Numerical Astronomy 1 – Part 2
Integral equations, inverse problems, and the loss of information

Norman Gray

9 November 1998*

1 Integral equations

You are familiar with *algebraic equations*, such as

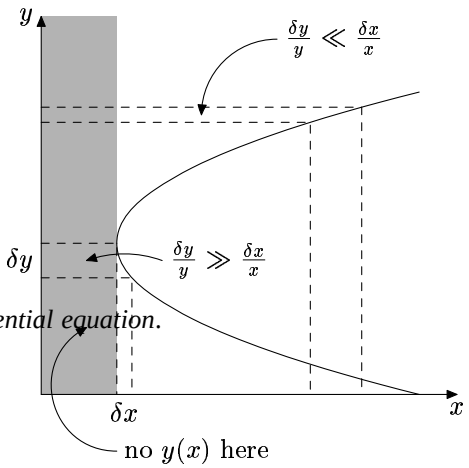
$$e^x = \sin y$$

or, more generally, $f(x) = g(y)$, which has the formal solution $x = f^{-1}[g(y)]$. This solution exists only for $x < 0$ and $y > 0$.

However, the solution might not *exist*, be *unique*, or be *stable* for all values of x and y. Consider limits on solution of $y = \arcsin \mathrm{e}^x$. Note

$$\frac{\delta y}{y} = \frac{\delta x}{x} \frac{x\mathrm{e}^x}{\sqrt{1-\mathrm{e}^{2x}}\arcsin \mathrm{e}^x},$$

illustrating that the solutions $y(x)$ or $x(y)$ can be very unstable.



An algebraic equation relates values of x and y. A *functional* equation relates *functions* of x and y:

$$g(x) = \Theta\{f(y); x\}. \qquad (2.1)$$

If the operator Θ is a differential operator, this is a (familiar!) *differential equation*. For example, if

$$\Theta = \left\{ \frac{\mathrm{d}}{\mathrm{d}y} + y^2 \frac{\mathrm{d}^2}{\mathrm{d}y^2} \right\}_{y=x},$$

then Eqn. (2.1) corresponds to

$$g(x) = \frac{\mathrm{d}f}{\mathrm{d}x} + x^2 \frac{\mathrm{d}^2 f}{\mathrm{d}x^2}.$$

In simple cases, we can solve such equations directly:

$$\frac{\mathrm{d}f}{\mathrm{d}x} = g(x) \qquad \text{implies} \qquad f = \int g(x)\,\mathrm{d}x.$$

In this course, we are interested in *integral equations*. We write

$$g(y) = \mathcal{K}\{u(x); \alpha, y\} = \alpha(x)u(x) + \lambda \int_{a(x)}^{b(x)} K(x,y)u(y)\,\mathrm{d}y, \qquad (2.2)$$

The formal solution to this *inverse problem* can be written

$$u(y) = \mathcal{K}^{-1}\{g(x); y\}. \qquad (2.3)$$

* Updated from notes of 23 October 1998

If the limits a and b are in fact constants, this is a *Fredholm equation*, if they are not, it is a *Volterra equation*. In either case, if the function $\alpha(x) \equiv 0$ the IE is termed *first kind*, if $\alpha(x) \equiv 1$ it is *second kind*.

Familiar examples of integral transforms are the laplace transform

$$\mathcal{L}\{u(y);x\} \longleftrightarrow K(x,y) = \mathrm{e}^{-xy} \qquad (a=0, b=\infty), \tag{2.4}$$

and the fourier transform

$$\mathcal{F}\{u(y);x\} \longleftrightarrow K(x,y) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{ixy} \qquad (a=-\infty, b=\infty). \tag{2.5}$$

2 Examples

2.1 Electron spectra from bremsstrahlung photon spectra

Observe an electron flux spectrum $\mathrm{d}F/\mathrm{d}E$ electrons $\mathrm{m}^{-2}\,\mathrm{s}^{-1}$ per unit electron energy E, in a plasma of ion density n. Rate of bremsstrahlung emission is $\mathrm{d}j/\mathrm{d}\epsilon$ photons s^{-1} per unit photon energy ϵ.

$$\frac{\mathrm{d}j}{\mathrm{d}\epsilon}(\epsilon) = n\int_\epsilon^\infty \frac{\mathrm{d}F}{\mathrm{d}E}(E)\frac{\mathrm{d}Q}{\mathrm{d}\epsilon}(E,\epsilon)\,\mathrm{d}E \tag{2.6}$$

where $\mathrm{d}Q/\mathrm{d}\epsilon$ is the bremsstrahlung cross-section, differential in ϵ.

