Shifted Base Functions: An Efficient and Versatile New Tool in Optics

K.–H. Brenner
Chair of Optoelectronics, University of Heidelberg
B6, 23-29, D 68131 Mannheim, Germany
brenner@ziti.uni-heidelberg.de

Abstract. In this paper, we are dealing with a new, accurate, fast and efficient way to describe and reconstruct smooth optical surfaces analytically. The applications of this representation cover almost all fields of optical metrology. Especially in the field of free-form surfaces, this method is applicable, since it offers sufficient flexibility. The method offers both, surface approximation and surface reconstruction. With surface approximation, measured data can be converted into an analytic description, which is useful to generate interpolated data or derived optical quantities. With surface reconstruction, arbitrary surfaces can be reconstructed only from slope data without any integration scheme.

1. Introduction
For the measurement of optical surfaces or wave fronts, there are numerous methods in the literature [1]. These methods can coarsely be divided into surface measurements, like e.g. mechanical scanning or interferometry and slope or gradient measurements, such as Hartmann-Shack [2] or Shape-from-Shading [3]. Many of these methods provide data on a regular grid, but there are also methods, where the data are spaced non-uniformly. In many cases, not all measurement samples can be used since there might be defects, making some of the samples inaccurate. A third and very important aspect concerns the measurement errors. An algorithm which describes or reconstructs optical surfaces must have an averaging property, since we are generally only interested in the shape of the surface and not in the high frequency measurement noise. This is especially important for gradient data, since measurement errors can result in a loss of integrability. This means that if the correct data are curl-free, errors in the measurement can introduce a curl, which makes the integral path dependent. Thus, in the most general case, we are dealing with a large set of $M$ non-uniformly sampled data, appearing as a list:

$$\left( \mathbf{r}_i, y_i \right), i = 1..M$$

(1)

where $\mathbf{r}_i$ is the position coordinate of the $i$-th measurement and $y_i$ is the measured value, which can be scalar or vector-valued. For height measurements or phase measurements, this quantity is clearly scalar. For slope measurements, this quantity is a vector with two dimensions. In even more general situations, like e.g. modelling of the diffraction efficiencies of a grating for different orders,
wavelength and polarisations, \( \mathbf{r} \) can be D-dimensional and may not have the meaning of a position. Likewise, \( \mathbf{y} \) can be F-dimensional, designating the measured quantities for the configuration \( \mathbf{r} \).

In industrial inspection, where accuracy demands are very high, the positional sampling can be in the micron range, generating data lists where \( M \) is in the order of millions of samples. With modern storage technology, there is no problem in storing these measurement lists in a computer, but it is clear, from an information theoretic point of view, that this representation is highly redundant. If we measure e.g. an optical asphere, the sequence of height values changes only slightly and in a predictable way. In other words, there is a high degree of correlation between adjacent points. Furthermore, if we want to derive analytic quantities from these measured data, we also need to be able to acquire interpolated values, i.e. values which do not fall on the measurement grid. In the example of the optical asphere, we may be interested e.g. in the zonal focal length, which is determined by the local radius of curvature. In summary, a method for representing or reconstructing optical surfaces should be able to handle the following situations:

- Data on a nonuniform grid
- Missing samples
- Noisy data
- Enforcement of integrability
- Large number of samples
- Arbitrary dimensions

For the problem of surface representation, there are a number of established methods which will be addressed later. These include the Zernike decomposition [4] and the Spline-interpolation [5]. For the problem of surface reconstruction from gradient data, the most widely used method is an algebraic integration scheme by Southwell [6]. Other approaches make use of the Laplace equation and Fourier transforms to realize the integration [7] and recent methods apply Radial basis function (RBF) interpolation [8] to solve this problem. In the following, we will briefly discuss the existing approaches from a very general point of view. We will look at RBF interpolation and we will discuss the issue of interpolation vs. approximation also from a general perspective. Then we will present the new method and we will apply it to surface approximation and surface reconstruction.

