Interpretations and estimates for ill-posed problems

Thomas I. Seidman
Department of Mathematics and Statistics
University of Maryland Baltimore County
Baltimore, MD 21228, USA
e-mail: ⟨seidman@math.umbc.edu⟩

ABSTRACT: Following a discussion of the relation of these problems to applications, intended to clarify the considerations which must be handled in order to obtain genuinely useful results, we consider techniques for determining optimal approximationss and consequent optimal error bounds for certain classes of ill-posed problems with appropriate *a priori* information.

KEY WORDS: *ill-posed, filtering, error estimates, partial differential equations.*
AMS classifications: 65J20, 35R25

1 Introduction

The canonical ill-posed problem requires us to 'evaluate' an unevaluable map **S**: we must 'solve', in some sense, an equation

$$\mathbf{F}(\bar{x}) = \bar{y} \tag{1.1}$$

in a context in which we have a continuous *direct map* **F**

$$\begin{array}{ccc} y & \xleftarrow{\quad\mathbf{F}\quad} & x \\ \mathcal{D} & \xrightarrow{\quad\mathbf{S}\quad} & \mathcal{P} \end{array} \tag{1.2}$$

from a *parameter space* $\mathcal{P}$ to a *data space* $\mathcal{D}$ for which the inverse $\mathbf{S} = \mathbf{F}^{-1}$ is not continuous. Our nominal 'problem' (1.1) views $\bar{x}$ as an unknown 'parameter' which we are to determine (estimate) from available data. Necessarily, since the data is to be obtained by 'measurement', there is some uncertainty, one is not given $\bar{y}$ in (1.1) but knows only that $\bar{y} \in \mathcal{Y} \subset \mathcal{D}$. The corresponding information for the solution will then be expressed by

$$\bar{x} \in \mathcal{U} = \mathcal{U}(\mathcal{Y}) := \{x \in \mathcal{P} : \mathbf{F}(x) \in \mathcal{Y}\} \tag{1.3}$$

so we may refer to $\mathcal{U}(\mathcal{Y}) = \mathbf{F}^{-1}(\mathcal{Y})$ as the 'region of uncertainty'; we emphasize that, with the information given, every $x \in \mathcal{U}(\mathcal{Y})$ is an equally valid candidate to be $\bar{x}$.

For well-posed problems, the usual practice is to represent $\mathcal{Y} = \mathcal{Y}_k$ — suggesting the possibility of considering a potential sequence of increasingly accurate measurements — by some nominal value y_k and then to take $x_k := \mathbf{S}(y_k)$ as a representative of $\mathcal{U}_k = \mathbf{S}(\mathcal{Y}_k)$. The measurement accuracy is typically indicated by $\delta = \delta_k$ so our available information is that

$$\bar{y} \in \mathcal{Y}_k \subset \{y \in \mathcal{D} : \|y - y_k\| \leq \delta_k\} \tag{1.4}$$

for some suitable choice of norm on $\mathcal{D}$; for simplicity one generally takes $\subset$ to be a defining equality. The uncertainty in the result is then correspondingly characterized by

$$\bar{x} \in \mathcal{U}_k \subset \{x \in \mathcal{P} : \|x - x_k\| \leq \varepsilon_k = \varepsilon(\delta_k)\} \tag{1.5}$$

with well-posedness defined by the property that

$$\varepsilon_k \to 0 \quad \text{as } \delta_k \to 0; \tag{1.6}$$

indeed, for well-posed problems one typically has a linear rate of convergence

$$\varepsilon(\delta) \leq K\delta \tag{1.7}$$

for some (local) Lipschitz constant $K = K(\bar{x})$.

For ill-posed problems one cannot use $x_k := \mathbf{F}^{-1}(y_k)$ as a plausible approximation for $\bar{x}$ as above since (a) this may not be defined as y_k need not be in the range of $\mathbf{F}$, (b) there are severe difficulties, which must be addressed, in constructing any computational approximation $\hat{\mathbf{S}}$ to $\mathbf{F}^{-1}$, and (c) (arbitrarily) small perturbations of y can give (arbitrarily) large perturbations of x so (1.5) typically requires $\varepsilon_k \equiv \infty$ even for arbitrarily small δ_k, not only precluding (1.6) but rendering (1.5) entirely uninformative. As a substitute for this, we seek some other notion of 'solution' which will be feasible in the context.

The intention here is to follow [9] to some extent in discussing the questions, "What is the problem? What is a solution?" — with some prejudice towards my own work in examples. The thrust of [9] was to note that

- despite the heuristic arguments of [4] in bestowing upon our concerns the pejorative title *ill-posed* and the technical arguments of [13] which demonstrate them *infinitely complex*, we nevertheless find ourselves successful in contributing usefully to applications and

- this can be understood (only) in terms of a consideration of ill-posed problems in an applied context rather than in isolation.

It is this last which involves the questions above: our notion of *success* depends on an interpretation of *'solution'* which, in turn, depends on a contextual interpretation of the *problem*. In general, the emphasis of this paper will be at least as much on pointing out 'things to worry about' as on specific algorithms or estimates. We do, however, use the perspective developed in the next section to introduce, in the following section, the paradigms, constructions, and estimates of the *'Method of Optimal Filtering'*.

2 Paradigms of interpretation

The first key question for interpretation for (1.1) is: *What do we mean by 'small'/'large' perturbations?* — i.e., What determines the choice of spaces $\mathcal{P}, \mathcal{D}$ and the topologies imposed on them for discussing the invertibility of $\mathbf{F}$?

