ON THE IDENTIFICATION OF DIFFUSION COEFFICIENTS*

GISBERT STOYAN

Akademie der Wissenschaften der DDR
Zentralinstitut für Mathematik und Mechanik, Berlin

1. The identification problem

1.1. Let us consider the following problem: Some chemical components are diffusing from outside into a plate of catalytic material, several reactions take place and the reaction products leave the plate by diffusion. Let there be \( n \) reactions. Then for \( n \) independent components with concentrations \( c_i \), \( i = 1, \ldots, n \), we have as a mathematical model of the whole process the system of equations:

\[
\begin{align*}
(1) \quad \frac{d}{dr} \left( d_i(r) \frac{dc_i}{dr} \right) &= q_i(r, c_1, \ldots, c_n), \quad -1 \leq r \leq 1, \quad i = 1, \ldots, n, \\
(2) \quad c_i(-1) &= c_i(1) = c_i(0).
\end{align*}
\]

In (1) \( d_i \) are the diffusion coefficients, \( q_i \) the reaction rates. (2) means that we are considering a process in which the outside concentrations \( c_i \) are equal on both sides of the catalytic plate (this is of no importance, of course; there may be different outer concentrations as it is the case for membranes). The process may be temperature dependent. In that case one of the \( c_i \), say \( c_i \), gives the value of the temperature and then \( d_i \) represents the heat conduction coefficient.

Let us now suppose that the coefficients \( d_i(r) \), \( i = 1, \ldots, n \), are unknown, but we are given a set of data, consisting, for instance, in the experimentally measured values \( \tilde{g} \) of

\[
\begin{align*}
(3) \quad g_{1i}(d) &= c_i(r_0), \quad g_{2i}(d) = d_i \frac{dc_i}{dr}(1), \quad i = 1, \ldots, n, \\
g_{3}(d) &= \sum_{i=1}^{n} \gamma_i \int_{-1}^{1} c_i(r) dr,
\end{align*}
\]

where \( d = (d_1, \ldots, d_n) \).

---

* Revised version of two lectures, held at the Banach Centre in May 1975.
\( g_1 \) gives the concentration (or temperature) measured at the place \( r_k \) of the plate, \( g_2 \) is the concentration flow across the boundary \( r = 1 \), \( g_3 \) is connected with the difference of the weight of the plate after and before reaction (\( y_1 \) is the specific weight of the \( i \)th component).

The question now is, under what conditions on the values \( \bar{y} \) and the rates \( q_i \), it is possible to determine (to identify) the diffusion coefficients \( d_i(r) \).

1.2. We now introduce the function space

\[
(\mathcal{W}^1)^n = \bigoplus_{i=1}^n \mathcal{W}^1[-1, 1]
\]

consisting of vectors \( y = (y_1, ..., y_n) \) of functions \( y_i \) belonging to the usual Sobolev space \( \mathcal{W}^1[-1, 1] \).

In \((\mathcal{W}^1)^n\) we define the scalar product

\[
[u, v] = \sum_{i=1}^n \int_{-1}^1 \left( \frac{d u_i}{d r} \cdot \frac{d v_i}{d r} \right) dr,
\]

and the corresponding norm \( \| \cdot \| \).

We shall pass from (1), (2) to an operator equation in \((\mathcal{W}^1)^n\) and look for conditions under which the latter has a solution in \((\mathcal{W}^1)^n\). To this end we restrict considerations to functions \( d_i(r) \) which belong to \( L^\infty_c \):

\[ d_i \in L^\infty_c = \{ d(r), r \in L^\infty([-1, 1]), \exists z_0 > 0, z \equiv z_0 \text{ a.e. in } [-1, 1] \}. \]

Furthermore, we consider only rate functions \( q_i \) satisfying the following three conditions:

(a) each \( q_i \) is continuous in \( r \in [-1, 1], |c| < \infty \), \( i = 1, ..., n \),

(b) \[ \sum_{i=1}^n \int_{-1}^1 q_i(r, c_1, ..., c_n) \frac{d r}{d r} > q_i > 0, \]

(c) \[ \sum_{i=1}^n (q_i(r, c_1, ..., c_n) - q_i(r, c_1, ..., c_n)) (c_1 - c_1) \geq 0 \]

for all \( r \in [-1, 1] \) (monotonicity condition).

