An equivalence between log-sum-exp approximation and entropy regularization in K-means clustering

井上，光平
九州大学大学院芸術工学研究院 ： 准教授

原，健二
九州大学大学院芸術工学研究院 ： 教授

出版情報：Nonlinear Theory and Its Applications, IEICE. 11 (4), pp.446-453, 2020-10-01.
Nonlinear Theory and Its Applications, IEICE
バージョン：

NOLTA, IEICE

Paper

An equivalence between log-sum-exp approximation and entropy regularization in K-means clustering

Kohei Inoue [1a)] *and Kenji Hara* [1]

[1] *Faculty of Design, Kyushu University*
4–9–1 Shiobaru, Minami-ku, Fukuoka 815-8540, Japan

[a)] *k-inoue@design.kyushu-u.ac.jp*

Received January 16, 2020; Revised April 15, 2020; Published October 1, 2020

Abstract: In this paper, we show an equivalence between log-sum-exp approximation and entropy regularization in K-means clustering, which is a well-known algorithm for partitional clustering. We derive an identical equation for updating centroids of clusters from the two formulations. Additionally, we derive an alternative equation suitable for another formulation of entropy regularization, maximum entropy method. We also show experimental results which support the theoretical results.

Key Words: K-means clustering, log-sum-exp approximation, entropy regularization, maximum entropy method

1. Introduction

Clustering is the task of grouping a set of objects in such a way that objects in the same group or cluster are more similar to each other than to those in other clusters [1]. In centroid-based clustering, each cluster is represented by a single mean vector or a centroid. K-means clustering [2] is one of the most popular algorithms in centroid-based clustering, and is categorized into hard clustering.

Dunn [3] developed a fuzzy version of K-means, fuzzy c-means clustering, and Bezdek [4] improved it [5]. Miyamoto and Mukaidono [6] proposed an entropy regularization of the crisp K-means clustering to derive a fuzzy c-means clustering, and showed the equivalence to a maximum entropy approach proposed by Li and Mukaidono [7, 8], which is briefly reviewed in this paper.

In this paper, we show an equivalence between the entropy regularization of K-means clustering by Miyamoto and Mukaidono [6] and a log-sum-exp approximation [9] of K-means clustering. Starting from the two different formulations, we derive an identical equation which is used for updating centroids of clusters. As a result, we conclude that the entropy regularization, the maximum entropy approach and the log-sum-exp approximation are equivalent to each other in K-means clustering. Additionally, we summarize the equivalence between entropy regularization [6] and maximum entropy method [7, 8] by using their Lagrange functions, and derive an equation suitable for the latter. We also show experimental results on a clustering benchmark dataset, which support our theoretical results.

The rest of this paper is organized as follows. Section 2 summarizes log-sum-exp approximation, entropy regularization and maximum entropy method in the context of K-means clustering, and shows the equiva-

Nonlinear Theory and Its Applications, IEICE, vol. 11, no. 4, pp. 446–453 DOI: 10.1587/nolta.11.446

lence of them. Section 3 shows experimental results, where an advantage of the log-sum-exp approximation and entropy regularization over the maximum entropy approach is demonstrated by showing a nonmonotonic behavior of total entropy. Finally, Section 4 concludes this paper.

2. *K*-means clustering

Given a set of points $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ in d-dimensional Euclidean space, K-means clustering aims to partition the n points in X into K sets $S_1, S_2, \cdots, S_K$ so as to minimize the objective function

$$J(\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_K) = \sum_{i=1}^{n} \min_{k \in \{1,2,\ldots,K\}} \|\boldsymbol{x}_i - \boldsymbol{c}_k\|^2 \tag{1}$$

with respect to K centroids $\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_K$ corresponding to $S_1, S_2, \cdots, S_K$, respectively.

In this section, we first summarize two methods for solving the above problem of K-means clustering: log-sum-exp approximation and entropy regularization. The latter has another expression called maximum entropy method, which is also summarized briefly. Then we show the equivalence between log-sum-exp approximation and entropy regularization.