Our first observation is that these choices, necessarily part of the specification of the problem, represent an approximation to a question of economics and technology: the ultimate goal of our analysis is the minimization (over a range of technologically feasible possibilities) of

$$\begin{aligned} \text{total cost} &= \text{cost of measurements (to some accuracy)} \\ &\quad + \text{cost of error/uncertainty in the estimate} \\ &\quad [+ \text{cost of computation, etc.}] \\ \mathcal{J} &= \varphi(\mathcal{Y}) + \psi(\mathcal{U})\ [+\gamma(\hat{\mathbf{S}})]. \end{aligned} \tag{2.1}$$

[In this last formula, $\hat{\mathbf{S}}$ denotes some computational algorithm for approximating $\mathbf{F}^{-1}$ and this cost may or may not be included.] Before any further analysis of (2.1), we expand the diagram (1.2) as

$$\begin{tikzcd} \mathcal{D} \arrow[d, "\mathbf{M}_k"] & \mathcal{P} \arrow[l, "\mathbf{F}"'] \\ \mathcal{D}_k & \mathcal{P}_k \arrow[l, "\hat{\mathbf{S}}_k"'] \arrow[u, "\mathbf{I}_k"'] \end{tikzcd} \tag{2.2}$$

to emphasize the facts that in the real world any actual implementation $\mathbf{M}_k$ of measurement *must* lead to a finite set of numbers (with finite precision) so $\mathcal{D}_k = \mathbb{R}^{n(k)}$ and, correspondingly, that computation must also be finitary so we must work with a finite dimensional approximating representation in $\mathcal{P}_k = \mathbb{R}^{m(k)}$ (e.g., using splines, etc., related to $\mathcal{P}$ by $\mathbf{I}_k$). As noted earlier, the index k suggests that we will later embed this schema in a sequence of such diagrams, e.g., with increasingly accurate measurements and increasingly refined computations as $m(k), n(k) \to \infty$.

Our second observation is that ill-posed problems of practical significance come with an applied context so our efforts are teleologically directed: we will assume now that there is some specified application in which a 'result' z depends on the unknown parameter $\bar{x}$ through a map

$$\mathcal{P} \xrightarrow{\quad \mathbf{A} \quad} \mathcal{R} : x \mapsto z \tag{2.3}$$

so that 'what we really want' is the result $\bar{z} = \mathbf{A}(\bar{x})$. From this point of view, **the solution of the ill-posed problem is only an intermediate construct intervening between the available data and the intended application.** In isolation, this problem (2.3) would be a 'standard' well-posed problem and we would eventally also introduce computational approximations $\mathbf{A}_k$ to $\mathbf{A}$.

Of course, we have an obvious difficulty — the 'value' of $\bar{x} \in \mathcal{P}$ is not 'given' *a priori* but is to be obtained from the data via the ill-posed problem (1.1). We concatenate (2.3) with (2.2) to obtain

$$\begin{tikzcd}
& \mathcal{D} \arrow[d, "\mathbf{M}_k"] & \mathcal{P} \arrow[l, "\mathbf{F}"'] \arrow[r, "\mathbf{A}"] & \mathcal{R} \\
\hat{\mathbf{G}}_k : & \mathcal{D}_k \arrow[r, "\hat{\mathbf{S}}_k"] & \mathcal{P}_k \arrow[r, "\mathbf{A}_k"] & \mathcal{R}_k \arrow[u]
\end{tikzcd} \tag{2.4}$$

in recognition of our awareness that $\bar{x}$ is only an 'instrumental variable'.

Our true concern, then, is with the approximation of $\bar{z} = \mathbf{G}(\bar{y})$ with $\mathbf{G} := \mathbf{A} \circ \mathbf{F}^{-1}$ by $\hat{\mathbf{G}}_k(y_k)$ with $y_k := \mathbf{M}_k(\bar{y})$. In this formulation, the ill-posed $\mathbf{F}^{-1}$ appears only as a factor of $\mathbf{G}$. The factorization here and also in taking $\hat{\mathbf{G}}_k = \mathcal{A}_k \circ \mathbf{S}_k$ is a convenience for our techniques of analysis and algorithm development; the ill-posedness of (1.1) is an artifact of this but may not be intrinsic to the real problem since it is entirely possible for $\mathbf{G}$ to be continuous even though *one* of its factors $\mathbf{F}^{-1}$ is not. To the extent that, at each stage, our computational procedures maintain this factorization, we do actually work with the ill-posed problem but the representation and topology for the *intermediate* variable $x \in \mathcal{P}$ are now seen to be somewhat arbitrary — subject to our convenient choice with respect to (2.3). [E.g., if $\mathbf{A}$ is linear, one convenient choice might be to use $x \mapsto \|\mathbf{A}x\|_{\mathcal{R}}$ as a norm for $\mathcal{P}$.]

EXAMPLE 1: The simplest example in which ill-posedness 'disappears' in our present framework of considering the composed problem is that in which $\mathcal{R} = \mathrm{I\!R}$ in (2.2), particularly when $\mathbf{A}$ is linear; compare [1].

Let us suppose that the (direct) map $\mathbf{F} : \mathcal{P} \to \mathcal{D}$ is also linear and write $\mathbf{A} = \boldsymbol{\alpha} \in \mathcal{P}^*$. Since we have $\mathbf{G} := \mathbf{A} \circ \mathbf{F}^{-1} = [\mathbf{F}^{-1}]^* \boldsymbol{\alpha} : \mathcal{D} \to \mathcal{R} = \mathbb{R}$ by definition, we now can solve the equation

$$\mathbf{F}^* \boldsymbol{\gamma} = \boldsymbol{\alpha} \tag{2.5}$$

to obtain $\boldsymbol{\gamma} \in \mathcal{D}^*$ — provided, of course, one has $\boldsymbol{\alpha} \in ran(\mathbf{F}^*)$. Since $\mathbf{F}^{-1}$ is unbounded (by assumption), so is $[\mathbf{F}^{-1}]^* = [\mathbf{F}^*]^{-1}$; the problem (2.5) is 'just as ill-posed' as the original ill-posed problem (1.1) *but*: while y is only 'given' approximately (by measurement), the functional $\boldsymbol{\alpha}$ is given analytically *a priori* in defining the application so we may use *analytic* methods to solve the new 'ill-posed problem' (2.5). Once one has precomputed $\boldsymbol{\gamma}$ in this way, the problem is no longer ill-posed: the genuinely significant map: $y \mapsto z$ is now easily realized as: $y \mapsto \langle \boldsymbol{\gamma}, y \rangle$ and *this* computation is well-posed.

