2017 IEEE Conference on Control Technology and Applications (CCTA)
August 27-30, 2017. Kohala Coast, Hawai'i, USA

Simultaneous Perturbation Stochastic Approximation for Clustering of a Gaussian Mixture Model under Unknown but Bounded Disturbances

Andrei Boiarov, Oleg Granichin, *Senior Member, IEEE*, and Hou Wenguang

***Abstract*— Multidimensional optimization holds a central role in many machine learning problems. When a model quality functional is measured with an almost arbitrary external noise, it makes sense to use randomized optimization techniques. This paper deals with the problem of clustering of a Gaussian mixture model under unknown but bounded disturbances. We introduce a stochastic approximation algorithm with randomly perturbed input (like SPSA) to solve this problem. The proposed method is appropriate for the online learning with streaming data, and it has a high speed of convergence. We study the conditions of the SPSA clustering algorithm applicability and show illustrative examples.**

***Index Terms*— Clustering, Gaussian mixture model, randomized algorithm, SPSA, unknown but bounded disturbances.**

I. Introduction

Many machine learning tools are reduced to multidimensional optimization problems. Such properties as accuracy, speed of processing and robustness are very important for these tools. In many applications (e.g. spam filter or streaming services) we get noisy observations in a sequential order and the decision must be taken in real time (online). Such systems requires online learning — the recurrent adaptive data processing algorithm. The stochastic optimization approach to clustering problem is widely used to achieve the desired properties.

Stochastic approximation was first introduced by Robbinson and Monro [1] and was further developed for optimization problems by Kiefer and Wolfowitz (KW) [2]. KW procedure was extended to the d-dimensional (multidimensional, $d > 1$) case in [3]. It is based on finite-difference approximations of the loss function gradient vector and uses $2d$ observations at each iteration to construct the sequence of estimates: two observations for approximations of each component of the gradient d-vector. Spall [4] 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. When observations data are corrupted with unknown but bounded disturbances, the performance of classical stochastic gradient based methods is dropping. But the performance of the SPSA-like algorithms is quite good [5]–[9].

Lloyd [10] studied the classical k-means algorithm. Its simplicity and stability lead to the wide popularity of this method. However it has some drawbacks: it processes all the data at a time and a large growth of data requires an increasing amount of a computer memory. Furthermore, the worst-case running time of Lloyd algorithm is exponential [11]. To improve these drawbacks several approaches were introduced which are based on online learning ideas. Algorithm [12] uses mini-batches to reduce the computation time required to converge to a local solution, while still attempting to optimize the same objective function. The obtained results are only slightly worse than the correspondence ones for the standard algorithm. A variant of streaming k-means is described in [11]. Another online clustering method was considered in [13]. It is based on an ensemble of learning agents. There is also numerically more stable vatiant of k-means such as Partition around medoids (PAM) [14].

A Gaussian mixture model (GMM) is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. We will consider a GMM as generalizing of k-means clustering. The well-known expectation-maximization (EM) algorithm [15] is traditionally used to find the unknown parameters of a GMM. It is based on a likelihood maximization, when model depends on latent variables. Variational Bayesian Gaussian mixture inference algorithm is an extension of EM that maximizes a lower bound on model evidence (see [16]). This method includes regularization by integrating the information from prior distributions which makes it more stable but slower then EM. Among online GMM clustering, we mention the streaming density-based clustering method [17], which is based on the theorem of estimator updating.

In this paper, the SPSA is considered in the context of clustering. The proposed new algorithm is an extension of the previous one for the k-means case from [18] for the case of a GMM. It is able to operate in online regime with streaming data and shows a high speed. The convergence of the cluster centers to their true values is guaranteed even if quality functionals are measured with unknown but bounded noises. We analyze the time complexity of the proposed method too.

This work was supported by RFBR (projects 16-07-00890 and 17-51-53053).

A. Boiarov and O. Granichin are with the Saint Petersburg State University (Faculty of Mathematics and Mechanics, and Research Laboratory for Analysis and Modeling of Social Processes), 7-9, Universitetskaya Nab., St. Petersburg, 199034, Russia. O. Granichin is also with the Institute of Problems in Mechanical Engineering, Russian Academy of Sciences, and with the ITMO University, St. Petersburg, Russia. Hou Wenguang is with the School of Life Science and Technology Huazhong university of Science and Technology, 1037 Luoyu Road, Wuhan, Hubei, 430074, China. E-mail: a.boiarov@spbu.ru, o.granichin@spbu.ru, houwenguang99@163.com.

This paper is organized as follows: in Section II, we formulate a problem setting of clustering of a Gaussian mixture model as a multidimensional optimization problem. Section III presents the SPSA clustering algorithm, its mathematical analysis and main properties. In Section IV, we demonstrate experiments results with and without several types of external noise and also the real life application of proposed method. Section V concludes the paper.

