Force curve segmentation by piecewise polynomial approximation: mathematical formulation and complete structure of the algorithm

Charles Soussen, Junbo Duan, David Brie, Pavel Polyakov, Gregory Francius, Jérôme D.F. Duval

Force curve segmentation by piecewise polynomial approximation: mathematical formulation and complete structure of the algorithm

Charles Soussen*, Junbo Duan*, David Brie*, Pavel Polyakov†,
Grégory Francius†, Jérôme Duval‡

*Centre de Recherche en Automatique de Nancy
†Laboratory of Physical Chemistry and Microbiology for the Environment
‡Laboratory of Environment and Mineral Processing

August 3, 2010

In this technical report, we give a detailed formulation of the force curve segmentation problem based on the fitting of the curve by a piecewise polynomial. The segmentation algorithm is a discrete search in which the unknowns are the positions D_j, $j = 1, \ldots, k$ of the discontinuity points. We first formulate the problem as the minimization of a least-square cost function with respect to $\boldsymbol{D} = [D_1, \ldots, D_k] \in \mathbb{R}^k$. Then, we describe the structure of the proposed optimization algorithm including the update of the list of discontinuity positions $\boldsymbol{D}$ when their number k increases.

Let us first introduce the main notations. The force curve to be segmented is a discrete signal (z_i, F_i), $i = 1, \ldots, n$ where n stands for the number of data samples. We denote by D_j, $j = 1, \ldots, k$ the discontinuity points, sorted in the ascending order ($D_1 < D_2 < \ldots < D_k$). A discontinuity position is actually a transition between two consecutive samples z_{i-1} and z_i (*e.g.,* a jump in the curve). We choose to define a discontinuity position D_j as the z-value of the sample on the right: $D_j = z_i$ (see Fig. 1).

Setting k discontinuity positions leads to a series of contiguous intervals $[D_0, D_1)$, $[D_1, D_2)$, ..., $[D_k, D_{k+1}]$ (where $D_0 \triangleq z_1$ and $D_{k+1} \triangleq z_n$ are set to the minimal and maximal z_i values) whose union yields the whole interval $[z_1, z_n]$. Each interval $[D_{j-1}, D_j)$ is right open, the last sample which belongs to this interval being the value z_{i-1} which is preceding D_j. $D_j = z_i$ is the lower bound of the next interval $[D_j, D_{j+1})$. Finally, we denote by $\mathcal{I}_j$ the set of indices i for which z_i belongs to the j-th interval $[D_j, D_{j+1})$: $i \in \mathcal{I}_j$ is equivalent to $D_j \leqslant z_i < D_{j+1}$.

1 Piecewise polynomial approximation

We first assume that the set of discontinuity positions $\boldsymbol{D} \in \mathbb{R}^k$ is given, and we define the quality of approximation $\mathcal{E}(\boldsymbol{D})$ as the least squared error obtained with a piecewise polynomial of degree r (r is given). Then, this definition will allow us to formulate the force curve segmentation problem as a minimization problem.

1.1 Known discontinuity positions

Assume that the discontinuity points $\boldsymbol{D}$ are given. On the j-th interval $[D_j, D_{j+1})$, the data $\{(z_i, F_i),\, i \in \mathcal{I}_j\}$ are being smoothed by a polynomial of degree r.

We denote by $F_j(z; \boldsymbol{a}_j) = \sum_{l=0}^{r} a_j^l z^l$ the polynomial yielding the best approximation in the least-square sense:

$$\boldsymbol{a}_j = \underset{\boldsymbol{a} \in \mathbb{R}^{r+1}}{\arg\min} \sum_{i \in \mathcal{I}_j} \big[F_i - F_j(z_i; \boldsymbol{a})\big]^2,$$

