11 Inverse Problems

Part II: Making the reconstructed solution smooth — the idea behind the Tikhonov regularization

In the previous Lecture we saw that inverse problems for certain integral equations (see Section 5 of Lecture 10) are *unstable*, in the sense that the primordial data reconstructed from the measurements can be very jagged and may have nothing in common with the true data. The first example considered in Lecture 10 was the problem of differentiation (of a noisy function). For it, we actually were able to ameliorate the instability. Our (partial) remedy consisted of choosing the value for the discretization step h to be in a certain relation to the noise amplitude. For the optimal choice $h_{\rm opt}$ of the discretization step, the jaggedness of the reconstructed data was small as long as the measurements' noise was small.

In this Lecture we will seek an alternative method of smoothing the reconstructed data. Our motivation here is that the method of Lecture 10 has (at least) two drawbacks. First, if the measurements come from an experiment, the value of h is fixed by the experimental setup (e.g., by how often the measurements are taken) and hence cannot really be optimized. Even if it so happens that $h_{\rm opt}$ is, say, close to $2h_{\rm experim}$, then a possibility would be to throw away every other measurement record and base the reconstructed data on the remaining set. Intuitively, however, this option is not appealing because the information thrown away may be of some importance. Second, and *more importantly*, the method of Lecture 10 cannot be used to ameliorate the instability of the "ultimately unstable" problem considered in Sections 4 and 5 of that Lecture. Therefore, an alternative method not having these two drawbacks, is needed.

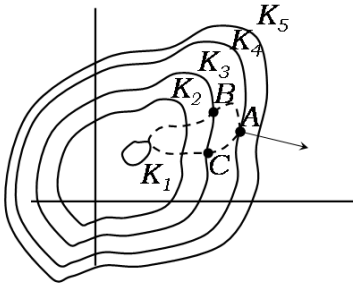
We will introduce such a method using as a starting point the problem of differentiation of a nonsmooth function. This new method belongs to a broad class of methods known as the ***Tikhonov regularization***. Then we will extend this technique to the "ultimately unstable" problem of Lecture 10. In both cases, we will highlight the role of certain *eigenvectors* as the ***"carriers of the instability"***. Before we introduce the idea of Tikhonov regularization, we need to review the elementary theory of constrained optimization using Lagrange multipliers.

11.1 Review of the constrained optimization problem; Lagrange multipliers

Consider the following problem:
Find a local extremum of function $f(x,y)$ *under the* ***constraint*** *that* x *and* y *must satisfy*

$$g(x,y) = c\,, \tag{11.1}$$

where $g(x,y)$ *is some other known function and* c *is a given constant.*



The ***level curve*** defined by Eq. (11.1), i.e. the curve where $g(x,y) = c$, is shown in the figure on the left by a dashed line. Several level curves $f(x,y) = K$ for $K_1 < K_2 < K_3 < K_4 < K_5$ are also shown in the figure by solid lines. The solution to the problem in question is point A where the level curves $g(x,y) = c$ and $f(x,y) = K_4$ are tangent to each other.

Indeed, for $K < K_4$, there exist solutions to the constraint equation (11.1) (e.g., points B and C), but they do not maximize $f(x,y)$. On the other hand, for $K > K_4$, there are no solutions to $g(x,y) = c$.

Now, at the common tangency point A, the normal vectors to both $g(x,y) = c$ and $f(x,y) = K_4$ are parallel. From Calculus III you know that the normal vector to a curve is pointed along the gradient

$$\vec{\nabla} f(x,y) \equiv \frac{\partial f}{\partial x}\vec{i} + \frac{\partial f}{\partial y}\vec{j}\,,$$

where $\vec{i}$ and $\vec{j}$ are the unit vectors along the x- and y-axes. Thus, the gradients $\vec{\nabla} f$ and $\vec{\nabla} g$ at point A must be parallel, i.e. differ by a scalar factor:

$$\vec{\nabla} f = -\ell \vec{\nabla} g\,. \tag{11.2a}$$

Here ℓ is the aforementioned scalar, called Lagrange multiplier, and the minus appears just as a matter of convention. Thus, we have three scalar equations:

$$\begin{aligned} \frac{\partial f}{\partial x} + \ell \frac{\partial g}{\partial x} &= 0\,, \\ \frac{\partial f}{\partial y} + \ell \frac{\partial g}{\partial y} &= 0\,, \end{aligned} \tag{11.2b}$$

and the constraint equation (11.1), from which one can determine a finite number of solution sets (x, y, ℓ). This method of finding a constrained optimum using Lagrange multipliers can be generalized to the case of more variables and more constraints. For example, for the function $f(x_1, \ldots, x_n)$ of n variables and for $m \le n-1$ constraints

$$g_k(x_1, \ldots, x_n) = c_k\,, \qquad k = 1, \ldots, m\,, \tag{11.3}$$

the counterpart of Eq. (11.2a) is:

