# An Introduction to Inverse Problems

Kurt Bryan

March 28, 2011

## 1 Introduction

Consider a pendulum of length $L$ in a gravitational field with acceleration $g > 0$; the situation is depicted in Figure 1. Suppose the pendulum is

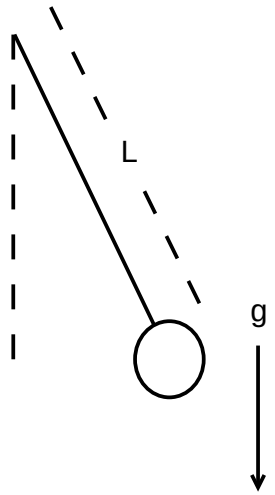


Figure 1: Pendulum.

displaced from its vertical rest position by a small amount and allowed to swing freely. A standard analysis done in most freshman physics courses shows that the period $P$ of the pendulum is given approximately by

$$P = 2\pi\sqrt{L/g} \tag{1}$$

assuming friction is negligible over the course of the experiment. A typical routine homework problem would be to compute the period $P$ for a given

pendulum of length $L$ and a fixed gravitational acceleration $g$. This typifies what would usually be called the *forward problem*: We are given a physical system (the pendulum) with certain physical parameters ($g$ and $L$) specified and the asked about some aspect of how the system behaves, in this case, the period of the motion if the pendulum is displaced a small amount.

But in real life the situation is often reversed: we may not know all the parameters that govern the system's physics, and they may not be easy to measure directly. They frequently have to be inferred from observation of the system. As an example, we may suppose that we don't know $g$, gravitational acceleration, but we do have a real pendulum at hand with a known length $L$, and we have the means to swing the pendulum and measure its period. Can we deduce $g$ from this information?

## 1.1 An Experiment

Your group has been given a pendulum; don't break it. Instead,

1. Use the yardstick to measure the length $L$ of the pendulum, in meters.
2. Use your stopwatch to measure the period of the pendulum, starting with a small displacement. You'll want to hold the string steady (maybe brace your hand on a table or something) and have someone time ten back-and-forth swings of the pendulum.
3. Compute $P$, the period, in seconds.

Now, figure out how to use equation (1) to estimate $g$, given that you have measurements of $L$ and $P$.

## 1.2 A General Framework for Inverse Problems

Inverse problems revolve around the estimation of unknown parameters that govern a physical system's behavior. The procedure above is fairly typical.

1. We start by developing a mathematical model of the system, treating the various parameters as unspecified constants. For example, for the pendulum we do the physics that leads to equation (1). The resulting mathematical model is called the *forward problem*.
2. If any parameters are easy to measure directly, we generally do that. For example, you measured $L$ for your pendulum.

3. We "stimulate" the real physical system of interest, then observe some aspect of its response. For the pendulum this means giving it a small displacement and then measuring the period of the motion.

4. From the data collected, we estimate the unknown parameter(s) in the model.

This framework makes it look pretty cut-and-dried, but in fact some real challenges can pop up, and they come in a variety of flavors.

## 1.3 Issues

The central questions of interest in any inverse problems are these:

- **Existence:** Given observed data for the system, is there some value for the unknown parameter(s) that actually yields the observed data? If not, the inverse problem has no solution. This is often the case for more complicated problems in which the data has noise or other error, but we can deal with this—more on that later. But in the case of the pendulum, for any observed (positive) period it's easy to see that we can find a suitable $g$ that makes equation (1) work.

- **Uniqueness:** Can the unknown parameter(s) in principle be uniquely determined from the measured data? Or could two different sets of values for the unknown parameters give rise to the same observations? For example, from equation (1) it's easy to see that if $P$ and $L$ are known then $g$ is determined uniquely. This isn't quite the same as asking for a formula for $g$. Silly as it sounds, you can see that $g$ can be determined without actually solving for it (see "Reconstruction" below, though).

- **Stability:** If the measured data contains small errors, will the error in the resulting estimates of the unknowns be correspondingly small? Of could small measurement errors lead to huge errors in our estimates? For the pendulum it's easy to see that, on a percentage basis, small errors in the measurements of $L$ or $P$ lead to comparable errors in the estimate of $g$ using equation (1); see Problem 1 below.

- **Reconstruction:** Is there a computationally efficient formula or procedure for actually finding the unknown parameter(s) from the data?

For example, for the pendulum we can explicitly solve equation (1) to find

$$g = \frac{4\pi^2 L}{P^2}. \tag{2}$$

An inverse problem for which existence, uniqueness, and stability hold is said to be *well-posed*. The alternative is an *ill-posed* inverse problem. Ill-posed problems tend to be the most common, and interesting!

**Problem 1** Let the true values for $g, L$, and $P$ be denoted $g_0, L_0$, and $P_0$, respectively (so from equation (2) $g_0 = 4\pi^2 L_0/P_0^2$, exactly). Suppose that in measuring $L$ and $P$ we make small errors $\Delta L$ and $\Delta P$; let $\Delta g$ be the resulting error in our estimate of $g$ from equation (2).

a. Use equation (2) to justify the approximation

$$\Delta g \approx \frac{4\pi^2 \Delta L}{P_0^2} - \frac{8\pi^2 L_0 \Delta P}{P_0^3}. \tag{3}$$

(Hint: Calculus 3).

b. Use equation (3) to justify the approximation

$$\frac{\Delta g}{g_0} \approx \frac{\Delta L}{L_0} - 2\frac{\Delta P}{P_0}. \tag{4}$$

How does this support the statement that "...on a percentage basis, small errors in the measurements of $L$ or $P$ lead to comparable errors in the estimate of $g$..."?

c. How accurate do you think your measurement of $L$ was? How accurate do you think your measurement of $P$ was? Use equation (3) and/or (4) to bound the (worst case) error in your estimate of $g$.

**Problem 2** Show that if both $g$ AND $L$ are unknown (maybe we have no tape measure) then we can't recover both from just a measurement of $P$.

## 1.4 Something More "Interesting"

### 1.4.1 The Forward Problem

Suppose you put $P_0$ dollars into some type of investment account. The account pays a variable annual interest rate $r(t)$, compounded continuously. How much money will you have at any given time? To answer this, let $P(t)$ denote the amount of money in the account at time $t$. If we earn money with an instantaneous interest rate $r(t)$ then (practically by definition) we have

$$P'(t) = r(t)P(t). \tag{5}$$

This DE is easy to solve; separation of variables yields

$$P(t) = P_0 e^{R(t)}$$

where

$$R(t) = \int_0^t r(s)\, ds.$$

Let's consider the above as the forward problem: given $P_0$ and $r(t)$ we are to determine $P(t)$. This means solving a differential equation.

### 1.4.2 The Inverse Problem

Now consider the corresponding inverse problem: If we have the function $P(t)$, can we determine $r(t)$? Equation (5) makes it clear that the answer is "yes", and indeed, provides the explicit reconstruction formula

$$r(t) = P'(t)/P(t). \tag{6}$$

This is an inverse problem for which uniqueness holds (existence too, if $P$ is differentiable and never zero). However, unlike in the pendulum problem, the unknown we seek here is a function $r(t)$, rather than a scalar like gravitational acceleration $g$. Inverse problems like this in which the unknown is a function are frequently ill-posed.

And this problem certainly fits into the ill-posed category. First note that we don't actually know the function $P(t)$ for all $t$, but rather we have measurements of $P$ at discrete time intervals. For example, we may have measurements of $P(t)$ only every quarter (3 months). Moreover, we don't have an exact value for $P$, but rather $P$ round to the nearest cent. How do we estimate $r(t)$ in this case, and does this affect the accuracy of our estimates?

In reality, we won't try to recover $r(t)$ for every $t$, but rather estimates of $r$ at select times. One way to estimate $r(t)$ at a time $t_0$ is to use a finite difference approximation to the derivative, say

$$P'(t_0) \approx \frac{P(t_0 + \Delta t) - P(t_0)}{\Delta t} \tag{7}$$

where $\Delta t$ is the time interval from one sample of $P$ to the next, e.g., $\Delta t = 0.25$ if we have quarterly data and measure time in years. Of course many other finite difference approximations to $P'(t)$ could be used, but this will do for the moment.

If we use the approximation (7) in equation (6) we find that $r(t)$ can be approximated as

$$r(t_0) \approx \frac{P(t_0 + \Delta t) - P(t_0)}{P(t_0)\Delta t}. \tag{8}$$

**Example 1** Suppose we start with $P(0) = 100$ dollars, and the interest rate $r(t)$ is given by

$$r(t) = 0.04(3 - 2\cos(2t) + t/3).$$

A graph of $r(t)$ is shown in Figure 2 below. We take measurements of $P(t)$ annually over the course of 5 years, as tabulated below. The data is rounded to the nearest cent.

| $t_0$ | $P(t_0)$ |
|---|---|
| 0.0 | 100 |
| 1.0 | 109.45 |
| 2.0 | 134.57 |
| 3.0 | 153.91 |
| 4.0 | 172.82 |
| 5.0 | 219.99 |

If we use equation (8) to estimate $r(t_0)$ at $t_0 = 0, 1, 2, 3, 4$ we obtain the graph on the left in Figure 2 (the solid red curve); the dashed (blue) line is the actual graph of $r(t)$.

The estimated $r$ seems to "lag" the true $r$, which makes sense. Intuitively, the quantity on the right in (7) should be a better estimate of $P'(t)$ at the point $t = t_0 + \Delta t/2$ rather than $t = t_0$, and hence the quantity on the right in (8) should be a better estimate of $r(t_0 + \Delta t/2)$. Thus to remedy the lag, we can replace equation (8) with

$$r(t_0 + \Delta t/2) \approx \frac{P(t_0 + \Delta t) - P(t_0)}{P(t_0)\Delta t}. \tag{9}$$

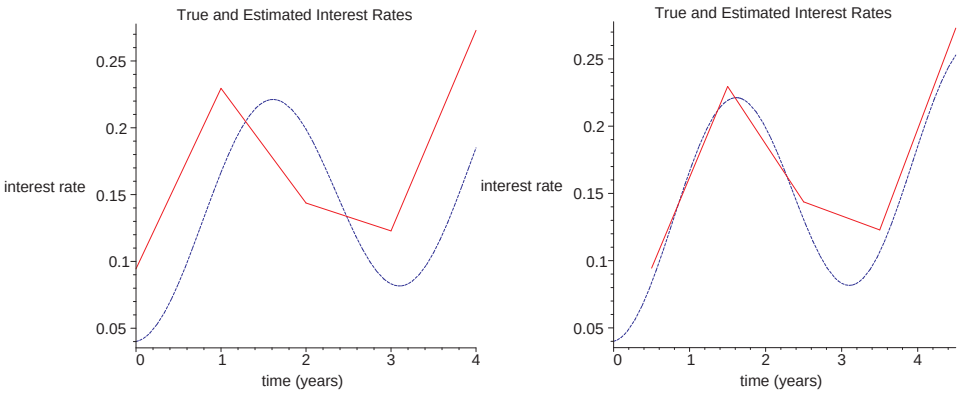


Figure 2: Estimated (solid,red) and true (dashed,blue) interest rate, for one-sided (left) and central (right) differences.

The result of applying equation (9) to the data is shown in the graph on the right in Figure 2. It looks a bit more accurate.

Of course, if we want to track variations in $r(t)$ more accurately, it stands to reason that we should obtain balance data more often. The table below is $P(t)$ sampled at quarterly intervals $t_0 = k/4$ for $0 \le k \le 20$, again rounded to the nearest cent:

| $t$ | $P(t)$ | $t$ | $P(t)$ | $t$ | $P(t)$ |
|---|---|---|---|---|---|
| 0.0 | 100 | 1.75 | 127.69 | 3.5 | 160.86 |
| 0.25 | 101.13 | 2.0 | 134.57 | 3.75 | 165.90 |
| 0.5 | 102.84 | 2.25 | 140.90 | 4.0 | 172.82 |
| 0.75 | 105.53 | 2.5 | 146.23 | 4.25 | 181.93 |
| 1.0 | 109.45 | 2.75 | 150.48 | 4.5 | 193.19 |
| 1.25 | 114.62 | 3.0 | 153.91 | 4.75 | 206.15 |
| 1.5 | 120.85 | 3.25 | 157.12 | 5.0 | 219.99 |

We use equation (9) to estimate $r(t_0 + \Delta t/2)$ (that is, at points of the form $t_0 = 0.125, 0.375, \ldots, 4.875$.) The result is shown in Figure 3 below; the dashed blue line is the actual plot of $r(t)$. More frequent sampling of $P(t)$ clearly yields better results.

It makes sense that tracking $P(t)$ even more frequently will result in still better estimates. Suppose we're willing to bother our broker every day, sampling $P(t)$ at time $t = k/365$, $k = 0, 1, 2, \ldots$ (ignore the fact that the market isn't open every day!) Estimating $r(t)$ with the procedure above produces the

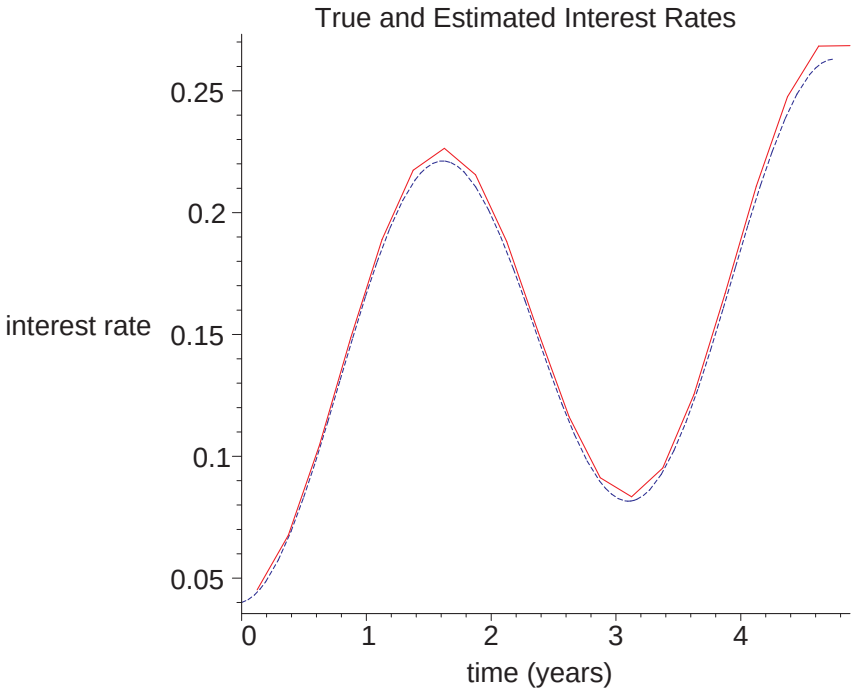


Figure 3: Estimated (solid) and true (dashed) interest rate.

graph on the left in Figure 4. If we could somehow obtain $P(t)$ ten times per day the result is even worse, as illustrated on the right in Figure 4. Clearly something has gone drastically wrong!

**Problem 3** Let $r(t) = 0.08 + 0.01t(7 - t)$ on the interval $0 \leq t \leq 5$, with $P(0) = 100$.

a. Compute $P(t)$.

b. Evaluate $P(t)$ annually, at each time $t = 0, 1, 2, 3, 4, 5$ (so $\Delta t = 1.0$). Round your answer to the nearest dollar (not cent).

c. Use equation (9) to estimate $r(t)$ at times $t = 0.5, 1.5, 2.5, 3.5, 4.5$, and compare to the true values of $r$.

d. Repeat (b) and (c) using $\Delta t = 0.25$ and $\Delta t = 0.1$; you'll want to use Excel or other software. Plot your estimates and the true values of $r$.

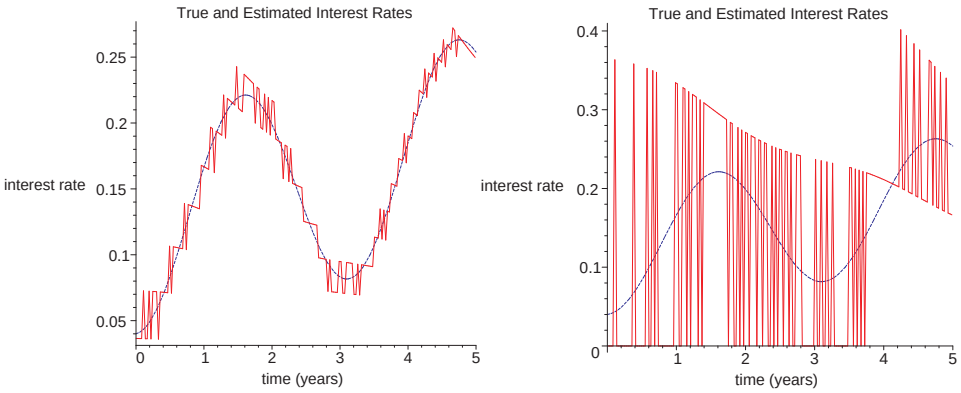


Figure 4: Estimated (solid) and true (dashed) interest rate, one sample per day (left), ten samples per day (right).

## 1.5 Ill-Posedness

The problem of estimating $r(t)$ is an example of an ill-posed inverse problem. Although the formulas (8) or (9) look good on paper (especially as $\Delta t$ gets smaller), the procedure fails if the data is noisy or contains other types of errors. In this case, the error stems simply from the process of rounding the balance $P(t)$ to the nearest cent!

To see what's going wrong in the example above, note that when we use equation (7) to estimate $P'(t_0)$, we don't really have the exact value of $P(t_0)$ or $P(t_0 + \Delta t)$, but rather rounded (and so slightly erroneous) values. Specifically, what we have instead of $P(t_0)$ is $P(t_0) + \epsilon_1$ where $-0.005 \leq \epsilon_1 \leq 0.005$ depending on which way the rounding went. Similarly we have $P(t_0 + \Delta t) + \epsilon_2$ with $-0.005 \leq \epsilon_2 \leq 0.005$. As a result, when we use equation (7) (or its successors, (8)/(9)) what we're really doing is estimating

$$P'(t_0) \approx \frac{P(t_0 + \Delta t) + \epsilon_2 - (P(t_0) + \epsilon_1)}{\Delta t}.$$

A little algebra yields

$$P'(t_0) \approx \frac{P(t_0 + \Delta t) - P(t_0)}{\Delta t} + \frac{\epsilon_2 - \epsilon_1}{\Delta t}. \tag{10}$$

When $\Delta t$ is small the first term on the right in (10) should be a good approximation to $P'(t_0)$. The second term, however, is a disaster. We might hope that $\epsilon_1$ and $\epsilon_2$ cancel, but that's unlikely—these rounding errors might just as

easily reenforce each other! All we can really be sure of is that $|\epsilon_2 - \epsilon_1| \leq 0.01$. Thus when $\Delta t$ is small the term on the right actually gets larger if $\epsilon_1 \neq \epsilon_2$. Of course the inaccuracy in estimating $P'(t_0)$ carries over into the estimate of $r(t_0)$ via (6).

To illustrate, in the example above we know that $P'(t_0)$ should be in the ballpark of $(0.1)(100) = 10$ (based on a typical 0.1 interest rate and 100 dollar balance). In the worst case rounding error we might have $|\epsilon_2 - \epsilon_1| = 0.01$. Based on these observations and equation (10) we expect the rounding error to be significant—say, 10 percent of the size of $r'(t_0)$ if $\Delta t$ is such that $(0.1)(10) \approx 0.01/\Delta t$. This yields $\Delta t \approx 0.01$. This is in total accordance with the graph on the left in Figure 4.

**Problem 4** Redo the computations in the paragraph above but under the assumption that we round to the nearest dollar.

**Problem 5** If $\Delta t$ is REALLY small (but still positive) then $P(t_0)$ and $P(t_0 + \Delta t)$ should be very close, within one, two, or a few cents of each other. Use this observation and equation (7) to argue that our estimate of $P'(t_0)$ will be either 0, $\pm 0.01/\Delta t$, $\pm 0.02/\Delta t$ (or maybe larger), depending on whether $P(t_0)$ and $P(t_0 + \Delta)$ round "together" or in opposite directions. Based on this, what will be our estimate of $r(t_0)$? Reconcile your conclusion with the graph on the right in Figure 4.