2.1 Log-sum-exp approximation

The log-sum-exp function is a differentiable approximation of the max function [9] as follows:

$$f(x_1, x_2, \ldots, x_n) = \log\left(\sum_{i=1}^{n} \exp(x_i)\right) \tag{2}$$

$$\approx \max\{x_1, x_2, \ldots, x_n\}, \tag{3}$$

which finds the maximum value in $\{x_1, x_2, \ldots, x_n\}$. Applying the log-sum-exp approximation to the objective function in Eq. (1), we have

$$J(\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_K) = -T \sum_{i=1}^{n} \max_{k \in \{1,2,\ldots,K\}} \left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{c}_k\|^2}{T}\right) \tag{4}$$

$$\approx -T \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{c}_k\|^2}{T}\right)\right) \tag{5}$$

$$= \tilde{J}(\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_K), \tag{6}$$

where T denotes a positive parameter. The necessary condition for optimality of the maximization of Eq. (6) is given by

$$\frac{\partial \tilde{J}}{\partial \boldsymbol{c}_k} = -2 \sum_{i=1}^{n} \frac{\exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{c}_k\|^2}{T}\right)}{\sum_{k'=1}^{K} \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{c}_{k'}\|^2}{T}\right)} (\boldsymbol{x}_i - \boldsymbol{c}_k) = \boldsymbol{0}, \tag{7}$$

where $\boldsymbol{0}$ denotes a d-dimensional zero vector having all components equal to zero. From Eq. (7), we have

$$\boldsymbol{c}_k = \frac{\displaystyle\sum_{i=1}^{n} \frac{\exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{c}_k\|^2}{T}\right)}{\sum_{k'=1}^{K} \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{c}_{k'}\|^2}{T}\right)} \boldsymbol{x}_i}{\displaystyle\sum_{i=1}^{n} \frac{\exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{c}_k\|^2}{T}\right)}{\sum_{k'=1}^{K} \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{c}_{k'}\|^2}{T}\right)}}. \tag{8}$$

Each centroid $\boldsymbol{c}_k$ is updated by Eq. (8) until all centroids converge.

2.2 Entropy regularization

The objective function J in Eq. (1) has another expression as follows:

$$J\left(\{\boldsymbol{c}_k\}, \{u_{ik}\}\right) = \sum_{i=1}^{n} \sum_{k=1}^{K} u_{ik} \left\|\boldsymbol{x}_i - \boldsymbol{c}_k\right\|^2, \tag{9}$$

where u_{ik} denotes a nonnegative variable indicating the membership of the ith point in the kth cluster. Using the expression in Eq. (9), the entropy regularization of K-means clustering is formulated as follows [6]:

$$\min_{\{\boldsymbol{c}_k\},\{u_{ik}\}} J\left(\{\boldsymbol{c}_k\}, \{u_{ik}\}\right) + T \sum_{i=1}^{n} \sum_{k=1}^{K} u_{ik} \log u_{ik} \tag{10}$$

$$\text{subj.to} \sum_{k=1}^{K} u_{ik} = 1, \quad \text{for} \quad i = 1, 2, \ldots, n, \tag{11}$$

where the constraint condition enforces that u_{ik} is a probability that the ith point belongs to the kth cluster. The Lagrange function for this constrained optimization problem is given by

$$L = J\left(\{\boldsymbol{c}_k\}, \{u_{ik}\}\right) + T \sum_{i=1}^{n} \sum_{k=1}^{K} u_{ik} \log u_{ik} + \sum_{i=1}^{n} \lambda_i \left(\sum_{k=1}^{K} u_{ik} - 1 \right), \tag{12}$$

where λ_i for $i = 1, 2, \ldots, n$ denote the Lagrange multipliers. Then we have the following necessary conditions for optimality:

$$\frac{\partial L}{\partial \boldsymbol{c}_k} = -2 \sum_{i=1}^{n} u_{ik} \left(\boldsymbol{x}_i - \boldsymbol{c}_k\right) = \boldsymbol{0}, \tag{13}$$

$$\frac{\partial L}{\partial u_{ik}} = \left\|\boldsymbol{x}_i - \boldsymbol{c}_k\right\|^2 + T \left(\log u_{ik} + 1\right) + \lambda_i = 0, \tag{14}$$

$$\frac{\partial L}{\partial \lambda_i} = \sum_{k=1}^{K} u_{ik} - 1 = 0. \tag{15}$$

Solving Eq. (13) for $\boldsymbol{c}_k$, we have

$$\boldsymbol{c}_k = \frac{\displaystyle\sum_{i=1}^{n} u_{ik} \boldsymbol{x}_i}{\displaystyle\sum_{i=1}^{n} u_{ik}}. \tag{16}$$