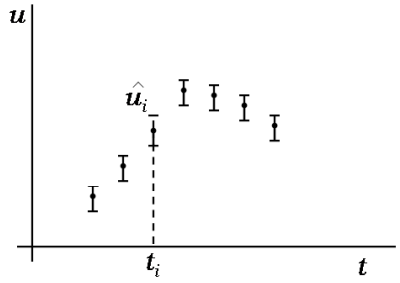
$$\vec{\nabla} f + \ell_1 \vec{\nabla} g_1 + \ldots + \ell_m \vec{\nabla} g_m = \vec{0}\,, \tag{11.4}$$

where, similarly to the first equation on this page, the gradient vector is defined as:

$$\vec{\nabla} f \equiv \left\langle \frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots \frac{\partial f}{\partial x_n} \right\rangle.$$

Verify that the number of unknowns in Eqs. (11.3) and (11.4) equals the number of equations.

11.2 The idea of regularization for a problem of fitting a smooth curve to nonsmooth data



Let $\hat{u}_i, i = 1, \ldots, N$, be experimental data scattered about an *unknown* smooth curve $u(t)$ within a measurement error of $\pm\varepsilon$. For simplicity, we assume that the measured points are evenly distributed along t : $t_{i+1} - t_i = h$. A *task* that a researcher often faces is to approximate the derivative of the unknown curve $u(t)$.

We, however, will be concerned only with finding a *smooth* curve $u(t)$ that is within the experimental error of the measurements $\hat{u}_i$. Once such a curve is obtained, then numerically differentiating it is a straightforward matter; see Eq. (10.5) of Lecture 10.

The condition that $u(t)$ not deviate from the $\hat{u}_i$'s by more than ε reads as:

$$\max_{1\le i\le N} |u(t_i) - \hat{u}_i| \le \varepsilon\,. \tag{11.5}$$

However, we will require a slightly weaker condition:

$$\frac{1}{N}\sum_{i=1}^{N}(u(t_i) - \hat{u}_i)^2 \le \varepsilon^2\,, \tag{11.6}$$

which is easier to analyze than (11.5).

In Linear Algebra, you may have studied a *deceptively* similar problem:
Find the best least-squares fit to a given set of points, such that

$$\sum_{i=1}^{N}(u(t_i) - \hat{u}_i)^2 = \min\,. \tag{11.7}$$

Above, we said 'deceptively similar', because one may think that in our problem we want to minimize the l.h.s. of (11.6). However, this is *not* the case! Let us now explain why problem (11.7) is *not* related to what we want to do with inequality (11.6). Then we will explain which problem we actually *will* solve and how it is related to the constrained minimization problem considered in Section 1.

One obvious difference between the least-squares problem (11.7) from Linear Algebra and our inequality (11.6) is that the r.h.s.'s of these relations look different. A more profound difference is that in (11.6), one does *not* know the functional form of $u(t)$, while in the least-squares problem (11.7) such a form is assumed (usually, as some polynomial of t). Therefore, we need some additional condition to determine the $u(t)$ in (11.6). Above, we have said that this $u(t)$ should be smooth. This requirement of smoothness of $u(t)$, put into a mathematical form, will be the cornerstone of our constrained minimization problem.

Let us emphasize **a very important point**: the requirement that $u(t)$ be smooth **does not** in any way follow from the measurements. It is an *arbitrary* requirement that we want to impose based either on some common sense or on our intuitive understanding of the solution's behavior. Therefore, the "truthfulness" of the solution that we will obtain depends on how well

we can guess the smoothness criterion. This is a fundamental limitation of the regularization method, but this is also the best one can do given the uncertainty that exists in finding a smooth $u(t)$ in this problem.

We now proceed with a mathematical formulation of the regularization technique. Let the measure of smoothness of $u(t)$ be represented by a condition

$$Q[u] = \min\,, \tag{11.8}$$

where $Q[u]$ is some nonnegative ***functional*** of $u(t)$. A functional of a function $u(t)$ is a scalar number that is *not* a function of t but that in some way (partially) accounts for the dependence of u on t. E.g., any of the following:

$$\max_{0\le t\le t_{\max}} |u(t)|\,; \qquad \int_0^{t_{\max}} u(t)dt\,; \qquad \int_0^{t_{\max}} (u'(t))^2 dt\,; \qquad Au(0) + Bu(t_{\max}/3)$$

is a functional of $u(t)$. Similarly, when $\underline{u} = (u_1, \ldots, u_N)^T$ is a finite-length vector, the corresponding functionals are:

$$\max_{1\le i\le N} |u_i|\,; \qquad \sum_{i=1}^{N} u_i\,; \qquad \sum_{i=1}^{N-1} \left(\frac{u_{i+1}-u_i}{h}\right)^2\,; \qquad Au_1 + Bu_{N/3}\,.$$

What functional $Q[u]$ is it meaningful to take *in our problem*? Intuitively, a smooth function probably does not have sharp turns or wiggles, i.e., sections with large curvature. Curvature is proportional to the second derivative of a function. Therefore, as a *plausible condition*, we will require that the mean square of $u''(t)$ be minimized:

$$\int_0^{t_{\max}} (u''(t))^2 dt = \min\,. \tag{11.9}$$

Again, **let us emphasize** that this condition **does not** follow from the measurements and was *arbitrarily* postulated based solely on our idea of *what we expect* the solution $u(t)$ to be.

