Supplementary Material

Theoretical study on analyzability of modified convex regression approach for radical reaction

Tomomi Shimazaki and Masanori Tachikawa

Quantum Chemistry Division, Yokohama City University, Seto 22-2, Kanazawa-Ku,
Yokohama 236-0027, Kanagawa, Japan.
tshima@yokohama-cu.ac.jp

S1. K-near method

The convex clustering method probabilistically assigns a data point $\mathbf{x}_k$ to multiple classes. Meanwhile, the K-means clustering method assigns each data point to a single class, which is represented by the mean (average) point obtained from $\bar{\mathbf{x}}_i = (1/N_i)\sum_{\mathbf{x}_k \in \Pi_i} \mathbf{x}_k$, where N_i represents the number of points in class Π_i. In this study, we discuss a slight modification of the K-means clustering method by selecting representative points directly from the training dataset, rather than the computed mean. To achieve this, we simply choose a data point that is closest to $\bar{\mathbf{x}}_i$. Typically, the nearest point is selected as the representative point, provided there are no duplicates. In cases where multiple classes would share the same representative point, we preferentially selected the class closest to the point. Based on this procedure, the second and third closest points may be assigned as representative points for some classes. We refer to this modified approach as the K-near method.

The K-near method follows a self-consistent process, similar to the modified convex clustering method. The process is summarized in **Figure S1**. Initially, the representative points of classes are randomly selected from the training dataset X. The number of clusters, C, is treated as a hyperparameter. Notably, the final clustering result can depend on the initial choice of representative points. Each data point in X is then assigned to the class of its nearest representative point. After this assignment, class means (averages) are recalculated, and new representative points are selected from nearby data points. These steps are repeated iteratively until the clustering results converge. The K-near method is a hard assignment clustering, meaning each data point is assigned to exactly one class.

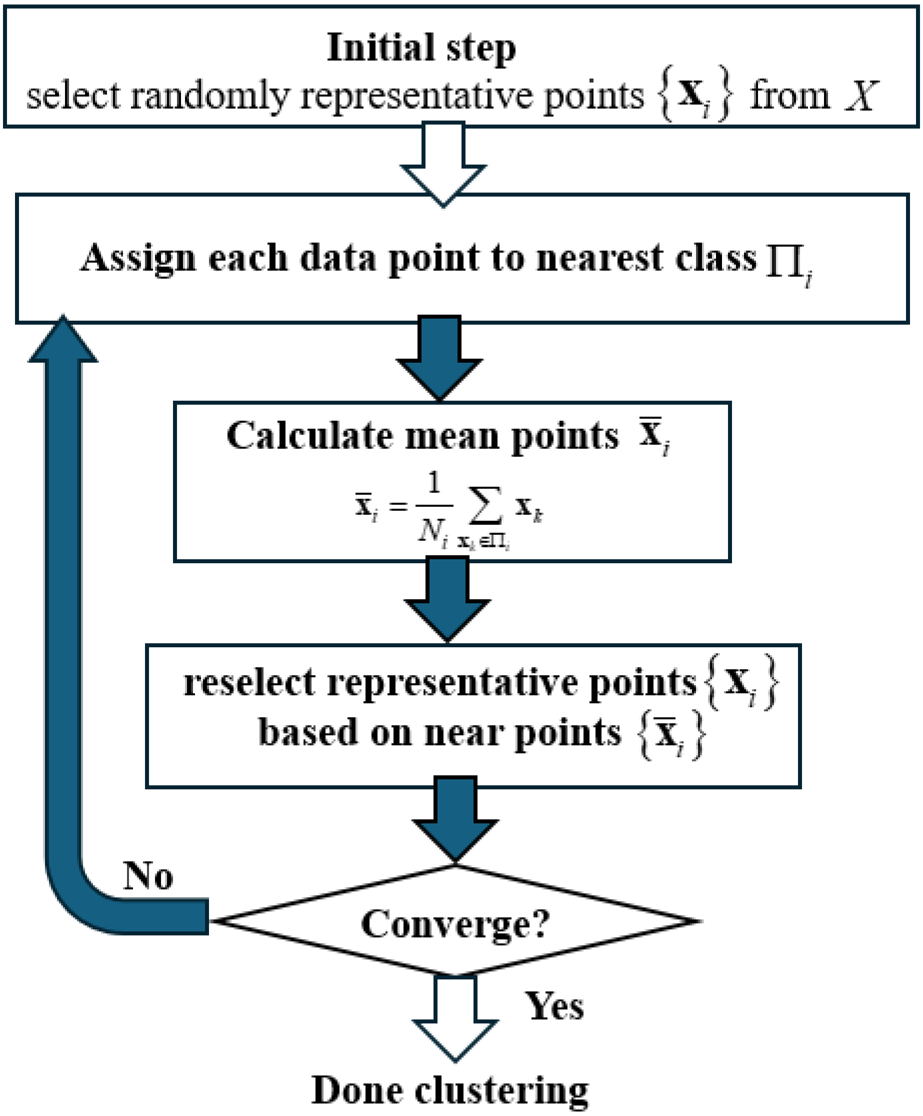
Initial step
select randomly representative points $\{\mathbf{x}_i\}$ from X
Assign each data point to nearest class Π_i
Calculate mean points $\bar{\mathbf{x}}_i$
$\bar{\mathbf{x}}_i = \frac{1}{N_i} \sum_{\mathbf{x}_k \in \Pi_i} \mathbf{x}_k$
reselect representative points $\{\mathbf{x}_i\}$
based on near points $\{\bar{\mathbf{x}}_i\}$
No
Converge?
Yes
Done clustering

Figure S1. Computational flow of the K-near method.

