Variations on Regularization

Daniel Keren and Michael Werman

Dept. of Computer Science
The Hebrew University of Jerusalem
91904 Jerusalem, Israel

Abstract

Regularization has become an important tool for solving many ill-posed problems in approximation theory, for example in computer vision: surface reconstruction, optical flow, shape from shading and more. This paper poses the following questions: is the approach taken in regularization always the correct one, and to what extent are the results of regularization reliable. For example, suppose regularization has been used to reconstruct a surface from sparse data. How strongly can the height of the surface at a certain point be relied upon? These questions are answered by defining a probability distribution on the class of surfaces considered, and computing its expectation and variance. The variance can be used, for instance, to construct a "safety strip" around the interpolated surface that should not be entered if collision with the surface is to be avoided.

1 Introduction

Regularization is used in image processing to reconstruct an object when partial data about it is given [15, 16, 7, 3, 17]. Reconstruction of surfaces given partial information has been studied in many other fields, for example geology [4], and electronics [2]. The data can be sparse - e.g. the height of a small number of points on a surface, or dense but incomplete - e.g. the case of optical flow and shape from shading [6] where data is available at many points but consists of the derivative's value in a certain direction only. The first difficulty in solving this problem stems from the multitude of possible solutions each agreeing with the partial data; which one should be chosen? Also, data instances which are not compatible with others can cause singularities in the solution. The regularization approach overcomes these difficulties by choosing among the possible objects one which approximates the given data and is also "smooth". This embodies an important assumption - that the "smoother" the object is, the more probable it is. Formally, a *cost function* $M(f)$ is defined for every object f by $M(f) = D(f) + \lambda S(f)$, where $D(f)$ measures the distance of f from the given data, $S(f)$ measures the smoothness of f, and λ is a parameter. The f chosen is the one minimizing $M()$. A typical example is surface reconstruction: the estimated (up to some degree of accuracy) height of the surface at a set of points $\{x_i, y_i\}_{i=1}^{l}$ (from now on called *sample points*) is given as $\{z_i\}_{i=1}^{l}$, and the problem is to interpolate these values to the rest of the surface. Then $D(f) = \sum_{i=1}^{l} \frac{1}{\sigma_i^2}[f(x_i, y_i) - z_i]^2$ and $S(f) = \iint (f_{xx}^2 + 2f_{xy}^2 + f_{yy}^2) dx dy$ are usually chosen, were σ_i is an indicator for the measurement reliability at the point (x_i, y_i) (for simplicity it will be assumed that $\sigma_i = 1$). As λ grows, the resulting surface tends to be smoother - possibly at the cost of not being close to the sample points. The f minimizing $M()$ can be found by variational methods. From now on we will deal with the surface reconstruction problem in its one and two-dimensional forms, but the ideas and results can be extended to other problems.

This work poses the two following questions:

1) Why should f be chosen as the function *minimizing* M? It is known that f is a random choice from a probability space that assigns higher probabilities to functions with small $M()$. But the element with the highest probability is not always the best representative of the probability space!

2) How reliable are the values of f at certain points? Suppose that f is the surface chosen and $f(x) = y$. Now, what if there are many functions g such that $M(g)$ is almost equal to $M(f)$ but $g(x)$ differs substantialy from $f(x)$? Naturally, the value $f(x)$ should be assigned a high degree of uncertainty. If x is close to many sample points, intuition tells us that its interpolated value is more reliable than that of an x that is situated far away from the sample points. Thus the measure of reliability should be pointwise and not a global one for f.

An attempt is made to answer these questions in a probabalistic setting. Each f is assigned a probability proportional to $e^{-M(f)}$ (for a justification of this choice, see [14]), thus generating a probability distribution over all functions. The distribution of the f-values at x is computed for each x. Note that this approach takes into account *all* functions and does not use a single function as the representative of the space.