One can also treat, in almost the same way, the situation in which $\mathcal{R}$ is finite dimensional by working with $\mathbf{A}$ 'component by component' and we again precompute $\mathbf{G} := \mathbf{A}\mathbf{F}^{-1}$ and proceed as above. This also gives some insight into the situation in which $\mathbf{F}$ has a nontrivial nullspace: if this happens to be contained in the nullspace of $\mathbf{A}$ (which is necessarily nontrivial if $\mathcal{R}$ is finite dimensional), then $\mathbf{G}$ is well-defined and we can either ignore the intermediate construct $\bar{x}$ entirely or else work with a representative of its equivalence class, factoring out the nullspace in some fashion (which may be fairly arbitrary if there are no genuine grounds for selection). Somewhat similar considerations apply to nonlinear problems although the analysis may then be rather more complicated. The issue, however is the same: to precompute (analytically) an effectively usable realization of the functional $\boldsymbol{\gamma} = \boldsymbol{\alpha} \circ \mathbf{F}^{-1}$ under hypotheses ensuring that this is continuous.

It is interesting that a variant of this approach [6], [8] can be used directly for an infinite dimensional ill-posed problem by noting the equivalence of (1.1) to

$$\boldsymbol{\gamma}_\nu(\bar{x}) = \langle \boldsymbol{\alpha}_\nu, \bar{y} \rangle \qquad \nu = 1, 2, \ldots \tag{2.6}$$

for a suitable sequence $\{\boldsymbol{\alpha}_\nu\}$ in $\mathcal{D}^*$ — although this no longer gives a well-posed problem except in approximating by truncation: (2.6) for $\nu = 1, \ldots, N$ with minimization of the solution norm determines x_N with convergence $x_N \to \bar{x}$, more precisely, to the minimum norm solution of (1.1) if that is unique. ■

EXAMPLE 2: Now suppose $\mathbf{A} : x \mapsto z : \mathcal{P} \to \mathcal{R}$ represents finding the solution z of a PDE with coefficients involving a parameter $x \in \mathcal{P}$ (system identification). Then the result may well be quite insensitive to certain (structured) perturbations of $x \in \mathcal{P}$. In this context it is very possible that $\mathbf{F}$ is roughly the same as $\mathbf{A}$, say, with a diffent choice of initial or boundary data for the PDE. This can be expected to make inversion of $\mathbf{F}$ ill-posed as we well know — **but**

the structure of the uncertainty is precisely such that this uncertainty probably does not really matter much! The general principle, here, is that if some 'aspects' of the unknown parameter x do not (significantly) affect the output y for purposes of identification, then those same aspects will not (significantly) affect the result z desired for the application — the compactness of $\mathbf{A}$ 'cancels' the compactness of $\mathbf{F}$ so one may reasonably hope that $\mathbf{G}$ is continuous. As in Example 1, this may also 'handle' a lack of injectivity in $\mathbf{F}$. Compare, also, the discussion in [2]. ■

At this point we note another highly significant effect of working in the context of an application. We have implicitly been viewing the parameter space $\mathcal{P}$ as a Banach space with only the implicit restriction from (1.1) that $\bar{x}$ is in the domain of $\mathbf{F}$. We now note that one often has additional auxiliary information about $\bar{x}$ — e.g., from prior measurements or from the physical context of the model. The effective use of such information (data fusion) is often computationally difficult but essential to getting reliable results. For the present, we represent this information by adjoining to (1.1) the *a priori* constraint

$$\bar{x} \in \mathcal{K} \tag{2.7}$$

for some specified set $\mathcal{K} \subset \mathcal{P}$. [It should be emphasized that, unlike occasional practice in the community, (2.7) is only legitimate in an applied context to the extent that it represents genuine knowledge about the unknown parameter (or, as above, imposition of a consciously arbitrary selection to handle nonuniqueness) and is not merely wishful thinking for the convenience of proving theorems of 'justification'.]

It is fortunate that in many applications the available information provides us with a *compact* set $\mathcal{K}$. As a first consequence, we note that it is this compactness which makes it possible for us to utilize the simplifying assumption (1.4) in working with an infinite dimensional problem with data given, as in (2.2), by a finite number of measurements of finite accuracy. As a second consequence, we also note that the ill-posedness of the problem may again disappear in some sense: a standard theorem of topology asserts that if the continuous map $\mathbf{F}$ is injective and is considered on a compact set $\mathcal{K}$, then $\mathbf{F}^{-1}$ is necessarily also continuous: if one now replaces $\mathbf{F} : \mathcal{P} \to \mathcal{D}$ by $\mathbf{F} : \mathcal{K} \to \mathcal{D}$, one will not only get ε_k finite in (1.5) but will have (1.6). It is this theorem and related considerations which suggests the importance for ill-posed problems of injectivity proofs *per se*, although we note that the significance of this injectivity is for the restriction of $\mathbf{F}$ to $\mathcal{K}$, possibly after the imposition of a 'selection'.

