Sparse Gaussian Mixture Model Clustering via Simultaneous Perturbation Stochastic Approximation

Andrei Boiarov * **Oleg Granichin** *

* *Faculty of Mathematics and Mechanics, Saint Petersburg State University, 7-9, Universitetskaya Nab., St. Petersburg, 199034, Russia, Institute of Problems in Mechanical Engineering, Russian Academy of Sciences. (e-mail: a.boiarov@spbu.ru, o.granichin@spbu.ru)*

Abstract: In this paper the problem of a multidimensional optimization in unsupervised learning and clustering is studied under significant uncertainties in the data model and measurements of penalty functions. We propose a modified version of SPSA-based algorithm which maintains stability under conditions such as a sparse Gaussian mixture model. This data model is important because it can be effectively used to evaluate the noise model in many practical systems. The proposed algorithm is robust to external disturbances and is able to process data sequentially, "on the fly". In this paper provides a study of this algorithm and its mathematical justification. The behavior of the algorithm is illustrated by examples of its use for clustering in various difficult conditions.

Keywords: Optimization, Stochastic approximation, Machine learning, Learning algorithms, Gaussian distributions.

1. INTRODUCTION

Standard machine learning algorithms work successfully when they are trained on a large amount of labeled data. There are relatively few such data sets available for a relatively small range of tasks (recognition and localization of objects, landmarks (Boiarov and Tyantov (2019)) recognition of faces, emotions and postures of a person, parallel cases for translating texts between languages, etc.) Goodfellow et al. (2016). Each new task requires the collection of a new dataset, which is a rather time-consuming task and often requires the efforts of a large number of people. However, in the world there is (mainly due to the development of the Internet, social networks and smartphones) a huge amount of unlabeled data. One of the most important tasks associated with this type of data is the unsupervised learning problem and one of its special cases is the clustering problem. The lack of a pre-known structure and markup of data is a source of uncertainties. To work in such conditions, it is necessary to develop new approaches.

For the successful operation of the main machine learning algorithms, a clear data model, the ability to calculate the gradient for the optimized loss function (quality functional), and as many training data as possible close to normally distributed are needed Polyak (1987). However, under real conditions, these requirements are often not fulfilled. For example, data can be inherently sparse (like Gaussian mixture model with sparse parameters Dahlin et al. (2018)). Therefore, standard universal methods receive a conservative estimate of the desired parameters. Thus, for such cases, it is necessary to develop new methods that can work in such non-standard conditions.

The main part of many machine learning methods is solving a multidimensional optimization problem (Polyak (1987)). Under conditions of substantially noisy observational data, standard gradient optimization algorithms demonstrate a significant deterioration in the quality of their operation. On the other hand, stochastic approximation algorithms with input randomization remain operational in many cases. Therefore, for training machine learning methods in such conditions, it makes sense to use recurrent adaptive data processing algorithms, among which one often uses approaches based on stochastic approximation (SA).

In Boiarov and Granichin (2019) the general problem of SPSA Gaussian mixture model (GMM) clustering was considered, and the sparse Gaussian mixture model statement was mentioned, but without mathematical justification. In this paper we focus only on the sparse GMM data model, which was first proposed in Dahlin et al. (2018), and give a mathematical result on the properties of estimates obtained by the method from Boiarov and Granichin (2019), as well as illustrate the performance of the proposed approach with several examples and compare it with some other methods.

The paper is organized as follows: Section 2 provides an overview of the main works related to the topic of this paper. In Section 3 we describe Gaussian mixture model with sparse parameters. Section 4 presents the SPSA clustering algorithm in a case of sparse GMM parameters and its mathematical analysis. In Section 5, we

[1] The work was supported IPME RAS by the Russian Science Foundation (project 16-19-00057).

provide results of experiments of the SPSA sparse GMM clustering. Section 6 concludes the paper.

2. RELATED WORKS

The SA algorithm was first proposed by Robbins and Monro in Robbins and Monro (1951) and was developed to solving the optimization problem by Kiefer and Wolfowitz (KW) in Kiefer et al. (1952). This approach, based on finite difference approximations of the gradient vector for the loss function, was extended to the d-dimensional (multidimensional with $d > 1$) case in Blum (1954). The method uses $2d$ observations on every iteration to construct a sequence of estimates: two observations to approximate each component of the d-dimensional gradient vector. Spall in Spall et al. (1992) introduced a simultaneous perturbation stochastic approximation (SPSA) algorithm with only two observations at each iteration which recursively generates estimates along random directions. It was turned out that for a large d the probabilistic distribution of appropriately scaled estimation errors is approximately normal and the SPSA algorithm has the same order of convergence rate as the KW-procedure, even though in the multidimensional case noticeably fewer (by the factor of d) observations are used. Granichin in Granichin (1992) and Polyak and Tsybakov in Polyak and Tsybakov (1990) proposed stochastic approximation algorithms with input randomization that use only one (or two) value of the function under consideration at a point (or points) on a line passing through the previous estimate in a randomly chosen direction. When unknown but bounded disturbances is added to the observed data, the quality of classical methods based on the stochastic gradient decreases. However, the quality of SA search algorithms remains high Granichin et al. (2015).

