Neural Network Interest Group

Título/*Title*:
Regularization. An Introduction

Autor(es)/*Author(s)*:
J.P. Marques de Sá

Relatório Técnico/*Technical Report* No. 2 /2007

FEUP/INEB, Rua Dr. Roberto Frias, 4200-465, Porto, Portugal

Título/*Title*:

Regularization. An Introduction

Autor(es)/*Author(s)*:

J.P. Marques de Sá

Relatório Técnico/*Technical Report* No. 2 /2007

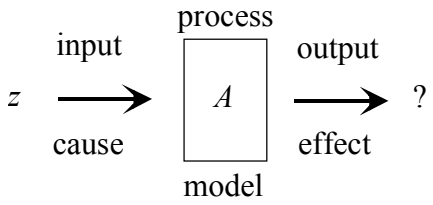
Publicado por/*Published by*: NNIG. http://paginas.fe.up.pt/~nnig/

Junho de 2007

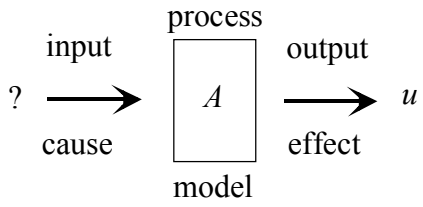
Contents

1 Inverse Problems

Direct Problem: $u = Az$



Inverse Problem: $z = A^{-1}u$



A can be a function, a functional or an operator (see Appendix).
Note that strictly speaking there is no distinction between direct and inverse problems. (Historically, inverse problems appear after the direct problems.)

Example 1.1

Multiplying two numbers is a direct problem. Factoring a number into two numbers is the corresponding inverse problem.

Example 1.2

Finding out where a projectile fired by a canon falls, given the angular *position* and muzzle *velocity* is the direct problem. Finding out which angular position, θ, and velocity, v, are possible so that the projectile falls at a certain point (the range R) is the inverse problem.
If θ and v are free to vary there are infinitely many solutions. If v is fixed there is a unique solution given by:

$$R(\theta) = \frac{v^2}{g}\sin 2\theta$$

Example 1.3

The computation of a centroid function given a density function, say in [0, 1], is a direct problem whose unique solution is:

$$C(u) = \frac{\int_o^u zf(z)dz}{\int_0^u f(z)dz}$$

Even a function not physically realizable, such as $f(z) = z^{-1/3}$ has a well-defined $C(u)$.
Moreover, the solution is *stable*:
Suppose $\{f_n\}$ is a sequence of densities converging uniformly to f:

$$\forall z \in [0,1] \quad \lim_{n\to\infty} |f(z) - f_n(z)| = 0$$

The sequence of centroids $\{C_n(u)\}$ will then converge to $C(u)$, for each $u \in]0,1]$.
The inverse problem is that of determining a density function given the centroid function. The solution:

$$f(z) = B(z)A(z) \ ,$$

where $B(z) = \dfrac{C'(z)}{z - C(z)}$ and $A(z) = \exp\int_1^z B(u)du$, is the mass of the density function, is *not unique* and is *not guaranteed to be stable*. For instance, $C(u) = u/2$ is the centroid function of $f(z) = 1$. Now, take for $n = 3,4,5,\ldots$

$$C_n(u) = \frac{u}{2} + \frac{1}{n}u^{n^2}$$

The uniform convergence of $\{C_n(u)\}$ holds true:

$$|C_n(u) - C(u)| = \frac{1}{n}u^{n^2} \underset{n\to\infty}{\to} 0$$

Now let us compute the solutions with the mass $A(z) = 1$:

$$f_n(z) = B_n(z) = \frac{C_n'(z)}{z - C_n(z)} = \frac{\frac{1}{2} + (n^2 - 1)z^{n^2-1}}{z\left(\frac{1}{2} - \frac{1}{n}z^{n^2-1}\right)}$$

It is now easily seen that $f_n(1) \underset{n\to\infty}{\to} \infty$ and hence $\{f_n\}$ does not converge to f.

The bottom line is: Inverse problems are generally harder to solve than direct problems, often with no unique and no stable solutions.

2 Supervised Learning Problems

A supervised learning machine observes n pairs $(x_1, t_1), \ldots, (x_n, t_n)$ i.i.d. with joint distribution $P(x, t) = P(x)P(t \mid x)$ and issues an appropriate function from a set of functions $\{y = g(x)\}$[1] hopefully minimizing the *risk functional*:

$$R(g(x)) = \int L(t, g(x))dP(x,t)$$

$L(.)$ is the *loss function* measuring how well the machine performs (imitates the supervisor) for a given pair (x_i, y_i).

$$R(g(x)) \textbf{ is the mathematical expectation of } L(t, g(x)).$$

The learning problem is the problem of finding an adequate function from a given set. We frequently deal with a parametrized set:

$$\{y_w = g(x; w)\}$$

The adequacy is assessed by *minimizing a risk functional*:

$$\text{Find } g \text{ such that } Ag \equiv \int L(g)dF = \min R.$$

The supervised learning problem is an inverse problem.

2.1 Pattern recognition (data classification) problem

The t (target) values are discrete values (labels).

The loss function for this setting is an indicator function:

$$L(t, y_w) = \begin{cases} 0 & \text{if} \quad y_w = t \\ 1 & \text{if} \quad y_w \neq t \end{cases}$$

With this loss function: $R(w) = \int L(t, y_w(x; w))dP(t, x) = P^{(w)}(\text{error})$

$$\min_w R(w) = P_{opt}^{(w)}(\text{error})$$

[1] For convenience we omit a possible dependency of g on t.

2.2 Regression problem

The t values are continuous values and there are stochastic dependencies between x and t describable by $P(t \mid x)$.

We are not interested in determining (estimating) $P(t \mid x)$. Only on the conditional mathematical expectation, the so-called regression function:

$$r(x) = \mathrm{E}[T \mid X] = \int t \, dP(t \mid x)$$

The machine issues $f(x,\alpha) \in \{f(x,\alpha)\}$.
If $\int t^2 dP(t,x) < \infty$ and $\int r^2(x) dP(t,x) < \infty$ (bounded second moments), the minimum of

$$R(w) = \int (t - f(x;w))^2 \, dP(t,x) \,,$$

if it exists, can be proved to be attained at $r(x)$ provided $r(x) \in \{f(x;w)\}$; otherwise is attained at the function closest to $r(x)$ in the metric L_2.
(See tutorial MLE, MSE et alia; see Vapnik, 1998.)

2.3 Density estimation problem

On the basis of empirical data $x_1, \ldots, x_l$, find a function $p(x; w_0) \in \{ p(x; w)\}$ such that

$$p(x; w_0) = \frac{dP(x)}{dx}$$

Consider the functional

$$R(w) = -\int \ln p(x; w) dP(x) = -\int p(x; w_0) \ln p(x; w) dx$$

It can be shown that:

1. The minimum of this functional (if it exists) is attained at the functions that may differ from $p(x;w_0)$ only on a set of zero measure.
2. The Bretagnolle-Huber inequality holds:

$$\int \left| p(x; w) - p(x; w_0) \right| dx \le 2\sqrt{1 - \exp\{R(w_0) - R(w)\}}$$

To the previous risk functional one can add the constant

$$\int p(x; w_0) \ln p(x; w_0) dx$$

and minimize the following risk functional:

$$R^*(w) = -\int p(x;w_0)\ln\frac{p(x;w)}{p(x;w_0)}dx,$$

the Kullback-Leibler divergence.