2. Different Approaches for Data Modelling

2.1. Model representation of data

For these reasons stated previously, it is desirable, to have an analytic description of data. In many cases, a surface \( f \) can be represented by a model function \( m \):

\[
f(\mathbf{r}) = m(\mathbf{r}, \mathbf{p})
\]

(2)

where \( \mathbf{p} \) is a parameter vector and the coefficients of this parameter vector can be determined by a least square fit, i.e. to minimize

\[
\Phi(\mathbf{p}) = \sum \| f(\mathbf{r}_i) - m(\mathbf{r}_i, \mathbf{p}) \| \rightarrow \min
\]

(3)

Then the coefficients follow by solving the algebraic equations:
With the dimension of the parameter vector, we determine the approximation order, i.e. how many coefficients are used for the approximation. Typical representatives of this approach are the Zernike polynomials and lens surface models as specified e.g. in ray trace software. The standard polynomials \( 1, x, x^2, ... \) are known to be very unsuitable for optical surfaces. For a simple sphere, e.g. it is easy to see that the approximation only improves up to an order of 8. At this order, the approximation is still not satisfactory for optical applications. For orders beyond 8, oscillations appear, which illustrates that a finite order approximation is not possible with simple polynomials. For Zernike polynomials we have to require, that the model has a rotational symmetry and we have to determine the rotational origin with high accuracy to obtain reliable data. Consequently, Zernike polynomials are only useful for modeling pupil distributions. For non-rotational symmetric applications, Zernike polynomials are inadequate.

2.2. Interpolation

Especially for free-form surfaces, there is in general no analytic model, which could be used for a continuous representation of surfaces. In this case, interpolation methods are applied. The most widely used interpolation method is Spline-interpolation or more prominently, NURBS (Non-Uniform Redundant B-Splines) modelling, which originated in the field of computer aided design for modelling large objects and is now also entering applications in optics. In this approach, a spline \( S \) is a piecewise polynomial function and the data area consists of polynomial pieces \( P \). Interpolation is suitable for uniform and also non-uniform data grids and has a certain locality, in that a change of one point only affects a finite neighbourhood. A general problem with interpolation methods - in contrast to least-square approximation - is that the data are assumed to be free of measurement errors. The interpolating polynomial will always reproduce the measurement data on the grid exactly. In the presence of errors, any interpolation method generates a (noisy) hyper sphere, which goes through all the measurement points. Consequently, the interpolated data maintain the errors from the measurement. This is different in approximation schemes, where a best fit is generated, which is smooth, but does not go through all measurement points. The best known example for this is linear regression, where a noisy cloud of dots is approximated by a best fit line. With interpolation through all the measurement points, we would achieve a jagged curve, reproducing all the measurement errors.

2.3. Radial Basis Function (RBF) Interpolation

Radial Basis Function Interpolation or short RBF-interpolation is a new interpolation approach, which has started in the mathematical literature and is now also entering the applied fields. In RBF-interpolation, a surface \( s \) is modeled by a linear superposition

\[ s(\mathbf{r}) = \sum c_i \varphi(|\mathbf{r} - \mathbf{r}_i|) \]  

(5)

where \( \varphi \) is a fixed, univariate function, \(|\cdot|\) is a norm of the vector (typically the Euclidean norm) and the coefficients \( c \) are real numbers. The basis positions \( \mathbf{r}_i \) should be spaced approximately equidistant, but there is no requirement for a regular grid. The interpolation condition \( s(\mathbf{r}_i) = y_i \) provides a linear system
Ac = b \hspace{1cm} (6)

for solving for the coefficient vector c. RBF-interpolation has recently been applied in the context of optics for the reconstruction of heights from slope values [8]. As with any interpolation method, the result is affected by measurement errors since the approach is constructed so that \( s(r_i) = y_i \).

3. SBF-Approximation of functions

The "Shifted base function" approach (SBF), presented here, is inspired by RBF-interpolation, with two significant differences: first: the approach is not an interpolation scheme but instead a least square approximation. Consequently measurement errors are reduced in the model. Second it does not require the basis function to be rotationally symmetric. Thus the extent of the base function can be adapted to the degree of smoothness of the function to be modelled. In the following, we will describe the approximation theory of Shifted Base Functions (SBF) for the most general case and then will consider a special examples. Finally we apply this approximation theory to gradient data and describe, how an efficient and noise tolerant reconstruction from slope date can be achieved with this approach.