For clustering problems, Lloyd first described the classical k-means method in Lloyd (1982), whose simplicity and stability made it popular. However, its main disadvantage is that it processes all data simultaneously, so increasing the amount of data will require an increase of the memory available in the computer. In order to alleviate these drawbacks, several approaches were proposed based on the idea of online learning. Algorithm Sculley (2010) uses mini-batches (subsamples from training data) to reduce the computational time required to converge to a local solution, while minimizing the same objective function. The results obtained in this way are only slightly worse than the corresponding results of the original algorithm. Another online clustering method based on an ensemble of trained agents is discussed in Katselis et al. (2014). A more robust variation of k-means is the k-medoid method (and its implementation called Partitioning Around Medoids, PAM), which is described in Kaufman and Rousseeuw (2009). A randomized search SA algorithm solving the k-means problem was proposed, justified and extended in Boiarov and Granichin (2019).

A Gaussian mixture model (GMM) is a probabilistic model that assumes that all data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. We will consider GMM as a generalization of clustering using the k-means method. The well-known EM (expectation-maximization) algorithm (Dempster et al. (1977)) is traditionally used to find unknown parameters of the GMM. It is based on maximizing the likelihood in case when the model depends on hidden variables. The Variational Bayesian Gaussian mixture inference algorithm is an extension of the EM algorithm that can also automatically find the number of components in a mixture (see Bishop (2006)). This algorithm includes regularization by integrating information on prior distributions, which makes it more robust but slower than EM. Among online GMM clustering methods we note the flow method based on density estimates (Song and Wang (2005)). An interesting new version of the Gaussian mixture model with sparse parameters (sparse GMM) was presented in Dahlin et al. (2018) and considered as a model describing various noises.

3. PROBLEM STATEMENT

Consider an input data set $\mathbb{X} = \{\mathbf{x}^1, \mathbf{x}^2, \ldots\}$, which is a subset of the Euclidean space $\mathbb{R}^d$, and the probability distribution $\mathrm{P}(\mathbb{X})$ defined on $\mathbb{X}$. We assume that the input dataset $\mathbb{X}$ is divided into $k, k > 0$ unknown subsets

$$\{\mathbf{X}_1^\star, \ldots, \mathbf{X}_k^\star\} : \mathbb{X} = \cup_{i \in 1..k} \mathbf{X}_i^\star$$

in such a way that the probability distribution of $\mathrm{P}(\mathbb{X})$ can be represented using a mixture of distributions:

$$\mathrm{P}(\mathbb{X}) = \sum_{i=1}^{k} p_i \mathrm{P}(\mathbf{X}_i^\star)$$

where p_i $(p_i > 0)$ and $\mathrm{P}(\mathbf{X}_i^\star)$, $i = 1, \ldots, k$, are the corresponding probabilities and distributions. The clustering problem is to find the optimal partition $\mathcal{X}$ of the input dataset $\mathbb{X}$ into k nonempty clusters

$$\mathcal{X}(\mathbb{X}) = \{\mathbf{X}_1, \ldots, \mathbf{X}_k\} : \mathbb{X} = \bigcup_{i=1}^{k} \mathbf{X}_i, \mathbf{X}_i \cap \mathbf{X}_j = \emptyset,\ i \neq j.$$

Denote the best such partition as $\mathcal{X}^\star = \{\mathbf{X}_1^\star, \ldots, \mathbf{X}_k^\star\}$.

To solve this clustering problem we introduce some penalty function (quality function) q_i that defines the "closeness" to cluster i, $i \in 1..k$. Denote vectors $\theta_i,\ i \in 1, \ldots, k$ as centers of clusters, or centroids, and matrices $\Gamma_i,\ i \in 1, \ldots, k$ as covariance matrices, then to obtain optimal clustering we need to minimize the functional

$$F(\mathcal{X}) = \sum_{i=1}^{k} \int_{\mathbf{X}_i} q_i(\theta_i, \Gamma_i, \mathbf{x}) \mathrm{P}(d\mathbf{x}) \to \min_{\mathcal{X}}. \qquad (1)$$

Here $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$ is $(d \times k)$ matrix, and Γ is a set consisting of k matrices $\Gamma_1, \Gamma_2, \ldots, \Gamma_k$, where $\Gamma_i \in \mathbb{R}^{d \times d}, i \in 1, \ldots, k$.