Related work was done on the problem of estimating linear functionals on Banach and Hilbert spaces given partial information [18, 11, 10]. Work more closely related are [14, 19]. In [19], an *a priori* distribution is assumed on the reconstructed surface. In [14], Monte-Carlo methods are heuristicly used to compute a measure of uncertainty. The novelty in the work presented lies in the probabalistic framework which allows definition and computation of

CH2898-5/90/0000/0093$01.00 © 1990 IEEE

variance, which is a natrual measure of uncertainty.
The organization of this paper is as follows: In Section 2 the problems are solved when f is constrained to a finite-dimensional subspace of the function space. In Section 3 the results will be extended to the complete function space by using the theory of *Hilbert space*. In Section 4 it will be shown that the values of the expectation and variance on certain "small" finite-dimensional subspaces approximate the values on the whole function space. In Section 5 the results will be extended to the 2-D surface interpolation problem. In Section 6 the problem of "active interpolation" will be presented, and in Section 7 experimental results are given. Due to space limitations, the proofs are omitted and can be found in [8].

2 The Case of a Finite Dimensional Space

In this section the two questions posed in the introduction are solved when the functions are restricted to a finite-dimensional subspace. The problem of one-dimensional surface reconstruction will be dealt with first. The input to this problem is a set of points $\{(x_i, y_i)\}_{i=1}^{l}$, $0 \leq x_i \leq 1$, and the regularization approach is to find the f minimizing the cost function

$$M(f) = \sum_{i=1}^{l}[f(x_i) - y_i]^2 + \lambda\int_0^1 [f''(v)]^2 dv \qquad (1)$$

In order to avoid drowning in a sea of indexes, the solution will be carried out for a 2-dimensional subspace, $\text{span}\{f, g\}$. For every function in this space $M(af+bg) = \sum_{i=1}^{l}[af(x_i)+bg(x_i)-y_i]^2+\lambda\int_0^1[af''(v)+bg''(v)]^2dv$. This is actually a bilinear form in a and b which can be written as $(a,b)\Delta(a,b)^T + C_a a + C_b b + C$, where Δ is a 2×2 matrix given by

$$\Delta_{11} = \sum_{i=1}^{l} f^2(x_i) + \lambda\int_0^1 [f''(v)]^2 dv, \quad \Delta_{22} = \sum_{i=1}^{l} g^2(x_i) + \lambda\int_0^1 [g''(v)]^2 dv,$$

$$\Delta_{12} = \Delta_{21} = \sum_{i=1}^{l} f(x_i)g(x_i) + \lambda\int_0^1 f''(v)g''(v)dv,$$

$$C_a = -2\sum_{i=1}^{l} y_i f(x_i), \quad C_b = -2\sum_{i=1}^{l} y_i g(x_i), \quad C = \sum_{i=1}^{l} y_i^2.$$

As noted before, the *probability* of $af+bg$ is defined to be $\frac{1}{K}e^{-M(af+bg)}$, and the *expectation at the point* x, denoted E_x, will therefore be $\frac{1}{K}\int_{\mathcal{H}} h(x)e^{-M(h)}dh$ where $\mathcal{H} = \text{span}\{f, g\}$ and $K = \int_{\mathcal{H}} e^{-M(h)}dh$ is the normalizing factor.

Substituting the explicit expression for $M()$ gives

$$E_x = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[af(x)+bg(x)]e^{-[(a,b)\Delta(a,b)^T+C_a a+C_b b]}dadb}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-[(a,b)\Delta(a,b)^T+C_a a+C_b b]}dadb}$$

This expression can be calculated by the substitution $(u,v)^T = A(a,b)^T$ where $A^2 = \Delta$. In the next section it will be shown that if f, g are linearly independent Δ is positive definite, and thus has a root. The result for E_x turns out to be $-\frac{1}{2}(C_a, C_b)\Delta^{-1}(f(x), g(x))^T$. However, this result can be obtained more easily - look for the a, b minimizing $M(af+bg)$; they are readily seen to be $(a,b) = -\frac{1}{2}(C_a, C_b)\Delta^{-1}$. So the value at x of the cost minimizing function is $-\frac{1}{2}(C_a, C_b)\Delta^{-1}(f(x), g(x))^T$, which is equal to the expectation. In the next section it will be shown that this equality is typical for *quadratic* cost functions $M()$.