We consider \( r \) as vector in a D-dimensional position space

\[
r = (x_0, x_1, \ldots, x_{D-1}) \in \mathbb{R}^D
\]

and the F-dimensional measurement vector for the \( i \)-th measurement:

\[
y_i = (y_{0,i}, y_{1,i}, \ldots, y_{F-1,i}) \quad i = 0..M - 1
\]

Then our model is

\[
f(r) = \sum_{k=0}^{F-1} c_k \psi(r - r_k, \sigma)
\]

Here \( c_k = (c_{0,k}, c_{1,k}, \ldots, c_{F-1,k}) \) can be considered as a set \( k = 0..B-1 \) of vectors of dimension \( F \) and \( r_k = (x_{0,k}, x_{1,k}, \ldots, x_{D-1,k}) \) is a set of \( k = 0..B-1 \) vectors of dimension \( D \), which we call the base points. Then our model states that the vector \( f \) is a weighted superposition of one single base-function \( \psi \), shifted to the base points \( r_k \). The vector \( \sigma \) has the same dimension as the position vector. For the choice of the base function \( \psi \), there are very little restrictions. Typical base functions are chosen to be continuous and nonzero over a reasonably large range. One example is a Gaussian function. The function we used in all our examples is essentially the amplitude of a Lorenzian function:

\[
\psi(r, \sigma) = \left(1 + \sum_{j=0}^{D-1} \left( \frac{x_j}{\sigma_j} \right)^2 \right)^{-1/2}
\]

\[
(10)
\]
It is also called "inverse multiquadric" in the literature. The assumption here is that the measurement values $y_i$ were taken at the measurement points $r_i = (x_{0,i}, x_{1,i}, \ldots x_{D-1,i})$. Consequently, we try to minimize the error function

$$
\Phi(c_i) = \sum_i (f(r_i) - y_i)^T (f(r_i) - y_i)
$$

by solving the set of equations

$$
\frac{1}{2} \frac{\partial \Phi}{\partial c_k} = 0
$$

By inserting the definition of the model function into eq. 11 and taking the partial derivative, we obtain a linear set of equations

$$
\sum_{k=0}^{B-1} c_{j,k} M_{k,k} = V_{j,k}, \quad j = 0..F-1
$$

where the matrix

$$
M_{k,k} = \sum_{i=0}^{M-1} \psi(r_i - r_k) \psi(r_i - r_k)
$$

is real and symmetric and the right-hand side vectors are

$$
V_{j,k} = \sum_{i=0}^{M-1} y_{j,i} \psi(r_i - r_k)
$$

The coefficients $c_{j,k}$ can thus be solved with standard algebraic techniques. Note, that the matrix elements in eq. 14 are only dependent on the base points $r_k$, and the measurement points $r_i$ and are the same for all the vector components. Thus this matrix has to be calculated only once. Another important observation is that both, the matrix $M$ and the right-hand vector in eq. 14 are accumulative in the sense that if e.g. $M$-1 measurements

$$(r_i, y_i), \ i = 0..M-2$$

are stored in the matrix and the vectors $V_j$, then the next measurement updates $M$ and $V_j$ according to

$$
M^N_{k,k} = M^{N-1}_{k,k} + \psi(r_{N-1} - r_k) \psi(r_{N-1} - r_k)
$$

$$
V^N_{j,k} = V^{N-1}_{j,k} + y_{j,N-1} \psi(r_{N-1} - r_k)
$$

From this property, it is clear that the measurement values can also be applied with weights. As we can see, if a certain measurement with index $i$ is added e.g. twice, then $V$ is increased by $2y_{j,i} \psi(r_i - r_k)$.
and $M$ will be increased by $2\psi(r_i - r_k)\psi(r_i - r_k)$. Likewise, if we enter a measurement with a weight $w_i$, the increment to $M$ and $V$ will be

$$
\Delta M_{k,i} = w_i \psi(r_i - r_k)\psi(r_i - r_k) \quad (17a)
$$

$$
\Delta V_{j,k} = w_j y_{j,k} \psi(r_i - r_k) \quad (17b)
$$

Thus, if the measurements are not equally reliable, the reliability measure could be added as a weight factor.

3.1. Application example: Approximation of a perfect focussing lens with SSB

In order to demonstrate the efficiency and accuracy of this type of approximation, we chose the example of a perfect focussing lens. A perfect focussing lens is constructed so that all rays with height $x$ propagating parallel to the $z$-axis should be focused to a single point. A plano-convex spherical lens in that sense is not a perfect focussing lens.

