ELSEVIER

Wear 231 (1999) 206–219

WEAR

www.elsevier.com/locate/wear

A numerical method for solving rough contact problems based on the multi-level multi-summation and conjugate gradient techniques

I.A. Polonsky, L.M. Keer *

Department of Civil Engineering, Northwestern University, Evanston, IL 60208-3109, USA

Received 10 August 1998; received in revised form 23 February 1999; accepted 23 February 1999

Abstract

An alternative numerical method for solving contact problems for real rough surfaces is proposed. The real area of contact and the contact pressure distribution are determined using a single-loop iteration scheme based on the conjugate gradient method, which converges for arbitrary rough surfaces. The surface deflections and subsurface stresses are computed using an alternative two-dimensional multi-level multi-summation algorithm, which allows the summation error to be kept under the discretization error for any number of contact points. The proposed method is fast: rough contact problems for surface samples with 10^5–10^6 data points are solved on a personal computer in a few hours. The numerical algorithms are described in full detail so that an interested reader can implement the new contact solver in a computer code. Numerical examples demonstrating the method advantages are presented. The method is compared with other fast contact solvers that have emerged in the last few years.

Keywords: Rough contact problems; Multi-level multi-summation; Conjugate gradient techniques

1. Introduction

The analysis of stresses, surface deflections, and microcontact areas generated by the contact of solids with rough surfaces has long been recognized as a fundamental problem of tribology. The importance of this problem stems from the fact that some of the common modes of wear, including sliding wear, abrasive wear, fatigue wear, and surface-initiated rolling contact fatigue, are driven by roughness-induced contact stresses. Furthermore, the microcontact stresses and the real area of contact are believed to play an essential role in most of the known mechanisms of dry friction. The real area of contact also determines the electrical and thermal conductivity of contacts, while the gap distribution between rough surfaces in contact has great significance to mechanical seals. Although purely mechanistic analyses of rough contacts are rarely sufficient to solve tribological problems, which are typically complex and multi-disciplinary, such analyses are necessary for obtaining theoretical estimates of friction and wear, and are thus an essential part of tribology.

The present concern is the dry contact of rough surfaces. It is understood that in common machines, contacts (particularly rolling and/or sliding ones) are usually lubricated. However, dry contacts often arise in such areas of engineering as magnetic data storage devices, electrical brushes, and some aerospace applications. Moreover, dry contact solutions generally provide upper bounds on the roughness-induced stresses generated by lubricated contacts. Taking into account that the lubricant properties under high pressure and/or high shear rate conditions are often uncertain [1,2], such upper bounds may be quite useful to many applications. In addition, it has been suggested that the stresses generated by heavily loaded and/or low-speed lubricated contacts can be approximated by the corresponding dry contact solutions (e.g., Ref. [3]).

There are a number of approaches to contact analysis of real rough surfaces. In the profilometric theory of contact (see Ref. [4] for an in-depth discussion and references), the contact pressure is assumed to equal the material hardness throughout the real area of contact, while the surface displacements at non-contact points are neglected. The

* Corresponding author. Fax: +1-847-491-4011; E-mail: 1-keer@nwu.edu

0043-1648/99/$ - see front matter
PII: S0043-1648(99)00113-1

shape of the contact area is then determined by a purely geometrical analysis of undeformed surface topography (as measured by a profilometric device). In statistical theories of contact [4–8], the real rough surface is represented by a large number of model asperities of different heights (and sometimes different curvatures) distributed randomly over an infinite plane. The contact behavior of these asperities is usually described by the Hertz elasticity solution, although the assumptions about the nature of individual asperity contacts are not essential to the statistical theories (see author's closure in Ref. [4]). In the fractal theory of contact [9,10], the real surface is replaced by a fractal function that, while being deterministic, has a random appearance resembling that of the actual surface. By applying the mathematical theory of fractals to the model surface, analytical formulae describing the contact size distribution are obtained. The determination of microcontact areas is based on the undeformed surface topography, as in the profilometric theory. The stresses generated at individual microcontacts are calculated using the Hertz equations with a plastic cutoff, similarly to the statistical theories.

However, all of the above approaches have a serious deficiency: they use the stress–displacement response of isolated microcontacts and do not take into account interactions between microcontacts. The most commonly recognized type of microcontact interaction stems from the long-range nature of the elastic displacements generated by concentrated contacts. As the nominal contact pressure increases, asperity contacts become more crowded and the contribution of surface deflections produced by each of them to the net surface deflection at the neighboring microcontacts becomes significant [11–14]. With a further increase in the contact load, microcontacts can start to coalesce. When this happens, the rough contact theories based on the Hertz solution break down [4], as do the theories based on plastic models of isolated contacts [15–17]. However, microcontact interaction is not limited to these 'neighbor' effects. Rough surfaces are essentially multi-scale in nature; in mathematical terms, they can be described as random processes with broad spectra [18,19]. Therefore, even in the contact of nominally flat surfaces, the nominal contact pressure applied to asperities on a certain scale can vary significantly for different surface locations, depending on the underlying large-scale roughness (or waviness). Small-scale asperities located near the summit of a large-scale one will experience a heavier load than similar asperities located in a large-scale valley [20,21]. Therefore, even at light contact loads, microcontacts can interact strongly and even coalesce at some locations within the macrocontact, as was observed in the experiments of O'Callaghan and Probert [22]. More recently, accurate experimental measurements of rough contacts performed by Hendriks and Visscher [23] demonstrated the limitations of both the statistical and the fractal theories, and the need to take into account microcontact interaction when modeling rough contacts. All of the above types of microcontact interactions were included in the semi-analytical model of Polonsky et al. [20], but their model was limited in other respects.

The only existing approach that accurately accounts for microcontact interaction effects, and hence can be used for accurate calculation of roughness-induced stresses, is the direct numerical solution of contact problems for samples of real rough surfaces. Such numerical analyses have been performed by Lai and Cheng [24], Webster and Sayles [25], Seabra and Berthe [26], Ren and Lee [27], Poon and Sayles [28], and many others. These analyses were based on the numerical methods developed earlier by Kalker and van Randen [29], Kubo et al. [30], and Francis [31]. However, practical application of this approach poses serious difficulties, which also stem from the multi-scale nature of surface roughness. To account for all of the roughness components that are important to the friction or wear mechanism(s) being analyzed (or at least, as many components as possible) roughness samples having both a sufficiently short sampling interval and a sufficiently long sample length must be considered. Consequently, the number of data in the roughness sample will often be extremely large, especially in the more realistic and technologically significant three-dimensional (3D) case. For example, modern devices for 3D surface topography analysis, such as optical profilometers and atomic force microscopes, typically produce data sets of about 500 by 500 points. Although the number of points in each direction is quite moderate, a system of about 2.5×10^5 equations arises from the corresponding contact problem. Solution of such large-scale systems of equations by conventional numerical algorithms requires unacceptably long times even on modern high-speed computers.

The numerical solution of realistic 3D rough contact problems became feasible when Brandt and Lubrecht [32] developed the multi-level multi-summation (MLMS) technique, which allows the time of surface deflection computation to be reduced by orders of magnitude, as compared to conventional algorithms. The MLMS technique, in combination with the full multigrid (FMG) iteration method, has been extensively used by Lubrecht, Venner, and coworkers in their numerical studies of both smooth and wavy EHL contacts (see Ref. [33] and references therein). However, an attempt to apply the same approach to rough contacts was unsuccessful, as the FMG scheme failed to converge for realistic rough surfaces [3]. It may be argued that the convergence problems reported in Ref. [3] are specific to the particular contact solver used in that study. However, to our best knowledge, no other applications of FMG to dry rough contacts have been published in the open literature. Moreover, the application of FMG to dry contact problems appears to lack a firm mathematical basis. FMG was originally developed for solving systems of equations. The corresponding convergence theorems do not automatically apply to problems with multiple inequality constraints, such as dry contact problems. However, no

attempt to justify the application of FMG to rough contacts was made in Ref. [3].

In the present work, an alternative method of fast numerical solution of rough contact problems is presented. This method combines MLMS with an iteration scheme based on the conjugate gradient (CG) method. The application of CG to rough contact problems is based on rigorous convergence theorems. The 2D MLMS algorithm used in the present work differs significantly from the one introduced in Ref. [32], as will be discussed below. It is demonstrated that our contact solver can be applied without difficulty to dry contacts of arbitrary rough surfaces.

2. Problem formulation and conventional methods of solution

2.1. Discretization

Numerical solution of 3D rough contact problems is commonly based on the following discrete formulation [30,31]. The contacting rough surfaces are described by two sets of surface height data, each corresponding to a uniformly spaced rectangular array of surface points (i.e.,, a surface grid). Such surface topography data can be obtained using a 3D surface imaging device (e.g., an optical profilometer, a relocation mechanical profilometer, or an atomic force microscope), or generated by a computer. The two surface grids are assumed to have the same spacings in both directions. They are also assumed to have the same numbers of columns and rows (if they do not, the smaller grid can be extended and the corresponding height array appropriately extrapolated).