Natural partition into clusters, which minimizes (1), leads to the following rule for assigning $\mathbf{x}$ to a particular cluster:

$$l = \mathrm{argmin}_{i=1,\ldots,k}\ q_i(\theta_i, \Gamma_i, \mathbf{x}),$$

where $l = l(\Theta, \Gamma, \mathbf{x})$ is a label function of the cluster to which the data point $\mathbf{x}$ is assigned. Denote $\mathbf{e}_l \in \mathbb{R}^k$ as a vector consisting of zeros, with one at position l. Functional (1) can be rewritten as follows:

$$F(\Theta, \Gamma) = \int_{\mathbb{X}} \mathbf{e}_l^{\mathrm{T}} \mathbf{q}(\Theta, \Gamma, \mathbf{x}) \mathrm{P}(d\mathbf{x}) \to \min_{\Theta, \Gamma}, \qquad (2)$$

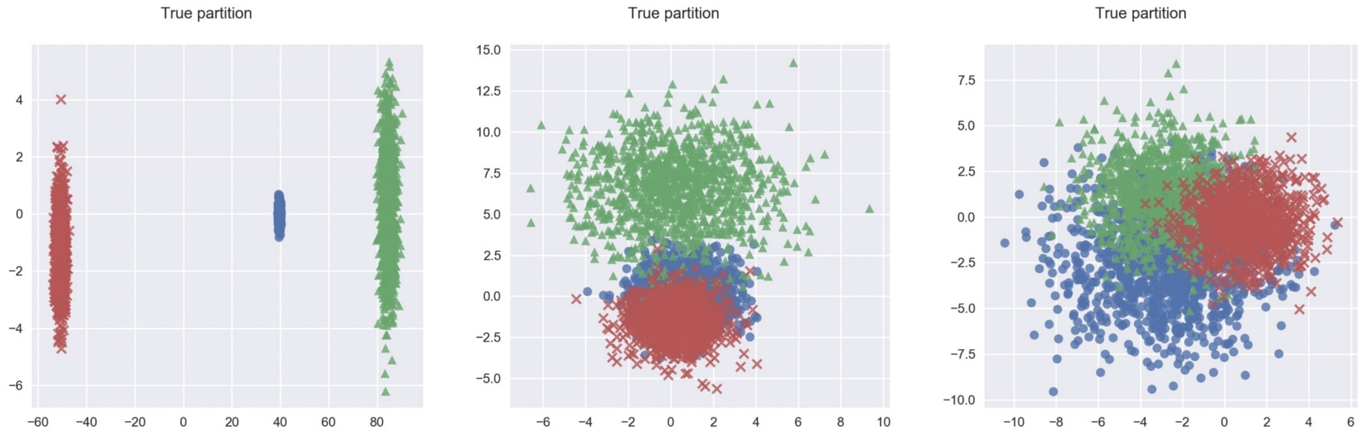

Fig. 1. Sparse GMM: Left: type 1; Center: type 2; Right: type 3.

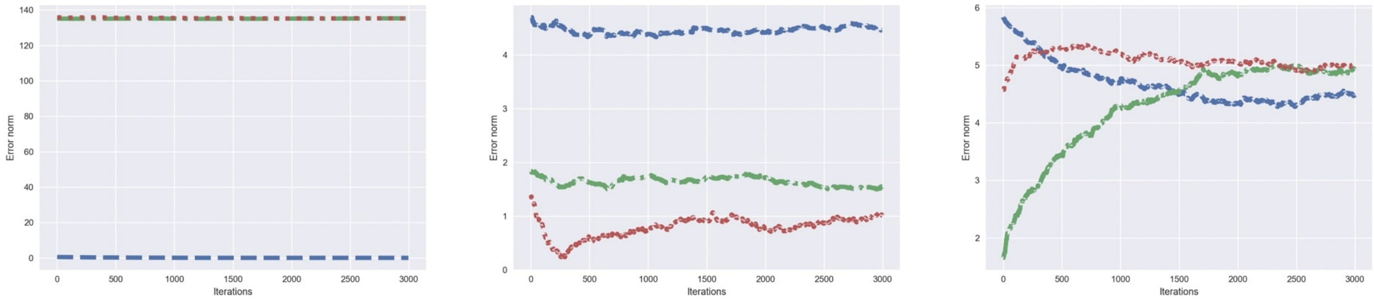

Fig. 2. Vanilla SPSA k-means centroids convergence: Left: type 1; Center: type 2; Right: type 3.