Solving Eq. (14) for u_{ik}, we have

$$u_{ik} = \exp\left(-\frac{\left\|\boldsymbol{x}_i - \boldsymbol{c}_k\right\|^2}{T} - \frac{\lambda_i}{T} - 1 \right). \tag{17}$$

Substituting this for u_{ik} in Eq. (15), we have

$$\exp\left(-\frac{\lambda_i}{T} - 1 \right) = \frac{1}{\displaystyle\sum_{k=1}^{K} \exp\left(-\frac{\left\|\boldsymbol{x}_i - \boldsymbol{c}_k\right\|^2}{T} \right)}. \tag{18}$$

Substituting this into Eq. (17), we have the final form of u_{ik} as follows:

$$u_{ik} = \frac{\exp\left(-\dfrac{\left\|\boldsymbol{x}_i - \boldsymbol{c}_k\right\|^2}{T} \right)}{\displaystyle\sum_{k'=1}^{K} \exp\left(-\frac{\left\|\boldsymbol{x}_i - \boldsymbol{c}_{k'}\right\|^2}{T} \right)}. \tag{19}$$

In the algorithm for this problem, $\boldsymbol{c}_k$ and u_{ik} are alternately updated by Eqs. (16) and (19), respectively, until they converge.

2.3 Maximum entropy method

Miyamoto and Mukaidono [6] showed that the entropy regularization is equivalent to the maximum entropy method [7, 8] formulated as follows:

$$\max_{\{u_{ik}\}} -\sum_{i=1}^{n}\sum_{k=1}^{K} u_{ik}\log u_{ik} \tag{20}$$

$$\text{subj.to}\sum_{k=1}^{K} u_{ik} = 1, \quad J\left(\{\boldsymbol{c}_k\},\{u_{ik}\}\right) = J_0, \tag{21}$$

where J_0 is a parameter, and pointed out the difficulty of determining J_0.

In fact, let L^{Ent} be the Lagrange function of the above constrained maximization problem as follows:

$$L^{\text{Ent}} = -\sum_{i=1}^{n}\sum_{k=1}^{K} u_{ik}\log u_{ik} + \sum_{i=1}^{n}\lambda_i^{\text{Ent}}\left(\sum_{k=1}^{K} u_{ik} - 1\right) + \mu\left(J\left(\{\boldsymbol{c}_k\},\{u_{ik}\}\right) - J_0\right), \tag{22}$$

where λ_i^{Ent} and μ denote the Lagrange multipliers. Then we have a relationship between L^{Ent} and L in Eq. (12) as follows:

$$L^{\text{Ent}} = \mu\left(L - J_0\right), \tag{23}$$

where it is assumed that $\mu T = -1$ and $\lambda_i^{\text{Ent}} = \mu\lambda_i$. Therefore, we arrive at the same necessary conditions for optimality in Eqs. (13)–(15) from the above maximum entropy formulation, and therefore, get the same equations as Eqs. (16) and (19), that concludes the equivalence between maximum entropy method and entropy regularization.

In Section 3, we will demonstrate that the entropy in Eq. (20) does not necessarily increase with the progress of the procedure.

2.4 The equivalence

The log-sum-exp approximation of K-means clustering described in Section 2.1 derives an equation in Eq. (8) for updating centroid $\boldsymbol{c}_k$. On the other hand, the entropy regularization described in Section 2.2 derives two equations in Eqs. (16) and (19) for updating centroid $\boldsymbol{c}_k$ and membership u_{ik}, respectively. Substitution of Eq. (19) into Eq. (16) gives Eq. (8). This proves the equivalence of the two methods.

In Section 2.3, we briefly showed the equivalence between maximum entropy method and entropy regularization. A drawback of maximum entropy method may be that $\{\boldsymbol{c}_k\}$ is not included in the objective function or entropy in Eq. (20). Therefore, the change in $\{\boldsymbol{c}_k\}$ will not affect the value of the objective function, i.e., updating c_k by Eq. (16) will not increase the entropy in Eq. (20). Substituting $\boldsymbol{c}_k$ in Eq. (16) for Eq. (19), we have an alternative equation to Eq. (8) as follows:

$$u_{ik} = \frac{\exp\left(-\dfrac{\left\|\boldsymbol{x}_i - \frac{\sum_{i'=1}^{n} u_{i'k}\boldsymbol{x}_{i'}}{\sum_{i'=1}^{n} u_{i'k}}\right\|^2}{T}\right)}{\sum_{k'=1}^{K}\exp\left(-\dfrac{\left\|\boldsymbol{x}_i - \frac{\sum_{i'=1}^{n} u_{i'k'}\boldsymbol{x}_{i'}}{\sum_{i'=1}^{n} u_{i'k'}}\right\|^2}{T}\right)}, \tag{24}$$

which does not include $\{\boldsymbol{c}_k\}$ explicitly. Therefore, when we use this equation instead of Eq. (8), we should initialize $\{u_{ik}\}$. We will show an example of this situation in the next section.

