_{t_2}$, если $t_1 > t_2$

Можно показать, что решение вариационной задачи минимизации функционала (7) на классе X_t дается следующей формулой

$$R_\delta[y] = \bar{x} = \int_t^{\infty} \frac{1}{\lambda} dE_\lambda y .$$

Покажем, что $R_\delta[y]$ является регуляризирующим алгоритмом в смысле А.Н.Тихонова.

Действительно

$$\|x - \bar{x}\| = \| \int_0^{\infty} \frac{1}{\lambda} dE_\lambda y - \int_t^{\infty} \frac{1}{\lambda} dE_\lambda (y+h) \| \leqslant$$

$$\leqslant \| \int_0^{t} \frac{1}{\lambda} dE_\lambda y \| + \| \int_t^{\infty} \frac{1}{\lambda} dE_\lambda h \| \leqslant \left[\int_0^{t} \frac{1}{\lambda^2} d(E_\lambda y, y) \right]^{\frac{1}{2}} + \frac{1}{t} \|h\| .$$

Отсюда следует: для любого $\varepsilon > 0$ существуют такие $\delta(\varepsilon)$ и $t(\varepsilon)$, что если $\|h\| < \delta(\varepsilon)$, то $\|x - \bar{x}\| \leqslant \varepsilon$.

В случае дискретного спектра регуляризирующий алгоритм эквивалентен определению решения в виде линейной комбинации собственных элементов, собственные числа которых больше t.

Покажем теперь, что регуляризация по А.Н.Тихонову эквивалентна применению принципов сложности.

Метод регуляризации, применительно к уравнению (I), сводится к решению уравнения

$$\lambda Bx + Ax = y,$$

где B - оператор Эйлера некоторого функционала $G(x)$, называемого регуляризирующим функционалом, λ - параметр регуляризации, выбираемый в зависимости от нормы погрешности исходных данных. Легко видеть, что уравнение (I8) получается при минимизации функционала

$$\lambda G(x) + J(x) . \qquad (I9)$$

Минимизацию функционала (I9) можно рассматривать как задачу минимизации функционала $G(x)$ при заданном значении функционала $J(x)$.

Рассмотрим множество $X_t = \{x | G(x) \leq t\}$. Минимизация функционала $G(x)$ эквивалентна сужению множества A_t в том смысле, что для

$$t_1 < t_2 \qquad A_{t_1} \subseteq A_{t_2} .$$

Отсюда следует, что регуляризация влечет минимизацию сложности. Рассмотрим несколько примеров применения принципов сложности. Пусть на вход линейной системы управления поданы:[3] управляющее воздействие, состоящее из заданного аналитически сигнала $g(t)$ и стационарного случайного сигнала $m(t)$ с нулевым средним значением и корреляционной функцией $R_m(t,\tau)$, а также помеха $n(t)$ с корреляционной функцией $R_n(t,\tau)$. Система должна наилучшим образом воспроизводить управляющий сигнал $y(t) = g(t) + m(t)$ и подавлять помеху $n(t)$. В качестве критерия качества системы управления примем функционал

$$J = \varepsilon_g^2 + \mu^2 \varepsilon_{ск}^2 , \qquad (20)$$

т.е. сумму квадратов динамической ошибки ε_g и среднеквадратической ошибки $\varepsilon_{ск}$ с некоторым весом.

Применяя несложные преобразования, легко получить выражение J через импульсную переходную функцию искомой системы в следующем виде:

$$J = q^2(t) + \mu^2(t) R_m(t,t) - 2\int_0^t \left[q(t)q(\tau) + \mu^2(t) R_m(t,\tau)\right] k(t,\tau)\, d\tau +$$

$$+ \int_0^t \int_0^t \left\{ q(\theta) q(\tau) + \mu^2(t) \left[R_m(t,\theta) + R_n(t,\theta)\right]\right\} k(t,\theta)\, k(t,\tau)\, d\tau\, d\theta . \qquad (21)$$

Минимизируя это выражение относительно $k(t,\tau)$, получим, что необходимое и достаточное условие минимума J сводится к тому, чтобы функция $k(t,\tau)$ удовлетворяла интегральному уравнению

$$\int_0^t \left\{ q(t)q(\tau) + \lambda^2(t) \left[R_m(\tau,\theta) + R_n(\tau,\theta)\right]\right\} k(t,\theta)\, d\theta = \qquad (22)$$

$$= q(t)\, q(\tau) + \mu^2(t)\, R_m(t,\tau) , \qquad t > \tau .$$

Это параметрическое интегральное уравнение Фредгольма первого рода. Решение такого уравнения может содержать δ - функции и вызывать поэтому существенные трудности при практической реализации.

Пусть функционал J может принимать в каждый момент времени заданное допустимое значение

$$J = \varepsilon_g^2 + \mu^2 \varepsilon_{ск}^2 = q(t). \tag{23}$$

В качестве функционала сложности примем следующий:

$$G = \int_0^t k^2(t,\tau)\, d\tau. \tag{24}$$

В этом случае необходимое и достаточное условие для $k(t,\tau)$ будет иметь вид параметрического интегрального уравнения Фредгольма второго рода:

$$\lambda k(t,\tau) + \int_0^t \{g(\theta)g(\tau) + \lambda^2(t)[R_m(\tau,\theta) + R_n(\tau,\theta)]\}\, k(t,\theta)\, d\theta =$$
$$= g(t)\, g(\tau) + \mu^2(t)\, R_m(t,\tau), \quad t > \tau. \tag{25}$$

Так как правая часть этого уравнения представляет собой ограниченную функцию, то и решение будет необходимо содержаться в классе ограниченных функций, т.е. не будет содержать δ - функций.

В случае, если бы функционал сложности был интегралом от квадрата некоторого дифференциального оператора вида:

$$G = \int_0^t \left[\sum_{i=0}^{n} a_i k^{(i)}(t,\tau)\right]^2 d\tau, \tag{26}$$

мы получили бы интегро-дифференциальное уравнение, решение которого имело бы ограниченную производную по крайней мере до порядка n.

Таким образом мы видим, что выбором функционала сложности можно управлять дифференциальными свойствами импульсной переходной функции.

Это обстоятельство становится чрезвычайно важным при определении структуры системы. Действительно, предположим, что любая импульсная переходная функция $k(t,\tau)$, принадлежащая к некоторому классу, может быть аппроксимирована со сколь угодной точностью суммами вида

$$P_m(t,\tau) = \sum_{i=1}^{m} c_i \varphi_i(t)\, \psi_i(\tau). \tag{27}$$

Как известно[3], в этом случае динамическая система может быть реализована в виде m звеньев первого порядка, причем дифференциальное уравнение каждого звена имеет вид:

$$D_i(p,t)r(t)=M_i(p,t)n(t),\quad i=1,\dots,n. \tag{28}$$

Дифференциальные операторы $D_i(p,t)$ и $M_i(p,t)$ определяются из выражений:

$$\left.\begin{aligned} D_i(p,t)r(t)&=\frac{dr(t)}{dt}-\frac{1}{\varphi_i(t)}\cdot\frac{d\varphi_i(t)}{dt}r(t),\\ C_i\psi_i(\tau)&=M_i^*(p,\tau)\frac{1}{\varphi_i(\tau)}\end{aligned}\right\} \tag{29}$$

где $M_i^*(p,\tau)$ - оператор, сопряженный $M_i(p,t)$.

Для получения максимально простых структур естественно искать среди множества сумм порядка m такие, которые дают наилучшее приближение: $E_m(k)=\inf\|k(t,\tau)-P_m(t,\tau)\|.$ (30)

Как известно[4] скорость убывания $E_m(k)$ целиком зависит от дифференциальных свойств $k(t,\tau)$ и тем выше, чем выше порядок дифференциальности.

В случае стационарных сигналов определение линейной стационарной системы, оптимальным образом преобразующей заданный сигнал в желаемый с функционалами сложности в виде интегралов от квадрата импульсной переходной функции и ее производных, приводит к синтезу систем с минимальной полосой пропускания. В работе[5] показано, что чем меньше полоса пропускания, тем проще физическая реализация системы.

В качестве другого примера использования принципов сложности, основанном на втором, указанном выше, методе построения шкалы сложности, рассмотрим задачу синтеза нелинейного дискретного фильтра с конечной памятью, наилучшим образом преобразующего заданный стационарный случайный сигнал в желаемый, также стационарный. Решение этой задачи в общем виде неизвестно и вряд ли может быть получено.

Рассмотрим шкалу сложности $\mathfrak{M}=\{F_m\}$, где F_m класс всевозможных полиномов вида

$$\sum_{i=0}^{n-1} h_i x_{t-i} + \sum_{0 \le i_1 \le i_2}^{n-1} h_{i_1 i_2} x_{t-i_1} x_{t-i_2} + \dots + \sum_{0 \le i_1 \le i_m}^{n-1} h_{i_1 \dots i_m} x_{t-i_1} x_{t-i_2} \dots x_{t-i_m} .$$

(3I)

Как нетрудно убедиться, такая система действительно является шкалой сложности и, в силу того, что любая непрерывная функция может быть сколь угодно точно аппроксимирована полиномом вида (3I) достаточно высокой степени, замыкание объединения классов UF_m содержит класс непрерывных функций. Фильтры вида (3I) называют фильтрами Колмогорова-Габора[6].

В соответствии с принципами сложности, синтез нелинейного фильтра сводится к решению вариационной задачи на классе F_m. Как легко видеть эта задача эквивалентна задаче определения проекции случайной величины y_t на подпространство образованное всевозможными линейными комбинациями случайных величин

$$x_{t-i}, x_{t-i_1} x_{t-i_2}, \dots, x_{t-i} x_{t-i_2} \dots x_{t-i_m} .$$

Решение этой задачи дается решением системы нормальных уравнений метода наименьших квадратов относительно весовых коэффициентов

$$h_i, h_{i_1 i_2}, \dots, h_{i_1 \dots i_m} .$$

Однако такой путь определения нелинейной системы все еще связан с рядом трудностей.

Первая трудность заключается в быстром росте числа подлежащих определению весовых коэффициентов с ростом степени полиномиального фильтра m и памяти n. Нетрудно подсчитать, что это число будет равно

$$N = \sum_{K=1}^{m} \frac{n(n+1)\dots(n+k-1)}{k!} .$$

(32)

Естественно, что большой объем весовых коэффициентов предъявляет повышенные требования к объему запоминающего устройства и быстродействию дискретного фильтра.

Кроме того, с ростом числа весовых коэффициентов быстро растет объем вычислительной работы, необходимой для их определения.

Вторая трудность заключается в том, что многие методы определения весовых коэффициентов становятся неустойчивыми относи-

тельно ошибок приближенных вычислений.
Так, например, определение весовых коэффициентов по методу нормальных уравнений приводит к плохо обусловленным системам линейных алгебраических уравнений, причем с ростом порядка системы обусловленность ухудшается.

В силу ограниченной точности приближенных вычислений,мы получаем практически некорректную задачу.

Таким образом, естественной мерой сложности фильтра Колмогорова-Габора является число определяемых весовых коэффициентов.