**Problem 6** Suppose a particle moves along the $x$-axis. The position of the particle at any time is $x = f(t)$ for some position function $f$. We are able to measure the position of the particle at discrete times $t = t_0, t_1, t_2, \ldots$ (so $t_k = k\Delta t$). However, our position measurements are subject a "random" error in the range $-\epsilon$ to $\epsilon$ for some positive number $\epsilon$. Our goal is to use this information to estimate the velocity of the particle at each time $t_k$.

a. Let $f_k$ denote our noisy measurement of the position of the particle at time $t = t_k$, so that $f_k = f(k\Delta t) + \epsilon_k$ for some error $\epsilon_k$ lying between $-\epsilon$ and $\epsilon$. We'll approximate the velocity $v_k = f'(t_k)$ of the particle at time $t_k$ as

$$v_k \approx \frac{f_{k+1} - f_k}{\Delta t}.$$

Write out $v_k$ explicitly in terms of $f$, $t_k$, $t_{k+1}$ and the errors $\epsilon_k, \epsilon_{k+1}$.

b. If we know that $\epsilon = 0.01$ and the true velocity of the particle is always between 1 and 5, about how small can we take $\Delta t$ before the worst case error in our estimate of $v_k$ is as large as the correct value of $v_k$ itself?

## 1.6 Summary

Many ill-posed inverse problems come down to something like what we encountered above: to recover the unknown of interest, often a function (e.g., $r(t)$), we have to somehow differentiate measured data (e.g., $P(t)$). Differentiation in the presence of noise is an unstable operation—small amounts of noise in the data become arbitrarily large errors in our estimate of the derivative. Collecting more data (e.g., more frequently in time) only magnifies the problem!

In Section 3 we'll talk about some approaches to combat ill-posedness.

# 2 Some General Techniques for Inverse Problems

Inverse problems are quite varied and there's certainly no "turn-the-crank" technique for solving all of them, any more than there's a general technique for solving all ODE's, PDE's, or even algebraic equations. But there are some general approaches that often work. A good way to see one in action is to consider the following specific problem.

## 2.1 Gravitational Prospecting

*Geodesy* is a branch of the earth sciences that deals with measuring and modeling the basic structure and geometry of the earth. One important tool that's been used in this science is *gravimetry*, precise measurements of the earth's gravitational field; the field is not uniform! From these kinds of measurements the internal structure (density variations) of the earth can be deduced. It's an inverse problem!

### 2.1.1 A Simplified Forward Problem

Let's examine a simplified but representative version of this problem. Consider a "one-dimensional" bar of length 1 meter stretching along the $x$-axis from the origin to the point $(1, 0, 0)$ in three-dimensional space. This bar has a variable linear density of $\lambda(x)$ kg per meter. Our goal is to determine $\lambda(x)$ without destroying the bar (e.g., cutting it up into pieces). The bar (red) is represented in Figure 5.

To do this, we place a "test mass" of 1 kg somewhere in space and measure the gravitational force exerted on this test mass by the mass of the bar. We repeat this experiment many times, moving the test mass around to various locations. For each location, we measure the gravitational force. We'll assume there are no other masses around!

To make things simple, let's confine our test mass to lie on the line parameterized by $x = t, y = 0, z = 1$ for $-\infty < t < \infty$, parallel to the $x$-axis, in the $xz$-plane; refer to Figure 5, with the bar in red and the line on which we'll place the test mass in blue (and dashed). At each such point we measure the gravitational force $\mathbf{F}$ (a vector) exerted on the test mass. Note $\mathbf{F}$ will depend on the point $(t, 0, 1)$ at which we place the test mass, so $\mathbf{F}$ can be thought of as a function of $t$.

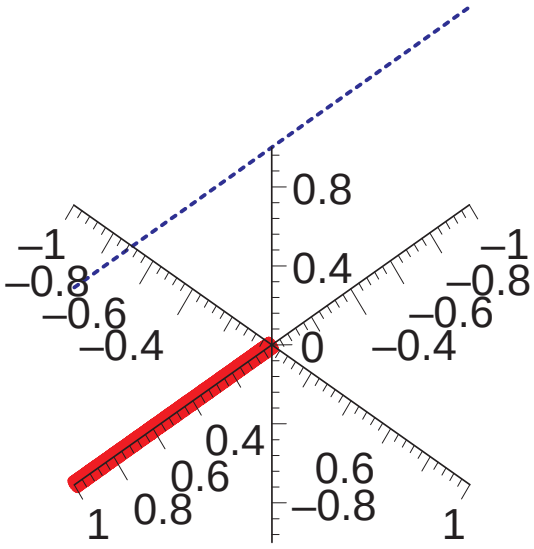


Figure 5: Bar (red) and line for test mass (blue).

The forward problem would be to compute $\mathbf{F}(t)$ given the density $\lambda(x)$. This can be done as follows: First, we (conceptually) slice the bar into small pieces, each of length $dx$; a typical slice lies at some position $(x, 0, 0)$. The mass of the slice is $\lambda(x)\,dx$ (locally "constant" density times length). The gravitational force $d\mathbf{F}$ that this little piece exerts on the test mass is given by

$$d\mathbf{F} = \frac{G\lambda(x)\,dx}{((x-t)^2+1)^{3/2}} < x-t, 0, 1 >$$

which is just Newton's Gravitational Law $\mathbf{F} = \frac{GMm}{r^3}\mathbf{r}$ with $M = \lambda(x)\,dx, m = 1, \mathbf{r} =< x, 0, 0 > - < t, 0, 1 >=< x-t, 0, 1 >$, and $r = |\mathbf{r}| = ((x-t)^2+1)^{1/2}$. Since the gravitational field of a collection of masses is just the sum of their individual fields, we find that the total gravitational force on the test mass is the integral

$$\mathbf{F}(t) = G\int_0^1 \frac{\lambda(x)}{((x-t)^2+1)^{3/2}} < x-t, 0, 1 > \, dx \tag{11}$$

(integrated component by component). This makes it easy to see that the $y$ component of force will always be zero, as expected from symmetry.

Let's suppose we only measure the vertical or $z$ component of the gravitational field (this is what many gravimeters do), so we pay attention to only the third component in equation (11). We then have

$$\mathbf{F}_z(t) = G\int_0^1 \frac{\lambda(x)}{((x-t)^2+1)^{3/2}}\,dx \tag{12}$$

for some range of $t$, where $\mathbf{F}_z(t)$ denotes the $z$-component of the gravitational field at the point $(t, 0, 1)$. Equation (12) quantifies the forward problem: given $\lambda(x)$ for $0 \le x \le 1$, we can compute the vertical component $\mathbf{F}_z(t)$ of the gravitational field exerted by the bar by evaluating the integral on the right in (12).

### 2.1.2 The Inverse Problem; Existence and Uniqueness

The inverse problem is to use equation (12) to recover $\lambda(x)$ from knowledge of $\mathbf{F}_z(t)$ where $t$ ranges over some interval, possibly even $(-\infty, \infty)$.

Equation (12) is an example of an *integral equation* (more specifically, a *Fredholm integral equation of the first kind*). Unfortunately, integral equations don't often make appearances in the undergraduate curriculum, but they are a vast area of research in higher mathematics. In the present case it's not at all clear how to find $\lambda(x)$ given the function $\mathbf{F}_z(t)$ in equation (12), or whether it can even be done.

The first question to ask is whether given the function $\mathbf{F}_z(t)$ we can expect a solution to equation (12) to exist. The answer is "yes" if $\mathbf{F}(t)$ decays to zero fast enough as $|t| \to \infty$ and is differentiable enough, but the proof is technical and not of great interest at the moment—we'll just assume an appropriate $\lambda(x)$ exists.

Slightly more relevant is the question of uniqueness: Is $\lambda$ uniquely determined from knowledge of $\mathbf{F}_z(t)$ (whether or not we know how to find $\lambda$)? The alternative is that two or more different choices for the density function $\lambda(x)$ could yield the same gravitational field (at least along the line we're measuring on). In this case the problem would be unsolvable. However, it turns out that $\lambda$ is uniquely identifiable from the given data, but the proof uses Fourier integral transforms. You'll have to take my word for it or work it out yourself.

So it comes down to this: If we have the function $\mathbf{F}_z(t)$ (or really, measurements of $\mathbf{F}_z(t)$ for some finite number of $t$) how can we determine $\lambda$? How stable is the inverse problem, and if it's not stable, what should we do about it?

**Example 2** Suppose $\lambda(x) = 2 + \sin(x)$ and we measure $\mathbf{F}_z(t)$ at 11 points of the form $t = -2, -1.6, -1.2, \ldots, 2$ (stepsize 0.4). The integral in (12) can be worked numerically for any given $t$; we'll just take $G = 1$ for simplicity. A graph of the resulting values is shown in Figure 6, but it's not easy to see how $\lambda$ is manifest in the data.

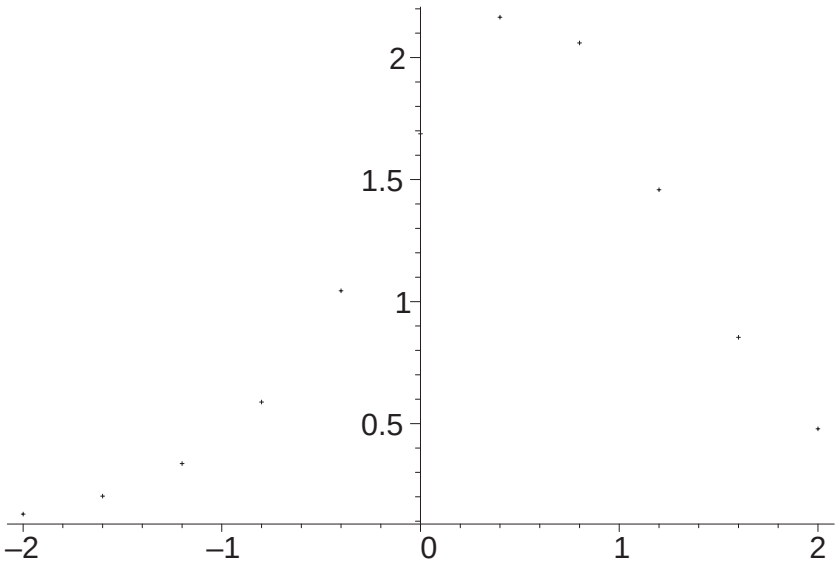


Figure 6: Vertical ($z$) component of gravitational field.

How can we estimate $\lambda(x)$ from these 11 data points?

## 2.2 Output Least-Squares

As mentioned at the start of this section, inverse problems tend to fall into a few different classes (like DE's), and each class has its own solution techniques and theoretical issues. There aren't many truly general techniques for attacking inverse problems, but one approach that's widely useful is the method of *output least-squares*.

The idea is this: We have a "real" physical system with some unknown physical parameters we'd like to estimate; think of $g$ for the pendulum or $\lambda(x)$ for the bar above. We observe or measure some aspect of the system's behavior, e.g., the period of the pendulum or the gravitational field of the bar, with the hope that the unknown is somehow encoded in this information. The situation is graphically depicted on the left in Figure 7, where we'll focus on the gravitational example. We'll use $\lambda^*$ to denote the true value of the physical parameter, in this case the density. We might think of $\lambda^*$ as an "input" to the real physical system, with the measured data as an output. To estimate $\lambda^*$ from data, we construct a mathematical model of the real

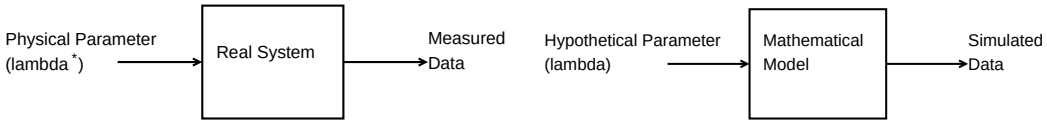


Figure 7: Real system and mathematical model.

physical system, in which we may plug hypothetical values for the density function; we'll use $\lambda$ to denote this hypothetical or "tunable" parameter. For any given choice of $\lambda$ we can "solve" the mathematical model and generate simulated output data, as illustrated on the right in Figure 7.

In *output least-squares estimation* we construct an estimate of $\lambda^*$ by adjusting our hypothetical $\lambda$ until the output from the mathematical model best matches the output from the real system. The resulting value of $\lambda$ is then taken as an estimate of $\lambda^*$. We quantify how well the real and hypothetical outputs match in a "least squares" sense. More precisely, we estimate $\lambda^*$ by minimizing the function

$$Q(\lambda) = \sum_{k=1}^{M} (y_k(\lambda) - y_k^*)^2 \tag{13}$$

where $y_k^*$ is an observed data point for the real system and $y_k(\lambda)$ is the corresponding output of the simulated system ($y_k$ depends on our hypothetical $\lambda$). The minimization of $Q$ is done using any standard optimization technique, though there are optimization techniques specifically designed for these types of problems.

Of course the hope is that if $Q(\lambda)$ is small (so $y_k(\lambda) \approx y_k^*$, the data are in close agreement) then $\lambda \approx \lambda^*$ in some reasonable sense. This is often the case, but numerous problems can arise. The best way to understand least-squares is to see it in action.

## 2.3 Applying Least-Squares to Gravitational Prospecting: Approach 1

Minimizing $Q$ as defined by equation (13), in which $\lambda$ can be "any" function, is too hard (it's a problem in the Calculus of Variations). To make things more tractable we usually assume some specific form for $\lambda$, by assuming $\lambda$ can be specified with finitely many parameters. For example, we might assume that $\lambda$ is a polynomial of given degree, or a Fourier expansion with some specified number of terms. Minimizing $Q$ then becomes a more standard problem of minimizing a function of finitely many variables.

Some examples are shown below.

### 2.3.1 Assuming $\lambda$ is a Polynomial

In order to estimate $\lambda(x)$ from the data, we're going to assume that $\lambda(x) = a_0 + a_1 x + a_2 x^2$, that is, we're going to do the best job we can explaining the observed data in Figure 6 under the assumption that $\lambda$ is a quadratic polynomial. The advantage of this assumption is that $\lambda$ is now characterized by just three numbers $a_0, a_1, a_2$ (instead of living in an infinite dimensional function space). Of course we can also use polynomials of higher degree for $\lambda(x)$.

The data points $y_1^*, \ldots, y_{11}^*$ for the true density in Figure 6 are

| $t_k$ | $-2.0$ | $-1.6$ | $-1.2$ | $-0.8$ | $-0.4$ | $0.0$ |
|---|---|---|---|---|---|---|
| $y_k^*$ | 0.12942 | 0.20281 | 0.33639 | 0.58837 | 1.04402 | 1.6875 |

| $t_k$ | 0.4 | 0.8 | 1.2 | 1.6 | 2.0 |
|---|---|---|---|---|---|
| $y_k^*$ | 2.1654 | 2.0597 | 1.4583 | 0.85354 | 0.47853 |

rounded to 5 significant figures. Plugging $\lambda(x) = a_0 + a_1 x + a_2 x^2$ into equation (12) and evaluating the integral at each $t$ value $t = -2, -1.6, \ldots, 2$ yields a bit of a mess. But, for example, when $t = -1.2$ we obtain

$$\mathbf{F}_z(-1.2) = 0.14215a_0 + 0.055807a_1 + 0.03292a_2.$$

A similar expression holds for the other $t$ values. Note this expression is linear in the $a_k$ because the forward problem is linear with respect to $\lambda$.

We now minimize

$$Q(a_0, a_1, a_2) = \sum_{k=1}^{11} (\mathbf{F}_z(t_k) - y_k^*)^2$$

as a function of $a_0, a_1, a_2$ (the dependence of $\mathbf{F}_z(t_k)$ on the $a_k$ isn't explicitly indicated). In this case $Q$ is a simple quadratic function of the $a_k$ and can be minimized with Calc 3 techniques—just find the critical point, which involves solving a system of LINEAR equations. In more complicated inverse problems, especially nonlinear problems, we'd have to use dedicated optimization software. For the present problem the optimal values are $a_0 = 1.9923, a_1 = 1.0916, a_2 = -0.23532$. The resulting estimate for $\lambda(x)$ and the correct density are shown together in Figure 8. The recovered density is

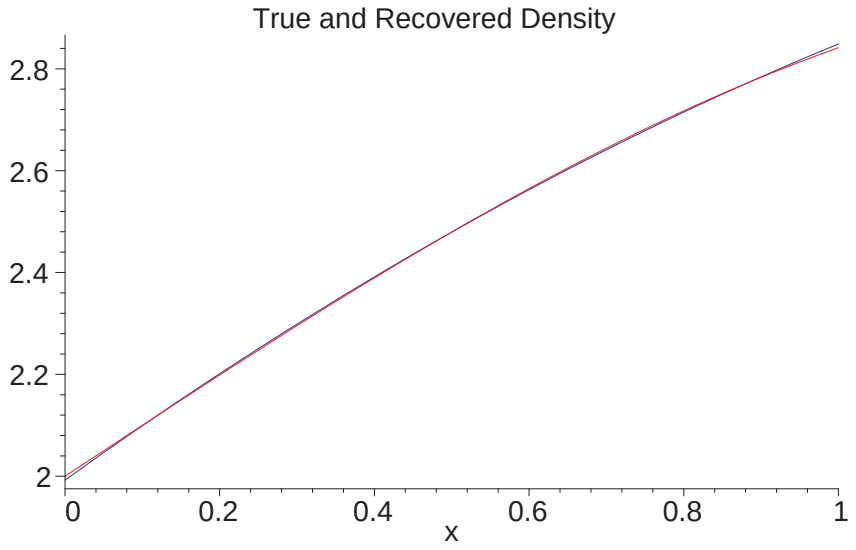


Figure 8: True (red,dashed) and recovered (blue, solid) densities.

very close to the correct density, almost indistinguishable.

Of course, this result is based on accurate data with no noise. Also, in this case the true density could be well-represented by a quadratic polynomial. If we add uniformly distributed independent noise in the interval $[-0.01, 0.01]$ to each data point above (but keep the same $\lambda$) the recovered density is shown in Figure 9. This answer is still pretty reasonable.

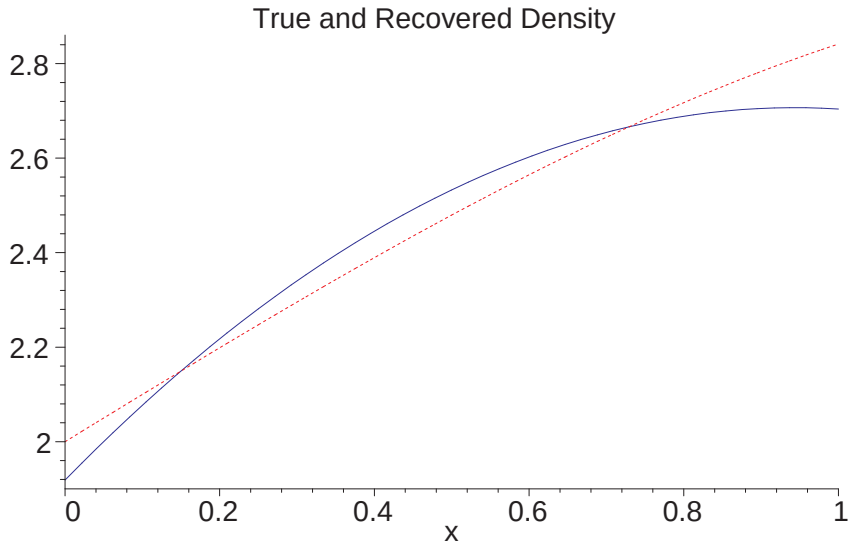


Figure 9: True (red,dashed) and recovered (blue, solid) densities with noisy data (left).

### 2.3.2 Assuming $\lambda$ is a Sum of Cosines

However, this inverse problem is more ill-posed then one might be led to believe by the above computations. One of the hallmarks of ill-posed inverse problems and integral equations is that highly oscillatory functions are hard to recover. To illustrate, let's change the true density function to $\lambda(x) = 2 + x\sin(8x)$; this function is graphed on the left in Figure 10. Also, to allow more flexibility in representing $\lambda$, let's assume that

$$\lambda(x) = \sum_{k=0}^{10} a_k \cos(k\pi x) \tag{14}$$

so $\lambda$ is characterized now by the 11 numbers $a_0, \ldots, a_{10}$. (Recall that any continuous function can be represented as a cosine series; we're just using the 11 basis functions $\cos(k\pi x)$ for $0 \le k \le 10$.) Using $Q$ as in equation (13), we thus have to minimize a quadratic function of 11 parameters, $a_0, \ldots, a_{10}$. Doing so with noise-free data (to 10 significant figures) produces the result on the right in Figure 10, a disaster. We can't even see the true density on the graph.