3 Well-Posed and Ill-Posed Problems

3.1 Definitions

Consider the solution $z = A^{-1}u = R(u)$ to an "initial" data u. Assume $z \in F$ and $u \in U$ with metrics $\rho_F(z_1, z_2)$ and $\rho_U(u_1, u_2)$.
The solution is said to be *stable* if:

$$\forall \varepsilon > 0, \exists \delta(\varepsilon) > 0, \quad \rho_U(u_1, u_2) \le \delta(\varepsilon) \Rightarrow \rho_F(z_1, z_2) \le \varepsilon$$

with $u_1, u_2 \in U, \quad z_1 = R(u_1), z_2 = R(u_2) \in F$. Note that this definition of stability is less strict than the previous one.

Definition: The problem of determining the solution z in the space F from the initial data u in the space U is said to be **well-posed (in the Hadamard sense) on the pair of metric spaces** (F, U) if:

1. *Existence*. There is a solution: $\forall u \in U, \quad \exists z = R(u) \in F$
2. *Uniqueness*. The solution is unique.
3. *Continuity (or stability)*. The problem is stable in the spaces (F, U).

Problems that do not satisfy any of the above conditions (that is, at least one of the above requirements is violated) are said to be **ill-posed**. Problems that satisfy 1 and 2, but do not satisfy 3 are said to be **ill-conditioned**.

3.2 Examples of Ill-Posed Problems

3.2.1 The Projectile Problem

The Example 1.2 problem with fixed v is well-posed. We have:

$$R(\theta) = \frac{v^2}{g}\sin 2\theta \text{ with } R(\theta) \in \Re^+, \theta \in [0, \pi/2]$$

Denoting $\alpha = v^2 / g$:

$$\left|R_1 - R_2\right| = \left|2\alpha\cos(\theta_1+\theta_2)\sin(\theta_1-\theta_2)\right| \le 2\alpha\sin|\,\theta_1-\theta_2\,| < 2\alpha\,|\,\theta_1-\theta_2\,|$$

Therefore $\forall\varepsilon > 0$ we choose $\delta(\varepsilon) = 2\alpha\varepsilon$.
Since the problem has a unique solution and is stable it is then well-posed.

3.2.2 The Centroid Problem

The centroid problem of Example 2.2 is ill-posed. The solution is not unique and is often not stable.

3.2.3 An ML Density-Estimation Problem

Consider the following density estimation problem. We are given ℓ points $x_1,\ldots,x_l$, of an unknown distribution that we want to model by the Gaussian mixture[2]:

$$p(x,a,\sigma) = \frac{1}{2\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right) + \frac{1}{2\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$$

Let the mean of the first distribution be set at one of the ℓ points, say x_1. The problem is to determine σ by a maximum likelihood (ML) procedure. That is, determine σ such that:

$$L(\sigma) = \sum_{i=1}^{l} \ln p(x_i, a = x_1, \sigma) \text{ is maximized}$$

However:

$$L(\sigma) > \ln\left(\frac{1}{2\sigma\sqrt{2\pi}}\right) + \sum_{i=2}^{l}\ln\left(\frac{1}{2\sqrt{2\pi}}e^{-x_i^2/2}\right) = -\ln\sigma - \text{const}$$

But the maximum of $L(\sigma)$ always occurs at $\sigma = 0$ and the ML method does not provide a solution. The problem is ill-posed.

3.2.4 The Convolution Problem

Consider the Fredholm equation/operator of type I, with *continuous* kernel $K(x, s)$ in $[a, b]^2$:

$$Az = u \quad \equiv \quad \int_a^b K(x,s)z(s)ds = u(x)$$

The Fredholm operator is linear,

$$\int_a^b K(t,x)\big(f(t)+g(t)\big)dt = F(x) + G(x),$$

[2] This example is based on a similar one in (Vapnik, 2000)

and continuous functions $z(s)$ in $[a, b]$ are mapped onto continuous functions $u(x)$ in $[a, b]$. A special case of this operator is the convolution operator:

$$g(x) = K(x) \otimes f(x) \equiv \int_{-\infty}^{\infty} K(x-s) f(s) ds$$

The inverse problem is the one of finding the function z which convoluted with the kernel K yields the function u: $z = A^{-1}u$ (e.g., deconvolution). Unfortunately the inverse problem is ill-posed (see the Appendix on the continuity of A^{-1}).
We illustrate with the special case:

$$\int_0^1 K(x,s) z(s) ds = u(x) \text{ with continuous kernel and } x, s \in [0,1].$$

Take the continuous function $u_\omega(x)$:

$$u_\omega(x) = \int_0^1 K(x,s) \sin(\omega s)\, ds \qquad \text{with the property} \qquad u_\omega(x) \underset{\omega \to \infty}{\to} 0$$

Now consider the integral equation:

$$\int_0^1 K(x,s) z^*(s) ds = u(x) + u_\omega(x)$$

Since the equation is linear, the solution is

$$z^*(s) = z(s) + \sin(\omega s)$$

But this solution is unstable. For sufficiently large ω, $u(x)$ and $u(x) + u_\omega(x)$ differ in the Euclidian norm by:

$$\left\{ \int_0^1 \left[\int_0^1 K(x,s) \sin(\omega s)\, ds \right]^2 \right\}^{1/2},$$

which for sufficiently large ω can be made arbitrarily small. However, in the Chebychev norm:

$$\rho_F(z_1, z_2) = \max_{[0,1]} \left| z(s) - z^*(s) \right| = 1$$

If we use the Euclidian norm for ρ_F the solution is also unstable.

A graphical illustration:

The Fredholm equation is the convolution with the [0,1] rectangular window.

Function $z(s)$ is also the [0,1] rectangular window.