В соответствии с принципами сложности, необходимо в пространстве, образованном всевозможными линейными комбинациями случайных величин

$x_{t-i}, x_{t-i_1} x_{t-i_2}, \ldots, x_{t-i_1} x_{t-i_2} \ldots x_{t-i_m}$ найти подпространство минимальной размерности, для которого удовлетворяется заданный уровень ошибки или, при заданной размерности, найти подпространство, для которого ошибка минимальна.Нахождение таких подпространств можно осуществить, применяя перебор. Однако такой путь связан с большим объемом вычислительной работы.

Для минимизации сложности фильтров вида (3I) можно воспользоваться теоретико-числовыми методами приближенного анализа[7].

Действительно, во многих случаях заданный дискретный случай сигнал x_t и желаемый дискретный случайный сигнал y_t могут быть связаны с некоторыми непрерывными сигналами $x(t)$ и $y(t)$. Это будет, например, в том случае, если x_t и y_t получены из случайных непрерывных сигналов $x(t)$ и $y(t)$ квантованием.

Если это не так, то всегда можно с помощью интерполяции построить сигналы $x(t)$ и $y(t)$, имеющие в дискретные моменты времени те же значения, что и рассматриваемые дискретные сигналы.

Без ограничения общности мы можем считать, что $(n-1)\Delta = 1$,

где Δ - интервал квантования случайного непрерывного сигнала, порождающего дискретный сигнал.

Рассмотрим одно из слагаемых степени $\mathfrak{d}$ в (3I)

$$x_{t-i_1} x_{t-i_2} \ldots x_{t-i_\mathfrak{d}}$$

В силу сделанного замечания это слагаемое можно рассматривать как значение случайного поля, заданного на единичном $\mathfrak{d}$ -

мерном гиперкубе, в точке $M=(\Delta i, \Delta i_2, \dots, \Delta i_s)$.

Само поле будет иметь вид

$$f(t,\tau_1,\dots,\tau_s)=x(t-\tau_1)\dots x(t-\tau_s) \quad (33)$$

Если непрерывный сигнал $x(t)$ имеет непрерывную в среднем производную порядка α, то можно доказать, что случайное поле (33) может быть аппроксимировано следующим выражением

$$x(t-\tau_1)\dots x(t-\tau_s)=\sum_{k=0}^{n-1} x(t-a_1 k(\mathrm{mod} n))\dots x(t-a_s k(\mathrm{mod} n))\,\theta_k(\tau_1,\dots,\tau_s)+R,\dots \quad (34)$$

где $\theta_k(\tau_1,\dots,\tau_s)$ - некоторые базисные функции, $a_1,\dots,a_s$ — оптимальные коэффициенты, выбираемые согласно[7], $p \bmod n$ - означает остаток от деления p на n. И R оценивается следующим неравенством

$$\overline{R^2} < C\,\frac{\ln^{\gamma} n}{n^{\alpha-1/2}} \quad (35)$$

Поскольку значение поля в любой точке может быть представлено в виде (34), то тем самым оно может быть представлено и в точках с координатами вида

$$\tau_1=\frac{1}{n-1}i_1,\ \tau_2=\frac{1}{n-1}i_2,\dots,\ \tau_s=\frac{1}{n-1}i_s.$$

Но это означает, что любое произведение s-й степени может быть выражено приближенно как линейная комбинация n произведений s-степени вида

$$x_{t-a_1 k(\mathrm{mod} n)}\,x_{t-a_2 k(\mathrm{mod} n)}\cdots x_{t-a_s k(\mathrm{mod} n)}, \quad (36)$$

то-есть

$$x_{t-i_1}\cdots x_{t-i_s}=\sum_{k=0}^{n-1}\theta_k(i_1,\dots,i_s)\,x_{t-a_1 k(\mathrm{mod} n)}\cdots x_{t-a_s k(\mathrm{mod} n)}+R. \quad (37)$$

Формула (37) показывает, что произведения, отличные от произведений вида (36), являются "лишними", так как могут быть приближенно получены из (37)

Таким образом оптимальным подпространством размерности будет подпространство, образованное всевозможными линейными комбинациями системы случайных величин

$$x_{t-k}, x_{t-a_1 k(\mathrm{mod} n)}, x_{t-a_2 k(\mathrm{mod} n)},\dots, x_{t-a_1 k(\mathrm{mod} n)}\cdots x_{t-a_m k(\mathrm{mod} n)}.$$

Иными словами, синтез нелинейного дискретного фильтра оказалось возможным свести к решению вариационной задачи

$$M\left\{\left|y_t-F\left[x_\tau, t-(n-1)\le\tau\le t\right]\right|^2\right\}=\min.$$

на классе фильтров вида $\sum_{K=0}^{n-1} h_K^1 x_{t-K} + \sum_{K=0}^{n-1} h_K^2 x_{t-a_1 K(modn)} x_{t-a_2 K(modn)} +$

$$+ \cdots + \sum_{K=0}^{n-1} h_K^m x_{t-a_1 K(modn)} \cdots x_{t-a_m K(modn)} \quad . \qquad (39$$

Для оценки эффективности минимизации сложности можно привести следующую таблицу числа весовых коэффициентов полного фильтра Колмогорова-Габора (N) и минимизированного фильтра (N_1) при длине памяти фильтра $n = 10$.

m	I	2	3	4	5	6
N	I0	65	285	I000	3002	8007
N_1	I0	20	30	40	50	60

Проведенные численные расчеты показали, что несмотря на существенную минимизацию сложности потери качества воспроизведения желаемого сигнала составляют доли процента. Причем увеличение сложности фильтра вида (39) не приводит к существенному улучшению качества воспроизведения желаемого сигнала, а во многих случаях приводит к ухудшению качества из-за ухудшения обусловленности системы нормальных уравнений.

В заключение следует отметить.

Развитие теории оптимального управления до настоящего времени в основном, в направлении разработки математического аппарата для решения тех или иных задач синтеза, в основном, без учета необходимости последующей реализации, без учета сложности алгоритмов определения оптимальных систем.

Но по существу эти две стороны процесса синтеза не отделимы, что особенно наглядно проявляется в самонастраивающихся системах.

Поэтому в настоящее время большую актуальность приобретают исследования по изменению традиционной постановки задач синтеза оптимальных систем с целью учета вопросов сложности как алгоритмов оптимизации так и физической реализации последних.

Литература

I. Тихонов А.Н. О решении некорректно поставленной задачи и методе регуляризации. ДАН.СССР, 1963г., 151, № 3.

2. В.В.Солодовников, В.Л.Ленский. Регуляризация некорректных задач статистической динамики систем автоматического управления. ДАН, № 6, 172, 1967г.

3. В.В.Солодовников. Статистическая динамика линейных систем автоматического управления, Физматгиз, 1960.

4. Н.И.Ахиезер. Лекции по теории аппроксимации Ф.М.1965г.

5. Ньютон Дж.К, Гулд Л.А., Кайзер Дж.Ф. Теория личейных следящих систем. Физматгиз. 1961 г.

6. А.Г.Ивахненко, В.Г.Лана. Кибернетические предсказывающие устройства. Наукова думка. 1965 г.

7. Н.М.Коробов. Теоретико-числовые методы в приближенном анализе. Ф.М.1963 г.

COMPUTATION OF OPTIMUM CONTROL FOR A ROBOT IN A PARTIALLY UNKNOWN ENVIRONMENT

W. G. Keckler and R. E. Larson
Stanford Research Institute
Menlo Park, California, U.S.A.

I INTRODUCTION

The optimization of systems in which stochastic effects are present has been studied extensively by a number of researchers.[1,2,3,4,5,6] An extremely general formulation of these problems has been called by Meier[6] the combined optimum control and estimations problem; a solution to this problem has been formulated using dynamic programming.[6,7,8] Even though several theoretical papers have been written on this subject, there have been very few examples worked out for any cases but the linear gaussian problem.[9,10]

This paper first describes a dynamic programming approach that was proposed for the problem of optimally controlling a robot, equipped with sensors, that is operating in an unknown environment.* A methodology is presented for formulating a class of stochastic control problems in which there are informational variables that specify the degree of knowledge about the physical state of the system as well as the physical variables of the type encountered in most control applications. These problems are present in a number of areas; the robot example discussed here is related to the general problem of unmanned exploration of a hostile, inaccessible environment, while another formulation of this type has been developed for mission reliability problems.[11] The detailed calculations required to implement this approach are also described. Dynamic programming is shown to be feasible for handling system equations, performance criteria, and constraints that simultaneously involve physical variables and informational variables. Another aim of this paper is to demonstrate the relationship of a number of concepts from system theory to the combined optimum control and estimation problem; among the concepts discussed are dual control[12] and value of information.[13]

In the robot problem, computational complexity increases exponentially with the number of physical and informational state variables. Thus, many problems of

---

*A robot project is under current study at SRI, "Application of Intelligent Automata to Reconnaissance," Contract AF 30(602)-4147, SRI Project 5953.

interest are too unwieldy to solve rigorously on present-day computers. In order to attack these problems, a heuristic method based on the optimization algorithm has been devised. It is thus possible in this paper to analyze the relation between the heuristic methods and optimization approaches for a concrete example. The results of heuristic methods are also compared with the performance of humans in some representative cases.

The remainder of the paper is organized as follows: Section II defines the mission of the robot and develops a state-space formulation for the problem of controlling it. Section III describes a dynamic programming solution method and presents results for some illustrative examples. The computational difficulties of this algorithm are also discussed. Section IV describes a heuristic solution method with greatly reduced computational requirements. The performance of the heuristic method is evaluated in terms of both the optimum performance and that obtained by a group of control theorists given the same information. Section V evaluates the potential of both optimization techniques and heuristic methods for solving practical problems involving stochastic effects.

II PROBLEM STATEMENT AND FORMULATION

The problem discussed in this paper can be stated as follows: A robot is attempting to perform a specified task in an unknown environment. As part of this task, it must travel to a particular spatial location, called the goal. There are a number of barriers that the robot cannot traverse; the presence or absence of these barriers is not known a priori, but their potential locations are all known. The robot has a sensor system that can detect the presence or absence of these barriers; however, there is a cost associated with using these sensors. The problem is to devise a policy for the robot that will find that path to the goal that minimizes the sum of the cost of using the sensor system and a cost that reflects the length of the path.

This problem can be formulated as an optimum-control problem in which the minimum-distance requirement and the sensor cost are joined in a common-cost criterion. Minimization of this criterion requires that state and control variables be defined that completely represent the situation and options of the robot. Complete specification of the state variables requires not only that the location of the robot in its environment be specified, but also that the state of knowledge about the barrier configuration at each position be considered. Control options

include both movements to a number of adjacent unblocked areas and use of the sensing equipment to determine the presence or absence of barriers at more distant sites. Since additional knowledge can prevent wasteful moves, the expected benefit of observing must be considered each time that a control decision is made. The control decision is a function of present location in the environment and of the present state of knowledge of barrier locations.