II. Problem Statement

Denote $1..k$ the set of indexes $\{1,2,\ldots,k\}$. Assume that k is a known value, input data set $\mathbb{X}$ is a subset of Euclidean space $\mathbb{R}^d$ and it is divided into k unknown subsets: $\mathbb{X} = \cup_{i\in 1..k}\mathbf{X}_i^\star$, and there is underlying probability distribution $\mathrm{P}(\mathbb{X})$ of set $\mathbb{X}$ which is represented as 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 $\mathrm{P}(\mathbf{X}_i^\star)$, $i \in 1..k$, are subset i probabilities and distributions.

Clustering problem is to find an optimal partition $\mathscr{X}$ of input data set $\mathbb{X}$ to k non-empty clusters $\mathscr{X}(\mathbb{X}) = \{\mathbf{X}_1,\ldots,\mathbf{X}_k\}$: $\mathbb{X} = \bigcup\limits_{i=1}^{k} \mathbf{X}_i$ and $\mathbf{X}_i \cap \mathbf{X}_j = \emptyset$, $i \neq j$. The partition $\mathscr{X}$ is defined by label function $\gamma_{\mathscr{X}} : \mathbb{X} \to 1..k$ which assigns to each point of $\mathbb{X}$ the number of data cluster. So, $\mathbf{X}_i = \{\mathbf{x} \in \mathbb{X} | \gamma_{\mathscr{X}}(\mathbf{x}) = i\}$. There are many possible partitions of the set $\mathbb{X}$ for any k. The clustering problem is to find the best of them which is coincide with $\mathscr{X}^\star = \{\mathbf{X}_1^\star,\ldots,\mathbf{X}_k^\star\}$.

Mathematically, the clustering problem can be described as follows: elements belonging to the same group (cluster) are more similar than the elements belonging to different groups (clusters). To solve this problem we introduce some distortion (penalty, quality) functions q_i determining "closeness" data point to the cluster i, $i \in 1..k$. Then the optimal clustering is a minimizer of mean risk functional:

$$F(\mathscr{X}) = Ef(\mathscr{X},\mathbf{x}) \to \min_{\mathscr{X}}, \tag{1}$$

where E is a symbol of mathematical expectation and

$$f(\mathscr{X},\mathbf{x}) = \sum_{i=1}^{k} \gamma_{\mathscr{X}}(\mathbf{x}) q_i(\mathscr{X},\mathbf{x}).$$

If vectors θ_i, $i \in 1..k$ are conveniently interpreted as *centers of clusters* or *centroids* and matrices Γ_i, $i \in 1..k$ — as *covariance matrices* then the functional of clustering quality (1) takes the form

$$F(\mathscr{X}) = \sum_{i=1}^{k} \int_{\mathbf{X}_i} q_i(\theta_i,\Gamma_i,\mathbf{x}) \mathrm{P}(d\mathbf{x}) \to \min_{\mathscr{X}}. \tag{2}$$

For $i \in 1..k$ and fixed $\mathbf{x} \in \mathbb{X}$ each function $q_i(\cdot,\cdot,\mathbf{x})$ depends only on θ_i and Γ_i, that is $q_i(\cdot,\cdot,\cdot) : \mathbb{R}^d \times \mathbb{R}^{d\times d} \times \mathbb{X} \to \mathbb{R}$. Then we can choose the partition rule

$$\mathbf{X}_i(\Theta,\Gamma) = \{\mathbf{x} \in \mathbb{X} : q_i(\theta_i,\Gamma_i,\mathbf{x}) < q_j(\theta_j,\Gamma_j,\mathbf{x}),\ j \in 1..i-1;$$
$$q_i(\theta_i,\Gamma_i,\mathbf{x}) \le q_j(\theta_j,\Gamma_j,\mathbf{x}),\ j \in i+1..k\},\ i \in 1..k,$$

which minimizes (1). Here $\Theta = (\theta_1,\theta_2,\ldots,\theta_k)$ is $d \times k$ matrix, and Γ is a set of k matrices $\Gamma_1,\Gamma_2,\ldots,\Gamma_k$, where $\Gamma_i \in \mathbb{R}^{d\times d}, i \in 1..k$. Thus we can rewrite (2) as follows:

$$F(\Theta,\Gamma) = \int_{\mathbb{X}} \langle \mathbf{j}(\Theta,\Gamma,\mathbf{x}), \mathbf{q}(\Theta,\Gamma,\mathbf{x}) \rangle \mathrm{P}(d\mathbf{x}) \to \min_{\Theta,\Gamma} \tag{3}$$

where $\mathbf{j}(\Theta,\Gamma,\mathbf{x}) \in \mathbb{R}^k$ is a vector consisting of ones and zeros, corresponding to values of characteristic functions $\mathbf{1}_{\mathbf{X}_i(\Theta,\Gamma)}(\Theta,\Gamma,\mathbf{x}), i \in 1..k$, and $\mathbf{q}(\Theta,\Gamma,\mathbf{x}) \in \mathbb{R}^k$ is a vector of values $q_i(\theta_i,\Gamma_i,\mathbf{x}), i \in 1..k$.

