Problem	Problem's name	Reference
1	Penalty 1	Gill and Murray (1973)
2	Trigonometric:	More et al. (1981)
3	Extended Rosenbrock	More et al. (1981)
4	Extended Powell	More et al. (1981)
5	Tridiagonal	Buckley and LeNir (1983)
6	QOR	Toint (1978)
7	GOR	Toint (1978)
8	PSP	Toint (1978)
9	Tridiagonal	Toint (1983a)
10	Linear Minimum Surface	Toint (1983a)
11	Exterked ENGVL1	Toint (1983a)
12	Matrix Square Root 1	
13	Matrix Square Root 2	
14	Extended Freudenstein and Roth	Toint (1983a)
15	Sparse Matrix Square Root	
16	u1ts0	Gilbert and Lemaréchal (1988)

Table 1: Set of test problems

Problems 12, 13 and 15, and the starting points used for them, are described in Liu and Nocedal (1988). They derive from the problem of determining the square root of a given matrix A, i.e. finding a matrix B such that $B^2 = A$. For all the other problems we used the standard starting points given in the references. All the runs reported in this paper were terminated when

$$||g_k|| < 10^{-5} \times \max(1, ||x_k||),$$
 (2.8)

where $\|\cdot\|$ denotes the Euclidean norm. We require low accuracy in the solution because this is common in practical applications.

Since we have performed a very large number of tests, we describe the results fully in an accompanying report (Liu and Nocedal (1988)). In this paper we present only representative samples and summaries of these results, and the interested reader is referred to that report for a detailed description of all the tests performed. We should note that all the comments and conclusions made in this paper are based on data presented here and in the accompanying report.

3 Comparison with the method of Buckley and LeNir

In this section we compare the method of Buckley and LeNir (B-L) with the L-BFGS method. In both methods the user specifies the amount of storage to be used, by giving a number m, which determines the number of matrix updates that can be stored. When m=1, the method of Buckley and LeNir reduces to Shanno's method, and when $m=\infty$ both methods are identical to the BFGS method. For a given value of m, the two methods require roughly the same amount of storage, but the L-BFGS method requires slightly less arithmetic work per iteration than the B-L method (as implemented by Buckley and LeNir (1985)).

In particular, consider pruning the training dataset by keeping only the examples with the smallest margin $|z^{\mu}| = |\mathbf{J}_{\text{probe}} \cdot \mathbf{x}^{\mu}|$ along a probe student $\mathbf{J}_{\text{probe}}$. The pruned dataset will follow some distribution p(z) along the direction of $\mathbf{J}_{\text{probe}}$, and remain isotropic in the nullspace of $\mathbf{J}_{\text{probe}}$. In what follows we will derive a general theory for an arbitrary data distribution p(z), and specialize to the case of small-margin pruning only at the very end (in which case p(z) will take the form of a truncated Gaussian). We will also make no assumptions on the form of the probe student $\mathbf{J}_{\text{probe}}$ or the learning rule used to train it; only that $\mathbf{J}_{\text{probe}}$ has developed some overlap with the teacher, quantified by the angle $\theta = \cos^{-1}\left(\frac{\mathbf{J}_{\text{retle}}\cdot\mathbf{T}}{\|\mathbf{J}_{\text{retle}}\|_2\|\mathbf{T}\|_2}\right)$ (Fig. 2A).

After the dataset has been pruned, we consider training a new student J from scratch on the pruned dataset. A typical training algorithm (used in support vector machines and the solution to which SGD converges on separable data) is to find the solution J which classifies the training data with the maximal margin $\kappa = \min_{\mu} \mathbf{J} \cdot (y^{\mu} \mathbf{x}^{\mu})$. Our goal is to compute the generalization error ε_g of this student, which is simply governed by the overlap between the student and the teacher, $\varepsilon_g = \cos^{-1}(R)/\pi$, where $R = \mathbf{J} \cdot \mathbf{T}/\|\mathbf{J}\|_2 \|\mathbf{T}\|_2$.