Furthermore, it is assumed that when the two surfaces come into contact, their reference planes become parallel and the two surface grids become aligned with each other so that all of the grid nodes match. Therefore, a single grid can be used to describe the shapes and elastic deflections of both surfaces, as well as the contact pressure acting between them. The nodes of the grid used in the contact analysis are denoted by (i,j), where the indices i and j refer to the grid columns and rows, respectively. The set of all nodes in the grid is denoted by I_g:

$$I_g = \{(i,j):\quad 0 \le i < M_x,\quad 0 \le j < M_y\},$$

where M_x and M_y are the numbers of columns and rows in the grid, respectively.

A Cartesian system $(x,\ y,\ z)$ is introduced in which the plane $z = 0$ coincides with the plane of contact, the z-axis passes through the grid center, and the x and y-axes are parallel to the grid sides. The horizontal coordinates of grid node $(i,\ j)$ in this system are denoted by $(x_i,\ y_j)$. The corresponding composite surface height is defined as the sum of the heights of the two surfaces at node $(i,\ j)$, and is denoted by h_{ij}. To avoid spurious stress concentration along the grid edges, a smooth convex shape can be added to the composite rough surface [31,34]. In any case, it is assumed that the slope of the composite surface is small throughout the grid area, so that

$$|h_{ij} - h_{kl}| \ll \sqrt{(x_i - x_k)^2 + (y_j - y_l)^2},$$

$$\forall (i,j) \in I_g, \forall (k,l) \in I_g, (k,l) \ne (i,j).$$

When the two solids come into contact, normal contact stress (i.e., contact pressure) is generated between them. This contact pressure causes the surface of each solid to deflect inwards. The composite surface deflection $u(x,\ y)$ is defined as the sum of the elastic deflections of the two contacting surfaces at the point $(x,\ y)$, where the deflection of each surface is measured in the direction of the corresponding outer normal. Since only composite surface deflections will be considered below, the word 'composite' will be omitted for brevity. The surface deflection distribution produced by a given contact pressure distribution $p(x,\ y)$ can be expressed as follows:

$$u(x,y) = -\int_{S_g}\int K(x - x', y - y')\, p(x', y')\, \mathrm{d}x'\, \mathrm{d}y', \quad (1)$$

where S_g is the grid area and $K(x,\ y)$ stands for the surface deflection distribution produced by a concentrated normal contact load of unit magnitude acting at the origin. For a pair of homogeneous elastic solids in contact, the kernel $K(x,\ y)$ is given by the Boussinesq formula [35]:

$$K(x,y) = \left(\frac{1-\nu_1^2}{\pi E_1} + \frac{1-\nu_2^2}{\pi E_2}\right)\frac{1}{\sqrt{x^2+y^2}}, \quad (2)$$

where E_1 and E_2 are the Young moduli of the two solids, and υ_1 and υ_2 are their Poisson's ratios.

The grid area S_g is divided into N rectangular surface elements S_{ij} centered at the grid nodes. The length and the width of each element are equal to the grid spacings in the x and y-directions, respectively, so that $N = M_x M_y$. The contact pressure distribution is approximated by a piecewise constant function, uniform within each surface element. Then, Eq. (1) can be re-written as

$$u_{ij} = -\sum_{(k,l)\in I_g} K_{i-k,j-l}\, p_{kl}, \quad (i,j) \in I_g, \quad (3)$$

where u_{ij} is the surface deflection at node $(i,\ j)$, p_{kl} is the uniform pressure acting within the element centered at node $(k,\ l)$, and K_{ij} are the influence coefficients, given by

$$K_{ij} = \int_{S_{00}}\int K(x_i - x', y_j - y')\, \mathrm{d}x'\, \mathrm{d}y', \quad (i,j) \in I_g. \quad (4)$$

In the case of homogenous elastic solids, the coefficients K_{ij} can be easily calculated using the well-known closed-form solution for a patch load acting on an elastic half-space [35]. The process of calculating the surface deflection sums

(3) for all of the grid nodes, or similar multiple sums for other functions, will be referred to as multi-summation.

The use of discrete multi-summation (3) instead of continuum integration (1) will generally lead to an error in the nodal deflections. The corresponding relative error will be referred to as the discretization error and denoted by ε_d. Generally, ε_d is the highest attainable accuracy for discrete problems based on Eq. (3) (e.g., Ref. [32]). For a differentiable $p(x, y)$ and the piecewise-constant pressure approximation used in the present analysis, it can be shown that $\varepsilon_d = O(M_{min}^{-2})$, where $M_{min} = \min(M_x, M_y)$. In most contact problems, $p(x, y)$ is not differentiable at the boundary points of the contact area. Consequently, the maximum attainable accuracy of numerical solution can be lower than $O(M_{min}^{-2})$ for such problems (see below).

Taking into account Eq. (3), the elastic contact problem can be described by the following system of equations and inequalities:

$$\sum_{(k,l)\in I_g} K_{i-k,j-l} p_{kl} = h_{ij} + \alpha, \quad (i,j) \in I_c; \tag{5a}$$

$$p_{ij} > 0, \quad (i,j) \in I_c; \tag{5b}$$

$$\sum_{(k,l)\in I_g} K_{i-k,j-l} p_{kl} \geq h_{ij} + \alpha, \quad (i,j) \notin I_c; \tag{5c}$$

$$p_{ij} = 0, \quad (i,j) \notin I_c; \tag{5d}$$

$$a_x a_y \sum_{(i,j)\in I_g} p_{ij} = P_0. \tag{5e}$$

Here α is the 'rigid-body' approach of the two solids, I_c is the set of all grid nodes that are in contact, a_x and a_y are the grid spacings, and P_0 is the total normal load supported by the contact.

The system of contact equations and inequalities (5) is to be solved for the elemental pressures p_{ij}. The set of contact nodes I_c, which is a discrete analogue of the real area of contact, is not generally known in advance and needs to be determined as part of problem solution. Usually, the normal load P_0 is specified, while the approach α is unknown.

After the elemental contact pressures p_{ij} have been determined for all of the grid nodes, the corresponding subsurface stresses at any given depth $z > 0$ can be computed as follows:

$$\sigma_{mn}(x_i, y_j, z) = \sum_{(k,l)\in I_g} \left[S^{mn}_{i-k,j-l}(z) + \mu T^{mn}_{i-k,j-l}(z) \right] p_{kl},$$

$$m,n = 1,2,3, \ (i,j) \in I_g. \tag{6}$$

Here μ is the coefficient of friction (or traction), and $S^{mn}_{ij}(z)$ and $T^{mn}_{ij}(z)$ are the stress influence coefficients for the normal and tangential contact loading, respectively. These coefficients play a role similar to that of the influence coefficients K_{ij} in Eq. (3): they describe stresses generated at subsurface points lying directly below grid nodes by the elemental contact pressures and the corresponding tangential contact stresses. Closed-form expressions for both $S^{mn}_{ij}(z)$ and $T^{mn}_{ij}(z)$ are available [36] for homogeneous elastic solids.

2.2. Conventional methods of solution

The discrete contact problems (5a)–(5e) is mathematically equivalent to the standard problem of quadratic optimization, and in principle, can be solved using the simplex method [29]. However, this method is only practical for relatively small N. Poon and Sayles [28] and some other workers used the Gauss elimination method to solve the system of Eq. (5a), but as the computation time for Gauss elimination is $O(N^3)$, this approach is also limited to small N.

The three-level iteration approach originally developed by Kubo et al. [30] is more efficient and is often used to solve rough contact problems. The solution begins by making initial guesses of I_c and α. For given I_c and α, the linear system of contact Eq. (5a) is solved using a standard iteration scheme, such as the Gauss–Seidel method [37]. Then, I_c is adjusted to satisfy the contact inequalities (5b), (5c), and the equality constraints (5d) are enforced outside I_c. After that, the contact equations are solved again for the modified I_c. This process continues until I_c no longer changes. Then, the value of α is corrected based on the error in the force balance Eq. (5e), and the previous steps are repeated until (5e) is satisfied to a desired precision. Alternatively, the inequality constraints (5b), (5c) can be enforced simultaneously with the Gauss–Seidel iterations [31], which tends to accelerate the iteration algorithm convergence [38].

The conventional method described above may be satisfactory for solving contact problems for model 3D surfaces defined on grids with relatively small N. However, the computation times become prohibitively long for 3D contact problems involving real rough surfaces. As was discussed above, roughness samples with very large numbers of data points ($N > 10^5$) need to be used in realistic analyses of 3D rough contacts. When N becomes large, two types of difficulties arise. First, the convergence of the iterative solution becomes slow: one has to perform large numbers of iterations to determine the set of contact nodes I_c, and to obtain p_i for each approximation of I_c. Second, the computational cost of even a single iteration becomes prohibitively high. The number of numerical operations, and hence the computation time, required for calculating the surface deflections for all elements by direct summation in Eq. (3) is $O(N^2)$. Consequently, solution of a rough contact problem by the conventional method may take days even on a modern computer and for a moderate N (cf., Ref. [38]). Computation of the subsurface stress

field by straightforward multi-summation in Eq. (6) also takes exceedingly long times.