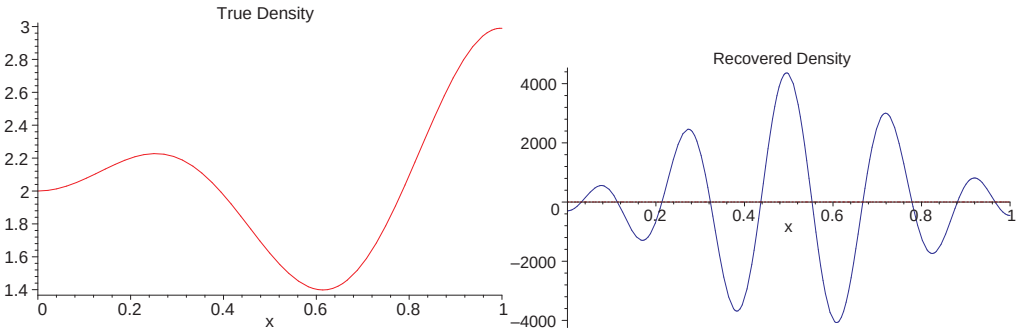


Figure 10: True (left) and recovered (right) densities with noisy data (left), 11 basis functions.

This is a reflection of the underlying instability in this inverse problem. If we're willing to confine our parameter selections to a low-dimensional space (e.g., quadratic polynomials) and if the true parameter is well-represented by such a selection, the problem can be solved accurately, even with noise. But as we increase the dimension of our parameter space the problem quickly becomes unstable, often resulting in excessively oscillatory estimates. This happens here even with "clean" data—the rounding error is enough. And unfortunately, to represent arbitrary functions accurately we may need to go to higher dimensional spaces, e.g., more Fourier basis functions or higher degree polynomials.

Some computational techniques for addressing instability in inverse problems will be discussed in Section 3.

## 2.4 Pros and Cons of Least-Squares

The attractions of least-squares are obvious. If we have a data $y_k^*$, $1 \le k \le N$ for some physical system and a mathematical model that depends on the unknown parameter $\lambda$, we just minimize $Q(\lambda)$ in equation (13). All we need

is the ability to compute $y_k(\lambda)$, which typically means solving an ODE, PDE, or integral equation, and we need some kind of optimization software. If you have these computational capabilities then least-squares provides a plug-and-play method for solving inverse problems, with no emotional or intellectual involvement needed!

However, least-squares has some drawbacks, especially if applied naively:

1. The least-squares function $Q$ may have MANY local minima. Most numerical optimization software will locate only one, and it need not be the best fit to the data. You may well end up with estimates that locally minimize $Q$ but have nothing to do with correct or physically relevant solutions (e.g., see Problem 11 above.)

   A good way to partially overcome this is to start the optimizer with a good initial guess. This requires some understanding of the inverse problem and where the solution is likely to be in the parameter space. You may need to use some other approximate technique on the inverse problem to get a good initial guess.

2. Least-squares doesn't address the ill-posedness of the inverse problem. With no understanding of the nature or severity of the ill-posedness, you don't know whether if your answers mean anything, especially in the presence of noise. Some kind of stabilization or "regularization" as discussed in Section 3 will be needed, but doing this properly probably requires some deeper understanding of the inverse problem.

3. Least-squares doesn't address the uniqueness of the inverse problem. If the data do not uniquely determine the unknown, the estimate you get may not be the true physical parameter(s) that govern the system. For example, in Problem 11 below, there are in fact infinitely many choices for $k$ that yield exactly the ten data points given. In the absence of uniqueness you'll need to incorporate additional information into the process in order to pick out the correct solution (e.g., in Problem 11 maybe we know the spring constant is in the range $15 \leq k \leq 20$ from other considerations).

**Problem 7** Suppose we are given a differentiable function $f$ on some interval and asked to solve the integral equation

$$\int_0^t \phi(s)\, ds = f(t)$$

for the unknown function $\phi$. Verify that $\phi = f'$. That is, the ill-posed problem of differentiating a function can be cast as that of solving an integral equation.

**Problem 8** A certain function $\phi(x)$ defined on $0 \leq x \leq 1$ satisfies the integral equation

$$\int_0^1 t^3 e^{-xt} \phi(x)\, dx = F(t) \tag{15}$$

for some function $F$, with $0 \leq t \leq 1$. We have data $F(0) = 0, F(1/4) = 0.01855, F(1/2) = 0.12990, F(3/4) = 0.38585$, and $F(1) = 0.80918$. Use this data to estimate $\phi(x)$ as follows.

a. Let's try constructing a solution of the form $\phi(x) = a_0 + a_1 x + a_2 x^2$, a quadratic polynomial. Plug this polynomial into the left side of equation (15) and evaluate the integral. You should obtain a function $g(t, a_0, a_1, a_2)$.

b. We'd like to arrange the $a_k$ so that $g(t, a_0, a_1, a_2) = F(t)$ for the given $t$ values $0, 0.25, 0.5, 0.75, 1.0$. This leads to 5 equations in just 3 unknowns, probably unsolvable (unless the data is perfect). Instead, minimize the function

$$Q(a_0, a_1, a_2) = \sum_{k=1}^{5} (g(t_k, a_0, a_1, a_2) - F(t_k))^2$$

where $t_k = (k-1)/4$.

c. Graphically compare your solution to the true solution $\phi(x) = 2 - \cos(\pi x/2)$.

**Problem 9** Let $p(t)$ be the population of some species at time $t$ and suppose $p$ obeys the logistic DE

$$p'(t) = kp(t)(1 - p(t)/M)$$

where $k$ is the *growth rate* for the species and $M$ is the *carrying constant* for the environment, and let's use initial population $p(0) = 100$. Our goal here is to estimate the constants $k$ and $M$ from data.

The solution to this DE is easy to find via separation of variables and is

$$p(t) = \frac{100M}{100 + e^{-kt}(M - 100)}. \tag{16}$$

Suppose we measure the population at times $t = 10, 20, \ldots, 100$ and obtain

$$234, 441, 694, 859, 943, 988, 992, 997, 996, 1005.$$

a. Write out the least-squares objective function

$$Q(k, M) = (p(10) - p_{10})^2 + (p(20) - p_{20})^2 + \cdots + (p(100) - p_{100})^2$$

where $p_k$ is the population at time $t = k$ from the list above. A computer algebra system like Mathematica will be helpful.

b. Minimize $Q$ by setting $\frac{\partial Q}{\partial k} = 0$ and $\frac{\partial Q}{\partial M} = 0$ and solving these two equations simultaneously. You'll need to solve numerically. Try giving the computer a reasonable initial guess at $k$ and $M$.

c. With the optimal values for $k$ and $M$, see how well $p(t)$ agrees with the measured data.

**Problem 10** One interesting area of inverse problems is that of *nondestructive testing*, using energy in various forms to "see" inside objects without damaging them. One technique that's being explored is the use of heat to find internal defects in objects, or to find internal sources of heat (which can indicate local defects).

Here's a simplified version of such an inverse problem. Consider a bar of "infinite" length stretching along the $x$-axis, initially at zero degrees. Sitting at the origin $x = 0$ is a point heat source that dumps heat (units of energy per time) into the bar at a rate $f(t)$. A little physics and PDE shows that (if we set a few physical constants equals to one for simplicity) the temperature $u(x, t)$ in the bar at position $x$ and any time $t > 0$ is given by the formula

$$u(x, t) = \frac{1}{2\sqrt{\pi}} \int_0^t f(s) \frac{e^{-\frac{x^2}{4(t-s)}}}{\sqrt{t - s}}\, ds. \tag{17}$$

Suppose we can take measurements of the temperature at position $x = 1$ (one unit away from the heat source) over some time interval $0 \leq t \leq T$.

Our goal is to determine the function $f(t)$ on this time interval, if possible. This means that we collect data $d(t)$ where

$$d(t) = \frac{1}{2\sqrt{\pi}} \int_0^t f(s) \frac{e^{-\frac{1}{4(t-s)}}}{\sqrt{t-s}}\, ds \tag{18}$$

(this is just $u(1,t)$ from equation (17).) In summary, given $d(t)$ in equation (18) we want to find $f$.

In particular, suppose we work on the time interval $0 \le t \le 1$, and have collected data

| $t$ | 0.05 | 0.15 | 0.25 | 0.35 | 0.45 |
|---|---|---|---|---|---|
| $d(t)$ | $6.843 \times 10^{-5}$ | $7.992 \times 10^{-3}$ | $2.985 \times 10^{-2}$ | $6.139 \times 10^{-2}$ | $1.003 \times 10^{-1}$ |

| $t$ | 0.55 | 0.65 | 0.75 | 0.85 | 0.95 |
|---|---|---|---|---|---|
| $d(t)$ | $1.452 \times 10^{-3}$ | $1.955 \times 10^{-1}$ | $2.502 \times 10^{-1}$ | $3.087 \times 10^{-1}$ | $3.701 \times 10^{-1}$ |

Let's try estimating $f$ using a quadratic polynomial $p(t) = a_0 + a_1 t + a_2 t^2$.

a. Use a computer algebra system to compute the function $\tilde{d}(t)$ given by

$$\tilde{d}(t) = \frac{1}{2\sqrt{\pi}} \int_0^t p(s) \frac{e^{-\frac{1}{4(t-s)}}}{\sqrt{t-s}}\, ds$$

in terms of $t$ and the coefficients $a_0, a_1, a_2$.

b. Write out the least-squares objective function

$$Q(a_0, a_1, a_2) = (\tilde{d}(0.05) - d(0.05))^2 + \cdots + (\tilde{d}(0.95) - d(0.95))^2$$

using your answer from part (a) and the data from the table above. You should obtain a quadratic function of $a_0, a_1, a_2$.

c. Minimize $Q$ by setting each partial derivative to zero and solving these three equations simultaneously.

d. With the optimal values for $a_k$, see how well $\tilde{d}(t)$ agrees with the measured data. Compare your estimate $p(t)$ to the true value for $f$, namely $f(t) = 1 + 2\sin(1.3t)$.

**Problem 11** Consider an unforced, underdamped spring-mass system with mass $m = 1$ and unknown spring constant and damping constants $k$ and $c$, respectively. Assume the system obeys the usual second order DE

$$mu''(t) + cu'(t) + ku(t) = 0 \tag{19}$$

where $u(t)$ is the displacement of the mass from equilibrium. Let's assume the system starts with initial conditions $u(0) = 1$ and $u'(0) = 0$. The goal is to identify $c$ and $k$ from measurements of the mass position at various times. Indeed, suppose we have the position data

| $t$ | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 |
|---|---|---|---|---|---|
| $u(t)$ | $-0.280$ | $-0.611$ | 0.624 | 0.0750 | $-0.545$ |

| $t$ | 6.0 | 7.0 | 8.0 | 9.0 | 10.0 |
|---|---|---|---|---|---|
| $u(t)$ | 0.299 | 0.237 | $-0.393$ | 0.070 | 0.266 |

a. The solution to the underdamped spring-mass DE with the given initials conditions can be written

$$u(t) = e^{-ct/2}\cos(\omega t) + \frac{c}{\sqrt{4k - c^2}} e^{-ct/2}\sin(\omega t) \tag{20}$$

where $\omega = \sqrt{4k - c^2}/2$. Use $u(t)$ as in equation (20) with unspecified values for $c$ and $k$ to write out the least-squares objective function

$$Q(c, k) = \sum_{k=1}^{10} (u(k) - u_k)^2 \tag{21}$$

where $u_k$ is the measured value of $u$ from the table above and $u(t)$ is given by equation (20). You should end up with a large and nonlinear objective function of $c$ and $k$. Of course, a computer algebra system will be necessary.

b. Plot the objective function $Q(c, k)$ on the range $0 < c < 1$, $2 < k < 10$. You should see a (local) minimum. Find the corresponding value of $c$ and $k$ (either by graphically estimating the location of the minimum or finding the critical point; if the latter, you'll need to use a numerical equation solver.) After you have reasonable values for $c$ and $k$, plot $u(t)$ from equation (20) on the range $0 \leq t \leq 10$ and compare with the tabled data.

c. Try plotting on the range $0 < c < 1, 15 < k < 20$; you should see another minimum. Find the relevant values for $c$ and $k$, and then plot and compare $u(t)$ as in part (b).

d. Repeat part (c) on the range $0 < c < 1, 65 < k < 70$.

e. Of the $c$ and $k$ estimates you found in parts (b)-(d), which best fits the data? (Hint: they all fit it perfectly).

# 3 Regularization

The idea behind regularization is to replace an unstable problem with a closely related stable problem, or really a family of stable problems. We refer to these as the "regularized" problems. The regularized problems should be indexed by some kind of "tunable" parameter $\alpha$. As $\alpha \to 0^+$ the regularized problem converges to the unstable original problem, and as $\alpha$ increases the regularized problem becomes more and more stable—but at the cost of differing markedly from the original problem we want to solve. This is best seen via an illustration.

## 3.1 Differentiation of Noisy Data

Suppose we have measurements $f(t_k)$ of a differentiable function $f(t)$ on the interval $[0, 1]$, where $t_k = k\Delta t$ for $k = 0, 1, 2, \ldots$ for some interval $\Delta t > 0$. Our goal is use this data to estimate $f'$. However, our measurements of $f$ are subject to some kind of error. If $f_k$ denotes our measured value for $f(t_k)$ then what we really have is $f_k = f(t_k) + \epsilon_k$, where $\epsilon_k$ is the error; $\epsilon_k$ might be random, or it might be something like round-off error (e.g., in the interest rate problem of Section 1.4). We can estimate $f'(t_k)$ as

$$f'(t_k) \approx \frac{f_{k+1} - f_k}{\Delta t} = \frac{f(t_{k+1}) - f(t_k)}{\Delta t} + \frac{\epsilon_{k+1} - \epsilon_k}{\Delta t}. \tag{22}$$

This is of course just equation (10) all over again. As $\Delta t$ goes to zero the first term on the right in (22) approaches $f'(t_k)$, exactly what we want, but the second term blows up (we can't count on $\epsilon_{k+1} = \epsilon_k$ if it's noise!)

To illustrate, let $f(t) = t + \sin(t)$ and let's take $\Delta t = 0.01$. For noise we'll choose $\epsilon_k$ to be uniformly distributed between $-0.005$ and $0.005$, all $\epsilon_k$ independent. The results are shown in Figure 11. The effect of the second

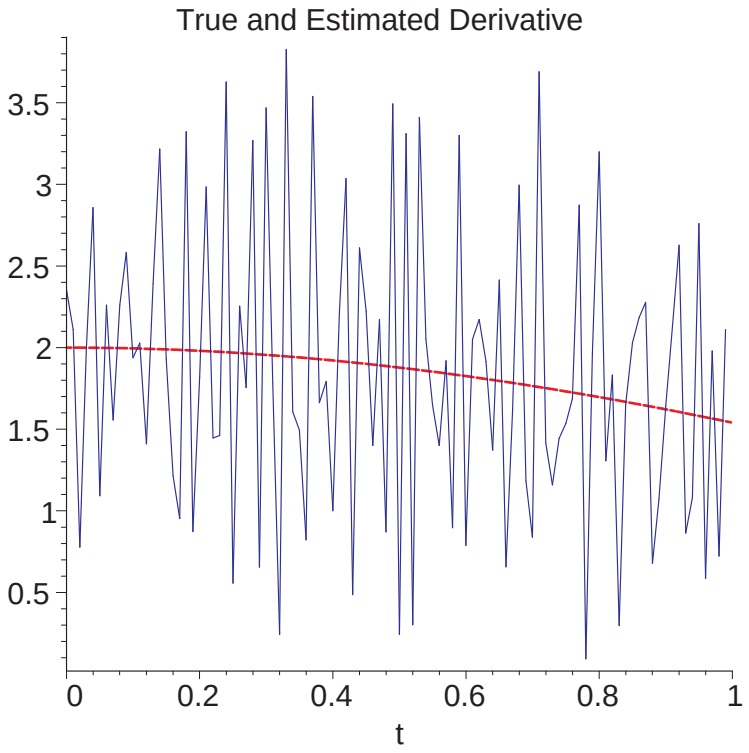


Figure 11: True (dashed) and estimated (solid) derivative of $f(t) = t + \sin(t)$, no regularization.

term on the right in equation (22) is to introduce wild oscillations into the estimate of $f'$, because the quantity $\epsilon_{k+1} - \epsilon_k$ is of "random" size and sign, so division by a small $\Delta t$ is a disaster.

## 3.2 Regularizing the Estimate of $f'$

In some sense we could argue that the estimate of $f'$ produced above is the best we can do—it agrees with the measured data perfectly—but in reality we often know that the underlying function we're trying to estimate should be smooth (think of the interest rate function $r(t)$). In light of this fact, and knowing that our data is noisy and the estimation procedure unstable, we should accept that the wild oscillations above are not real, but artifacts. What we want is some way to produce estimates $d_k$ of $f'(t_k)$ in which the $d_k$ are in reasonable agreement with the measured data, but at the same time

the $d_k$ are "encouraged" to not oscillate wildly from index $k$ to index $k+1$.

Here's one approach. Take $\Delta t = 1/M$ for some integer $M$, so we have measurements of $f$ at points $t_k = k\Delta t$, where $0 \leq k \leq M$. Our goal is to produce estimates $d_k$ of $f'(t_k)$ at times $t_0, t_1, \ldots, t_{M-1}$ (our one-sided difference formula (22) doesn't cover the right endpoint $t_M$.) Consider finding the $d_k$ by minimizing the quadratic functional

$$Q(d_1, \ldots, d_{M-1}) = \frac{1}{2} \sum_{k=0}^{M-1} \left( d_k - \frac{f_{k+1} - f_k}{\Delta t} \right)^2$$

with respect to $d_1, d_2, \ldots, d_{M-1}$ (the $1/2$ in front isn't important, it just makes things prettier later.) It's easy to see that the optimal choice for the $d_k$ is $d_k = (f_{k+1} - f_k)/\Delta t$, exactly the horrible result produced by equation (22). So let's modify $Q$ to encourage the minimizing $d_k$ to not oscillate so wildly, by taking

$$Q(d_1, \ldots, d_{M-1}) = \frac{1}{2} \sum_{k=0}^{M-1} \left( d_k - \frac{f_{k+1} - f_k}{\Delta t} \right)^2 + \frac{\alpha}{2} \sum_{k=0}^{M-2} \left( \frac{d_{k+1} - d_k}{\Delta t} \right)^2 \quad (23)$$

where $\alpha \geq 0$ is a parameter we can choose. As before, we'll produce our estimates $d_k$ of $f'(t_k)$ by minimizing $Q$ in (23). The addition of the second term on the right in (23) is an example of what is called *Tikhonov* regularization.

To see why this might produce better estimates, look at the second term on the right in (23). When $\alpha > 0$ this term "wants" to take $d_{k+1} = d_k$ when we minimize. That is, this terms gets large if the estimates $d_k$ vary wildly. When we minimize $Q$ for a given $\alpha > 0$ the wild oscillations in the $d_k$ tend to be damped out (depending on how big $\alpha$ is), at the expense of making the first term on the right larger (the $d_k$ don't agree with the data as well.) The trade-off between fidelity to the raw data and smoothness is illustrated in figures below. The second term on the right in (23) is often called a *penalty term*, since in the optimization this term penalizes solutions in which the $d_k$ vary wildly.