A simple problem that illustrates the features of this formulation is shown pictorially in Fig. 1. In this example, an automaton is attempting to find its way through the environment shown in Fig. 1 to a goal located at the upper left-hand square, which has coordinates ($x_1 = 1$, $x_2 = 1$). As shown in Fig. 2, the automaton can move one square either up, down, left, or right.

There are two squares on which it is possible that barriers are present; namely ($x_1 = 1$, $x_2 = 2$) and ($x_1 = 2$, $x_2 = 3$). The automaton is not allowed to pass through these barriers. Initially, the automaton does not know if these barriers are present; instead, it knows that there is a barrier in ($x_1 = 1$, $x_2 = 2$) with probability 0.4 and a barrier in ($x_1 = 2$, $x_2 = 3$) with probability 0.5. The robot can always "see" one move ahead--i.e., if it is within one move of a barrier location, it can find out if the barrier is there or not. For a certain price, which is expressed as a specified fraction of a move (.3 moves), the automaton can make an observation of all squares that are two moves away (see Fig. 2). The objective is to find the policy for the automaton that reaches the goal square ($x_1 = 1$, $x_2 = 1$) from any initial square while minimizing the expected value of the sum of moves and penalties for making observations.

This problem can most easily be put into the desired framework by defining the complete state description of the system (information state) to be four-dimensional vector $\underline{x}$,

$$\underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}, \tag{1}$$

where

x_1 = horizontal coordinate in Fig. 1

x_2 = vertical coordinate in Fig. 1

x_3 = state of knowledge about barrier at ($x_1 = 1$, $x_2 = 2$)

x_4 = state of knowledge about barrier at ($x_1 = 2$, $x_2 = 3$).

The first two variables are quantized to the values

$$x_1 = 1, 2, 3$$
$$x_2 = 1, 2, 3, 4 \quad . \tag{2}$$

The latter two variables are quite different from the usual state variables associated with dynamic systems. Each variable can take on three different values as follows

$$x_3 = P, A, Q$$
$$x_4 = P, A, Q \quad , \tag{3}$$

where

P = barrier is known to be present

A = barrier is known to be absent

Q = absence or presence of barrier is not known.

The control vector, $\underline{u}$, has three components. They are

$$\underline{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} \quad , \tag{4}$$

where

u_1 = negative change in x_1*

u_2 = positive change in x_2*

u_3 = decision to make an observation.

The variable u_3 takes on two values

$$u_3 = L, N \quad , \tag{5}$$

where

L = an observation is made

N = no observation is made.

The set of admissible controls is thus

$$U = \left\{ \begin{bmatrix} 1 \\ 0 \\ N \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ N \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ N \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \\ N \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ L \end{bmatrix} \right\}, \tag{6}$$

corresponding to move right, move left, move up, move down, and make an observation. The first two system equations can be written as

---

*For historical reasons, the convention was adopted that a positive move was up and to the right in Fig. 1, while the direction of position x_1 and x_2 was as shown there.

$$x_1(k+1) = x_1(k) + u_1(k)$$
$$x_2(k+1) = x_2(k) - u_2(k) \qquad (7)$$

The uncertainty about the barriers is taken into account by defining a random forcing function vector $\underline{w}$ with two components,

$$\underline{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \qquad (8)$$

w_1 = presence or absence of barrier at ($x_1 = 1$, $x_2 = 2$)

w_2 = presence or absence of barrier at ($x_1 = 2$, $x_2 = 3$) .

These variables can take on the values

$$w_1 = B, R$$
$$w_2 = B, R \quad , \qquad (9)$$

where

B = barrier is present

R = barrier is absent .

This vector affects only the state variables x_3 and x_4. In writing the system equations for these two variables, it is useful to define two auxiliary variables, m_1 and m_2. The variable m_1 takes on the value 1, if the control is such that the presence or absence of the barrier at ($x_1 = 1$, $x_2 = 2$) will be determined; $m_1 = 0$ otherwise. The variable m_2 is 1 if the control will determine the presence or absence of the barrier at ($x_1 = 2$, $x_2 = 3$); $m_2 = 0$ otherwise.

For $u_3 = N$, m_1 is equal to 1, if the move chosen causes the next square to be within one move of the barrier at ($x_1 = 1$, $x_2 = 2$). For $u_3 = L$, m_1 is equal to 1, if the barrier at ($x_1 = 1$, $x_2 = 2$) is within two moves of the present square. Otherwise, $m_1 = 0$. Similar conditions can be written for m_2.

The system equation for x_3 can thus be written in the form

$$x_3(k+1) = f_3[x_3(k), u_3(k), m_1(k), w_1(k)] \quad , \qquad (10)$$

where f_3 is defined by Table 1, and where

$$p(w_1 = B) = 0.4$$
$$p(w_1 = R) = 0.6$$
$$p(w_2 = B) = 0.5$$
$$p(w_2 = R) = 0.5$$

Table 1: Next Value of x_3 for Automaton

$x_3(k)$	$u_3(k)$	$m_1(k)$	$w_1(k)$	$x_3(k+1)$
P	-	-	-	P
A	-	-	-	A
Q	N	0	-	Q
Q	N	1	B	P
Q	N	1	R	A
Q	L	0	-	Q
Q	L	1	B	P
Q	L	1	R	A

- = value of $x_3(k+1)$ is the same for all values of this variable

The equation for x_4 is

$$x_4(k+1) = f_4[x_4(k), u_4(k), m_2(k), w_2(k)] \quad , \tag{12}$$

where f_4 is specified by a table similar to that in Table 1.

The performance criterion, which is to be minimized, is the sum of moves and penalties for observations. This criterion can be written as

$$J = \sum_{k=0}^{\infty} \ell[\underline{x}(k), \underline{u}(k)] \quad , \tag{13}$$

where $\ell[\underline{x}(k), \underline{u}(k)]$ is specified as in Table 2. In this table, it is assumed that the penalty for an observation is 0.3 moves.

Table 2: Value of $\ell[\underline{x}(k), \underline{u}(k)]$ for Automaton

$x_1(k)$	$x_2(k)$	$u_3(k)$	$\ell[\underline{x}(k), \underline{u}(k)]$
1	1	-	0
1	≠1	N	1.0
1	≠1	L	0.3
≠1	1	N	1.0
≠1	1	L	0.3
≠1	≠1	N	1.0
≠1	≠1	L	0.3

The constraints are that the set of admissible controls and the set of admissible states are as defined above. Also, if a barrier is present, a move to that square is forbidden. The extension of this formulation to a larger problem is straightforward, and the optimum solution for the general case is presented in the next section.

III SOLUTION AND RESULTS FOR THE COMPLETE-STATE REPRESENTATION

The problem formulation of the preceding section satisfies the conditions for use of the technique "approximation in policy space."[1,14] Two of these conditions are that the system equations and constraints have no explicit dependence on time and that the performance criterion be the sum of time-invariant terms over an infinite number of stages. Another condition is that there exist a state and control that have a single-stage cost of zero. Finally, there must exist a finite-length path from this state to any other state that is unblocked with probability 1.0. If the latter condition is not met, the expected cost of reaching the goal is infinite, and it is necessary to use the technique "iteration in policy space."[15]

Because of the stochastic nature of this problem, the particular version of approximation in policy space described below is used. A proof that this method converges to the optimum solution can be obtained by using the results shown in Ref. 14.

In order to start the procedure, an initial guess of the optimal policy, $\underline{\hat{u}}^{(0)}(\underline{x})$, is made. One such control policy that is easy to compute is to pick the control that is optimal if all barriers are present, the worst possible situation. The corresponding minimum cost function, $I^{(0)}(\underline{x})$, is found by solving a deterministic minimum-path-length problem. Formally, this function is obtained from solving

$$I^{(0)}(\underline{x}) = \ell[\underline{x}, \underline{\hat{u}}^{(0)}(\underline{x})] + I^{(0)}\left\{\underline{f}[\underline{x}, \underline{\hat{u}}^{(0)}(\underline{x}), \underline{w}^{*}]\right\} , \qquad (14)$$

where $\underline{f}$ represents the system equation vector defined in Eqs. (7), (10), and (12), and where $\underline{w}^{*}$ corresponds to $w_1 = B$, $w_2 = B$. A particularly efficient method for solving this equation is to note that $I^{(0)}(\underline{x}) = 0$ for all states corresponding to physical location at the goal, and then to compute outward from this square.

When $I^{(0)}(\underline{x})$ has been found for all $\underline{x}$, a new policy $\underline{\hat{u}}^{(1)}(\underline{x})$ is found by solving

$$I'(\underline{x}) = \min_{\underline{u}} \left[\ell(\underline{x}, \underline{u}) + E_{\underline{w}} \left\{ I^{(0)} [\underline{f}(\underline{x}, \underline{u}, \underline{w})] \right\} \right] \quad . \tag{15}$$

Since $\underline{w}$ is not allowed to be a stochastic variable, it is necessary to take the expected value of $I^{(0)}$ in Eq. (15). The policy $\underline{\hat{u}}^{(1)}(\underline{x})$ is the value of $\underline{u}$ for which the minimum in Eq. (15) is attained. However, $I'(\underline{x})$ is not $I^{(1)}(\underline{x})$, the minimum cost corresponding to policy $\underline{\hat{u}}^{(1)}(\underline{x})$, because $I^{(0)}$ appears inside the braces. The function $I^{(1)}(x)$ is found from

$$I^{(1)}(\underline{x}) = \ell[\underline{x}, \underline{\hat{u}}^{(1)}(\underline{x})] + E_{\underline{w}} \left\{ I^{(1)} [\underline{f}(\underline{x}, \underline{\hat{u}}^{(1)}(\underline{x}), \underline{w})] \right\} \quad . \tag{16}$$

This equation can be solved iteratively as in Ref. 14.

In general, a new policy $\underline{\hat{u}}^{(j+1)}(\underline{x})$ is formed from knowledge of $I^{(j)}(\underline{x})$, using

$$I'(\underline{x}) = \min_{\underline{u}} \left[\ell(\underline{x}, \underline{u}) + E_{\underline{w}} \left\{ I^{(j)} [\underline{f}(\underline{x}, \underline{u}, \underline{w})] \right\} \right] \quad . \tag{17}$$

The corresponding minimum cost function $I^{(j+1)}(\underline{x})$, is found by solving

$$I^{(j+1)}(\underline{x}) = \ell[\underline{x}, \underline{\hat{u}}^{(j+1)}(\underline{x})] + E_{\underline{w}} \left\{ I^{(j+1)} [\underline{f}[\underline{x}, \underline{\hat{u}}^{(j+1)}(\underline{x}), \underline{w}]] \right\} \quad . \tag{18}$$

As already indicated, convergence to the true optimum can be proved for this case.

When this method is applied to the problem described in the preceding section, the results shown in Table 3 are obtained. Several interesting effects can be observed in Table 3. The first is Feldbaum's dual control effect;[12] this effect is said to occur whenever the optimal control is used to gain more information about the system instead of optimizing the performance directly. The effect is illustrated in this example by the optimal controls of looking instead of moving toward the goal. By imposing a penalty for the observation, the system is able to make a decision whether to gather more information or to continue moving on the basis of the information gathered so far. This occurs at $(x_1, x_2, x_3, x_4) = (1, 4, Q, P)$ and $(x_1, x_2, x_3, x_4) = (1, 4, Q, Q)$.