3. New method of rough contact problem solution

3.1. Fast computation of surface deflections using 2D MLMS

The numerical method presented here computes the surface deflections using an alternative 2D MLMS algorithm in which intergrid operations are applied simultaneously in the x and y-directions. The general concept of MLMS is due to Brandt and Lubrecht [32]. The MLMS technique exploits analytical properties of integral kernels to enable fast numerical evaluation of the corresponding integrals. The following discussion of MLMS is limited to the case of convolution-type integrals with potential-type kernels, such as those arising in contact problems (see Eqs. (1) and (2)). The basic idea of MLMS is to transfer the sums resulting from integral discretization, such as the ones appearing in Eq. (3), to a very coarse grid, on which multi-summation can be performed in just $O(N)$ operations. The resulting coarse-grid sums are then transferred back to the original fine grid. Both the fine-to-coarse and the coarse-to-fine transfers are performed stepwise; with each step, the grid is coarsened (refined) by a factor of two. The corresponding intergrid transfer operators (often called the reduction and the prolongation operators, respectively) are based on a Lagrange polynomial interpolation of the kernel. The kernel value at each fine grid node is interpolated from the $2t$ nearest coarse grid nodes. The number $2t$ is often referred to as the transfer order. For a fixed t, the combined computational cost of all transfers is also $O(N)$.

Since the common potential-type kernels are singular, and polynomial interpolation is invalid near the singularity point (the point $(i, j) = (k, l)$ in Eq. (3)), a special correction procedure has to be applied. The correction term for each node is computed by summing the error of kernel interpolation over the neighboring fine grid nodes. The correction region needs to be made sufficiently large to keep the remaining kernel interpolation error under the discretization error. Then, the use of MLMS will not lead to a significant precision loss, as compared to direct summation. In the case of one-dimensional (1D) MLMS with a logarithmic kernel, Brandt and Lubrecht [32] derived a rigorous formula for the minimum size of the correction region. For the discretization used in the present analysis, this formula reduces to

$$m_c = 0.7tN^{1/t} - 1, \tag{7}$$

where $2m_c + 1$ are the number of nodes in the correction region.

It was suggested in Ref. [32] that the MLMS algorithm can be extended to 2D by alternately applying the 1D transfer operators in the x and y-directions. This method was used in the subsequent applications of MLMS to contact problems (e.g., Ref. [3]). However, such an alternating method has a serious drawback, in our opinion: when using this method, it is not clear how the dimensions of the correction region are to be selected. In their 2D examples, Brandt and Lubrecht [32] used $m_c = 3 + 0.5\ln N$ in the current direction of interpolation and $m_c = 2$ in the perpendicular direction. However, these ad hoc formulae lack mathematical justification. Furthermore, both in Ref. [32] and in the subsequent applications of MLMS to contact problems, such logarithmic formulae for m_c were used with fixed-order transfers. However, it is clear from Eq. (7) that such a combination can only be valid for a limited range of N. Note that in some of the numerical examples of Lubrecht and Ioannides [3], the desired precision was not achieved initially, and the order of transfers had to be increased a posteriori. In most practical applications, however, no analytical solution that can be used to control the numerical error is available. Hence, a rigorous formula for the minimum m_c in the 2D case is essential.

Here, a different approach to MLMS in 2D is introduced, where the intergrid transfers are performed simultaneously for the x and y-directions, and the corresponding 2D transfer operators are based on a bivariate Lagrange polynomial interpolation of the kernel. A brief description of the resulting algorithm is presented below, including explicit formulae for the transfer operators of an arbitrary order and the corresponding correction terms. Such formulae (which were not included in Ref. [32]) may be of use to readers who wish to implement MLMS in computer codes. The derivations will be omitted, however, as they are generally analogous to those in the 1D case, and the reader is referred to Ref. [32] for a justification of the general concept of MLMS.

The first step of the algorithm is to calculate the array of coefficients appearing in the transfer operators:

$$s_i^t = \prod_{\substack{k=l \\ k\neq i}}^{2t} \left(\frac{2(t-k)+1}{2(i-k)} \right), \quad 1 \le i \le 2t. \tag{8}$$

This array is sometimes referred to as the transfer stylus in the multigrid literature; it is used in both 1D and 2D MLMS algorithms. The transfer styli of orders 2, 4, and 6 were listed in Ref. [32].

The specified array of elemental pressures is sequentially transferred to a very coarse grid by repeatedly applying the reduction operator. On each step, the grid spacings are increased by a factor of two: $a_x^c = 2a_x^f$, $a_y^c = 2a_y^f$, where the superscripts c and f refer to the coarse and the fine grids, respectively. The dimensions of the coarse grid are chosen as follows:

$$M_x^c = M_x^f/2 + 2t - 1, \quad M_y^c = M_y^f/2 + 2t - 1, \tag{9}$$

where a slash mark denotes integer division. One sees from Eq. (9) that on each level the coarse grid is extended beyond the fine grid. This is necessary to enable the use of the same transfer operators for all grid nodes, including the ones along the grid edges. On each level, the elemental pressures for the coarse grid are computed as follows:

$$p^{\mathrm{c}}_{mn} = p^{\mathrm{f}}_{ij} + \sum_{k=1}^{2t} s^{\mathrm{t}}_k\, p^{\mathrm{f}}_{i+2(k-t)-1,\,j} + \sum_{l=1}^{2t} s^t_l\, p^{\mathrm{f}}_{i,\,j+2(l-t)-1} + \sum_{k=1}^{2t}\sum_{l=1}^{2t} s^t_k s^t_l\, p^{\mathrm{f}}_{i+2(k-t)-1,\,j+2(l-t)-1},$$

$$(m,n) \in I^{\mathrm{c}}_{\mathrm{g}}, \tag{10}$$

where $i = (m - t + 1)$ and $j = 2(n - t + 1)$. When applying Eq. (10), the contact pressure is assumed to vanish outside the fine grid: $p^{\mathrm{f}}_{ij} = 0$, $\forall (i,\ j) \notin I^{\mathrm{f}}_{\mathrm{g}}$.

The transfer operations (9), (10) are repeated until the number of nodes in the coarse grid becomes sufficiently small: $M^{\mathrm{c}}_x M^{\mathrm{c}}_y \approx \sqrt{N}$, where N is the number of nodes in the finest (i.e., the original) grid. The nodal deflections for the coarsest grid are then computed by direct summation:

$$u^{\mathrm{c}}_{ij} = -\sum_{k=0}^{M^{\mathrm{c}}_x - 1}\ \sum_{l=0}^{M^{\mathrm{c}}_y - 1} K^{\mathrm{c}}_{i-k,\,j-l}\, p^{\mathrm{c}}_{kl},$$

$$0 \le i < M^{\mathrm{c}}_x, \quad 0 \le j < M^{\mathrm{c}}_y. \tag{11}$$

The computational cost of this operation is obviously $O(N)$. The influence coefficients for the coarse and the fine grids on the current level are defined as follows:

$$K^{\mathrm{c}}_{ij} = K_{2di,2dj}; \quad K^{\mathrm{f}}_{ij} = K_{di,dj}. \tag{12}$$

Here K_{ij} are the influence coefficients for the original grid, and $d = 2^q$, where q is the current level and $q = 0$ corresponds to the finest grid.

Next, the computed coarse-grid sums are sequentially transferred back to the original grid. Each transfer involves an interpolation operation and two steps of correction. First, correction coefficients for the coarse-grid sums are pre-computed for all pairs of indices such that $-m_{\mathrm{c}} \le i \le m_{\mathrm{c}}$ and $-m_{\mathrm{c}} \le j \le m_{\mathrm{c}}$:

$$C^{(1)}_{ij} = 0, \quad \text{even } i,j; \tag{13a}$$

$$C^{(1)}_{ij} = K^{\mathrm{f}}_{ij} - \sum_{k=1}^{2t} s^t_k K^{\mathrm{f}}_{i-2(k-t)+1,\,j}, \quad \text{odd } i, \text{even } j; \tag{13b}$$

$$C^{(1)}_{ij} = K^{\mathrm{f}}_{ij} - \sum_{l=1}^{2t} s^t_l K^{\mathrm{f}}_{i,\,j-2(l-t)+1}, \quad \text{even } i, \text{odd } j; \tag{13c}$$

$$C^{(1)}_{ij} = K^{\mathrm{f}}_{ij} - \sum_{k=1}^{2t}\sum_{l=1}^{2t} s^t_k s^t_l K^{\mathrm{f}}_{i-2(k-t)+1,\,j-2(l-t)+1},$$

$$\text{odd } i,\ j. \tag{13d}$$

The coarse-grid deflections are then corrected by using these coefficients:

$$u^{\mathrm{c}}_{ij} \leftarrow u^{\mathrm{c}}_{ij} + \sum_{k=-m_{\mathrm{c}}}^{m_{\mathrm{c}}}\ \sum_{l=-m_{\mathrm{c}}}^{m_{\mathrm{c}}} C^{(1)}_{kl}\, p^{\mathrm{f}}_{2(i-t+1)-k,\,2(j-t+1)-l},$$

$$(i,j) \in I^{\mathrm{c}}_{\mathrm{g}}, \tag{14}$$

where a left arrow denotes re-calculation of a variable using its old value. This notation is adopted to avoid the use of cumbersome superscripts denoting the old and new values. After the correction step (14), the fine-grid deflections are interpolated from the coarse-grid deflections:

$$u^{\mathrm{f}}_{ij} = u^{\mathrm{c}}_{mn}, \quad \text{even } i,j; \tag{15a}$$

$$u^{\mathrm{f}}_{ij} = \sum_{k=1}^{2t} s^t_k u^{\mathrm{c}}_{m+k-t-1,\,n}, \quad \text{odd } i, \text{even } j; \tag{15b}$$

$$u^{\mathrm{f}}_{ij} = \sum_{l=1}^{2t} s^t_l u^{\mathrm{c}}_{m,\,n+l-t-1}, \quad \text{even } i, \text{odd } j; \tag{15c}$$

$$u^{\mathrm{f}}_{ij} = \sum_{k=1}^{2t}\sum_{l=1}^{2t} s^t_k s^t_l u^{\mathrm{c}}_{m+k-t-1,\,n+l-t-1}, \quad \text{odd } i,j. \tag{15d}$$