The *variance at* x, V_x, is equal to

$$V_x = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[af(x)+bg(x)-E_x]^2 e^{-[(a,b)\Delta(a,b)^T+C_a a+C_b b]}dadb}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-[(a,b)\Delta(a,b)^T+C_a a+C_b b]}dadb}$$

After some cumbersome manipulations this turns out to be $\frac{1}{2}(f(x), g(x))\Delta^{-1}(f(x), g(x))^T$. Similarly it is seen that this result is true for any finite-dimensional subspace:

Lemma 1 *If* $\mathcal{H} = span\{f_1, f_2 ... f_n\}$ *where* $\{f_1, f_2 ... f_n\}$ *are linearly independent and the* $n \times n$ *matrix* Δ *is defined by* $\Delta_{i,j} = \sum_{k=1}^{l} f_i(x_k)f_j(x_k) + \lambda\int_0^1 f_i''(v)f_j''(v)dv$, *and* $C_j = -2\sum_{i=1}^{l} y_i f_j(x_i)$, *then*

$$E_x = -\frac{1}{2}(C_1, C_2 ... C_n)\Delta^{-1}(f_1(x), f_2(x) ... f_n(x))^T$$

$V_x = \frac{1}{2}(f_1(x), f_2(x) ... f_n(x))\Delta^{-1}(f_1(x), f_2(x) ... f_n(x))^T$ *where* $\mathcal{H}$ *is the probability space and* $M()$ *if defined as in Equation 1.*

It is interesting to observe that V_x does not depend on the set $\{y_i\}_{i=1}^{l}$.

3 The Infinite Dimensional Case

3.1 Some Notions of Hilbert Space

In order to extend the definitions of expectation and variance to the set of all functions the notion of Hilbert space is used, as it is a generalization of finite dimensional spaces. A *real Hilbert space* H is a vector space over the reals $\mathcal{R}$ with an *inner product* $(\cdot,\cdot) : H \times H \rightarrow \mathcal{R}$. The *norm* of $x \in H$, $\|x\|$, is defined as $\sqrt{(x,x)}$. For an introduction to Hilbert space, see [20].
A *linear functional* on H is a linear mapping $L{:}H \rightarrow \mathcal{R}$. The

norm of L, $\|L\|$, is defined as $\sup_{x \in H} \frac{|L(x)|}{\|x\|}$. If it is finite L will be called *bounded*. An important theorem about bounded linear functionals is the -

Riesz representation theorem: if $L: H \to \mathcal{R}$ is a bounded linear functional there exists a unique element y of H such that $(x, y) = L(x)$ for all $x \in H$. Furthermore, $\|L\| = \|y\|$.

In the previous section it was needed to compute integrals on finite-dimensional spaces. In order to generalize these results, the theory of integration theory on Hilbert spaces is used [9]. Without delving into the mathematics involved, it suffices to say that it is possible to define a measure on a Hilbert space that assigns low values to sets consisting of elements with a large norm and thus suits our purposes.

3.2 Expectation and Variance in Hilbert Space

The Hilbert space suiting the framework of regularization is the space $L_2^{(2)}$ of all functions f with a square integrable distributional second derivative. This is because the cost function $M()$ of Equation 1 contains the integral of the square of the second derivative. This is an example of a *Sobolev space* [1]. As a result of the *Sobolev imbedding theorem* it follows that these functions are continuous, hence defined at every point. The inner product on this space is defined as $(f, g) = \sum_{i=1}^{l} f(x_i)g(x_i) + \lambda \int_0^1 f''(v)g''(v)dv$.

Lemma 2 *For $l \geq 2$ this defines an inner product on $L_2^{(2)}$.*

This inner product actually appears in Lemma 1, as the (i, j)-th element of the matrix Δ is (f_i, f_j). An easy exercise in Hilbert space shows that such a matrix, called "gram matrix" is positive definite iff the corresponding elements are linearly independent.
Another result needed in the sequel is -

Lemma 3 *For every $0 \leq x \leq 1$, the linear functional $L_x : L_2^{(2)} \to \mathcal{R}$ defined by $L_x(f) = f(x)$ is bounded by $\frac{4}{(x_2 - x_1)\min\{1, \sqrt{\lambda}\}}$.*

By the Riesz representation theorem, there exists for every $0 \leq x \leq 1$ a function $h_x \in L_2^{(2)}$ such that $(f, h_x) = f(x)$ for every $f \in L_2^{(2)}$. Using this fact the cost function $M()$ can be expressed as $M(f) = \sum_{i=1}^{l} y_i^2 + \|f\|^2 - 2(f, \sum_{i=1}^{l} y_i h_{x_i})$.