Another effect that can be observed is Howard's value of information.[15] This effect is related to the cost that should be paid in order to gain information. Again, this is illustrated here in the decision of whether or not to make an observation; if the cost of 0.3 moves does not pay for the increase in performance, then the observation should not be made. At the two points where the observation is made, it is interesting to note the optimal control and minimum cost for the

Table 3: Solution to Automaton Problem

Present State				Optimal Control			Minimum Cost
x_1	x_2	x_3	x_4	u_1	u_2	u_3	1
1	1	-	-	-	-	-	0
2	1	-	-	-1	0	N	1
3	1	-	-	-1	0	N	2
1	2	-	-	0	1	N	1
2	2	-	-	0	1	N	2
3	2	-	-	0	1	N	3
1	3	A	-	0	1	N	2
1	3	P	A	1	0	N	4
1	3	P	P	0	-1	N	8
2	3	-	-	0	1	N	3
3	3	-	-	0	1	N	4
1	4	A	-	0	1	N	3
1	4	P	A	0	1	N	5
1	4	P	P	1	0	N	7
1	4	P	Q	1	0	N	6
1	4	Q	A	0	1	N	3.8
1	4	Q	P	0	0	L	4.9
1	4	Q	Q	0	0	L	4.5
2	4	-	A	0	1	N	4
2	4	A	P	-1	0	N	4
2	4	P	P	1	0	N	6
2	4	Q	P	-1	0	N	5.9
3	4	-	-	0	1	N	5

- = optimal control and minimum cost are the same for all values of this variable

Table 4: Comparison of Optimal Control and Minimum Cost With and Without the Observation Control

x_1	x_2	x_3	x_4	Minimum Cost With Observation	Minimum Cost Without Observation	Optimal Control With No Observation		
						u_1	u_2	u_3
1	4	Q	P	4.9	5.4	0	1	N
1	4	Q	Q	4.5	4.6	0	1	N

case where the decision to observe is <u>not</u> allowed. It is seen from Table 4 that at $(x_1, x_2, x_3, x_4) = (1, 4, Q, P)$, the net profit of making the observation is 0.5 moves, while at $(x_1, x_2, x_3, x_4) = (1, 4, Q, Q)$, the net profit is 0.1 moves.

The technique described above has been implemented in a computer program, and other larger examples have been solved. One of these is the one shown in Fig. 3.

There are four squares on which barriers are possible; the presence of each one has a probability of .7. The price of a look two moves ahead is still .3. The state vector of this system has six state variables -- the two position coordinates and four information variables. Since this problem has 1620 possible states, the complete solution cannot be presented here. However, in only 29 of these situations is the look option chosen when it is available. The improvement resulting from the use of this option at a cost of .3 moves ranges from .3 to 1.2 moves. When the a priori probability of the presence of each barrier is reduced to .5, 25 situations require looks, and the maximum saving is .8 moves. When the barrier probability is reduced to .3, the number of look situations declines to 12, but the maximum gain is again 1.2. A discussion of the reasons for the situations is interesting, but not central to the theme of this paper.

Though the complete state description does lead to optimal solutions, the procedure does have one very serious shortcoming: the number of states that must be considered builds up very rapidly with the number of potential barriers --

$$N = n_{x_1} \cdot n_{x_2} \cdot 3^{n_b}, \tag{19}$$

where

N - total number of states

n_b - number of barriers

$n_{x_1} \cdot n_{x_2}$ - number of states which are used to represent the environment

Thus, a problem which requires a 10 x 10 grid to represent the environment and includes 20 barriers requires 1.9×10^8 storage locations. Although information states which do not affect the control (e.g. all but one of the states at a location one move from the goal) clearly are unnecessary, elimination of these will not reduce the problem to manageable proportions. The next section describes a formulation that can yield control policies based on fewer computations than the procedure described in this section.

IV REDUCED-STATE SPACE FORUMLATION

The problem stated in Sec. II does not require that the control policy for all possible states be computed. If the problem is attacked in the context of the particular example the robot is facing, considerable savings in computer time are obtained. The solution is initiated with the robot located at a par-

ticular position in the environment and equipped with a specific a priori knowledge about the barrier configuration it must overcome. This includes knowledge of the possible locations of barriers and the probability of each one being present. The information about the barriers is not formalized as state variables, but instead is used as parameters for the decision-making process. The state of the system is represented by only the quantized position coordinates, x_1 and x_2. The control vector has the same five options given in Eq. (6), and the state equations are given by Eq. (7). If the control option "Look" is chosen, or if a move yields additional information about the barrier configuration, the control policy is revised to include the new information. The performance criterion continues to be the sum of moves and penalties for observations.

The computation procedure is much like that in Sec. III. The initial cost estimate $J^{(0)}(\underline{x})$ is obtained from control policy $\underline{u}^{(0)}(\underline{x})$, which assumes that all barriers are present. The first iteration of the control $\underline{u}^{(1)}(x)$ is found by solving Eq. (15) at all states. However, only transitions to other accessible locations are considered when a move option is examined. The look option is evaluated by determining the value of each possible barrier configuration and the probability of this configuration occurring. The control option chosen at each state is that which tentatively minimizes the cost of reaching the goal. The resulting value of the cost, $J^{(1)}(\underline{x})$, is associated with this location. Iterations are repeated in the state space until the process converges. Convergence to a final value is assured in the procedure and occurs in four or five iterations.

This procedure results in actions on the part of the robot simulation that appear to be a learning process. When it is positioned at a particular cell in the environment, it is able to determine its next move through consideration of the entire barrier and goal configuration. If this move increases its knowledge of the environment, the computational procedure is reapplied to determine if a better control policy can be found. Thus, as the robot moves through the environment searching for the shortest path to the goal in a systematic manner, it is also mapping the environment. If it were again placed in this same environment, this "experience" would enable it to reach the goal much more quickly than the first time.

Since this heuristic is not optimal, it is difficult to determine how effective it is. However, in one test, the performance of a group of control

theorists was compared with the performance of the heuristic in the barrier configurations shown in Figs. 4 and 5. The barriers were placed according to the probabilities shown in these figures. Solutions for six cases based on the configuration of Fig. 4 were obtained by the heuristic and by four people. In the resulting 24 comparisons of control theorists with the heuristic, the heuristic performed better 11 times, poorer 7 times, and there were 6 ties. Five cases based on Fig. 5 were presented to 5 people and to the heuristic. The heuristic won 16 of the resulting 25 trials, lost 6 and tied 3. Tables 5 and 6 compare the performance of the heuristic with the average performance of the humans in each case.

Table 5

Comparison of Performances by Heuristic
and the Control Theorists on the Configuration of Figure 4

Case	Heuristic		Average of 4 Control Theorists	
	Moves	Looks	Moves	Looks
1	12	2	13.75	.75
2	12	2	13.00	1.25
3	14	2	14.50	.75
4	12	0	12.00	1.00
5	14	0	14.00	0
6	14	2	14.50	2.00

Heuristic's Combined Record: Win 11, Lose 7, Tie 6

Table 6

Comparison of Performances by Heuristic
and the Control Theorists on the Configuration of Figure 5

Case	Heuristic		Average of 4 Control Theorists	
	Moves	Looks	Moves	Looks
1	12	1	13.2	.8
2	22	1	21.6	1.0
3	18	2	18.8	1.6
4	12	1	18.0	1.4
5	12	2	14.0	.8

Heuristic's Combined Record: Win 16, Lose 6, Tie 3

The configuration shown in Fig. 5 is considerably more complex than that in Fig. 4; there are 100 states instead of 64 and the probabilities are no longer the simple .5 that humans can deal with intuitively. As shown above, the heuristic, which considers each configuration in the same systematic manner, scores better relative to the control theorists as the problems become more complex. Even though the heuristic is not optimal, it is able to deal with these problems more successfully than can humans.

V CONCLUSION

The paper shows how additional state variables can be defined in a dynamic programming formulation to include information-gathering considerations in the decision-making process. An example illustrates the change in control policies that result from these considerations. Although the complete state description of the system does provide a good example of the value of information, the number of possible states increases so fast that solutions are not feasible for moderate-sized problems.

A heuristic that includes much of the viewpoint of the complete description has been devised and applied to several grid and barrier configurations. Instead of considering all possible barrier configurations, this procedure just deals with the situation actually facing the robot. As new information is obtained, the solution of the problem is repeated. The performance of the route-finding heuristic has been compared with the performance of a group of control theorists, and it has been found to be superior even on moderate-sized problems. As the problems become larger, this heuristic will be far superior. Procedures motivated by the same viewpoint could be used to automate many searches now done by humans.

REFERENCES

1. R. Bellman, Adaptive Control Processes (Princeton, N. J., Princeton University Press, 1961).

2. H. J. Kushner, "Some Problems and Recent Results in Stochastic Control," 1965 IEEE Conv. Record.

3. H. J. Kushner, Stochastic Stability and Control, (Academic Press, Inc.) New York/London, 1967).

4. W. M. Wonham, "Stochastic Problems in Optimal Control," 1963 IEEE National Conv. Record, Vol. 11 (2), pp. 114-124.

5. M. Aoki, Optimization of Stochastic Systems, (Academic Press, Inc., New York/London, 1967).

6. L. Meier, III, "Combined Optimum Control and Estimation," Proc. 3rd Ann. Allerton Conf. on Circuit and System Theory, (Urbana, Ill., October 1965) pp. 109-120.

7. R. Sussman, "Optimal Control of Systems with Stochastic Disturbances," Electronics Research Lab., University of California, Berkeley, Ser. 63, No. 20, (November 1963).

8. M. Aoki, "Optimal Bayesian and Min-Max Controls of a Class of Stochastic and Adaptive Dynamic Systems," Preprints, IFAC Tokyo Symp. on Systems Engrg. for Control System Design (1965), pp. 11-21-11-31.

9. R. S. Ratner, "Optimum Design of Reliable and Maintainable Systems" (to be published).

10. T. L. Gunckel, II, and G. F. Franklin, "A General Solution for Linear Sampled-Data Control," Trans. ASME, J. Basis Engr., Ser. D. Vol. 85, pp. 197-201 (March 1963).

11. P. D. Joseph and J. T. Tou, "On Linear Control Theory," Trans. AIEE (Applications and Industry), Vol. 80, pp. 193-196 (September 1961).

12. A. A. Feldbaum, "Dual Control Theory--I," Avtom. i Telemekh., Vol. 21, pp. 1240-1249 (September 1960).

13. R. A. Howard, "Information Value Theory," IEEE Trans. Systems Science Cybernetics, Vol. SSC-2, pp. 22-26 (August 1966).

14. R. E. Larson, "A Survey of Dynamic Programming Computational Procedures," IEEE Transactions on Automatic Control (Survey Papers), Vol. AC-12, pp. 767-774 (December 1967).