Assuming we have chosen an appropriately application-related norm for $\mathcal{P}$, we may now change notation and rewrite (2.1) as

$$\mathcal{J} = \varphi(\delta) + \psi(\varepsilon(\delta))\ [+\gamma(\hat{\mathbf{S}})]. \tag{2.8}$$

Here, $\varphi(\delta)$ is the cost of making measurement accurate enough to ensure, utilizing the constraint (2.7), that

$$\bar{y} := \mathbf{F}(\bar{x}) \in \mathcal{Y}_\delta := \{y \in \mathcal{D} : \|y - \hat{y}\| \leq \delta\} \tag{2.9}$$

where $\hat{y} = y_k$ is the nominal data value; $\varepsilon(\delta)$ is the radius of $\mathcal{U}(\mathcal{Y}_\delta) \cap \mathcal{K}$ so

$$\bar{x} \in \mathcal{U}(\mathcal{Y}_\delta) \cap \mathcal{K} \subset \{x \in \mathcal{P} : \|x - \hat{x}\| \leq \varepsilon(\delta)\} \tag{2.10}$$

where $\hat{x}$ is a 'center' for $\mathcal{U}(\mathcal{Y}_\delta) \cap \mathcal{K}$.

The original ill-posedness of the problem reasserts itself in that one now has no hope of (1.7), but can expect some extremely slow rate of convergence $\varepsilon(\delta) \to 0$ as $\delta \to 0$. This is further exacerbated on noting that, since we are using the same compactness of $\mathcal{K}$ and so of $\mathbf{F}(\mathcal{K})$ to control the relation between $n(k)$ in (2.2) and δ_k (even with idealized measurements of infinite accuracy), we may expect the cost $\varphi(\mathcal{Y}_\delta)$ in (2.1) to increase quite rapidly as one might hope to let $\delta = \delta_k$ become small. This may, of course, mean that the practical application is simply infeasible: even neglecting computational costs, one may have $\psi(\varepsilon(\delta))$ unacceptably large unless δ becomes small enough that one has $\varphi(\delta)$ unacceptably large.

One interpretation of practical importance is that of '*a posteriori* analysis', in which the measurements have already been made so $\mathcal{Y}$ is fixed. A 'solution' now means some effective determination of the *set* $\mathcal{U} = \mathcal{U}(\mathcal{Y})$. The analyst's first task is to use the *a priori* specification of $\mathcal{K}$, an understanding of $\mathbf{M}_k$, and the measured data to represent $\hat{y}$ in some convenient form with a bound δ on its uncertainty. The second task is then to construct some algorithm denoted by $\hat{\mathbf{S}}$, enabling us to compute an approximation, and a bound ε such that **reliably** one has

$$\bar{x} \in \mathcal{U}(\mathcal{Y}) \subset \{x \in \mathcal{P} : \|x - \hat{x}\| \leq \varepsilon\} \qquad [\hat{x} := \hat{\mathbf{S}}(\hat{y})]. \tag{2.11}$$

In this case the cost will be $\mathcal{J} = \psi(\varepsilon) + \gamma(\hat{\mathbf{S}})$. Note that the data for the construction of $\hat{\mathbf{S}}$ actually consists of $\mathcal{K}$ and δ as well as $\hat{y}$. While much is often made of the 'regularized' continuity properties of $\hat{y} \mapsto \hat{x}$ as compared with $\mathbf{S}$, we assert that the *only* significance of $\hat{\mathbf{S}}$ comes through (2.11) and, indeed, that one should be careful not to let such regularization obscure the genuine uncertainty involved in replacing $\bar{x}$ by $\hat{x}$, which is intrinsic to the problem.

In contrast, one may consider the interpretation as '*a priori* analysis'. While eventually this must be interpreted as above, one now has an opportunity to analyze the problem *before* the measurements have been already made and, within the limitations of technological feasibility, to treat (2.8) as an optimization problem, with (2.10) replaced by (2.11) in defining $\varepsilon = \varepsilon(\delta)$. It is this mode of analysis which leads to the considerations of our next section and which is closely related to to the suggestion above of embedding the problem in a sequence of potential computations of increasing accuracy.

Without further comment, we note — as an obvious parallel to adaptive control theory and the sequential decision theory of statistical analysis — an intermediate mode of analysis in which one might incrementally optimize (2.8) for the proposal of new measurements based on a sensitivity analysis, etc., in the light of the new $\mathcal{K}$ determined by the computations thus far.

The formulation already implicit in (1.3) — viewing $\mathcal{Y}$ as an undifferentiated fixed set — implies a deterministic context with specified error bounds for measurements and the treatment of *a priori* auxiliary information as a constraint (2.7) is consistent with that. In this paper we confine ourselves to this deterministic setting. A brief comment is in order here, however, as to the related nature of interpretation in contexts involving statistical considerations for these uncertainties. We first note that any probability distribution for the noise determines a corresponding probability distribution for the uncertainty in the solution by

$$Prob\{\bar{x} \in \mathcal{S}\} = Prob\{\bar{y} \in \mathbf{F}(\mathcal{S})\} \tag{2.12}$$

and that this may be combined with an uncertainty distribution in the *a priori* information by a relatively standard Bayesian computation. A plausible choice for $\hat{\mathbf{S}}$ for the *a posteriori* mode of interpretation might now be to use a 'maximum likelihood estimator' (i.e., selecting $\hat{x}$ to maximize the relative probability of observing the given data); we will, however, view our goal in the statistical context to be much as above for the deterministic setting, except that the total cost of (2.1) would now be replaced by an 'expected total cost' with respect to the Bayesian distribution or one might seek an assurance that the *risk* (= *'probability that the total cost might exceed some acceptable level'*) is acceptably small. To the extent that the relevant distributions can be reasonably expressed simply in terms of distributions for the norms of uncertainties, it may be possible to use the deterministic estimates below for the relevant statistical computations.

3 Estimates and asymptotics

For well-posed problems the usual analysis leads to the result that:

> *One can find the desired result to any degree of specified precision*
> *— provided one is prepared to work hard enough.*

Here, what we mean by 'working hard enough' is: making enough measurements with enough accuracy and then doing enough computation. The history of analysis of ill-posed problems has consisted primarily of demonstrations that this remains true for ill-posed problems even without considering the composed framework introduced in (2.4). As already noted, the principal difference is that the ill-posedness shows up in an explosive growth of the accuracy requirements for the measurements as one seeks more and more precision in the result. We observe that *ill-posedness resides primarily in the asymptotics* so that for some specific level of desired precision it is quite possible that the problem may still be entirely reasonable.