Defining $f_0 = \sum_{i=1}^{l} y_i h_{x_i}$ gives $M(f) = \sum_{i=1}^{l} y_i^2 + \|f\|^2 - 2(f, f_0) = \sum_{i=1}^{l} y_i^2 + \|f - f_0\|^2 - \|f_0\|^2$

Lemma 4 *$M()$ attains its minimal value at f_0.*

Theorem 3.1 *for all $0 \leq x \leq 1$,*

$$V_x = \frac{1}{2}\|h_x\|^2 = \frac{1}{2}h_x(x) = \frac{1}{2} \sup_{f \in L_2^{(2)}} \frac{f^2(x)}{\sum_{i=1}^{l} f^2(x_i) + \lambda \int_0^1 [f''(v)]^2 dv}$$

From this theorem and Lemma 3 lower and upper bounds on the variance can be derived:

Lemma 5 *for all $0 \leq x \leq 1$,*

$$\frac{1}{2l} \leq V_x \leq \frac{8}{\max_{1 \leq i,j \leq n} (x_j - x_i)^2 \min\{1, \lambda\}}$$

It can be shown that

$$\nu\{f : a \leq f(x) \leq b\} = \frac{1}{\sqrt{2\pi V_x}} \int_a^b e^{\frac{-(v - E_x)^2}{2V_x}} dv$$

and so $f(x)$ is normally distributed. This allows to give confidence intervals on the height of the interpolated surface.
The rightmost expression in Theorem 3.1 has geometric meaning, and indeed could have been taken as a *definition* of the uncertainty at x. Suppose x is at a great distance from the $\{x_i\}$. In that case, even if the $\{f(x_i)\}$ and $\int_0^1 [f''(v)]^2 dv$ are small then $f(x)$ can be large, because although the function varies slowly the distance of x from the x_i allows f to increase. This agrees with the intuitive observation in the introduction - points surrounded by many sample points should be associated with lower uncertainty than points which are situated far away from the sample points.
The most direct way to compute V_x is to find h_x. From Theorem 3.1 it is enough to compute $h_x(x)$.

Theorem 3.2 *h_x is a cubic spline with knots at $\{x_i\}_{i=1}^{l}$ and x.*

4 Approximation

In this section it is shown that V_x can be approximated on a finite dimensional subspace of $L_2^{(2)}$. This allows faster computations in some cases. Since the norm of an operator is to be evaluated, it is desirable to find an increasing sequence of subspaces S_n such that the norm of the operator on the "supplement" between S_{n-1} and S_n is small. Such a sequence is $S_n = \text{span}\{1, x, \sin(\pi x), \sin(2\pi x) \ldots \sin(n\pi x)\}$. $S = \bigcup S_n$ is *dense* in $L_2^{(2)}$, meaning that every element of the space $L_2^{(2)}$ can be approximated in norm by an element of S. This follows from Fourier theory - the odd trigonometric functions are dense in the integral part of the norm, and by integrating twice the completeness in the norm follows. This integration adds the functions 1 and x. The norm of an operator when restricted to a dense subspace of a Hilbert space is the same as the norm on the space. Thus it suffices to compute the norm on S.

Lemma 6 *In order to approximate the norm up to a factor of $1 + c$, it is enough to evaluate it in a subspace of dimension $\frac{1}{c}\sqrt{\frac{l}{\lambda}}$.*

5 The 2-D Case

The 2-D case, e.g surface interpolation, is similar to the 1-D problem in principle. Given a set of sample points $\{(x_i, y_i)\}_{i=1}^{l}$ with z_i the height at the point (x_i, y_i), a function f that minimizes $M(f) = \sum_{i=1}^{l}[f(x_i, y_i) - z_i]^2 + \lambda\int_0^1\int_0^1 (f_{xx}^2 + 2f_{xy}^2 + f_{yy}^2)dxdy$ is sought. Proceeding as in the 1-D case, Lemmas 1,4 and Theorem 3.1 go unchanged, except for the obvious change in the Hilbert space concerned and the inner product definition. Lemma 2 is replaced by -

