Data Engineering by the Best \( \ell_1 \) Convex Data Fitting Method

S. S. Papakonstantinou and I. C. Demetriou

Abstract—We consider the application of a method that makes the least sum of absolute value change to measured values of a convex function contaminated with random errors to achieve convexity. Thus, we present how to analyze, summarize and interpret these data. The method uses an algorithm of descent direction that employs Karush-Kuhn-Tucker like parameters that are both important to the characterization of the solution and useful for sensitivity analysis. Convexity is expressed in terms of non-negative second divided differences of the smoothed data, which gives a linear programming calculation that is subsequently solved by this algorithm. The data set that is employed for illustration is the time series of the annual Gini coefficients in the U.S.A. from 1947 to 1996. The results are analyzed and the interpretation capability of the method is demonstrated.

Index Terms—convex data fitting, diminishing return, divided difference, \( \ell_1 \) approximation, linear programming

I. INTRODUCTION

What hypotheses can be made about the nature of a convex function, which is known only by a set of measured values that have lost convexity due to errors of measurement? The answer to this question has strong ties with many applications of economics, science, engineering and medicine. Applications in economics, for example, arise when assuming diminishing rates of change of certain demand and production relations [14]. In medical imaging and robotic vision applications, convex polygonal curves are recovered from measurements of convex sets that include errors by optimizing some measure of performance [11]. Other examples arise from univariate convex data fitting, from decision making [13] and biology [8], for instance. In particular, the method may be used in order to explore possible convex relationships between the variables in multivariate data analysis, thus providing a useful companion to data exploration methodologies.

The purpose of this paper is to present how to analyze, summarize and interpret a set of measured values of a convex function \( f(x) \) contaminated with random errors by the use of a \( \ell_1 \) data approximation method that minimizes the sum of the moduli of the errors subject to the condition that the second divided differences of the smoothed data are nonnegative. There are many advantages to using \( \ell_1 \) data approximation techniques in practice (see, for example, [15]) and this paper is an aid to application of a specific \( \ell_1 \) algorithm for achieving convexity.

The data are the pairs \( (x_i, \phi_i) \), \( i = 1, 2, \ldots, n \), where the abscissae \( x_i \), \( i = 1, 2, \ldots, n \) satisfy the inequalities \( x_2 < \cdots < x_n \), and \( \phi_i \) is the measurement \( f(x_i) \). We assume that \( \phi_i = f(x_i) + \epsilon_i \), where \( \epsilon_i \) is a random number. We also assume that there are some gross errors in the data due to blunders.

The authors [17] addressed the problem of calculating numbers \( y_i \), \( i = 1, 2, \ldots, n \), from the measurements that are smooth and closer than the measurements to the true function values. They regarded the original data and the smoothed values as vectors \( \phi \) and \( y \) respectively in \( \mathbb{R}^n \) and considered the problem of minimizing the objective function

\[
\|\phi - y\|_1 = \sum_{i=1}^{n} |\phi_i - y_i|, \quad y \in \mathbb{R}^n,
\]

subject to the convexity constraints

\[
y[x_{i-1}, x_i, x_{i+1}] \geq 0, \quad i = 2, 3, \ldots, n - 1,
\]

where

\[
y[x_{i-1}, x_i, x_{i+1}] = \frac{y_{i-1}}{(x_{i-1} - x_i)(x_{i+1} - x_i)} + \frac{y_{i+1}}{(x_{i+1} - x_i)(x_{i+1} - x_i)} - \frac{y_i}{x_{i+1} - x_i}, \quad i = 2, 3, \ldots, n - 1.
\]

Also, we let \( a_{ij} \) be the \( i \)th component of \( a_{ij} \). Since each divided difference depends on only 3 adjacent components of \( y \), it immediately follows that the constraints have linearly independent normals.

Since (1) is continuous in \( y \) and tends to infinity as \( \|y\|_1 \to \infty \) for feasible \( y \) and the set of feasible vectors is closed, a finite solution exists. We call it a best \( \ell_1 \) convex fit to \( \phi \). Since (1) is not strictly convex, the solution need not be unique.