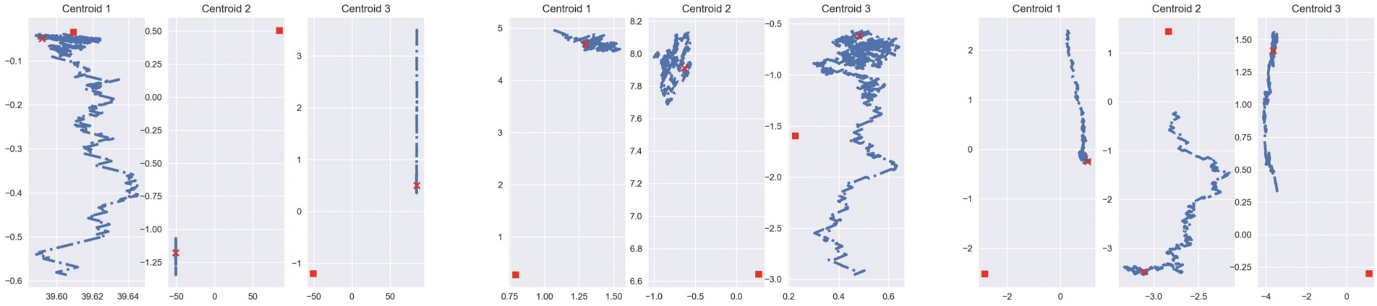

Fig. 3. Vanilla SPSA k-means traces of centroids estimates: Left: type 1; Center: type 2; Right: type 3.

where $\mathbf{q}(\Theta, \Gamma, \mathbf{x}) \in \mathbb{R}^k$ is the vector of values $q_i(\theta_i, \Gamma_i, \mathbf{x})$, $i \in 1, \ldots, k$.

An important special case corresponds to the uniform distribution of $\mathrm{P}(\cdot)$ and the penalty function, which is the square of the Mahalanobis distance

$$q_i(\theta_i, \Gamma_i, \mathbf{x}) = (\mathbf{x} - \theta_i)^{\mathrm{T}} \Gamma_i^{-1} (\mathbf{x} - \theta_i). \tag{3}$$

3.1 Gaussian Mixture Model

As the model for describing the data, we will use one of the most common such models, namely the Gaussian Mixture Model (GMM):

$$f(\mathcal{X}, \mathbf{x}) = f(\Theta, \Gamma, \mathbf{x}) = \sum_{i=1}^{k} p_i G(\mathbf{x}|\boldsymbol{\theta}_i, \Gamma_i), \tag{4}$$

where $G(\mathbf{x}|\boldsymbol{\theta}_i, \Gamma_i)$ is the density of the Gaussian distribution with mean $\boldsymbol{\theta}_i \in \mathbb{R}^d$ and covariance matrix Γ_i, $i \in 1, \ldots, k$.

Consider the following problem: By an input data sequence $\{\mathbf{x}^1, \mathbf{x}^2, \ldots\}$ and a given value k, find parameters $\boldsymbol{\theta}_i \in \mathbb{R}^d$ and Γ_i, $i \in 1, \ldots, k$ of Gaussian distributions whose mixture has generated the input data sequence. This definition fits the clustering problem introduced above and thus functional (2) according to (3) takes the form

$$F(\Theta, \Gamma) = \sum_{i=1}^{k} \sum_{\mathbf{x}^j \in \mathbf{X}_i} (\mathbf{x}^j - \boldsymbol{\theta}_i)^{\mathrm{T}} \Gamma_i^{-1} (\mathbf{x}^j - \boldsymbol{\theta}_i) \to \min_{\Theta, \Gamma}. \tag{5}$$

3.2 Sparse Gaussian Mixture Model

According to proposition in Dahlin et al. (2018) consider sparse Gaussian Mixture Model (sparse GMM) with mean:

$$\theta_i \in \mathbb{R}^d, \theta_{il} \sim \mathcal{N}(0, \sigma_i^2), \sigma_i \sim \mathcal{C}_+(0,1), l \in 1, \ldots, d, \quad (6)$$

where $\mathcal{C}_+(0,1)$ denotes the Cauchy distribution restricted to be positive with location 0 and scale 1.

And diagonal covariance matrices is given by

$$\Gamma_i = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2), \sigma_j \sim \mathcal{C}_+(0, 0.5), j \in 1, \ldots, d, \quad (7)$$

where $\mathcal{C}_+(0,0.5)$ is the half-Cauchy distribution.

Weights $p_i \sim \mathcal{D}(e_0, \ldots, e_0), i \in 1, \ldots, k$ assumes a Dirichlet prior, where parameter $e_0 \sim \mathcal{G}(\alpha_p, k\alpha_p)$ is from a Gamma distribution with mean k^{-1} and variance $(\alpha_p k^2)^{-1}$, $\alpha_p = 10$.

In sparse GMM we can distinguish three main types of the behavior of clusters (consider case with $k = 3, d = 2$) (Fig. 1). In this models often there is a situation when one cluster lies inside another what makes the clustering procedure difficult.