![Figure 1 Model geometry for a perfect focussing lens](image)

The back surface is assumed to be flat and for the front surface there is no simple analytic description in the form $z(x)$, but we once derived a parameterized analytic description in the form:

$$
x(s) = -\frac{s}{n} (\lambda_1 + n\lambda_2)
$$

$$
z(s) = D - \lambda_1 \sqrt{1 - (s/n)^2}
$$

where $s$ ranges from $-NA$ to $NA$ and the various quantities are:

$$
Q = \frac{R}{NA}, w = Q\sqrt{1-NA^2}, D = \frac{Q - w}{n-1}, \lambda_2 = \frac{w}{\sqrt{1-s^2}}, \lambda_1 = \frac{Q - \lambda_2}{n - \sqrt{1 - (s/n)^2}}
$$

with the parameters $n$, $R$, $w$ as shown in figure 1 and NA being the numerical aperture of the lens. Note that the front surface is neither spheric nor polynomial. For the following example, we chose a
numerical aperture of NA=0.5 and fig. 2 shows the corresponding surface sag, which is defined to be zero on axis and roughly 2.68 mm at the lens edge.

Figure 2 Surface sag of a perfect focussing lens with a diameter of 10 mm and NA = 0.5

Applying the described approximation scheme, we have to distinguish between the number of base points $B$ and the number of measurement points $M$. In all our examples, we chose $M=32$ measurement points, spaced equidistantly and a value $\sigma = 8$. The number of base points determines the number of shifted base functions used for modeling. In fig. 3 we used only $B=7$ points, which means that only the mid point and three points on each side of the axis are used as base points. The maximum approximation error here is below 1 µm in a height range of 2600 µm as shown in fig. 3.

Figure 3: Approximation error in µm for 3 points on each side

Figure 4: Approximation error in µm for 15 base points
For 15 base points and the same 32 measurement points equally spaced, the maximum approximation error is around 500 nm, again in a height range of 2600 µm as shown in fig. 4. This example illustrates, that in order to describe a surface with a precision of 0.5 µm, only 15 real coefficients are necessary.

4. Gradient approximation with shifted base functions (SBF)

Gradient approximation is a very important field of optics, since it enables the reconstruction of an optical surface only from gradient data. As we will see, in contrast to standard integration schemes, no explicit integration is needed here.

For gradient approximation, we also are dealing with position vectors in a D-dimensional position space, but the measurement values

$$y_i = (y_{0,i}, y_{1,i}, \ldots, y_{D-1,i})\quad i = 0..M - 1$$

(20)

are now to be identified with the gradient components of a scalar function $f(r)$. Therefore, the $y$-vector has the same dimension as the position vectors. If we denote $f_j = \frac{\partial f}{\partial x_j}$ and $\psi_j = \frac{\partial \psi}{\partial x_j}$ our model for gradient approximation is:

$$f_j(r) = \sum_{k=0}^{B-1} c_k \psi_j(r - r_k, \sigma)$$

(21)

In this case $c_k$ is a simple sequence of real coefficients and the base positions $r_k$ are again a set of $k = 0..B-1$ vectors of dimension $D$. Now, the model states that the gradient vector $f$ is a weighted superposition of the gradient of one single base-function $\psi$, shifted to the base points $r_k$. The base function $\psi$ can be the same as in function approximation.

Once, the optimal coefficients $c_k$ are determined, the height function $f$ can be determined without any explicit integration by:

$$f(r) = \sum_{k=0}^{B-1} c_k \psi(r - r_k, \sigma)$$

(22)

Assuming again, that the measurement values $y_i$ were taken at the measurement points $r_i$, we try to minimize exactly the same error function as before:

$$\Phi(c_k) = \sum_i (f(r_i) - y_i)^T (f(r_i) - y_i)$$

(23)

Inserting the definition of the model function (eq. 21) and taking the partial derivative, we now obtain:
\[ \sum_{k=0}^{B-1} c_j \cdot M_{k,k^*} = V_k \]  

(24)

with \( \mathbf{M} \) and \( \mathbf{V} \) now appearing as a sums over \( j \):

\[ M_{k,k^*} = \sum_{j=0}^{B-1} \sum_{i=0}^{M-1} \left( \psi_j \left( \mathbf{r}_i - \mathbf{r}_k \right) \psi_j \left( \mathbf{r}_i - \mathbf{r}_k \right) \right) \]  

(25a)

\[ V_k = \sum_{j=0}^{B-1} \left( \sum_{i=0}^{M-1} y_{j,i} \psi_j \left( \mathbf{r}_i - \mathbf{r}_k \right) \right) \]  

(25b)