In view of [12], this problem may also be derived when the data come from processes that show increasing rates of change (cf. convexity), but one does not have sufficient information to set up a parametric form for the underlying function. Thus, by writing the \( i \)th second divided difference in the form

\[
y[x_{i-1}, x_i, x_{i+1}] = \frac{y[x_{i+1}, x_{i+1}] - y[x_{i-1}, x_i]}{x_{i+1} - x_{i-1}}, \quad i = 2, \ldots, n - 1,
\]

where

\[
y[x_{i-1}, x_i] = \frac{y_i - y_{i-1}}{x_i - x_{i-1}}.
\]
is the $i$th first divided difference, the inequalities on the rates of change of the sequence $\{y_i : i = 1, 2, \ldots, n\}$

$$\frac{y_i - y_{i-1}}{x_i - x_{i-1}} \leq \frac{y_{i+1} - y_i}{x_{i+1} - x_i}, \quad i = 2, 3, \ldots, n - 1$$

imply the inequalities (2). Therefore, an alternative expression of the constraints (2) is that we require increasing rates of change on $[x_1, x_n]$, a property that is quite common in describing processes, where a potential shape for the underlying function is that of a convex curve.

The piecewise linear interpolant through the points $(x_i, y_i), i = 1, 2, \ldots, n$ provides some useful geometric description. If all the differences (2) are zero, then the smoothed values lie on a straight line. Otherwise some divided differences are positive, which makes this interpolant be a convex polygonal curve. It is interesting to note that the knots of this polygon are a subset of the abscissae which are not known in advance, but they are calculated automatically from the process. Further, if $n > 4$ and if the data lie on two straight lines in the shape of letter “V” with one of the data at the vertex, then no other convex function can interpolate them [5].

Besides that the shape of convexity is likely to strike immediately a user’s eye when he inspects the data, two properties of this calculation, which provide some advantages over other smoothing techniques [3], are as follows. There is no need to choose a set of approximating functions, because the missing property of convexity is imposed as a smoothing condition, namely inequalities (2), in an optimization calculation that undertakes the process. The approximation process is a projection because, if the data satisfy the convexity constraints, then the data provide the required approximation.

Similar problems are studied and characterized by [23], where (1) is replaced by the supremum norm

$$\| \phi - y \|_\infty = \max_{1 \leq i \leq n} | \phi_i - y_i |$$

and by [7], where (1) is replaced by the least squares norm

$$\| \phi - y \|_2^2 = \sum_{i=1}^{n} (\phi_i - y_i)^2.$$  

Expression (8) is appropriate when the data errors have a uniform distribution, while expression (9) is appropriate when the data errors have a normal distribution. Methods that rely upon (1) are well suited to long tailed error distributions, like Cauchy or Laplace, and have the remarkable property of ignoring some gross errors in the data that makes it “markedly superior among the $L_p$ norms” as [21] states. For a general reference on $\ell_1$ approximation from finite dimensional subspaces see [18].

The paper is organized as follows. In Section II characterization conditions are stated, which resemble the Karush-Kuhn-Tucker conditions, that are important both in theory and in developing efficient algorithms. The associated Lagrange multipliers as well as some other parameters derived from the optimal fit are highly informative for practical analyses and applications of the problem. In Section III the method is applied to the times series of the Gini coefficients in the U.S.A. for the period 1947-1996. An optimal fit is calculated, the values of the mentioned multipliers and parameters are considered and several features of the data are revealed. This example is worked out as an illustration of the optimality conditions that a best $\ell_1$ convex fit satisfies.

A similar analysis may well be applied to a variety of situations which may arise in several fields. Finally, some concluding remarks are presented in Section IV. Based on the characterization conditions, the authors have developed an algorithm of descent direction and one of the authors [16] has implemented the method in Matlab. The program consists of about 200 lines including a simple driver, which gives an idea of the size of the required calculation in this environment.