Authors of Dahlin et al. (2018) used this approach to simulate a variety of different interesting noise distributions. SPSA for clustering, evaluating only the centroids of the clusters Θ (similar to the k-means algorithm), shows weak results for data obtained from the sparse GMM model (see Section 5). Fig. 2 shows the plots of L_2-norms of distances between true centers of the clusters and their estimates obtained at each step of the algorithm. Fig. 3 demonstrates traces of these centroids estimates. As we can see from these plots, SPSA k-means shows weak convergence.

4. SPARSE GMM SPSA CLUSTERING

To improve the quality of the SPSA clustering algorithm on sparse GMM, we offer the following modifications. Firstly, we define the clustering algorithm from Boiarov and Granichin (2019). For all input point $\mathbf{x}^n$ and for any chosen pair Θ, Γ we can get noisy observation of penalty functions

$$y_i^n(\Theta, \Gamma) = q_i(\theta_i, \Gamma_i, \mathbf{x}^n) + v_i^n, \ i \in 1, \ldots, k,$$

where noise v_i^n is bounded: $|v_i^n| \leq c_v$, or if it is random then it does not depend on our choice of Θ, Γ and $E\{v_i^n\} < \infty$, $E\{{v_i^n}^2\} \leq (\sigma^n)^2$. Denote k-vectors of values $y_i^n(\Theta, \Gamma)$ and v_i^n as $\mathbf{y}^n(\Theta, \Gamma)$ and $\mathbf{v}^n$ respectively; $\widehat{\Theta}^n$, $\widehat{\Gamma}^n$ are estimates of centers and covariance matrices of the clusters on the n-th step of the algorithm (i.e. for $\mathbf{x}^n$) respectively; l^n is an index of the cluster to which the data point $\mathbf{x}^n$ is assigned.

Let $\Delta^n \in \mathbb{R}^d$, $n = 1, 2, \ldots$ be vectors consisting of independent random variables with Bernoulli distribution, called the *test randomized perturbation*, k is the number of clusters, $\widehat{\Theta}^0 \in \mathbb{R}^{d \times k}$ is the matrix of centroids initial values, $\widehat{\Gamma}^0$ is the set of initial covariance matrices, $\{\alpha^n\}$ and $\{\beta^n\}$ are sequences of positive numbers. Let λ be a natural number and ω^n is also a sequences of positive numbers. Then the SPSA clustering algorithm builds the following estimates

$$\begin{cases} \mathbf{y}_{\pm}^n = \mathbf{y}^n(\widehat{\Theta}^{n-1} \pm \beta^n \Delta^n \mathbf{e}_{l^n}^{\mathrm{T}}, \widehat{\Gamma}^{n-1}), \\ \widehat{\Theta}^n = \widehat{\Theta}^{n-1} - \mathbf{e}_{l^n}^{\mathrm{T}} \alpha^n \dfrac{\mathbf{y}_+^n - \mathbf{y}_-^n}{2\beta^n} \Delta^n \mathbf{e}_{l^n}^{\mathrm{T}}. \end{cases} \quad (8)$$

$$\Xi_{l^n} = \begin{cases} \omega^n \dfrac{(\widehat{\theta}_{l^n}^{n-1} - \mathbf{x}^n)(\widehat{\theta}_{l^n}^{n-1} - \mathbf{x}^n)^{\mathrm{T}} - \widehat{\Gamma}_{l^n}^{n-1}}{n}, & n > \lambda, \\ I_d, \quad otherwise. \end{cases}$$

$$\widehat{\Gamma}_{l^n}^n = \widehat{\Gamma}_{l^n}^{n-1} + \Xi_{l^n}, \quad (9)$$

where I_d is identity $d \times d$ matrix.

Secondly, we add a L_2 regularizer Hoerl and Kennard (1970) to the quality function q_i to make centroids $\theta_i, i = 1, \ldots, k$ closer to centers in sparse GMM:

$$\psi^n \left\| \theta_{l^n} - \xi \right\|^2, \quad (10)$$

where

$$\xi \in \mathbb{R}^n, \xi_l \sim \mathcal{N}(0, \sigma^2), \sigma \sim \mathcal{C}_+(0,1), l \in 1, \ldots, d;$$

ψ^n is a sequences of increasing positive numbers.

Third, we add a L_1 regularizer Tibshirani (1996) to the quality function q_i for each dimension of centroids to make them closer to a sparse mean (6):

$$\tau^n \sum_{i=1}^{d} \sum_{j=1}^{k} |\widehat{\Theta}_{ij}|, \quad (11)$$

where τ^n is a sequences of increasing positive numbers.