Main result and overview

Our main result is a set of self-consistent equations which can be solved to obtain the generalization error $\varepsilon(\alpha, p, \theta)$ for any α and any data distribution p(z) along a probe student at any angle θ relative to the teacher. These equations take the form,

$$\frac{R - \rho \cos \theta}{\sin^2 \theta} = \frac{\alpha}{\pi \Lambda} \left\langle \int_{-\infty}^{\kappa} dt \, \exp\left(-\frac{\Delta(t, z)}{2\Lambda^2}\right) (\kappa - t) \right\rangle_z \tag{1}$$

$$1 - \frac{\rho^2 + R^2 - 2\rho R \cos \theta}{\sin^2 \theta} = 2\alpha \left\langle \int_{-\infty}^{\kappa} dt \frac{e^{-\frac{(t-\rho\mu)^2}{2}}}{\sqrt{2\pi}\sqrt{1-\rho^2}} H\left(\frac{\Gamma(t,z)}{\sqrt{1-\rho^2}\Lambda}\right) (\kappa - t)^2 \right\rangle_z \tag{2}$$

$$\frac{\rho - R\cos\theta}{\sin^2\theta} = 2\alpha \left\langle \int_{-\infty}^{\kappa} dt \frac{e^{-\frac{(t-\rho\mu)^2}{2(1-\rho\mu)^2}}}{\sqrt{2\pi}\sqrt{1-\rho^2}} H\left(\frac{\Gamma(t,z)}{\sqrt{1-\rho^2}\Lambda}\right) \left(\frac{z-\rho t}{1-\rho^2}\right) (\kappa - t) + \frac{1}{2\pi\Lambda} \exp\left(-\frac{\Delta(t,z)}{2\Lambda^2}\right) \left(\frac{\rho R - \cos\theta}{1-\rho^2}\right) (\kappa - t) \right\rangle$$

Where,

$$\Lambda = \sqrt{\sin^2 \theta - R^2 - \rho^2 + 2\rho R \cos \theta},\tag{4}$$

(3)

$$\Gamma(t, z) = z(\rho R - \cos \theta) - t(R - \rho \cos \theta), \tag{5}$$

$$\Delta(t,z) = z^2 \left(\rho^2 + \cos^2 \theta - 2\rho R \cos \theta\right) + 2tz(R \cos \theta - \rho) + t^2 \sin^2 \theta. \tag{6}$$

Where $\langle \cdot \rangle_z$ represents an average over the pruned data distribution p(z) along the probe student. For any $\alpha, p(z), \theta$, these equations can be solved for the order parameters R, ρ, κ , from which the generalization error can be easily read off as $\varepsilon_g = \cos^{-1}(R)/\pi$. This calculation results in the solid theory curves in Figs 1,2,3, which show an excellent match to numerical simulations. In the following section we will walk through the derivation of these equations using replica theory. In Section A.6 we will derive an expression for the information gained per training example, and show that with Pareto optimal data pruning this information gain can be made to converge to a finite rate, resulting in at least exponential decay in test error. In Section A.7, we will show that super-exponential scaling eventually breaks down when the probe student does not match the teacher perfectly, resulting in power law scaling at at a minimum pruning fraction $f_{\min}(\theta)$.

Replica calculation of the generalization error

To obtain Eqs. 1,2,3, we follow the approach of Elizabeth Gardner and compute the volume $\Omega(\mathbf{x}^{\mu}, \mathbf{T}, \kappa)$ of solutions J which perfectly classify the training data up to a margin κ (known as the Gardner volume) [29, 25]. As κ grows, the volume of solutions shrinks until it reaches a unique solution at a critical κ , the max-margin solution. The Gardner volume Ω takes the form,

and the rule is proved that

$$\frac{du^*}{dx} = nu^{*-1}\frac{du}{dx},$$

where n is a positive fraction whose numerator and denominator are integers. This rule has already been used in the solution of numerous exercises.