Because sometimes it would be better to employ nonpositive instead of nonnegative second divided differences, the method may be applied after a change of sign of the components of $\phi$, which implies diminishing rates of change (cf. concavity) of the sequence $\{y_i : i = 1, 2, \ldots, n\}$, giving the inequalities

$$\frac{y_i - y_{i-1}}{x_i - x_{i-1}} \geq \frac{y_{i+1} - y_i}{x_{i+1} - x_i}, \quad i = 2, 3, \ldots, n - 1.$$  

II. BEST $\ell_1$ DATA FITTING SUBJECT TO NONNEGATIVE SECOND DIVIDED DIFFERENCES

The general $\ell_1$ linear approximation problem may be formulated as a primal or a dual linear programming calculation [1], [2]. These formulations allow certain characterization theorems and specific numerical methods [9] for implementing the simplex method [6]. Therefore, it is straightforward to minimize the objective function (1) subject to the convexity constraints (2) by using a standard simplex method. However, because several thousand data points may occur in many smoothing calculations, a special technique for this problem has been developed by the authors using search of directions of descent [22] and taking into account the structure of the constraints. This method gains an order of magnitude over the classical simplex approaches, both in storage and in number of operations during the iterations.

Let $A$ be a nonempty subset of $\{2, 3, \ldots, n-1\}$ and let $|A|$ be the number of elements of $A$. The problem that minimizes (1) subject to the equality or active constraints

$$y_i[x_{i-1}, x_i, x_{i+1}] = 0, \quad i \in A,$$

has particular interest to the inequality constrained problem of Section I. It is a discrete $\ell_1$ approximation problem from the linear subspace defined by the equations (11) to the finite set of values $\{\phi_i, i = 1, 2, \ldots, n\}$. It always has a solution, which could be determined by interpolation to some components of $\phi$, as the following theorem shows.

**Theorem 1:** There exists a vector $y$ that minimizes (1) subject to the constraints (11) and that has the property

$$y_i = \phi_i, \quad i \in I \subseteq \{1, 2, \ldots, n\},$$

with set $I$ containing at least $n - |A|$ indices.

**Proof:** A proof that is based on [20] is provided by [17].

The theorem states that a best $\ell_1$ fit $y$ to $\phi$ subject to (11) may be calculated by seeking a set $I$ that allows $y$ to be obtained by interpolation to the points $\{\phi_i : i \in I\}$. Since set $I$ is not known in advance, a method of searching is
needed that should also test the optimality of a trial set of interpolation points [19]. The following theorem provides optimality conditions for this problem.

**Theorem 2**: Let $s_i$ be the sign of $y_i - \phi_i$,

$$s_i = \begin{cases} -1, & \phi_i > y_i \\ 0, & \phi_i = y_i, \ i = 1, 2, \ldots, n \\ 1, & \phi_i < y_i. \end{cases}$$

A vector $\mathbf{y} \in \mathbb{R}^n$ minimizes (1) subject to (11) if and only if there exists a vector $\mathbf{v}$ in

$$V = \{ \mathbf{v} \in \mathbb{R}^n : |v_i| \leq 1, \ i \in I; v_i = s_i, \ i \notin I \}$$

such that

$$\mathbf{y}^T \mathbf{v} = 0.$$  

**Proof**: See Theorem 6.1 of [24].

Therefore in order that $y$ minimizes (1) subject to (11), it suffices to find a vector $\mathbf{v}$ in $V$ that is orthogonal to $y$.

Furthermore, in order to minimize (1) subject to (2), Theorem 1 suggests searching for a best $\ell_1$ fit among feasible (i.e. convex) vectors that satisfy the interpolation conditions (12). Therefore several applications of Theorem 2 may be needed before a solution is reached. Theorem 3 below provides conditions for testing whether a feasible vector that satisfies the conditions of Theorem 1 is optimal [17]. It is remarkable that although the objective function (1) is non-differentiable, these conditions are given in terms of Karush-Kuhn-Tucker description.