Lemma 7 *If the set* $\{(x_i, y_i)\}_{i=1}^{l}$ *contains three non-collinear points, the inner product*

$(f, g) = \sum_{i=1}^{l} f(x_i, y_i)g(x_i, y_i) + \lambda\int_0^1\int_0^1 (f_{xx}g_{xx} + 2f_{xy}g_{xy} + f_{yy}g_{yy})dxdy$ *defines an inner product on the Hilbert space of 2-D functions concerned.*

The analogue of Lemma 5 is -

Lemma 8 *If the variance at the point* (x, y) *is denoted* $V_{x,y}$, *then*

$$\frac{1}{2l} \leq V_{x,y} \leq \frac{C}{\max_{1\leq i,j,k\leq l}(S_{P_iP_jP_k}/E_{P_iP_jP_k})^2 \min\{1, \lambda\}}$$

where $S_{P_iP_jP_k}$ *is the area of the triangle with vertices* P_i, P_j, P_k, *and* $E_{P_iP_jP_k}$ *the length of the longest edge of that triangle.*

This is an interesting result, which agrees with intuition: the bound on the variance is larger as the three points approach collinearity. It is seen that not only the *area* of the triangle is related to the variance, but also its *shape*: an equilateral triangle and a "flat" triangle give rise to greatly different variances even if their areas are the same. An example of this is given in Section 7.

We would like to have an analogue of Theorem 3.2; however, the *reproducing kernel* [5] for the inner product defined in Lemma 7 can be described only as an infinite series.

Just as the variance in the 1-D case could be computed on a "small" subspace, it can be in the 2-D case as well. The formalism is the same as in Section 4, and the base used is a combination of the bases used in Section 4 in the x and y directions: $\{1, x, y, xy, \sin(\pi x), y\sin(\pi x), \sin(\pi x)\sin(\pi y), x\sin(\pi y), \sin(\pi y)...\}$. However, it turns out that in order to approximate the norm up to a factor of $1 + c$, it is necessary to use $\frac{l}{c\sqrt{\lambda}}$ basis functions.

6 Active Interpolation

The newly developing field of "Active Vision" [13] deals with the possibility of improving the knowledge on some observed phenomena by taking measurements designed to decrease the uncertainty related with it. In the case of regularization, this could be interpreted as sampling the data in points that would make the variance as small as possible. The measure of variance used in this work was the integral of the point variances over the domain in question (the unit interval or unit square). Specifically, the question posed was: for some l, what are the l points that will minimize this integral? This problem poses a numerical difficulty because the function connecting the points to the variance is rather complicated. Nonetheless it can be solved using methods that do not require computing the function's derivatives, such as Powell's method [12]. The answers depend considerably on λ, the weight given to the smoothing part of $M()$: as λ decreases the points tend to be closer, while for large λ the points tend to be spread out and frequently lie on the region's boundary. This can be explained as follows: as λ increases non-smooth functions are assigned very low probabilities, and thus have less effect on the variance. Smooth functions, however, can be sampled at points spatially spread apart without losing too much information.

The active interpolation paradigm can also be applied when some sampling has already been done, and a new sample point - or points - are to be chosen so the resulting variance is minimal.

The next tables give, for some values of l and λ, the best points in which to sample one and two-dimensional functions.

- One-dimensional results:

	$\lambda = 1$	$\lambda = 0.1$	$\lambda = .001$
$l = 2$	0, 1	0.02, 0.98	0.2, 0.8
$l = 3$	0, 0.5, 1	0, 0.5, 1	0.09, 0.5, 0.91
$l = 4$	0, 0, 1, 1	0, 0.18, 0.82, 1	0.05, 0.35, 0.65, 0.95

- Two-dimensional results:

	$\lambda = 1$
$l = 3$	(0.72,1), (0,0.3), (1,0)
$l = 4$	(0,0), (0,1), (1,0), (1,1)

7 Experimental Results

Some tests were run in the one and two-dimensional cases to demonstrate the concept of variance introduced in this work. The results are presented in the 1-D case by graphs in which the horizontal axis stands for the points and the vertical axis for the variance at those points, and similarly for the 2-D results where the variance is described as a surface.

