Development of a free-form tooth flank optimization method to improve pitting resistance of spur gears

Christos Kalligeros 1*, Panagiotis Koronaios 1, Panteleimon Tzouganakis 2, Christos Papalexis 2, Antonios Tsolakis 2, and Vasilios Spitas 1

1Laboratory of Machine Design, National Technical University of Athens, 9 Iroon Polytechniou, 15780 Zografou, Greece
2Laboratory of Machine Elements and Vehicles, University of West Attica, 250 Thivon & P. Ralli, 122 44 Egaleo, Greece

Abstract. Although steel involute gears are the standard solution for gear transmissions, they tend to suffer from poor pitting resistance. Pitting typically occurs when the gear tooth flanks have high equivalent curvature at the contact point and/or when the equivalent curvature is not constant across the contact path leading to high contact pressures and the development of surface fatigue. In this paper a new optimization method is presented to produce spur gear tooth flanks with improved pitting performance compared to involute ones. The tooth flanks are represented as B-spline curves, the control points of which are the variables for the optimization problem. The constraints were designed to ensure that all the examined profiles satisfy the law of gearing and do not contain any cusps or C1 discontinuities. Deterministic and stochastic algorithms were implemented and both closed and open path of contact gear sets were examined to determine the optimum tooth profile. The optimization results show that the maximum equivalent curvature of the optimum profiles is reduced by 83% compared to the corresponding standard profiles, while the deviation from the mean value is reduced by 98%. Both the standard and the optimized gears where examined comparatively also through finite element analysis. For the case selected the maximum contact pressure developed on the optimized gear set was 77% of the respective maximum contact pressure on the standard gear set whereas the corresponding deviation from the mean value was 5%. At the same time, the bending stresses developed in the optimized gear are slightly lower than those in the standard one.

1 Introduction

Involute spur gears have been the standard type of gear transmissions in the majority of industrial applications. They combine line-contact meshing, constant pressure angle and insensitivity to center distance variation while they are easy-to-manufacture and easy-to-find in the market. However their load capacity is restricted to some extent due to the development of high contact stresses that lead to pitting [1-5]. Historically, the main alternatives to involute gears include the cycloid and the circular-arc or Wildhaber-Novikov (W-N) gears. Cycloid gears present good efficiency, less interference and lower pressure angle values but they suffer from low load capacity and sensitivity to manufacturing errors. On the other hand, W-N gears can withstand higher loads due their profile conformity and rolling contact but they are highly sensitive to manufacturing and assembly errors, a disadvantage that hinders their use in most applications [1].

The optimization of gear tooth flanks has been a primary field of interest for the research community and the gear industry. By analyzing the published work on this field in the last few decades (starting from the 90s) it can be seen that gear tooth flank optimization generally follows three main trends. The first trend refers to the optimization around the involute curve focusing on the selection of its optimal parameters or the introduction of optimal profile modifications (i.e., tip and root reliefs). The main objectives of the optimization are the minimization of the weight and the volume of the gears [6-8], the increase of their efficiency and load capacity [9-11] and the improvement of their dynamic and NVH characteristics (mainly through optimal profile modifications) [12-14].

The second trend moves away from the involute curve and experiments on new profile geometries anticipating better performance compared to standard gears. Apart from the cycloid and W-N gears that fall into this category, there is a large number of alternative profiles proposed in the literature. Ariga and Nagata [15] developed a gear profile that has an addendum of circular arc and a dedendum of an involute to combine the high load capacity of circular arc (i.e., W-N) gears with less sensitivity to center distance variations. Tsay and Fong [16] analyzed a helical gear drive with a pinion with

* Corresponding author: c.kalligeros@mail.ntua.gr

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).
circular arc teeth and a gear with involute teeth to achieve point contact (instead of line contact) between the meshing profiles to reduce their sensitivity to misalignments. Hlebanja [17] presented the so-called S-gear design that has an S-shaped contact path, which ensures a significantly larger equivalent radius of curvature at the beginning and the end of meshing compared to the involute. Luo et al. [2] introduced a novel cosine-based gear profile that has lower sliding coefficient and lower bending and contact stresses compared to the involute gears. Litvin et al. [18] investigated a pinion profile that is conjugate to a parabolic rack-cutter to reduce contact pressures and improve the bearing contact between the conjugate profiles. Litvin and Lu [19] proposed a tooth profile generated by a double circular-arc rack profile (in extension of the idea of single-arc W-N gears) to obtain two meshing zones (thus reducing the bending stresses) and to retain the beneficial profile conformity of W-N gears. In the same way, Yang [20] developed a triple circular-arc rack cutter improving the concept of the stepped triple circular-arc profile that comes from the standard Chinese double circular-arc and involute profiles. In addition, apart from new pinion and rack profiles, researchers experimented with new contact path trajectories such as parabolic [21], cubic [22], forth-order [23] or combinative curves (i.e., straight line and circular arc [24]).

All the above cases that belong to the second optimization trend refer to the evaluation of a predetermined geometry (either the pinion profile, the rack profile or the contact path) to access the performance of the produced gear pair and determine its characteristics. Conversely, the third and final optimization trend seeks to design an arbitrary, free-form tooth profile that is based on a predetermined desirable performance characteristic (i.e., maximum efficiency or minimum contact pressures). In this way the designer has direct control on the performance characteristic that needs to be optimized. Following this trend, in order to achieve an infinite number of points where the equivalent curvature and the specific sliding take zero values, Komori et al. [3] developed the so-called ‘LogiX’ profile that is produced from a basic rack that constitutes many continuously divided minute involute curves. Tsai and Tsai [25] designed a high contact ratio gear profile that satisfies a given second-order polynomial function of the pressure angle. Wang et al. [26] proposed a method for the preliminary tooth flank design based on given sliding coefficients. Liu et al. [5] developed a mathematical model that enables the design of a tooth profile with constant relative curvature. Yeh et al. [1] proposed an optimization method to produce free-form tooth flanks implementing non-uniform rational B-Splines to achieve profiles with better average profile conformity and decreased contact and bending stresses. Yu and Ting [27] introduced a free-form conjugation modeling technique to optimize spur gears with the master geometry being represented with B-Splines.