**Theorem 3**: Let $y^*$ be a vector that minimizes (1) subject to (11) and let $s_{i^*}$ be the sign of $y_{i^*}^* - \phi_i$. Then $y^*$ minimizes...
(1) subject to (2), if and only if there exist nonnegative multipliers \( \lambda_j, j \in A \), such that

\[
s^*_i = \sum_{j \in A} \lambda_j a_{ij}, \quad i \notin I.
\]

Motivated by this theorem we write \( y \) as a linear combination of the constraint normals \( a_j, j \in A \),

\[
y = \sum_{j \in A} \lambda_j a_j,
\]

which, in view of (11), satisfies the orthogonality condition (14) and, in view of (15), satisfies the relations \( v_i = s_i, \ i \notin I \) that appear in (13). In addition, by invoking the separating hyperplane theorem (see, for example, [25]), it can be proved that a vector \( y^* \) minimizes (1) subject to (11) if and only if

\[-1 \leq v_i \leq 1, \ i \in I.
\]

Hence, the construction of \( y \) is complete. Further, for notational purposes, we let \( \lambda_i = v_i \), for \( i = 1, 2, \ldots, n \).

The parameters \( \lambda_j \) and \( \lambda_i \) are important to the development of an algorithm of descent direction, both for adding and deleting interpolation points, and for deleting from and adding to the active set constraints iteratively until the conditions of Theorem 3 are satisfied.

### III. DATA ENGINEERING BY THE BEST \( \ell_1 \) CONVEX FIT METHOD

In this section we analyze, summarize and interpret the results obtained by a best \( \ell_1 \) convex fit to a particular data set, which shows a convex pattern. Specifically, the data are the values of the Gini coefficient and its evolution in the U.S.A. for the time period 1947 to 1996. The same data set is used by [10]. The Gini coefficient is commonly used as a measure of inequality of income or wealth [4], where a value of 0 expresses total equality and a value of 100 maximal inequality. Fifty pairs of data were retrieved from the World Income Inequality Database of the U.S. Bureau of Census 1997 and presented in the second \( (x_i) \) and third \( (\phi_i) \) column of Table I. The interest here does not lie on any theoretical assumption of convexity nor on any underlying relation that has to be validated, but on explaining numerically what is happening with the optimality conditions of the calculated best \( \ell_1 \) convex fit to the data. Furthermore, we are not interested in the physical details of the process, but only in what they imply for the convex relationship.

The data were fed to the computer program without any preliminary analysis. The initial components for \( y \) were set to the line that interpolates the points \((x_1, \phi_1)\) and \((x_{50}, \phi_{50})\) and the solution was reached with the set of data interpolation indices \( I = \{1, 6, 13, 19, 29, 30, 32, 41, 50\} \) and the set of indices of active constraints \( A = \{2, 3, \ldots, 49\} \setminus \{7, 9, 23, 29, 30, 33, 45\} \). The components of the calculated best \( \ell_1 \) convex fit are presented in the fourth column \( (y_i) \) of Table I and displayed in Fig.1. Index 6 \( \in I \) implies the interpolation condition \( y_6 = \phi_6 = 36.80 \). Index 6 \( \in A \) implies the active constraint \( y[x_5, x_6, x_7] = 0 \). To the contrary, since index 7 \( \notin A \), the 7th constraint is inactive, giving \( y[x_6, x_7, x_8] = 0.0231 > 0 \). Moreover, the sequences of the first and the second divided differences of the best convex fit are presented in the fifth and sixth column respectively. They are useful to the analysis of the results, because they identify and reveal local trends, such as linearities and convexities. The Lagrange multipliers and the components of vector \( \lambda \) associated with the best fit are presented in the seventh \( (\lambda_j) \) and eighth column \( (\lambda_i) \) respectively.

Having the sets \( A \) and \( I \) available, the components of \( y \) are defined by interpolation to some components of \( \phi_i \), in view of (12), and by solving the system of the divided differences that are equal to zero, in view of (11). In case of degeneracy, the inequality \( |A| + |I| > n \) may hold, because set \( I \) may contain more than \( n - |A| \) elements. Since degeneracy can be avoided in practice, the method arranges the calculation so that \( |A| + |I| = n \). The solution has given \( |A| = 41 \) and \( |I| = 9 \). Hence the components \( \{y_i : i \notin I\} \) are calculated by solving a \((n - |I|) \times (n - |I|)\) system of equations whose coefficient matrix elements are obtained by deleting a row and column of the \( n \times n \) coefficient matrix of the linear equations (11) and (12) for each \( i \in I \). This matrix is considered later on when discussing the Lagrange multipliers.