15. R. A. Howard, "Dynamic Programming and Markov Processes," (John Wiley & Sons, Inc., New York, 1960).

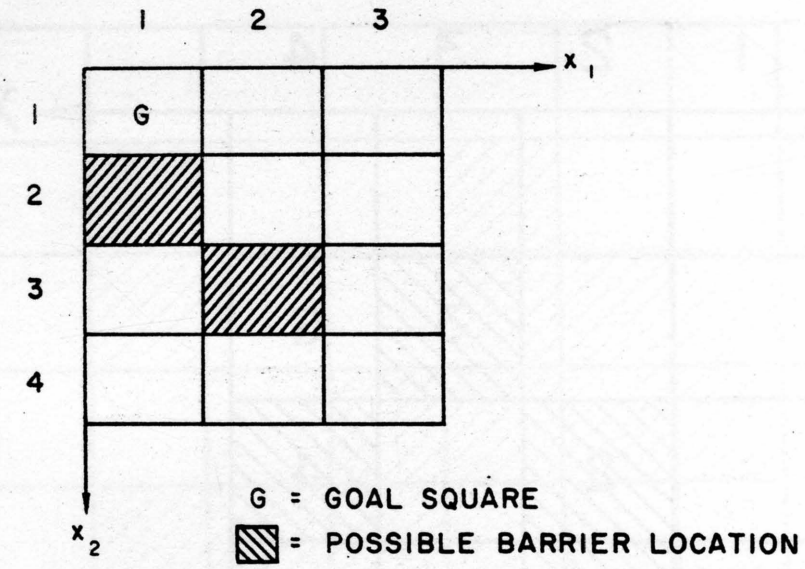

Fig. 1 Automaton Environment

	2	
2	1	2
1	0	1
2	1	2

0 = PRESENT SQUARE
1 = SQUARES THAT CAN BE REACHED IN ONE MOVE
2 = SQUARES THAT CAN BE REACHED IN TWO MOVES

TA-742531-6

Fig. 2 Automaton Moves

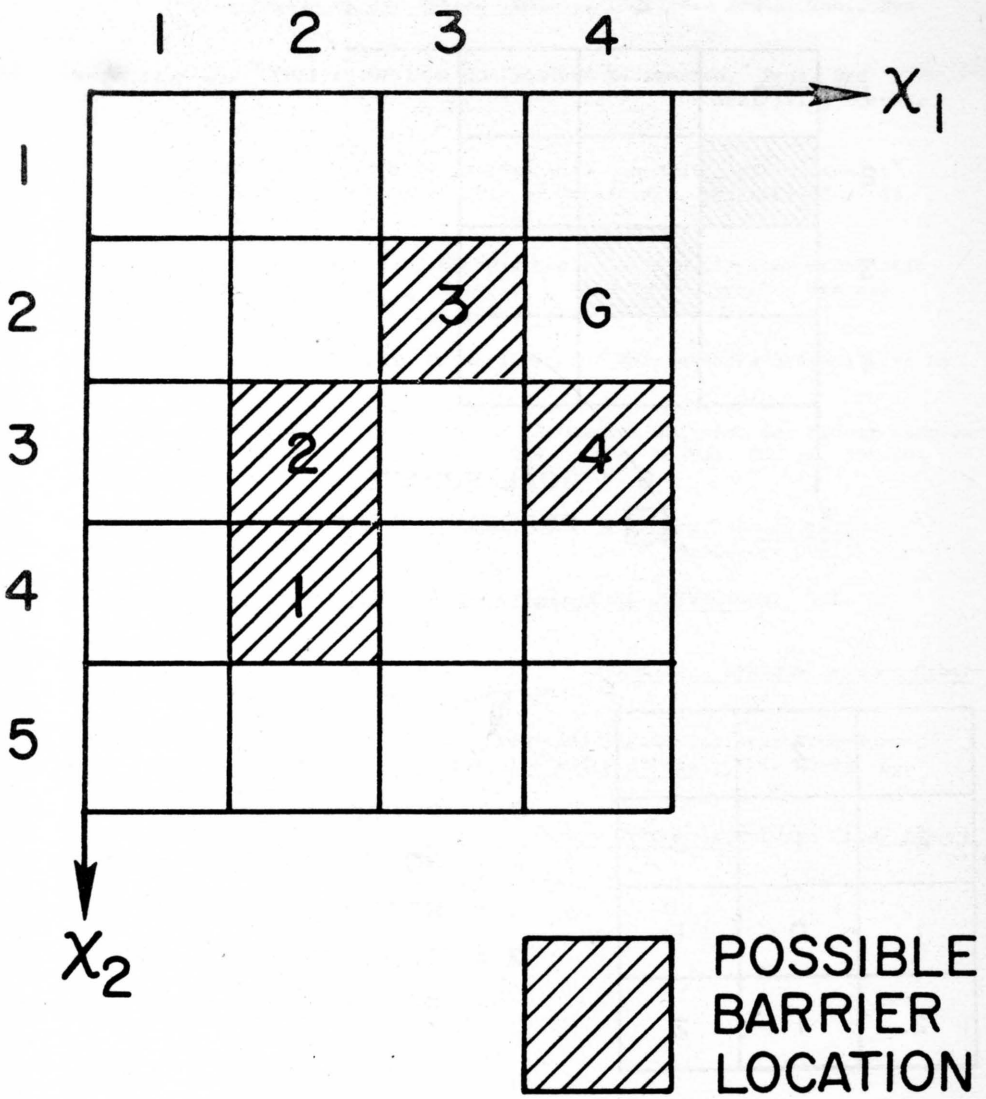

Fig. 3 Enlarged Automaton Environment

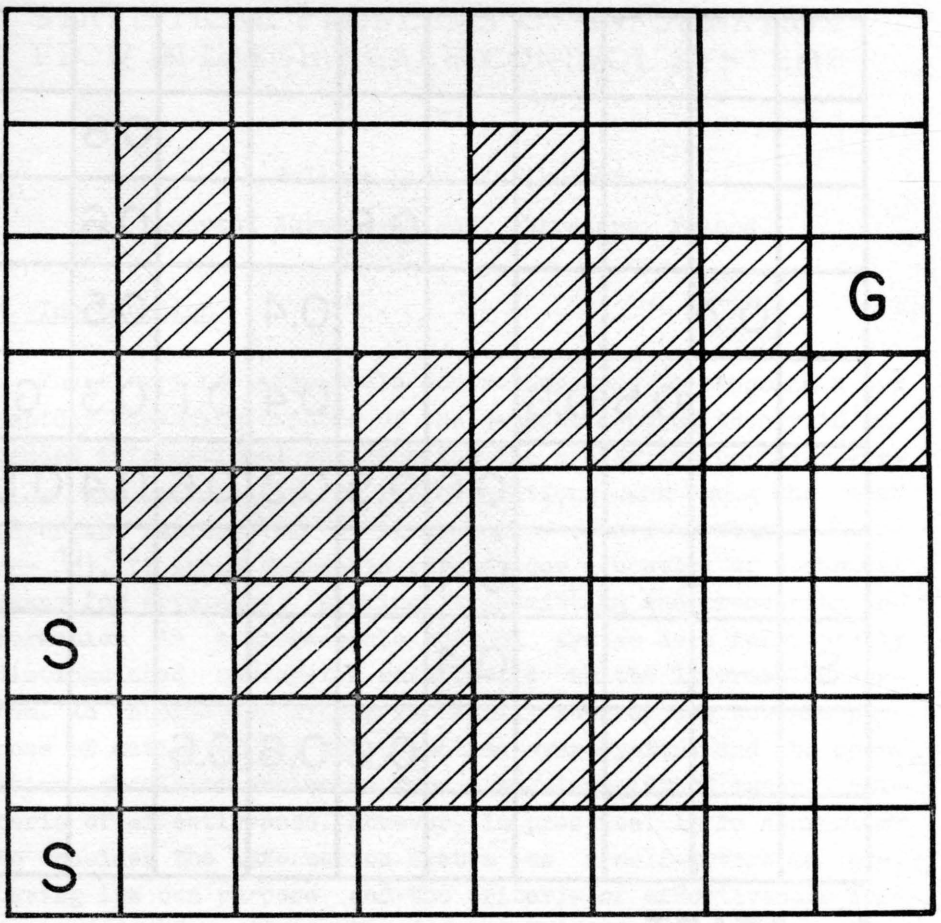

Fig. 4 First Environment Presented to Control Theorists. Probability of a barrier on a shaded square is .5.

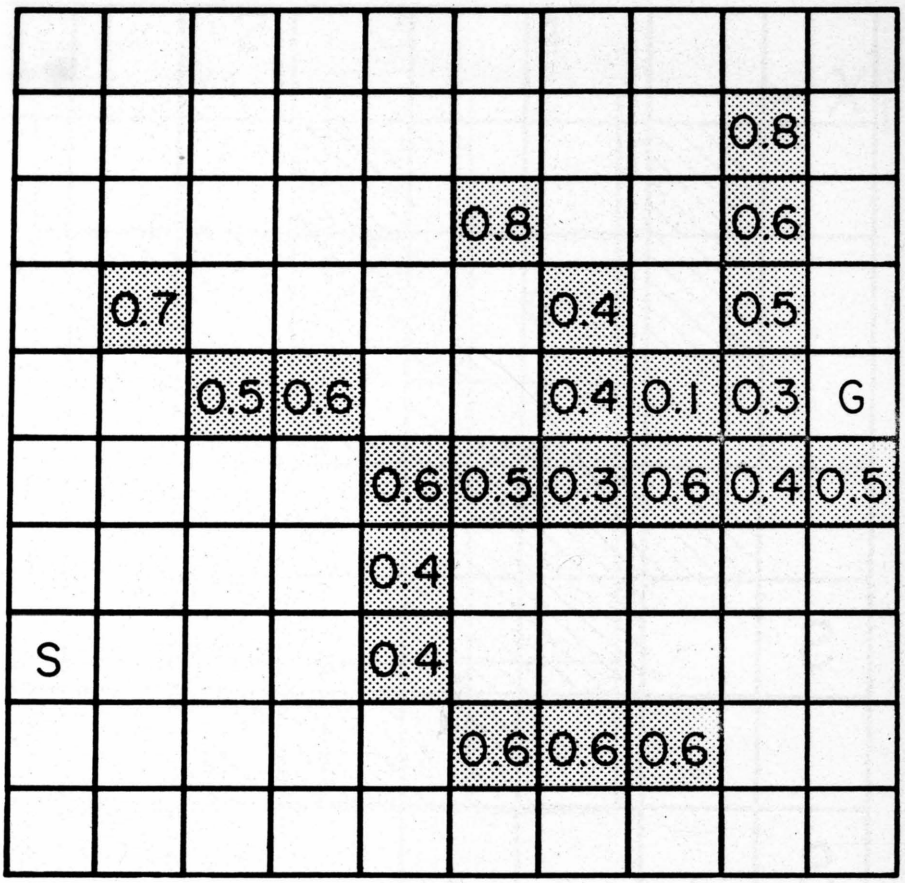

Fig. 5 Second Environment Presented to Control Theorists. Probability of a barrier on a shaded square equals the number in the square.