Even if it is advantageous for the designer to deduce the tooth profiles based on good performance that is given in advance rather than experimenting on different profiles trying to attain it [26], less than 5% of the published work on gear tooth flank optimization follows this direction. The main obstacle discouraging researchers on further investigating free-form tooth flanks appears to be the lack of a comprehensive optimization process that includes an adequate representation of the gear tooth profile (or the master geometry in general) and a carefully formulated set of constraints to ensure that all the produced profiles satisfy the law of gearing and contain no cusps or other discontinuities.

In this paper a free-form tooth flank optimization method is presented that employs a B-splines representation of the master geometry (either the pinion or the rack profile), introduces a number of constraints that ensure that the produced profiles are suitable to be used as gear tooth flanks and investigates both deterministic and stochastic algorithms to find the optimal solution. The objective of the optimization was selected to be the improvement of spur gear pitting resistance, which is one of the primary limitations of standard steel gears. By introducing tooth profiles with minimum and constant equivalent curvature, the contact pressures will be reduced along with the probability of the initiation and propagation of cracks under the surface of the tooth flank. The optimized profiles could achieve significant improvements in this area, displaying a reduction of 83% to the maximum value of the equivalent curvature and a reduction of 98% to its standard deviation. The improved performance of the optimized gears compared to the standard involute ones was also proved through finite element analysis.

2 Optimization method

The development of an optimization method requires the determination of the following ingredients:

a. The optimization variables
b. The objective function
c. The optimization constraints
d. The optimization algorithm

2.1 Optimization variables

A free-form optimization method requires the revision of the tooth flank profile for every calculation of the objective function. In order to achieve this efficiently, it is necessary to introduce a robust representation of the tooth flank to be able to control its profile by changing only a few variables. Therefore, during the optimization process the tooth profile is represented by an interpolation curve (i.e., B-splines or Bezier), which is typically described by a set of polynomial functions and a number of control points. The control points of the interpolation curve serve as the variables of the optimization process. By adjusting the coordinates of the control points, the interpolation curve and hence the tooth profile can be actively modified. In this study, B-splines interpolation is implemented since it is a widely-used representation method for CAD systems and offers better local control
of the generated profile compared to other methods (i.e., Bezier).

### 2.1.1 B-splines formulation

The most common formulation of B-splines curves in modern software is based on de Boor’s method [28]. De Boor introduced the use of a knot vector that is constructed based on the breakpoints (i.e., the interpolation points) and the preferable continuity order (i.e., C² continuity) and a recursive expression for the calculation of the B-splines base functions that is given in Eq. (1).

\[ N_{i,p}(\xi) = \begin{cases} 
1 & \text{if } \xi_i \leq \xi < \xi_{i+1} \\
0 & \text{otherwise}
\end{cases} \]

\[ N_{i,p}(\xi) = \frac{\xi - \xi_i}{\xi_{i+p} - \xi_i} N_{i,p-1}(\xi) + \frac{\xi_{i+p+1} - \xi}{\xi_{i+p+1} - \xi_{i+1}} N_{i+1,p-1}(\xi) \quad (1) \]

where \( N_{i,p} \) is the \( i \)-th base function of order \( p \), \( \xi_i \) are the knots and \( 0 \leq \xi \leq 1 \) is the scaled coordinate of the B-spline curve.

The main advantage of de Boor’s formulation is that the B-splines curve is not controlled through the polynomial coefficients of the base functions (as in the case of the Schoenberg formulation [29]) but through a number of control points \( P_i \) (\( i = 0, ..., n_p \)). The coordinates of the B-splines curve are given in Eq. (2).

Forth order B-splines was used instead of cubic ones since it has been shown that increasing the order of the B-spline base functions results to smoother curves, an advantageous trait for a tooth flank profile [1].

\[
\begin{align*}
\begin{cases}
x(\xi) = \sum_{i=0}^{n_p} N_{i,p}(\xi) x_{P_i} \\
y(\xi) = \sum_{i=0}^{n_p} N_{i,p}(\xi) y_{P_i}
\end{cases} & \quad (2)
\end{align*}
\]

where \((x, y)\) the coordinates of the profile and \((x_{P_i}, y_{P_i})\) the coordinates of the control points.

In Fig. 1 an example of the interpolation of a tooth profile with 6 control points is presented. In this case there are 10 optimization variables, namely the two coordinates of each control point except for the first and the last one where one coordinate is free and the other is derived to ensure that the tooth flank extends from the base circle to the addendum circle. In example, for the case of Fig. 1 the optimization variables could be: \( \{X_{CP1}, X_{CP2}, Y_{CP2}, X_{CP3}, Y_{CP3}, X_{CP4}, Y_{CP4}, X_{CP5}, Y_{CP5}, X_{CP6}, Y_{CP6}\} \), while \( X_{CP1} \) and \( Y_{CP6} \) would be derived from Eq. (30).

\[ \begin{align*}
X_{CP1} + Y_{CP1} &= r_k^2 \\
X_{CP6} + Y_{CP6} &= r_g^2
\end{align*} \quad (3) \]

where \( r_k \) the tip radius and \( r_g \) the base radius.