In (1) we now transform \( c_i \) into \( x_i \),

\[ x_i = c_i - c_{i-1}, \quad i = 1, ..., n, \]

possessing homogeneous boundary values. Multiplying the resulting equations by \( v \in (\mathcal{W}^1)^n \) and integrating by parts we get

\[
\sum_{i=1}^n \int_{-1}^1 \left( d_i(r) \frac{d x_i}{d r} \cdot \frac{d v_i}{d r} \right) dr + \int_{-1}^1 p_i(r, x_1, ..., x_n) v_i dr = 0,
\]

where

\[ p_i(r, x_1, ..., x_n) = q_i(r, x_1 + c_{i-1}, ..., x_n + c_{i-1}). \]

By the Riesz theorem we obtain

\[
\sum_{i=1}^n \int_{-1}^1 d_i(r) \frac{d x_i}{d r} \cdot \frac{d v_i}{d r} dr = [Dx, v],
\]

where \( D \) is a linear self-adjoint operator which is, by (4), positive-definite

\[ D: (\mathcal{W}^1)^n \to (\mathcal{W}^1)^n, \quad D = D^* \geq z_0 \cdot I, \]

where \( z_0 = z_0(d) = \text{ess min} \ d_i(r) > 0 \).

In the same manner, using (5a), we have

\[
\sum_{i=1}^n \int_{-1}^1 p_i(r, x_1, ..., x_n) v_i dr = [P(x), v],
\]

where \( P(x) \) is a (completely) continuous operator, nonnegative in view of (5c):

\[ [P(x) - P(y), x - y] \cdot
\]

\[
= \sum_{i=1}^n \int_{-1}^1 \left( q_i(r, x_1 + c_{i-1}, ..., x_n + c_{i-1}) - q_i(r, y_1 + c_{i-1}, ..., y_n + c_{i-1}) \right) \times
\]

\[ \times \left( (x_1 + c_{i-1}) - (y_1 + c_{i-1}) \right) dr \geq 0.
\]

Thus (7) is equivalent to the operator equation

\[ Q(d, x) = Dx + P(x) = 0. \]

Now, \( Q \) being continuous and monotone,

\[ [Q(d, x) - Q(d, y), x - y] \geq z_0(d) \| x - y \|^2, \]

equation (9) has a unique solution \( x = x(d) \) in \((\mathcal{W}^1)^n\).

(For the assertions stated in this section see [1], [2] and [3]).

We show now that this solution is locally lipschitzian in \( d \): Following [4] we write

\[ [(D_i - D) x(d, i), x(d) - x(d, i)] = [Q(d, x(d, i)) - Q(d, x(d))], x(d) - x(d, i)]
\]

\[ \geq z_0(d) \| x(d) - x(d, i) \| \| x - x(d) \| \| x(d, i) \|.
\]

but

\[ [(D_i - D) x(d, i), x(d) - x(d, i)]
\]

\[ = \sum_{i=1}^n \int_{-1}^1 \left( d_i(r) d_i(r) - d_i(r, i) \right) \left( \frac{d x_i}{d r} - \frac{d x_i}{d r} \right) dr
\]

\[ \leq \| d_i - d_i \| \| x(d, i) \| \| x(d) - x(d, i) \|.
\]