We can immediately notice the non-decreasing property of the sequence of the first divided differences in column 5 as stated by conditions (7). These differences are negative in the interval [1947,1969] and positive subsequently. Therefore, the smoothed Gini coefficients decrease in [1947,1969] down to the value \( y_{27} = 35.27 \) and increase subsequently, with a rate of change of income inequality that increases gradually from negative to positive values, as it is shown in Table II.

![Fig. 1. Best \( \ell_1 \) convex fit (+) of the Gini coefficients of Table I for the years 1947-1996. The piecewise linear interpolant illustrates the fit](image-url)

### TABLE II

<table>
<thead>
<tr>
<th>Period</th>
<th>1st divided differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1947 - 1953</td>
<td>−0.1600</td>
</tr>
<tr>
<td>1953 - 1955</td>
<td>−0.1033</td>
</tr>
<tr>
<td>1955 - 1969</td>
<td>−0.0833</td>
</tr>
<tr>
<td>1969 - 1975</td>
<td>0.0722</td>
</tr>
<tr>
<td>1975 - 1976</td>
<td>0.1000</td>
</tr>
<tr>
<td>1976 - 1979</td>
<td>0.2500</td>
</tr>
<tr>
<td>1979 - 1991</td>
<td>0.3438</td>
</tr>
<tr>
<td>1991 - 1996</td>
<td>0.3650</td>
</tr>
</tbody>
</table>

The nonnegativity of the sequence of the second di-

---


**ISBN:** 978-988-19252-7-5  
**ISSN:** 2078-0958 (Print); **ISSN:** 2078-0966 (Online)
vided differences in column 6 shows the convexity of the piecewise linear interpolant to the smoothed values. Since points with zero second divided differences lie on a straight line and since the positive second divided differences are centered at the abscissae with indices \{7, 9, 23, 29, 30, 33, 45\}, the calculated smoothed values lie on a convex polygonal line that consists of eight consecutive line segments that join the smoothed values at the abscissae \(x_1, x_7, x_9, x_{23}, x_{29}, x_{30}, x_{33}, x_{45}\) and \(x_{50}\). The rates of change shown in Table II are the slopes of the piecewise linear interpolant to the smoothed values.

All the Lagrange multipliers in column 7 are nonnegative, according to Theorem 3. We see that whenever \(\lambda_j > 0\), then \(y_j(x_j - 1) = 0\). In words, if the \(j\)th Lagrange multiplier is strictly positive, the \(j\)th constraint is by necessity an active constraint at the solution. However, if the \(j\)th constraint is strictly positive, then the method ignores the associated \(j\)th multiplier. For each vector \(y = y^*\), we obtain the system of equations (15) and we let the corresponding Lagrange multipliers be obtained by solving this system. In words, the columns of the \((n - |I|) \times (n - |I|)\) matrix of system (15) are the normals of the active constraints after deleting the rows for \(i \in I\). It is remarkable that the matrix that occurs in the definition of \(\lambda\) is the transpose of the matrix that is used to define \(y\). In Fig. 2 we see the coefficient matrix of (15), emphasizing its sparsity. The first row of numbers displays the indices of the active constraints, namely \(A\). The first column of numbers displays the data indices after excluding the interpolation point indices, namely \(\{1, 2, \ldots, 50\} \setminus I\). Because the abscissae are equally spaced and because multiplication of this matrix by a constant does not change the problem, the matrix elements in Fig. 2 come from the coefficients of the second difference giving a quint diagonal band and similarly if the abscissae are not equally spaced. The higher the value of a Lagrange multiplier, the stronger the linear tendency of the corresponding constraint is. The symmetric values of the Lagrange multipliers observed in the intervals [1956, 1968] and [1980, 1990] are due to the local symmetries of the coefficient matrix of system (15) and the distribution of the \(\pm 1\) signs at the constant side of this system.