### 2.1.2 Analytical theory of gearing

Apart from the pinion tooth flank, the geometry of the conjugate tooth profile should be given in order to assess the pitting performance of a spur gear pair. Buckingham [30] proved that there is an one-to-one correspondence between the tooth flank profile of the pinion, the tooth flank profile of the conjugate gear, the rack profile and the contact path so that when a single one of the above is given along with the pitch circles, all the other conjugate geometries can be directly calculated.

Let \( y_r = F(x_r) \) be the rack profile that is given and \( r_o \) and \( r_{o2} \) the pitch radius for the pinion and the conjugate gear correspondingly. If the pinion is rotated by an angle \( \theta \), the displacement \( K \) of the rack is given according to the law of gearing by Eq. (4).

\[ K = \theta r_o \quad (4) \]

As for the coordinates \((x_{CP}, y_{CP})\) of the contact path and its inclination \( a \) at each contact point:

\[ \tan a = \frac{y_{CP}}{x_{CP}} = -\frac{dy_{CP}}{dx_{CP}} \quad (5) \]

From Eqs. (4) and (5):

\[ K = -\left( y_{CP} \frac{dy_{CP}}{dx_{CP}} + x_{CP} \right) \quad (6) \]

Finally, the coordinates \((x, y)\) of the pinion gear tooth flank are calculated in Eq. (7), while the coordinates \((x_2, y_2)\) of the conjugate tooth profile are given in Eq. (8) [30]. It is noted that the origin of the coordinate system is located at the pitch point of the gear pair.

\[ \begin{align*}
x &= (x_r + K)\cos \theta - (y_r + r_o)\sin \theta \\
y &= (x_r + K)\sin \theta + (y_r + r_o)\cos \theta - r_0
\end{align*} \quad (7) \]

\[ \begin{align*}
x &= (x_r + K)\cos \theta_2 - (y_r - r_{o2})\sin \theta_2 \\
y &= (x_r + K)\sin \theta_2 + (y_r - r_{o2})\cos \theta_2 + r_o
\end{align*} \quad (8) \]

\[ r_{o2}\theta_2 = -r_o\theta \]

Similarly, if the contact path is given instead of the rack profile, the coordinates of the pinion and the conjugate profiles are given in Eqs. (9) and (10) [30].

\[ \begin{align*}
x &= x_{CP}\cos \theta - (y_{CP} + r_o)\sin \theta \\
y &= x_{CP}\sin \theta + (y_{CP} + r_o)\cos \theta - r_0
\end{align*} \quad (9) \]

\[ \begin{align*}
x_2 &= x_{CP}\cos \theta_2 - (y_{CP} - r_{o2})\sin \theta_2 \\
y_2 &= x_{CP}\sin \theta_2 + (y_{CP} - r_{o2})\cos \theta_2 + r_{o2}
\end{align*} \quad (10) \]

\[ \begin{align*}
r_{o2}\theta_2 &= -r_o\theta
\end{align*} \]
Even though for the aforementioned cases the conjugate geometries can be calculated explicitly, when the pinion tooth profile is the given geometry, the calculation of the conjugate profiles requires the solution of the following implicit equation:

$$\tan^{-1}\left[\frac{x \cos \beta + (y + r_0) \sin \beta}{-x \sin \beta + (y + r_0) \cos \beta - r_0}\right]$$

$$+ \beta = \frac{\pi}{2} + \frac{dy}{dx}$$

(11)

Eq. (11) introduces complexity and a non-negligible computational cost to the calculation that is magnified when it becomes part of an optimization process.

### 2.1.3 Theory of involutization

To overcome the implicit calculation of the conjugate geometries when the pinion tooth profile is given, Spitas et al. [31, 32] developed the theory of involutization that enables the explicit calculation of the conjugate profiles. According to this method each point of the tooth profile is attributed to a local involute curve. By implementing the law of gearing the contact path for every point of the tooth contact point that correspond to the contact path) and by implementing the law of gearing the local involute will have a local base circle equal to the radius of curvature along the contact points of the flanks of the two bodies and has negligible elasticity modulus.

Choose more compliant materials (i.e., reduced elasticity modulus).

According to Hertz theory the maximum pressure that can develop between two contact bodies is:

$$p_{\text{max}} = \frac{3F}{2\pi ab}$$

(18)

$$\alpha = \frac{3\pi F(k_1 + k_2)}{4A + B}$$

(19)

$$b = \frac{3\pi F(k_1 + k_2)}{4A + B}$$

(20)

where \(F\) is the load, \(k_1\) is a function of the materials’ elasticity modulus and Poisson ratios, \(A + B\) is a function of the radii of curvature of the two bodies and \(m, n\) are coefficients that are also a function of the radii of curvature. According to Eqs. (18), (19) and (20) the following measures can be taken to reduce the magnitude of contact pressures:

1. Reduce the load \(F\)
2. Choose more compliant materials (i.e., with reduced elasticity modulus).
3. Increase the radii of curvature of the two bodies at the contact area.

For the specific case of steel spur gears the aforementioned measures can be formed as:

1. Reduce the load \(F\) between the working flanks, i.e., by increasing the contact ratio.

$$x_p = r_{gg} \left[ \frac{r_0^2 - r_{gg}^2}{r_0^2} - \sqrt{1 - \left(\frac{r_{gg}}{r_0}\right)^2} \right]$$

(17)

$$y_p = x_p \left(\frac{r_0}{r_{gg}}\right) - 1$$

### 2.2 Objective function

Pitting is initiated by small cracks that are formed under the contact zone due to the periodic action of Hertzian shear stresses or at the transition point between micropitting and non-micropitting areas at the surface of the teeth. The cracks that originate below the tooth surface do not have a significant effect on gear performance as long as they do not propagate towards the surface. However, due to the action of different mechanisms such as high contact pressures, maldistribution of contact pressures along the tooth flank or lubricant entering the cracks, they can reach the tooth surface causing detachment of small chunks of material [33].