We can approximate $\mathrm{d}Q/\mathrm{d}\epsilon = Q_0 mc^2/\epsilon E$, so that Eqn. (2.6) becomes

$$\int_\epsilon^\infty \frac{1}{E}\frac{\mathrm{d}F}{\mathrm{d}E}\,\mathrm{d}E = \frac{1}{nQ_0mc^2}\,\epsilon\frac{\mathrm{d}j}{\mathrm{d}\epsilon}, \tag{2.7}$$

which we can differentiate to get

$$\frac{\mathrm{d}F}{\mathrm{d}E} = \frac{E}{nQ_0mc^2}\left[-\frac{\mathrm{d}}{\mathrm{d}\epsilon}\left(\epsilon\frac{\mathrm{d}j}{\mathrm{d}\epsilon}\right)\right]_{\epsilon=E}. \tag{2.8}$$

2.2 Instrument convolutions – spectrometer

We wish to measure a photon spectrum $\mathrm{d}F/\mathrm{d}\lambda$; what we actually obtain from a specrometer measurement is a function $G(l)\mathrm{d}l$ of position l on a plate or CCD image. We separately know the spectrometer's *response function* $R(l-\lambda)$. Thus

$$\frac{\mathrm{d}G}{\mathrm{d}l}(l) = \int_{-\infty}^\infty \frac{\mathrm{d}F}{\mathrm{d}\lambda}(\lambda)R(l-\lambda)\,\mathrm{d}\lambda. \tag{2.9}$$

This is a Fredholm equation (of the first kind).

There are numerous other examples, many much more elaborate, in Craig and Brown.

3 Error amplification and ill-posedness

In Eqn. (1.3), we can write $y = r$, and $f(y) = 2yn(y)/\sqrt{y^2-x^2}$, so that

$$N(s) = \int_s^\infty f(y)\,\mathrm{d}y. \tag{2.10}$$

Analytically, a problem of the form

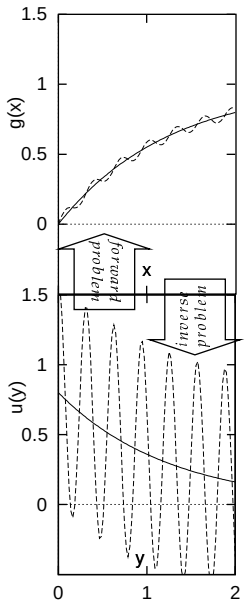
$$g(x) = \int_0^x u(y)\,\mathrm{d}y \tag{2.11}$$

has the immediate solution

$$u(y) = \left(\frac{\mathrm{d}g}{\mathrm{d}x}\right)_{x=y}, \tag{2.12}$$

but things are not as simple as this.

The functions $g_a(x) = 1 - \exp -\alpha x$ and $g_b(x) = 1 - \exp -\alpha x + \beta \sin \omega x$ are close together in function space because the superimposed 'ripple' has a small amplitude. However, their derivatives $u_a(y) = (\mathrm{d}g/\mathrm{d}x)_{x=y}$ and $u_b(y)$ are very far apart, and *become further apart as the frequency of the ripple, ω, increases*. A very small perturbation (ie, measurement or random error) on the data, $g(x)$, can result in a very large perturbation on the recovery, $u(y)$. A data function with many fourier components (such as a saw-tooth, with infinitely many) can produce arbitrarily large components in $u(y)$ even for arbitrarily small β. We will come back to this later.



Another way of seeing this instability (which will, again, be of importance later) is to consider solving Eqn. (2.11) numerically. Using a very simple quadrature for the integral, with $\Delta x = 2h$, $x = 2hi, (i = 1, 2, \ldots)$, we have

$$g(2hi) = g(i\Delta x) = \sum_{j=1}^{i} u[(2j-1)h]\Delta x \tag{2.13}$$

Setting $g_i = g(i\Delta x)$ and $u_j = u[(2j-1)h]$, this can be represented as

$$\mathbf{g} = \Delta x \mathsf{K}\mathbf{u},$$

or

$$g_i = \Delta x K_{ij} u_j,$$

where K is the lower-triangular matrix

$$K_{ij} = \begin{cases} 1 & \text{if } j \le i \\ 0 & \text{otherwise} \end{cases}.$$