Here, $m = (i + 1)/2 + t - 1$, $n = (j + 1)/2 + t$-1, where a slash mark denotes integer division. Then, correction coefficients for the fine-grid sums are pre-computed for all pairs of indices such that $-m_{\mathrm{c}} \le i \le m_{\mathrm{c}}$ and $-m_{\mathrm{c}} \le j \le m_{\mathrm{c}}$:

$$C^{(2)}_{ij} = K^{\mathrm{f}}_{ij} - \sum_{k=1}^{2t} s^t_k K^{\mathrm{f}}_{i+2(k-t)-1,\,j}; \tag{16a}$$

$$C^{(3)}_{ij} = K^{\mathrm{f}}_{ij} - \sum_{l=1}^{2t} s^t_l K^{\mathrm{f}}_{i,\,j+2(l-t)-1}; \tag{16b}$$

$$C^{(4)}_{ij} = K^{\mathrm{f}}_{ij} - \sum_{k=1}^{2t}\sum_{l=1}^{2t} s^t_k s^t_l K^{\mathrm{f}}_{i+2(k-t)-1,\,j+2(l-t)-1}. \tag{16c}$$

These coefficients are used to correct the fine-grid deflections:

$$u^{\mathrm{f}}_{ij} \leftarrow u^{\mathrm{f}}_{ij} + \sum_{k=-m_{\mathrm{c}}}^{m_{\mathrm{c}}}\ \sum_{l=-m_{\mathrm{c}}}^{m_{\mathrm{c}}} C_{kl}\, p^{\mathrm{f}}_{i-k,\,j-l}, \quad (i,j) \in I^{\mathrm{f}}_{\mathrm{g}}. \tag{17}$$

Here $C_{kl} \equiv 0$ if both i and j are even; $C_{kl} = C^{(2)}_{kl}$ if i is odd and j is even; if $C_{kl} = C^{(3)}_{kl}$ if i is even and j is odd; and $C_{kl} = C^{(4)}_{kl}$ if both i and j are odd.

The correction coefficients (13), (16) are based on the same bivariate Lagrange polynomial interpolation of the kernel as the 2D transfer operations (10), (15). The corresponding correction terms in Eq. (14), Eq. (17) are computed over squares of $(2m_{\mathrm{c}} + 1) \times (2m_{\mathrm{c}} + 1)$ nodes. For this fully 2D MLMS algorithm, a rigorous formula for m_{c} is derived in the same way as the corresponding 1D

formula. In the case of an inverse-distance kernel, such as the Boussinesq kernel $K(x, y)$, this 2D formula is very similar to Eq. (7):

$$m_c = 0.7tM_{min}^{1/t} - 1. \quad (18)$$

For $m_c < 2t$, this formula becomes invalid, in which case m_c is set to $2t$. The use of the rigorous formula (18) in the present MLMS algorithm guarantees that the summation results will be accurate to M_{min}^{-2} for any number of nodes in the grid. As was mentioned in Section 2.1, this accuracy may exceed the highest attainable numerical accuracy for most contact problems. Nevertheless, formula (18) is used in the present contact solver to ensure that the highest possible accuracy is achieved for any type of problem. In any case, an important advantage of formula (18) is that the summation error does not grow indefinitely with increasing N.

In the present MLMS algorithm, t is set to the value minimizing the computational work W for a given N. It is seen from Eqs. (8) and (17) that for large N,

$$W \approx \beta_1 Nt^2 + \beta_2 Nm_c^2, \quad (19)$$

where $\beta_{1,2}$ are implementation-dependent constants. Substituting m_c from Eq. (18) and minimizing the result with respect to t, one obtains

$$t \approx \beta \ln M_{min}, \quad (20)$$

where t needs to be rounded to the nearest integer number, and β is an implementation-dependent constant that can be determined by numerical experimentation. For the computer program developed by the authors, near-optimum performance was achieved with $\beta \approx 0.84$. Substituting Eq. (20) into Eq. (18) yields $m_c \sim t$, so that Eq. (19) reduces to $W = O(Nt^2)$. Substituting t from Eq. (20), one obtains $W = O(N(\ln N)^2)$, which represents a major improvement over direct summation.

To illustrate the above point, the nodal deflections corresponding to the pressure distribution $p_{ij} = (x_i^2 + y_j^2)^{1/2}$ have been computed first by direct summation (Eq. (3)), and then by using the proposed 2D MLMS algorithm. Such numerical tests have been performed on a personal computer for several grid sizes. The corresponding computation times T and relative errors ε (calculated using the exact solution) are listed in Table 1. It is seen from Table 1 that the present MLMS algorithm is virtually as precise as direct summation. The computation times for MLMS, although somewhat scattered, appear consistent with the estimate $T = O(N(\ln N)^2)$. For $N = 512^2$, our MLMS algorithm is more than 400 times as fast as direct summation.

It is seen from Eq. (20) that t increases monotonically, although very slowly, with increasing N. One might question the numerical stability of the present MLMS algorithm for large values of t. It is indeed well known that the Newton–Cotes quadrature formulae, which are also based on polynomial interpolation with uniformly spaced nodes, become unstable as their order increases. To investigate this issue, a series of numerical tests was run in which t was not calculated from Eq. (20), but was 'manually' set to very large numbers. No sign of numerical instability was detected in any of these tests, even for t as large as 50. The relative error of MLMS never exceeded M_{min}^{-2} (and was often much lower). Thus, the present MLMS algorithm appears to be stable for all practical purposes.

Table 1
Computation times and relative errors for direct summation and MLMS

N	Direct		MLMS	
	T (s)	ε	T (s)	ε
64^2	2.0	5.4×10^{-4}	0.4	5.1×10^{-4}
128^2	34.8	1.1×10^{-4}	1.9	1.3×10^{-4}
256^2	791.2	3.2×10^{-5}	10.2	3.4×10^{-5}
512^2	1.67×10^4	8.9×10^{-6}	38.4	8.5×10^{-6}
1024^2	–	–	237.4	1.8×10^{-6}

3.2. Contact problem solution using CG-based iteration scheme

The iteration scheme used in the present analysis to solve the discrete contact problem (5) is based on the CG method [39]. Although it was originally developed for unconstrained optimization, this method can be extended to quadratic optimization problems with linear inequality constraints, such as the contact problem (5) [39–41]. Importantly, there exists a rigorous mathematical proof of the method convergence for such problems [39,40].

When solving contact problems by the extended CG method, the contact area is determined in the course of iteration with respect to the contact pressure. Hence, there is no need for an outer level of iteration with respect to the contact area. The contact inequalities (5b), (5c) are enforced on each iteration step, so that the current approximation to the solution always remains inside the domain of feasibility.

A distinctive feature of the present iteration scheme is that the force balance Eq. (5e) is also enforced in the course of contact pressure iteration. Hence, there is no need for iteration with respect to the normal approach, which is present in most contact solvers as the outermost level of iteration, and the contact problem is solved using just one level of iteration.

Another advantage of the CG method is that it is a simultaneous (or explicit) iteration method [37]. This feature makes the CG method compatible with the MLMS technique, which is essential for the present analysis. It can be seen from Eqs. (10) and (15d) that the numbers u_{ij} are calculated simultaneously in the MLMS algorithm, and that all of the numbers p_{ij} are needed for this computation.

Therefore, the successive iteration methods such as the Gauss–Seidel method are generally incompatible with MLMS.

Furthermore, the CG method has a comparatively high (superlinear) rate of convergence. Hence, the number of iterations required for solving the problem (5) to the desired precision (as determined by the discretization error) usually remains moderate even for huge N. This rapid convergence will be demonstrated by the numerical examples presented in the next section. Although FMG and similar multigrid methods can solve systems of linear equations faster than the CG method [42], direct application of such methods to problems with inequality constraints lacks a firm mathematical basis, as discussed in Section 1.