Even though pitting triggered by surface originated cracks can be dealt mainly with better tooth surface finishing, pitting by under-surface cracks can be reduced by optimal design of the gear tooth profiles. Since the origin of these cracks are Hertzian pressures, improving the radius of curvature of the tooth flanks can reduce the contact pressures thus preventing the initiation of under-surface cracks. In addition, a uniform distribution of the radius of curvature along the contact points of the flanks will smoothen the Hertzian pressures thus preventing both the initiation and the propagation of cracks towards the tooth surface.

According to Hertz theory the maximum pressure that develops between two contact bodies is:

$$p_{\text{max}} = \frac{3F}{2\pi ab}$$

(18)

$$\alpha = \frac{3\pi F(k_1 + k_2)}{4A + B}$$

(19)

$$b = \frac{3\pi F(k_1 + k_2)}{4A + B}$$

(20)

where \(F\) is the load, \(k_1\) is a function of the materials’ elasticity modulus and Poisson ratios, \(A + B\) is a function of the radii of curvature of the two bodies and \(m, n\) are coefficients that are also a function of the radii of curvature. According to Eqs. (18), (19) and (20) the following measures can be taken to reduce the magnitude of contact pressures:

1. Reduce the load \(F\)
2. Choose more compliant materials (i.e., with reduced elasticity modulus).
3. Increase the radii of curvature of the two bodies at the contact area.

For the specific case of steel spur gears the aforementioned measures can be formed as:

1. Reduce the load \(F\) between the working flanks, i.e., by increasing the contact ratio.
2. Reduce the curvature term $A + B$, which for spur gears is defined as:

$$A + B = \frac{1}{2} \left( \frac{1}{R_1} + \frac{1}{R_2} \right)$$  \hspace{1cm} (21)

where $R_1, R_2$ the radii of curvature of the two meshing gears.

3. Maintain a near-constant magnitude of contact pressures along the contact path to limit the rise of cracks to the tooth surface. This can be achieved when the product $P(A + B)$ is constant at every contact point during meshing.

In order to maintain a near-constant product of $P(A + B)$ at every contact point it is necessary to achieve a contact ratio that is very close to an integer value (i.e., 1, 2 or 3), otherwise there will be abrupt load changes at the working teeth that cannot be compensated by equally abrupt changes in the radii of curvature. A type of spur gears that has by definition a contact ratio equal to 1 is the closed contact path spur gears (i.e., the gears produced by a sinusoidal rack profile), which are examined by the optimization process along with the more widely-used open contact path ones.

In this study the design of the objective function focuses on the radii of curvature of the meshing tooth flanks. The main goal is to achieve a maximum and constant equivalent radius of curvature along the contact path. The objective function is constructed as:

$$G = w_1 \ mean \left( \frac{1}{R} \right) + w_2 \ max \left( \frac{1}{R} \right)$$

$$+ w_3 \ \frac{1}{2} \int \left( \frac{1}{R} \right) \ dx$$  \hspace{1cm} (22)

$$+ w_4 \ \text{std} \left( \frac{1}{R} \right)$$

where $R$ is the equivalent radius of curvature that is calculated as:

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2}$$  \hspace{1cm} (23)

The radius of curvature at each point is calculated as:

$$\frac{1}{R_i} = \frac{d^2y_i}{dx_i^2} \left( 1 + \left( \frac{dy_i}{dx_i} \right)^2 \right)^{\frac{3}{2}}$$  \hspace{1cm} (24)

The objective function constitutes four terms that refer to the average, the maximum value, the integral along the contact path and the standard deviation of the equivalent curvature (namely the reverse of the equivalent radius of curvature). The first two terms are used to exclude any local curvature maxima, while the last two deal with the whole curvature distribution attempting to keep it constant and close to zero. The values of the weights $w_i$ vary for each case to achieve the best optimization result possible.

### 2.3 Optimization constraints

A critical factor for the development of a successful free-form optimization process for gears is the assurance that every produced curve can be used as a tooth flank profile. In general, in order for an arbitrary curve to be considered for a tooth profile it should satisfy the following principle conditions (along with its conjugate profile):

1. Each point of the profile satisfy the law of gearing.
2. Ensure pitch compatibility.
3. Allow continuous meshing (depends on the contact ratio).

The first constraint of the optimization process refer to the produced curves that have to be monotous, i.e. for ascending x-coordinates, the y-coordinates to be descending. Monotonous curves satisfy the law of gearing and provide conjugate geometries that are smooth, without any cusps or C1 discontinuities. For non-monotonous curves, conjugate geometries like in Fig. 2 can mathematically occur that cannot constitute a gear tooth flank. In order to enhance the robustness of the optimization process, a second constraint refers to the conjugate profile, which is also examined for the appearance of any discontinuities similar to Fig. 2.

Another important constraint concerns the contact ratio. The contact ratio $\varepsilon$ should be at least greater than 1, while for more reliable operation especially in high-speed applications the minimum value is usually set even higher (i.e., at least 1.2 or 1.3). The requirement regarding the contact ratio can be enforced through a strong constraint or it can be incorporated in the objective function with an additional term as in Eq. (30). This term serves as a penalty factor that gets higher values for low contact ratios. The effect of the penalty is controlled by the value of the weight $w_5$.

$$G = w_1 \ mean \left( \frac{1}{R} \right) + w_2 \ max \left( \frac{1}{R} \right)$$

$$+ w_3 \ \frac{1}{2} \int \left( \frac{1}{R} \right)^2 \ dx$$  \hspace{1cm} (25)

$$+ w_4 \ \text{std} \left( \frac{1}{R} \right) + w_5 \ \frac{1}{\varepsilon}$$

As it was mentioned in Par. 2.1, only one of the two coordinates of the first and the last control points is considered a free variable, while the other is derived. In reality, additional constraints should be enforced to the free variables to ensure a minimum and a maximum tooth thickness in the addendum and the base circle respectively. If such constrains are neglected, results similar to Fig. 3 can occur where the two flanks of a single
tooth overlap. The $x$-coordinate of the first control point cannot be lower from a certain value related to the minimum thickness of the top land, which is $0.25m$, $m$ being the module of the gear pair [34]. A similar constrain can be applied to the $x$-coordinate of the last control point regarding a maximum thickness at the base circle.