It is traditional to be concerned about the convergence of an approximation scheme, that

$$x_k := \hat{\mathbf{S}}_k(y_k) \to \bar{x} \quad \text{as } y_k \to \bar{y} \ (\delta_k \to 0), \tag{3.1}$$

and in many cases one can prove this. It is not at all clear, however, that this has much relevance to actual computation, with or without a 'rate of convergence' estimate as in (1.6). In practice one never actually approaches the asymptotic limit, so such results are significant primarily as a guide — e.g., providing an error estimate at any stage so as to select appropriate computational parameters, such as a stopping criterion. Note, too, that one is only *guessing* the magnitudes of any 'constants' appearing in an asymptotic expression unless it is reasonable to view these as available *a priori* from the model — there is a big difference between knowing that some constant exists and knowing (even a rough estimate of) its value. [Ironically, one may also have an easily computable estimate yet have no reasonable way to determine whether one is in its domain of validity.] This is a consideration which applies equally well to instances of the same tradition for standard (well-posed) problems and techniques of *a posteriori* error analysis have been developed for those concerned by this. These techniques are typically based on the asymptotics of the residuals and are inapplicable to ill-posed problems. We have not observed, however, any development of corresponding modes of analysis which *would* be applicable in addressing this difficulty.

Our emphasis, then, is on the estimation of $\varepsilon = \varepsilon(\delta)$ — for fixed δ and with computable constants. This addresses both the context of *a posteriori* analysis and also permitting the optimization involved in an *a priori* analysis, while *a fortiori* providing the asymptotics as $\delta \to 0$. Note that the ε we seek is a bound on the 'worst case error':

$$\varepsilon = \varepsilon(\delta) := \inf_{\hat{\mathbf{S}}} \max_{\hat{y}} \sup_{x} \{ \|x - \hat{\mathbf{S}}(\hat{y})\| : x \in \mathcal{K},\ \|\mathbf{F}(x) - \hat{y}\| \leq \delta \} \tag{3.2}$$

with the implicit assumptions that the $\mathcal{D}$-norm used to measure data perturbations reasonably corresponds to the realistic uncertainties of measurement and that, similarly, the $\mathcal{P}$-norm reasonably corresponds to the related uncertainties in the result when $\hat{x}$ is to be used for $\bar{x}$ in the application under consideration. While one might possibly obtain a sharper bound dependent on $\hat{y}$ (e.g., by allowing for a possible preprocessing step of projecting $\hat{y}$ to $\mathbf{F}(\mathcal{K})$), we will content ourselves with estimating $\varepsilon(\delta) = \varepsilon(\delta; \mathcal{K})$ so (3.2) includes, on the right, a '$\max_{\hat{y}}$' (which may be taken formally over all $\hat{y} \in \mathcal{D}$, but actually reduces to considering only $\hat{y} \in \mathbf{F}(\mathcal{K}) \subset \mathcal{D}$). We must consider the '$\sup_x$' in (3.2) since this is necessary to ensure a bound on $\|\bar{x} - \hat{x}\|$: as already noted, with the information given, every $x \in \{x \in \mathcal{K} : \|\mathbf{F}(x) - \hat{y}\| \leq \delta\}$ is an equally valid candidate to be $\bar{x}$. While a true '$\min_{\hat{x}}$' is attained for each $\mathcal{Y}_\delta$ when $\hat{x}$ is the unknown center of $\mathcal{U}(\mathcal{Y}_\delta)$, the minimization here is actually taken as an inf over feasibly implementable quasi-inverse maps

$$\hat{\mathbf{S}} = \hat{\mathbf{S}}(\cdot;\mathcal{K},\delta) : \hat{y} \mapsto \hat{x} = \hat{\mathbf{S}}(\hat{y}). \tag{3.3}$$

In general this is an 'inf' rather than a 'min', but we will be able to determine an optimal $\hat{\mathbf{S}}(\cdot;\mathcal{K},\delta)$ in certain special cases for which an explicit construction is possible; cf., e.g., [12], [11] and below.

A fairly typical form for the constraint (2.7) is that

$$\mathcal{K} = \mathcal{K}_M := \{\mathbf{E}\xi : \|\xi - \xi_*\| \leq M\} \tag{3.4}$$

for a suitable (linear) map $\mathbf{E} : \tilde{\mathcal{X}} \to \mathcal{P}$ and some $\xi_* \in \tilde{\mathcal{X}}$ (which, for simplicity, we will take to be $\xi_* = 0$); for example, $\mathbf{E}$ might be the embedding map of $\tilde{\mathcal{X}} = C^2(\bar{\Omega})$ into $\mathcal{P} = L^2(\Omega)$, representing a regularity condition on the unknown parameter $\bar{x}$ with an explicit bound[1] M in the stronger norm. With $\bar{x} = \mathbf{E}\bar{\xi}$ and assuming $\hat{\mathbf{S}}$ is Lipschitzian so

$$\|\hat{\mathbf{S}}(y_1) - \hat{\mathbf{S}}(y_2)\| \leq \beta\|y_1 - y_2\| \tag{3.5}$$

for some β, we have, using (1.1), the estimate

$$\|\bar{x} - \hat{\mathbf{S}}(\hat{y})\| = \|[\mathbf{1} - \hat{\mathbf{S}} \circ \mathbf{F}](\mathbf{E}\bar{\xi}) - [\hat{\mathbf{S}}(\bar{y}) - \hat{\mathbf{S}}(\hat{y})]\|. \tag{3.6}$$

Then, setting

$$\mathbf{\Gamma}(\beta) := \inf_{\hat{\mathbf{S}}}\{\| \left[\mathbf{1} - \hat{\mathbf{S}} \circ \mathbf{F}\right] \circ \mathbf{E}\|_{Lip} : \|\hat{\mathbf{S}}\|_{Lip} \leq \beta\} \tag{3.7}$$