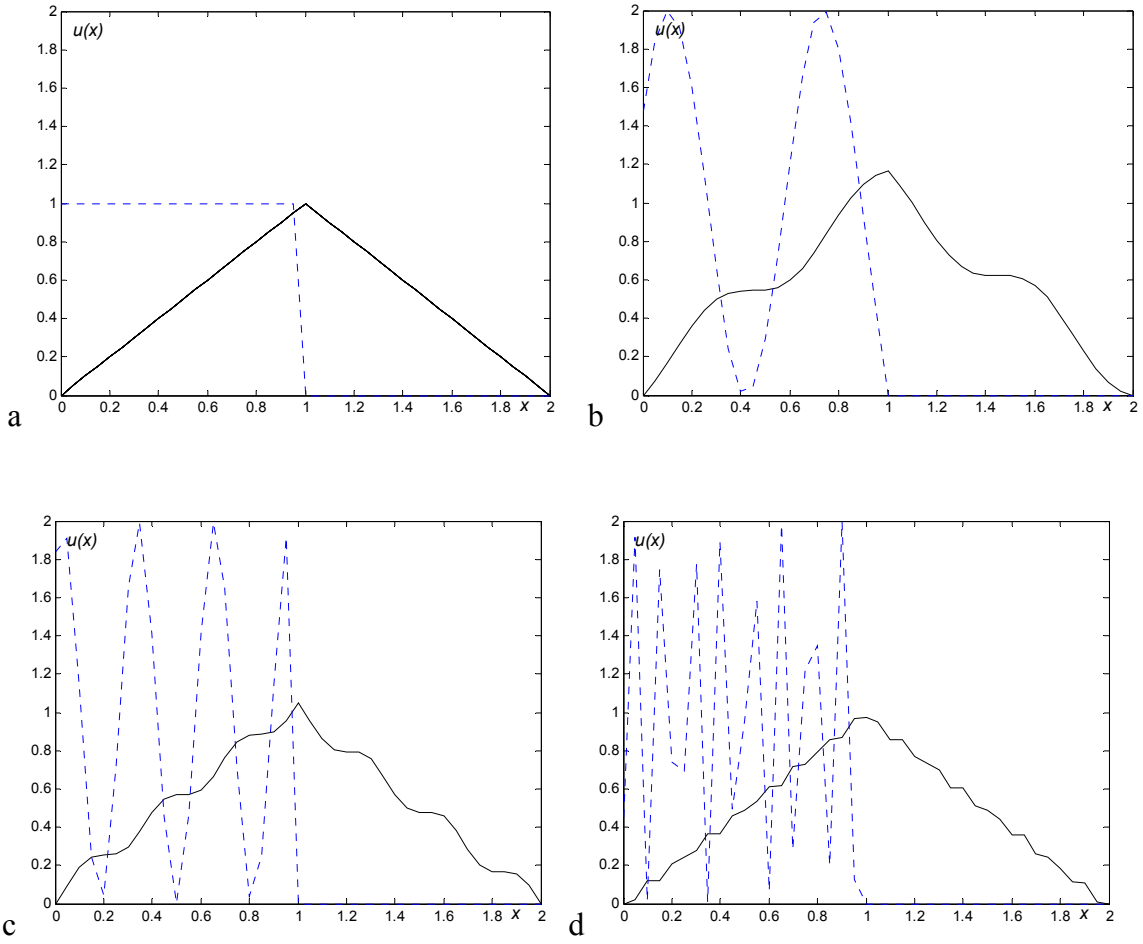


Fig. 3.1. Convolution (black) of a rectangular function with superimposed sinusoid (blue), defined in [0,1] using the rectangular kernel [0,1]: a) $\omega = 0$; b) $\omega = 10$; c) $\omega = 20$; d) $\omega = 200$. The jagged contour is due to the finite resolution.

3.2.5 The Derivative Estimation Problem

The problem of estimating derivatives is also ill-posed in general.

The problem is formulated as follows: given the measurements of a smooth function $F(x)$ at n points in [0,1] find an estimate of the derivative $f(x)$ of $F(x)$ at x.

The solution corresponds to solving the Volterra integral:

$$\int_0^x f(t)dt = F(x) - F(0)$$

or equivalently, the Fredholm integral:

$$\int_0^1 \theta(x-t) f(t)dt = F(x) - F(0) \text{ with } \theta(u) = \begin{cases} 1 & \text{if } u > 0 \\ 0 & \text{otherwise} \end{cases}$$

Special case: $F(x)$ is a monotonically increasing function satisfying $F(0) = 0$; $F(1) = 1$: density estimation.
We have seen already that the Fredholm problem is ill-posed in general.

3.2.6 The Fourier Series Problem

Consider the (inverse) problem of determining the coefficients of a Fourier series:

$$f_1(t) = \sum_{n=0}^{\infty} a_n \cos nt$$

Suppose that instead of a_n we compute $c_0 = a_0$ and $c_n = a_n + \varepsilon$ for $n \geq 1$. We get:
$f_2(t) = \sum_{n=0}^{\infty} c_n \cos nt$. Now, in the l_2 metric the difference of the coefficients

$$\left\{ \sum_{n=0}^{\infty} (c_n - a_n)^2 \right\}^{1/2} = \varepsilon \left\{ \sum_{n=1}^{\infty} \frac{1}{n^2} \right\}^{1/2} = \varepsilon \sqrt{\pi^2 / 6}$$

can be made as small as we wish, whereas the difference

$$| f_2(t) - f_1(t) | = | \varepsilon \sum_{n=1}^{\infty} \frac{1}{n} \cos nt |$$

is arbitrarily large.
However, if we take the difference between functions in the L_2 metric the problem is well-posed (Parseval's theorem).
Therefore, we have here an illustration that a problem may be well-posed using some metrics and ill-posed using other metrics.

3.2.7 The Linear Equation System Problem

Consider A to be a non-singular square matrix. The inverse problem corresponding to solving the system of linear equations $z = A^{-1}u$ is ill-conditioned whenever the matrix A *condition number* $k(A) = \|A\| \|A^{-1}\|$ is large, where $\|A\|$ is a matrix norm with submultiplicative property ($\|AB\| \leq \|A\| \|B\|$), e.g. induced vector p-norms (see Appendix).

Consider the system:

$$\begin{bmatrix} 0.835 & 0.667 \\ 0.333 & 0.266 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} 0.168 \\ 0.067 \end{bmatrix}$$

The exact solution is: $z_1 = 1$; $z_2 = 1$
If u_2 is perturbed so that $u_2 = 0.066$ the exact solution is: $z_1 = -666$; $z_2 = 834$

Why does this happen?

First note that a_{11} / a_{21} = 2.507508 ≈ a_{12} / a_{22} = 2.507519, that is, both lines are practically *collinear*.
For collinear lines a small perturbation can yield a large deviation of the results.

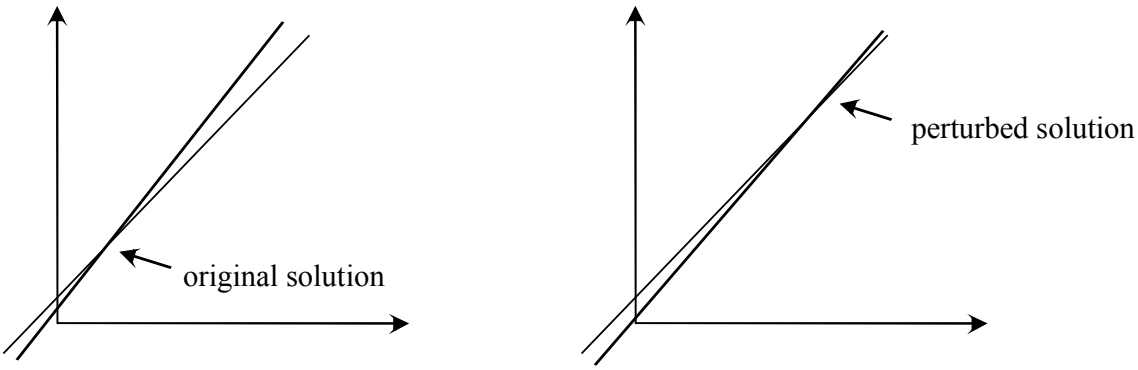


Fig. 3.2. Original and perturbed solutions with almost collinear lines.

Next, notice that the eigenvalues of A are (MATLAB computed):

$\lambda_{\max}$ = 1.10100090826446
$\lambda_{\min}$ = –0.00000090826446

Therefore, $\lambda_{\min}$ is practically zero. In other words, the matrix data defines only one direction instead of two. This is precisely the collinearity issue.

The condition number of A using 2-norm (see Appendix) is:

$$k(A) = \frac{|\lambda_{A,\max}|}{|\lambda_{A,\min}|} = 1.2\times 10^6$$

A perturbation on u is amplified as:

$$\frac{\|\Delta z\|}{\|z\|} \lesssim k(A)\frac{\|\Delta u\|}{\|u\|}$$

Using 2-norm and the values of the example, we get:

$$\frac{\|\Delta u\|}{\|u\|} = \frac{0.001}{\sqrt{0.168^2 + 0.067^2}} = 0.00553 \qquad \frac{\|\Delta z\|}{\|z\|} \lesssim k(A)\frac{\|\Delta u\|}{\|u\|} \approx 6635$$

Let us now perturb matrix A as follows:

$$\begin{bmatrix} 0.835 & 0.666 \\ 0.3329 & 0.266 \end{bmatrix}$$

The exact solution for the unperturbed u is: $z_1 = 0.1656$; $z_2 = 0.04465$

If the matrix coefficients are perturbed to produce the system $(A + \Delta A)z = u$, we have:

$$\frac{\|\Delta z\|}{\|z\|} \lesssim k(A)\frac{\|\Delta A\|}{\|A\|}$$

Now, we have: $\Delta A = \begin{bmatrix} 0 & 0.001 \\ 0.0001 & 0 \end{bmatrix} \Rightarrow \lambda_{\Delta A,\max} = 0.316\times10^{-3}$

$$k(A)\frac{\|\Delta A\|}{\|A\|} = 1.2\times10^{6}\,\frac{0.316\times10^{-3}}{1.1} = 273$$

3.2.8 The Normal Equation System Problem

We now consider $u = Az$ as a linear model of order p (p – 1 predictors plus an independent term), where z is a $p\times1$ matrix of the predictor values and u is a $n\times1$ matrix of observations, with $p \le n$.
The *linear* regression problem corresponds to:

$$u = Az + \varepsilon, \quad \text{with } A \equiv A_{np}$$

The inverse problem is solved in the least-squares sense by solving the system of *normal equations*[3]:

$$A'Az = A'u$$

$A'A$ is now a square matrix $n\times n$ and the system can then be solved as:

$z = (A'A)^{-1}A'u = A^{*}u$, where A^{*} is the pseudo-inverse of matrix A.