Fig. 3. Overlapping tooth flank when the constraints are neglected.

2.4 Optimization algorithm

Optimization algorithms can be classified generally into two types; the deterministic and the stochastic ones. Deterministic optimization methods are based on the gradual shifting of an initial solution to the optimal one through stepwise corrections. They typically converge fast to the final solution but they run the risk of not converging to the optimum but being trapped in local extrema, depending on the choice of initial solution. Stochastic methods, on the other hand, are based on the almost random search for new solutions that are better than the existing one, which ultimately leads to the best. Their main feature is that they are not easily trapped in local extrema and are therefore preferred for solving convoluted optimization problems. Their primary drawback is that they often have a high computational cost compared to deterministic methods.

In this study both deterministic and stochastic algorithms are implemented for the optimization problem. The deterministic algorithms are based on the steepest descent method and the *fmincon* function that is incorporated into the Matlab software. For a stochastic algorithm the genetic algorithm *ga* of Matlab is used.

Steepest descent is one of the most popular optimization methods. The optimum solution results from the minimization of an objective function of the form:

$$
\nabla f^T s = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} s_i
$$

where $s$ is the unit vector at a $n$-dimensional space that:

$$
\begin{align*}
\bullet s^T s &= 1 \\
\bullet s &= \frac{\nabla f}{\|\nabla f\|}
\end{align*}
$$

The minimization of the objective function $f$ is achieved through the following formulation:

$$
\begin{align*}
x_{k+1} &= x_k + \alpha s \\
\text{the step } \alpha \text{ is calculated as:} \\
\alpha^* &= \frac{(x_k^T + b^T)s}{(s^T Q s)}
\end{align*}
$$

where the symmetry of the Hessian matrix $Q$ has been used.

The algorithm based on the *fmincon* function is constructed in Matlab. *fmincon* is a nonlinear programming solver that finds the minimum of an optimization problem including various types of constraints, both linear and nonlinear. The results of this algorithm are compared with the ones from the steepest descent method to evaluate the algorithms’ convergence.

Apart from the deterministic algorithms, the genetic algorithm *ga* of Matlab is implemented. Genetic Algorithms are based on the existence of populations that are potential solutions of the problem. Population varies from one generation to another while it undergoes (a) selection that depends whether it is suitable, (b) crossover in order to transfer its features to the next generations and (c) mutation in order to cover the whole space of potential solutions. In genetic algorithms, exploitation is achieved through the parent selection mechanism while exploration is achieved through crossover and mutation [35].

The final flowchart of the optimization process is presented in Fig. 4. It is divided into two paths depending on the initial gear geometry that is interpolated with B-splines. For closed contact path the B-splines represent the rack profile, while for open contact path gears they represent the pinion’s tooth flank profile (unless stated otherwise). Both paths are equivalent and can be used conversely.
3 Results

In order to demonstrate the capabilities of the optimization process a case study is selected that refer to an automotive-graded gear pair. The main parameters of the selected gear pear are given in Table 1.

Table 1. Case study gear parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of teeth - pinion</td>
<td>19</td>
</tr>
<tr>
<td>Number of teeth - gear (wheel)</td>
<td>55</td>
</tr>
<tr>
<td>Transmission ratio</td>
<td>2.895</td>
</tr>
<tr>
<td>module (mm)</td>
<td>2.5</td>
</tr>
<tr>
<td>Addendum coefficient</td>
<td>1.0</td>
</tr>
<tr>
<td>Dedendum coefficient</td>
<td>1.25</td>
</tr>
</tbody>
</table>

As mentioned above gears with both closed and open contact paths are examined. Even though gears with closed contact path are not widely-used to power transmission applications mainly because their marginal contact ratio ($\varepsilon = 1$) cannot ensure reliable operation (i.e., dynamical instabilities, meshing at points outside the theoretical contact path), they are useful for the development of the optimization process since for the same amount of control points, the free variables are fewer compared to the respective case of open contact path gears.

3.1. Closed contact-path gears

In order to produce spur gears with closed contact path the first and the last point of the rack profile should be fixed and have a derivative equal to zero. Consequently, the position of the first and the last control point (that coincides with the first and the last point of the profile) is fixed. In addition, the $y$-coordinates of the second and the penultimate control point are also fixed and equal to those of the first and the last control point respectively to ensure a zero derivative at the extremities of the rack profile as shown in Fig. 5. In this way, for a case of five control points the free variables are only four (i.e., $\{X_{CP2}, X_{CP3}, Y_{CP3}, X_{CP4}\}$), while for the respective case of an open contact path the free variables would be eight (i.e., $\{X_{CP1}, X_{CP2}, Y_{CP2}, X_{CP3}, Y_{CP3}, X_{CP4}, Y_{CP4}, X_{CP5}\}$).

![Fig. 4. Flowchart of the free-form tooth flank optimization process.](image)

![Fig. 5. Interpolated gear rack profile for the production of spur gears with closed contact path.](image)
equivalent curvature distribution of the produced free-form (‘optimal’) profiles as shown in Fig. 6. This attribute is not desirable since even though the average equivalent curvature is improved compared to the involute (Fig. 6, yellow line) and the sinusoidal profiles (Fig. 6, red line), this local, abrupt change can contribute to both the initiation and propagation of under-surface cracks. To avoid such results a careful design of the objective function should be done. The most effective terms in avoiding local maxima of curvature were proved to be the $\text{std} \left( \frac{1}{R} \right)$ and the $\text{max} \left( \frac{1}{R} \right)$.