The iteration scheme used in the present work is generally similar to one of the algorithms described in Ref. [40]. For those readers who wish to implement the new method in a computer code, a detailed description of this scheme is provided next.

Before the start of iteration, the process is initialized, as follows. Initial values are chosen for all variables p_{ij}. The initial p_{ij} must be non-negative and satisfy Eq. (5e), but otherwise can be arbitrary. Auxiliary variables δ and G_{old} are defined and initialized by setting $\delta = 0$ and $G_{\text{old}} = 1$. The accuracy goal ε_0 is set to the desired value, and the iteration process begins.

On each step of iteration, the following operations are performed. First, the nodal deflections u_{ij} produced by the current p_{ij} are computed using the MLMS algorithm (8)–(17) with t and m_c given by Eq. (20) and Eq. (18), respectively. The resulting u_{ij} will satisfy Eq. (3) to a sufficient accuracy, as discussed above. Then, the gap distribution g_{ij} is computed and its mean value is adjusted, as follows:

$$g_{ij} = -u_{ij} - h_{ij}, \quad (i,j) \in I_g; \tag{21a}$$

$$\bar{g} = N_c^{-1} \sum_{(k,l)\in I_c} g_{kl}; \tag{21b}$$

$$g_{ij} \leftarrow g_{ij} - \bar{g}, \quad (i,j) \in I_g. \tag{21c}$$

Here N_c denotes the number of nodes in I_c. In this context, I_c is understood as the current contact area, i.e. the set of all (i, j) for which $p_{ij} > 0$. For the new g_{ij}, the sum

$$G = \sum_{(i,j)\in I_c} g_{ij}^2 \tag{22}$$

is calculated, which is used to compute the new conjugate direction t_{ij}:

$$t_{ij} \leftarrow g_{ij} + \delta(G/G_{\text{old}})t_{ij}, \quad (i,j) \in I_c; \tag{23a}$$

$$t_{ij} = 0, \quad (i,j) \notin I_c. \tag{23b}$$

The new t_{ij} is the direction in which the next step will be made in the multidimensional space of elemental pressures. Note that if $\delta = 0$, t_{ij} will coincide with the steepest descent direction. The current value of G is then stored for the next iteration: $G_{\text{old}} = G$. A convolution of K_{ij} with t_{ij} is then computed:

$$r_{ij} = \sum_{(k,l)\in I_g} K_{i-k,j-l} t_{kl}, \quad (i,j) \in I_g. \tag{24}$$

Since Eq. (24) has the same structure as Eq. (3), the distributions r_{ij} and u_{ij} are computed using the same MLMS algorithm. The mean value of r_{ij} is then adjusted as follows:

$$\bar{r} = N_c^{-1} \sum_{(k,l)\in I_c} r_{kl}; \tag{25a}$$

$$r_{ij} \leftarrow r_{ij} - \bar{r}, \quad (i,j) \in I_g. \tag{25b}$$

The distribution r_{ij} is used to calculate the length of the step that will be made in the direction t_{ij}:

$$\tau = \frac{\sum_{(i,j)\in I_c} g_{ij} t_{ij}}{\sum_{(i,j)\in I_c} r_{ij} t_{ij}}. \tag{26}$$

At this point, the current p_{ij} are stored for the subsequent error estimation:

$$p_{ij}^{\text{old}} = p_{ij}, \quad (i,j) \in I_g; \tag{27}$$

after which the solution is updated by making a step of the length τ in the direction t_{ij}:

$$p_{ij} \leftarrow p_{ij} - \tau t_{ij}, \quad (i,j) \in I_c. \tag{28}$$

Next, the inequalities (5d) are enforced by setting all negative p_{ij} to zeros. The set of all non-contact nodes at which the two surfaces overlap is then determined:

$$I_{\text{ol}} = \{(i,j) \in I_g : \quad p_{ij} = 0,\ g_{ij} < 0\}. \tag{29}$$

If $I_{\text{ol}} = \varnothing$, then δ is set to unity. Otherwise, δ is set to zero, and the pressures at the overlapping nodes are corrected:

$$p_{ij} \leftarrow p_{ij} - \tau g_{ij}, \quad (i,j) \in I_{\text{ol}}. \tag{30}$$

It can be shown that $\tau > 0$ at all times. Hence, all nodes in I_{ol} will enter I_c after the correction (30), thus enforcing the inequality (5c). Next, the current contact load is calculated and the force balance condition (5e) is enforced:

$$P = a_x a_y \sum_{(i,j)\in I_g} p_{ij}; \tag{31a}$$

$$p_{ij} \leftarrow (P/P_0)\, p_{ij}, \quad (i,j) \in I_g. \tag{31b}$$

Then, the current relative error is estimated:

$$\varepsilon = a_x a_y P_0^{-1} \sum_{(i,j)\in I_g} |p_{ij} - p_{ij}^{\text{old}}|. \tag{32}$$

If $\varepsilon \geq \varepsilon_0$, another iteration is performed. Otherwise, the iteration process is stopped.

The computational cost of the multi-summations (3) and (24) is $O(N(1n\ N)^2)$, since they are performed using

MLMS (see Section 3.1). The cost of the operations (21)–(23) and (25)–(32) is $O(N)$. Therefore, the net cost of a single iteration is $O(N(\ln N)^2)$.

The rigid-body approach α is not determined in the present scheme. The load-balance equation is enforced and the contact pressure updated without using α (see Eqs. (31a) and (31b)). In most practical applications, the value of α is not needed as long as p_{ij} can be computed.

The choice of ε_0 is based on the following considerations. On one hand, the accuracy of the present MLMS algorithm is $M_{\min}^{-2}$ (see Section 3.1). Hence, one should use $\varepsilon_0 \geq M_{\min}^{-2}$, as the iteration scheme may not converge for smaller values of ε_0. On the other hand, even if a converged solution has been obtained, the actual error in some of the solution parameters may significantly exceed ε_0. Indeed, in a discretized contact problem, the contact size can only change in steps equal to the grid spacings, and there are no more than $M_{\min}$ grid nodes per contact length in the x-direction and/or the y-direction. Hence, the relative error in the contact size will generally be in excess of $M_{\min}^{-1}$ even for well-conditioned smooth contact problems. An estimate shows that the corresponding error in the contact pressure is $O(M_{\min}^{-3/2})$. Thus, $\varepsilon_0 = M_{\min}^{-3/2}$ appears to be a good choice. The situation is more complicated for rough contact problems. The error in short-wavelength components of the calculated contact stress distributions, which correspond to small-scale roughness components, can be very large. However, a much higher accuracy can often be achieved in the long-wavelength (i.e., smooth) solution components. To ensure that the maximum precision possible for the given pair of surfaces is always achieved, $\varepsilon_0 = M_{\min}^{-3/2}$ can be used for rough contact problems as well.

In numerical analyses of rough contacts, an upper limit on the contact pressure is commonly imposed to account for the possibility of plastic yielding at the most stressed asperity microcontacts (e.g., Refs. [27,28]). The limiting pressure p_u is usually set to either the hardness of the softer material or the corresponding yield stress. This is a rather crude elastic–plastic model, but for surfaces with sharp asperities, it tends to produce more realistic results than a purely elastic contact analysis. The iteration scheme described above can be modified rather easily to include an upper limit on the contact pressure. The additional inequality $p_{ij} \leq p_u$ is enforced on each iteration step, in the same way as the condition $p_{ij} \geq 0$. The condition (5e) is enforced by modifying p_{ij} for the nodes that are in elastic contact. The implementation-specific details will be omitted for brevity.

3.3. Fast computation of subsurface stresses

The sums appearing in their right-hand sides of Eqs. (3) and (6) have similar structures. Therefore, the distributions of the subsurface stresses are computed in the same way as the surface deflections, using the MLMS algorithm described in Section 3.1. In the computation of the stress component σ_{mn} at the depth z, the influence coefficients K_{ij} are replaced by the quantities $S_{ij}^{mn}(z) + \mu T_{ij}^{mn}(z)$.

The correction terms are still necessary when computing subsurface stresses, since the stress influence functions vary abruptly near the origin. In principle, smaller values of m_c can be used for subsurface stresses than for surface deflections; for very large z, one can even set $m_c = 0$. However, our experience has showed that the resulting reduction in the computation time is usually moderate, while the determination of the proper value of m_c for each depth is not an easy task. As a simple and safe alternative, the value of m_c given by Eq. (18) can be used for all depths.

The number of depths at which the stresses are computed depends on the nature of the contacting surfaces and the purpose of the contact stress analysis. Generally, the depth interval is made small (on the order of the grid spacings) near the surface, where the stresses vary rapidly due to the influence of surface roughness. On the other hand, a much longer depth interval is used for deep subsurface layers, where the stress field is much smoother. In the rare cases when full detail is desired even for deep layers, stress distributions at additional depths can be obtained by polynomial interpolation.