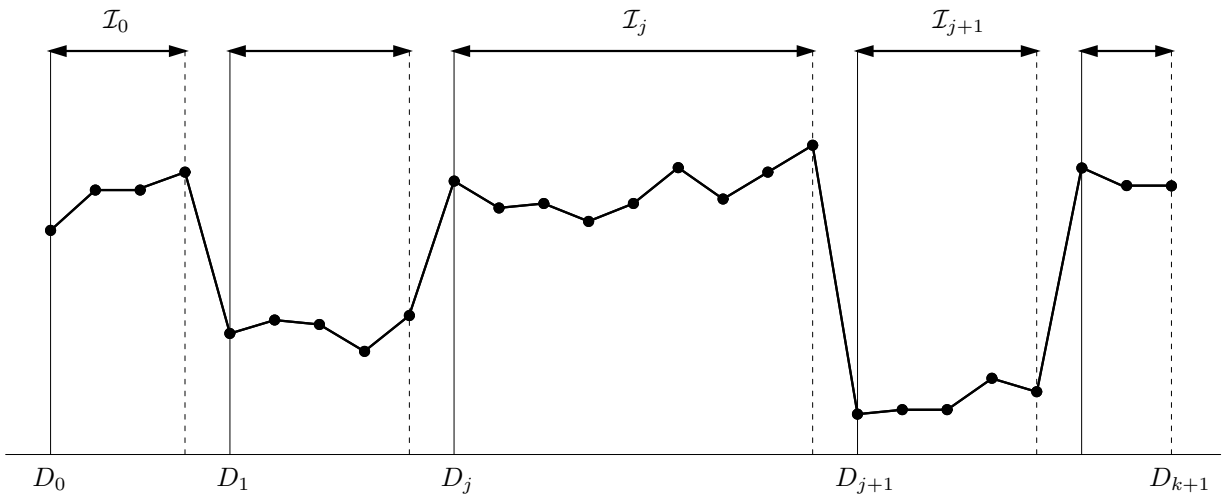

Figure 1: Definition of the discontinuity points D_j. Each "discontinuity point" is actually related to a transition between two consecutive samples z_{i-1} and z_i of the signal, and D_j is set to z_i. The plain vertical bars refer to the D_j positions (beginning of the j-th interval) while the dashed vertical bars refer to the samples z_{i-1} preceding the D_j positions (end of the $(j-1)$-th interval). When the D_j positions are given, the signal is smoothed on each interval $[D_j, D_{j+1})$ independently.

and by

$$\mathcal{E}_j = \sum_{i \in \mathcal{I}_j} \big[F_i - F_j(z_i; \boldsymbol{a}_j)\big]^2$$

the associated squared error. Finally, the piecewise-polynomial approximation of degree r is defined by the series of polynomials $F_j(z; \boldsymbol{a}_j)$ on the $k+1$ intervals $\mathcal{I}_j$, $j = 0, \ldots, k$, and the global approximation error reads

$$\mathcal{E}(\boldsymbol{D}) = \sum_{j=0}^{k} \mathcal{E}_j. \tag{1}$$

Notice that the polynomial coefficients $\boldsymbol{a}_j$ can be easily computed by linear regression:

$$\boldsymbol{a}_j = (\boldsymbol{A}_j^t \boldsymbol{A}_j)^{-1} \boldsymbol{A}_j^t \boldsymbol{F}_j, \tag{2}$$

where $\boldsymbol{A}_j$ is the Vandermonde matrix of $(r+1) \times (r+1)$ whose rows are formed of vectors $[1, z_i, \ldots, z_i^r]$ for all $i \in \mathcal{I}_j$ and $\boldsymbol{F}_j$ is the vector of size $\mathrm{Card}\,[\mathcal{I}_j] \times 1$ formed by gathering the values F_i, $i \in \mathcal{I}_j$.

1.2 Research of the discontinuity positions

The key issue is to estimate as precisely as possible the position of the discontinuity points D_k. We formulate this problem as the following optimization problem:

$$\boldsymbol{D}^\star = \underset{\boldsymbol{D}}{\arg\min}\ \mathcal{E}(\boldsymbol{D}), \tag{3}$$

where $\boldsymbol{D}^\star$ stands for the optimal solution, *i.e.,* the set of k discontinuity points yielding the least squared error. It is important to notice that (3) is actually a *discrete* optimization problem: all discontinuity positions D_k are equal to one of the n original positions z_j: $\forall j$, $D_j \in \{z_1, \ldots, z_n\}$.

Finding the optimal solution $\boldsymbol{D}^\star$ requires to perform an exhaustive search by testing all possible configurations of k discontinuity positions and computing the cost $\mathcal{E}(\boldsymbol{D})$ for each configuration. This strategy cannot be carried out in a reasonable computation time (less than 1 minute) for large signals (thousands of samples). Instead, we develop a sub-optimal optimization algorithm which allows for a fast implementation.

Inputs: data signal $\{z_i, F_i\}$, polynomial degree r, number of iterations K.
Set $k = 0$, $\boldsymbol{D} = \emptyset$.
For $k = 1$ to K,
Search for the k-th discontinuity position D_k.
Do $\boldsymbol{D} = \boldsymbol{D} \cup \{D_k\}$.
Sort $\boldsymbol{D}$ in the ascending order.
Update the list of intervals $\mathcal{I}_0, \ldots, \mathcal{I}_k$.
Update the list of coefficients $\boldsymbol{a}_j$ and the global approximation error $\mathcal{E}(\boldsymbol{D})$ according to (1).
End For.
Outputs: K discontinuity points $\boldsymbol{D} = \{D_j,\ j = 1, \ldots, K\}$, $K+1$ polynomials $F(z; \boldsymbol{a}_j)$, $j = 0, \ldots, K$.

Table 1: Segmentation algorithm (structure). The critical task is the research of the next discontinuity position. It is also the most time consuming.