The ill-conditioning of the regression problem can be assessed using the condition number of $A'A$, as was done in the previous example. As an illustration, let us take:

$$A_{63} = \begin{bmatrix} 1 & 1 & 6.25 \\ 1 & 2 & 2.25 \\ 1 & 3 & 0.25 \\ 1 & 4 & 0.25 \\ 1 & 5 & 2.25 \\ 1 & 6 & 6.25 \end{bmatrix}; \quad u = \begin{bmatrix} 7.25 \\ 4.25 \\ 3.25 \\ 4.25 \\ 7.25 \\ 12.25 \end{bmatrix}$$

[3] It is a well-known fact that the normal equations yield an MSE solution, which for the linear model and proper assumptions of the errors ε, corresponds to the regression function of 2.2.

There is no multicollinearity. The solution is: z = [0 1 1]'.

If we perturb u, for instance u_3 = 12.2, the solution is: z = [0.0297 0.9929 0.9955]'.

If we add a new predictor that is a linear combination of existing ones we will get zero eigenvalues; i.e., the matrix is singular. Instead, let us add a linear combination with noise ($0.5A_2 + 0.2\,A_3 + N(0,0.1)$):

$$A_{63} = \begin{bmatrix} 1 & 1 & 6.25 & 1.86 \\ 1 & 2 & 2.25 & 1.47 \\ 1 & 3 & 0.25 & 1.63 \\ 1 & 4 & 0.25 & 2.00 \\ 1 & 5 & 2.25 & 2.82 \\ 1 & 6 & 6.25 & 4.41 \end{bmatrix}$$

The eigenvalues are: 0.027, 0.888, 29.250, 194.888. Therefore, the condition number is $k(A)$ = 194.888/0.027 = 7218.

The solutions for the unperturbed and perturbed u are: z = [0 1 1 0]' and z = [0.0292 1.0435 1.0186 -0.1032]'.

As the prediction order p grows the conditioning issue becomes more serious. The following figure shows averages (in 10000 experiments) of the condition number of noise ($N(0,1)$) matrices $n \times p$. Note that the ill-conditioning always increases with p, although for the same p it decreases with n.

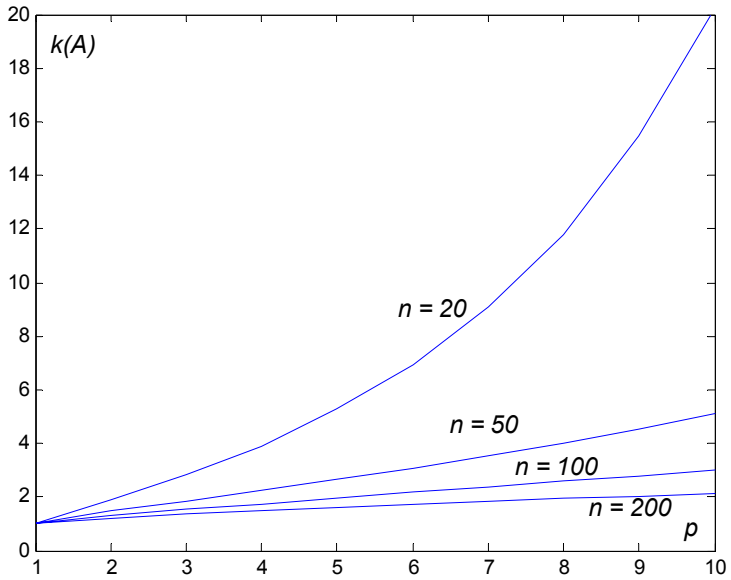


Fig. 3.3. Average (in 10000 experiments) condition number of noise ($N(0,1)$) matrices $n \times p$. Note that the ill-conditioning increases with p, and for the same p decreases with n.

Let us consider solving the following regression problem. A machine with unknown polynomial response function is fed with the x values of Fig.3.2. The corresponding y

values are obtained with a certain amount of noise. Two noise situations are shown in Fig.3.2a and 3.2b. Identify the system.

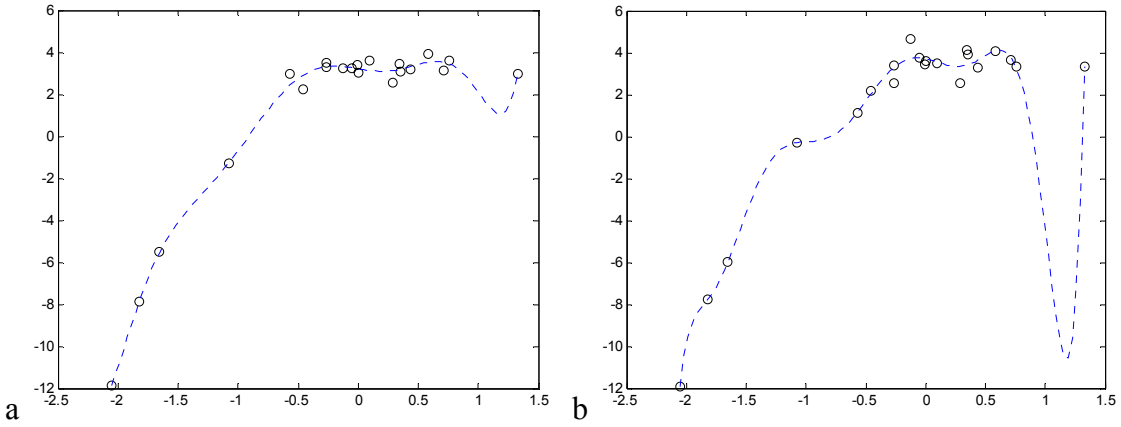


Fig. 3.4. A set of 21 points (circles) with 9th order polynomial fits (dotted blue lines). In both cases the x values are the same; only the y values correspond to different noise values from the same distribution.

Assume we choose $p = 9$. We get the fits shown in Fig. 3.2. A small variation of the noise produces large variations of the polynomial fits (in both cases the R^2 is high):

Polynomial coefficients	a_0	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9
Figure 7.17a	3.21	–0.93	0.31	8.51	–3.27	–9.27	–0.47	3.05	0.94	0.03
Figure 7.17b	3.72	–1.21	–6.98	20.87	19.98	–30.92	–31.57	6.18	12.48	2.96

Therefore, ill-conditioning means that a large dataset (in the example, many predictors) may contain a very small amount of useful information.

3.2.9 The Mapping Reconstruction Problem

We are given n observations $(x_1, t_1), \ldots, (x_n, t_n)$ of an unknown mapping $f : X \to T$ and want to find the function f.

As in the previous problem we may try to find a solution in the least-square sense, by minimizing

$$\sum_{i=1}^{n} \| t_i - f(x_i) \|^2$$

The difference now is that there is no assumption of linearity of f and no assumptions on the distribution of the errors $t_i - f(x_i)$.