4. Numerical examples

4.1. Smooth contact

The numerical method described in the Section 3 was implemented in a computer program written in C++

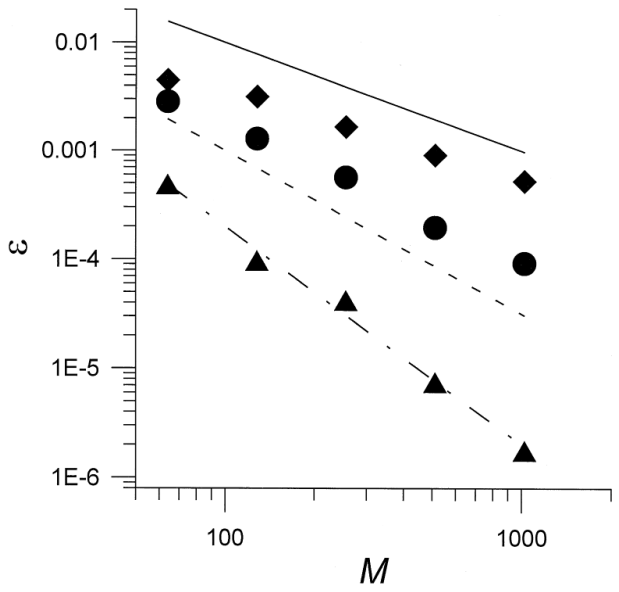

Fig. 1. The relative numerical error ε vs. the number of grid rows M for a Hertz contact problem. The errors in the contact radius (diamonds), contact pressure (circles) and subsurface Mises stress (triangles) are shown. The curves $\varepsilon = M^{-1}$ (solid line), $\varepsilon = M^{-3/2}$ (dashed line), and $\varepsilon = 2M^{-2}$ (chain line) are also shown.

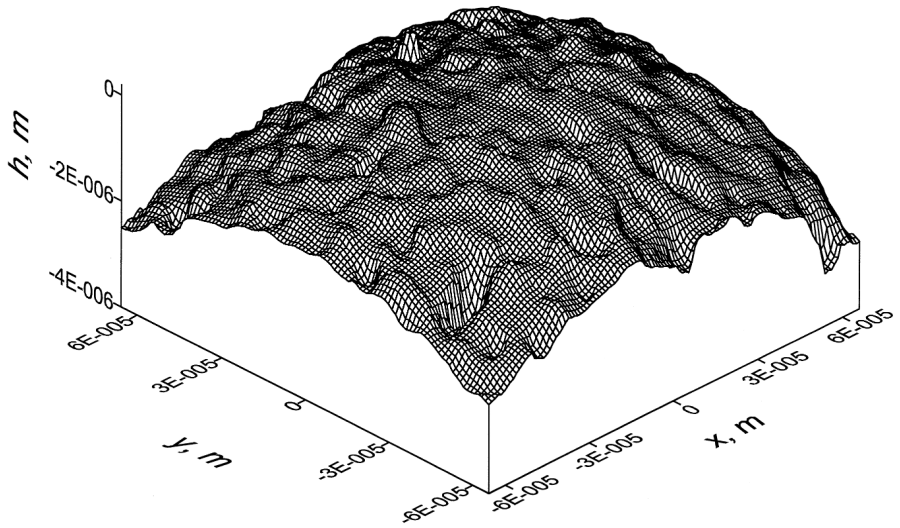

Fig. 2. The composite shape of the contacting surfaces for a rough contact problem. Grid dimensions: 121 × 103 nodes.

programming language. To test the method accuracy, a Hertz contact problem for a pair of smooth elastic spheres was solved numerically. Since this problem has an exact solution (e.g., Ref. [35]), it is suitable for estimating the numerical error of contact solvers.

Square grids with $a_x = a_y$ and $M_x = M_y = M$ were used in this numerical study. The contact problem was solved for $M = 64$, 128, 256, 512, and 1024. All five grids had the same length, slightly exceeding the contact diameter. For each grid, the accuracy goal ε_0 was set to $M^{-3/2}$. The computations were performed on a personal computer with a 200 MHz processor. For all values of M considered, the iteration scheme rapidly converged. Even for $M = 1024$, the accuracy goal was attained in just 34 iteration steps, despite the huge number of nodes ($N \approx 10^6$).

The distributions of the six subsurface stress components and the Mises stress were then computed using the MLMS algorithm described in Section 3.3. It was assumed that in addition to the calculated contact pressure distribution, there is a proportional tangential contact stress distribution. The friction coefficient μ was set to 0.15, which is a typical value for rolling/sliding contacts working under the boundary lubrication conditions.

The following three error parameters were calculated for the obtained numerical solution:

$$\varepsilon_r = \frac{|\sqrt{A/\pi} - r_H|}{r_H};$$

$$\varepsilon_p = \frac{\sum_{(i,j)\in I_c} |p_{ij} - p_H(x_i, y_j)|}{\sum_{(i,j)\in I_c} p_H(x_i, y_j)};$$

$$\varepsilon_s = \frac{\sum_{(i,j)\in I_g} |s(x_i, y_j, z_m) - s_H(x_i, y_j, z_m)|}{\sum_{(i,j)\in I_g} s_H(x_i, y_j, z_m)}.$$

Here $A = N_c a_x a_y$ is the numerically computed area of contact, r is the contact radius, s is the Mises stress, z_m is the depth where the maximum Mises stress occurs, and the subscript H denotes the exact Hertz solution. The distribution $s_H(x_i, y_j, z_m)$ was calculated using the subsurface stress solution of Sackfield and Hills [43]. The parameters ε_r, ε_p, and ε_s characterize the relative error in the contact radius, the contact pressure, and the Mises stress, respectively. These three error parameters are plotted versus M in Fig. 1. As expected, ε_r turns out to be the most sensitive of them. Nevertheless, it is seen from Fig. 1 that $\varepsilon_r < M^{-1}$ for all M considered. The plot also shows that ε_p is on the order of $M^{-3/2}$ in the range considered, which justifies our choice of ε_0. However, ε_p appears to grow slightly faster than $M^{-3/2}$. This slight deviation from the anticipated trend has not been explained and is left to future studies. On the other hand, it is seen from Fig. 1 that $\varepsilon_s \approx 2M^{-2}$, and there appears to be no indica-

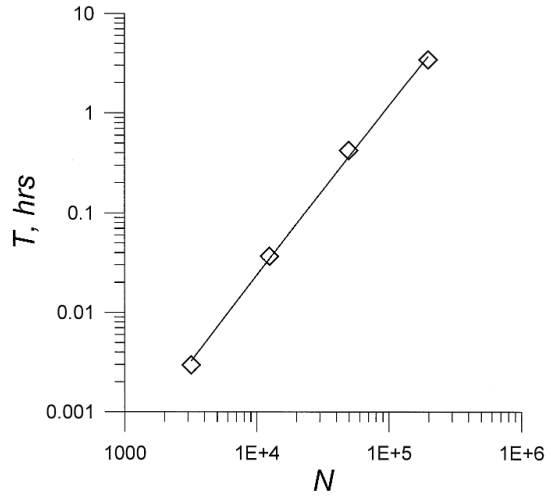

Fig. 3. The solution time T vs. the number of grid nodes N for a set of rough contact problems (diamonds). The analytical curve $T = 2.8 \times 10^{-10} N^{3/2} (\ln N)^2$ is also shown (solid line).

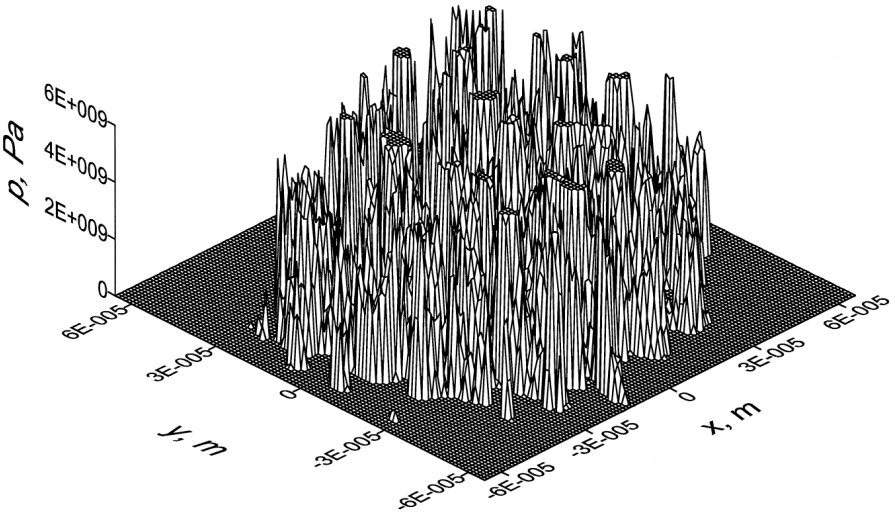

Fig. 4. The contact pressure distribution for a rough contact problem. Grid dimensions: 121 × 103 nodes.

tion of a faster growth for larger M. Thus, the subsurface stresses can be computed to a much higher relative accuracy than the main surface quantities. This is not surprising, as the kernels in the integral relations for subsurface stresses are regular, unlike the Boussinesq kernel $K(x, y)$ appearing in Eq. (1), and integration with a regular kernel tends to suppress the error in the contact pressure.