Thus, the problem we consider is:
Find the minimum of

$$Q[u] = \int_0^{t_{\max}} (u''(t))^2 dt \tag{11.10}$$

subject to the constraint (11.6).
This is not yet exactly the constrained minimization problem considered in Section 1 because (11.6) is an inequality rather than an equation, like (11.1). However, it can be shown (see "Problem I" on p. 514 in the posted article by M. Hanke and O. Scherzer) that unless $u(t)$ happens to be a straight line, then condition (11.6) *must* be satisfied with the " = " sign, not with the " $\le$ ". Then, (11.6) becomes:

$$\sum_{i=1}^{N} (u(t_i) - \hat{u}_i)^2 = N\varepsilon^2\,. \tag{11.11}$$

Now our problem has exactly the form of the constrained minimization problem considered in Section 1: The quantity that we minimize, i.e. the counterpart of function $f(x_1, x_2, \ldots, x_n)$, is given by the l.h.s. of Eq. (11.9), while the constraint equation (the counterpart of (11.3) with $m = 1$) is given by Eq. (11.11). The role of unknowns, which we want to solve for, is

played by $u_i \equiv u(t_i),\ i = 1, \ldots, N$. Therefore, the solution $u(t)$ of our problem must satisfy (11.11) and also a counterpart of Eq. (11.2) (or (11.4) with $m = 1$):

$$\frac{\partial}{\partial u_i}\left[\alpha \int_0^{t_{\max}} (u''(t))^2 dt + h\sum_{i=1}^{N}(u_i - \hat{u}_i)^2\right] = 0, \tag{11.12}$$

where we have denoted $\alpha = 1/\ell$. For the convenience of subsequent calculations, we included the factor h in front of the second term to make it look like a discretized version of $\int_0^{t_{\max}} (u(t) - \hat{u}(t))^2\, dt$. Before proceeding to explain how to solve the combined problem (11.11) and (11.12), let us explain the reason behind the rescaling $\alpha = 1/\ell$.

The size of α determines which term in (11.12),

$$\int_0^{t_{\max}} (u''(t))^2 dt \qquad \text{or} \qquad h\sum_{i=1}^{N}(u_i - \hat{u}_i)^2,$$

we assign more significance to. Our primary goal is to have the curve $u(t)$, and hence the points $u_i \equiv u(t_i)$ on it, to be representative of the measured data points $\hat{u}_i$. This is expressed by condition (11.11). On the contrary, the other condition, (11.9), was postulated by us somewhat arbitrarily, i.e. based on our *expectation* about the smoothness of $u(t)$ but *not on the actual measurements*. Therefore, it makes sense to assign more significance to the second term in (11.12) than to the first one. Mathematically, we assign more significance to (11.11) than to (11.9) if we take α in (11.12) to be *small*. (If $\alpha = 0$, then we would simply ignore our postulated condition (11.9).) It is simply the matter of convention to assign more significance to a term (the second term in (11.12) in this case) by multiplying the *other* term (the first one in (11.12)) by a small number, rather than multiplying the former term by a large number $\ell = 1/\alpha$.

The next issue that we need to address is how to solve the combined problem (11.11) and (11.12). The difficulty here is in the necessity to solve Eq. (11.11), because it is nonlinear, whereas Eq. (11.12) is, as we will see later, linear. A common way to resolve this issue is to *omit* Eq. (11.11) and consider *only* Eq. (11.12) where one *picks* a value for α and verifies *aposteriori* whether the resulting solution satisfies the original inequality (11.6). So, this is how we will proceed: we will consider only Eq. (11.12) with a small α, with the idea that a verification of inequality (11.6) can be done aposteriori. (As a side note, we mention that a computational algorithm of solving *both* (11.11) and (11.12) was developed in the paper by C. Reinsch posted on-line. However, understanding that algorithm requires nontrivial effort.)

We now need to rewrite the integral term in (11.12) in such a form that it would explicitly depend on u_i and hence we could take the derivative $\partial/\partial u_i$. We employ the finite-difference approximation for $u''(t)$ (see Eq. (9.4) of Lecture 9 and Eq. (10.15) of Lecture 10), which is fairly accurate when $u(t)$ is smooth. Then, for any of the *inner* points:

$$u''(t_i) \approx \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2}\,, \qquad 2 \le i \le N-1\,. \tag{11.13}$$

For the endpoints, we *arbitrarily* postulate that

$$u''(t_2) = u''(t_{N-1}) = 0\,. \tag{11.14}$$

(Other choices of conditions at the endpoints do not qualitatively change the final result as long as N is sufficiently large.) Equations (11.14) and (11.13) yield two conditions on u_1 and

u_N:

$$\begin{aligned} u''(t_2) = 0 \quad &\Rightarrow \quad u_1 = 2u_2 - u_3 \\ u''(t_{N-1}) = 0 \quad &\Rightarrow \quad u_N = 2u_{N-1} - u_{N-2}\,. \end{aligned} \tag{11.15}$$