The mapping reconstruction can be attempted with a neural network (NN) with weights w:

$$\min_w E_w = \sum_{i=1}^{n} \| t_i - f_w(x_i) \|^2$$

This is a discrete version of the supervised learning problem using the square norm as loss function. As a matter of fact, assuming $p(x,t)$ is the joint pdf of (X,T), we have for the finite discrete dataset:

$$p(x,t) = \frac{1}{n} \sum_i \delta(x - x_i)\delta(t - t_i)$$

Therefore:

$$R(f_w(x)) = \frac{1}{n} \iint \| t - f_w(x) \|^2 \sum_i \delta(x - x_i)\delta(t - t_i) dx dt = \frac{1}{n} E_w$$

In terms of the formalism $Az = u$, the direct mapping problem can be cast as:

$$f_w(x) = t \ , \quad \text{with } A = I,\ z = f_w(x)$$

The inverse problem – the mapping reconstruction problem – is then cast as:

$$\text{Find } f_w(x) \text{ such as } A^{-1}t = t \equiv f_w(x)$$

In other words, $\forall t$ one can reconstruct the mapping using a domain point. Maybe there is more than one such point. Here, we are more interested in the possibility of ill-conditioning, as in the case of the normal equations.

3.3 Problems Well-Posed in Tikhonov's Sense

Definition: The problem of solving $Az = u$ is *well-posed (correct) in Tikhonov's sense* on the set $M \subset F$ if:

- The solution of $Az = u$ exists for each $u \in AM = N$ and belongs to M.
- The solution belonging to M is unique for any $u \in N$.
- The solutions belonging to M are stable (with respect to $u \in N$)

Therefore, correctness in Tikhonov's sense corresponds to a restriction of Hadamard's correctness to subsets of solutions: $M \subset F$.

Lemma: If A is a continuous one-to-one operator defined on a *compact* set $M \subset F$, then the inverse operator A^{-1} is continuous on the set $N = AM$ (for the proof see Tikhonov AN, Arsenin VY, 1977 or Vapnik V, 1998).

This lemma provides the so-called *selection method* of solving ill-posed problems: narrow the solution set to a compact set.

Example 3.1

Consider again Example 3.4 and suppose that $\{z(s)\}$ is compact. However, $\{z^*(s) = z(s) + \sin(\omega s)\}$ is not compact since $M = \{\sin\omega s\}$ is not compact: no subsequence of M converges to a function of M.

Now, suppose that we had $M = \{\alpha\sin\omega t\}$, ω = constant, M is compact and the problem is well-posed.

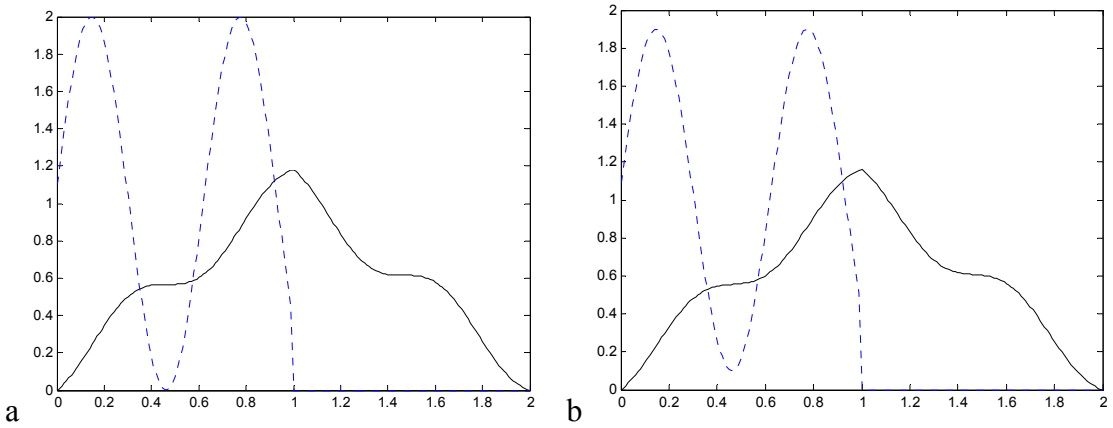


Fig. 3.5. Two close convolutions as in Example 3.4, but now with $M = \{\alpha\sin\omega t\}$ and $\omega = 10$: a) $\alpha = 1$; a) $\alpha = 0.9$.

4 The Regularization Method

There are two categories of regularization methods aiming at converting an ill-posed problem into a well-posed one:

1. Regularization by data correction
2. Regularization by operator correction

The first category of methods convert $Az = u$ into $Az = f(u)$, so that continuity conditions hold. The problem of course is to find an appropriate smoothing $f(u)$. In the following we consider the second category of methods.

4.1 The Regularizing Operator

The regularization method is one of the methods of dealing with ill-posed problems, applicable even when the set F of all possible solutions is not compact (*genuinely ill-posed problems*).

The regularization method provides an approximate solution for a perturbation δ of the exact right-hand member u_T. Let us denote by u_δ the perturbed right-hand member:

$$\rho_U(u_\delta, u_T) \le \delta$$

Since the problem is ill-posed, the solution is not $z_\delta = A^{-1}u_\delta$. However, it may be possible to find an operator $R(u,\delta)$ providing a value $z_\delta = R(u_\delta,\delta)$ that is as close as we wish of the exact solution z_T with $Az_T = u_T$.

Definition: An operator $R(u,\delta)$ is said to be a *regularizing operator* for the equation $Az = u$ in a neighborhood of u_T if

1. There exists a positive number δ_1 such that $R(u,\delta)$ is defined for every δ in $[0, \delta_1]$ and every u_δ in U such that
$$\rho_U(u_\delta, u_T) \le \delta$$
2. For every $\varepsilon > 0$ there exists a $\delta_0 = \delta_0(\varepsilon, u_T) \le \delta_1$ such that
$$\rho_U(u_\delta, u_T) \le \delta \le \delta_0 \quad \Rightarrow \quad \rho_F(z_\delta, z_T) \le \varepsilon$$
where $z_\delta = R(u_\delta,\delta)$.

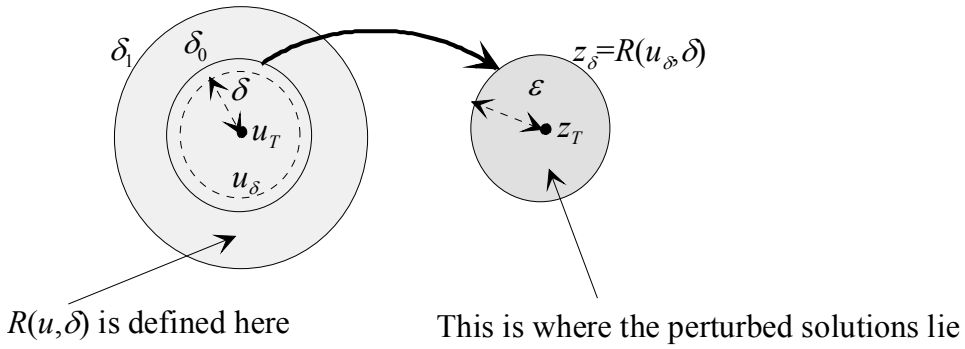


Fig. 4.1. **The regularizing operator is such that for any deviation ε of the exact solution one can find a neighborhood of u_T such that any δ-perturbed solution does not deviate more than ε from the exact solution.**

It is often convenient to write the regularizing operator as $R(u,\alpha)$ depending on a *regularizing parameter* $\alpha(\delta)$.