3. Experimental results

In this section, we show experimental results for confirming the above theoretical results numerically. We used a synthetic 2-dimensional dataset, S1, with $n = 5000$ points and $K = 15$ Gaussian clusters with different degree of cluster overlap, which is publicly available at the website "Clustering basic benchmark" [10]. Figure 1 shows the data with blue points and 15 initial centroids with red points.

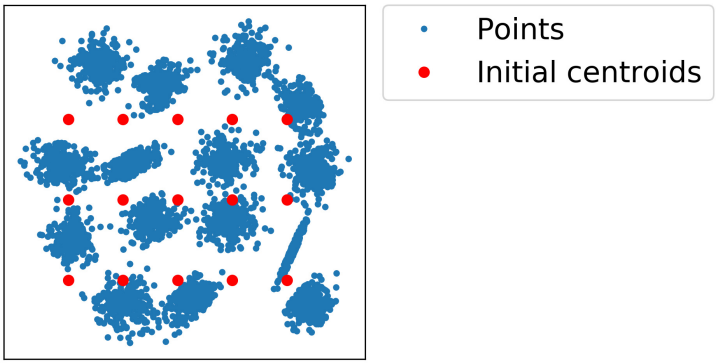

Fig. 1. Two-dimensional data and initial centroids.

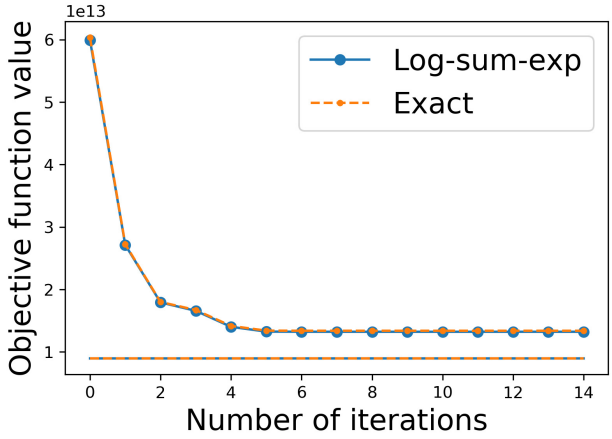

Fig. 2. Objective function values with the number of iterations started from the initial placement of centroids shown with red points in Fig. 1.

Figure 2 shows the transition of the objective function values, where the vertical and horizontal axes denote the objective function value and the number of iterations of the procedure for updating centroids, respectively. In this figure, the solid blue line with points denotes the value of the objective function of the log-sum-exp approximation $\tilde{J}$ in Eq. (6), and the broken orange line with points denotes that of entropy regularization J in Eq. (1). In both methods, we set the parameter T as $T = 10^9$. This figure shows that the log-sum-exp approximation gives the similar objective function values $\tilde{J}$ to the original objective function values J for K-means clustering.

Figure 3 shows the obtained centroids with red points after 14 iterations and the ground truth with yellow points which are globally optimal centroids. As shown in the above section, both log-sum-exp approximation and entropy regularization give the same result as each other in this example. Note that the obtained red points do not coincide with the yellow points exactly. The objective function values of the log-sum-exp approximation and the entropy regularization for the ground truth are shown in Fig. 2 with solid blue and broken yellow lines without points, which are lower than the corresponding lines of obtained solutions. That is, the obtained solutions are locally optimal ones.

Figure 4 shows the transition of total entropy in Eq. (20), where the vertical and horizontal axes denote the total entropy and the number of iterations of the procedure for updating centroids, respectively. Although the maximum entropy method [7, 8] is intended to maximize the total entropy as formulated in Eq. (20), the derived procedure fails to increase it monotonically as shown in Fig. 4.