Therefore, the endpoints u_1, u_N should be considered as known once the remaining unknowns $u_2, \ldots, u_{N-1}$ will have been determined. Then Eqs. (11.13), (11.14) can be written in matrix form for these unknowns:

$$\begin{pmatrix} u''(t_2) \\ u''(t_3) \\ u''(t_4) \\ \vdots \\ u''(t_{N-2}) \\ u''(t_{N-1}) \end{pmatrix} \approx \frac{1}{h^2} \begin{pmatrix} 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \ldots & 0 & 0 & 0 \\ \vdots & & & & & & & \\ 0 & 0 & 0 & 0 & \ldots & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} u_2 \\ u_3 \\ u_4 \\ \vdots \\ u_{N-2} \\ u_{N-1} \end{pmatrix}, \tag{11.16a}$$

or, equivalently,

$$\underline{u''} \approx -H\underline{u}\,, \tag{11.16b}$$

where $(-H)$ is the matrix appearing in (11.16a). The integral in (11.12) can then be written as a finite sum:

$$\begin{aligned} \int_0^{t_{\max}} (u''(t))^2 dt \quad &\approx \quad h \sum_{i=2}^{N-1} (u''(t_i))^2 = h\, (\underline{u''})^T \underline{u''} \\ &= \quad h \cdot (-H\underline{u})^T \cdot (-H\underline{u}) = h\underline{u}^T (H^T H)\underline{u}\,. \end{aligned} \tag{11.17}$$

Next, the finite sum term in (11.12) can be rewritten as:

$$\sum_{i=2}^{N-1} (u_i - \hat{u}_i)^2 + (u_1 - \hat{u}_1)^2 + (u_N - \hat{u}_N)^2\,. \tag{11.18}$$

In principle, one can carry on the calculations with expression (11.18) while using Eqs. (11.15) for u_1 and u_N. However, as we noted already, the endpoints contribute little compared to the sum of the inner $(N-2)$ terms, and hence for simplicity we omit the last two terms in (11.18). Then

$$\sum_{i=1}^{N} (u_i - \hat{u}_i)^2 \approx \sum_{i=2}^{N-1} (u_i - \hat{u}_i)^2 \equiv (\underline{u} - \underline{\hat{u}})^T (\underline{u} - \underline{\hat{u}})\,. \tag{11.19}$$

Substituting (11.17) and (11.19) into (11.12), we obtain:

$$\frac{\partial}{\partial u_i} \left[\alpha \underline{u}^T (H^T H)\underline{u} + (\underline{u} - \underline{\hat{u}})^T (\underline{u} - \underline{\hat{u}}) \right] = 0\,. \tag{11.20}$$

To perform the $\partial/\partial u_i$ operation, we will write this equation as a linear system. Since the first term in (11.20) has the form

$$\underline{u}^T M \underline{u}$$

where $M = \alpha H^T H$. For future use we note that, as shown in elementary Linear Algebra, such a matrix is always symmetric, i.e. $M^T = M$. We will now calculate

$$\frac{\partial}{\partial u_i} \underline{u}^T M \underline{u}\,. \tag{11.21a}$$

Using the definition of matrix multiplication, we rewrite this further as

$$\frac{\partial}{\partial u_i}\left(\sum_{k=2}^{N-1} u_k \sum_{m=2}^{N-1} M_{km}u_m\right). \tag{11.21b}$$

In the double sum in (11.21b), we need only the terms that contain u_i, because $\partial u_j/\partial u_i = 0$ for $j \neq i$. There are two such terms: with $k = i$ and $m = i$. Then:

$$\begin{aligned}\frac{\partial}{\partial u_i}\left(\underline{u}^T M \underline{u}\right) &= 1\cdot\sum_{m=2}^{N-1} M_{im}u_m + \sum_{k=2}^{N-1} u_k M_{ki}\cdot 1\\ &= \sum_{k=2}^{N-1} M_{ik}u_k + \sum_{k=2}^{N-1} u_k M_{ki}\\ &= \sum_{k=2}^{N-1}(M_{ik}+M_{ki})u_k = (M+M^T)\underline{u}\,. \end{aligned}\tag{11.22}$$

In particular, if M is symmetric, as it is in our case, then

$$\frac{\partial}{\partial u_i}(\underline{u}^T M \underline{u}) = 2M\underline{u}\,. \tag{11.23}$$

Using this result, one transforms Eq. (11.20) into:

$$(\underline{u}-\underline{\hat{u}}) + \alpha\cdot(H^T H)\underline{u} = \underline{0}\,, \tag{11.24a}$$

or, equivalently,

$$(I + \alpha\cdot(H^T H))\underline{u} = \underline{\hat{u}}\,. \tag{11.24b}$$

Recall that α is an empirically selected (small) parameter and H is defined in Eq. (11.16). As we have discussed, the choice for neither α nor H is unique but depends on our understanding of what the solution $\underline{u}$ should look like. In practice, experimentation may be needed to find values of α and a form of H that yield a "reasonable" $\underline{u}$.