Fig. 6. Undesirable steep change of the equivalent curvature of the produced free-form profiles.

In Table 2 a selected group of optimization results are presented regarding spur gears with closed contact path. Cases #1 and #2 were run with a steepest descent algorithm, whereas cases #3 and #4 with a genetic algorithm. Even though a variety of different objective functions were examined (i.e., different values of the weights), the cases of Table 2 refer to a specific expression with three terms that was proved to provide the best results. An investigation of the effect of the different weights of the objective function on the results is included in the following section regarding the open contact path gears.

The steepest descent algorithm converged fast to the solutions (i.e., after 100 iterations) whereas the genetic algorithm required significant more time to run (i.e., in some cases 40 times slower). The results are compared with the sinusoidal profile (the most common example of gears with closed contact path) based on the average, the maximum value and the standard deviation of the equivalent curvature of the conjugate gear profiles. The produced profiles of cases #1, #2 and #3 have a very similar distribution of the equivalent curvature as presented in Fig. 7 for case #2. Even though the equivalent curvature are significantly closer to zero in the worst regions of the sinusoidal profile (thus improving significantly the overall metrics of the profile compared to the sinusoidal one), not only it is not constant along the whole contact path (only at certain regions) but also it is inferior in some areas. Similar results occur also when more control points are used (i.e., in case #2 instead of 5 in case #1) or when a different optimization algorithm is implemented (i.e., steepest descent or ga).

In case #4 only the central part of the rack profile was modeled in order to optimize the equivalent curvature of the contact points around the pitch point, since only this part of the theoretical eight-shaped contact path is actually been followed by the meshing gears during operation. As it can be seen in Fig. 9, the equivalent curvature at the optimized region is sufficiently constant along the contact path, with smaller maximum value but with regions that again are inferior to the sinusoidal profile.

In summary the optimization of spur gears with closed contact path has provided profiles with significantly improved curvature characteristics, especially in the regions where the common sinusoidal profile suffer the most. However, due to the problematic contact that takes place between the meshing gears at the extremities of the contact path, the optimization could not ensure both minimum and constant equivalent curvature at the entirety of the contact path mainly since the minimization of the curvature at these regions hindered the characteristics of the other parts of the profile. This can be also seen from case #4 where the extremities of the contact path were excluded from the optimization and the result was significantly more constant than before. Nonetheless, the examination of these gears allowed the development of the optimization process and the investigation of the effect of some important parameters such as the weights of the objective function, the number of control points, the optimization algorithms and their parameters.

<table>
<thead>
<tr>
<th>Case</th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
<th>#4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of control points</td>
<td>5</td>
<td>8</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Objective function</td>
<td>$\text{max} \left( \text{abs} \left( \frac{1}{R} \right) \right) + \text{mean} \left( \text{abs} \left( \frac{1}{R} \right) \right) + \text{std} \left( \frac{1}{R} \right)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Algorithm</td>
<td>Steepest Descent</td>
<td>Genetic Algorithm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population size</td>
<td>-</td>
<td>-</td>
<td>300</td>
<td>200</td>
</tr>
<tr>
<td>Number of generations</td>
<td>-</td>
<td>-</td>
<td>900</td>
<td>900</td>
</tr>
<tr>
<td>Results (% Improvement from sinusoidal profile)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean of eq.curvature</td>
<td>94.69%</td>
<td>93.25%</td>
<td>94.61%</td>
<td>93.14%</td>
</tr>
<tr>
<td>Max of eq. curvature</td>
<td>75.73%</td>
<td>79.27%</td>
<td>76.17%</td>
<td>81.48%</td>
</tr>
<tr>
<td>Standard deviation of eq. curvature</td>
<td>68.21%</td>
<td>73.42%</td>
<td>67.64%</td>
<td>85.58%</td>
</tr>
</tbody>
</table>
3.2 Open contact-path gears

Gears with open contact path and especially involute gears are the most widely-used type of gears in power transmissions. Unlike the case of closed contact path gears, the contact ratio of these gears are not by definition equal to 1 but it depends on the tooth profile geometry. For this reason, the term referring to the contact ratio is added to the objective function as presented in Eq. (25).

Generally, in order to minimize and stabilize the equivalent curvature the tooth profile has a tendency to stretch as in Fig. 10. For this result, the equivalent curvature is almost exactly zero except from a minor region of the contact path and displays a remarkable stability. However, apart from the fact that the working and the coasting flanks are overlapping (Fig. 10), the contact ratio is only equal to 0.74. As it was seen from the optimization results in general, it can be concluded that there is a ceiling in the optimization results preventing the realization of a uniform, near-zero distribution of the equivalent curvature. The goal of the optimization should then be to achieve a constant curvature with the minimum possible value.

In Table 3 a selected group of optimization results using the steepest descent method are presented. For cases #5 and #6 the only differences are found on the weights of the objective function. Both cases achieve excellent results regarding the stability of the equivalent curvature while avoiding the high curvature values of the involute profile at the end of the contact path (Fig. 11). Case #5 maintains lower equivalent curvature values at the entirety of the contact path but comes with a marginal contact ratio of 1.04. On the other hand, the gears of case #6 have a contact ratio of 1.45, an excellent curvature stability but slightly worse values at a substantial part of the contact path. This compromise between the magnitude of the equivalent curvature and the contact ratio has been a pattern of the optimization results and can be sufficiently handled by the values of the weights of the objective function. Furthermore, as it was discussed above, the inclusion of additional control points does not necessarily deliver better results as shown in case #7, where 6 control points were implemented (instead of 5 for cases #5 and #6) but the stability of the equivalent curvature distribution was compromised as it is evident from its standard deviation increase (i.e., drop in the % improvement compared to the involute).