Thus, the new quality function takes the form:

$$\begin{cases} g_{l^n} = q_{l^n} + \psi^n \left\| \theta_{l^n} - \xi \right\|^2 + \tau^n \displaystyle\sum_{i=1}^{d} \sum_{j=1}^{k} |\widehat{\Theta}_{ij}|, \\ g_t = q_t, \ t = 1, \ldots, k, \ t \neq l_n. \end{cases} \quad (12)$$

Thus, in a modified SPSA clustering algorithm, the noisy measurement of the penalty function is defined as

$$y_i^n(\Theta, \Gamma) = g_i(\theta_i, \Gamma_i, \mathbf{x}^n) + v_i^n, \ i \in 1, \ldots, k.$$

Assumption. As can be seen from the Fig. 1, in the sparse GMM model, clusters are not always strictly separable. For this condition to be satisfied, the quality functions must have a different form. We consider here the case when they depend only on θ_i and Γ_i (for example (3)). We are interested in the convergence of estimates $\{\widehat{\Theta}^n\}$. Therefore, we will consider a model in which assumption 3 for the theorem from Boiarov and Granichin (2019) is satisfied.

Theorem 1. Let assumptions from Boiarov and Granichin (2019) and following conditions hold
(1) The learning sequence $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n, \ldots$ consists of identically distributed independent random vectors that take values in each of k classes in the attribute space $\mathbb{X}$ with a nonzero probability;
(2) $\forall n \geq 1$ the random vectors $\mathbf{v}^1, \mathbf{v}^2, \ldots, \mathbf{v}^n$ and $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^{n-1}$ do not depend on $\mathbf{x}^n$ and Δ^n, and the random vector $\mathbf{x}^n$ does not depend on Δ^n;
(3) $\sum_n \alpha^n = \infty$ and $\alpha^n \to 0$, $\beta^n \to 0$, $\alpha^n \beta^{n-2} \to 0$ as $n \to \infty$, $\omega^n \to 1$ as $n \to \infty$, $\lambda < C$.

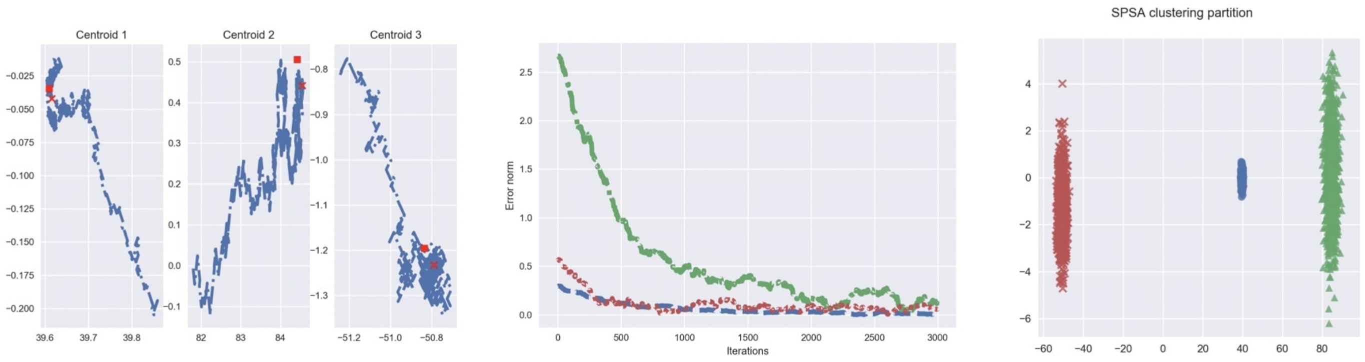

Fig. 4. Type 1: Left: centroids convergence; Center: traces of centroids estimates; Right: Sparse GMM SPSA clustering partition.