Minimizing $Q$ is straightforward, since $Q$ is quadratic. At a minimum we have $\partial Q/\partial d_j = 0$ for $j = 0, \ldots, M-1$, and it's easy enough to compute that for $1 \leq j \leq M-2$ we obtain

$$\frac{\partial Q}{\partial d_j} = (d_j - (f_{j+1} - f_j)/\Delta t) + \alpha/\Delta t^2(-d_{j-1} + 2d_j - d_{j+1}) = 0$$

while when $j = 0$ or $j = M - 1$ (oddball endpoint cases) we have

$$\begin{aligned} \frac{\partial Q}{\partial d_0} &= (d_0 - (f_1 - f_0)/\Delta t) + \alpha/\Delta t^2(d_0 - d_1) = 0 \\ \frac{\partial Q}{\partial d_{M-1}} &= (d_{M-1} - (f_M - f_M - 1)/\Delta t) + \alpha/\Delta t^2(d_{M-1} - d_{M-2}) = 0. \end{aligned}$$

This is a system of LINEAR equations for $d_0, \ldots, d_{M-1}$, and it can be written conveniently in matrix form as $\mathbf{Ad} = \mathbf{b}$ where $\mathbf{A}$ is the $M \times M$ matrix

$$\mathbf{A} = \mathbf{I} + \frac{\alpha}{\Delta t^2} \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ & & & \vdots & & \\ 0 & \cdots & 0 & -1 & 2 & -1 \\ 0 & 0 & \cdots & 0 & -1 & 1 \end{bmatrix} \tag{24}$$

and $b_j = (f_j - f_{j-1})/\Delta t$ with $1 \leq j \leq M$.

In a nutshell, we solve $\mathbf{Ad} = \mathbf{b}$ for some choice of $\alpha > 0$ and take $d_j$ as the estimate of $f'(t_j)$. What we should take for $\alpha$ is discussed a bit below.

Here are some examples of this regularization procedure in action. For the choices $\alpha = 0.001$ and $\alpha = 0.01$ we obtain the graphs in Figure 12. The

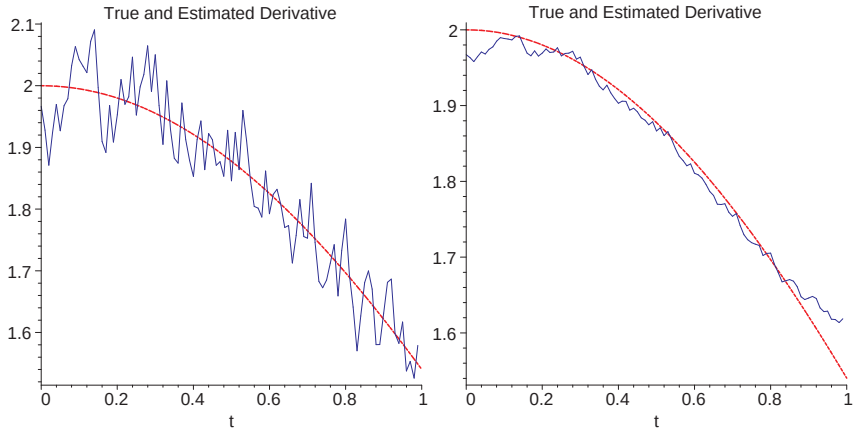


Figure 12: True (dashed) and estimated (solid) derivative of $f(t) = t + \sin(t)$, $\alpha = 0.001$ (left) and $\alpha = 0.01$ (right).

choice $\alpha = 0.001$ is an example of "under-regularization," while $\alpha = 0.01$ looks pretty good, perhaps still a bit under-regularized.

Figure 13 shows two examples of over-regularization, with $\alpha = 0.1$ and $\alpha = 1.0$ on the left and right, respectively. With larger values of $\alpha$ the minimization of $Q$ puts a premium on keeping the second term on the right in equation (23) small (non-oscillatory $d_k$) at the expense of making the first term small (good agreement with the data).

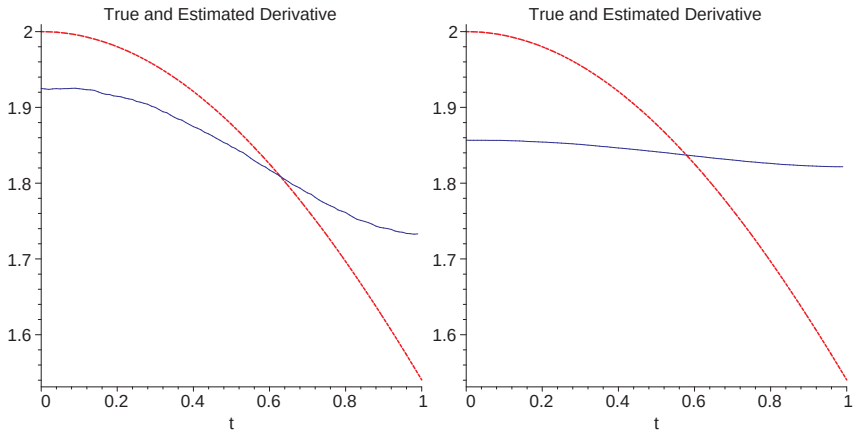


Figure 13: True (dashed) and estimated (solid) derivative of $f(t) = t + \sin(t)$, $\alpha = 0.1$ (left) and $\alpha = 1.0$ (right).

## 3.3 Regularization for the Gravitational Prospecting Problem

Let's revisit the gravitational prospecting problem, in particular, the cosine expansion approach of Section 2.3.2, to see how these ideas might be applied there. As with differentiating data, the goal here is to regularize the reconstruction by "encouraging" the estimate not to oscillate so much. To the objective function $Q(a_0, \ldots, a_{10})$ we obtain from equation (13) (and $\lambda$ of the form in equation (14)) we add a regularizing or "penalty term" to encourage the solution not to oscillate excessively. An appropriate term would be something like

$$R = \int_0^1 \left[ \frac{d}{dx} \left( \sum_{k=0}^{10} a_k \cos(k\pi x) \right)^2 \right] dx = a_0^2 + \frac{\pi^2}{2} \sum_{k=1}^{10} k^2 a_k^2$$

(the right side above follows easily from working the integral). The integral that defines $R$ is large if the derivative of the function $\lambda$ is large in a

mean square sense, so adding a multiple $\alpha R$ to the objective function should encourage the minimizing solution to be smoother. Moreover, since $R$ is quadratic the function $Q + R$ is quadratic and the normal equations are still linear, so easy to solve. Again, the addition of the term "$R$" above to the objective function is an example of Tikhonov regularization.

Adding this regularization or penalty term with $\alpha = 10^{-8}$ produces the result in Figure 14, a vast improvement over that on the right in Figure 10. If we then add a small amount of noise as above, the same regularization parameter yields the reconstruction in Figure 15, still pretty good.

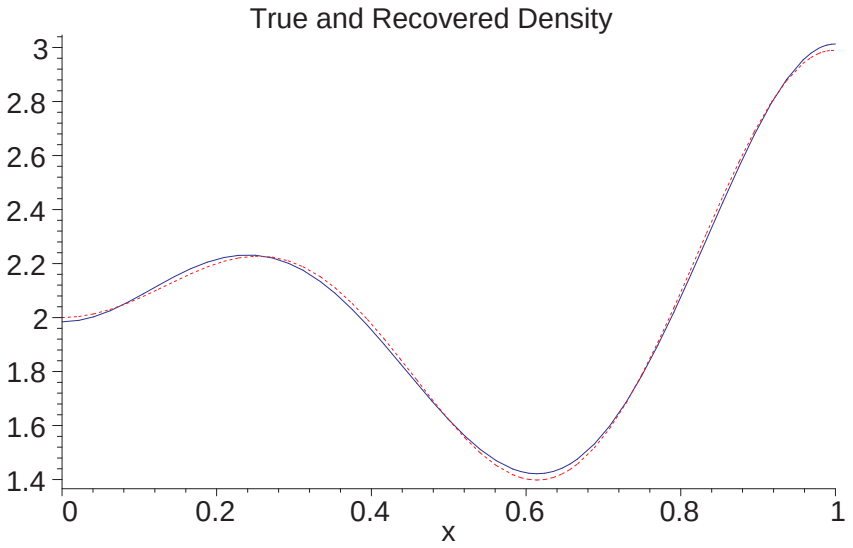


Figure 14: True (red,dashed) and recovered (blue, solid) densities with noiseless data (left), 11 basis functions.

## 3.4 Choosing the Regularization Parameter

In each example above the natural question is "what should we choose for the regularization parameter $\alpha$?" There are a large number of techniques designed to get at this problem—people devote entire careers to the selection of regularization parameters. It depends on what a priori information we have about the unknown function we're trying to estimate, e.g., "its derivative is always between 1.2 and 2.3," and the noise level in the data is also an important factor. Such information or assumptions can be used to help

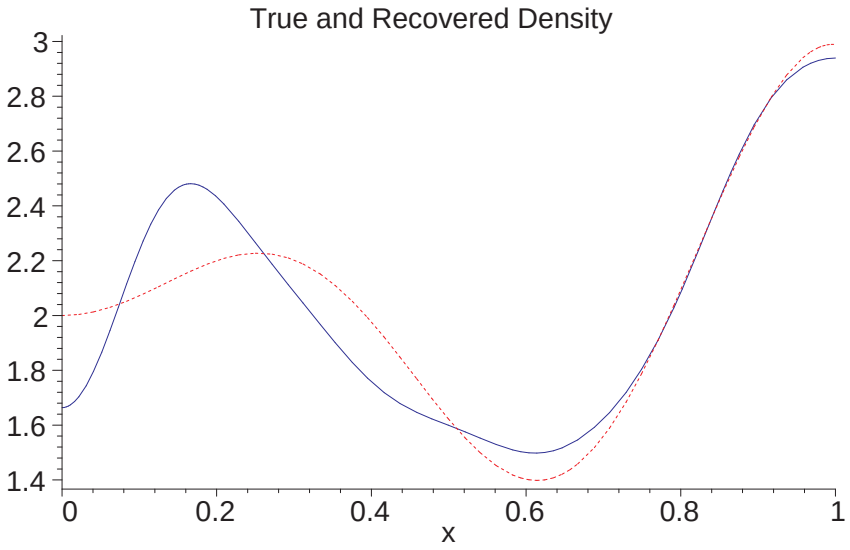


Figure 15: True (red,dashed) and recovered (blue, solid) densities with noisy data (left), 11 basis functions.

dictate the value of $\alpha$. In practice, one often does simulations with "made up" functions and realistic noise levels to see what range of $\alpha$ works best. We then use this value of $\alpha$ on the "real" problem.

**Problem 12**

a. Consider the interest rate problem of Section 1.4 with true interest rate $r(t) = 0.1(2 + \sin(t))$ and $P(0) = 100$. In this case the true balance at any time is

$$P(t) = 100e^{0.1(1+2t-\cos(t))}.$$

We can simulate measuring $P(t)$ over a 5 year period at $M = 100$ time intervals of width $\Delta t = 5/M$ and rounding to the near cent with the Matlab commands

```
M = 100
dt = 5/M
P = inline('100 * exp(0.1 * (1 + 2. * t - cos(t)))')
tk = [0 : M] * 5/M;
Pdat = P(tk);
```

We can then round the $P(t_k)$ to the nearest cent with

$$\texttt{Pdatr} = \texttt{round}(\texttt{Pdat} * 100)/100;$$

We can estimate the interest rate $r(t_k)$ for $0 \le k \le M-1$ by using equation (9) with the command

$$\texttt{rest} = (\texttt{Pdatr}(2:\texttt{M}+1) - \texttt{Pdatr}(1:\texttt{M}))/\texttt{dt}./\texttt{Pdatr}(1:\texttt{M});$$

To compare the estimated and true interest rates graphically execute

$$\begin{aligned}&\texttt{r} = \texttt{inline}('0.1 * (2 + \sin(\texttt{t}))')\\ &\texttt{rtru} = \texttt{r}(\texttt{tk}(1:\texttt{M}));\\ &\texttt{plot}(\texttt{tk}(1:\texttt{M}), \texttt{rtru}, \texttt{tk}(1:\texttt{M}), \texttt{rest})\end{aligned}$$

Repeat this for $M = 500, 1000, 5000$.

b. The graph in part (a) should look pretty bad when $M \ge 1000$. Of course, the heart of the problem is the estimation of $P'(t)$ lurking in equation (9). Regularize the results of part (a) as follows. First, form "raw" estimates of $P'(t_k)$ with

$$\texttt{Pdest} = (\texttt{Pdatr}(2:\texttt{M}+1) - \texttt{Pdatr}(1:\texttt{M}))/\texttt{dt};$$

then choose a value of $\alpha$, e.g., $\texttt{alpha} = 0.01$. Construct the regularized estimates $d_0, \ldots, d_{M-1}$ of $P'(t_k)$ by solving $\mathbf{Ad} = \mathbf{b}$ where $\mathbf{A}$ is as in equation (24). This can be done with

$$\begin{aligned}&\texttt{A0} = 2 * \texttt{diag}(\texttt{ones}(1,\texttt{M})) - \texttt{diag}(\texttt{ones}(1,\texttt{M}-1),1) - \texttt{diag}(\texttt{ones}(1,\texttt{M}-1),-1);\\ &\texttt{A0}(1,1) = 1; \texttt{A0}(\texttt{M},\texttt{M}) = 1;\\ &\texttt{A} = \texttt{eye}(\texttt{M}) + \texttt{alpha}/\texttt{dt}^2 * \texttt{A0};\\ &\texttt{d} = \texttt{A}\backslash\texttt{Pdest}';\end{aligned}$$

Finally, a regularized estimate of $(t_k)$ for $0 \le k \le M-1$ is obtained and plotted as

$$\begin{aligned}&\texttt{rreg} = \texttt{d}./\texttt{Pdatr}(1:\texttt{M})';\\ &\texttt{plot}(\texttt{tk}(1:\texttt{M}), \texttt{rtru}, \texttt{tk}(1:\texttt{M}), \texttt{rreg})\end{aligned}$$

Experiment with different values for $\alpha$.

# 4 The Singular Value Decomposition and Regularization

## 4.1 Introduction

Many inverse problems, for example, the gravitational prospecting problem, lead to linear systems of equations, $\mathbf{Ax} = \mathbf{b}$. The ill-posedness of the inverse problem is reflected in the fact that the resulting matrix $\mathbf{A}$ is almost singular. This means that small changes in the right hand side vector $\mathbf{b}$ (e.g., noise in the data) produce huge changes in the solution $\mathbf{x}$. Even inverse problems that lead to nonlinear equations ultimately require the numerical solution of linear systems, since nonlinear solution methods (e.g., Newton's method) require the solution of linear systems as a subproblem. Regularizing the solution of almost singular linear systems of equations is thus an important topic in inverse problems. One elegant approach is to use the *singular value decomposition*.

Consider the prototypical problem in linear algebra, the analysis and solution of a system of $m$ equations in $n$ unknowns,

$$\mathbf{Ax} = \mathbf{b}$$

where $\mathbf{A}$ is an $m \times n$ matrix, $\mathbf{b} \in \mathbb{R}^m$, and the vector $\mathbf{x} \in \mathbb{R}^n$. It may be the case that the system is consistent but underdetermined (usually the case if $m < n$, i.e. more unknowns than equations), or the system may be overdetermined (usually the case if $m > n$, i.e., more equations than unknowns.) Even if the system is uniquely solvable, it may be "close" to underdetermined or overdetermined, if the matrix $\mathbf{A}$ is close to singular.

If the system is consistent but underdetermined then we have infinitely many solutions. Which one should we pick? In many cases an argument can be made that we should take that solution $\mathbf{x}$ of minimal norm. On the other hand, if the system is inconsistent there is no solution. An argument can often be made that we should then settle for that vector $\mathbf{x}$ that "comes closest" to solving the system, in that $\mathbf{x}$ minimizes the usual Pythagorean norm $\|\mathbf{Ax} - \mathbf{b}\|$. If there are many vectors that minimize $\|\mathbf{Ax} - \mathbf{b}\|$ then we take that (unique) such $\mathbf{x}$ of minimal norm.

There is a powerful tool in linear algebra that lets us tackle all of these issues in a unified framework, and do a great deal more. It's called the *singular value decomposition* or "SVD" for short.

## 4.2 The Singular Value Decomposition

Let $\mathbf{A}$ be an $m \times n$ matrix with (for simplicity) real-valued entries. The singular value decomposition of $\mathbf{A}$ is the factorization

$$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T \tag{25}$$

where the "$T$" superscript denotes matrix transpose. Here $\mathbf{U}$ is an $m \times m$ orthogonal matrix, $\mathbf{V}$ an $n \times n$ orthogonal matrix, and $\mathbf{S}$ is an $m \times n$ matrix (same shape as $\mathbf{A}$) with all diagonal entries $S_{ii}$ non-zero; we usually just use one subscript to refer to these entries, i.e., $S_i$ for the entry in the $(i, i)$ position. These diagonal entries are called the *singular values* of $\mathbf{A}$. Conventionally the singular values are non-negative and appear in descending order, i.e., $S_i \geq S_{i+1}$. Under this convention the singular values of $\mathbf{A}$ are uniquely determined, though the matrices $\mathbf{U}$ and $\mathbf{V}$ are not (for example, we could flip the sign of all elements in both $\mathbf{U}$ and $\mathbf{V}$ without changing the value of the product on the right in (25). Nonetheless, we'll refer to "the" SVD of a matrix, rather than "an" SVD. A proof that the SVD exists for any matrix and a description of efficient algorithms for computing the SVD can be found in [3].

Depending on the relative values of $m$ and $n$, the matrix $\mathbf{S}$ may take several slightly different forms. If $m < n$, e.g., $m = 3, n = 5$ then $\mathbf{S}$ would look like

$$\mathbf{S} = \begin{bmatrix} S_1 & 0 & 0 & 0 & 0 \\ 0 & S_2 & 0 & 0 & 0 \\ 0 & 0 & S_3 & 0 & 0 \end{bmatrix} \tag{26}$$

while if $m > n$, e.g., $m = 5, n = 3$, then $\mathbf{S}$ would look like

$$\mathbf{S} = \begin{bmatrix} S_1 & 0 & 0 \\ 0 & S_2 & 0 \\ 0 & 0 & S_3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{27}$$

In the square case $\mathbf{S}$ would simply be a diagonal matrix. In any of these cases we may some $S_i = 0$.

**Problem 13** Let

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \end{bmatrix}.$$

Verify that the SVD of $\mathbf{A}$ is given by

$$\mathbf{U} = 1, \quad \mathbf{S} = \begin{bmatrix} \sqrt{2} & 0 \end{bmatrix}, \quad \mathbf{V}^T = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}.$$

**Problem 14** If $\mathbf{A}$ has SVD $\mathbf{A} = \mathbf{USV}^T$, what is the SVD of $\mathbf{A}^T$? Use this with the information in Problem 13 to compute the SVD of the transpose of $\mathbf{A}$ in that problem.

**Problem 15** If an $m \times n$ matrix $\mathbf{A}$ has SVD $\mathbf{A} = \mathbf{USV}^T$ and $\mathbf{Q}$ is an orthogonal $m \times m$ matrix, what the SVD of $\mathbf{QA}$?

### 4.2.1 Geometry of the SVD

Before discussing the geometric significance and utility of the SVD, it's worth recalling a couple special properties of orthogonal matrices. First, if $\mathbf{M}$ is a $k \times k$ orthogonal matrix (that is, $\mathbf{M}^T\mathbf{M} = \mathbf{I}$) we have

1. $\mathbf{M}$ is invertible (immediate from $\mathbf{M}^T\mathbf{M} = \mathbf{I}$.)
2. $\|\mathbf{Mx}\| = \|\mathbf{x}\|$ for any $\mathbf{x} \in \mathbb{R}^k$. The proof is simply that

   $$\|\mathbf{Mx}\|^2 =< \mathbf{Mx}, \mathbf{Mx} >=< \mathbf{x}, \mathbf{M}^T\mathbf{Mx} >=< \mathbf{x}, \mathbf{x} >= \|\mathbf{x}\|^2.$$

   Thus multiplication of a vector by an orthogonal matrix preserves length.
3. More generally, $< \mathbf{Mx}, \mathbf{My} >=< \mathbf{x}, \mathbf{y} >$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^k$. The proof is almost the same as above, but easier.

   $$< \mathbf{Mx}, \mathbf{My} >=< \mathbf{x}, \mathbf{M}^T\mathbf{My} >=< \mathbf{x}, \mathbf{y} > .$$

   Thus multiplication by an orthogonal matrix preserves inner products or angles.
4. The columns of $\mathbf{M}$ (or $\mathbf{M}^T$) form an orthonormal basis for $\mathbb{R}^k$. This is immediate from $\mathbf{M}^T\mathbf{M} = \mathbf{I}$.