where

\[ \| d_i \| = \text{ess max} \ d_i(r), \quad d_i \in L^\infty_c. \]
Hence
\[ \|x(d) - x(d_1)\| \leq \frac{\|x(d)\|}{L_x} |d - d_1|, \]
(11)
where \(L_x\) is the Lipschitz constant of \(x\). To get an estimate for \(\|x(d)\|\), we set in (10) \(x = x(d_1), d = d_1, \) and \(y = 0\):
\[ z_0(d_1) |x(d_1)|^2 \leq \left[ -Q(d_1, 0), x(d_1) \right] = -[F(0), x(d_1)] \]
\[ = -\sum_{i=1}^{5} \int_{-1}^{1} p_i(r, 0, \ldots, 0) x(d_1) \, dr \]
\[ \leq \sum_{i=1}^{5} \left( \int_{-1}^{1} q_i(r, c_i, \ldots, c_i) \, dr \right)^{1/2} \]
\[ \leq q_0 \|x(d_1)\|. \]
Here \(q_0\) is the number occurring in (5b). Using the well-known inequality
\[ |e_i|_{(s,r)} = \left( \sum_{i} \left( \int_{-1}^{1} |e_i|^2 \, dr \right)^{1/2} \right)^{1/2} \leq \frac{2}{\pi} \|e_i\|, \]
we obtain
\[ z_0(d_1) |x(d_1)|^2 \leq \frac{2}{\pi} q_0 \|x(d_1)\|, \]
i.e.
\[ \|x(d_1)\| \leq \frac{2q_0}{\pi z_0(d_1)}. \]
(Here the sense of (3b) becomes clear: This condition excludes the trivial solution \(x = 0\), i.e., \(c_i = c_i, V_{i1}\). Thus, by (11),
\[ \|x(d) - x(d_1)\| \leq L_x |d - d_1|, \]
(12)
where \(L_x = 2z_0 / \pi z_0(d_1)\).
We remark that, by the Sobolev imbedding theorem, we have \(x \in C^0([a, b], \mathbb{R}^{N_n})\), \(s \leq \frac{1}{2}, \) and that from
\[ |e_i(r)\| \leq \left( \frac{1 - r^2}{2} \right)^{1/2} |e_i|, \]
\(r \in (\tilde{W}^{1, 2}), \)
follows also the local Lipschitz continuity of \(x_i(d, r)\) (and likewise of \(c_i(d, r)\)) for all \(i\) and \(r\).

Theorem 1. Let \(d_i(r), i = 1, \ldots, n, \) be positive continuous functions on \([-1, 1]\. Then, if (5) is satisfied, the boundary value problem (1), (2) has the unique (generalized) solution \(c_i, c_i = x_i + \xi_i, i = 1, \ldots, n, x \in (\tilde{W}^{1, 2})\), the dependence of this solution on \(d\) is locally lipchitzian:
\[ |c_i(d_1, r) - c_i(d_2, r)| \leq \frac{2q_0}{\pi z_0(d_1) z_0(d_2)} \left( 1 + r^2 \right)^{1/2} \text{ess}\sup_{V \ni V_r} |d_i(r) - d_{11}(r)|, \]
where \(z_0(d) = \text{ess}\sup_{V \ni V_r} d_i(r), k = 1, 2. \)

1.3. Now take \(d\) from a compact subset of \((L_\infty)^n\), assuming, for instance, that the components of the elements of this subset are representable by finite linear combinations of continuous coordinate functions with bounded coefficients. Define
\[ f(d) = \|c(d) - \tilde{c}\|^2 \]
where \(c\) denotes the (generalized) solution of (1), (2) corresponding to a given \(d\) and \(\tilde{c}\) is a continuously differentiable vector-function constructed from the given values \(\tilde{c}\) of the functions (3).
By the generalized theorem of Weierstrass (see [5] and [6]) the continuous function \(f(d)\) will attain a minimum, say \(d^0\), on the chosen compact subset. This minimum \(d^0\) can be considered as a solution of the identification problem (1)–(3) for the given data \(\tilde{c}\). \(d^0\) is a quasi-solution in the sense of Ivanov [7], [8].

Remark. Indeed, \(f(d)\) attains its minimum also on closed bounded balls in \((L_\infty)^n\) [6]

By better using the properties of the operator \(Q(d, x)\), it can be shown that \(f(d)\) is weakly* lower semicontinuous on \((L_\infty)^n = (L_\infty)^n\), in the space (the adjoint of the Banach space \((L_\infty)^n\)) bounded closed balls are also closed in the weak* topology and therefore, again by the generalized Weierstrass theorem, \(f(d)\) takes on its minimum on such sets.

1.4. As an illustration we turn to a rather special case: There is only one reaction \((n = 1)\), the reaction rate depends on the chosen independent component only and the diffusion coefficient is assumed to be constant.
We shall see that even in this case the identification problem (1)–(3) may have no unique solution.