Equation (11.24) is the main result of this section. It highlights the key idea of the ***Tikhonov regularization***, which is the following. Consider a matrix equation

$$A\underline{w} = \underline{\mu}\,, \tag{11.25}$$

where, as before, $\underline{\mu}$ is the known vector of measurements and the solution $\underline{w}$ needs to be reconstructed from (11.25). (In the particular case of Eq. (11.24b), $\underline{\mu} = \underline{\hat{u}}$, $\underline{w} = \underline{u}$, and $A = I$.) If this reconstruction problem produces a jagged solution $\underline{w}$, consider also a *different* problem:

$$(A + \alpha B)\underline{w_\alpha} = \underline{\mu}\,, \tag{11.26}$$

where B is some matrix such that the solution $\underline{w_\alpha}$ of (11.26) is, in some sense, smooth. An optimal (range of) value(s) of α is determined by *two* conditions. On one hand, $\underline{w_\alpha}$ is not too different from the true solution of the original problem (11.25) (or, actually, from our guess of what that "true" solution should be). On the other hand, $\underline{w_\alpha}$ must be smooth. *In practice, satisfying both of these conditions requires that "α be small, but not too small".*

In the next Section we will discuss how this rather vague statement can be related to the eigenvectors of matrix A that "are responsible" for the jaggedness of the solution $\underline{w}$.

11.3 Eigenvectors as "carriers of the instability"

Here we will explain why the solution u of Eq. (11.24b) can be smoother than the original measurements vector $\underline{\hat{u}}$. Before we begin, however, we will slightly modify matrix H, so as to simplify the calculations and not to obscure the main idea by minor technical details. Namely, instead of

$$H = \frac{1}{h^2}\begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 & 0 \\ & & & \ddots & & & \\ 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 \end{pmatrix},$$

as given by Eq. (11.16), we will consider

$$H_{\rm mod} = \frac{1}{h^2}\begin{pmatrix} 3 & -1 & 0 & \dots & 0 & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 & 0 \\ & & & \ddots & & & \\ 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & 0 & \dots & 0 & -1 & 3 \end{pmatrix}, \tag{11.27}$$

which is different from H only by the "near-boundary" entries in the first and last rows. The reasons for this modification are as follows.

(i) Unlike H, $H_{\rm mod}$ is symmetric, and hence for the matrix product appearing in (11.24),

$$H_{\rm mod}^T H_{\rm mod} = H_{\rm mod}^2\,;$$

(ii) It will be easy to identify an eigenvector of $H_{\rm mod}$ that will play a crucial role in our analysis.

For a practically important case when the number of measurements N is large, it is also intuitively clear that modifying the behavior of $\underline{u}$ just at the endpoints will not significantly affect the "bulk" of $\underline{u}$. Moreover, we have already made an assumption (see Eq. (11.14)) and an approximation (see Eq. (11.19)) concerning the endpoints of $\underline{u}$, so another such approximation will not qualitatively affect our results.

Thus, proceeding with $H_{\rm mod}$ instead of H, we have a modified Eq. (11.24b):

$$(I + \alpha H_{\rm mod}^2)\underline{u} = \underline{\hat{u}}\,. \tag{11.28}$$

Next, any $(N-2)$-vector is an eigvenvector of the $(N-2)\times(N-2)$ identity matrix I, so in this particular case we will expand the r.h.s. of (11.28) into a linear combination of the $(N-2)$ eigenvectors of the symmetric matrix $H_{\rm mod}$:

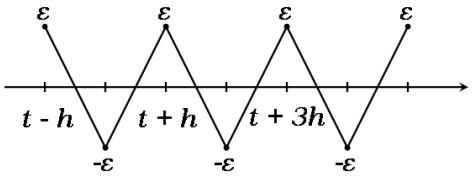
$$\underline{\hat{u}} = \sum_{j=1}^{N-2} b_j \underline{v_j}\,, \quad \text{where} \quad H_{\rm mod}\underline{v_j} = \lambda_j \underline{v_j}\,. \tag{11.29}$$

(However, in the general case of Eq. (11.26), **the expansion must be over the eigenvectors of A, not B.**) We then solve (11.28) for $\underline{u}$:

$$\begin{aligned} \underline{u} &= (I + \alpha H_{\rm mod}^2)^{-1}\underline{\hat{u}} \\ &= \sum_{j=1}^{N-2} b_j (I + \alpha H_{\rm mod}^2)^{-1}\underline{v_j} = \sum_{j=1}^{N-2} \frac{b_j}{1+\alpha\lambda_j^2}\,\underline{v_j}\,, \end{aligned} \tag{11.30}$$

where we have used the fact that the eigenvectors of any function $f(H_{\rm mod})$ are also $\underline{v_j}$, with the corresponding eigenvalues being $f(\lambda_j)$.