To see the geometric significance of the SVD, consider the product $\mathbf{Ax} = \mathbf{USV}^T\mathbf{x}$. The first step in computing the product $\mathbf{y} = \mathbf{Ax}$ is the computation of the product $\mathbf{V}^T\mathbf{x}$. Let's define a vector $\mathbf{x}'$ as

$$\mathbf{x}' = \mathbf{V}^T\mathbf{x}. \tag{28}$$

Because $\mathbf{V}$ (and so $\mathbf{V}^T$) is orthogonal, the operation preserves norms. In fact, since $\mathbf{x} = \mathbf{V}\mathbf{x}'$ we see that $\mathbf{x}'$ is the representation of the vector $\mathbf{x}$ with respect to the orthogonal basis formed by the columns of $\mathbf{V}$, and $\|\mathbf{x}\| = \|\mathbf{x}'\|$.

In the same vein, let's define the vector $\mathbf{y}' := \mathbf{U}^T\mathbf{y}$, so that $\mathbf{y}'$ is the representation of the vector $\mathbf{y}$ with respect to the orthogonal basis formed by the columns of $\mathbf{U}$. With this notation the equation $\mathbf{y} = \mathbf{A}\mathbf{x}$ can be written as

$$\mathbf{y}' = \mathbf{S}\mathbf{x}'. \tag{29}$$

Equation (29) lays bare how multiplication by $\mathbf{A}$ maps a vector in $\mathbb{R}^n$ to a vector in $\mathbb{R}^m$: after an orthogonal changes of coordinates in each space, the matrix multiplication is simply a rescaling of some of the components of $\mathbf{x}'$ by the diagonal entries of $\mathbf{S}$.

### 4.2.2 Range and Nullspace of A

Let's take a look at how multiplication by $\mathbf{S}$ affects the vector $\mathbf{x}'$ in a few simple cases. Consider the case in which $\mathbf{S}$ (and hence also $\mathbf{A}$) is a $3 \times 5$ matrix like that in equation (26). Under multiplication by $\mathbf{S}$ the vector $\mathbf{x}' = (x'_1, x'_2, x'_3, x'_4, x'_5)$ becomes the vector $\mathbf{y}' = (S_1x'_1, S_2x'_2, S_3x'_3)$. The values of $x'_4$ and $x'_5$ are "lost" in the multiplication. Moreover, if any of the singular values $S_j = 0$ then the corresponding $x'_j$ component is also zeroed out and the corresponding entry in $\mathbf{y}$ is zero. It's easy to see that dimension of the range of $\mathbf{S}$ is the number of non-zero singular values. The dimension of the nullspace of $\mathbf{S}$ will be 5 minus the number of NON-ZERO singular values. Similar remarks will hold whenever $m \leq n$.

Now consider the case in which $\mathbf{S}$ (and hence also $\mathbf{A}$) is a $5 \times 3$ matrix like that in equation (27). Under multiplication by $\mathbf{S}$ the vector $\mathbf{x}' = (x'_1, x'_2, x'_3)$ becomes the vector $\mathbf{y}' = (S_1x'_1, S_2x'_2, S_3x'_3, 0, 0)$.The last two components of $\mathbf{y}'$ are zero regardless of $\mathbf{x}'$. Moreover, if any of the singular values $S_j = 0$ then the corresponding $x'_j$ component is also zeroed out and the corresponding entry in $\mathbf{y}$ is zero. Again, it's easy to see that the dimension of the range of $\mathbf{S}$ is the number of non-zero singular values, while the dimension of the nullspace of $\mathbf{S}$ will be 5 minus the number of NON-ZERO singular values. Similar observations holds whenever $m \geq n$.

The remarks just made about the dimension of the range and nullspace of $\mathbf{S}$ also hold for $\mathbf{A}$. For example, if $\mathbf{x}' = \mathbf{V}^T\mathbf{x}$ is in the nullspace of $\mathbf{S}$ then, by definition, $\mathbf{S}\mathbf{x}' = \mathbf{0}$. But this is equivalent to $\mathbf{S}\mathbf{V}^T\mathbf{x} = \mathbf{0}$, or $\mathbf{U}\mathbf{S}\mathbf{V}^T\mathbf{x} = \mathbf{0}$, that is, $\mathbf{A}\mathbf{x} = \mathbf{0}$. In short, the nullspace of $\mathbf{A}$ is just the image of the

nullspace of $\mathbf{S}$ under the action of the orthogonal (hence invertible) matrix $\mathbf{V}$. Both nullspaces thus have the same dimension. Similar remarks hold for the dimensions of the range (see Problem 17).

In summary

1. The dimension of the range of $\mathbf{A}$ equals the number of non-zero singular values for $\mathbf{A}$.

2. The dimension of the nullspace of $\mathbf{A}$ equals the number of columns of $\mathbf{A}$ minus the number of non-zero singular values.

**Problem 16** Based on the SVD given in Problem 13, what is the dimension of the nullspace of the matrix $\mathbf{A}$ in that problem? What is the dimension of the nullspace of $\mathbf{A}$?

**Problem 17** Show that the range of $\mathbf{S}$ and $\mathbf{A}$ are of the same dimension. Hint: argue that $\mathbf{y}' = \mathbf{U}\mathbf{y}$ is an orthogonal transformation of the range of $\mathbf{A}$ to the range of $\mathbf{S}$.

**Problem 18** Suppose $\mathbf{A}$ is square and invertible with SVD $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$. Show that all of the singular values of $\mathbf{S}$ are nonzero, and write out the SVD of $\mathbf{A}^{-1}$ in terms of the matrices $\mathbf{U}, \mathbf{S}, \mathbf{V}$.

## 4.3 Solving $\mathbf{A}\mathbf{x} = \mathbf{b}$ with the SVD

Consider the problem of solving the equation $\mathbf{A}\mathbf{x} = \mathbf{b}$ for $\mathbf{x}$. From the SVD we can write this as

$$\mathbf{U}\mathbf{S}\mathbf{V}^T\mathbf{x} = \mathbf{b}. \tag{30}$$

Multiply both sides above by $\mathbf{U}^T$ on the left and as above let $\mathbf{x}' = \mathbf{V}^T\mathbf{x}$. We obtain

$$\mathbf{S}\mathbf{x}' = \mathbf{b}' \tag{31}$$

where $\mathbf{b}' = \mathbf{U}^T\mathbf{b}$. If $\mathbf{x}'$ that satisfies equation (31) then $\mathbf{x} = \mathbf{V}\mathbf{x}'$ will satisfy $\mathbf{A}\mathbf{x} = \mathbf{b}$, and conversely. Solutions to $\mathbf{A}\mathbf{x} = \mathbf{b}$ and (31) are thus in a one-to-one correspondence via $\mathbf{x} = \mathbf{V}\mathbf{x}'$. Also note that $\|\mathbf{x}\| = \|\mathbf{x}'\|$.

**Problem 19** Let

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}.$$

The SVD of $\mathbf{A}$ is given by

$$\mathbf{U} = \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} \sqrt{2} & 0 \\ 0 & \sqrt{2} \end{bmatrix}, \quad \mathbf{V}^T = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Use the procedure above (in particular, equation (31)) to solve $\mathbf{Ax} = [1, 2]^T$.

### 4.3.1 Consistent Equations

Consider first the case in which $m \leq n$. Equation (31) yields $m$ equations of the form $S_i x'_i = b'_i$, $1 \leq i \leq m$. If all of the $S_i > 0$ then the $x'_i$ are uniquely determined as $x'_i = b'_i/S_i$, while $x'_{m+1}, \ldots, x'_n$ are free variables. In this case there are infinitely many solutions to (31), except when $m = n$; then there are no free variables and the solution is uniquely determined. For the case $m < n$, among all solutions, the solution $\mathbf{x}'_0$ that has minimum norm is clearly obtained by take all free variables equal to zero,

$$\mathbf{x}'_0 = (b'_1/S_1, b'_2/S_2, \ldots, b'_m/S_m, 0, \ldots, 0)$$

($n - m$ zeros at the end.) This means that the minimum norm solution to $\mathbf{Ax} = \mathbf{b}$ is obtained as $\mathbf{x}_0 = \mathbf{V}\mathbf{x}'_0$.

Another way to write this is to first note that $\mathbf{x}'_0 = \mathbf{S}^\dagger \mathbf{b}'$ where $\mathbf{S}^\dagger$ is the $n \times m$ matrix

$$\mathbf{S}^\dagger = \begin{bmatrix} 1/S_1 & 0 & 0 \\ 0 & 1/S_2 & 0 \\ 0 & 0 & 1/S_3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

at least for the $m = 3, n = 5$ case, and assuming all $S_i > 0$. In general the matrix $\mathbf{S}^\dagger$ has dimensions $n \times m$. Now since $\mathbf{x}'_0 = \mathbf{S}^\dagger \mathbf{b}'$ as well as $\mathbf{b}' = \mathbf{U}^T \mathbf{b}$ and $\mathbf{x}_0 = \mathbf{V}\mathbf{x}'_0$ we see that the minimum norm solution to $\mathbf{Ax} = \mathbf{b}$ is given by

$$\mathbf{x} = \mathbf{A}^\dagger \mathbf{b} \tag{32}$$

where

$$\mathbf{A}^\dagger = \mathbf{V}\mathbf{S}^\dagger \mathbf{U}^T \tag{33}$$

The matrix $\mathbf{A}^\dagger$ is called the *Moore-Penrose pseudoinverse* of the matrix $\mathbf{A}$.

**Problem 20** Let $\mathbf{A}$ as in Problem 13. Find the Moore-Penrose pseudoinverse for $\mathbf{A}$, and use the pseudoinverse to find the minimum norm solution to $\mathbf{Ax} = b$ where $b$ is any scalar. Verify directly (using elementary calculus) that the resulting solution is the minimum norm solution.

**Problem 21** Suppose $\mathbf{A}$ is a square invertible matrix. Show that $\mathbf{A}^\dagger = \mathbf{A}^{-1}$.

**Problem 22** In the argument above we assumed that all $S_i > 0$. Suppose now that $\mathbf{A}$ has only $r$ non-zero singular values, so $S_{r+1} = S_{r+2} = \cdots S_m = 0$. Suppose also that all of $b'_{r+1}, \ldots, b'_m$ are zero (so the system (31) is consistent.) Show that the same solution procedure still works to find the minimum norm solution, if we modify $\mathbf{S}^\dagger$ by replacing entries $1/S_k$ with zero when $k \geq r+1$.

### 4.3.2 Inconsistent Equations

Generically, the case $m > n$ yields an inconsistent system of equations (though certainly not necessarily.) In any case, as above the system $\mathbf{Ax} = \mathbf{b}$ can be written in the form of (31) with $\mathbf{b}' = \mathbf{U}^T\mathbf{b}$ and $\mathbf{x}' = \mathbf{V}^T\mathbf{x}$. The matrix $\mathbf{S}$ will look something like that in equation (27). The system (31) is inconsistent unless it so happens that $b'_{n+1} = \cdots b'_m = 0$; if any $S_k = 0$ for $k \leq n$ then we also need the corresponding $b'_k = 0$.

Consider the case when the system is inconsistent, say with $m > n$ and non-zero singular values $S_1, \ldots, S_r$. We want to find a "best solution" to $\mathbf{Ax} = \mathbf{b}$ by finding a vector $\mathbf{x}$ that minimizes the quantity $\|\mathbf{Ax} - \mathbf{b}\|^2$. Note that

$$\mathbf{Ax} - \mathbf{b} = \mathbf{USV}^T\mathbf{x} - \mathbf{b} = \mathbf{U}(\mathbf{Sx}' - T\mathbf{b}') \tag{34}$$

where as previously, $\mathbf{x}' = \mathbf{V}^T\mathbf{x}$ and $\mathbf{b}' = \mathbf{U}^T\mathbf{b}$. Since $\mathbf{U}$ is orthogonal we conclude from equation (34) that

$$\|\mathbf{Ax} - \mathbf{b}\|^2 = \|\mathbf{Sx}' - \mathbf{b}'\|^2. \tag{35}$$

Based on equation (35) we see that a vector that minimizes $\|\mathbf{Ax} - \mathbf{b}\|^2$ can be obtained as $\mathbf{x} = \mathbf{Vx}'$ with $\mathbf{x}'$ as a minimizer of $\|\mathbf{Sx}' - \mathbf{b}'\|^2$.

Minimizing $\|\mathbf{Sx}' - \mathbf{b}'\|^2$ as a function of $\mathbf{x}'$ is easy. This quantity can be written explicitly as

$$Q(x'_1, \ldots, x'_n) = \sum_{j=1}^{r}(S_j x'_j - b'_j)^2 + \sum_{j=r+1}^{m}(b'_j)^2.$$

Clearly we can't do anything with $\mathbf{x}'$ to change the value of the second summation. The best we can do to minimize $Q$ is to take $x'_j = b'_j/S_j$ for all $j \leq r$ (so the first sum is zero), but the remaining $x'_j$ can be "anything". However, if we choose the remaining $x'_j = 0$ for $j > r$ we obtain that least-squares solution $\mathbf{x}'_0$ to $\mathbf{S}\mathbf{x}' = \mathbf{b}'$ that has the smallest possible norm. If we then take $\mathbf{x}_0 = \mathbf{V}\mathbf{x}'_0$ (so $\|\mathbf{x}_0\| = \|\mathbf{x}'_0\|$) we see that $\mathbf{x}_0$ will minimize $\|\mathbf{A}\mathbf{x}_0 - \mathbf{b}\|$, and if such a minimizer is not unique then $\mathbf{x}_0$ will be the minimizer with the smallest possible norm.

If you chase through the previous argument, you see that in fact the vector $\mathbf{x}'_0$ can be obtained as $\mathbf{x}'_0 = \mathbf{S}^\dagger\mathbf{b}'$, just as in the consistent case. By the argument given there we see that equation (32), in particular, the Moore-Penrose inverse, again provides the solution.

In summary we have shown (at least after you do Problem 23 below)

**Theorem 1** *If an $m \times n$ matrix $\mathbf{A}$ has singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ then the minimum-norm least-squares solution to $\mathbf{A}\mathbf{x} = \mathbf{b}$ is given by $\mathbf{x} = \mathbf{A}^\dagger\mathbf{b}$ where $\mathbf{A}^\dagger = \mathbf{V}\mathbf{S}^\dagger\mathbf{U}^T$. Here $\mathbf{S}^\dagger$ is the $n \times m$ matrix with entries $S^\dagger_{ij} = 0$ for $i \neq j$, $S^\dagger_{i,i} = 1/S_i$ if $S_i \neq 0$, and $S^\dagger_{i,i} = 0$ if $S_i = 0$.*

**Problem 23** Verify that the argument in this section works perfectly well if $m \leq n$.

**Problem 24** Suppose that

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \\ 1 & 0 \end{bmatrix}.$$

Verify that the SVD of $\mathbf{A}$ has factors

$$\mathbf{U} = \begin{bmatrix} -1/\sqrt{3} & -1/\sqrt{2} & -1/\sqrt{6} \\ 1/\sqrt{3} & -1/\sqrt{2} & 1/\sqrt{6} \\ -1/\sqrt{3} & 0 & 2/\sqrt{6} \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} \sqrt{3} & 0 \\ 0 & \sqrt{2} \\ 0 & 0 \end{bmatrix}, \quad \mathbf{V}^T = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Compute $\mathbf{A}^\dagger$. Use this to find the least-squares solution to $\mathbf{A}\mathbf{x} = \mathbf{b}$ with $\mathbf{b} = [1, 2, 4]^T$.

**Problem 25** Suppose an $m \times n$ matrix $\mathbf{A}$ has singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$. Verify that we can write the SVD of $\mathbf{A}$ in the alternate form

$$\mathbf{A} = \sum_{j=1}^{r} S_j \mathbf{u}_j \mathbf{v}_j^T \tag{36}$$

where $r$ is the number of non-zero singular values, $\mathbf{u}_j$ denotes the $j$th column of $\mathbf{U}$ (as an $m \times 1$ matrix) and similarly for $\mathbf{v}_j$ (so $\mathbf{v}_j^T$ is a $1 \times n$ matrix).

## 4.4 Stability and Regularization Using the SVD

### 4.4.1 Ill-Posed Linear Systems and the SVD

The Moore-Penrose pseudoinverse in equation (33) as computed via the SVD allows us to solve $\mathbf{Ax} = \mathbf{b}$ in a very general and flexible way. As it turns out, this procedure is also very useful even when $\mathbf{Ax} = \mathbf{b}$ has a unique solution (e.g., $\mathbf{A}$ is square, invertible), especially when $\mathbf{A}$ is very close to singular. This is often the case if the equation $\mathbf{Ax} = \mathbf{b}$ arises as the discretization of some inverse problem; often $\mathbf{A}$ is singular or nearly so, and the vector $\mathbf{b}$ is usually generated from data, which may be noisy. In this case small amounts of noise in $\mathbf{b}$ may yield huge changes in $\mathbf{x}$.

### 4.4.2 Ill-Posedness Example

Let

$$\mathbf{A} = \begin{bmatrix} 1.01 & 100 \\ 0.01 & 1 \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} 201.01 \\ 2.01 \end{bmatrix}.$$

(This matrix is nearly singular). The exact solution to $\mathbf{Ax} = \mathbf{b}$ is the vector $\mathbf{x}_0 = [1, 2]^T$. Suppose that we change $\mathbf{b}$ by a small amount, to simulate "noise" in the problem. We let

$$\mathbf{b}_e = \mathbf{b} + \begin{bmatrix} 0.01 \\ 0.02 \end{bmatrix} = \mathbf{b} + \mathbf{e} \tag{37}$$

(a change in the vector norm of about one part in 100), where $\mathbf{e} = [0.01, 0.02]^T$. In this case the solution to $\mathbf{Ax} = \mathbf{b}_e$ is vector $\mathbf{x}_e = [202, -0.03]^T$. The difference $\|\mathbf{x}_e - \mathbf{x}_0\| \approx 201.01$ is rather large, despite the fact that $\|\mathbf{e}\| \approx 0.0224$ is a rather small change in the right side.

The heart of the problem in this example is illustrated well by looking at

the SVD of $\mathbf{A}$, which has approximate factors

$$\begin{aligned}
\mathbf{U} &= \begin{bmatrix} -0.99995 & -0.0099995 \\ -0.0099995 & 0.99995 \end{bmatrix}, \\
\mathbf{S} &= \begin{bmatrix} 100.01 & 0 \\ 0 & 9.999 \times 10^{-5} \end{bmatrix}, \\
\mathbf{V}^T &= \begin{bmatrix} -0.0101 & -0.99995 \\ -0.99995 & 0.0101 \end{bmatrix}.
\end{aligned}$$

If we let $\mathbf{x}'_e = \mathbf{U}\mathbf{x}_e$, $\mathbf{x}'_0 = \mathbf{U}\mathbf{x}_0$, $\mathbf{b}' = \mathbf{V}\mathbf{b}$, and $\mathbf{e}' = \mathbf{V}\mathbf{e}$ then the equation $\mathbf{A}\mathbf{x}_e = \mathbf{b}_e$ can be written as

$$\mathbf{S}\mathbf{x}'_e = \mathbf{b}' + \mathbf{e}'. \tag{38}$$

Of course without the error term $\mathbf{e}'$ we obtain the exact solution for $\mathbf{x}'_0$ (and hence $\mathbf{x}_0$), with magnitude $\|\mathbf{x}'_0\| \approx 2.236$. Now $\|\mathbf{e}'\| = \|\mathbf{V}\mathbf{e}\| = \|\mathbf{e}\| \approx 0.0224$ is small in relation to $\mathbf{x}'_0$, but consider what happens to this error when we solve equation (38). We compute the second component $(x'_e)_2$ of $\mathbf{x}'_e$ as

$$(x'_e)_2 = \frac{b'_2 + e'_2}{10^{-4}}.$$

Though $e'_2$ is small (comparable to $\|\mathbf{e}'\| \approx 0.0224$) the division by $S_2 = 10^{-4}$ inflates this small error dramatically. The result is that $(x'_e)_2$ is way off, perhaps by as much as $10^4(0.0224) = 224.0$. Of course this error will be reflected in $\mathbf{x}_e = \mathbf{V}\mathbf{x}'_e$.