STATISTICAL PROBLEMS OF INFORMATION FLOW IN LARGE-SCALE CONTROL SYSTEMS

by

Juliusz Lech Kulikowski

Instytut Automatyki PAN, Warszawa, Poland

1. Introduction

The theory of large-scale control systems is becoming a new rapidly expanding branche of the general control theory. An extended informational set securing the system, its normal operation and optimization of its operations rank among the most important features of a large-scale control system (Lerner [10]). It seems desirable to consider a complex of technical means for obtaining, storing, transmitting and processing information in a large-scale control system as a functionally distinguished sub-system referred to as the information system. An information system is usually subordinate to the purpose of action of the corresponding over-system, and its operations should be evaluated from the viewpoint of general criteria of effectiveness. However, in practice, it is convenient to consider the information system as a self-governing one, having its own purpose and the criteria of effectiveness formulated in the specific language of the information theory. The development of the information systems theory as a branch of science situated between the large-scale control systems theory and the information theory is a logical consequence of this fact.

The development of rational methods of choosing an organization for the information system designed, which wolud be suitable for reaching the purposes of the over-system, is one of the basic practical problems of the information system theory. Optimization of such systems is a task which may be solved approximatively, step by step only, since the information systems are highly composite, and usually include sets of universal and special electronic computers, autonomous memory

units, data transmission lines, commutators and a considerable number of auxiliary devices. The effectiveness of such an optimization method depends to great extent on the ability of estimating the loss or even gain which can be reached through such or other reorganization of the system. A theoretical analysis of the system as well as its numerical model realized by a computer or observation of a real object can be accepted as a basis of such an estimation.The order in which the methods have been named corresponds to the increasing costs involved, but, simultaneously, to the concreteness of the results they afford. Since the projecting of an informational system is itself an economical problem, it is necessary to develop uniformly all the methods mentioned above: theoretical, numerical and experimental. This paper deals with the first group of methods only. An attempt is made to propose a methamatical tool suitable for describing the statistic - dynamical properties of information and its value flowing through the system.

2. General Problems

(a) Information system organization.There are many concepts of interpreting the system organization.Hence the information system theory is considered here to be a branch of science,and the organization of the informational system is interpreted here in a narrower but more concrete sense.It is considered to consist of three components:

- a generalized graph (di-multi-graph) describing the spatial structure of the system;
- an algebraic structure describing the functional structure of the system;
- a class of operational rules governing the information processing and transmission as a function of time.

The choice of an optimum system organization thus becomes a strictly mathematical problem. However, the actual state of applied mathematics including operational research permits partial optimization of selected system parameters only;the optimization procedure is a kind of iterative solution of a multi-

-variable functional problem. If the other factors of the system are fixed, the spatial location of the information-processing points (the nodes of the di-multi-graph) and the channel capacities of the transmission-lines (the weights corresponding to the directed borders of the di-multi-graph) may, for instance, be optimized; the linear programming methods including some modifications of the transport-algorithm may be used here (Ford, Fulkerson [2]). The optimization of functional structure and of the operational rules are considerably less homogeneous problems from the mathematical point of view. The problems become more complicated if the spatial and functional structures and the operational rules are subordinate to some additional requirements of the system reliability. However, it does not seem necessary to consider a structural redundance of the system as a fourth factor of its organization, because reliability must be taken into account and optimized together with other factors.

The functional structure and the operational rules of the system will be considered in a more detailed form after some new definitions are introduced.

(b) The information value. In order to subordinate the information system criteria of effectiveness to those governing the over-control system, a measure of the information value will be introduced. This value was first defined by Kharkevitsh [5], Bongard [1] and others authors. The definition given here is based on the present author's concept published in an earlier paper [7].

A basic element of the information process, as it will be considered in the informations systems theory, is the part distinguished for its spatial, temporary and contents features, and called the message. Unlike the general theory of information, the micro-structure of messages, their subdivision into phrases, words, symbols, etc. will not be considered here. It will be supposed that every operation performed on a message: its generation in a source, storage in the memory unit, transmition to another point, processing the information it brings and so on, implies some costs on the one hand, and some gain due to the increase of the control effectiveness index r of

the over-system the message belongs to on the other. The value v_ξ of the information contained by the massage ξ can then be measured by finding the difference between the statistical means of this index calculated before and after receiving the message:

$$v_\xi \overset{\text{def}}{=} \underset{(r|\xi)}{E}\{r\} - \underset{(r)}{E}\{r\} . \qquad (1)$$

The control effectiveness index r (the net gain, for example) should be taken assuming a fixed algorithm of action of the over-system in both complete and uncomplete information situations. The symbol $\underset{(r)}{E}$ denotes a statistical averaging over the random variable r , and sometimes may be considered to be the conditional random variable; the condition is then indicated behind a stroke.

If it is assumed that the index r equals 1 when a fixed aim is reached, and equals 0 other wise, it may be proved that the above defined information value is a logarithm of the one defined by Kharkevitsh [5] . It follows that the definition given here is more general. However, it is based on the assumption that the messages are random events and the corresponding gains reached by the system are measurable in the probabilistic sense.

Now, let us suppose that several operations are to be performed by the information system on the messages, as for example:

(i) the message is to be obtained from an information source $\xi_1 = a_1(\xi_0)$,

(ii) the message ξ_1 is to be stored in a memory unit up to a fixed moment of time $\xi_2 = a_2(\xi_1)$,

(iii) some data ξ_3 necessary for further processing are to be selected $\xi_3 = a_3(\xi_2)$,

(iv) the data ξ_3 are to be transmitted to another point of the system $\xi_4 = a_4(\xi_3)$,

(v) the message ξ_4 is to be delivered to the over-system $\xi_5 = a_5(\xi_4)$,

where the input ξ_0 and the output ξ_5 will not be con-

sidered in the context of the information system. Every one of the operations named above involves a certain amount of costs. On the other hand, the gain is realized only after the final operation a_5 is performed. Optimizing such a series of operations would be possible in a total sense only, with the well known disadvantages from the point of view of calculations. This difficulty can be reduced if the value of information is distributed over the sequence of functionally related operations in some arbitrary manner, as for example:

messages:		ξ_0		ξ_1		ξ_2		ξ_3		ξ_4		ξ_5
operations:	$\longrightarrow$	a_1	$\longrightarrow$	a_2	$\longrightarrow$	a_3	$\longrightarrow$	a_4	$\longrightarrow$	a_5	$\longrightarrow$	
gain:		-		-		-		-		r		
costs:		c_1		c_2		c_3		c_4		c_5		
forthcoming costs:		C_1		C_2		C_3		C_4		C_5		
values:		$r-C_1$		$r-C_2$		$r-C_3$		$r-C_4$		$r-C_5$		r

where:

$$C_i \overset{\text{def}}{=} \sum_{j=i}^{5} c_j \,, \quad i = 1, \ldots, 5 \,.$$

Although the information value has been defined in a general form, it cannot be calculated without some further assumptions; this ressembles the situation connected with defining the value of a product component when only the final product brings a real income.

The assumption of additivity of information values:

$$v_{\xi' \cap \xi''} \equiv v_{\xi'} + v_{\xi''} \tag{2}$$

(where the $\cap$ denotes a logical alternative of the message) is also desirable for simplifying the considerations. This assumption puts some restrictions to the distinction of messages which should be independent both in logical and in statistical sense, but it is usually true if the information system serves

a large number of independent clients, as is the case with post-office communication systems, information storage and retrieval systems, open data-processing systems and so on. However, it is untrue in strongly-centralized control systems, where the information carried by the messages concerns functionally dependent objects. In such a case the assumption of additivity of information values cannot be justified but by simplicity reasons.

Whatever may be the restrictions, the mathematical methods make it possible to generalize the results, namely, it is possible to define the information value both additive and comparable in a generalized sense. This is possible supposing that the information value is an element of a partially ordered linear space (Kulikowski [7]).

Let $\mathcal{X}$ be a linear system with the operations of adding its elements and multiplying its elements by real numbers defined on it and satisfying the well known conditions of the commutativity, associativity, etc.

Let θ be a null-element of $\mathcal{X}$, then for every x

$$0 \cdot x = \theta \quad , \tag{3}$$

the multiplication of x by the real number 0 being interpreted in the sense of the linear system $\mathcal{X}$. Let us suppose that a property of generalized "positiveness" of some elements of $\mathcal{X}$ exists

$$x \succ \theta \quad , \tag{4}$$

which satisfies the following conditions (Kantorovitsh, Vulich, Pinsker [4]):

(i) $x \succ \theta$ excludes $x = \theta$;

(ii) if $x \succ \theta$ and $y \succ \theta$ then $x + y \succ \theta$, the adding of elements being interpreted in the linear system $\mathcal{X}$ sense;

(iii) for every $x \succ \mathcal{X}$ there exists an element $y \succ \theta$ such that $y \succ x$, it is $y + (-1) \cdot x \succ \theta$;

(iv) if $x \succ \theta$ and $a > 0$, a being a given real number, then $a \cdot x \succ \theta$;

(v) for every up-limited subset $\{x\} \subset \mathcal{X}$ there exists a strong upper-bound $\sup \{x\}$.

A set $\mathcal{X}$ which satisfies the above-given conditions will be called a K-space (a partially ordered linear space of Kantorovitsh). If the "positiveness" $x \succ \theta$ is suitably defined, the real-axis, the complex-variable plane, any euclidean space, the space of real or complex matrices of given dimensions, the space of scalar or vector random variables, the space of scalar or vector stochastic processes and so on, may be considered to be some particular cases of the K-space. Since the only condition a well-defined "value" should satisfy is its possibility to be added, multiplied by real numbers, and, at least in particular cases, to be compared, it becomes possible to give the information value idea a considerably wide sense.

It becomes therefore possible to apply the theory considered to the multi-aims control systems. It is also possible to consider the information value v_ξ as a pair (r_ξ, c_ξ) containing a net gain r_ξ and a cost c_ξ corresponding to a message ξ, instead of considering their difference $r_\xi - c_\xi$ only, as it was done above. In this case the information system can be optimized in both absolute and relative gain senses. Let us remark that, as it has been recently shown by Tshernikov [11] the well-known linear programming technique can be applied in the partially ordered linear spaces.

(c) The functional structure and operational rules. Any message that occurs in the information system carries some "proper" information concerning the over-system as well as some "auxiliary" information concerning: (1) the statistical measure of the "proper" information contained by the message, (2) its value, (3) its actual address in the spatial structure, (4) the codes of operations that are to be performed, etc. These data are in some sense more important from the information system management point of view than the "proper" information itself. Thus, a message ξ shall be considered as a pair (i_ξ, m_ξ) consisting of some "proper" information (data) i_ξ, and some "meta-information" m_ξ, the latter being the basic information used in the information system control. Consequently, the general information f..owing a large-scale control system can be imagined as shown in Fig. 1. This makes it possible to analyse the information systems as self-govering ones, as it was

already mentioned above.