34 The Derivative of a Constant

Let y = c, where c is a constant. Corresponding to any Dx, Dy = 0, and consequently

$$\frac{\Delta y}{\Delta x} = 0,$$

and

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = 0,$$

or

$$\frac{dy}{dx} = 0.$$

The derivative of a constant is zero. Interpret this result geometrically.

35 The Derivative of the Sum of Two Functions

Let

$$y = u + v$$
,

where u and v are functions of x. Let Du, Du, and Dy be the increments of u, v, and y, respectively, corresponding to the increment Dx.

$$y + \Delta y = u + \Delta u + v + \Delta v$$

$$\Delta y = \Delta u + \Delta v$$

$$\frac{\Delta y}{\Delta x} = \frac{\Delta u}{\Delta x} + \frac{\Delta v}{\Delta x}$$

$$\frac{dy}{dx} = \frac{du}{dx} + \frac{dv}{dx},$$

or

$$\frac{d(u+v)}{dx} = \frac{du}{dx} + \frac{dv}{dx}.$$

The derivative of the sum of two functions is equal to the sum of their derivatives.

On the other hand, if we denote by Tb and Tb the times in which light covers the "arm" of the "parallel" clock "there" and "back" (i.e., along the direction of propagation of the clock, and against this direction), we shall have for these two cases

$$cT_B^* = d + vT_B^*, \qquad cT_B^* = d - vT_B^*,$$

from where

$$T_B = T_B^* + T_B^* = \frac{2d}{c(1 - v^2/c^2)},$$

Hence it will be

$$T_L = T_B(1 - v^2/c^2)^{1/2}.$$

Now if for a certain time t the "perpendicular" light clock makes n1 ticks and the "parallel" nH ticks, we shall have

$$t = n_A T_A, \qquad t = n_B T_B,$$

and from (6) and (7) we obtain (1).

Our tenth axiom asserts, however, that instead of (1) one must have

$$n_B = n_A$$

and thus the periods of the light clocks must be equal

$$T_B = T_A$$
.

This empirical fact was first proved by the Michelson-Morley experiment (see SS49).

In the next sections of this chapter we shall see which conclusions are to be drawn from the assertion (9) of the tenth axiom and which must be the transformations of the space and time coordinates resulting from this axiom.

SS3. Transformation of Coordinates

A. The Galilean transformation

All transformations of the space and time coordinates which we consider in this section are between a frame K attached to absolute space and a frame K' moving with a constant velocity \overrightarrow{V} . To avoid trivial constants, we shall consider the so-called HOMOGENEOUS TRANSFOPMATION (cf. I, p. 201), at which at the initial time (t = 0) the origins of both frames coincide (see fig. 3.1 wherefor the sake of simplicity a two-dimensional case is presented). The transformation shown in fig. 3.1 is called GENER. TRANSFORMATION. If the axes of the moving frame K' are parallel to the axes of the rest frame K and if the velocity of K' is parallel to one of these axes (as a rule to the x-axis), this is called a SPECIAL TRANSFORMATION.

In fig. 3.2 a special transformation between frames K and K' is presented, where again a twodimensional case is given. Let point P (see fig. 3.2) be at rest in the moving frame

The heat energy is measured by the heat content, and the kinetic energy by the expression $\frac{V^2}{2q}$. We may then write:

$$JH_1 + \frac{{V_1}^2}{2g} = JH_2 + \frac{{V_2}^2}{2g} = \cdots JH_5 + \frac{{V_5}^2}{2g}.$$

As written, the expression refers to one pound of the gas, H representing the heat content of the gas per pound, and $\frac{V^2}{2g}$, the kinetic energy per pound. The above equation may be called the equation of the continuity of energy. It is a special case of

the First Law of Thermodynamics, or the general law of the Conservation of Energy.