### 4.4.3 Ill-Posedness and Singular Values

Let's now consider the stability issue in a more general setting. The essence of the problem in the previous example is the disparity between the largest and smallest singular values. To see this, consider the equations

$$\begin{aligned}
\mathbf{S}\mathbf{x}'_0 &= \mathbf{b}' && (39) \\
\mathbf{S}\mathbf{x}'_e &= \mathbf{b}' + \mathbf{e}'. && (40)
\end{aligned}$$

in which the first equations contains the true "noiseless" right hand side vector $\mathbf{b}'$ from which we can obtain the "true" solution $\mathbf{x}'_0$; for simplicity assume $\mathbf{x}'_0$ is uniquely determined. In the second equation the right side

vector contains noise $\mathbf{e}'$, and this yields an erroneous solution $\mathbf{x}'_e$. We'd like to obtain a bound on the proportional error $(\|\mathbf{x}'_e - \mathbf{x}'_0\|)/\|\mathbf{x}'_0\|$ in terms of $\|\mathbf{b}\|$ and $\|\mathbf{e}\|$.

If we subtract (39) from (40) we obtain $\mathbf{S}(\mathbf{x}'_e - \mathbf{x}'_0) = \mathbf{e}'$. If $S_r$ denotes the smallest singular value in $\mathbf{S}$ then it's easy to see that $\|\mathbf{x}'_e - \mathbf{x}'_0\| \leq \|\mathbf{e}'\|/S_r$. From (39) it's easy to see that $\|\mathbf{x}'_0\| \geq \|\mathbf{b}'\|/S_1$. We thus have an upper bound

$$\frac{\|\mathbf{x}'_e - \mathbf{x}'_0\|}{\|\mathbf{x}'_0\|} \leq \frac{S_1}{S_r}\frac{\|\mathbf{e}'\|}{\|\mathbf{b}'\|}. \tag{41}$$

That is, the proportional error in $\mathbf{x}'_e$ is bounded by the ratio $S_1/S_r$ times the proportional error in the right side. If $S_1/S_r$ is small then the error in $\mathbf{x}'_e$ will be comparable to the error in $\mathbf{b}'$, on a proportional basis. If $S_1/S_r$ is large, e.g., as in the last example in which $S_1/S_2 \approx 10^6$, then the worst case error in the solution can be comparable to this ratio times the proportional error in $\mathbf{b}'$. All of these error estimates hold for the original problem $\mathbf{Ax} = \mathbf{b}$ and its true and approximate solutions $\mathbf{x}_0$ and $\mathbf{x}_e$, since $\mathbf{U}$ and $\mathbf{V}$ are orthogonal.

The ratio $S_1/S_r$ is called the *condition number* of the matrix $\mathbf{A}$. If the vector $\mathbf{x}$ in the system $\mathbf{Ax} = \mathbf{b}$ is uniquely determined but the ratio $S_1/S_r$ is large (here "large" depends on the context) then the linear system is said to be *ill-conditioned*. Thus, for example, if $S_1/S_r \approx 10^{15}$, even round-off error in the vector $\mathbf{b}$ (in double precision arithmetic, 15 to 16 significant figures) may render the solution $\mathbf{x}$ meaningless. In the example of subsection 4.4.2 above the condition number of that matrix was around $10^6$; the error of order $10^{-2}$ in the right side $\mathbf{b}$ was inflated dramatically, but not quite as much as the worst case amount of $10^{-2} \times 10^6 = 10^4$. Nonetheless, the resulting error in the right side completely destroys our ability to accurately recover $\mathbf{x}$.

### 4.4.4 Regularization with the SVD

The singular value decomposition can be used to combat ill-conditioned systems, by allowing us to solve the linear system in a controllably regularized fashion.

Consider the alternate version of the SVD in equation (36) of Problem 25. Since the Moore-Penrose pseudoinverse for $\mathbf{A}$ is given by $\mathbf{A}^\dagger = \mathbf{V}\mathbf{S}^\dagger\mathbf{U}^T$ we see from equation (36) that

$$\mathbf{A}^\dagger = \mathbf{V}\mathbf{S}^\dagger\mathbf{U}^T = \sum_{j=1}^{r} \frac{1}{S_j}\mathbf{v}_j\mathbf{u}_j^T \tag{42}$$

where the sum is over only the non-zero $S_j$, which we assume are indexed by $1 \leq j \leq r$.

When we solve $\mathbf{Ax} = \mathbf{b}$ we obtain $\mathbf{x} = \sum_{j=1}^{r} \frac{1}{S_j} \mathbf{v}_j \mathbf{u}_j^T \mathbf{b}$, which can be written as

$$\mathbf{x} = \sum_{j=1}^{r} c_j \mathbf{v}_j \tag{43}$$

where $c_j = \mathbf{u}_j^T \mathbf{b}/S_j$. The solution $\mathbf{x}$ is thus synthesized as a linear combination of the vectors $\mathbf{v}_j$, the columns of $\mathbf{V}$. However, as discussed above, if a singular value $S_k$ is very close to zero and $\mathbf{b}$ contains any error then the solution $\mathbf{x}$ may be way off. What can we do about this? One method is to omit from the sum on the right (43) any term for which the singular value $S_j$ is too small, to prevent a large and highly erroneous multiple of $\mathbf{v}_j$ from being incorporated into the estimate of $\mathbf{x}$. Here "too small" depends on the size of $\mathbf{x}$ and the amount of error in $\mathbf{b}$.

Let's quantify all of this. Let $\mathbf{x}_0$ be the true solution to $\mathbf{Ax} = \mathbf{b}$ when no noise is present and let $\mathbf{x}_e$ be the solution when the right side is $\mathbf{b} + \mathbf{e}$ for some error vector $\mathbf{e}$. Then $\mathbf{Ax}_0 = \mathbf{b}$ and $\mathbf{Ax}_e = \mathbf{b} + \mathbf{e}$. In light of equation (43) we conclude

$$\mathbf{x}_0 = \sum_{j=1}^{r} c_j \mathbf{v}_j \text{ and } \mathbf{x}_e = \sum_{j=1}^{r} \tilde{c}_j \mathbf{v}_j \tag{44}$$

where $c_j = \mathbf{u}_j^T \mathbf{b}/S_j$ and $\tilde{c}_j = \mathbf{u}_j^T (\mathbf{b} + \mathbf{e})/S_j$. Subtract the above equations to find

$$\mathbf{x}_e - \mathbf{x}_0 = \sum_{j=1}^{r} (\tilde{c}_j - c_j) \mathbf{v}_j.$$

Since the $\mathbf{v}_j$ form an orthonormal set we have

$$\|\mathbf{x}_e - \mathbf{x}_0\|^2 = \sum_{j=1}^{r} (\tilde{c}_j - c_j)^2.$$

Use $\tilde{c}_j - c_j = \mathbf{u}_j^T \mathbf{e}/S_j$ to conclude that

$$\|\mathbf{x}_e - \mathbf{x}_0\|^2 = \sum_{j=1}^{r} (\mathbf{u}_j^T \mathbf{e}/S_j)^2. \tag{45}$$

Now in most applications we'd have SOME idea of $\|\mathbf{x}_0\|$; let $X_0$ denote this value. Let's suppose we're willing to tolerate a proportional error of no

more than a fraction $q$ in our estimate of $\mathbf{x}_0$, i.e., an absolute error of size $qX_0$. We'd also likely have some idea of the magnitude of the error in our data vector $\mathbf{b}$; if each $b_k$ might be off by an amount at most $\sigma$ (that is, if each $e_k$ satisfies $|e_k| \leq \sigma$) then we can estimate $\|\mathbf{e}\| \leq \sigma\sqrt{m}$. Finally, note that $\mathbf{u}_j^T\mathbf{e} \leq \|\mathbf{u}_j\|\|\mathbf{e}\| = \sigma\sqrt{m}$ (since each $\mathbf{u}_j$ is a unit vector.) We can use this with equation (45) to regularize the solution of $\mathbf{Ax} = \mathbf{b}$ as follows: Replace $\mathbf{u}_j^T\mathbf{e}$ on the right in (45) with the upper bound $\sigma\sqrt{m}$ to obtain

$$\|\mathbf{x}_e - \mathbf{x}_0\|^2 \leq m\sigma^2 \sum_{j=1}^{r} \frac{1}{S_j^2}. \tag{46}$$

Since we want to be sure that $\|\mathbf{x}_e - \mathbf{x}_0\|^2 \leq q^2X_0^2$ let's try to enforce the inequality

$$m\sigma^2 \sum_{j=1}^{p} \frac{1}{S_j^2} \leq q^2X_0^2 \tag{47}$$

by taking the upper summation limit $p$ as large as possible while maintaining the inequality. (It's possible it fails even when $p = 1$, if the data is noisy enough or the problem sufficiently ill-conditioned.) The idea is that by omitting the really small singular values we omit those terms in (43) that could inflate the error in $\mathbf{b}$ enough to violate $\|\mathbf{x}_e - \mathbf{x}_0\| \leq qX_0$. The maximum value for $p$ in (47) depends on $m$ (the number of equations) but more importantly on the error level $\sigma$ (bigger $\sigma$ clearly causes $p$ to decrease), as well as $q$ and $X_0$. And of course the singular values play an essential role. Note that in the noiseless case $\sigma = 0$ we can take $p = r$, all the singular values, and get a perfect solution.

### 4.4.5 Regularization Example

Consider the problem of solving a first kind integral equation of the form

$$\int_a^b K(s,t)u(s)\,ds = f(t) \tag{48}$$

for an unknown function $u$ defined on the interval $[a, b]$, where the kernel function $K(s, t)$ is known for $a \leq s, t \leq b$ and the right side function $f(t)$ is known for $a \leq t \leq b$. In particular, let's consider the specific case $K(s, t) = e^{-st}$ and $f(t) = \frac{2\pi(1-e^{-t})}{t^2+4\pi^2}$, in which case the true solution is $u(s) = \sin(2\pi s)$

(but pretend we don't know that.) We thus want to find (or approximate) a function $u(s)$ that satisfies

$$\int_0^1 e^{-st} u(s)\, ds = \frac{2\pi(1 - e^{-t})}{t^2 + 4\pi^2}. \tag{49}$$

We discretize the problem and turn it into a finite set of linear equations as follows. First, we'll approximate $u$ with a piecewise constant function $\tilde{u}$ of the form $\tilde{u}(s) = u_j$ if $(j-1)/N \le s < j/N$, where $1 \le j \le N$. In this case equation (48) becomes

$$\sum_{j=1}^{N} A_j(t) u_j = f(t) \tag{50}$$

where

$$A_j(t) = \int_{(j-1)/N}^{j/N} e^{-st}\, dt = \frac{e^{-jt/N}(e^{t/N} - 1)}{t}.$$

We'll now substitute in each of the values $t = 1/2N, 3/2N, 5/2N, \ldots, (2N - 1)/2N$ into equation (50) to obtain a system of $N$ linear equations in $N$ unknowns, of the form

$$\sum_{j=1}^{N} A_{jk} u_j = b_k \tag{51}$$

where $A_{jk} = A_j((2k-1)/2N)$ and $b_k = f((2k-1)/2N)$. We now solve the $N \times N$ linear system (51) to obtain $u_1, \ldots, u_N$, in which $u_k$ is considered to be an estimate of $u((2k-1)/2N)$.

The graph on the left in Figure 16 shows the result with $N = 4$ when we solve the system (51) using traditional unregularized methods. The choice $N = 4$ gives a rather crude, low resolution estimate of $u$. The natural thing to do to improve the resolution is to increase $N$. On the right in Figure 16 is the estimate recovered using $N = 10$ with no regularization. This is pretty clearly a disaster. The problem is that when $N = 10$ the singular values for the $10 \times 10$ matrix $\mathbf{A}$ in equation (51) are

$$0.809, 0.0641, 2.05 \times 10^{-3}, 4.00 \times 10^{-5}, 5.33 \times 10^{-7},$$
$$5.05 \times 10^{-9}, 3.52 \times 10^{-11}, 7.48 \times 10^{-12}, 2.36 \times 10^{-12}, 1.46 \times 10^{-13}.$$

The ratio of the largest to smallest is $S_1/S_{10} \approx 5.54 \times 10^{12}$. Since the arithmetic was carried out using 10 significant figures, it's not surprising that the results are poor.

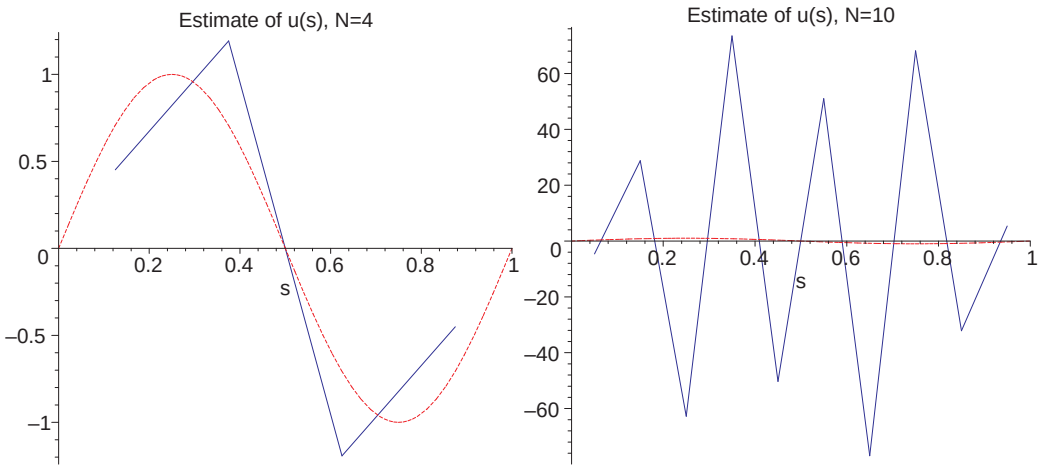


Figure 16: True solution (red, dashed) and recovered estimate (blue, solid), $N = 4$.

Consider now what happens if we regularize the solution to (51) using the ideas above. We'll assume the true solution has magnitude of about $X_0 = 1$ (anything in the ballpark will do), that $\sigma = 10^{-11}$ (the values $f((2k-1)/2N)$ are around 0.1, so 10 significant figures corresponds to error of about $10^{-11}$). Finally, let's take $q = 0.1$ for the allowable fractional error. In this case the largest value of $p$ we can take in (47) is $p = 6$. We estimate the solution using equation (43) but with upper summation limit 6. The result is shown on the left in Figure 17. The result for $N = 20$ is shown on the right, where again we have $p = 6$.

### 4.4.6 Another Approach to Regularization with the SVD

Let's return to the problem of solving $\mathbf{Ax} = \mathbf{b}$ in which the right side vector $\mathbf{b}$ is corrupted by noise. Let $\mathbf{x}_0$ be the true solution i.e., $\mathbf{Ax}_0 = \mathbf{b}$, and let $\mathbf{x}_e$ be the solution to the noisy problem $\mathbf{Ax}_e = \mathbf{b} + \mathbf{e}$. Then $\mathbf{x}_0$ and $\mathbf{x}_e$ are given by equation (44) with $c_j = \mathbf{u}_j^T\mathbf{b}/S_j$ and $\tilde{c}_j = \mathbf{u}_j^T(\mathbf{b} + \mathbf{e})/S_j$. However, as we've seen, when $S_j$ is close to zero it's likely that $\tilde{c}_j$ is far from the correct value of $c_j$. We're going to regularize the solution in the presence of noise by instead taking

$$\tilde{c}_j = \frac{\mathbf{u}_j^T(\mathbf{b} + \mathbf{e})}{\tilde{S}_j} \tag{52}$$

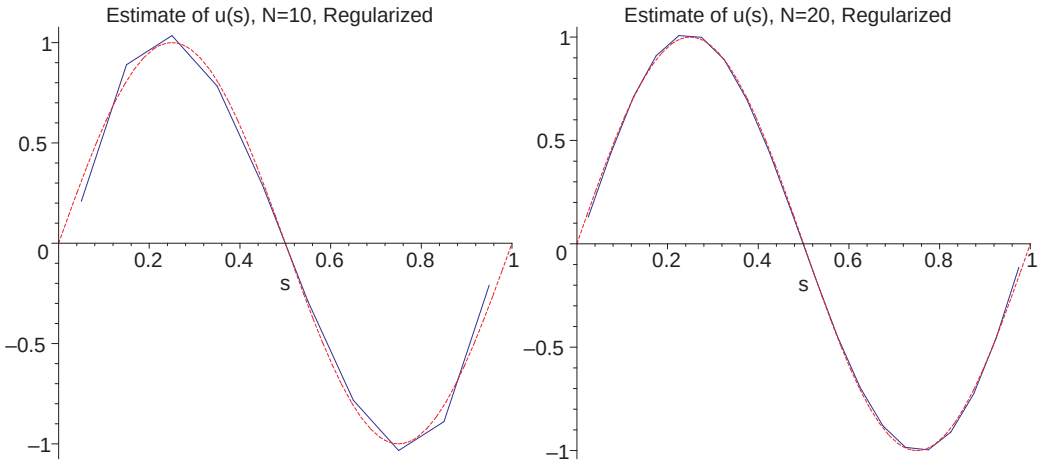


Figure 17: True solution (red, dashed) and recovered estimate (blue, solid), $N = 10$ and $N = 20$, regularized.

in equation (44), in which the singular values $S_j$ have been altered; we'll use all $r$ terms, however. This is slightly different than what we did in Section 4.4.4, in which the sum for $\mathbf{x}_e$ in equation (44) was simply truncated when the singular values got too small.

How should we alter the singular values? For simplicity we'll assume that the components of the vector $\mathbf{e}$ are independent normal random variables with zero mean and variance $\sigma^2$. In this case the error $\|\mathbf{x}_e - \mathbf{x}_0\|^2$ is itself a random variable whose distribution depends on the choice of $\tilde{S}_j$. We'll choose the $\tilde{S}_j$ to minimize $E(\|\mathbf{x}_e - \mathbf{x}_0\|^2)$, the expected mean square error in the reconstruction.

It's easy to see that since the $\mathbf{v}_j$ are orthonormal vectors we have

$$\begin{aligned}
\|\mathbf{x}_e - \mathbf{x}_0\|^2 &= \sum_{j=1}^{r} (\tilde{c}_j - c_j)^2 \\
&= \sum_{j=1}^{r} \left( \frac{\mathbf{u}_j^T \mathbf{b}}{\tilde{S}_j} - \frac{\mathbf{u}_j^T \mathbf{b}}{S_j} \right)^2 + \sum_{j=1}^{r} \left( \frac{\mathbf{u}_j^T \mathbf{e}}{\tilde{S}_j} \right)^2 \\
&+ \sum_{j \neq k}^{r} \left( \frac{\mathbf{u}_j^T \mathbf{b}}{\tilde{S}_j} - \frac{\mathbf{u}_j^T \mathbf{b}}{S_j} \right) \left( \frac{\mathbf{u}_k^T \mathbf{e}}{\tilde{S}_k} \right). \qquad (53)
\end{aligned}$$

Because the $e_k$ are independent with zero mean and variance $\sigma^2$ it's easy to

compute

$$E(\mathbf{u}_k^T\mathbf{e}) = 0 \text{ and } E((\mathbf{u}_k^T\mathbf{e})^2) = \sigma^2 \sum_{j=1}^{m} (\mathbf{u}_k)_j^2 = \sigma^2$$

since the $\mathbf{u}_k$ are unit vectors. If we take the expected value of both sides of equation (53) we then obtain

$$E(\|\mathbf{x}_e - \mathbf{x}_0\|^2) = \sum_{j=1}^{r} \left[ \left( \frac{\mathbf{u}_j^T\mathbf{b}}{\tilde{S}_j} - \frac{\mathbf{u}_j^T\mathbf{b}}{S_j} \right)^2 + \frac{\sigma^2}{\tilde{S}_j^2} \right]. \tag{54}$$

In particular, the cross terms from (53) drop out. Minimizing $E(\|\mathbf{x}_e - \mathbf{x}_0\|^2)$ as a function of the $\tilde{S}_j$ is easy: simply differentiate both sides of equation (54) with respect to each $\tilde{S}_j$ to obtain

$$\frac{\partial}{\partial \tilde{S}_j} \left[ \left( \frac{\mathbf{u}_j^T\mathbf{b}}{\tilde{S}_j} - \frac{\mathbf{u}_j^T\mathbf{b}}{S_j} \right)^2 + \frac{\sigma^2}{\tilde{S}_j^2} \right] = \frac{2}{\tilde{S}_j^2} \left( \frac{(\mathbf{u}_j^T\mathbf{b})^2}{S_j} - \frac{(\mathbf{u}_j^T\mathbf{b})^2}{\tilde{S}_j} - \frac{\sigma^2}{\tilde{S}_j} \right) = 0.$$

A little algebra shows that the optimal choice for $\tilde{S}_j$ is

$$\tilde{S}_j = \frac{(\mathbf{u}_j^T\mathbf{b})^2 + \sigma^2}{(\mathbf{u}_j^T\mathbf{b})^2} S_j. \tag{55}$$

Of course with no noise ($\sigma = 0$) we obtain $\tilde{S}_j = S_j$. If it so happens that $\mathbf{u}_j^T\mathbf{b} = 0$ then we could take "$\tilde{S}_j = \infty$," that is, omit the corresponding terms in (44), since then $\tilde{c}_j = 0$.