The important partial case is corresponded the uniform distribution $\mathrm{P}(\cdot)$ and squares of Mahalanobis distances as penalty functions

$$q_i(\theta_i,\Gamma_i,\mathbf{x}) = (\mathbf{x}-\theta_i)^{\mathrm{T}} \Gamma_i^{-1} (\mathbf{x}-\theta_i). \tag{4}$$

A. K-means algorithm

Consider one of the most popular clustering techniques: k-means algorithm [10]. It looks for a partition $\mathscr{X}$ that minimizes the sum of squares of intracluster distances. Each cluster is characterized by the relevant centroid. All of the matrix from Γ are assumed unit.

Algorithm 1 k-means

Input: $\mathbb{X}$, k, the maximum number of iterations
Output: estimate centroids $\widehat{\Theta}$, $\mathscr{X}$

1: *Initialization* : $n := 0$. Select randomly k initial centroids $\widehat{\Theta}^0 = (\widehat{\theta}_1^0, \widehat{\theta}_2^0, \ldots, \widehat{\theta}_k^0)$ from the elements of $\mathbb{X}$
2: *Classification* : $i \in 1..k$
$\mathbf{X}_i^n = \{\mathbf{x} \in \mathbb{X} : \|\widehat{\theta}_i^n - \mathbf{x}\|^2 \le \|\widehat{\theta}_j^n - \mathbf{x}\|^2,\ j \in 1..k\}$
3: *Minimization* : $\widehat{\theta}_i^{n+1} = \frac{1}{|\mathbf{X}_i^n|} \sum_{\mathbf{x}_j \in \mathbf{X}_i^n} \mathbf{x}_j$
4: *Iteration* : $n := n+1$. Steps 2 and 3 are repeated until centroids do not change or the maximum number of iterations is not reached.

B. Gaussian Mixture Model

One of the most widespread input data model is the *Gaussian Mixture Model* (GMM) of data fitting:

$$f(\mathscr{X},\mathbf{x}) = f(\Theta,\Gamma,\mathbf{x}) = \sum_{i=1}^{k} p_i G(\mathbf{x}|\theta_i,\Gamma_i), \tag{5}$$

where $G(\mathbf{x}|\theta_i,\Gamma_i)$ is the Gaussian density with the mean θ_i and covariance matrix Γ_i, $i \in 1..k$.

Likelihood maximization approaches are traditionally used to solve the correspondence clustering problem. Expectation-maximization (EM) algorithm [15] or Variational Bayesian Gaussian mixture inference algorithm [16] are based on the estimation of latent variables. Gaussian mixture models is a generalization of k-means clustering which can consider the information about the covariance structures of the data as well as the centers of the latent Gaussians.

III. SPSA Clustering

Let $\mathbf{x}^1,\mathbf{x}^2,\ldots,\mathbf{x}^n,\ldots$ be a sequence of input data which generates according the probability distribution (5) with nominal parameters $\Theta^\star$ and $\Gamma^\star$. Hereandafter the upper index n is used as iteration index.

In this section we study the SPSA-like clustering algorithm which minimizes mean risk functional (3) with penalty function determined by (4) (a squared Mahalanobis distance).

Lemma 1. Under the conditions described above, the functional (3) reaches a minimum at $\Theta^\star$ and $\Gamma^\star$.

Proof: If the sequence $\{\mathbf{x}^n\}$ is generated by the GMM with parameters $\Theta^\star$ and $\Gamma^\star$, then with these parameters a minimum of the minus log-likelihood function $-\sum_{\mathbf{x}\in\mathbb{X}}\ln\left(\sum_{i=1}^k p_i G(\mathbf{x}|\theta_i,\Gamma_i)\right)$ is reached. By definition of the Gaussian density it can be rewritten as follows

$$\sum_{i=1}^{k}\sum_{\mathbf{x}_i\in\mathbf{X}_i}\left(-\ln p_i(2\pi)^{-\frac{d}{2}}|\Gamma_i|^{-1}+1/2(\mathbf{x}-\theta_i)^{\mathrm{T}}\Gamma_i^{-1}(\mathbf{x}-\theta_i)\right). \tag{6}$$

Denote the left-hand side of (6) as L and the right-hand side of (6) as R. Then $\arg\min_{\Theta,\Gamma}L=\arg\min_{\Theta,\Gamma}R=(\Theta^\star,\Gamma^\star)$.