Next, we show an example of another situation in terms of the initialization of K-means clustering procedure, where memberships are initialized and updated by Eq. (24). We assigned each point listed in a text file for the same data as above from top to bottom to almost equally-sized initial clusters, in which the first cluster has 338 points selected from the top 338 rows in the text file, and the remaining 14 clusters have 333 points each (the total number of points is 5000), e.g., if the ith point is assigned to the 1st cluster, then $u_{i,1} = 1$ and $u_{ik} = 0$ for

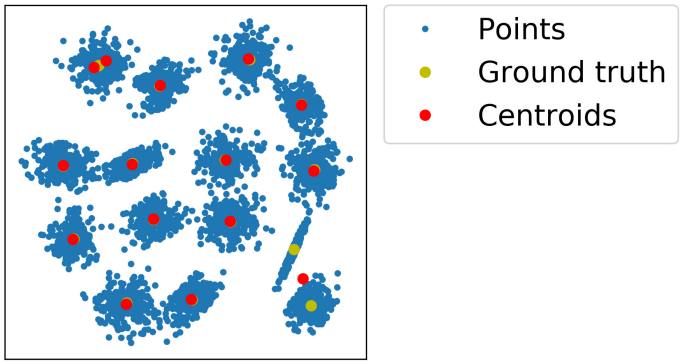

Fig. 3. Ground truth and obtained centroids after 14 iterations started from the initial centroids shown in Fig. 1.

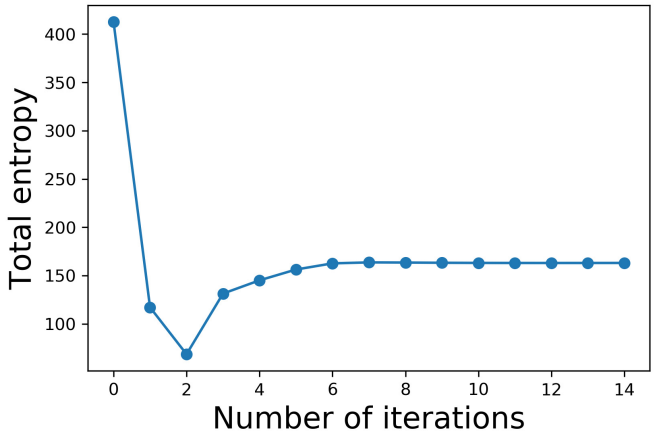

Fig. 4. Total entropy with the number of iterations started from the initial centroids shown in Fig. 1.

$k \neq 1$ which satisfy the constraint $\sum_{k=1}^{K} u_{ik} = 1$.

Figure 5 shows the initial assignment of points to 15 clusters, where different colors and markers indicate different clusters.

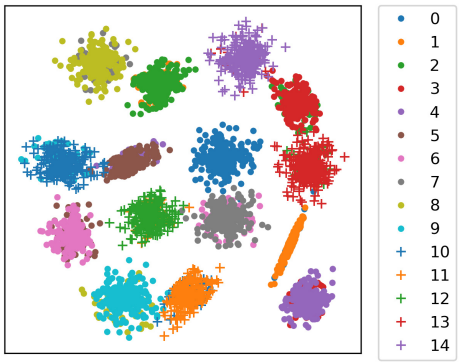

Fig. 5. Classification based on initial memberships.

Figure 6 shows the transition of the objective values similarly to Fig. 2. In this example, the objective function values reached the minimum in a smaller number of iterations than that of formaer example in Fig. 2.

Figure 7 shows the obtained centroids with red points after 9 iterations, which coincide with the ground truth indicated with yellow points, which are hidden behind the red points.

Figure 8 shows the transition of total entropy in Eq. (20) similarly to Fig. 4, where we observe the nonmonotonic behavior of total entropy again. Thus, the increase and decrease of the entropy are in conflict with the

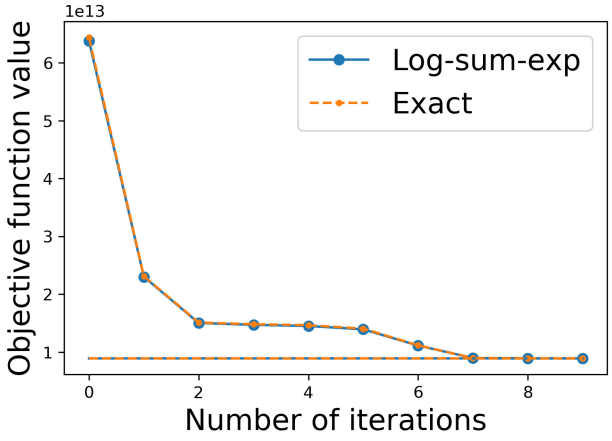

Fig. 6. Objective function values with the number of iterations started from the initial assignment of 5000 points to 15 clusters shown with different colors and markers in Fig. 5.