The approximate solution $z_\alpha = R(u_\delta, \alpha(\delta))$ is called the *regularized solution*.
This method of constructing approximate solutions is called *regularization method*.

4.2 The Stabilizing Functional

The problem of building a regularizing operator may not be an easy task. The so-called *variational principle* uses the concept of stabilizing functional in order to achieve this task.

Definition: Let $\Omega(z)$ denote a *continuous nonnegative* functional defined on a subset F_1 of F that is everywhere dense in F. Suppose that:

1. z_T belongs to the domain of definition of $\Omega(z)$.
2. For every $d > 0$, the set of elements z of F_1 for which $\Omega(z) \le d$ is a compact subset of F_1.

The functional $\Omega(z)$ is then called a *stabilizing functional*.

It can be shown (see Tikhonov AN, Arsenin VY, 1977) that the following method is a regularizing method (proposed by Tikhonov in 1963):

1. Consider: $Q_\delta \equiv \{z;\ \rho_U(Az, u_\delta) \le \delta\}$
2. Take only those elements of Q_δ where $\Omega(z)$ is defined: $F_{1,\delta} \equiv Q_\delta \cap F_1$
3. Minimize $\Omega(z)$ in $F_{1,\delta}$

Using further results (see details in Tikhonov AN, Arsenin VY, 1977) and imposing some mild conditions on $\Omega(z)$ (quasimonotonicity[4]), the regularized solution is obtained by solving the following constrained extremum problem

$$\text{Determine } \min \Omega(z) \text{ on } F_1 \text{ subject to } \rho_U(Az, u_\delta) = \delta$$

This constrained extremum problem is solved with the Lagrange multipliers method by minimizing the following *smoothing functional*:

$$M^\alpha(z, u_\delta) = \rho_U^2(Az, u_\delta) + \alpha\Omega(z)$$

The solution of this problem, z_α, can be seen as $z_\alpha = R(u_\delta, \alpha)$, that is $\min M^\alpha(z, u_\delta)$ produces stable solutions . The regularizing parameter α can be chosen based on the discrepancy $\rho_U(Az_\alpha, u_\delta) = \delta$.

[4] $\Omega(z)$ is quasimonotonic if for every z outside infimum values one can always find in any neighborhood of z an element z_1 such that $\Omega(z_1) < \Omega(z)$.

4.3 Examples of stabilizing functionals

Example 4.1

F is the space of continuous functions $z(x)$ on the interval $[a, b]$ with the C-metric:

$$\rho_F(z_1, z_2) = \sup_{x\in[a,b]} |z_1(x) - z_2(x)|$$

Tikhonov stabilizer of order p:

$$\Omega(z) = \int_a^b \sum_{r=0}^{p} q_r(x) \left(\frac{d^r z}{dx^r}\right)^2 dx$$

where $q_r(x) \geq 0$ for $r = 0, 1, \ldots, p-1$ and $q_p(x) > 0$. For instance, a Tikhonov stabilizer of order 1 is:

$$\Omega(z) = \int_a^b \left\{ q_0(x) z^2(x) + q_1(x) \left(\frac{dz}{dx}\right)^2 \right\} dx$$

The functions $q_r(x)$ can be constants, for instance:

$$\Omega(z) = \int_a^b \left(\frac{dz}{dx}\right)^2 dx = \|Dz\|^2$$

where D is a *linear differential operator*.

Example 4.2

Suppose F and U are Hilbert spaces and that A is a continuous operator from F into U. Let F_1 denote a Hilbert subspace of F with a norm such that, for every $d > 0$, the set of elements z of F_1 for which $\|z\| \leq d$ is compact. In this case we can take for the stabilizer the functional:

$$\Omega(z) = \|z\|^2$$

As a special case, assume: $\rho_U^2(Az, u) = \|Az - u\|^2$.

The minimization of the regularizer $M^\alpha(z, u) = \|Az - u\|^2 + \alpha\|z\|^2$ corresponds to equating to zero its derivative (in order to z); that is:

$$A^*(Az - u) + \alpha z = 0 \quad \Rightarrow \qquad A^* Az + \alpha z = A^* u$$

where A^* is the adjoint operator[5] of A.
Now, A^*u can be written as a series:

$$A^*u = \sum_{n=1}^{\infty} c_n \varphi_n$$

where φ_n are the eigenfunctions with eigenvalues λ_n. If we seek a solution in the form

$$z = \sum_{n=1}^{\infty} b_n \varphi_n$$

we just have to compute the coefficients as

$$b_n = \frac{c_n}{\lambda_n + \alpha}$$

4.4 Tikhonov's Regularization is a Nutshell

We illustrate with the mapping reconstruction problem.
Tikhonov's regularization based on the variational principle involves:

1. **The standard error term ($E_s = \rho_U^2(Az, u)$)**

$$E_s = \sum \left\| f(x_i) - t_i \right\|^2$$

2. **The regularizing term ($\Omega(z)$)**

$$E_r = \left\| Df \right\|^2$$

3. **The minimization of the total error**

$$E_t = E_s + \alpha E_r \qquad (\alpha > 0)$$

Comments:

1. The linear differential operator imposes (obviously) some *smoothing* condition (therefore, model complexity) to the admissible mapping *f*. The same applies to other types of stabilizing functionals.
2. The exact form of the stabilizing functional is problem-dependent; i.e., it depends on *prior information* we may have on the problem.

[5] A^* is the adjoint operator of A if $\langle Ax, y\rangle = \langle x, A^*y\rangle$ for all $x, y \in F$

3. The regularizing term E_r can be interpreted as a *model complexity-penalty*.
4. The amount of complexity penalization is tuned with α, the *regularization parameter*.

   $\alpha = 0$: no penalization.

   $\alpha = \infty$: the smoothness constraint specifies the solution (the data are unreliable).

   Usually α is chosen experimentally (see figure). Theoretical results on the choice of α (see literature) are not easily applied in practice.
5. Regularization can be viewed as providing a practical solution to the *bias-variance dilemma*.

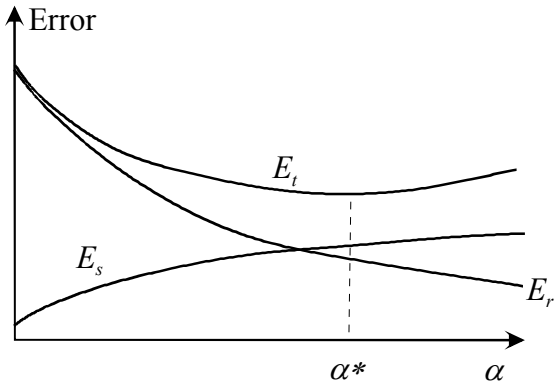


Fig. 4.2. Regularization error terms.

5 Regularized Solutions of Some Problems

5.1 Regularized Training of NN

5.1.1 Weight Decay

Weight decay is inspired on the stabilizing functional $E_r = \|Df\|^2$. Assume, for instance, a linear perceptron:

$$f(x_i) = \sum_i w_i x_i$$

The discrete version of a possible stabilizing functional is then:

$$\Omega(z) = \sum_i \left(\frac{df}{dx_i} \right)^2 = \sum_i w_i^2$$

When minimizing the total error the penalizing function "encourages" the weights to be small, and therefore "encourages" less complex functions.

For quadratic error functions it can be shown that weights along principal directions with $\lambda_j >> \alpha$, the weights remain practically unchanged; for $\lambda_j << \alpha$, the weights become neglectable.

5.1.2 Early Stopping

There is no theory of early stopping.

Qualitatively early stopping often works because training first progresses along the main principal components (e.g., maximum gradient descent) and progresses on less important components. Halting the training before reaching a minimum corresponds to limiting the network complexity.