Functional (3) can be rewritten by the conditions through (2) and (4) as follows

$$F(\Theta,\Gamma)=\sum_{i=1}^{k}|\mathbf{X}_i|^{-1}\sum_{\mathbf{x}\in\mathbf{X}_i}(\mathbf{x}-\theta_i)^{\mathrm{T}}\Gamma_i^{-1}(\mathbf{x}-\theta_i).$$

Thus $\arg\min_{\Theta,\Gamma}F(\Theta,\Gamma)=\arg\min_{\Theta,\Gamma}R=(\Theta^\star,\Gamma^\star)$. ■

We make the following assumptions: **As1.** For all input point $\mathbf{x}^n$ and for any chosen pair Θ,Γ we can get noisy observation of penalty functions

$$y_i^n(\Theta,\Gamma)=q_i(\theta_i,\Gamma_i,\mathbf{x}^n)+v_i^n,\ i\in 1..k, \tag{7}$$

and noise v_i^n is bounded: $|v_i^n|\le c_v$, and if it is random then it does not depend on our choice of Θ,Γ and $E\{v_i^n\}<\infty$, $E\{v_i^{n2}\}\le\sigma^{n2}$.

As2. All matrices Γ_i, $i\in 1..k$ are symmetric positive semidefinite and their eigenvalues are bounded $\lambda_i^j\le C_\lambda$, $i\in 1..k, j\in 1..d$.

As3. The clusters $\mathbf{X}_i(\Theta^\star,\Gamma^\star)$, $i\in 1..k$, are divided among themselves significantly in the sense that: if for some $i\in 1..k$, $\mathbf{x}\in\mathbf{X}_i(\Theta^\star,\Gamma^\star)$, and θ_i, Γ_i the inequality

$$|q_i(\theta_i,\Gamma_i,\mathbf{x})|\le d_{\max}=\max_{i\in 1..k}\max_{\mathbf{x}\in\mathbf{X}_i(\Theta^\star,\Gamma^\star)}|q_i(\theta_i^\star,\Gamma_i^\star,\mathbf{x})|$$

holds, then $\forall j\ne i,\ j\in 1..k$, the following inequity is satisfied:

$$|q_j(\theta_j,\Gamma_j,\mathbf{x}')|>d_{\max}+2c_v\ \ \forall\mathbf{x}'\in\mathbf{X}_j(\Theta^\star,\Gamma^\star). \tag{8}$$

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; $\mathbf{j}^n(\Theta,\Gamma)$ is k-vector of characteristic functions $(\mathbf{j}(\Theta,\Gamma,\mathbf{x}))$, found with noisy measurements $\mathbf{y}^n(\Theta,\Gamma)$, $n=1,2,\ldots$; $\widehat{\Theta}^n$, $\widehat{\Gamma}^n$ 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; $l^n=argmax\ \mathbf{j}^n(\Theta,\Gamma)$ 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.

For estimating a covariance matrix of the cluster we will use scatter matrix from maximum likelihood estimation [19] and parameterized cumulative moving of covariance matrices. Let λ is a natural number and ω^n is also a sequences of positive numbers. Then the SPSA clustering algorithm builds the following estimates

$$\begin{cases}\mathbf{y}_{\pm}^n=\mathbf{y}^n(\widehat{\Theta}^{n-1}\pm\beta^n\Delta^n\mathbf{j}^{n\mathrm{T}},\widehat{\Gamma}^{n-1}),\\ \widehat{\Theta}^n=\widehat{\Theta}^{n-1}-\mathbf{j}^{n\mathrm{T}}\alpha^n\frac{\mathbf{y}_+^n-\mathbf{y}_-^n}{2\beta^n}\Delta^n\mathbf{j}^{n\mathrm{T}}.\end{cases} \tag{9}$$

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

$$\widehat{\Gamma}_{l^n}^n=\widehat{\Gamma}_{l^n}^{n-1}+\Xi_{l^n}, \tag{10}$$

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

Theorem 1: Let assumptions **As1–3** 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) $\forall n\ge 1$ 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 $n\to\infty$, $\lambda<C$.

If estimate sequences $\{\widehat{\Theta}^n\}$ and $\{\widehat{\Gamma}^n\}$ generated by algorithm (9) and (10) satisfy the relation

$$\overline{\lim_{n\to\infty}}\left\langle\mathbf{j}^n(\widehat{\Theta}^{n-1},\widehat{\Gamma}^{n-1}),\mathbf{q}(\widehat{\Theta}^{n-1},\widehat{\Gamma}^{n-1},\mathbf{x}^n)\right\rangle\le d_{\max}+c_v, \tag{11}$$

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** $\sum_n\alpha^n\beta^{n2}+\alpha^{n2}\beta^{n-2}<\infty$,

then $\widehat{\Theta}^n\to\Theta^\star$ as $n\to\infty$ with probability 1.