Norming the concentration to the outside concentration we consider the boundary value problem
\[ d \frac{d^2 c}{dr^2} = \phi(c), \quad c(-1) = c(1) = 1. \]
Let \(g(d)\) be defined by
\[ g_1(d) = c(0), \quad g_2(d) = \frac{d c}{dr}(1), \quad g_3(d) = \frac{1}{2} \phi(c) dr. \]
The problem now is to find the positive number \(d\) from
\[ g(d) = \tilde{g}_1 \]
for given values \(\tilde{g}_1\).
We assume \(\phi(c)\) to be defined for \(c \geq 0\) and to satisfy:
\begin{enumerate}
\item \(\phi\) is continuous in \(c \geq 0\),
\item \(\phi(0) = 0, \phi(c) > 0, c > 0,\)
\item \(\phi\) is nondecreasing in \(c \geq 0\).
\end{enumerate}
First we replace \( q \) by \( \tilde{q} \).

\[
\tilde{q}(c) = \begin{cases} 
q(c), & c > 0, \\
-q(c), & c < 0.
\end{cases}
\]

To the boundary value problem

\[
dc^2 \over dt^2 = \tilde{q}(c), \quad c(0) = 0, \quad c(1) = 1
\]

we now can apply Theorem 1 establishing the existence of a unique generalized solution to (17) which, in this case, is the classical two times differentiable solution. Of course, for deriving the existence and uniqueness of a classical solution to (17) under conditions (15) it is not necessary to apply the monotone operator theory but rather refer to [9] or [10].

For systems (1) with differentiable \( d(c) \), the results of [11] and the references cited there could be useful. From the maximum principle and from (15), (16) it follows that this solution is nonnegative. Hence, it is a nonnegative solution of (13), and (13) has only this nonnegative solution.

We remark that by symmetry (13) is equivalent to

\[
dc^2 \over dt^2 = q(c), \quad dc \over dt (0) = 0, \quad c(1) = 1.
\]

Integrating once we get (\( d > 0 \) by assumption)

\[
dc \over dr = c^1 \over r^1 \int_0^r q(c(\tau))d\tau,
\]

i.e., by (15),

\[
dc \over dr (r) \geq 0 \text{ for } r \geq 0, \quad dc \over dt (r) \leq 0 \text{ for } r \leq 0.
\]

Thus

\[
0 \leq c(0) \leq c(1) \leq 1, \quad r \in [-1, 1].
\]

Note that (15) does not exclude the case \( c(0) = 0 \), but this may occur only if \( q \) belongs to the set of continuous functions for which the Cauchy problem

\[
dc^2 \over dt^2 = q(c), \quad c(0) = 0, \quad dc \over dt (0) = 0
\]

is not uniquely solvable (see [12] and [13]), e.g. if

\[
q(c) = ac^2, \quad 0 < b < 1.
\]

If, on the other hand, this Cauchy problem is uniquely solvable (see [14]; for instance, if \( q \) is lipschitzian in \( c \)) then, of course, necessarily \( c(0) > 0 \); for \( c(0) = 0 \) would then imply \( c \equiv 0 \) and hence \( c(1) = 1 \) would not be fulfilled.

But if \( c(0) > 0 \), then, repeating the above argument, we get

\[
dc \over dr (r) > 0, \quad r > 0, \quad 0 < c(0) < c(r) < 1, \quad 0 < r < 1.
\]

Further use of explicit representations similar to (19) leads to the following result [15]:

**Theorem 2.** Consider the identification problems

\[
P_i = \{(13), (14), (14i)\}, \quad i = 1, 2, 3.
\]

Assume (15) to be fulfilled, and, for \( i \neq 3 \), also (15):

\[
(15) \quad q(c) \text{ is strictly increasing on } (c_3, 1) \text{ for some } c_3 \in (0, 1).
\]

Then \( P_i \) has a solution \( d \in (0, \infty) \) only if

\[
\tilde{r}_i \in (0, 0) \quad \text{where} \quad s_1 = 1, \quad s_2 = q(1), \quad s_2 = 2.
\]

This solution is unique.