[where $\|\cdot\|_{Lip}$ denotes the minimal Lipschitz constant of a map], we have

$$\varepsilon(\delta) \leq \mathbf{\Gamma}(\beta)M + \beta\delta \tag{3.8}$$

for any β. Clearly we can minimize the right hand side of (3.8) over β and, setting

$$\tilde{\mathbf{\Gamma}}(\tilde{\beta}) := \begin{cases} -\mathbf{\Gamma}(-\beta) & \text{if } \tilde{\beta} \leq 0 \\ \infty & \text{else,} \end{cases}$$

this gives the fundamental estimate of *'Optimal Filtering'*

$$\varepsilon(\delta;\mathcal{K}_M) \leq M \min_{\beta \geq 0}\{\mathbf{\Gamma}(\beta) + \beta\delta/M\} = -M\tilde{\mathbf{\Gamma}}^*(\delta/M) \tag{3.9}$$

where $\tilde{\mathbf{\Gamma}}^*$ denotes the Legendre-Fenchel conjugate of $\tilde{\mathbf{\Gamma}}$. In many cases this estimate is 'asymptotically sharp', in that (with M fixed) there is a sequence $\delta_k \to 0$ and a sequence of pairs $\bar{x}_k, \bar{x}'_k \in \mathcal{K}_M$ such that

$$\|\mathbf{F}(\bar{x}) - \hat{y}_k\|, \|\mathbf{F}(\bar{x}) - \hat{y}_k\| \leq \delta_k \quad \|\bar{x}_k - \bar{x}'_k\| = M\tilde{\mathbf{\Gamma}}^*(\delta_k/M).$$

[1]We remark that, although the assumptions are comparable, the approach presented here, which we term the 'Method of Optimal Filtering', is quite different from the weighted 'Least Squares' optimization used in [5]. As we shall see, this approach gives an estimate which is genuinely optimal in a specified sense.

Our next problem, of course, is to show how to compute $\mathbf{\Gamma}(\cdot)$ (or, better, $\tilde{\mathbf{\Gamma}}(\cdot)$) and the optimizing $\hat{\mathbf{S}}$. The principal result of [10] is that a fairly explicit construction is possible in the Hilbert space setting (in which case we may identify the spaces $\tilde{\mathcal{X}}$, $\mathcal{P}$, $\mathcal{D}$), provided $\mathbf{E}$ and the map $\mathbf{F} \circ \mathbf{E}$ are commuting normal linear operators. This is a considerable restriction, of course, but still admits a variety of interesting and significant examples. For simplicity, we describe the construction assuming discrete spectra so there is an orthonormal basis $\{w_k\}$ with

$$\mathbf{FE}w_k = \lambda_k w_k \qquad \mathbf{E}w_k = \epsilon_k w_k$$

and we take $\hat{\mathbf{S}}$ in the form $\hat{\mathbf{S}} : w_k \mapsto \rho_k w_k$ so

$$\hat{\mathbf{S}}\hat{y} = \sum_k \rho_k c_k w_k \quad \text{with } c_k := \langle \hat{y}, w_k \rangle. \tag{3.10}$$

From (3.7) we wish to minimize

$$\|\mathbf{E} - \hat{\mathbf{S}}\mathbf{FE}\| = \max_k \{|\epsilon_k - \rho_k \lambda_k|\} \tag{3.11}$$

by our choice of the coefficients $\{\rho_k\}$, subject to $|\rho_k| \le \beta$. Treating each k independently and noting that λ_k/ϵ_k is precisely the corresponding eigenvalue φ_k of $\mathbf{F}$, this minimization is clearly realized by taking

$$\rho_k = \begin{cases} 1/\varphi_k & \text{if } \beta|\varphi_k| \ge 1 \\ \beta \arg\{1/\varphi_k\} & \text{else} \end{cases} \tag{3.12}$$

so $|\epsilon_k - \rho_k \lambda_k| = (|\epsilon_k| - \beta|\lambda_k|)_+$. At this point we introduce the spectral function of $|\mathbf{E}|$ relative to $|\mathbf{FE}|$:

$$s(\lambda) := \max\{|\epsilon_k| : |\lambda_k| = \lambda\} \tag{3.13}$$

with the interpretation that this 'max' is $-\infty$ if λ is not in the spectrum of $\mathbf{FE}$. We now have

$$\mathbf{\Gamma}(\beta) = \max_k\{|\epsilon_k| - \beta|\lambda_k|\} = \sup_\lambda\{s(\lambda) - \beta\lambda\}. \tag{3.14}$$

With slight manipulation, this gives both

$$\beta = \beta(\tilde{\nu}) = \operatorname{argmin}\{\mathbf{\Gamma}(\beta) + \beta\tilde{\nu}\} = \operatorname{argmin}\{\beta\tilde{\nu} - [-s]^*(-\tilde{\nu})\} \tag{3.15}$$

with $\tilde{\nu} := \delta/M$ for use in (3.12) and the new version of our fundamental estimate:

$$\tilde{\mathbf{\Gamma}}(\tilde{\nu}) = [-s]^*(\tilde{\nu}) \qquad \varepsilon(\delta; \mathcal{K}_M) = -M[-s]^{**}(\delta/M) \tag{3.16}$$

where * again indicates the Legendre-Fenchel conjugate, whence

$$\varepsilon(\delta; \mathcal{K}_M) = M s(\delta/M) \qquad \text{if } s(\cdot) \text{ is concave.} \tag{3.17}$$

It is important to note that we may often have a fairly simple expression $\tilde{s}(\cdot)$, giving $s(\lambda) = \tilde{s}(\lambda)$ when $\lambda = \lambda_k$, and we may then use (3.15), (3.16), (3.17) with s replaced by $\tilde{s}$ and retain asymptotic correctness.