This can be solved simply by 'forward substitution':

$$g_1 = \Delta x u_1, \qquad g_2 = \Delta x(u_1 + u_2) \Rightarrow u_2 = (g_2 - g_1)/\Delta x.$$

so that the general term is

$$u_j = \frac{g_j - g_{j-1}}{\Delta x} = \frac{g[j\Delta x] - g[(j-1)\Delta x]}{\Delta x}. \tag{2.14}$$

The right hand side is a finite-difference approximation for the derivative g'. This is correct to $O(h^2)$ (the quadrature in Eqn. (2.13) is obtained by truncating a Taylor series at the term in h^2), so this expression would appear to improve as $h \to 0$. However, the data samples g_i will have measurement errors, so that as the discretisation parameter $h \to 0$, the difference $g_i - g_{i-1}$ will become smaller, and dominated by the random variations in g_i. Even in the absence of this, Eqn. (2.14) would be vulnerable to numerical roundoff error, so that this inversion method is not particularly well-conditioned. We will see this effect in more detail when we study condition numbers in the next part.

Just as there are much more sophisticated integration techniques, there are much more sophisticated techniques for addressing inverse problems However, the same problems of instability must be addressed.

4 Stability and the Riemann-Lebesgue lemma

Recall from Sect. 1 the distinct forms of the Fredholm equations of the first and second kind:

$$g(x) = \int_a^b K(x,y)u(y)\,\mathrm{d}y \qquad g(x) = u(x) + \lambda \int_a^b K(x,y)u(y)\,\mathrm{d}y. \tag{2.15}$$

Volterra equations have the upper limit $b(x) = x$. Equations of the second kind are easier to solve since, if λ is small, $u_0(x) = g(x)$ is an approximate solution. From this, it is an easy step to an iterative solution, of the form

$$u_{n+1}(x) = g(x) - \lambda \int_a^b K(x,y)u_n(y)\,\mathrm{d}y \qquad , n = 0, 1, 2, \ldots. \tag{2.16}$$

It is a theorem that this process will converge if $\lambda(b-a)\max_{x,y}|K(x,y)| < 1$.

Volterra equations are sufficiently different that this iteration has better properties: the iteration converges for all values of λ. Also, differentiating a Volterra equation of the *first* kind leads to

$$\frac{\mathrm{d}g}{\mathrm{d}x} = K(x,x)u(x) + \int_a^x \frac{\partial K}{\partial x}(x,y)u(y)\,\mathrm{d}y,$$

which exists *if* $K(x,y)$ is non-singular, and which is a (typically well-behaved) Volterra equation of the *second* kind. This can be an improvement *despite* the potential for instability due to a (numerical) differentiation of the data function $g(x)$.

The instability problems of the Fredholm equation are illustrated by the Riemann-Lebesgue Lemma.

4.1 The Riemann-Lebesgue lemma

If

$$\int_a^b |K(x,y)|\,\mathrm{d}y \quad \text{is bounded, } \forall x, \tag{2.17}$$

then

$$\int_a^b K(x,y)\mathrm{e}^{-imy}\,\mathrm{d}y \to 0, \qquad \text{as } m \to \infty. \tag{2.18}$$

That is, for an integrable kernel, high-frequency components in the underlying function (exemplified by e^{-imy}) make small-amplitude contributions to the data. Which is to say that data with a given measurement error is consistent with a recovered function $u(x)$ with arbitrarily many components above a certain frequency.

Note that the fourier-transform kernel ($K(x,y) = \exp ixy$, $(a,b) = (-\infty,\infty)$) is not integrable, so the lemma does not apply.