From a practical standpoint, note that when we actually seek a regularized solution to $\mathbf{Ax} = \mathbf{b}$ we almost always have some reasonable idea of the noise level $\sigma$. Also, we can compute $\mathbf{u}_j^T\mathbf{b}$ (or really, $\mathbf{u}_j^T(\mathbf{b} + \mathbf{e})$ as a stand in) when we do the SVD. As a result, we can compute the $\tilde{S}_j$ in equation (55).

## 4.5 Example: Gravitational Prospecting Again

Let's revisit the gravitational prospecting problem of Section 2.3.2 using the technique of Section 4.4.6. In particular, our tentative $\lambda$ will be represented as in equation (14), which we use in equation (13) to form a least-squares function $Q(a_0, \ldots, a_{10})$. We form the normal equations $\partial Q/\partial a_j = 0$, a system of 11 equations in 11 unknowns. The result obtained by simply solving this

system without regularization was shown on the right in Figure 10. In this case the $11 \times 11$ matrix $\mathbf{A}$ that governs the system has largest singular value $S_1 \approx 5.443$ and smallest singular value $S_{11} \approx 2.79 \times 10^{-15}$, for a condition number of about $1.95 \times 10^{15}$, pretty bad. The computation of the right side vector $\mathbf{b}$ was carried out to 10 digit precision, so we shouldn't be surprised to see this error magnified to values in the range $10^3$ to $10^4$, much larger than the solution itself.

However, if we apply the technique of Section 4.4.6 with estimated noise level $\sigma = 10^{-10}$ the results are much better, as illustrated on the left in Figure 18. On the right in Figure 18 is the reconstruction obtained from data to which independent normal random noise with standard deviation $10^{-4}$ has been added. The reconstruction was done using $\sigma = 10^{-4}$ in equation (55).

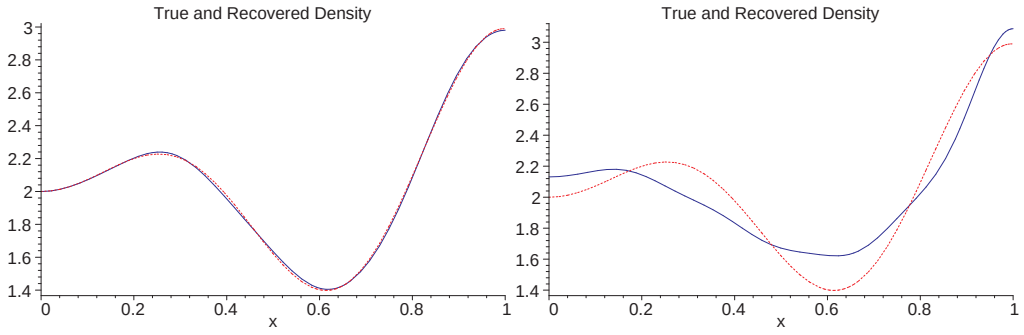


Figure 18: True (red, dashed) and recovered (blue, solid) densities, for noiseless and noisy data.

# 5 The Radon Transform and CT Scans

## 5.1 Introduction

The first practical CT ("computed tomography", or "computed axial tomography" for the "CAT" acronym) scanners were developed in the late 1960's. This technology allows physicians to create detailed images of two-dimensional cross sections of a human body, non-invasively. Moreover, modern computer power allows such images to be amalgamated into detailed 3D images that can be manipulated in real-time, and provide a powerful diagnostic tool in medicine. The mathematical ideas behind CT scans date back to

the early part of the 20th century and were developed by Johann Radon. It's yet another example of mathematics that greatly pre-dated its incorporation into technology.

## 5.2 Attenuation of x-rays

A CT scan works by passing x-rays through the body at many different angles and positions. The situation is depicted in Figure 19. Depending on the

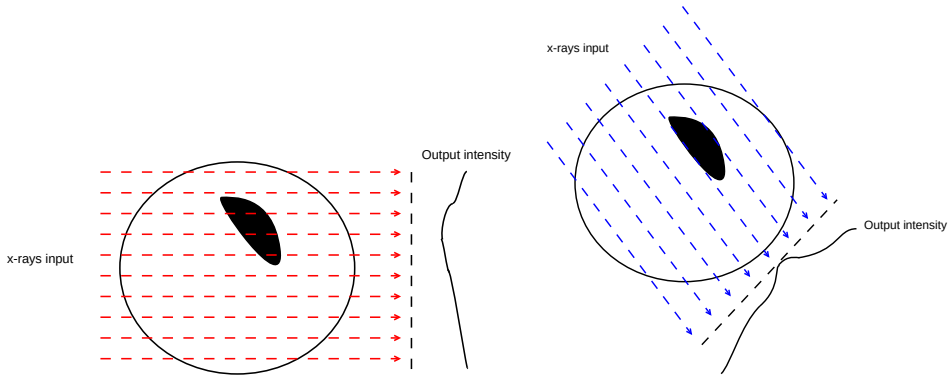


Figure 19: X-rays passing through at various angles, offsets.

tissues and length of path encountered, the beam is attenuated (i.e., dimmed or diminished) more or less. By measuring the attenuation of the beam for "all" angles and positions, we seek to form an image of the interior of the body by using the fact that different tissues attenuate x-rays differently.

Let's begin with a simple model of how x-rays are attenuated as they pass through the body. Let an x-ray move along some line parameterized by $\mathbf{L}(s)$ for $a \leq s \leq b$. We'll assume the parameterization is with respect to arc length, so $|\mathbf{L}'(st)| = 1$. The standard model for the attenuation of the x-ray is

$$I'(s) = -\lambda(\mathbf{L}(s))I(s) \tag{56}$$

where $\lambda \geq 0$ is a function of position and is called the *attenuation coefficient* of the tissue, while $I(s) \geq 0$ denotes the intensity of the x-ray. Equation (56) posits that the x-ray is attenuated in proportion to its current intensity, with the constant of proportionality dependent on the type of tissue through which the beam is passing at that point.

If the intensity of the x-ray at $s = a$ is $I_a$ then the solution to (56) is

$$I(s) = I_a \exp\left(-\int_a^s \lambda(\mathbf{L}(t))\, dt\right). \tag{57}$$

Let us parameterize the line so that $\mathbf{L}(a)$ corresponds to a point outside the body, before the x-ray enters, and we'll assume the attenuation in the air is zero. Let $s = b$ correspond to a point after which the x-ray has emerged and $I_b = I(b)$ denote the intensity of the x-ray after it emerges from the body. If we plug $s = b$ into equation (57) and solve for the value of the integral we obtain

$$-\ln(I_b/I_a) = \int_a^b \lambda(\mathbf{L}(s))\, ds \tag{58}$$

If we know the "input" $I_a$ and measure the "output" $I_b$, we can recover the line integral of the attenuation function $\lambda$ (the right side of (58)) over the line $\mathbf{L}$ through the body.

**Problem 26**

a. Suppose $\lambda \equiv 0$. What is the solution $I(t)$ to (56) with initial condition $I(a) = I_a$? Why does that make perfect sense?

b. Suppose $I(b) = I(a)$ (with $a < b$). What can we conclude about $\lambda$ on the line between the points $\mathbf{L}(a)$ and $\mathbf{L}(b)$, especially given that $\lambda \geq 0$?

## 5.3 The Radon Transform

### 5.3.1 Geometry

For convenience, let's assume that the cross-section of the body we seek to image is contained in the unit disk $D$ in $\mathbb{R}^2$ (but the cross section need not BE the unit disk!) Refer to Figure 20 below. Let $\mathbf{L}(s)$ parameterize a line through $D$ with unit normal vector $\mathbf{n} =< \cos(\theta), \sin(\theta) >$; note $\mathbf{n}$ lies at an angle $\theta$ with respect to horizontal. Then the vector $< -\sin(\theta), \cos(\theta) >$ (which is just $\mathbf{n}$ rotated 90 degrees counter-clockwise) is parallel to the line. We'll take $\theta$ to lie in the range $0 \leq \theta < \pi$. For any given $\theta$ in this range that point $\mathbf{p}$ on the line closest to the origin is of the form $\mathbf{p} = r\mathbf{n} = r < \cos(\theta), \sin(\theta) >$ where $-1 < r < 1$ (if the line goes through the disk $D$). This line $\mathbf{L}(s)$ can

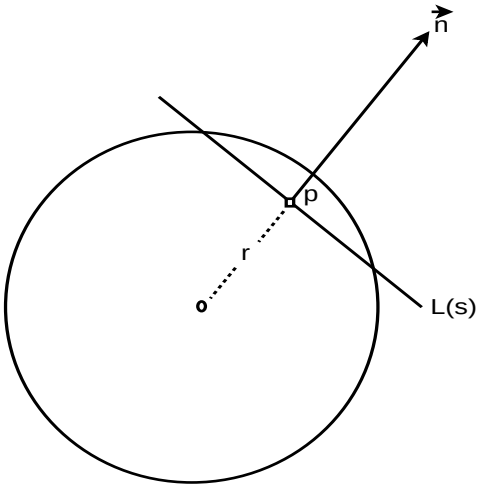


Figure 20: Geometry for CT scan computations.

thus be parameterized as

$$\mathbf{L}(s) = r < \cos(\theta), \sin(\theta) > + s < \sin(\theta), -\cos(\theta) > . \tag{59}$$

In non-parametric form the line is given by $\mathbf{n} \cdot < x, y >= r$, or more explicitly,

$$\cos(\theta)x + \sin(\theta)y = r. \tag{60}$$

Indeed, it's easy to see that there's a one-to-one correspondence between lines through $D$ and parameterizations of the form (59) with $-1 \leq r < 1$, $0 \leq \theta < \pi$.

With this geometry and notation we can take integration limits $a = -\sqrt{1-r^2}$ and $b = \sqrt{1-r^2}$ in equations (57)/(58).

### 5.3.2 The Forward Transform and Sinogram

To perform a CT scan, mathematically anyway, we fire x-rays through $D$ for all $r \in (-1, 1)$ and $\theta \in [0, \pi)$. For each such x-ray we measure or control the input intensity $I_a$ and measure the output intensity $I_b$. From this information and equation (58) we can recover the integral

$$d(\theta, r) = \int_{\mathbf{L}_{r,\theta}} \lambda \, ds = \int_{-\sqrt{1-r^2}}^{\sqrt{1-r^2}} \lambda(r\cos(\theta) + s\sin(\theta), r\sin(\theta) - s\cos(\theta)) \, ds \tag{61}$$

where $\mathbf{L}_{r,\theta}$ denotes the line parameterized by (59) and $ds$ is arc length. Equation (61) defines the *Radon Transform* in two dimensions. This transform takes a function $\lambda(x, y)$ of two variables and turns it into a new function $d(\theta, r)$ in two variables.

**Problem 27**

a. Suppose $\lambda(x, y) = 1 + x^2$. Use equation (61) to compute $d(\theta, r)$ explicitly.

b. Suppose the unit disk $D$ has zero attenuation everywhere except for a circular "block of lead" of radius $R < 1$ centered at the origin (that completely blocks the x-rays). What $\lambda$ would model this situation? Find a formula for $d(\theta, r)$ in this case.

Of course when we perform a CT scan we're really implementing a Radon Transform physically. The goal in CT imaging is to invert the transform, that is, use $d(r, \theta)$ for $r \in (-1, 1)$, $\theta \in [0, \pi)$ to reconstruct the function $\lambda(x, y)$.

Here's an illustration. On the left in Figure 21 is a synthetic "target" for our CT scan. The default attenuation in the disk is 0, while the darker areas indicate a higher attenuation (up to a maximum of 1). Even though the only area of interest is the interior of the disk $D$, I colored the outside of $D$ black solely for contrast. On the right is a plot of $d(\theta, r)$ ($\theta$ down the vertical axis, $r$ across.) This plot is called the *sinogram* for the target on the left. The lighter areas indicate a larger value for the line integral over the line corresponding to the $(\theta, r)$ pair. If you exam the sinogram closely you can "see" the higher attenuation regions.

But does the data in the sinogram (or the function $d(\theta, r)$) uniquely determine the attenuation function $\lambda(x, y)$? If so, how do we quantitatively invert the Radon transform to construct an actual image?

## 5.4 Inverting the Radon Transform I

Let's start by looking at an easy special case for inverting the Radon Transform. Specifically, we'll assume that the attenuation function $\lambda$ is a function of radius only, that is, $\lambda(x, y) = \lambda(\sqrt{x^2 + y^2})$, so $\lambda$ is effectively a function of a single variable. It's then easy to see that $d(\theta, r)$ in equation (61) will not

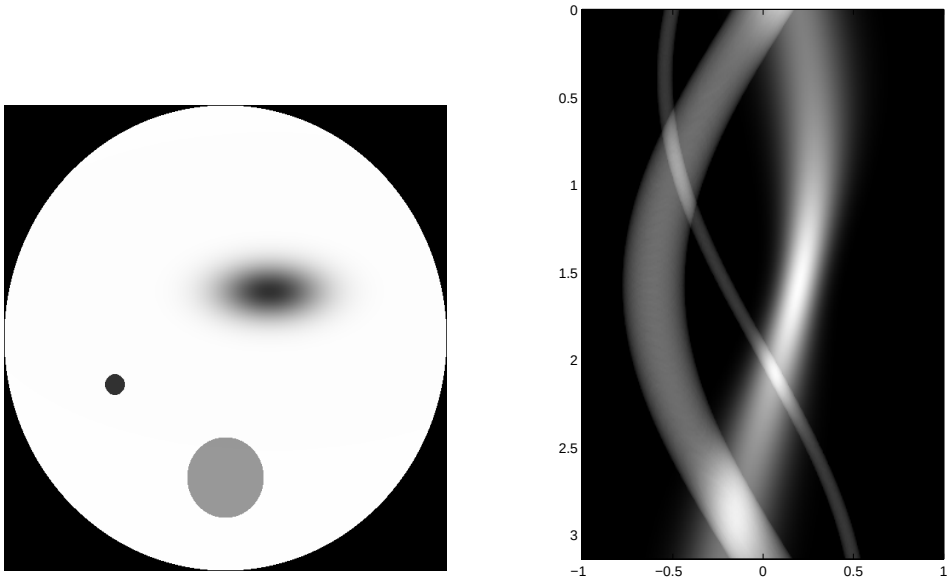


Figure 21: CT target (left) and sinogram (right).

depend on $\theta$ (think about it geometrically). In this case equation (61) can be written

$$d(r) = \int_{-\sqrt{1-r^2}}^{\sqrt{1-r^2}} \lambda(\sqrt{r^2+s^2})\, ds \tag{62}$$

since $(-r\sin(\theta) + s\cos(\theta))^2 + (r\cos(\theta) + s\sin(\theta))^2 = r^2 + s^2$. Our goal is to recover the function $\lambda$ on the interval $[0, 1]$ from knowledge of the quantity $d(r)$ for $0 \leq r \leq 1$.

First, symmetry makes it clear that (62) can be written as

$$d(r) = 2\int_{0}^{\sqrt{1-r^2}} \lambda(\sqrt{r^2+s^2})\, ds. \tag{63}$$

Make a substitution $u = \sqrt{r^2+s^2}$, so that $s = \sqrt{u^2-r^2}$, $ds = u/\sqrt{u^2-r^2}\, du$. Equation (63) becomes

$$d(r) = 2\int_{r}^{1} \frac{u\lambda(u)}{\sqrt{u^2-r^2}}\, du \tag{64}$$

if we adjust the limits in the integral appropriately. Write equation (64) in the form

$$d(r) = 2\int_r^1 \frac{u\lambda(u)}{\sqrt{(1-r^2)-(1-u^2)}}\,du. \tag{65}$$

Now define $z = 1 - r^2$ (so $r = \sqrt{1-z}$) and substitute $t = 1 - u^2$ (so $u = \sqrt{1-t}, du = -\frac{1}{2\sqrt{1-t}}\,dt$.) Equation (65) becomes

$$d(\sqrt{1-z}) = \int_0^z \frac{\lambda(\sqrt{1-t})}{\sqrt{z-t}}\,dt \tag{66}$$

again, after changing the limits in the integral and replacing all $r$'s on the left and right with $\sqrt{1-z}$.

Finally, define functions

$$g(z) = d(\sqrt{1-z}) \text{ and } \phi(t) = \lambda(\sqrt{1-t}) \tag{67}$$

(note if we know $g$ we know $d$ and vice-versa, ditto for $\lambda$ and $\phi$. Equation (66) becomes a first kind Volterra integral equation,

$$\int_0^z \frac{\phi(t)}{\sqrt{z-t}}\,dt = g(z) \tag{68}$$

for $0 \le z \le 1$. In fact, equation (68) is called *Abel's equation.*

As it turns out, Abel's equation has a simple closed form solution! You can verify directly that

$$\phi(t) = \frac{1}{\pi}\frac{d}{dt}\left(\int_0^t \frac{g(z)\,dz}{\sqrt{t-z}}\right). \tag{69}$$

In summary then, in the radial case we can in principle recover the attenuation function $\lambda(r)$ from the function $d(r)$ as follows:

1. Let $g(z) = d(\sqrt{1-z})$ for $0 \le z \le 1$.
2. Use equation (68) to compute the function $\phi(t)$ for $0 \le t \le 1$.
3. Recover $\lambda$ as $\lambda(r) = \phi(1-r^2)$.

## 5.5 Inversion Example

Let's use $\lambda(z) = 2 - z + \sin(4z)$ as an example target attenuation function, and pretend we've measured the quantity $d(r)$ at $n = 10$ offsets of the form $r_k = k/n$ for $0 \leq k \leq n-1$. The data (computed by evaluating the right side of equation (62) numerically) is given in the table below, to three significant figures.

| $r$ | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|---|---|---|---|
| $d(r)$ | 3.827 | 3.843 | 3.797 | 3.624 | 3.297 | 2.822 | 2.234 | 1.593 | 0.972 | 0.445 |

We can also use the value $d(1) = 0$, a given. But to use equation (69) what we really need are the values of $g(z) = d(\sqrt{1-z})$, which yields the data below (just use $r = \sqrt{1-z}$, so $z = 1 - r^2$ to transform the independent variable):

| $z$ | 1.0 | 0.99 | 0.96 | 0.91 | 0.84 | 0.75 | 0.64 | 0.51 | 0.36 | 0.19 |
|---|---|---|---|---|---|---|---|---|---|---|
| $g(z)$ | 3.827 | 3.843 | 3.797 | 3.624 | 3.297 | 2.822 | 2.234 | 1.593 | 0.972 | 0.445 |

We also have $g(0) = 0$. Our goal is to use this data to estimate $\phi$ with equation (69).

One simple approach is to assume that $g$ is piecewise linear, or perhaps a cubic spline, in between the $z$ values listed above. We plug such a $g$ into equation (69), evaluate the integral, differentiate to obtain $\phi(t)$, then change variables back to $\lambda$ as $\lambda(u) = \phi(1 - u^2)$. The result of interpolating the data for $g(z)$ above with a piecewise linear function and carrying out this procedure is shown in Figure 22.

**Problem 28** Suppose the attenuation is radial and given by $\lambda(t) = 1 + t^2/3$.

a. Compute the sinogram data $d(r)$ using equation (62), in closed form.

b. Find the function $g(z)$ explicitly using equation (67).

c. Use formula (69) to compute $\phi(t)$, then use (67) to find $\lambda$. Of course, you should get back what you started with.