If (15) is not satisfied, \( P_i (i = 1, 2) \) has a solution if \( \tilde{r}_i \in (0, 1) \) and \( \tilde{r}_3 \in (0, q(3)) \), but for \( \tilde{r}_i = 0 \) and \( \tilde{r}_3 = q(3) \) the solution is the set of \( (0, d^*) \) and \( (d^*, \infty) \), respectively, where \( d^* \) and \( d^* \) are some positive numbers.

**Remark.** If

\[
a_1 c^b < q(c) < a_2 c^b \quad \text{with} \quad 0 < a_1, \quad 0 < b < 1,
\]

then one can prove

\[
d_1 < d' < d_2, \quad d_j = \frac{a_j (1-b_j)^2}{2(1+b_j)}, \quad j = 1, 2;
\]

if \( q(c) = q(c_i) = q(1) \) for \( 1 \geq c \geq c_i \) and \( q(c) < q(c_i) \) for \( c < c_i \), then

\[
d^* = \frac{q(1)}{2(1-c_i)}.
\]

Theorem 2 remains true if \( q(c) \) satisfies (15b), is lipschitzian in \( [0, 1] \) for some \( \epsilon \in (0, 1) \), piecewise lipschitzian in \( [-1, 1] \) possessing there a finite number of jumps, and strictly monotone. This can be shown by using (21).

2. The practical solution of the identification problem

2.1. Firstly, a stable method of good approximation is needed for solving the direct problem (1), (2) — to calculate \( c(d, r) \) for given \( d(r) \).

(a) Discretizing the two-point boundary value problem (see [16]) we arrive at a system of algebraic equations. The latter may be solved by the method of conjugated gradients (see [17] and the references there) or Newton-like methods [18]. Here it is advantageous to use gradient-free algorithms as [19], [20]. In all cases, difficulties caused by strong nonlinearities may be attacked with use of continuation processes [18], [35]. A natural way of introducing a continuation parameter \( r \) is to replace \( q_i \) in (1) by \( q_i r \), \( r \in [0, 1] \). Another difficulty for the numerical solution is the usual stiffness of equations (1), the occurrence of boundary layers. This can necessitate the use of nonequally spaced grids for the discretization (21).
An alternative way of solving (1), (2) is the use of direct shooting methods [22] (see also [23]). This consists in solving a series of initial-value problems for (1), where the starting values at $\pm 1$ are adjusted so as to satisfy the conditions at $\pm 1$. That is, as above a system of nonlinear algebraic equations has to be solved but one of much less unknowns. Moreover, for the solution of the initial value problems a number of good methods is available which are developed for stiff systems [24].

First approximations to the starting values needed at $\pm 1$ may be found from the measurement data $\tilde{g}$ or from previous calculations with a nearby $d(r)$. In difficult cases again a continuation process may help where the continuation parameter continuously changes the length of the integration interval from 0 to 2 and, simultaneously, the right-hand sides from 0 to $g$. Storage requirements are much less for the shooting method: Only vectors of dimension 2n must be handled. For the over-all discretization considered in (a) it would be economic to hand over the set of all c-values as first approximations to the next calculations corresponding to a perturbed $d(r)$. But this will turn out to be too expensive with respect to storage space, for instance, if a series of experiments was carried out for sets of different outer concentrations and temperatures (2), and the measurement data for only one set (2) are too poor for the identification of $d(r)$.

Therefore, shooting methods look quite advantageous, but not every problem can be solved by them. (This occurs if the initial value problems for (1) are unstable or nonuniquely solvable, which is not excluded by (5) and indeed happens for some source terms used in practice.)

(c) In the later stages of the search for $d(r)$ (described below), where $c(d, r)$ changes only slowly, a serious competitor to the above mentioned methods of solution of (1), (2) is a simple Gauss-Seidel type iteration with relaxation parameter. The latter may be chosen by some additional calculations minimizing a cost functional including the number of iterations carried out and the value of defect reached then.

To end this point we mention that stiffness problems can be softened by elimination of the temperature equation (if present), because diffusion coefficients and heat conduction coefficient differ usually by some magnitudes. Such an elimination is always possible owing to a special structure of (1) in chemical problems.

2.2. The minimization of the functional (12) is the next task once the solution of the boundary value problem (1), (2) has been settled.