Let us now recall what we want to do: We want to obtain a vector $\underline{u}$ representing a *smooth* curve, given that the measured vector $\underline{\hat{u}}$ contains some experimental noise. In Sections 1 and 3 of Lecture 10, we saw that the "worst" profile of noise that causes the largest error in the numerically calculated derivative is the ripple of the form:



Note that this is the *fastest* ripple that it is experimentally possible to measure with a given discretization step h. Can one represent this worst-case ripple as a linear combination of the eigenvectors $\underline{v_j}$, so as to analyze its transformation given by Eq. (11.30)? The answer turns out to be remarkably simple: this ripple *is* the eigenvector $\underline{v_{N-2}}$ with the largest eigenvalue λ_{N-2}. We will now show that this is the case and then explain why having the term $\alpha H_{\rm mod}^2$ in Eq. (11.26) smoothens the solution $\underline{u}$.

For simplicity, let us assume that N is even. (For an odd N the calculations are similar and the conclusion is the same.) Then the aforementioned ripple is:

$$\begin{pmatrix} \text{fastest} \\ \text{ripple} \end{pmatrix} = \varepsilon \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \\ \vdots \\ 1 \\ -1 \end{pmatrix} . \tag{11.31}$$

It is then a straightforward matter to verify (do it) that

$$H_{\rm mod} \cdot \begin{pmatrix} \text{fastest} \\ \text{ripple} \end{pmatrix} = \frac{4}{h^2} \cdot \begin{pmatrix} \text{fastest} \\ \text{ripple} \end{pmatrix} , \tag{11.32}$$

which confirms our assertion about the fastest ripple being an eigenvenctor of $H_{\rm mod}$. We have not yet shown that it is the eigenvector $\underline{v_{N-2}}$ with the largest $\lambda = \lambda_{N-2}$. This follows from the fact (which in this course you have to accept on faith) that for symmetric matrices related to the operation of second derivative, the eigenvectors corresponding to larger eigenvalues have a greater number of zero crossings; note that the ripple in question has the largest possible number of such crossings. Then it follows from (11.32) that

$$\lambda_{N-2} = \frac{4}{h^2} ; \tag{11.33}$$

compare this with a similar term in Eq. (10.17) of Lecture 10. Now, substituting (11.33) into the last term in the sum in Eq. (11.30), we find that the corresponding expansion coefficient is:

$$\frac{\varepsilon}{1 + \alpha \left(\dfrac{4}{h^2} \right)^2} . \tag{11.34}$$

When $\alpha = 0$, this coefficient equals ε, i.e. this ripple in $\underline{u}$ is the same as in $\underline{\hat{u}}$ (which is not surprising given that for $\alpha = 0$, $\underline{u} = \underline{\hat{u}}$). However, when

$$\alpha \ll 1\,, \tag{11.35a}$$

but

$$\alpha \cdot \left(\frac{4}{h^2}\right)^2 \gg 1\,, \tag{11.35b}$$

the denominator in (11.34) is large, and hence the overall coefficient of the ripple term is greatly reduced. **This is the mechanism** that suppresses the experimental noise of $\underline{\hat{u}}$ and makes the solution $\underline{u}$ of (11.26) smooth. In a homework problem you will verify that a similar suppression also takes place for the eigenvector with the second-largest eigenvalue. Note that the strong inequalities in (11.35a) and (11.35b) are compatible with one another because for a small discretization step h, the number $(4/h^2) \gg 1$. Also, given Eq. (11.33), conditions (11.35a,b) can be written as

$$\frac{1}{\max(|\lambda|^2)} \ll \alpha \ll 1\,, \tag{11.35c}$$

which is a quantitative form of the statement from the end of Section 2 that "α must be small, but not too small".

Let us now summarize the main findings of this Section. First, we identified the ripple profile in the measurements $\underline{\hat{u}}$ that causes the largest error in the numerical approximation to the derivative of $\underline{\hat{u}}$. Second, we related that particular ripple with the eigenvectors of the "smoothing matrix" $H^2_{\rm mod}$. (As we have noted, in the general case, the eigvenvectors must be those of the other matrix, A in Eq. (11.26).) One can say that we looked for a "carrier of the instability" in the differentiation problem, similarly to how a doctor is looking for a carrier of the disease. Third, we showed the mechanism by which this "instability carrier" can be strongly suppressed in the solution of Eq. (11.26). Fourth and last, we determined the range, Eq. (11.35c), for the smoothing parameter α. In the next Section, we will follow similar steps to develop a regularization technique for the "ultimately unstable" problem of Lecture 10.