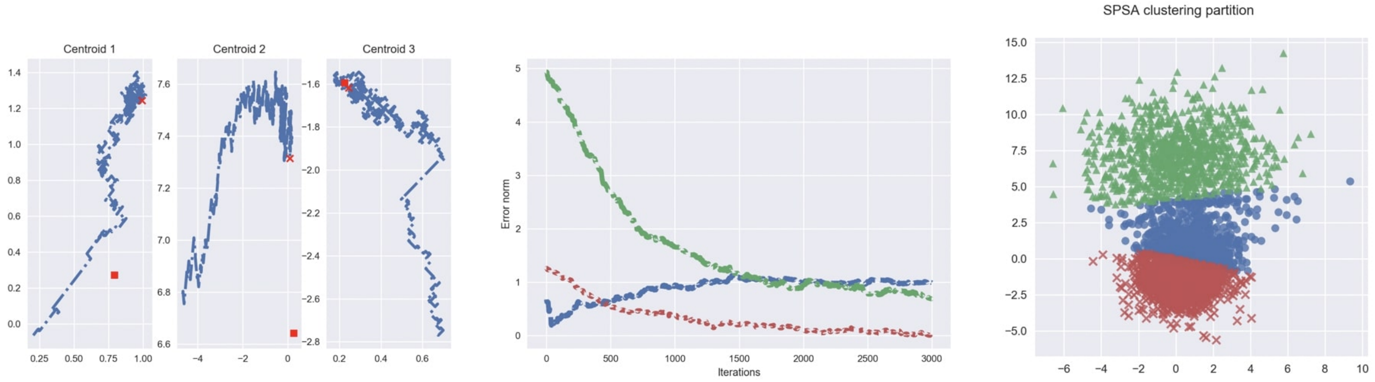

Fig. 5. Type 2: Left: centroids convergence; Center: traces of centroids estimates; Right: Sparse GMM SPSA clustering partition.

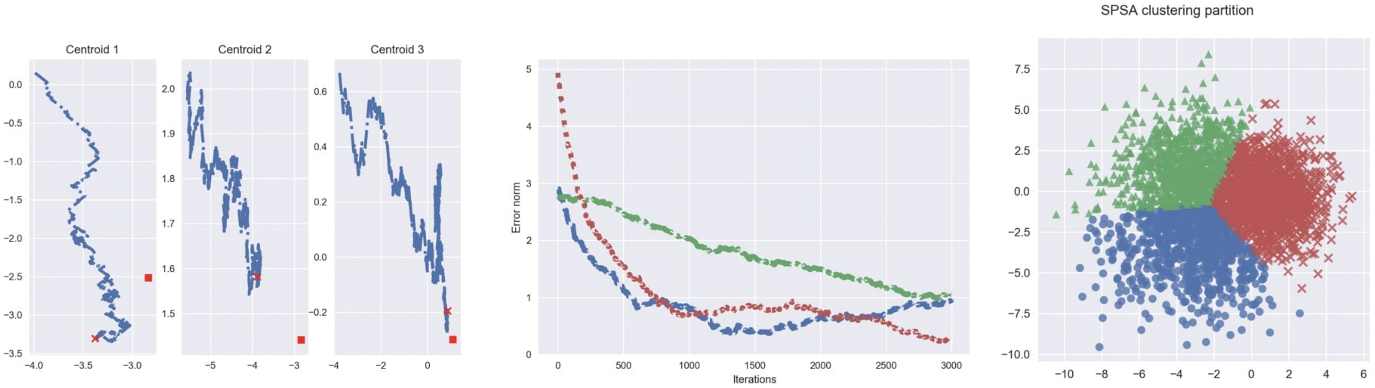

Fig. 6. Type 3: Left: centroids convergence; Center: traces of centroids estimates; Right: Sparse GMM SPSA clustering partition.

If estimate sequences $\{\widehat{\Theta}^n\}$ and $\{\widehat{\Gamma}^n\}$ generated by algorithm (8) and (9) with penalty function (12) satisfy the relation

$$\overline{\lim_{n\to\infty}} \left(\mathbf{e}^{\mathrm{T}}_{l^n}, \mathbf{g}(\widehat{\Theta}^{n-1}, \widehat{\Gamma}^{n-1}, \mathbf{x}^n) \right) \le d_{\max} + c_v,$$

then $\{\widehat{\Theta}^n\}$ converges in the mean-square sense: $\lim_{n\to\infty} E\{\|\widehat{\Theta}^n - \Theta^\star\|^2\} = 0$ and $\{\widehat{\Gamma}^n\}$ converges in probability: $\widehat{\Gamma}^n \xrightarrow{p} \Gamma^\star$.

Furthermore, **if** $\sum_n \alpha^n {\beta^n}^2 + {\alpha^n}^2 {\beta^n}^{-2} < \infty$, **then** $\widehat{\Theta}^n \to \Theta^\star$ as $n \to \infty$ with probability 1.

Proof.

1. Let's proof that the gradients of g satisfy the Lipschitz condition. Denote (10) as $L_2(\cdot)$ and (11) as $L_1(\cdot)$. Then $\|\nabla_\theta g(\theta_1) - \nabla_\theta g(\theta_2)\| =$
$= \|\nabla_\theta q(\theta_1) + \nabla_\theta L_2(\theta_1) + \nabla_\theta L_1(\theta_1) - \nabla_\theta q(\theta_2) - \nabla_\theta L_2(\theta_2) -$
$- \nabla_\theta L_1(\theta_2)\| \le \|\nabla_\theta q(\theta_1) - \nabla_\theta q(\theta_2)\| + \|\nabla_\theta L_2(\theta_1) -$
$- \nabla_\theta L_2(\theta_2)\| \le MC\|\theta_1 - \theta_2\|$,
with some constants M and C independent of $\mathbf{x}$. Moreover g is a convex function as sum of convex functions.