The matrix \( \mathbf{M} \) is still real and symmetric and can be solved with standard algebraic techniques. Again, by interchanging the order of the summation, we observe that the matrix elements are accumulative. If \( i = 0, \ldots, M - 2 \) measurements are stored in the matrix \( \mathbf{M} \) and in the vector \( \mathbf{V} \), the next measurement updates \( \mathbf{M} \) and \( \mathbf{V} \) according to

\[ M^N_{k,k^*} = M^{N-1}_{k,k^*} + \sum_{j=0}^{B-1} \psi_j \left( \mathbf{r}_{N-1} - \mathbf{r}_k \right) \psi_j \left( \mathbf{r}_{N-1} - \mathbf{r}_k \right) \]  

(26a)

\[ V^N_k = V^{N-1}_k + \sum_{j=0}^{B-1} y_{j,N-1} \psi_j \left( \mathbf{r}_{N-1} - \mathbf{r}_k \right) \]  

(26b)

Therefore, exactly as in function approximation, the measurements can also be added with weight factors.

4.1. Application example: gradient approximation of a perfect focusing lens with SSB

Just for comparison, we first take the same example of a perfect focusing lens as before. We start with the same configuration, a perfect focusing lens with diameter 10 mm, a NA of 0.5. For the gradient we take the analytic expression:

\[ z_s = \frac{-s}{\sqrt{n^2 - s^2 - 1}} \]  

(27)

and the shifted base function is given in eq. 10. For the number of base points \( B \), we start with only seven points, i.e. the centre point and three symmetric base points spaced equally around the centre. The result still appears similar, but a look at the axis shows that the accuracy is now significantly better compared to height approximation. Also the error is now only in one direction. The reason for the latter is very simple: A function, reconstructed only from gradient information, like every integration method, has an undetermined bias value. Here we adjust the bias so that the midpoint is zero. Thus the (oscillatory) errors are shifted to the minimum. More importantly, the maximum deviation is now below 100 nm!

By increasing the number of base points from 7 to 15, the error drops to P/V-values below 5 nm in a height range of 0..2.679.491 nm, which demonstrates again, that for optical surfaces, the gradient information is more important than the height information.
4.2. Gradient approximation for 2D-surfaces

In order to test the method on 2D-data, we took the result of a Hartmann Shack measurement of a surface. As can be seen in fig. 6, the raw data were very unreliable and thus noisy. In this case the number of measurement points was 16x12. Nevertheless, by choosing only 5x5 base points, the surface shape could be reconstructed reliably. The reconstructed height, shown in fig. 7 is encoded as grey value with dark being the lowest height. The results agree very well with a reconstruction using the standard integration scheme by Southwell [6].

5. Conclusion

We have presented a new, accurate, fast and efficient way to describe and reconstruct smooth surfaces in arbitrary dimensions analytically. The method is applicable especially in the field of free-form surfaces, since it offers sufficient flexibility. It can handle data on a non-uniform grid, deal with missing samples and noisy data. For surface reconstruction from slope data, it also enforces integrability. The method can be used for both, surface approximation and surface reconstruction. Furthermore, the results demonstrated that the method also provides a very efficient description of large amounts of data and can thus also be considered as a scheme for data-compression.
Figure 6: Hartmann-Shack measurement of a surface

Figure 7: Smooth reconstruction of the height values using this method

References
[1] D. Malacara, ed., Optical Shop testing (Wiley 1978)
[2] R.K. Tyson, ed., Adaptive Optics Engineering Handbook (Marcel Dekker, 2000)
[3] R.T. Frankot, R. Chellappa, "A method for enforcing integrability in shape from shading algorithms", IEEE Trans. Pat. Anal. Mach. Intel. 10(4) (1988), 439-451
[4] Wang, J.Y. and Silva, D.E. "Wave-Front Interpretation with Zernike Polynomials." Appl. Opt. 19, 1510-1518, 1980.
[5] G. Farin, "Trends in Curve and Surface Design," Computer-Aided Design, vol. 21, no. 5, pp. 293-296, 1989.
[6] W. H. Southwell, "Wave front estimation from wave front slope measurements", J. Opt. Soc. Am. 70, (1980) 998-1006
[7] K. R. Freischlad, C. L. Koliopoulos, "Modal estimation of a wave front from difference measurements using the discrete Fourier transform", J. Opt. Soc. Am. A 3, 1852-1861 (1986)
[8] S. Ettl, J. Kaminski, M.C. Knauer, G. Häusler, "Generalized Hermite interpolation with radial basis functions considering only gradient data", in Curve and Surface fitting: Avignon 2006, pp. 141-149 (2007)