At this point it is interesting to compare the estimates obtained here with those obtainable by some other standard schemes. Following [3], we note that the assumption there (p. 18) that $Kg \in \mathcal{R}(\tilde{K}^\nu)$ is, in our present notation, just our requirement that $\bar{x} \in \mathcal{K}_M$ as in (3.4) with $\mathbf{E} = |\mathbf{F}|^{2\nu}$ and $\xi_* = 0$, although with M unspecified. One then has $\tilde{s}(\lambda) = \lambda^{1-\vartheta}$ so (3.17) gives

$$\varepsilon(\delta; \mathcal{K}_M) = M^\vartheta \delta^{1-\vartheta} \text{ with } \vartheta = \frac{1}{2\nu + 1} \tag{3.18}$$

if one uses (3.10), (3.12) with the correct choice of β. Note that the assumption on $\mathcal{K}_M$ makes (3.18) and the choice of β independent of the particular problem (i.e., of $\mathbf{F}$). In comparison, Groetsch notes accuracy $\mathcal{O}\left(\delta^{1-1/2\nu}\right)$ for the Landweber-Fridman iteration, which is reasonably comparable for large $\tilde{\nu}$ and notes exactly the same asymptotic *order* as in (3.18) for Tikhonov regularization when $\nu \leq 1$ but shows that this never gives better accuracy than $\mathcal{O}\left(\delta^{2/3}\right)$ for larger ν.

Actually, it is quite easy to obtain such results by the mode of analysis we have used above. For Tikhonov regularization, for example, one is effectively using (3.10) with $\rho = \rho_T := \varphi/[\varphi^2 + \alpha]$ so maximization of $(1 - \rho_T\varphi)\epsilon$ (again with $\epsilon = \varphi^{2\nu}$ for $0 < \nu < 1$) easily gives $\mathbf{\Gamma}_T(\alpha) = \left[\nu^\nu(1-\nu)^{1-\nu}\right]\alpha^\nu$ and $\beta(\alpha) = \|\hat{\mathbf{S}}_T\| = 1/2\sqrt{\alpha}$ whence optimization of $M\mathbf{\Gamma}_T(\alpha) + \delta\beta_T(\alpha)$ over α gives the value: $\alpha = \alpha_{opt} = [C\delta/M]^{2\vartheta}$ (ϑ as in (3.18)) with $C = 1/4\nu^{\nu+1}(1-\nu)^{1-\nu}$. If we consider the approximation method based on truncated singular value decomposition, we are again essentially using (3.10), now with $\rho_{SV} := \{1/\varphi \text{ if } \varphi \geq \bar{\varphi};\ 0 \text{ else}\}$ so $\beta = 1/\bar{\varphi}$, etc., whence we get $\bar{\varphi}_{opt} = [(\delta/M)/2\nu]^\vartheta$ by elementary calculus and then $\varepsilon_{SV} = CM^\vartheta\delta^{1-\vartheta}$ for comparison with (3.18) with $C = \nu^\vartheta(1 + 1/2\nu)$ for $\nu > 0$.

EXAMPLE 3: We sketch the application of these ideas to the Backwards Heat Equation (BHE). Thus, we consider, say,

$$u_t = \Delta u \text{ on } \mathcal{Q} := (0,T) \times \Omega \text{ with } u\Big|_{\partial\Omega} = 0 \tag{3.19}$$

and take $\bar{x} = \bar{x}_\tau := u\Big|_{t=\tau}$ (for some $0 \leq \tau < T$) as the unknown 'parameter' and $\bar{y} := u\Big|_{t=T}$ as data. Our treatment here is based on the fact that there is a semigroup associated with the forward problem:

$$\mathbf{H}(s) : u\Big|_{t=\sigma} \longmapsto u\Big|_{t=\sigma+s} \tag{3.20}$$

and that this semigroup can be expressed, following Fourier, in terms of the orthonormal spectral expansion of the generator $(-\Delta + \mathrm{BC})$:

$$-\Delta w_k = \mu_k w_k \text{ on } \Omega \text{ with } w_k = 0 \text{ on } \partial\Omega. \tag{3.21}$$

We distinguish two quite distinct cases: $\tau = 0$ and $0 < \tau < T$ with an assumed explicit bound M on the unknown initial data $\xi = u(0,\cdot)$ in $H_0^{2\vartheta}(\Omega)$ for the first case and in $\mathcal{H}$ for the second; in each case we take $\mathcal{P} = \mathcal{D} = \mathcal{H} := L^2(\Omega)$.

For the first case, $\mathbf{F}$ is just $\mathbf{H}(T)$ and $\mathbf{E}$ is the embedding: $H_0^{2\vartheta}(\Omega) \hookrightarrow L^2(\Omega) = \mathcal{H}$. If we use the norm

$$\|\xi\|_{2\vartheta} := \|(-\Delta)^{\vartheta}\xi\|_{\mathcal{H}} = \left[\sum_k \mu_k^{2\vartheta} c_k w_k\right]^{1/2} \qquad \left(\xi = \sum_k c_k w_k\right),$$

then $\mathbf{E}$ and $\mathbf{FE}$ commute and we have

$$\varphi_k = e^{-T\mu_k} \qquad \epsilon_k = \mu^{-\vartheta} \tag{3.22}$$

for use in the formulas above. Note that we here can obtain $\tilde{s}(\cdot)$ on $\mathbb{R}_+$ as the (concave) inverse of the convex monotone function $\lambda(\epsilon) := \epsilon \exp\left[-T/\epsilon^{\vartheta}\right]$. In principle the optimizations in (3.14) and (3.9) require only elementary calculus, but there is no simple formula so actual solution must be numeric. A straightforward, although somewhat messy, computation gives the asymptotic result (for very small δ) that one can approximate the optimal β by solving