In Figures 1-4, the sample points are {0,1} and λ is changing from 1 to 0.001. It can be seen that not only the height but the *shape* of the variance map changes completely as λ varies. For large λ, smooth functions are given much higher probability than oscillating functions

and thus the linear functions influence the variance considerably. This reflects in the fact that 0.5 is the most reliable point, because for linear functions the value at 0.5 is the average of the value at the sample points; but of all the convex combinations of two identical random variables, the average is the one of lowest variance. However, as λ decreases, the variance increases as one moves away from the sample points.

In Figures 5-7 the sample points are {0,0.1,0.2,0.3...0.9,1} and also here λ is decreasing. The same phenomena as for the previous examples occures, although for different values of λ.

In Figures 8-11 (2-D results) $\lambda = 1$ and the sample points change. The variance maps can be computed for any number of points; the results are presented for three points as they are intuitively appealing. In Figure 8, the sample points are {(0,0), (1,0), (0,1)} and the increase in the variance towards the point (1,1) is obvious. In Figure 9, the sample points are the optimal ones from the table in Section 6: Indeed, the variance map is much more balanced and the total variance is lower. In Figure 10 The sample points are {(0,0), (1,1), (0.4,0.6)}: This produces a variance map that has low values across the main diagonal where the points are situated but increases rapidly as one moves away from it. In Figure 11 the sample points are {(0.3,0.3), (0.747,0.3), (0.3,0.747)}: the area of the triangle bounded by them is the same as in Figure 10, but the more compact form of the triangle results in a more balanced variance map and in smaller total variance, as implied by Lemma 8.

References

[1] R. A. Adams. *Sobolev Spaces.* Academic Press, 1975.

[2] H. Akima. Bivariate interpolation and smooth surface fitting based on local procedures. *Comm. ACM*, 17:26–31, 1974.

[3] M. Bertero, T.A Poggio, and V. Torre. Ill-posed problems in early vision. *Proceedings of the IEEE*, 8:869–889, 1988.

[4] R.J Chorley. *Spatial Analysis in Geomorphology.* Methuen and Co., 1972.

[5] E. Hille. Introduction to the general theory of reproducing kernels. *Rocky Mountain Journal of Mathematics*, 2:321–368, 1972.

[6] B. Horn. *Robot Vision.* MIT Press, 1986.

[7] B.K.P Horn and B.G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[8] D. Keren and M. Werman. *Variations on Regularization.* TR No. 90-6, Department of Computer Science, Hebrew University of Jerusalem, Israel, 1990.

[9] J. Kuelbs, F.M. Larkin, and J.A. Willaimson. Weak probability disributions on reproducing kernel hilbert spaces. *Rocky Mountain Journal of Mathematics*, 2:369–378, 1972.

[10] F.M. Larkin. Gaussian measure in hilbert space and applications in numerical analysis. *Rocky Mountain Journal of Mathematics*, 2:379–421, 1972.

[11] D. Lee. Approximation of linear operators on a wiener space. *Rocky Mountain Journal of Mathematics*, 16:641–659, 1986.

[12] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. *Numerical Recipes.* Cambridge University Press, 1986.

[13] A. Shmuel and M. Werman. Active vision: 3d from an image sequence. In *ICPR*, 1990.

[14] R. Szeliski. Regularization uses fractal priors. In *National Conference on Artificial Intelligence*, pages, 1987.

[15] D. Terzopoulos. Multi-level surface reconstruction. In A. Rosenfeld, editor, *Multiresolution Image Processing and Analysis*, Springer-Verlag, 1984.

[16] D. Terzopoulos. Regularization of visual problems involving discontinuities. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8:413–424, August 1986.

[17] A.N Tikhonov and V.Y Arsenin. *Solution of Ill-Posed Problems.* Winston and Sons, 1977.

[18] G.W. Wasilkowski. Optimal algorithms for linear problems with gaussian measures. *Rocky Mountain Journal of Mathematics*, 16:727–749, 1986.

[19] G. Whaba. *Bayesian Confidence Intervals for the Cross-Validated Smoothing Spline.* TR No. 645, Department of Statistics, University of Wisconsin, Madison, 1981.

[20] N. Young. *An Introduction to Hilbert Space.* Cambridge Mathematical Textbooks, 1988.