For all i such that $z_i \notin \boldsymbol{D}$,
Test $D_k = z_i$: compute $\mathcal{E}(\boldsymbol{D} \cup \{z_i\})$.
End For.
Set D_k according to (4).

Table 2: Research of the next discontinuity position.

2 Proposed segmentation algorithm

2.1 Principle

The principle of the algorithm is to iteratively increase by one element the list of discontinuity positions. At the beginning, this list is empty. Then, the algorithm sequentially includes a new discontinuity point into the list, and then refines the piecewise-smooth approximation whenever the list is modified by one element. The structure of the algorithm is presented in Table 1. For simplicity reasons, the stopping condition consists of a maximal number of K iterations (the use of another stopping condition is discussed in paragraph 2.3 below). As output, the algorithm yields a sequence of K discontinuity positions $D_1, \ldots, D_K$ and the polynomial coefficients $\boldsymbol{a}_j$ on each interval (polynomial approximation of the force curve for $z_i \in [D_j, D_{j+1})$).

In this algorithm, the key issue is to select the next discontinuity point when the list of discontinuity points has to be increased by one element. We detail this research in the following paragraph.

2.2 Search for the next discontinuity position

At a given iteration k, $k-1$ discontinuities have already been included ($\boldsymbol{D} \in \mathbb{R}^{k-1}$). The next position to be included into the list is defined by:

$$D_k = z_{i_k}, \text{ where } i_k = \arg\min_i \mathcal{E}(\boldsymbol{D} \cup \{z_i\}). \tag{4}$$

In practice, it is computed in two steps:

1. first, all possible inclusions $\boldsymbol{D} \cup \{z_i\}$ are tested (for all[1] positions z_i which have not already been included in $\boldsymbol{D}$). For each trial, the approximation error $\mathcal{E}(\boldsymbol{D} \cup \{z_i\})$ is computed;
2. then, the sample z_{i_k} yielding the least error is selected ($D_k = z_{i_k}$).

[1] Actually, the positions z_i which are too close to an existing discontinuity position D_j must not be considered: if an interval $\mathcal{I}_j$ contains less than $r+1$ samples, fitting a polynomial of degree r is not possible in this interval since the linear system (2) is singular.

If $k = K$ or if $\mathcal{E}(\boldsymbol{D})/n \leqslant \mathcal{E}_{\text{MIN}}$,
Exit from the segmentation loop.
End If.

Table 3: Alternative stopping condition relying on a threshold on the mean approximation error. K is an upper bound on the possible number of desired discontinuities. We set the threshold $\mathcal{E}_{\text{MIN}}$ to a ratio of the empirical variance of the noise (v_n), *e.g.,* $\mathcal{E}_{\text{MIN}} = v_n$ or $1.5\, v_n$. When processing a force curve (approach or retraction), v_n can be easily estimated since there are flat regions at the end of the curve, with only noise samples.

Table 2 summarizes this inclusion process. Once an inclusion is performed, the list of discontinuity points is updated ($\boldsymbol{D}$ is sorted in the ascending order) and the list of intervals $\mathcal{I}_j$ is updated as well. Actually, all intervals are unchanged except the interval containing D_k which is being bisected.

When computing the new position D_k, all errors $\mathcal{E}(\boldsymbol{D} \cup \{z_i\})$ have to be computed for all possible inclusions (inclusion tests). This is clearly the main part of the cost of an iteration of the segmentation algorithm. However, one can easily implement a fast algorithm by taking into account the fact that for each insertion trial $\boldsymbol{D} \cup \{z_i\}$, the piecewise polynomial fitting is unchanged except in the two sub-intervals containing z_i. Thus, the cost of a discontinuity insertion trial amounts to two polynomial fitting operations. Similarly, the approximation errors $\mathcal{E}_j$ are unchanged except for the two (new) sub-intervals.

2.3 Practical settings

Let us discuss the practical settings, namely the setting of the desired number of discontinuity points. A first possibility is to set a maximal number (K discontinuity points are being inserted: see Table 1). Setting K is done manually by the user, *e.g.,* after viewing the signal and counting the number of discontinuity points. An alternative and more automated stopping condition consists in setting a threshold on the mean approximation error $\mathcal{E}(\boldsymbol{D})/n$ (*i.e.,* the average of the squared error $[F_i - F_j(z_i; \boldsymbol{a}_j)]^2$ for all samples i). If the mean approximation error is lower than $\mathcal{E}_{\text{MIN}}$, then the algorithm terminates.

This stopping condition is illustrated on Table 3. When processing experimental force curves, we set the threshold $\mathcal{E}_{\text{MIN}}$ to a ratio of the empirical variance of the noise (v_n), *e.g.,* $\mathcal{E}_{\text{MIN}} = v_n$ or $1.5\, v_n$. In both approach and retraction cases, v_n can be easily estimated since there are flat regions at the end of the curve, with only noise samples.