In Fig. 12 the tendency of the tooth flank to expand and create a more flat profile is clear compared to the involute. This result to a much smaller top land (Fig. 13) and a larger thickness below the pitch circle. The path of contact does not follow a straight line as expected.
Table 3. Optimization results for spur gears with open contact path using a steepest descent algorithm.

<table>
<thead>
<tr>
<th>Case</th>
<th>#5</th>
<th>#6</th>
<th>#7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
<td>Steepest Descent</td>
<td>Steepest Descent</td>
<td>Steepest Descent</td>
</tr>
<tr>
<td>Number of control points</td>
<td>5</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>$w_1$ - $\frac{1}{\rho} \int \frac{1}{\rho} , dx$</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$w_2$ - $\frac{1}{\rho} \int \frac{1}{\rho} , dx$</td>
<td>0.5</td>
<td>0.5</td>
<td>1.4</td>
</tr>
<tr>
<td>$w_3$ - $\frac{1}{\rho} \int \frac{1}{\rho} , dx$</td>
<td>2.5</td>
<td>0.5</td>
<td>2.5</td>
</tr>
<tr>
<td>$w_4$ - $\frac{1}{\rho} \int \frac{1}{\rho} , dx$</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$w_5$ - $\frac{1}{\rho}$</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Results (% Improvement from involute profile)

| Mean of eq. curvature | 36.62% | 8.73% | 26.38% |
| Max of eq. curvature | 82.89% | 75.33% | 78.39% |
| Standard deviation of eq. curvature | 98.18% | 97.64% | 88.87% |
| Contact ratio | 1.04 | 1.45 | 1.22 |

In Table 4 three optimization cases using an fmincon algorithm are presented. The results are very similar to those from the steepest descent algorithm and display the same compromise between the magnitude of the equivalent curvature and the contact ratio. Cases #8, #9 and #10 have almost identical results with cases #5, #6 and #7 respectively both in terms of the equivalent curvature (as indicated by comparing the metrics of Tables 3 and Table 4) and the contact ratios. As it can be seen also from Fig. 14 the equivalent curvature distribution of case #8 is practically the same with case #5 (Fig. 11a) even if the results were reached with different optimization results. The same applies with the other cases. This observation enhances the reliability of the optimization results.

Table 4. Optimization results for spur gears with open contact path using an fmincon algorithm.

<table>
<thead>
<tr>
<th>Case</th>
<th>#8</th>
<th>#9</th>
<th>#10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
<td>Fmincon</td>
<td>Fmincon</td>
<td>Fmincon</td>
</tr>
<tr>
<td>Number of control points</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>$w_1$ - $\frac{1}{\rho} \int \frac{1}{\rho} , dx$</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$w_2$ - $\frac{1}{\rho} \int \frac{1}{\rho} , dx$</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>
Apart from the deterministic algorithms, the genetic algorithm ga of Matlab was also implemented to optimize the gear tooth flanks. Similarly to the previous cases, a parametric analysis regarding the values of the weights of the objective function was performed and different numbers of control points were used to achieve the best possible results. In addition, in this case not only the pinion tooth flank was used as the main geometry (i.e., the interpolated profile) but also the rack profile to investigate if there is any influence to the produced profile.

In Table 5 optimization cases referring to the genetic algorithm are presented. The results show that on the one hand the average equivalent curvature is improved compared to the previous cases (as indicated by the mean values) but on the other hand its stability was compromised (as indicated by the standard deviation values and Fig. 15). Similar results were produced even with different number of control points, with different population sizes and different initial geometries (i.e., pinion or rack profile). The from the analysis two main conclusions can be drawn. Firstly, besides the compromise between the magnitude of the equivalent curvature and the contact ratio, there is also a tradeoff between the magnitude and the stability of the equivalent curvature. If the main requirement for the equivalent curvature is to be constant, then there is a ceiling to its minimum value. Otherwise, if a limited fluctuation is permitted as in Fig. 15, significant improvements can be achieved in several regions of the contact path. Secondly, as the number of control points and thus the number of optimization variables increases, the genetic algorithm needs a lot more generations (i.e., more than 2,000) to converge. Consequently, a significant amount of computational resources is required but also it is more difficult to control the outcome of the solution (i.e., prioritizing minimum values or stability) through the weights of the objective function compared to the deterministic algorithms.

In summary, the optimization process of gears with open contact path produced tooth profiles with lower and more constant equivalent curvature than standard involute gears. Depending on the requirement of the application, the maximum value or the stability of the equivalent curvature can be prioritized through the appropriate selection of the weights of the objective function. In order to achieve both constant and low equivalent curvature, the

### Table 5. Optimization results for spur gears with open contact path using a genetic algorithm.

<table>
<thead>
<tr>
<th>Case</th>
<th>#11</th>
<th>#12</th>
<th>#13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
<td>Genetic algorithm</td>
<td>Genetic algorithm</td>
<td>Genetic algorithm</td>
</tr>
<tr>
<td>Number of control points</td>
<td>6</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>w₁ - mean ( \left( \frac{1}{R} \right) )</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>w₂ - std ( \left( \frac{1}{R} \right) )</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>w₃ - ( \frac{1}{2} \int R \left( \frac{1}{R} \right)^2 dx )</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>w₄ - ( \max \left( \frac{1}{R} \right) )</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>w₅ - ( \frac{1}{R} )</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Population size</td>
<td>400</td>
<td>400</td>
<td>800</td>
</tr>
<tr>
<td>Number of generations</td>
<td>1900</td>
<td>2300</td>
<td>2300</td>
</tr>
</tbody>
</table>