5.1.3 Training with Noise

Training with noise with zero mean and uncorrelated between different inputs can be shown to add the following term to the error formula presented in 2.2:

$$\Omega = \sum_i \iint \left(\frac{\partial f}{\partial x_i} \right)^2 p(t,x) dt dx$$

This is a Tikhonov stabilizing functional.

The derivation assumes sufficiently small noise so that the Taylor series expansion may neglect higher than second-order terms, as well as a neglectable contribution of the second-order term near the minimum.

If cross-entropy is used as cost function a similar conclusion is derived.

For details see (Bishop C, 1995b).

5.2 Ridge Regression

We are in the conditions of Example 4.2 and use as stabilizer $\|u\|^2 = u'u$. That is, we try to minimize

$$E = \|Az - u\|^2 + \alpha \|u\|^2 = (Az - u)'(Az - u) + \alpha u'u$$

Taking the derivatives equal to zero we get

$$z = (A'A + \alpha I)^{-1} A'u$$

That is we add a penalization term to the diagonal of $A'A$, α, known as the *ridge factor*.

The ridge factor can be chosen based on:

- The error curves (see Fig. 4.2 and following example)
- The so-called *ridge traces* (evolution of a_{ij} with α)
- Cross- validation

We now present the influence of the ridge factor using the very simple dataset shown in Fig, 5.1 with a fitted first-order and a second-order model.

In Fig. 5.1a the ridge factor is zero; therefore, the parabola passes exactly at the 3 points, is tightly attached to the "training set" and unable to generalise. The z vector is in this case [0 3.5 −1.5]'.
The remaining pictures of Fig. 5.1 show what happens with a regularizer.
Fig. 5.2 shows for $r \in [0, 2]$ the Sum of Squares Error curve together with the curve of the following error:

$$\mathrm{SSE(L)} = \sum (\hat{u}_i - \hat{u}_{iL})^2 ,$$

where the $\hat{u}_i$ are the predicted values (second-order model) and the $\hat{u}_{iL}$ are the predicted values of the linear model, which is the preferred model in this case. Note the *ridge* aspect of this curve. The minimum of SSE(L) (L from Linear) occurs at $\alpha = 0.6$, where the SSE curve starts to saturate.

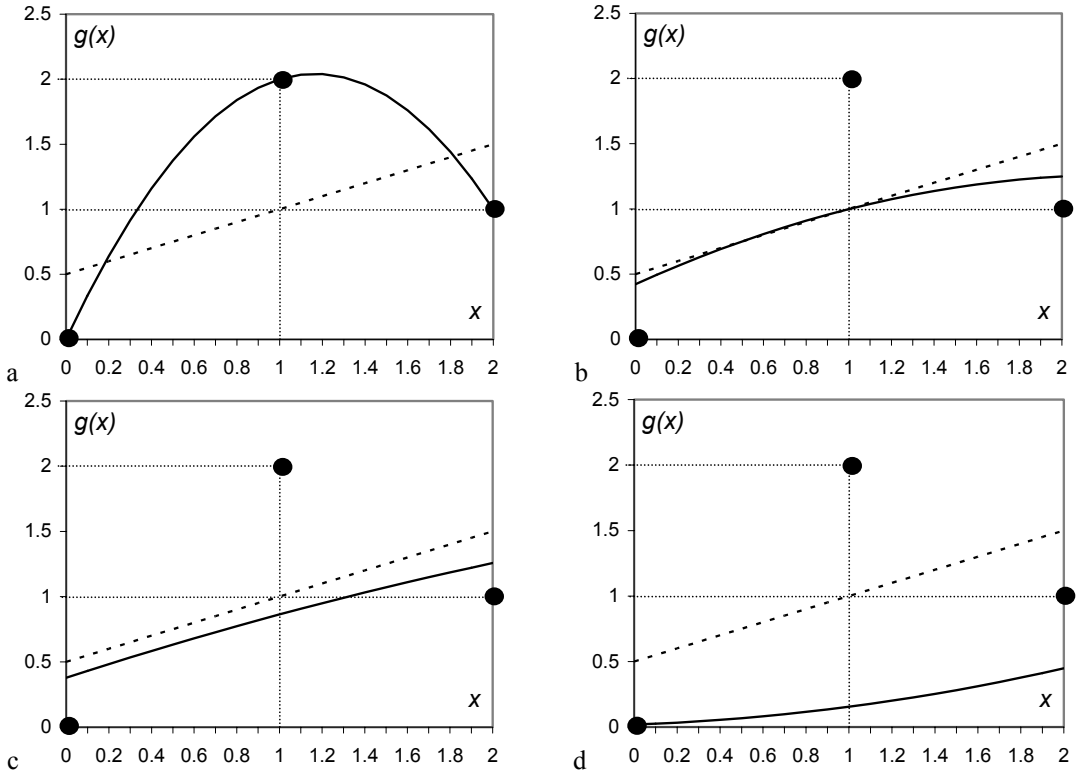


Fig. 5.1. Fitting a second-order model to a very simple dataset (3 points represented by solid circles) with ridge factor: a) 0; b) 0.6; c) 1; d) 50.

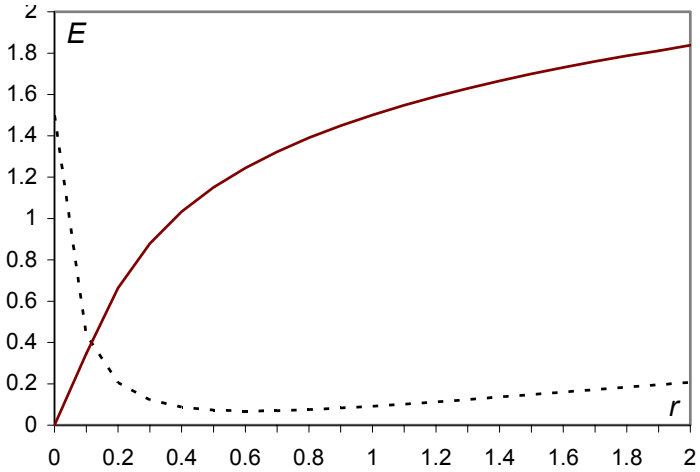


Fig. 5.2. SSE (solid line) and SSE(L) (dotted line) curves for the ridge regression solutions of Fig. 5.1 dataset (the ridge factor is denoted *r*).

Fig. 5.3 and following table show the *ridge regression* solution for the polynomial dataset of Example 3.8.

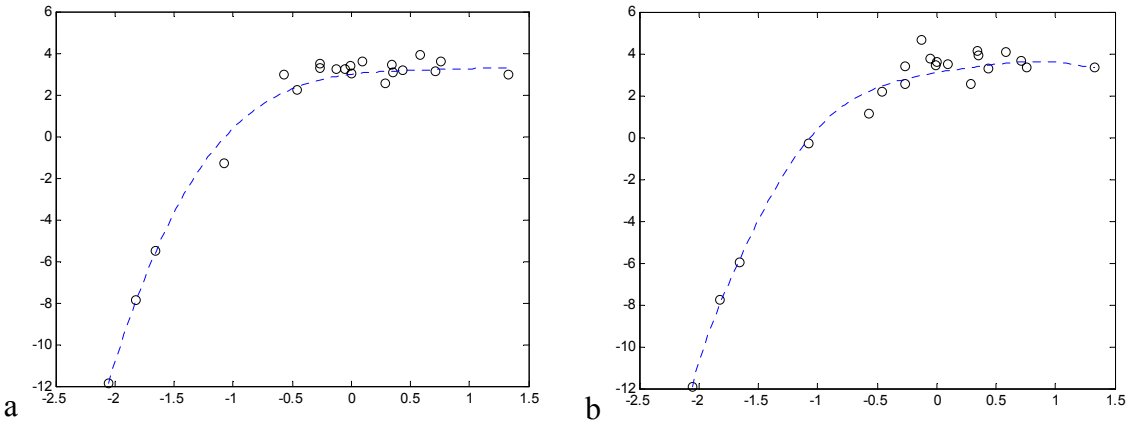


Fig. 5.3. Ridge regresion solutions for diferent noise values and $\alpha = 1$.