The proof of the theorem is given in the Appendix.

Note that the advantages properties of the SPSA clustering described above are:

- The algorithm is iterative, i.e. implementing the idea of online learning:
  - adaptability, processing of a new data on the fly;
  - memory savings, it is not necessary to store the entire data set in memory.
- High speed of the algorithm;
- The algorithm remains in operation during the growth of the dimension of the estimated parameters (unlike for example k-means);
- Resistance to an almost arbitrary external noise in the measurement of penalty function at the points of the input data.

Compare the time complexity of the simple basic version of k-means algorithm, described in (**??**) (as a special case of *EM*), and the SPSA clustering. In general k-means is *NP*-Hard to optimize. Let t be the number of iterations of the algorithm (the maximum setting, or necessary for convergence). For one iteration of k-means (classification, minimization) it is necessary kd operations. The time complexity

is $\mathcal{O}(tNkd)$ since the algorithm involves all N observations at each iteration.

For comparison with k-means we will assume that all matrices $\Gamma_i, i \in 1..k$, in (4) are unit. It is necessary kd operations to calculate $\mathbf{j}^{n\mathrm{T}}$. It is also need kd operations for calculations of each $\mathbf{y}_-^n$ and $\mathbf{y}_+^n$. Thus, the time complexity of the each iteration of the SPSA clustering is $\mathcal{O}(3kd)$. The number of iterations is equal to N. Then estimate the time complexity is $\mathcal{O}(3Nkd)$.

From these estimates of time complexities, we can conclude that the SPSA clustering algorithm is faster than k-means under considered conditions for $t > 3$. Note that t is significantly greater than 3 often in real practical applications. Method of Γ estimation (10) adds $\mathcal{O}((N-\lambda)d^2)$ to the complexity of main algorithm.

IV. Experiments

We carry out a series of experiments for a comparison of the proposed algorithm with the classical approaches. We consider a synthetic data set of small dimension and the case of a real data of high dimensionality. We use adjusted Rand index (ARI) [20] as clustering performance metric in experiments.

At first, we take Gaussian mixture with the following parameters: amount of data $N = 5000$, $\Theta = \begin{pmatrix} 0 & 2 & -3 \\ 0 & 2 & 6 \end{pmatrix}$, $\Gamma_1 = \begin{pmatrix} 1 & -0.7 \\ -0.7 & 1 \end{pmatrix}$, $\Gamma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\Gamma_3 = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$, mixture probabilities are $\mathbf{p} = col(0.4, 0.4, 0.2)$.

We chose the following parameters for the SPSA clustering: $\gamma = 1/6$, $\alpha^n = 0.25/n^\gamma$, $\beta^n = 15/n^{\frac{\gamma}{4}}$ and $\omega^n = \tanh(\frac{n}{\lambda})$, based on the results concerning fastest rate of convergence of stochastic algorithm from [21].

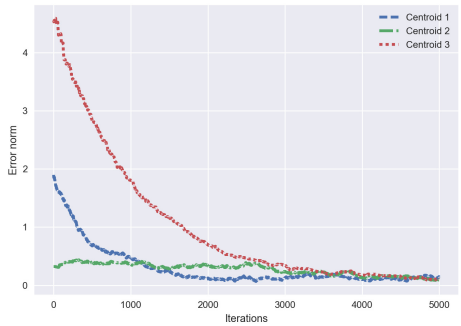

Fig. 1: L_2-norm of the convergence of centroids estimates obtained by the SPSA clustering in the k-means experiment.

Fig. 1–2 represents results of using the SPSA clustering for the case identity matrices $\Gamma_i, i \in 1..k$ (k-means case). Fig. 1 shows L_2-norm of the distance between the true centroid and its estimate obtained at each step of the algorithm. Fig. 2 displays traces of centroids estimates by algorithm steps.

We use mini-batch k-means from [12] with size of batch 1 (so it can be called online k-means) for the performance comparison. Table I shows the average values of adjusted

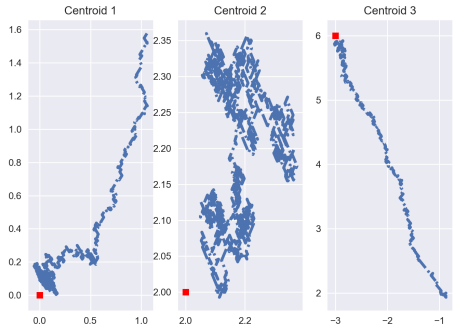

Fig. 2: Traces of centroids estimates from the SPSA clustering (blue dots) and true values of cluster centers (red square) in the k-means experiment.