In view of the zero Lagrange multipliers \(\lambda_{32} = \lambda_{46} = 0\) and the corresponding zero differences \(y[x_{31}, x_{32}, x_{33}] = y[x_{45}, x_{46}, x_{47}] = y[x_{48}, x_{49}, x_{50}] = 0\), we deduce that the obtained best \(\ell_1\) convex fit is degenerate. Hence, the minimization of the objective function (1) subject to the active constraints that are indexed in set \(A\) leads to the same best fit, whether any of the constraints with index 32, 46 or 49 is present or not. For example, whether we impose the constraint \(y[x_{31}, x_{32}, x_{33}] \geq 0\) or not and then proceed to minimize (1) subject to the remaining constraints indexed in set \(A\), we obtain a solution which happens to satisfy the equation \(y[x_{31}, x_{32}, x_{33}] = 0\). Similarly, whether we impose \(y[x_{45}, x_{46}, x_{47}] \geq 0\) or not, the result is the same.

The best fit interpolates the points \((x_i, y_i)\), for \(i \in \{1, 6, 13, 19, 29, 30, 32, 41, 50\}\). The associated parameters \(\lambda_i\) in column 8 of Table I have the values \(\lambda_1 = -0.10, \lambda_9 = 0.60, \lambda_{13} = 1.00, \lambda_{19} = 1.00, \lambda_{28} = -0.50, \lambda_{30} = -0.50, \lambda_{32} = -0.50, \lambda_{41} = 0.00\) and \(\lambda_{50} = 0.00\), which are all in the interval \([-1, 1]\) as it follows from the optimality conditions (17). The underlined components of \(\phi\) and \(\tilde{\lambda}\) in Table I indicate the interpolation point positions and the corresponding \(\lambda_i\). The components \(\{\lambda_i : i \notin I\}\) have the values of the signs of the residuals \(y_i - \phi_i\), according to the equations contained in (13), and the components \(\{\lambda_i : i \in I\}\) are obtained from formula (16) for \(i \in I\). The corresponding 9 \times 41 coefficient matrix in (16) for \(i \in I\), say it is \(\Sigma^T\), is presented below in transposed form. The first row of numbers displays the indices in set \(I\) and the first column of numbers displays the indices in set \(A\). In words, \(\Sigma^T\) consists of the rows excluded from the formation of the matrix in Fig. 2. Further, the underlined components of \(\lambda_i, i \in I\) in column 8 of Table I, in view of (16) for \(i \in I\), are obtained by multiplying \(\Sigma^T\) with the vector whose components are the active constraint multipliers \(\lambda_j, j \in A\) (see column 7 of Table I).

<table>
<thead>
<tr>
<th>(\lambda_1)</th>
<th>(\lambda_6)</th>
<th>(\lambda_{13})</th>
<th>(\lambda_{19})</th>
<th>(\lambda_{28})</th>
<th>(\lambda_{30})</th>
<th>(\lambda_{32})</th>
<th>(\lambda_{41})</th>
<th>(\lambda_{50})</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>-0.60</td>
<td>1.00</td>
<td>1.00</td>
<td>-0.50</td>
<td>-0.50</td>
<td>-0.50</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

IV. CONCLUDING REMARKS

If measurements of function values show some gross errors and away from them the function seems to be convex, then the least sum of absolute change to the data that provides nonnegative second divided differences may be required. This problem is a highly structured constrained \(\ell_1\) approximation calculation, which can be solved by standard linear programming techniques. However, we solve it by an iterative algorithm of descent direction that employs Karush-Kuhn-Tucker like optimality conditions, despite the non-differentiable objective function. The associated parameters are important to the characterization of the solution and the development of the algorithm as well as useful for sensitivity analysis. The primary purpose of this work is to illustrate the optimality conditions by a numerical example, so as to guide the application of the method and the interpretation of the results. The data that are used are real measurements of the Gini coefficient in the U.S.A. Besides that the analysis investigates particular aspects of the evolution of this coefficient, such situations are common and similar analyses of
the general approach may arise in several fields. Moreover, the subject deserves study in the area of inferential statistics in order to bring about further possibilities for applications.

ACKNOWLEDGMENT

This work was supported in part by the University of Athens under Research Grant 11105.

REFERENCES