The above results demonstrate that the new numerical method can solve contact problems to the highest order of accuracy attainable with the present discretization. Note that such ability was not demonstrated for the FMG-based contact solver of Lubrecht and Ioannides [3].

4.2. Contact of real rough surfaces

Two sets of surface roughness data were collected from a pair of tribological specimens using an optical profilometer. Both data sets contained 503 × 469 points, and had the horizontal spacings $a_x = 1.1$ μm and $a_y = 1.29$ μm. Normal contact problems were solved for the following four grid sizes: $M_x = 483$, $M_y = 412$; $M_x = 242$, $M_y = 206$; $M_x = 121$, $M_y = 103$; and $M_x = 61$, $M_y = 52$. The roughness data for each grid were obtained by clipping the edges of the original data sets. Each roughness sample was superimposed onto a smooth sphere. The sphere diameter was 30 mm for the 483 × 412 grid, and was scaled to the grid length. The composite surface shape for the 121 × 103 grid is shown in Fig. 2. The material parameters were chosen as follows: $E_1 = E_2 = 210$ GPa, $\upsilon_1 = \upsilon_2 = 0.28$, $p_u = 6$ GPa. The normal load P_0 was set to 240 N for the 483 × 412 grid, and was scaled to the grid area. As in the previous example, $\varepsilon_0 = M_{\min}^{-3/2}$ was used.

The iteration process converged steadily for all of the grids considered. Although the convergence was noticeably slower than in the smooth contact case, it was sufficiently fast to solve the rough contact problem (5) in reasonable times (less than 3.5 h for the 483 × 412 grid). The solution time T (i.e., the computation time required for attaining the accuracy goal with respect to p_{ij}) is plotted as a function of N in Fig. 3. The time data appear to fit the curve $T = 2.8 \times 10^{-10}\ N^{3/2}\ (\ln N)^2$, which is also shown in the plot, rather closely. Since the cost of a single iteration is $O(N(\ln N)^2)$ (see Section 3.2), this fit indicates that the number of iterations required for attaining the accuracy goal is $O(N^{1/2})$ for the rough surfaces and the loading level considered.

The computed contact pressure distribution for the 121 × 103 grid is shown in Fig. 4. As one might expect, the small-scale features of this distribution are dominated by surface roughness and are highly irregular. The pressure spikes are quite high: at a few locations, the pressure

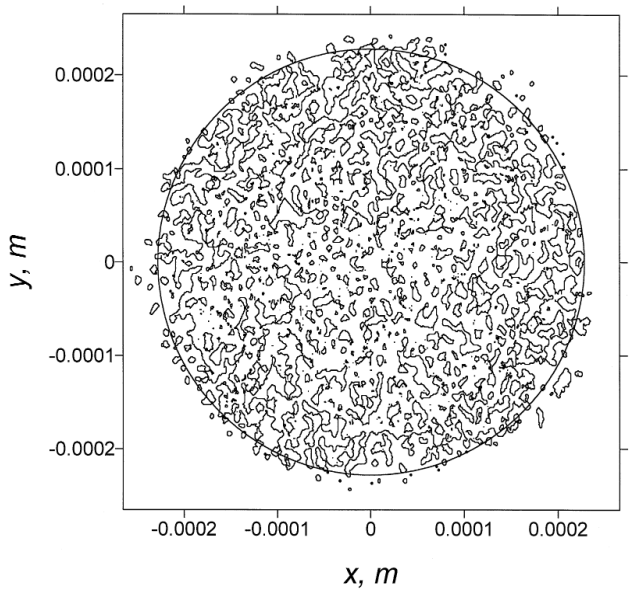

Fig. 5. The boundary of the real area of contact for a rough contact problem. Grid dimensions: 483×412 nodes. The circular boundary of the corresponding Hertz contact is also shown.

reaches the cutoff value (6 GPa). For comparison, the maximum pressure for the corresponding Hertz contact is only 2.2 GPa. The rough contact boundary for the 483×412 grid is shown in Fig. 5, together with the boundary of the corresponding Hertz contact. Parts of the rough contact are seen to lie beyond the Hertz contact boundary, which agrees with the predictions of Greenwood and Tripp [44]. However, the difference in contact dimensions is relatively small for this particular problem. On the other hand, the real area of contact is much smaller than the nominal (Hertzian) area of contact: $A/A_H \approx 0.55$. It is also seen from Fig. 5 that the real area of contact becomes more disconnected near the nominal contact boundary.

The subsurface stress distributions were also calculated for the grids considered, under an assumption of frictionless contact. The stress computation times were somewhat longer than the times required for solving the contact problem (5), primarily due to the number of stress components, but were still reasonable. For example, for the 483×412 grid, the stress distributions at 45 depths were computed in 4.3 h (5.7 min per depth). The distributions of the Mises stress in the vertical planes $x = 0$ and $y = 0$ for the 121×103 grid are shown in Fig. 6. It is seen that at relatively shallow depths, the subsurface stress field is dominated by the roughness-induced stress perturbations, while in deeper subsurface layers, the smooth stress field

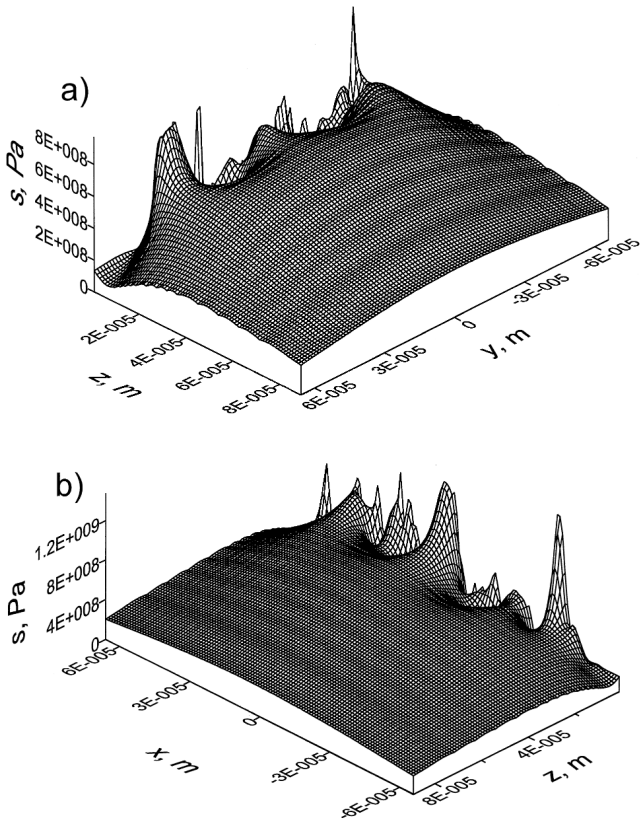

Fig. 6. The distributions of the Mises stress in the vertical planes $x = 0$ (a) and $y = 0$ (b) for a rough contact problem. Grid dimensions: 121×103 nodes.

corresponding to the macrocontact takes over. The maximum stress occurs at the surface and is significantly higher than the macrocontact-generated stress.

Finally, it can be mentioned that the authors have applied the present rough contact solver to rough surfaces arising from various tribological applications. In all cases studied, the method converged without difficulty, and the computation times were reasonably short.

5. Other fast methods for solving rough contact problems

Recently, Nogi and Kato [45] developed a numerical method for solving 3D rough contact problems based on the fast Fourier transform (FFT) technique. The basic idea of their approach is to transfer Eq. (3) from the space domain to the wave number domain. In the latter domain, the discrete convolution in the right-hand side of Eq. (3) reduces to a simple product of the Fourier transforms of K_{ij} and p_{ij}. The transform of p_{ij} is first computed using FFT, followed by the transform multiplication. Then, u_{ij} is restored from its transform using the inverse FFT. The combined cost of these operations is $O(N \ln N)$. The system of contact equations and inequalities (5) is solved by iteration, in the same way as in the MLMS-based approach.

However, the FFT technique is strictly applicable to periodic discrete functions only. On one hand, this makes the FFT-based approach ideal for solving contact problems for nominally flat rough surfaces that can be modeled as periodic. On the other hand, concentrated contact problems for rough surfaces that are significantly curved on the scale of interest (such as the problem considered in Section 4.2) are essentially non-periodic. When FFT is applied to such a problem, an error is introduced in the solution. This error can be reduced by extending the surface grid, but the increase in N tends to offset the computational efficiency of FFT rather severely. Consequently, the MLMS-based approach turns out to be more advantageous for non-periodic rough contact problems. This assessment is based on the results of a comparative numerical study by these authors, which will be published separately.

A quite different approach to solving rough contact problems was presented very recently by Chekina and Keer [46]. It is based on an integral formula inverse to the relation (1), i.e., expressing $p(x, y)$ as an integral involving $u(x, y)$. The surface deflections are determined by a specially designed iteration scheme. On each step, the integral is computed numerically using a set of non-uniform grids. Such integration is fast because the kernel appearing in the inverse relation decays very rapidly with the distance. The inverse method is more advantageous than the present one for nearly complete rough contacts, while the opposite is true in situations where the real area of contact is small. However, a rigorous error estimate was

not presented for the fast integration procedure of Chekina and Keer [46], unlike the present MLMS algorithm.