TABLE I: ARI of the k-means experiment

Algorithm	Mean ARI
k-means	0.858
Online k-means	0.819
PAM	0.84
SPSA clustering	0.857

Rand index for each algorithm after 100 runs. It illustrates that the performance of our method is similar to the performances of k-means, online k-means and PAM algorithms.

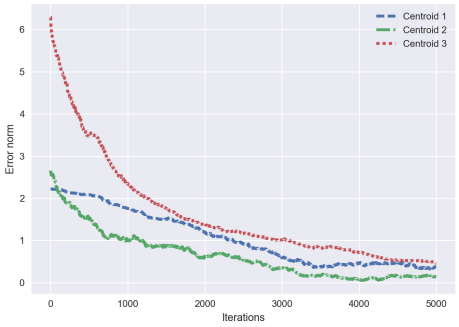

Fig. 3: L_2-norm of the convergence of centroids estimates obtained by the SPSA clustering in the GMM experiment.

Fig. 3 shows L_2-norm of the distance between the true cluster centers and its estimate obtained from the SPSA clustering and Fig. 4 displays traces of centroids estimates for the GMM case with $\lambda = 1000$.

We compare the SPSA clustering with expectation-maximization (EM) algorithm and Variational Bayesian Gaussian mixture inference algorithm. Table II represents the average values of adjusted Rand index for each algorithm after 100 runs. One can see from these results that our method is close to two classical approaches of fitting GMM, but we note again that it has a significantly lower time complexity.

We considered the external noise v_i^n in (7) as zero in previ-

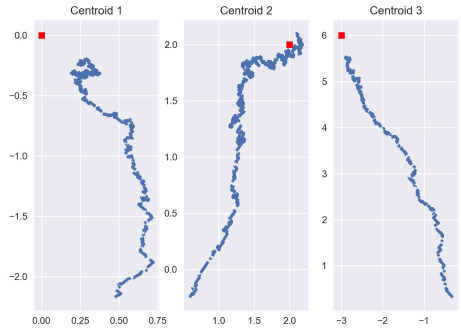

Fig. 4: Traces of centroids estimates from the SPSA clustering (blue dots) and true values of cluster centers (red square) in the GMM experiment.

TABLE II: ARI of the GMM experiment

Algorithm	Mean ARI
EM	0.903
Variational Bayesian Gaussian mixture	0.915
SPSA clustering	0.909

ous experiments. We examine also various types of noise for all $i \in 1..k$, $n = 1,2,\ldots$: $v_i^n \sim \mathcal{N}(0,1)$; $v_i^n \sim \mathcal{N}(0,\sqrt{2})$; $v_i^n \sim \mathcal{N}(1,1)$; $v_i^n \sim \mathcal{N}(1,\sqrt{2})$; random: $v_i^n = 10 \cdot (\text{rand}() \cdot 4 - 2)$; irregular: $v_i^n = 0.1 \cdot \sin(n) + 19 \cdot \text{sign}(50 - \text{n} \mod 100)$; constant: $v_i^n = 20$.

We compare four algorithms under this conditions: k-means (as more stable than online version), our method with identity covariance matrices (denote as SPSA identity), EM algorithm and our method with estimation of covariance matrices with $\lambda = 3000$ (denote as SPSA cov). The average values of adjusted Rand index for each algorithm after 10 runs are presented in Table III.

As can see from this results the SPSA clustering algorithm greatly outperform k-means almost in all noise cases. Version of our method with estimation of covariance matrices also almost in all cases outperform both EM algorithm and the version with identity matrices.

To illustrate the work of the SPSA clustering in the case of a high dimensional real life data we apply our algorithm to the subset of well-known base of handwritten digits MNIST [22]. This subset contains 10000 gray-scale

TABLE III: Mean ARI of noise experiments

Noise	k-means	SPSA identity	EM	SPSA cov
$\mathcal{N}(0,1)$	0.608	0.768	0.654	0.815
$\mathcal{N}(0,\sqrt{2})$	0.246	0.546	0.463	0.738
$\mathcal{N}(1,1)$	0.612	0.829	0.657	0.774
$\mathcal{N}(1,\sqrt{2})$	0.246	0.601	0.371	0.612
Random	0	0.418	0.318	0.434
Irregular	0.758	0.854	0.863	0.856
Constant	0.812	0.861	0.807	0.860

normalized images of size 28×28. Resulting centroids are presented on Fig. 5.

Fig. 5: MNIST centroids obtained by the SPSA clustering.

Our algorithm demonstrates adequate results but centroids of 3 and 9 stuck with each other that can be explained by the fact that the images of these digits are similar.