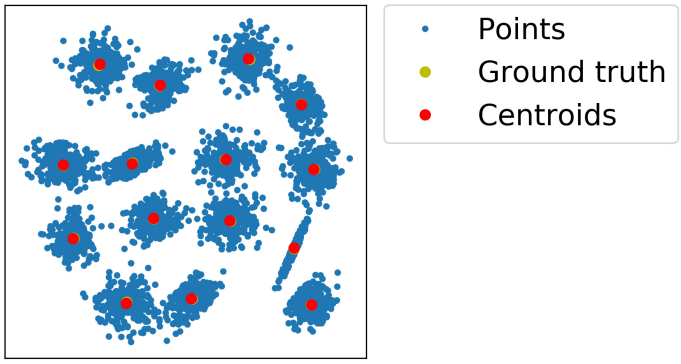

Fig. 7. Ground truth and obtained centroids after 9 iterations started from the initial assignment of 5000 points to 15 clusters shown in Fig. 5.

objective of maximum entropy method. These examples suggest the superiority of entropy regularization and log-sum-exp approximation to maximum entropy method for K-means clustering.

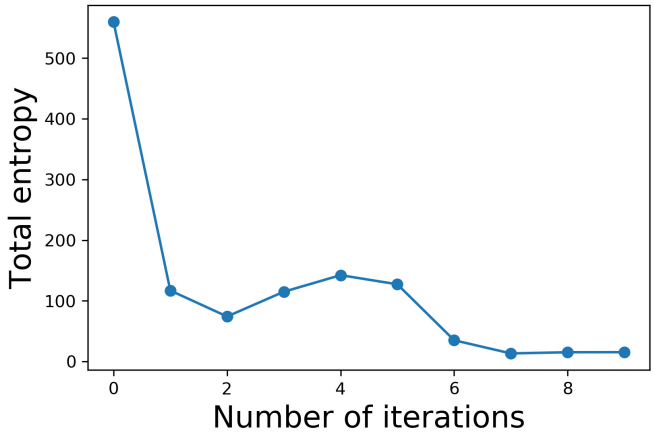

Fig. 8. Total entropy with the number of iterations started from the initial assignment of 5000 points to 15 clusters shown in Fig. 5.

4. Conclusion

In this paper, we showed an equivalence between the log-sum-exp approximation and the entropy regularization in K-means clustering by deriving the same equation for updating centroids from the two formulations, and

demonstrated that the centroids converged to the same local optimum by the two methods using a synthetic 2-dimensional dataset. Furthermore, we also demonstrated that the derived procedure does not necessarily increase total entropy monotonically in spite of the equivalence between the entropy regularization method and maximum entropy method which is formulated as a constrained maximization problem of the entropy.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP16H03019.

References

[1] Cluster analysis. In *Wikipedia: The Free Encyclopedia*. https://en.wikipedia.org/wiki/Cluster_analysis This page was last edited on 20 April 2019, at 21:19 (UTC).

[2] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob.*, vol. 1, pp. 281–297, 1967.

[3] J.C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57, 1973.

[4] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Springer, 1981.

[5] Fuzzy clustering. In *Wikipedia: The Free Encyclopedia*. https://en.wikipedia.org/wiki/Fuzzy_clustering This page was last edited on 4 April 2019, at 06:21 (UTC).

[6] S. Miyamoto and M. Mukaidono, "Fuzzy c-means as a regularization and maximum entropy approach," *The proceedings of the seventh International Fuzzy Systems Association World Congress (IFSA'97)*, vol. 2, pp. 86–92, 1997.

[7] R.-P. Li and M. Mukaidono, "Gaussian clustering and its application to rock classification," *Proc. of Eleventh Fuzzy System Symposium*, Japan Society of Fuzzy Theory and Systems, pp. 697–698, 1995.

[8] R.-P. Li and M. Mukaidono, "A maximum entropy approach to fuzzy clustering," *Proc. of the 4th IEEE Intern. Conf. on Fuzzy Systems (FUZZ-IEEE/IFES'95)*, Yokohama, Japan, pp. 2227–2232, 1995.

[9] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[10] P. Fränti and S. Sieranoja, "K-means properties on six clustering benchmark datasets," *Applied Intelligence*, vol. 48, no. 12, pp. 4743–4759, 2018. http://cs.joensuu.fi/sipu/datasets/