$$\beta(\log\beta)^{\vartheta+1} = \vartheta T^{\vartheta} M/\delta$$

and still realize through use of (3.12) the optimal error estimate

$$\varepsilon(\delta;\mathcal{K}_M) = MT^{\vartheta}(\log 1/\delta)^{-\vartheta} + \ldots \tag{3.23}$$

where '...' denotes something 'negligible in comparison for δ sufficiently small'. We note that (3.23) is better than an order estimate for the error — it contains the correct constant, i.e., it is the first term in an asymptotic expansion for $\delta \to 0$ — but would still be useless for either our *a posteriori* or *a priori* interpretations until we would actually specify satisfactorily the meaning of 'sufficiently small'. For example, if we were to 'solve' (3.23) with some margin, prescribing a measurement accuracy $\delta = \exp\left[-T(2M/\hat{\varepsilon})^{1/\vartheta}\right]$, we would have a guarantee that we achieve the desired accuracy $\hat{\varepsilon}$ — once we would be in the domain of validity (as yet unspecified!) of the bound $\varepsilon(\delta;\mathcal{K}_M) \leq 2MT^{\vartheta}(\log 1/\delta)^{-\vartheta}$.

For the second case we follow [11]. Note that we then have $\mathbf{E} = \mathbf{H}(\tau)$ and $\mathbf{F} = \mathbf{H}(T-\tau)$ so

$$\varphi_k = e^{-(T-\tau)\mu_k}, \qquad \epsilon_k = e^{-\tau\mu_k}$$

so $\tilde{s}(\lambda) = \lambda^{\tau/T}$. Using this in (3.14), (3.15) gives, by elementary calculus, the explicit formula

$$\beta = \beta(\tilde{\nu}) := \frac{\tau}{T}(1/\tilde{\nu})^{1-\tau/T} \qquad (\tilde{\nu} = \delta/M) \tag{3.24}$$

for (3.12). We now recover the well-known 'log-convexity' estimate from (3.17), which here gives

$$\varepsilon(\delta; \mathcal{K}_M) = M^{1-\tau/T} \delta^{\tau/T}. \tag{3.25}$$

At this point we note that the construction (3.10) involves an infinite series expansion which cannot actually be implemented without some further approximation: we can only work computationally with a finite sum and also will typically have to compute the eigenvalues and eigenfunctions of (3.21). A 'brute force' approach to this error analysis would lead to completely unrealistic estimates but, as was noted in [11], the same analysis used here for (3.19) also applies to any plausible discretization. Such a discretization will certainly give a finite sum in (3.10) and (3.21) will be replaced by computation of the eigenvalues and eigenvectors of a finite (self-adjoint) matrix, so the total computation is implementable. Fortunately, the error can be estimated without bounding the errors in the spectral approximation, only considering the approximation error $\delta_*(\cdot)$ for the discretization of the forward problem (using the *a priori* bound on the initial data). When $\hat{x}$ is computed by this approximation procedure we thus get the composite error bound

$$\|u(\tau, \cdot) - \hat{x}\| \leq \delta_*(\tau) + M^{1-\tau/T} \left[\delta + \delta_*(T)\right]^{\tau/T}, \tag{3.26}$$

which, as expected, reduces to (3.25) as the computation is refined. [This assumes we have computed exactly the spectral data for the discretized problem and the coefficients for $\hat{y}$ with respect to those eigenvectors, but these additional sources of error can also be handled.] This treatment of discretization applies equally well to the first case ($\tau = 0$). ■

ACKNOWLEDGMENTS: I am pleased to thank H. Engl, H. Haario, W. Rundell, and UMBC for their different contributions, which made it possible for me to participate in the excellent and enjoyable conference at St. Wolfgangsee.

References

[1] R.S. Anderssen, *The linear functional strategy for improperly posed problems*, in (J.R. Cannon and U. Horning, eds.) *Inverse Problems* [ISNM #77], Birkhäuser, Basel, 1987, pp. 11–30.

[2] W. Chewning and T.I. Seidman, *A convergent scheme for boundary controls for the heat equation,* SIAM J. Control Opt. **15** (1977), pp. 64–72.

[3] C.W. Groetsch, *The theory of Tikhonov regularization for Fredholm equations of the first kind*, Pitman, London, 1984.

[4] J. Hadamard, *Lectures on the Cauchy Problem in Linear Partial Differential Equations*, Yale Univ. Press, New Haven, 1923.

[5] K. Miller, *Least-squares methods for ill-posed problems with a prescribed bound,* SIAM J. Math. Anal. **1** (1970), pp. 52–74.

[6] T.I. Seidman, *The solution of singular equations, I: linear equations in Hilbert space*, Pac. J. Math. **61** (1975), pp. 513–520.

[7] T.I. Seidman, *Computational approaches to ill-posed problems: general considerations,* in *Proc. 1976 Conf. on Inf. Sci. and Systems,* Johns Hopkins Univ. (1976), pp. 258–262.

[8] T.I. Seidman, *Convergent approximation methods for ill-posed problems, Part I: general theory*, Control and Cybernetics **10** (1981), pp. 31–49.

[9] T.I. Seidman, *Some 'complexity' issues for ill-posed problems,* in *Inverse Problems in Partial Differential Equations* (D. Colton, R. Ewing, W. Rundell, eds.), SIAM (1990), pp. 17–35.

[10] T.I. Seidman, *The method of 'optimal filtering' for a class of ill-posed problems*, in *Wave Propagation and Inversion* (W. Fitzgibbon and M. Wheeler, eds.) SIAM (1992), pp. 108–123.

[11] T.I. Seidman, *Optimal filtering for the backward heat equation*, SIAM J. Num. Anal. (1994), to appear.

[12] T.I. Seidman and L. Eldén, *Optimal filtering for the sideways heat equation*, Inverse Problems **6** (1990), pp. 681–696.

[13] A.G. Werschulz, *What is the complexity of ill–posed problems?* Numer. Fnl. Anal. and Optim. **9** (1987), pp. 945–967.