6. Conclusion

The new 3D contact solver, based on a combination of the MLMS and CG techniques, is sufficiently fast to solve contact problems with large numbers of surface points ($N \sim 10^5$–10^6) in reasonable times. Hence, the method can be applied to roughness samples of practical sizes, such as the ones collected by modern 3D profilometric devices. The use of a fully 2D MLMS algorithm allows a rigorous estimate of the summation error to be obtained. Consequently, the MLMS parameters are automatically set to proper values for any grid size, which ensures that discrete contact problems are always solved to the highest possible accuracy. Since the CG method is strictly applicable to problems with inequality constraints, the contact solver converges without difficulty for arbitrary rough surfaces, even when the contact area is highly disconnected. The new method appears to be the most advantageous for solving concentrated contact problems for which the macrocontact dimensions are comparable to those of the roughness sample.

Acknowledgements

We thank Dr. O.G. Chekina of the Institute for Problems of Mechanics, Moscow, Professor H.S. Cheng of Northwestern University, and Dr. S.C. Lee of Ohio State University for valuable discussions. We are indebted to Dr. J.D. Cogdell of Timken for providing us with roughness data. Thanks to Dr. S.J. Harris of General Motors for using early versions of our computer program in his work, which helped us to improve the code. The financial support of Caterpillar, General Motors, and Timken under the Advanced Technology Program of NIST is gratefully acknowledged.

References

[1] K.L. Johnson, J.L. Tevaarwerk, Shear behaviour of elastohydrodynamic oil films, Proc. R. Soc. London A 356 (1977) 215–236.
[2] J.A. Greenwood, K.L. Johnson, The behaviour of transverse roughness in sliding elastohydrodynamically lubricated contacts, Wear 153 (1992) 107–117.
[3] A.A. Lubrecht, E. Ioannides, A fast solution of the dry contact problem and the associated subsurface stress field, using multilevel techniques, ASME J. Tribol. 113 (1991) 128–133.
[4] J.A. Greenwood, The area of contact between rough surfaces and flats, ASME J. Lubrication Technol. 89 (1967) 81–91.
[5] J.A. Greenwood, J.B.P. Williamson, Contact of nominally flat surfaces, Proc. R. Soc. London A 295 (1966) 300–319.
[6] D.J. Whitehouse, J.F. Archard, The properties of random surfaces of significance in their contact, Proc. Roy. Soc. Lond. A 316 (1970) 97–121.
[7] R. Nayak, Random process model of rough surfaces in plastic contact, Wear 26 (1973) 305–333.
[8] J.I. McCool, Comparison of models for the contact of rough surfaces, Wear 107 (1986) 37–60.
[9] A. Majumdar, B. Bhushan, Fractal model of elastic-plastic contact between rough surfaces, ASME J. Tribol. 113 (1991) 1–11.
[10] B. Bhushan, A. Majumdar, Elastic–plastic contact model for bifractal surfaces, Wear 153 (1992) 53–64.
[11] D. Berthe, Ph. Vergne, An elastic approach to rough contact with asperity interactions, Wear 117 (1987) 211–222.
[12] I.G. Goryacheva, M.N. Dobychin, Multiple contact model in the problems of tribomechanics, Tribol. Int. 24 (1991) 29–35.
[13] H. Aramaki, H.S. Cheng, Y.-W. Chung, The contact between rough surfaces with longitudinal texture: Part I. Average contact pressure and real contact area, ASME J. Tribol. 115 (1993) 419–424.
[14] Y. Leng, G. Yang, Y. Huang, L. Zheng, The characteristics of elastically contacting ideal rough surfaces, ASME J. Tribol. 118 (1996) 90–97.
[15] J.B.P. Williamson, R.T. Hunt, Asperity persistence and the real area of contact between rough surfaces, Proc. R. Soc. London A 327 (1972) 147–157.
[16] J. Pullen, J.B.P. Williamson, On the plastic contact of rough surfaces, Proc. R. Soc. London A 327 (1972) 159–173.
[17] T. Wanheim, Friction at high normal pressures, Wear 25 (1973) 225–244.
[18] R.S. Sayles, T.R. Thomas, Surface topography as a nonstationary random process, Nature 271 (1978) 431–434.
[19] J.A. Greenwood, Problems with surface roughness, in: I.L. Singer, H.M. Pollock (Eds.), Fundamentals of Friction: Macroscopic and Microscopic Processes, Kluwer, Dordrecht, 1992, pp. 57–76.
[20] I.A. Polonsky, T.P. Chang, L.M. Keer, W.D. Sproul, An analysis of the effect of hard coatings on near-surface rolling contact fatigue initiation induced by surface roughness, Wear 208 (1997) 204–219.
[21] M.P.F. Sutcliffe, Flattening of random rough surfaces in metal forming process, ASME J. Tribol., 1999, in press.
[22] P.W. O'Callaghan, S.D. Probert, Real area of contact between a rough surface and a softer optically flat surface, J. Mechan. Eng. Sci. 12 (1970) 259–267.
[23] C.P. Hendriks, M. Visscher, Accurate real area of contact measurements on polyurethane, ASME J. Tribol. 117 (1995) 607–611.
[24] W.T. Lai, H.S. Cheng, Computer simulation of elastic rough contacts, ASLE Trans. 28 (1985) 172–180.
[25] M.N. Webster, R.S. Sayles, A numerical model for the elastic frictionless contact of real rough surfaces, ASME J. Tribol. 108 (1986) 314–320.
[26] J. Seabra, D. Berthe, Influence of surface waviness and roughness on the normal pressure distribution in the Hertzian contact, ASME J. Tribol. 109 (1987) 462–470.
[27] N. Ren, S.C. Lee, Contact simulation of three-dimensional rough surfaces using moving grid method, ASME J. Tribol. 115 (1993) 597–601.
[28] C.Y. Poon, R.S. Sayles, Numerical contact model of a smooth ball on an anisotropic rough surface, ASME J. Tribol. 116 (1994) 194–201.
[29] J.J. Kalker, Y.A. van Randen, A minimum principle for frictionless elastic contact with application to non Hertzian problems, J. Eng. Math. 6 (1972) 193–206.
[30] A. Kubo, T. Okamoto, N. Kurokawa, Contact stress between rollers with surface irregularity, ASME J. Mechan. Design 103 (1981) 492–498.
[31] H.A. Francis, The accuracy of plane strain models for the elastic contact of three-dimensional rough surfaces, Wear 85 (1983) 239–256.
[32] A. Brandt, A.A. Lubrecht, Multilevel matrix multiplication and fast solution of integral equations, J. Comp. Phys. 90 (1990) 348–370.

[33] C.H. Venner, A.A. Lubrecht, Numerical analysis of the influence of waviness on the film thickness of a circular EHL contact, ASME J. Tribol. 118 (1996) 153–161.
[34] S.C. Lee, N. Ren, Behavior of elastic–plastic rough surface contacts as affected by surface topography, load, and material hardness, STLE Tribol. Trans. 39 (1996) 67–74.
[35] K.L. Johnson, Contact Mechanics, Cambridge Univ. Press, Cambridge, 1985, Chaps. 3, 4.
[36] J.J. Kalker, Numerical calculation of the elastic field in a half-space, Comm. Appl. Numer. Methods 2 (1986) 401–410.
[37] R.S. Varga, Matrix Iterative Analysis, Prentice-Hall, Englewood Cliffs, NJ, 1962, Chap. 3.
[38] X. Tian, B. Bhushan, A numerical three-dimensional model for the contact of rough surfaces by variational principle, ASME J. Tribol. 118 (1996) 33–43.
[39] M.R. Hestenes, Conjugate Direction Methods in Optimization, Springer, New York, 1980, Chaps. 2, 3.
[40] B.N. Pshenichny, Yu.M. Danilin, Numerical Methods in Optimization Problems, Nauka, Moscow, 1975, Chap. 3, in Russian.
[41] G. Zoutendijk, Mathematical Programming Methods, North Holland, Amsterdam, 1976, Chap. 16.
[42] W. Hackbusch, Multi-Grid Methods and Applications, Springer, Berlin, 1985, Chap. 2.
[43] A. Sackfield, D. Hills, A note on the Hertz contact problem: a correlation of standard formulae, J. Strain Anal. 18 (1983) 195–197.
[44] J.A. Greenwood, J.H. Tripp, The elastic contact of rough spheres, ASME J. Appl. Mechan. 34 (1967) 153–159.
[45] T. Nogi, T. Kato, Influence of a hard surface layer on the limit of elastic contact: Part I. Analysis using a real surface model, ASME J. Tribol. 119 (1997) 493–500.
[46] O.G. Chekina, L.M. Keer, A new approach to calculation of contact characteristics, ASME J. Tribol. 121 (1999) 20–27.