2. Assumption 2 of the theorem from Boiarov and Granichin (2019) is satisfied by definition of matrices $\Gamma_i, i \in 1, \ldots, k$.

3. Assumption 3 of the theorem from Boiarov and Granichin (2019) is satisfied by Assumption before this Theorem.

Thus, all conditions of the theorem from Boiarov and Granichin (2019) are satisfied.

5. EXPERIMENTS

For experiments sets $k = 3, d = 2, N = 3000$. Parameters for the SPSA clustering:

$$\gamma = 1/6, \alpha^n = 0.25/n^{\gamma}, \beta^n = 15/n^{\frac{\gamma}{4}},$$
$$\omega^n = \tanh(\tfrac{n}{\lambda}), \psi^n = 1e^{-5}, \tau^n = 1e^{-4}\tanh(\tfrac{n}{\lambda}).$$

As metrics for algorithms comparison was chosen Adjusted Rand index (ARI) and mean L_2-distance between centriods of algorithm and true centers. Comprasion of vanilla SPSA k-means, SPSA clustering for sparse GMM and some standard clustering algorithms after 1000 experiments is presented in Table 1 (Mean ARI — high is better; Mean L_2 distance — smaller is better).

Table 1. Sparse GMM clustering: ARI and L_2 distance

Algorithm	Mean ARI	Mean centers L_2
k-means	0.480	2.487
Online k-means	0.482	2.338
PAM	**0.512**	2.152
SPSA k-means	0.134	1.830
Sparse GMM-SPSA	**0.518**	**1.617**

Consider results of using modified SPSA clustering algorithm in various types of sparse GMM. For Type 1 sparse GMM the new algorithm get: ARI=1.0, L_2 centers distance=0.130 (Fig. 4).

For Type 2 sparse GMM the new algorithm get: ARI=0.521, L_2 centers distance=0.923 (Fig. 5).

For Type 3 sparse GMM the new algorithm get: ARI=0.277, L_2 centers distance=2.051 (Fig. 6).

6. CONCLUSIONS

In this work, we have presented a new modification of SPSA clustering algorithm for the case of sparse parameters of the Gaussian mixture model. Such type of model can be used to simulate a variety of different noise distributions. We have demonstrated that the algorithm remains operational under conditions of sparse GMM.

The few-shot learning task Lake et al. (2015) is close to the considered in this paper problem. Therefore, a further promising area of research is the study of the application of SPSA-based methods to this problem.

REFERENCES

Bishop, C.M. (2006). *Pattern recognition and machine learning.* Springer.

Blum, J.R. (1954). Multidimensional stochastic approximation methods. *The Annals of Mathematical Statistics*, 737–744.

Boiarov, A. and Granichin, O. (2019). Stochastic approximation algorithm with randomization at the input for unsupervised parameters estimation of gaussian mixture model with sparse parameters. *Automation and Remote Control*, 80(8), 1403–1418.

Boiarov, A. and Tyantov, E. (2019). Large scale landmark recognition via deep metric learning. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 169–178. ACM.

Dahlin, J., Wills, A., and Ninness, B. (2018). Sparse bayesian arx models with flexible noise distributions. *IFAC-PapersOnLine*, 51(15), 25–30.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 1–22.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning.*

Granichin, O., Volkovich, Z., and Toledano-Kitai, D. (2015). Randomized algorithms in automatic control and data mining.

Granichin, O. (1992). Procedure of stochastic approximation with disturbances at the input. *Automation and Remote Control*, 53(2), 232–237.

Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67.

Katselis, D., Beck, C.L., and Van Der Schaar, M. (2014). Ensemble online clustering through decentralized observations. In *53rd IEEE Conference on Decision and Control*, 910–915. IEEE.

Kaufman, L. and Rousseeuw, P.J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons.

Kiefer, J., Wolfowitz, J., et al. (1952). Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3), 462–466.

Lake, B.M., Salakhutdinov, R., and Tenenbaum, J.B. (2015). Human-level concept learning through probabilistic program induction. *Science*, 350(6266), 1332–1338.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2), 129–137.

Polyak, B.T. (1987). *Introduction to optimization.* Optimization Software, Inc, New York.

Polyak, B.T. and Tsybakov, A.B. (1990). Optimal order of accuracy of search algorithms in stochastic optimization. *Problemy Peredachi Informatsii*, 26(2), 45–53.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, 400–407.

Sculley, D. (2010). Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web*, 1177–1178. ACM.

Song, M. and Wang, H. (2005). Highly efficient incremental estimation of gmms for online data stream clustering. In *Proc. of SPIE Conference on Intelligent Computing.*

Spall, J.C. et al. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE transactions on automatic control*, 37(3), 332–341.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.