**Problem 29** Here's an alternate way to estimate $\lambda$ in the example above, using equation (68) directly. Assume that $\phi(t) = a_0 + a_1 t + a_2 t^2$, a quadratic

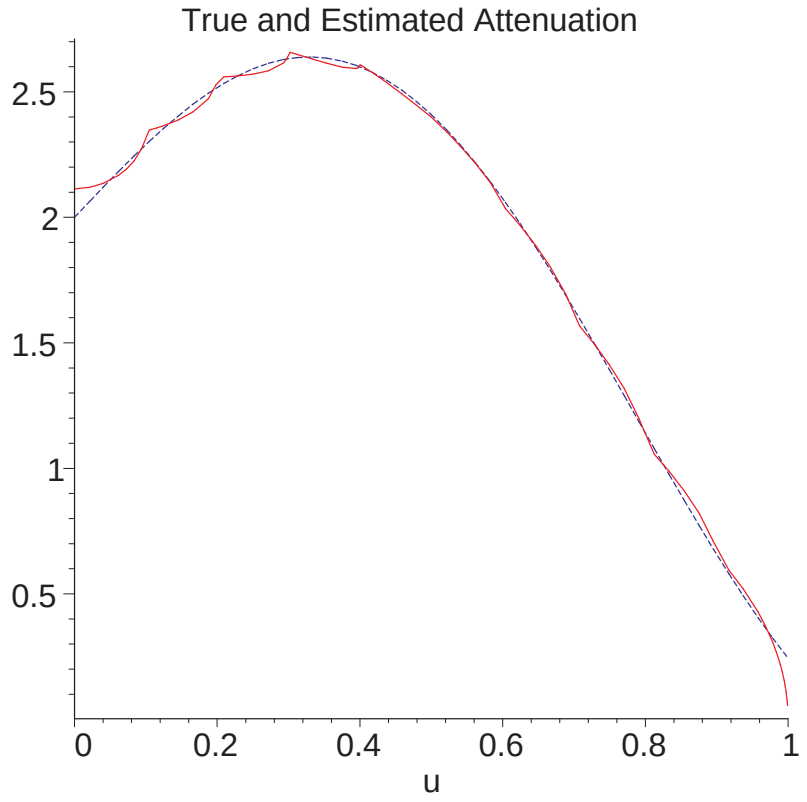


Figure 22: True (blue, dashed) and estimated (red, solid) attenuation functions.

polynomial (you can go higher if you like). Plug this choice for $\phi(t)$ into equation (68) and evaluate the integral in terms of $a_0, a_1, a_2$, and $z$; indeed, let

$$\tilde{g}(a_0, a_1, a_2, z) = \int_0^z \frac{a_0 + a_1 t + a_2 t^2}{\sqrt{t - z}}\, dt.$$

In computing $\tilde{g}$ you might find is useful to note that

$$\int_0^z \frac{t^k}{\sqrt{t - z}}\, dt = c_k z^{k+1/2}$$

with $c_0 = 2, c_1 = 4/3, c_2 = 16/15$. Then choose the $a_k$ to minimize the least-squares functional

$$Q(a_0, a_1, a_2) = \sum_{j=1}^{10} (\tilde{g}(a_0, a_1, a_2, z_j) - g_j)^2$$

a quadratic optimization problem, where the $z_j, g_j$ are from the tabled values for $g$ above. Finally, form the estimate $\lambda(u) = \tilde{g}(a_0, a_1, a_2, 1 - u^2)$ and compare to the true $\lambda$.

## 5.6 Inverting the Radon Transform II

Let's start with a very simple example that leads to an intuitive idea for inverting the Radon Transform, as defined by equation (61). Consider the target as illustrated below on the left in Figure 23, consisting of a zero attenuation background in the unit disk with a single circular "inclusion" of higher attenuation. The sinogram is shown on the right.

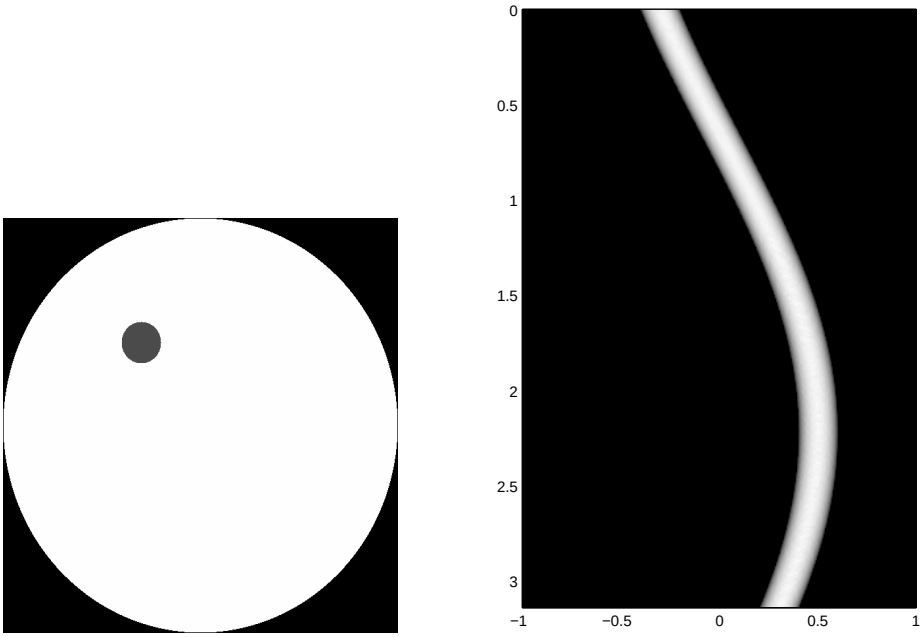


Figure 23: Sample CT target and sinogram.

Forget for a moment about the fact that we know where the inclusion is, and instead think about what we'd see if we fired x-rays through the unit disk. Each x-ray that passes through this inclusion is attenuated more than comparable x-rays that don't pass through the inclusion. The result is

that the corresponding line integrals of the attenuation function $\lambda$ are larger (refer back to equation (58)) whenever the line passes through this area of high attenuation, smaller when the line passes through region of smaller attenuation. The situation in illustrated in Figure 24, where the (solid) red lines pass through the high-attenuation inclusion while the (dashed) green lines do not. We expect the line integrals for the solid red lines to be larger than those for the green dashed lines.

This suggests a way to invert the Radon transform, at least approximately: For any give point $(x_0, y_0)$ in the disk, estimate $\lambda(x_0, y_0)$ by doing some kind of average of all the line integrals corresponding to lines that pass through $(x_0, y_0)$. That is, average the value of $d(\theta, r)$ over all $\theta, r$ pairs that correspond to a line that passes through $(x_0, y_0)$. The intuitive idea is that the line integrals for lines that pass through high attenuation regions will be larger, and so will have a larger average. Conversely, the line integrals for lines that pass through low attenuation regions will be smaller, and so will have a smaller average.

From equation (60) it's easy to see that for any given angle $\theta$ the line with normal vector $\mathbf{n} =< \cos(\theta), \sin(\theta) >$ through the point $(x_0, y_0)$ has $r$ value

$$r = \cos(\theta)x_0 + \sin(\theta)y_0. \tag{70}$$

Averaging $d(\theta, r)$ over all such $\theta, r$ pairs gives

$$\tilde{\lambda}(x_0, y_0) = \frac{1}{\pi}\int_0^\pi d(\theta, \cos(\theta)x_0 + \sin(\theta)y_0)d\,\theta. \tag{71}$$

I call the integral on the right in (71) "$\tilde{\lambda}(x_0, y_0)$" because it's an estimate of the value of $\lambda$ at $(x_0, y_0)$.

The procedure above—in particular, the integral on the right in (71)—is called the *back projection* of $d(\theta, r)$. This terminology stems from another interpretation of the operation above: For any given choice of $\theta$ and $r$ we back project ("smear") the value of $d(\theta, r)$ uniformly over the line $x\cos(\theta) + y\sin(\theta) = r$. We repeat this for all lines through the disk, overlaying each smeared value on top of the others. The result is that we build up an image that's an estimate of the true attenuation function $\lambda$.

**Example 3** We'll use the attenuation function on the left in Figure 23, and the accompanying data for the sinogram on the right. The picture on the left in Figure 25 shows what we obtain if, at each point $(x_0, y_0)$ in the disk,

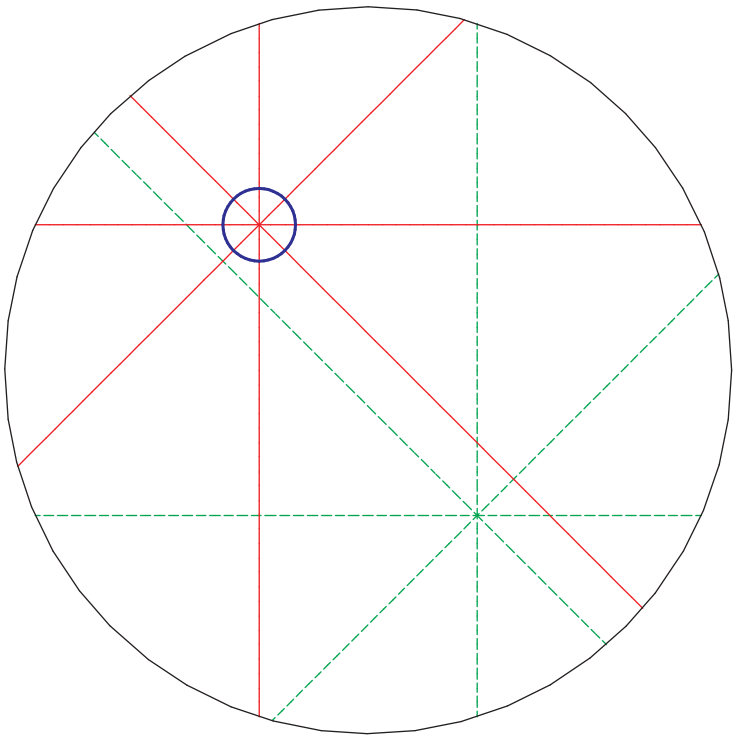

Figure 24: Solid (red) lines pass through high attenuation inclusion (blue), while dashed (green) lines do not.

choose 10 lines through the point at angle $\theta = 0, \pi/10, 2\pi/10, \ldots, 9\pi/10$, compute the corresponding value of $r$ from equation (70), then average the corresponding values of $d(\theta, r)$ (a discrete approximation to the back projection integral.) The picture on the left in Figure 25 shows the result for 20 integrals. Figure 26 shows the results for 50 and 300 line integrals through each point.

Back projection seems very promising, but it looks like the final image is a slightly blurred version of the original attenuation function (it doesn't get better after 300) . This is in fact correct—back projection provides only an approximate inversion for the Radon transform. The result is a slightly blurred version of the original.

The exact inverse for the Radon transform is obtained by applying a "high-pass" or "ramp" filter to the sinogram $d(\theta, r)$, in the $r$ variable, prior to performing back projection. It's a very simple operation, quite akin to

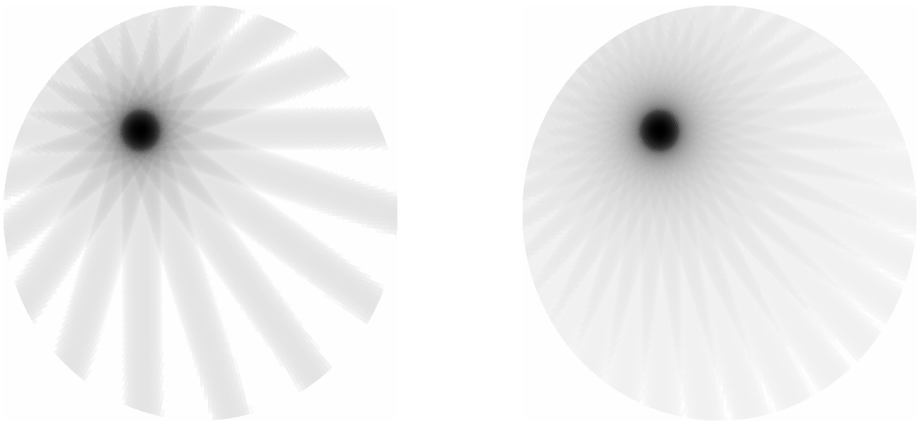

Figure 25: Reconstruction from 10 line integrals (left) and 20 (right).

differentiating $d(\theta, r)$ with respect to $r$, but the proof that this then yields an exact inverse requires the use of the Fourier integral transform, a bit beyond the scope of the present discussion. However, this is the standard procedure for inverting the Radon transform and is called, not surprisingly, *filtered back projection*.

To illustrate, in Figure 27 we show on the left the result of back projecting the sinogram from Figure 21, with no filtering. On the right is the filtered back projected image.

As a last example, on the left in Figure 28 is an attenuation function, my version of a standard simulated CT target called the *Shepp-Logan phantom* (it's supposed to look vaguely anatomical). On the right is the sinogram for the target on the left. Figure 29 shows the unfiltered and filtered back projected reconstructions.

**Problem 30** Let $\lambda(x, y) = 1$ in the unit disk.

a. Compute $d(\theta, r)$ explicitly.

b. Use equation (71) to back project your $d(\theta, r)$ from part (a) and compute $\tilde{\lambda}(x_0, y_0)$. To make things easier, note that since $\lambda$ is radial, $d$ will depend only on $r$. Thus you may as well stick to computing $\tilde{\lambda}(x_0, 0)$. But a computer algebra system is still helpful!

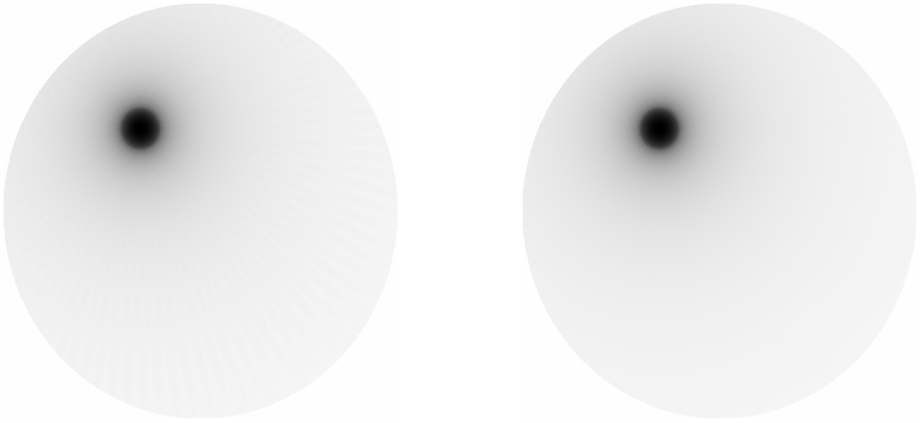

Figure 26: Reconstruction from 50 line integrals (left) and 300 (right).

c. If back projection were the exact inverse we'd hope in this case that $\tilde{\lambda}(x_0, 0)$ would equal $\lambda(x_0, 0)$, or at least be a constant multiple of $\lambda(x_0, 0)$. Show that this is not the case (i.e., $\tilde{\lambda}$ isn't constant).

**Problem 31** Based on the above discussion, where do you think ill-posedness might pop up in inverting a Radon transform?

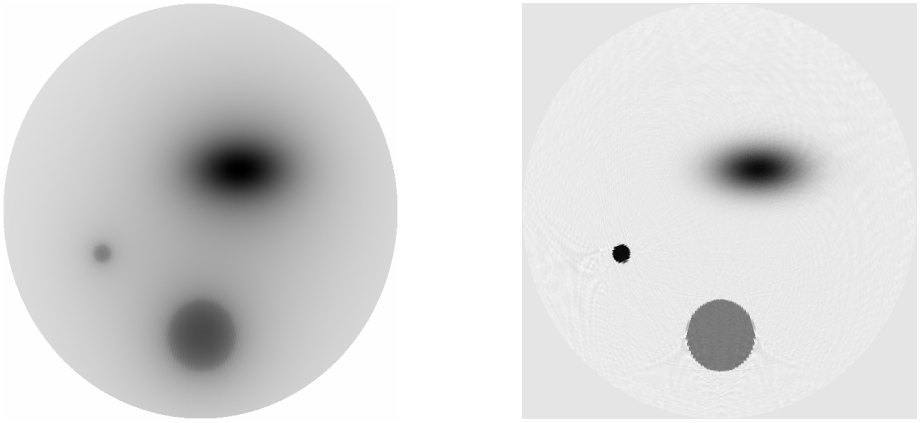

Figure 27: Back projection image (left) and filtered back projection (right) for sinogram in Figure 21.

# References


[1] C.W. Groetsch, Inverse Problems in the Mathematical Sciences, Viewig Verlag, Wiesbaden/Braunschweig, 1993.

[2] C.W. Groetsch, Inverse Problems: Activities for Undergraduates, Mathematical Association of America, Washington D.C., 1999.

[3] L.N. Trefethen and D. Bau, Numerical Linear Algebra, SIAM, Philadelphia, 1997.

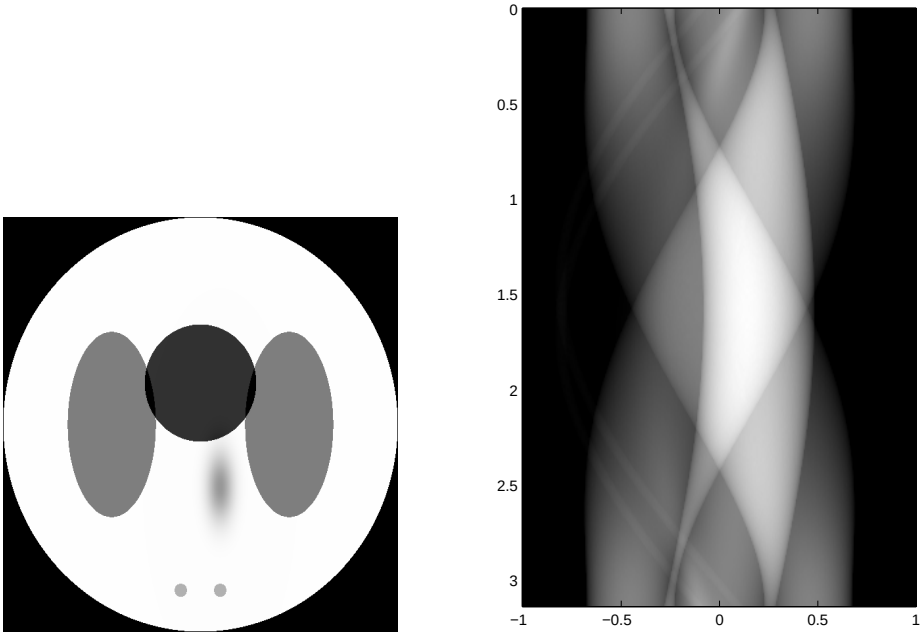


Figure 28: Original target and sinogram.

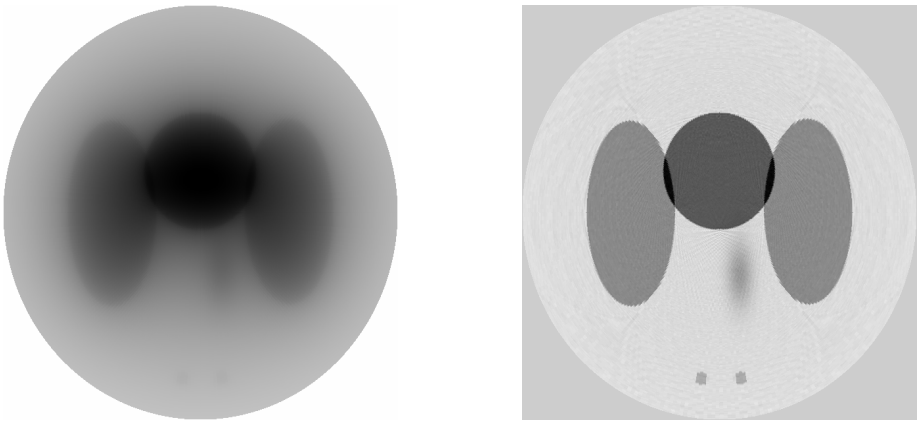

Figure 29: Unfiltered back projection and filtered back projection.