Polynom. coeffs	Fig. 5.3a	Fig 5.3b
a0	2.9557	3.0914
a1	0.6229	0.9740
a2	-0.4268	-0.5294
a3	0.7946	0.5240
a4	-0.5544	-0.4400
a5	0.3571	0.2263
a6	-0.1701	-0.2073
a7	-0.3214	-0.1868
a8	0.0802	0.0952

a9	0.0678	0.0515

5.3 Parzen Window Estimate as a Regularized Density Estimate

By using appropriate distance metrics and regularizing functions one can obtain the classic nonparametric density estimators: Parzen window estimator; projective estimator; spline estimator, etc.
In the following we only consider the Parzen window estimator.

Let us consider the above regularization functional with

1. $\rho_2(F_l, F) = \sqrt{\int_{-\infty}^{\infty} (F_l(x) - F(x))^2 \, dx}$

2. $W(f) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} K(z-x) f(x) dx \right)^2 dz$

Then the estimators f_l that minimize the regularized functional $R_{\gamma l}(f, F_l)$ are Parzen window estimators:

$$f_l(x) = \frac{1}{l} \sum_{i=1}^{l} G_{\gamma l}(x - x_i)$$

with[6]: $G_{\gamma l}(x) = \mathcal{F}^{-1}(g_{\gamma l}(\omega))$

$$g_{\gamma l}(\omega) = \frac{1}{1 + \gamma_l \omega^2 k(\omega) k(-\omega)}$$

$$k(\omega) = \mathcal{F}(K(x))$$

If the density has a finite support the Parzen window estimator has a bias. (The estimate needs corrections for the ending points of the finite support.)

Vapnik (2000) compares and discusses the similarity between Parzen window estimation and SVM estimation of a density function.

[6] $\mathcal{F}$ denotes the Fourier transform operator

6 Appendix

6.1 Function, Functional, Operator

Let M and N be two sets of elements connected by a functional dependence:

$$A : M \to N \quad (AM = N)$$
$$f \to F$$

- M and N are sets of numbers: A is a **function**
- $M = \{f(t)\}$ and N is a set of numbers: A is a **functional**

$$\text{Example: } H = -\int f(t) \log f(t) dt \,;\; A = H(f)$$

- $M = \{f(t)\}$ and $N = \{F(x)\}$: A is an **operator**

$$\text{Example: } F(x) = K(t) \otimes f(t)\big|_x = \int K(x-t) f(t) dt \,; \qquad A = K \otimes$$

6.2 Matrix Norms and Perturbation Formulas

A matrix norm, $||A||$ must obey:

1. $||A|| \geq 0$ and $||A|| = 0$ iff $A = 0$
2. $||\alpha A|| = |\alpha| \, ||A||$
3. $||A + B|| \leq ||A|| + ||B||$ for matrices of the same size
4. $||AB|| \leq ||A||.||B||$ for conformable matrices

A family of matrix norms satisfying the above conditions is the family induced vector p-norms:

$$\|A\|_p = \max_{\|x\|_p = 1} \|Ax\|_p$$

- Matrix 1-norm: $\|A\|_1 = \max_{\|x\|_1 = 1} \|Ax\|_1 = \max_j \sum_i |a_{ij}|$, the largest absolute column sum
- Matrix ∞-norm: $\|A\|_\infty = \max_{\|x\|_\infty = 1} \|Ax\|_\infty = \max_i \sum_j |a_{ij}|$, the largest absolute row sum

- Matrix 2-norm: $\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 = \sqrt{|\lambda_{A,\max}|}$. When A is nonsingular $\left\|A^{-1}\right\|_2 = \frac{1}{\sqrt{|\lambda_{A,\min}|}}$ (λ are eigenvalues)

Suppose the $Ax = b$ systems is perturbed by adding ΔA and Δb. The following results can be proved for sufficiently small ΔA:

- $\|\Delta x\| \le \frac{\left\|A^{-1}\right\|\left(\|\Delta b\| + \|\Delta A\|\|x\|\right)}{1 - \left\|A^{-1}\right\|\|\Delta A\|}$
- For $\Delta b = 0$: $\frac{\|\Delta x\|}{\|x\|} \lesssim k(A)\frac{\|\Delta A\|}{\|A\|}$
- For $\Delta A = 0$: $\frac{\|\Delta x\|}{\|x\|} \lesssim k(A)\frac{\|\Delta b\|}{\|b\|}$

where $k(A)$, the condition number of A, is defined as:

$$k(A) = \left\|A^{-1}\right\|\|A\|,$$

which for the 2-norm is:

$$k(A) = \frac{|\lambda_{A,\max}|}{|\lambda_{A,\min}|}$$

The above results are based on the following result: $(A+B)^{-1} \approx A^{-1} - A^{-1}BA^{-1}$

6.3 Completely Continuous Operator and the Fredholm Integral

A linear operator A, defined in a linear normed space E_1, with range in a linear normed space E_2, is *completely continuous* if each infinite bounded sequence in E_1:

$$f_1, f_2, ..., f_i, ... \quad \|f_i\| \le c$$

is mapped into a sequence

$$Af_1, Af_2, ..., Af_i, ...$$

such that a convergent subsequence can be extracted from it (is a compact set).

- If E_1 contains bounded noncompact sets the inverse operator A^{-1} need not be continuous (proof in e.g. Vapnik, 1998).

- When the kernel $K(t,x)$ is continuous in $[a, b]^2$, the Fredholm integral is a *completely continuous* operator (proof in e.g. Kolmogorov and Fomin, 1970).
- Therefore the Fredholm integral need not be continuous.

References

Bishop C (1995a) Neural Networks for Pattern Recognition. Oxford University Press.

Bishop C (1995b) Training with Noise is Equivalent to Tikhonov Regularzation. Neural Computation 7, pp. 108-116.

Bjorkstrom A (2001) Ridge regression and inverse problems. Stockholm Univ., Sweden.

Golub HG, Van Loan CF (1989) Matrix Computations. The John Hopkins Univ. Press.

Groetsch CW (1999) Inverse Problems. The Mathematical Association of America.

Haykin S (1999) Neural Networks. A Comprehensive Foundation. Prentice Hall.

Kirsch A (1996) An Introduction to the Mathematical Theory of Inverse Problems. Springer Verlag, Inc..

Kolmogorov A, Fomin S (1999) Elements of the Theory of Functions and Functional Analysis. Dover Pub. Inc.

Lauter H, Liero H (1997) Ill-posed inverse problems and their global optimization. Institute of Mathematics, Univ. Potsdam, Germany.

Marques de Sá JP (2007) Applied Statistics, Using SPSS, STATISTICA, MATLAB and R. Springer-Verlag.

Meyer CD (2000) Matrix Analysis and Applied Linear Algebra. SIAM.

Ozturk F, Akdeniz F (2000) Ill-Conditioning and Multicollinearity. Linear Algebra and its Applications, 321, pp. 295-305.

Ramm AG (2005) Inverse Problems. Mathematical and Analytical Techniques with Applications to Engineering. Springer Verlag, Inc.

Tikhonov AN, Arsenin VY (1977) Solutions of Ill-Posed Problems. John Wiley & Sons.

Vapnik V (1998) Statistical Learning Theory. John Wiley & Sons, Inc.

Vapnik V (2000) The Nature of Statistical Learning Theory. Springer-Verlag.