Let Z denote a set of all the possible meta-informations concerning the messages occurring in the system, and let Z^{*} stand for a set of all the possible finite series formed by different elements picked out of the set Z . The functional structure of an information system can be in general defined as a class Ψ of all the admissible functional mappings of Z^{*} into Z^{*} itself. In order to simplify the considerations it will be assumed that Z is a countable set,and hence the corresponding information system will also be called a countable one.Further considerations will be restricted to the countable information systems only.

The problem of planning the tasks in information systems and of their operational management has been given general consideration in several papers, e.g. in the paper [8]. However, the problem requires further investigations. The main results published deal with some simplified information systems,as for example transmission networks or central data processing systems. However, the theory of information systems is interested in the general methods of systems organization optimization,including information systems of the most general class. Operational rules, for example, should be taken such that an optimal decision concerning the sequence of data processing, the time intervals for every operation, the technical means (computers, memory units, etc.) will be, in general, chosen so as to maximize the total value of the information processed by the system during a long time-interval. It is almost impossible to get a strong-optimum solution of this problem, except for some particular cases[6] . On the other hand, if an algorithm of information processing and the spatial and functional structures of the system are fixed, several variants of the sets of operational rules can usually be proposed and their effectiveness can be compared. Such a comparison needs a suitable mathematical tool. The latter can be based on the mathematical theory of queues, suitable modified, as has been shown below.

3. The Markov Processes Describing Information Flow in the Systems

(a) Basic assumptions. In the theory of queues a crowd of "clients" arriving to a serving system is considered under the following preassumptions[3]:

(1) the probability of a number Δk of clients arriving in the time-interval $(t, t + \Delta t)$ depends on the amount Δt only, and does not depend on the time t itself (the arrivals stationnarity condition);

(2) the probability in (1) does not depend on the number k of clients just waiting in the system (the independence of arrivals condition);

(3) the probability $P_{\Delta t}\{\Delta k > 1\}$ of more than one client arriving in the time-interval $(t, t + \Delta t)$ satisfies the condition

$$\lim_{\Delta t \to 0} \frac{P_{\Delta t}\{\Delta k > 1\}}{\Delta t} = 0 . \qquad (5)$$

Under these preassumptions the flow of clients gets the form of Poisson. However, when information systems are considered it is desirable to make some further assumptions, since the phenomena are considerably more complicated, the messages flow, their informations interact and their values increase. The information measure (no matter how defined) and its value are components of a stochastic vector process with,in general, an uncountable set of states. In fact,it should be considered as a K-space. The realizations of the process suffer jumping changes at random time-instants corresponding either to new messages entering the information system or to some informational operation ending. The process of interest is thus a general kind of the Kolmogorov-Feller stochastic process.However, the analysis of the probabilistic properties of such a class of processes leads to the systems of integro-differential equations, a numerical solution of which usually cannot be obtained in a simple form[9]. This difficulty can be overcome if an approximate description of the information system is permis-

sible. This can be obtained if the set of states is quantized. However, the meaning of the states of such a process will be different to that used in the theory of queues.

Let us suppose, for example, that in a given information processing point (i.p.p.) of a system messages having the following meta-informational properties may occur: (1) messages needing a "great", a "mean" or a "little" number of logic and arithmetic operations to be performed, (2) messages which occupy a "great", a "mean" or a "little" number of cells if stored in the memory units, (3) messages of a "common" or a "considerable" value or importance for the over-system. This gives a total of 18 meta-informational classes of messages in the system considered. In order to describe a current state of this i.p.p. it suffices to say that a number k_1 of the first--class messages, a number k_2 of the second-class, and so on, are present at a given time-instant, and that a message belonging to class s is being processed actually. If the information system contains N i.p.p.-s, the total number M of meta--informational classes should be multiplied by N. Unfortunately, this is a very hard restriction put on the theoretical analysis of the highly composed systems.

The following additional assumptions will now be made.

(4) A discrete vector process is considered:

$$(\bar{K}(t), \bar{S}(t)) = (K_1(t), \ldots, K_m(t), \ldots, K_M(t), S_1(t), \ldots, \ldots, S_n(t), \ldots, S_N(t)), \qquad (6)$$

where:

- the component $K_m(t)$ is a scalar process taking values of a countable set $\{0, 1, 2, 3, \ldots\}$, which indicates the number of messages of the m-th meta-informational class;

- the component $S_n(t)$ is a scalar process taking values of a $(M + 1)$-element set $[0, M]$, which indicates to what meta-informational class belongs the message actually processed in the n-th i.p.p.

(5) The new messages coming from the over-system for processing belong to one of the M meta-informational classes.

The arrivals form a multidimensional Poisson process with

statistically independent components:

$$P_{\Delta t}(\Delta k_1, \ldots, \Delta k_M) = \prod_{m=1}^{M} \frac{(\beta_m \Delta t)^{\Delta k_m}}{\Delta k!} e^{-\beta_m \Delta t}, \quad \Delta t > 0, \quad (7)$$

where the coefficients $\beta_m > 0$ characterize the intensities of arrivals of the given class messages.

(6) The operational rules governing the information system are given by a deterministic vector-function $\bar{\Phi}$ with the components

$$s_n = \varphi_n(\bar{k}), \quad n \in [1, N], \quad (8)$$

where $\bar{s} = (s_1, \ldots, s_N)$ and $\bar{k} = (k_1, \ldots, k_M)$ are the realizations of the vector processes $\bar{S}(t)$ and $\bar{K}(t)$, respectively. The function-component φ_n describes what class message is going to be processed next in the n-th i.p.p. providing the so called priority rules in the system.

(7) The jumps of value of the processes $\bar{K}(t)$, $\bar{S}(t)$ are of two types:

- spontaneous, and
- stimulated.

The spontaneous changes occur in one of the components..... $k_{n'}(t)$, where

$$(n - 1)\frac{M}{N} + 1 \leqslant n' \leqslant n\frac{M}{N}, \quad n \in [1, N], \quad (9)$$

(these components will be gathered into a sub-vector $\bar{k}_{(n)}(t)$ connected with the n-th i.p.p.), if the message comes to the n-th i.p.p. from the over-system or if the n-th i.p.p. starts with another operation the former being ended. It will be supposed that the spontaneous changes occuring in different i.p.p.-s are statistically independent.

The stimulated change is an immediate consequence of a spontaneous change of a component functionally related with the given one. Let us suppose that component of the vector $\bar{k}_{(n)}(t)$ has been changed spontaneously because of the coming of a new message into the n-th i.p.p. Thus, a stimulated jump

of the $s_n(t)$ component will be induced according to the priority rules. On the other hand, if the component $s_n(t)$ is changed spontaneously as described above, the stimulated changes of a component of the vector $\bar{k}_{(n)}(t)$ as well as those of some component of the vectors $\bar{k}_{(\nu)}(t)$ will occur, ν being any index of a i.p.p. functionally dependent on the n-th one, that is satisfying the equation

$$\nu = \Psi(n) , \quad n \in [1, N] , \tag{10a}$$

where Ψ describes the functional structure of the system. It follows that to perform any operation in the system it is necessary to carry out some other operations. However, at this level of abstraction, the sequences of logically related informational operations performed in the system are considered but as statistically averaged.

The slight difference between the spontaneous and stimulated changes can be reflected by supposing a right-continuity of the spontaneously changing realizations, and a left-continuity of the realizations changing by stimulation, as illustrated in Fig. 2. However, in probabilistic analysis of the processes both types of jumps must be taken into account together, which enables us to consider the process on the left-closed right-opened time intervals only.

(8) Let $\{\nu\}_n^-$ denote a set of indices ν satisfying relation (10a) for a given n, and $\{\nu\}_n^+$ - a set of indices satisfying a reciprocal relation

$$n = \psi(\nu) , \quad \nu \in [1, N] . \tag{10b}$$

The corresponding sets of i.p.p.-s will be called the input and output areas of the n-th i.p.p., the latter not being included in its input and output areas.

Let $Q^{(-)}(\bar{k}_{(\nu)}, \bar{s}_{(\nu)} | \bar{k}'_{(\nu)}, \bar{s}'_{(\nu)}; \bar{k}'_{(n)}, \bar{s}'_{(n)})$ denote a conditional probability distribution (p.d.) of stimulated states in the i.p.p. belonging to $\{\nu\}_n^-$, supposing that the states proceding the jump were $\bar{k}'_{(\nu)}, \bar{s}'_{(\nu)}, \bar{k}'_{(n)}, \bar{s}'_{(n)}$.

Let $\mathcal{H}_n(\bar{k}_{(n)}, s_n | \bar{k}'_{(n)}, s_n, m)$ be a conditional p.d. of the n-th i.p.p. states after receiving a message of the m-th class from the over-system.

Let $\lambda_n(\bar{k}_{(n)}, s_n \bar{k}'_{(n)}, s'_{(n)})$ be a conditional p.d. of the n-th i.p.p. states after a spontaneous change caused by the ending of an operation. The following condition should be satisfied according to the priority rules governing the n-th i.p.p.:

$$\sum_{\{\bar{k}_{(n)}\}} \mathfrak{Ae}_n(\bar{k}_{(n)}, s_n|\bar{k}'_{(n)}, s'_n, m) = \mathfrak{Ae}_n(s_n|\bar{k}'_{(n)}, s'_n, m) \equiv 1, \tag{11a}$$

$$\sum_{\{\bar{k}_{(n)}\}} \lambda_n(\bar{k}_{(n)}, s_n|\bar{k}'_{(n)}, s'_{(n)}) = \lambda_n(s_n|\bar{k}'_{(n)}, s'_n) \equiv 1, \tag{11b}$$

for $s_n = \varphi_n(\bar{k}_{(n)})$ only, and $\equiv 0$ otherwise.

(9) In order to derive the basic equation, a conditional p.d. $q_n(\bar{k}_{(n)}, s_n|\bar{k}'_{(n)}, s'_n, \bar{k}'_{(\nu)}, \bar{s}'_{(\nu)})$ of the states $\bar{k}_{(n)}$, s_n immediately after a jump of any type, the states $\bar{k}'_{(n)}$, s'_n, $\bar{k}'_{(\nu)}$, $\bar{s}'_{(\nu)}$, $\nu \in \{\nu\}_n^+$ before it given, is necessary. Since four typical situations are possible: the spontaneous jump in the n-th i.p.p. or stimulated jump caused by a spontaneous one in the input area $\{\nu\}_n^+$ on the one hand, and the arrival of a new message from the over-system or the ending of some information processing inside the system on the other hand, the conditional p.d. shall consist of four parts:

$$\begin{aligned} q_n(\bar{k}_{(n)}, s_n|\bar{k}'_{(n)}, s'_n, \bar{k}'_{(\nu)}, \bar{s}'_\nu) \cdot \Delta t = \\ = \lambda_n(\bar{k}_{(n)}, s_n|\bar{k}'_{(n)}, s_n) \cdot \gamma_n(s'_n) \cdot \Delta t + \\ + \sum_{\{\mu\}_n} \mathfrak{Ae}_n(\bar{k}_{(n)}, s_n|\bar{k}'_{(n)}, s'_n, \mu) \cdot \beta_\mu \cdot \Delta t + \\ + \sum_{\nu \in \{\nu\}_n^+} Q^{(-)}(\bar{k}_{(n)}, s_n|\bar{k}'_{(n)}, s'_n; k'_{(\nu)}, s'_\nu) \times \\ \times \Big[\sum_{\{\mu\}_\nu} \beta_\mu + \gamma_\nu(s'_\nu)\Big] \cdot \Delta t + o(\Delta t), \end{aligned} \tag{12}$$

where: $\{\mu\}_\nu$ denotes a set of indices corresponding to the meta-informational classes of the messages belonging to the ν-th i.p.p.-s, and $\gamma_\nu(s'_\nu)$ is the intensity coefficient of operations taking place in the ν-th i.p.p.