### Results (% Improvement from involute profile)

<table>
<thead>
<tr>
<th></th>
<th>Case</th>
<th>#11</th>
<th>#12</th>
<th>#13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean of eq. curvature</td>
<td>63.58%</td>
<td>64.93%</td>
<td>65.39%</td>
<td></td>
</tr>
<tr>
<td>Max of eq. curvature</td>
<td>81.70%</td>
<td>81.16%</td>
<td>80.96%</td>
<td></td>
</tr>
<tr>
<td>Standard deviation of eq. curvature</td>
<td>45.44%</td>
<td>47.97%</td>
<td>48.17%</td>
<td></td>
</tr>
<tr>
<td>Contact ratio</td>
<td>1.14</td>
<td>1.17</td>
<td>1.19</td>
<td></td>
</tr>
</tbody>
</table>

**Fig. 14.** The equivalent curvature of the ‘optimal’ gears of case #8 (fmincon).

[Image of equivalent curvature graph]

**Fig. 15.** The equivalent curvature of the ‘optimal’ gears of case #11 (genetic algorithm).

[Image of equivalent curvature graph]
contact ratio tend to become marginal (i.e., close to 1), while to further minimize the equivalent curvature, its stability should be compromised. However, the produced results already display an improved performance compared to standard gears as it was also proved by a finite element analysis (FEA) that was carried out.

4 Finite element analysis

The FE analysis was performed in ANSYS Workbench and two gear pairs were investigated; an involute gear pair with the properties of Table 1 and the optimized gear pair of case #5. The main parameters of the analysis are presented in Table 6.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Load (Nmm)</td>
<td>5,000</td>
</tr>
<tr>
<td>Material</td>
<td>Steel</td>
</tr>
<tr>
<td>Young’s modulus (MPa)</td>
<td>200,000</td>
</tr>
<tr>
<td>Poisson ratio</td>
<td>0.3</td>
</tr>
<tr>
<td>Type of analysis</td>
<td>2D</td>
</tr>
</tbody>
</table>

To design the mesh distribution on the contact area of the meshing gears the theoretical contact width was first calculated from Hertz theory and it was divided into approximately 100 elements implementing the Edge Sizing and Refining tools of ANSYS Static Structural. The mesh distribution is presented in Fig. 16.

The results showed that the optimized tooth profiles produced consistently lower contact pressure values than their standard involute counterparts, achieving about 33% improvement in some cases. In Fig. 17 the equivalent Von Mises stresses of the two profiles are presented when they are meshing at their pitch point. The maximum contact pressures at this case are identical to the maximum Von Mises stresses (±0.5%). Furthermore, the maximum pressure of the optimized gears is almost constant along the contact path (±2%), though this would change for bigger contact ratios at the transition between single and double tooth contact. Apart from the contact stresses, the optimized gears produce slightly lower bending stresses as well (in the order of 20% lower, but it depends on the selection of the rack tip radius).

5 Conclusions

In this study a free-form tooth flank optimization process was presented aiming to improve the pitting performance of spur gears by controlling the equivalent curvature of the conjugate tooth profiles. The main objective of the optimization was the realization of an equivalent curvature distribution that is constant along the contact path and has the minimum possible value to prevent both the initiation and propagation of under-surface cracks due to the presence of Hertzian pressures. The optimization included cases with both open and closed contact path gear pairs and the produced profiles were compared with standard involute and sinusoidal gears to evaluate the performance of the optimization process.

The variables of the optimization process were the control points of a B-splines curve that represented either the pinion tooth flank or the rack profile. For every new profile that was produced, its conjugate geometries were calculated by the analytical theory of gearing and the theory of involutization [30-32]. With the meshing tooth profiles known, the equivalent curvature of the gears were calculated and been imported to the objective function that aimed to minimize and stabilize its distribution along the contact path. The produced profiles should comply with a number of constraints that ensured the satisfaction of the law of gearing and the prevention of the appearance of any cusps or discontinuities on the tooth profiles. Both deterministic (steepest descent and Matlab’s fmincon) and stochastic (Matlab’s ga) algorithms were implemented for the convergence of the optimization process.
The optimization results showed that there is a compromise between the magnitude of the equivalent curvature and the contact ratio of the tooth profiles. As the equivalent curvature becomes lower, the contact ratio decreases until its value becomes marginal (i.e., close to 1) thus forming a ceiling to the minimum possible equivalent curvature values. In addition, there is also a tradeoff between the minimization and the stability of the equivalent curvature distribution along the contact path. In order to achieve a near-perfect constant distribution, the average value of the equivalent curvature should be sacrificed.

A number of optimization cases were presented using both deterministic and stochastic optimization algorithms. Steepest descent and Matlab’s fmincon algorithms provided fast convergence and results with near-perfect constant equivalent curvature. Matlab’s ga required significantly more computational resources (i.e., even 100 times more) but provided results with lower equivalent curvature values by compromising slightly the stability of the equivalent curvature distribution.

The two main objectives of the optimization is the minimization and the stabilization of the equivalent curvature of the gear pair. When the first is the main priority, the average equivalent curvature value of the optimized profiles can be reduced by as much as 39% compared to standard involute gears, its maximum value can be reduced by 83% and its standard deviation by 98%. When the latter is the prime objective the average equivalent curvature can be reduced by 65%, the maximum value by 82% and the standard deviation by 48%.

The improved equivalent curvature characteristics of the optimized profiles lead to the development of lower contact pressures as proved by a finite element analysis carried out using ANSYS Workbench. The optimized gears displayed consistently lower pressures than standard involute gears and in some cases a reduction of 33% was achieved. In addition, the bending stresses were also improved by as much as 20%.

References

557. https://doi.org/10.1016/j.mechmachtheory.2010.11.005