V. Conclusion

We present the SPSA-based clustering algorithm for the Gaussian mixture model and its theoretical justification. We demonstrate robustness of this method in a case when data are measured with an almost arbitrary external noise. In addition to this property, the proposed algorithm implements the idea of online learning and has a high speed. Series of experiments demonstrate advantages of our method over classical approaches. We investigate the application of the proposed algorithm in the task of processing medical data using convolutional neural network.

VI. Appendix

Proof: **(1)** At first, we fix Γ and prove the convergence of $\{\widehat{\Theta}^n\}$. Choose some r from $1,2,\ldots,k$ and select a subsequence $\{\widehat{\Theta}^{n_j}\}$, when constructing which have been adjusted only estimates of r-th class center.

We need to prove that $\exists\ N_r$ such that $\mathbf{x}^{n_j} \in \mathbf{X}_r(\Theta^\star,\Gamma^\star)\ \ \forall j \geq N_r$. Suppose this is not true. Hence, $\exists$ an infinite increasing subsequence $\{n_{j_t}\}$ for which $\mathbf{x}^{n_{j_t}} \in \mathbf{X}_r(\Theta^\star,\Gamma^\star)$ and $\mathbf{x}^{n_{j_t+1}} \notin \mathbf{X}_r(\Theta^\star,\Gamma^\star)$. Let the function $F_i(\theta_i,\Gamma_i) = \int_{\mathbf{X}_i} q_i(\theta_i,\Gamma_i,\mathbf{x})\mathrm{P}(d\mathbf{x})$, $i \in 1..k$ is strongly convex, then due to the algorithm (9), the mean value theorem and a compactness of the set, we have $\|\widehat{\theta}_r^{(n_{j_t}-1)} - -\widehat{\theta}_r^{n_{j_t}}\| \leq \alpha^{n_{j_t}} m(2C_v + \max_{s\in[-1;1]}|\nabla_\theta q(\mathbf{x}^{n_{j_t}}, \widehat{\theta}_r^{n_{j_t}-1} + +s\alpha^{n_{j_t}}\Delta^{n_{j_t}},\Gamma_r)|) \leq \alpha^{n_{j_t}} C$, with some constant C. Consequently, for sufficiently large t the difference $|q(\mathbf{x}^{n_{j_t+1}}, \widehat{\theta}_r^{n_{j_t}},\Gamma_r) - q(\mathbf{x}^{n_{j_t+1}}, \widehat{\theta}_r^{(n_{j_t}-1)},\Gamma_r)| < C_v/3$. Since $\mathbf{x}^{n_{j_t+1}} \notin \mathbf{X}_r(\Theta^\star,\Gamma^\star)$, then by (8) $|q(\mathbf{x}^{n_{j_t+1}}, \widehat{\theta}_r^{(n_{j_t}-1)},\Gamma_r)| > > d_{\max} + 2C_v$. From these relations for sufficiently large t, we have $|q(\mathbf{x}^{n_{j_t+1}}, \widehat{\theta}_r^{n_{j_t}},\Gamma_r)| \geq |q(\mathbf{x}^{n_{j_t+1}}, \widehat{\theta}_r^{(n_{j_t}-1)},\Gamma_r)| - -|q(\mathbf{x}^{n_{j_t+1}}, \widehat{\theta}_r^{n_{j_t}},\Gamma_r) - q(\mathbf{x}^{n_{j_t+1}}, \widehat{\theta}_r^{(n_{j_t}-1)},\Gamma_r)| > d_{\max} + 5/3C_v$. On the other hand, by the condition (11) for sufficiently large t $|q(\mathbf{x}^{n_{j_t}}, \widehat{\theta}_r^{(n_{j_t}-1)},\Gamma_r)| < d_{\max} + \frac{4}{3}C_v$. This is a contradiction.

Next we enumerate for the convenience of the sequence $\{\widehat{\theta}_r^{n_j}\}$ with $j = N_k$ to $\{\widehat{\theta}_r^i\}$. By (9) we have $\|\widehat{\theta}_r^i - \theta_r^\star\|^2 \leq \leq \|\widehat{\theta}_r^{(i-1)} - \theta_r^\star - \frac{\alpha^i}{2\beta^i}\left(y_+^i - y_-^i\right)\Delta^i\|^2$.

Denote $\mathscr{F}_r^{n-1}$ - σ-algebra of events, generated by the random variables $\{\widehat{\theta}_r^{(n-1)}\}$, constructed by (9). Then we get

$$\mathrm{E}_{\mathscr{F}_r^{n-1}}\|\widehat{\theta}_r^i - \theta_r^\star\|^2 \leq \|\widehat{\theta}_r^{i-1} - \theta_r^\star\|^2 - \alpha^i\langle\widehat{\theta}_r^{(i-1)} - \theta_r^\star, \quad (12)$$

$\beta^{i-1}\mathrm{E}_{\mathscr{F}_r^{n-1}}\Delta^i\left(y^i_{r,+}-y^i_{r,-}\right)\rangle+\frac{\alpha^{i2}}{4\beta^{i2}}\mathrm{E}_{\mathscr{F}_r^{n-1}}\|\Delta^i\|^2\left(y^i_{r,+}-y^i_{r,-}\right)^2$.