(b) The basic equation of information system. Let $P(\bar{k}, \bar{s}; t)$

denote a p.d. of the instaneous states of the processes $\bar{K}(t)$, $\bar{S}(t)$. A set of differential equations describing the above mentioned p.d. as a basic characteristic of the information system behaviour, shall be derived. Let the distribution P at a time-instant $t + \Delta t$ be considered as a function of its value at a time instant t :

$$P(\bar{k}, \bar{s}; t + \Delta t) = P(\bar{k}, \bar{s}; t) \cdot \prod_{(n,m)} \left[1 - \beta_m \cdot \Delta t + O(\Delta t)\right] \times$$

$$\times \left[1 - \gamma_n(s_n') \cdot \Delta t + O(\Delta t)\right] + \sum_n \sum_{\{\nu\}_n^+} \sum_{\{\bar{k}', s'\}} P(\bar{k}', \bar{s}'; t) \times$$

$$\times \left[q_n(\bar{k}_{(n)}, s_n | \bar{k}'_{(n)}, s_n', \bar{k}'_{(\nu)}, s_\nu') \cdot \Delta t + O(\Delta t)\right] . \qquad (13)$$

If we substitute the expression (12) into (13), open the parantheses, take the therm $P(\bar{k}, \bar{s}; t)$ over to the left side, divide both sides of the equation by Δt and pass to the limit at $\Delta t \to 0$, we obtain the result:

$$\frac{d}{dt} P(\bar{k}, \bar{s}; t) = -P(\bar{k}, \bar{s}; t) \cdot \left[\sum_m \beta_m + \sum_n \gamma_n(s_n')\right] +$$

$$+ \sum_n \sum_{\{\nu\}_n} \sum_{\{\bar{k}', \bar{s}'\}} P(\bar{k}', \bar{s}'; t) \cdot q_n(\bar{k}_{(n)}, s_n | \bar{k}'_{(n)}, s_n', \bar{k}'_{(\nu)}, s_\nu'), \qquad (14)$$

which should be completed by the requirement

$$\sum_{\{\bar{k}, \bar{s}\}} P(\bar{k}, \bar{s}; t) \equiv 1 \quad \text{for every} \quad t \in [0, \infty) \qquad (14a)$$

4. Final Remarks

Despite its complicated form it can be remarked that the basic equation (14) is equivalent to the matrix equations of the general form

$$\frac{d}{dt} \tilde{P}(t) = \tilde{Q} \cdot \tilde{P}(t) , \qquad (15)$$

where $\tilde{P}(t)$ is a column-vector of the probabilities $P(\bar{k}, \bar{s}; t)$, $\bar{k}, \bar{s} \in \{\bar{k}, \bar{s}\}$ (in general, of a countable set of components),

and $\tilde{Q}$ - a corresponding infinite square-matrix of the constant coefficients depending on the properties of the information system, as is evident from the equation (14). A possibility of getting a numerical solution depends on the form of the particular matrix $\tilde{Q}$: which seems unreal if the system is very large and very compact, that is if the input and output areas of every i.p.p. cover the whole system. However, the depth of the functional relations is usually limited if systems with hierarchical structures are considered.

A formal solution of equation (15) is a matrix function

$$\tilde{P}(t) = \tilde{P}_o \, e^{\tilde{Q}(t - t_o)} = \tilde{P}_o \sum_{\rho=0}^{\infty} \frac{1}{\rho!} \tilde{Q}^{\rho}(t - t_o)^{\rho} , \qquad (16a)$$

where

$$\tilde{P}_o \equiv \tilde{P}(t)_{t=t_o} \qquad (16b)$$

is the initial state. The practical sense of this formal solution consists in that it is possible to get some approximate numerical solu' ons, if necessary. On the other hand, the general form (14) or (15) makes it possible to get some asymptotical general results. We are usually interested in estimating the system behaviour at infinity, $t \to \infty$. Putting $(d/dt)\tilde{P}(t) = \tilde{O}$ in (15), where $\tilde{O}$ is a null-vector, we obtain

$$\tilde{Q} \cdot \tilde{P} = \tilde{O} . \qquad (17)$$

Since we are interested in the non-trivial solutions of this set of linear equations only, the requirement

$$\mathrm{Det}\, \tilde{Q} = 0 \qquad (17a)$$

follows immediately. The latter can be regarded as a necessary condition of the information system's probabilistic asymptotic stability.

Another problem arises if a solution of the basic equation (14) is known, and it may be formulated as follows: what information concerning the continuous random variables I, V, describing the information measure and value, can be extracted from the solution given in a discrete form.

Let $\{I\} \times \{V\}$ denote an uncountable set of all possible

information measures and values: the set $\{I\}$ is a non-negative half real axis, and $\{V\}$ is a "non-negative" cone of a K-space. Let a partition of $\{I\} \times \{V\}$ into M measurable non-overlapping cells $W_1, W_2, \ldots, W_M$ be given in the form of a conditional p.d.f. $u(\bar{k}|i, v)$. Let use consider a random vector-variable:

$$(I, V)_{\bar{K}} = \sum_{\mathcal{K}_1=0}^{K_1} (I, V)_{\mathcal{K}_1} + \ldots + \sum_{\mathcal{K}_M=0}^{K_M} (I, V)_{\mathcal{K}_M} , \qquad (18)$$

where: $(I, V)_{\mathcal{K}_M}$ denotes the conditional vector variable (I,V) supposing its value belongs to the cell W_m, and $K_1, \ldots, K_M$ are discrete random variables described by the p.d. $P(\bar{k}, \bar{s})$.

The p.d. sought for can be obtained in the form of a density function by applying the Bayes formula

$$w(i, v|\bar{k}) = \frac{p(i, v) \cdot u(\bar{k}|i, v)}{P(\bar{k})} , \qquad (19)$$

where

$$P(\bar{k}) = \sum_{\{\bar{s}\}} P(k, s) , \qquad (19a)$$

$p(i, v)$ is a p.d.f. of the random variables I, V,which is given a priori on the basis of observations of the input messages coming from the over-system. The solution of the basic equation (14) given, the original continuous variables I, V can be reconstructed on the basis of their quantized representations.

References

1. Bongard M.M.: O ponjatii "poleznana informacja". Problemy kibernetiki. Wyp. 9. Moscow 1963.
2. Ford L.R., Fulkerson D.R.: Flows in networks.Princeton 1962
3. Gnedenko B.V., Kovalenko I.B.: Vvedenje v teorju massovogo obsluživanja. Moskov 1967.
4. Kantorovitsh L.V., Vulich B.Z., Pinsker A.G.: Funkcjonalnyj analiz v poluuporjadotshennych prostranstvach. Moscov 1950.
5. Kharkevitsh A.A.: O cennosti informacji. Problemy kibernetiki. Wyp. 4. Moscov 1960.
6. Kulikowski J.L.: On the optimum organization of calculations in a decision system. Kybernetika, Prague, 1965 No. 6.
7. Kulikowski J.L.: Podstawowe zagadnienia organizacji systemów przetwarzania informacji. Problemy cybernetyki techni-

cznej. Warszawa 1967. In polish.
8. Kulikowski J.L.:Nekotoryje voprosy upravlenja informacjonnymi sistemami. Teorja upravlenja v bolshich sistemach.Sophia 1968.
9. Kulikowski J.L.: Markowskie procesy na strukturach i niektóre ich zastosowania. Archiwum Automatyki i Telemechaniki 1969. In polish.
10. Lerner A.J.: Zadania teorii sterowania wielkimi systemami. Problemy sterowania wielkimi systemami. Wrocław 1964. In Polish.

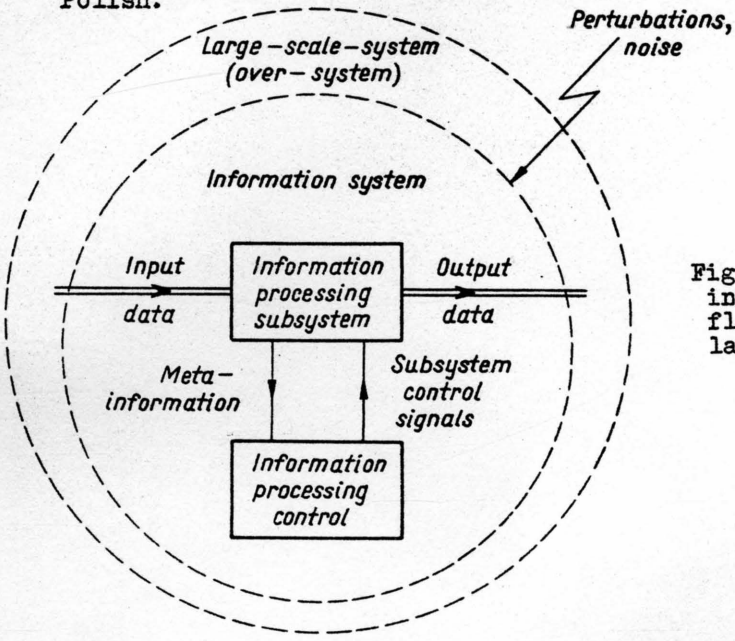

Fig. 1. Main information flows in a large-scale control system

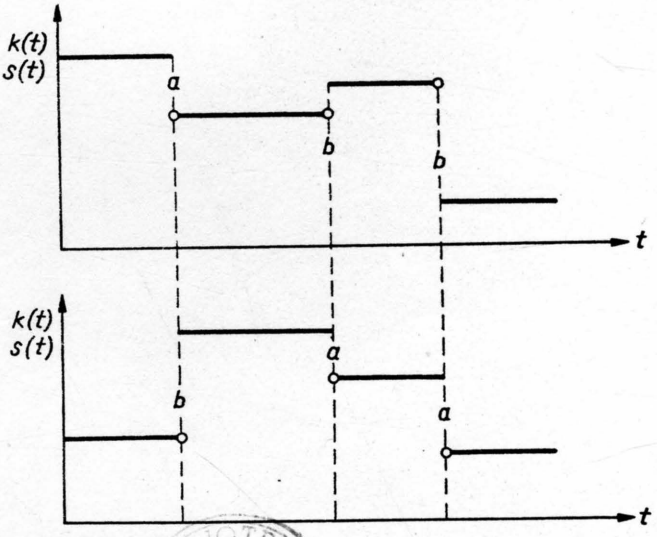

Fig. 2. Several examples of the spontaneous a and stimulated b jumps of instaneous value of the process describing behaviour of an information system