179. The Equation of the Continuity of Mass.-In Fig. 109, for any section designated, the following relation holds;

$$\overline{W} = \frac{AV}{S}$$

where \overline{W} = weight passing the section per second;

A =area of the section in sq. ft.;

V = velocity of the gas at the section in ft. per sec.; and S = specific volume of the gas at thesection in cu. ft. per pound. (The change from V to S, to represent volume, is done to allow the representation of velocity by V.)

Although A, V, and S may have different values at other sections along the channel, \overline{W} is the same at every point. Hence,

$$\overline{W} = \frac{A_1 V_1}{S_1} = \frac{A_2 V_2}{S_2} = \cdots \frac{A_5 V_5}{S_5}$$

This equation may be called the equation of the Continuity of Mass.

180. Contour of a Nozzle Passage in Longitudinal Section.—A nozzle is an element whose primary function

then

$$\underset{x \pm \infty}{L} \frac{f(x)}{\phi(x)}$$

exists and is equal to k.4

Footnote 4: This and the following theorem are due to O. Stolz, who generalized them from the special cases (stated in our corollaries) due to Cauchy. See Stolz und Gmeiner, Functionentheorie, Vol. 1, p. 31. See also the reference to Bortolotti given on page 82.

Proof. Let $V_1(k)$ and $V_2(k)$ be a pair of vicinities of k such that $V_2(k)$ is entirely within $V_1(k)$. By hypothesis there exists an k and an k2 such that if k2,

$$\frac{f(x+h) - f(x)}{\phi(x+h) - \phi(x)} \tag{1}$$

is in $V_2(k)$. Since this is true for every $x > X_2$,

$$\frac{f(x+2h) - f(x+h)}{\phi(x+2h) - \phi(x+h)} \tag{2}$$

is also in $V_2(k)$. From this it follows by means of the lemma that

$$\frac{f(x+2h) - f(x)}{\phi(x+2h) - \phi(x)},\tag{3}$$

whose value is between the values of (1) and (2), is also in $V_2(k)$. By repeating this argument we have that for every integral value of n, and for every $x > X_2$,

$$\frac{f(x+nh) - f(x)}{\phi(x+nh) - \phi(x)}$$

is in $V_2(k)$.

By Theorem 65, for any x

$$\mathop{L}_{n \pm \infty} \frac{f(x+nh) - f(x)}{\phi(x+nh) - \phi(x)} = \frac{f(x)}{\phi(x)}.$$

Hence for every x and for every ε there exists a value of $n, N_{x\varepsilon}$, such that if $n > N_{x\varepsilon}$

$$\left| \frac{f(x+nh) - f(x)}{\phi(x+nh) - \phi(x)} - \frac{f(x)}{\phi(x)} \right| < \varepsilon.$$

Taking ε less than the distance between the nearest end-points of $V_1(k)$ and $V_2(k)$ it is plain that for every $x > X_2$, $\frac{f(x)}{\phi(x)}$ is on $V_1(k)$, which, according to Theorem 26, proves that

$$\underset{x \pm \infty}{L} \frac{f(x)}{\phi(x)} = k.$$

the center, the axis of z horizontal and the axis of y positive downward. The element of pressure is

and the total pressure is

$$P = 2k \int_0^6 yx dy.$$

z is expressed in terms of y by means of the equation of the ellipse,

$$\frac{x^2}{64} + \frac{y^2}{36} = 1.$$

Then

$$P = 2k \, 3 \int_0^6 y \sqrt{36 - y^2} \, dy.$$

Exercises

- 1. Find the pressure on the vertical parabolic gate, Fig. 51: (a) if the edge AB lies in the surface of the water; (b) if the edge AB lies 5 feet below the surface.
- 2. Find the pressure on a vertical semicircular gate whose diameter, 10 feet long, lies in the surface of the water.
 - 73. Arithmetic Mean. The arithmetic mean, A, of a series of n numbers, a1, a2, a3,...,...,...,...,...