11.4 The Tikhonov regularization for Fredholm integral equations of the first kind

In Section 4 of Lecture 10 we showed that the matrix equation (11.25), i.e.

$$A\underline{w} = \underline{\mu}\,, \tag{11.25$_{\rm repeated}$}$$

obtained by the discretization of the integral equation

$$\int_0^1 \frac{\gamma D \cdot w(s) ds}{\left(\sqrt{D^2 + (x-s)^2}\right)^3} = \mu(x)\,, \tag{11.36}$$

had an extremely jagged solution $\underline{w}$ even when $\underline{\mu}$ was smooth. We determined that a reason for such an unstable behavior was an extremely large condition number of the matrix A in (11.25). Because of this (see Eq. (10.27) in Lecture 10), a tiny error (ripple) in the measured vector $\underline{\mu}$ results in a humongous error in the solution $\underline{w}$.

We then showed that *for symmetric matrices*, which A in this problem is,

$$\mathrm{cond}(A) = \frac{\max(|\lambda|)}{\min(|\lambda|)}\,. \tag{11.37}$$

This number can be large either because $\max(|\lambda|)$ is large or because $\min(|\lambda|)$ is small. Accordingly, either the eigenvector with the largest eigenvalue or that with the smallest eigenvalue will be the "instability carrier" and will need to be kept at bay in order to stabilize the problem i.e., to make the solution $\underline{w}$ "smooth". We have taken the word 'smooth' in quotes because it will turn out to mean a different thing than the 'smooth' in Section 2. We will elaborate on this later in this Section, and you will explore it further in a homework problem.

Thus, we first need to look at the size of the eigenvalues of A. Those eigenvalues which are either *extremely small or extremely large* are the most likely "trouble makers". Once we have these "suspects", we then verify our suspicion by looking at the shapes of the corresponding eigenvectors. Indeed, you know from a Problem in Homework 10 that the humongous error in solving the "ultimately unstable" problem (11.36) has a very jagged shape. Therefore, we expect that the "trouble making" eigenvectors will have a jagged shape, too.

With Matlab one can easily verify that for $5 \le N \le 16$, an eigenvalue λ_j of the matrix A in question has the order of magnitude $O(10^{-(N-j)})$, where N is the dimension of vectors $\underline{\mu}$ and $\underline{w}$ and the eigvenvalues are arranged as before:

$$0 < \lambda_1 < \lambda_2 < \ldots < \lambda_N \,.$$

(For $N > 16$, the smallest eigvenvalues are at or below the Matlab's accuracy of 10^{-16}.) In a homework problem, you will verify that the shape of the eigenvectors corresponding to a few smallest eigenvalues looks like a very fast ripple. Thus, it is these eigenvectors, corresponding to the *smallest eigvenvalues*, that need to be controlled in this problem[12].

To see why the most rapidly varying eigenvectors have the smallest eigenvalues in this problem, recall that in the original non-discretized form, the eigenvectors satisfy the integral equation

$$\int_0^1 \frac{\gamma D}{(\sqrt{D^2+(x-s)^2})^3} \cdot v_j(s)ds = \lambda_j v_j(x)\,, \qquad \lambda_j \ll 1\,.$$

Also recall that the integral of a fast ripple is a small number (see Section 4 of Lecture 9 and Problem 1c in HW 9). In a homework problem, you will verify that the shape of $\underline{v_1}$ is qualitatively similar to that of the worst-case ripple shown after Eq. (11.30).

We will now follow the lines of the analysis in Sections 2 and 3. However, as it will turn out, our goal here will be more modest: Rather than *suppressing* the eigenvectors with $\lambda_j \ll 1$, we will merely need *not to amplify* them "too much". The meaning for this vague criterion will be given later. So, in place of Eq. (11.25), we consider Eq. (11.26) where now we do *not* specify the form of matrix B but, *only* to simplify the calculations and not to obscure the key idea with technical details, we ask that A and B have a common set of eigenvectors with

$$A\underline{v_j} = \lambda_j \underline{v_j} \quad \text{and} \quad B\underline{v_j} = \beta_j \underline{v_j}\,. \tag{11.38}$$

For a practical implementation of regularization, such an assumption is not needed. Now, similarly to Eqs. (11.29) and (11.30), we expand $\underline{\mu}$ over the set of the N eigenvectors of A and solve Eq. (11.26) for $\underline{w}_\alpha$:

$$\underline{w}_\alpha = (A+\alpha B)^{-1}\underline{\mu} = \sum_{j=1}^{N} \frac{m_j}{\lambda_j + \alpha\beta_j}\,\underline{v_j}\,; \tag{11.39}$$

[12] Was this also the case in Sec. 11.3? You will explore this further in a homework problem.

here m_j are the expansion coefficients of $\underline{\mu}$. Note that when $\underline{\mu}$ is smooth, these expansion coefficients corresponding to fast-oscillating eigenvectors (like $\underline{v_1}$) are very small, i.e.

$$m_j \ll 1 \text{ for the first few } j \ (j \gtrsim 1)\,. \tag{11.40}$$

Indeed, if they were not small, one would have seen a fast ripple in a smooth vector $\underline{\mu}$, which would have been a contradiction of terms.