(a) The first question arising is whether (12) is a useful form of functional or if another one should be used instead. If many experimental data $\tilde{g}$ are given for one set of outer concentrations (2) such that a good smoothing fit (25)-(27) $\tilde{c}$ can be computed from $\tilde{g}$, then one can take (12) (or rather a discretized version of it, using a quadrature formula). This has the advantage that, using fairly smooth functions, one tries to approximate another smooth function by varying $d(r)$. On the other hand, this means that one tries to come close to a function $\tilde{c}$ which, possibly, is not a solution of any differential equation (1). Therefore it seems better to carry out the process of densifying of the available information not up to the end, but to perform only the first natural step: to calculate mean values in the case where several experimental results are given for the same quantity. As the functional for the identification of $d(r)$ we then take

\[
(f/d(r) = \sum_{k=1}^{n} \left( \tilde{g}_{k} - c(d, r) \right)^{2} \omega_{k} + \sum_{k=1}^{n} \left( \tilde{g}_{k} - d_{k} \frac{d_{k}}{x} \right)^{2} \omega_{k} + \sum_{k=1}^{n} \gamma \omega_{c(d, r), \omega} \left( \tilde{c} - \sum_{k=1}^{n} \gamma \omega_{c(d, r), \omega} \right)^{2}.
\]

Here $\omega_{k}$'s are weights proportional to the measurement accuracy of the corresponding measurement value (and, if mean values are used, the number of experimental values given for that quantity should be taken into account, too).

If (12) is used, then fitting is done first (not looking at the mathematical model (1) of the process) and identification next. If (22) is taken, then fitting is carried out through solutions of (1) and identification is done at the same time.

In any case, to get fuller information about the possibilities of identifying some quantity one has to formulate several minimum problems, to interpret and compare the results.

2.3. The next question is, in what form to take $d(r)$. From experience, and theory as well, it is known that one should not minimize the chosen functional on the domain of all possible $d(r)$ but on a compact subset. The smaller this subset, the better. Therefore a good approach is to search for a $d(r)$ possessing some parametric form (and not to try to find some general function). If additional knowledge does not suggest any such form, take cubic splines. As the first step in finding $d(r)$, determine only one parameter for every $d(r)$. Starting from this it is easier to determine more parameters. Continuing in this way, the number of parameters can be raised step by step. An important question (actually, a very difficult one of stochastics) is where to stop this kind of projection-iteration method (1). Of course, one should not try to determine more parameters than the number, say $m$, of measurement data occurring in (22) (i.e. the number of squares in (22)). Naturally, a reason to stop is the occurrence of nonuniqueness — if this can be detected during computing. From experience in fitting of equidistant experimental data by spline functions there was gained the rule of thumb [27] to take about four points per spline. Carried over to our fitting (and identification) problem, this would mean that the upper number of parameters to be determined could be about a fourth of the number $m$ of measurement data.

It may happen very well that already the determination of only one parameter for each $d(r)$ is not possible because of nonuniqueness. In such a case only some...
further information helps. Either more experimental results must be delivered or there must be a reason to select a special solution between the possible ones, e.g. that with a minimal distance from a given point. Then Tychonov regularization [28] is helpful. This method can be applied to determine any given number of parameters, but rather should not be applied without an interpretation for why and how it is done.

2.4. After the parametrization of $d(t)$, we arrive at a well-known situation: the sum of $m$ nonlinear squares has to be minimized by varying $p$ parameters, $p \leq m$.

Hence, one of the newer developments of the Marquardt algorithm [31], [32] will ensure an efficient solution of the problem. Derivatives needed should be calculated approximately from finite differences. Usually a projection has to be carried out to meet conditions as $0 < \delta_{m+k} \leq \delta_{k} \leq \delta_{m+k}$ and, possibly, smoothness conditions. The choice of the Marquardt step-length and parameter for linear side conditions is discussed in [33].

An alternative way is the minimization of (12) using a gradient procedure. Once the numerical solution of the equation adjoint to (1) and (12) is managed, the gradient of (12) can be calculated [34] (and this can be done flexibly for several functionals). But the gradient method becomes less efficient in the vicinity of the minimum and in the presence (usual for such identification problems) of deep valleys in the relief of the functional.

References


Presented to the Seminar Mathematical Models and Numerical Methods (February 3-June 14, 1975)