S2. Radical reaction dataset

Table S1. Reaction energy barrier ΔE_{TS} [kcal/mol], energy difference between reactant and product ΔE_{RP} [kcal/mol] calculated by DFT, and dummy variable DP(*m*) representing ACR or MA for the radical reaction $X^{\cdot} + Y \rightarrow XY^{\cdot}$ (see also **Figure 3a** in the main manuscript).

X	Y	DP(X)	DP(Y)	ΔE_{RP}	ΔE_{TS}
1	1	0	0	−17.30	4.94
1	2	0	1	−19.46	4.26
1	3	0	0	−17.00	5.19
1	4	0	1	−18.71	4.56
1	5	0	0	−16.54	5.58
1	6	0	1	−18.48	4.64
1	7	0	0	−17.31	5.16
1	8	0	1	−19.69	4.25
1	9	0	0	−16.38	5.58
1	10	0	1	−18.56	4.35
1	1	1	0	−12.60	6.17
2	2	1	1	−13.11	6.76
2	3	1	0	−10.84	7.94
2	4	1	1	−12.50	7.19
2	5	1	0	−11.55	7.15
2	6	1	1	−12.54	7.01

2	7	1	0	−12.53	6.36
2	8	1	1	−13.57	6.45
2	9	1	0	−11.70	6.95
2	10	1	1	−13.03	6.46
3	1	0	0	−17.55	4.77
3	2	0	1	−19.28	4.61
3	3	0	0	−17.22	5.08
3	4	0	1	−18.64	5.01
3	5	0	0	−16.74	5.51
3	6	0	1	−18.04	5.34
3	7	0	0	−17.49	5.05
3	8	0	1	−19.14	4.99
3	9	0	0	−16.65	5.42
3	10	0	1	−18.74	4.40
4	1	1	0	−13.11	5.80
4	2	1	1	−13.79	6.29
4	3	1	0	−10.66	8.28
4	4	1	1	−12.52	7.38
4	5	1	0	−11.80	7.08
4	6	1	1	−12.23	7.57
4	7	1	0	−12.86	6.17
4	8	1	1	−13.68	6.55
4	9	1	0	−11.98	6.85
4	10	1	1	−13.13	6.60

5	1	0	0	−17.74	4.57
5	2	0	1	−19.98	3.91
5	3	0	0	−17.38	4.90
5	4	0	1	−18.84	4.81
5	5	0	0	−16.26	5.36
5	6	0	1	−18.67	4.88
5	7	0	0	−18.17	4.88
5	8	0	1	−20.41	4.25
5	9	0	0	−16.07	5.36
5	10	0	1	−18.09	4.37
6	1	1	0	−13.45	5.46
6	2	1	1	−14.12	5.95
6	3	1	0	−13.14	5.80
6	4	1	1	−13.16	6.74
6	5	1	0	−12.18	6.48
6	6	1	1	−12.54	7.25
6	7	1	0	−10.64	8.85
6	8	1	1	−12.57	8.07
6	9	1	0	−12.14	6.45
6	10	1	1	−13.33	6.12
7	1	0	0	−15.08	6.14
7	2	0	1	−17.62	4.38
7	3	0	0	−14.95	6.26
7	4	0	1	−17.28	4.40

7	5	0	0	−13.77	6.52
7	6	0	1	−14.74	4.48
7	7	0	0	−14.82	6.38
7	8	0	1	−15.68	4.37
7	9	0	0	−13.81	6.27
7	10	0	1	−13.74	4.92
8	1	1	0	−9.57	8.54
8	2	1	1	−10.34	8.89
8	3	1	0	−9.25	8.89
8	4	1	1	−9.87	9.20
8	5	1	0	−9.21	8.78
8	6	1	1	−9.42	9.18
8	7	1	0	−8.44	10.06
8	8	1	1	−11.09	8.99
8	9	1	0	−9.76	8.17
8	10	1	1	−10.84	7.90
9	1	0	0	−17.82	4.56
9	2	0	1	−20.10	3.90
9	3	0	0	−17.49	4.89
9	4	0	1	−18.77	4.86
9	5	0	0	−16.56	5.35
9	6	0	1	−19.16	4.91
9	7	0	0	−18.53	4.89
9	8	0	1	−20.86	4.26

9	9	0	0	−15.91	5.77
9	10	0	1	−18.39	4.40
10	1	1	0	−13.74	5.36
10	2	1	1	−13.86	6.46
10	3	1	0	−13.44	5.71
10	4	1	1	−13.32	6.83
10	5	1	0	−12.36	6.45
10	6	1	1	−12.70	7.58
10	7	1	0	−10.76	9.02
10	8	1	1	−13.03	8.01
10	9	1	0	−12.40	6.36
10	10	1	1	−13.65	6.03