When in Eq. (11.39) $\alpha = 0$, the first few terms with $\lambda_j = O(10^{-(N-j)}) \ll 1$ are *very large* unless $m_j \ll \lambda_j$. Thus, the observed **enormous amplification of a small ripple was due to the small denominators in Eq. (11.39)**. Now, let us take α such that

$$\alpha\beta_j \gg m_j \quad \text{for the first few } j \ (j \gtrsim 1) \tag{11.41a}$$

but at the same time,

$$\alpha\beta_j \ll \lambda_j \quad \text{for the last few } j \ (j \lesssim N)\,. \tag{11.41b}$$

Combining the last two inequalities, we obtain a counterpart of condition (11.35c) for the Tikhonov regularization of Fredholm integral equations of the first kind:

$$m_1 \ll \alpha\beta_1 \quad \text{and} \quad \alpha\beta_N \ll \lambda_N\,. \tag{11.41c}$$

Note that a small ripple in $\underline{\mu}$ is still likely to be amplified in the reconstructed vector $\underline{w}$, because $\alpha\beta_j$ may be less than 1. However, thanks to (11.41a), it will *not* be amplified *as much* so as to overshadow the smooth solution. This is the key point of the Tikhonov regularization for Fredholm integral equations of the first kind, like Eq. (11.36).

To conclude this Lecture, we comment on a couple of possible choices for matrix B. The simplest choice is, surprisingly, $B = I$, the identity matrix. In this case, Eq. (11.39) becomes:

$$\underline{w}_\alpha = (A + \alpha I)^{-1}\underline{\mu} = \sum_{j=1}^{N} \frac{m_j}{\lambda_j + \alpha}\,\underline{v_j}\,. \tag{11.42}$$

(Note that $B = I$ would *not* work for the problem considered in Sections 2 and 3. Why?) For this simplest choice, we also can give an explanation of the regularization technique in terms of the condition number of matrix $(A + \alpha I)$. First, we have:

$$\text{cond}(A + \alpha I) = \frac{\lambda_{\max} + \alpha}{\lambda_{\min} + \alpha} < \frac{\lambda_{\max} + \alpha}{\alpha} \ll \frac{\lambda_{\max}}{\lambda_{\min}}\,,$$

provided that $\alpha \gg \lambda_{\min}$. Under this condition,

$$\text{cond}(A + \alpha I) \ll \text{cond}(A)\,. \tag{11.43}$$

Recall that in Section 5 of Lecture 10, we showed (see (10.27)) that the relative error of the solution $\underline{x}$ of $M\underline{x} = \underline{b}$ is proportional to cond(M). Therefore, the vector $\underline{w}_\alpha$ of (11.26) (with $B = I$), obtained by inverting the matrix $(A+\alpha I)$, is much less sensitive to small, but in practice unavoidable, errors in $\underline{\mu}$ than the vector $\underline{w}$ of (11.25), obtained by inverting the matrix A. Also, as long as $\alpha \ll \lambda_{\max}$, then matrix $(A+\alpha I)$ is not "too far" from the original matrix A. Then the expansion coefficients $m_j/(\lambda_j + \alpha)$ of vector $\underline{w}_\alpha$ corresponding to the eigenvectors with the larger λ_j are not "too different" from the respective expansion coefficients m_j/λ_j of the "original" vector $\underline{w}$. Thus, to summarize this discussion of the case $B = I$, we can state that a condition

$$\lambda_{\min} \ll \alpha \ll \lambda_{\max} \tag{11.44}$$

may be considered as a *weaker counterpart* of condition (11.41c) when $B = I$. The part of (11.44) that is weaker than the corresponding part of (11.41c) is the left-hand inequality. Indeed, the left-hand strong inequality in (11.41c) guarantees that the amplitude of the fast ripple in $\underline{w}_\alpha$ will be small (why?). In contrast, the left-hand strong inequality in (11.44) only guarantees that the fast ripple in $\underline{w}_\alpha$ will be much smaller than that in $\underline{w}$, although not necessarily small (again, why?).

The regularization (i.e. the smoothing of the ripple in the solution $\underline{w}$) can be made more efficient if the "smoothing matrix" B amplifies the first few eigenvectors, like $\underline{v_1}, \underline{v_2}, \ldots$, more than the remaining ones. This will make

$$\mathrm{cond}(A + \alpha B) < \mathrm{cond}(A + \alpha I) \tag{11.45}$$

and will further reduce the instability of the original problem (11.25). For example, if in (11.38)

$$\beta_1 > 1 > \beta_N \,, \tag{11.46}$$

then

$$\mathrm{cond}(A + \alpha B) = \frac{\lambda_N + \alpha\beta_N}{\lambda_1 + \alpha\beta_1} < \frac{\lambda_N + \alpha}{\lambda_1 + \alpha} = \mathrm{cond}(A + \alpha I) \,,$$

which is precisely the inequality (11.45). You will verify this in a homework problem by taking $B = H_{\mathrm{mod}}$ from Eq. (11.27). Even though this B does not have the same eigenvectors as the matrix A obtained by the discretization of (11.36), the results of the above analysis will still hold qualitatively.