As an example of the Riemann-Lebesgue Lemma in action, consider the solution of the convolution $g(x) = \int K(x-y)u(y)\,\mathrm{d}y$. We can solve this using fourier transforms. Transform both sides of the equation to get

$$\begin{aligned}
\mathcal{F}[g(x)] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \mathrm{e}^{-i\omega x}g(x)\,\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}}\int_{x=-\infty}^{\infty} \mathrm{d}x\mathrm{e}^{-i\omega x}\int_{y=-\infty}^{\infty} K(x-y)u(y)\,\mathrm{d}y \\
&= \frac{1}{\sqrt{2\pi}}\int_{\xi=-\infty}^{\infty} \mathrm{d}\xi\mathrm{e}^{-i\omega\xi}K(\xi)\int_{-\infty}^{\infty} \mathrm{d}y u(y)\mathrm{e}^{-i\omega y}
\end{aligned}$$

substituting $\xi = x - y$. Thus

$$\overline{g}(\omega) = \sqrt{2\pi}\overline{K}(\omega)\overline{u}(\omega)$$

so that

$$u(y) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{\overline{g}(\omega)}{\overline{K}(\omega)}\mathrm{e}^{i\omega y}\,\mathrm{d}y. \tag{2.19}$$

Consider adding a noise function $\delta g(x)$ to the data, which will result in an error $\delta u(x)$ in the recovered solution. Since the fourier transform is linear, the transform applied to the total data function $g(x) + \delta g(x)$ results in

$$\overline{\delta u}(\omega) = \frac{1}{\sqrt{2\pi}}\frac{\overline{\delta g}(\omega)}{\overline{K}(\omega)}. \tag{2.20}$$

Now, if the kernel is a physically realistic one, then the Riemann-Lebesgue lemma will be satisfied:

$$|\overline{K}(\omega)| \to 0 \qquad \text{as } |\omega| \to \infty.$$

As a result, the recovery error $\overline{\delta u}$ will be bounded only if

$$|\overline{\delta g}(\omega)| \to 0 \qquad \text{as } |\omega| \to \infty,$$

and does so sufficiently quickly. This will not be true in practice. As a result, δu will not be bounded in practice, unless we either filter out high frequency noise in the data (forcing $|\overline{\delta g}| \to 0$), or else we force the recovered solution $u(x)$ to be smooth.

4.2 The null space and nearly-singular opearators

We can write Eqn. (2.15) as

$$\mathcal{K}u = g \tag{2.21}$$

and

$$(\mathcal{I} - \lambda\mathcal{K})u = g. \tag{2.22}$$

If we discretise these as in Eqn. (2.13), then we can consider the behaviour of the corresponding linear systems. If λ is small, then Eqn. (2.22) is diagonally dominant (close to the unit matrix) and can be inverted stably. When λ becomes large, however, and if the matrix K is strongly smoothing, then the rows of K can become approximately linearly dependent. If any of the rows of a matrix K become precisely linearly dependent, then (i) the matrix becomes singular (non-invertible), and will have some zero eigenvalues, and (ii) the matrix will map a subspace of its domain to zero.

The extent to which (i) becomes true is measured by the condition number; the effect of (ii) is expressed in the notion of the *null space*. This is

$$Z_{\mathcal{K}} = \{\phi : \mathcal{K}\phi = 0\}, \tag{2.23}$$

with the property that

$$\mathcal{K}u = g \quad \text{implies} \quad \mathcal{K}(u + \alpha\phi) = g \qquad , \phi \in Z_{\mathcal{K}}, \alpha \in R. \tag{2.24}$$

This obviously leads to non-uniqueness in the solution $u(x)$ for the inverse problem. In general, we'll assume below that operators are non-singular (so that solutions are at least formally unique), but the point of the study of inverse problems is that we wish to deal with operators which are *nearly* singular.

Another way of talking about the null space is to note that there will be zero eigenvalues, $\lambda_u = 0$ (remember that u is an eigenfunction with eigenvalue λ_u if $\exists\lambda_u : \mathcal{K}u = \lambda_u u$). That is, $\lambda_\phi = 0, \forall\phi \in Z_K$.

'Nearly singular' means that there is a large volume of $U = \mathrm{domain}(\mathcal{K})$ which maps to a small volume of $G = \mathrm{range}(\mathcal{K})$, which means that tiny errors in your measured position in that small volume will turn into huge errors in the inverted position in U. If $\mathcal{K}$ is nearly singular, then λ_u will be 'small' for u in the 'almost-null-space' of $\mathcal{K}$.

For example, consider the simple (separable) kernel $K(x, y) = \cos x \cos y$ over a symmetric interval, so that

$$g(x) = \cos x \int_{-a}^{a} \cos y u(y)\,\mathrm{d}y.$$

This will have solutions only if $g(x) \propto \cos x$, but if there are any solutions, there will be an infinite number, since to any solution $u(y)$ we can add an arbitrary antisymmetric function without changing the right-hand-side.