Let us estimate last term on the right-hand side (12). Let the function $q(\mathbf{x}_i,\cdot,\Gamma_i)$ satisfies the Lipschitz condition, $\|\Delta^i\|^2=m$, using the mean value theorem and the Cauchy-Schwarz inequality we derive $\mathrm{E}_{\mathscr{F}_r^{n-1}}\|\Delta^i\|^2\left(y^i_{r,+}-y^i_{r,-}\right)^2 = 2m(\mathrm{E}_{\mathscr{F}_r^{n-1}}\left(v^i_{r,+}-v^i_{r,-}\right)^2+$ $+\mathrm{E}_{\mathscr{F}_r^{n-1}}\big(q(\mathbf{x}^i,\widehat{\theta}_r^{(i-1)}+\beta^i\Delta^i,\Gamma_r)-q(\mathbf{x}^i,\widehat{\theta}_r^{(i-1)}-\beta^i\Delta_i,\Gamma_r)\big)^2\big)\le 2m\mathrm{E}_{\mathscr{F}_r^{n-1}}(v^i_r)^2+2m\mathrm{E}_{\mathscr{F}_r^{n-1}}(|q(\mathbf{x}^i,\widehat{\theta}_r^{(i-1)}+$ $+\beta^i\Delta^i,\Gamma_r)-q(\mathbf{x}^i,\theta^\star_r,\Gamma^\star_r)|+|q(\mathbf{x}^i,\widehat{\theta}_r^{(i-1)}-\beta^i\Delta^i,\Gamma_r)-$ $-q(\mathbf{x}^i,\theta^\star_r,\Gamma^\star_r)|)^2\le 2m\mathrm{E}_{\mathscr{F}_r^{n-1}}\big((M+0.5)\left(\|\widehat{\theta}_r^{(i-1)}+\Delta^i\beta^i\|^2+\right.$ $\left.+\|\widehat{\theta}_r^{(i-1)}-\Delta^i\beta^i\|^2\right)+\|\nabla_\theta q(\mathbf{x}^i,\theta^\star_r,\Gamma^\star_r)\|^2\big)^2+2m\mathrm{E}_{\mathscr{F}_r^{n-1}}(v^i_r)^2$. In view of the uniform boundedness $\nabla_\theta q(\cdot,\theta^\star_r,\Gamma^\star_r)$, we get $\mathrm{E}_{\mathscr{F}_r^{n-1}}\|\Delta^i\|^2\left(y^i_{r,+}-y^i_{r,-}\right)^2\le C_1\mathrm{E}_{\mathscr{F}_r^{n-1}}(v^i_r)^2+C_2\beta^{i2}\|\widehat{\theta}_r^{(i-1)}-\theta^\star_r\|^2+o(\beta^{i^2})$, where C_i, $i=1,2\ldots$ is some positive constants. Let us turn to the second term of the right-hand side (12)

$$\beta^{i-1}\mathrm{E}_{\mathscr{F}_r^{n-1}}\Delta^i\left(y^i_{r,+}-y^i_{r,-}\right)=\beta^{i-1}\mathrm{E}_{\mathscr{F}_r^{n-1}}\Delta^i v^i_r+\beta^{i-1}\mathrm{E}_{\mathscr{F}_r^{n-1}}\Delta^i\times$$
$$\times\big(q(\mathbf{x}^i,\widehat{\theta}_r^{(i-1)}+\beta^i\Delta^i,\Gamma_r)-q(\mathbf{x}^i,\widehat{\theta}_r^{(i-1)}-\beta^i\Delta^i,\Gamma_r)\big).\quad(13)$$

The first term on the right-hand side of (13) is equal to zero by the properties of Δ^i and independence of Δ^i and $v^i_{r,\pm}$. Using the mean value theorem, uniform boundedness of the function $\nabla_\theta q(\cdot,\theta,\Gamma)$ and definition of $F_r(\Theta,\Gamma)$, convert the second term on the right of the (13) to the form $\beta^{i-1}\mathrm{E}_{\mathscr{F}_r^{n-1}}(\Delta^i\left(y^i_{r,+}-y^i_{r,-}\right)=2\nabla F_r(\widehat{\theta}_r^{(i-1)})+$ $+\mathrm{E}_{\mathscr{F}_r^{n-1}}(\Delta^i\big\langle\Delta^i,\nabla_\theta q(\mathbf{x}^i,\theta_r^{(i-1)^+},\Gamma_r)-\nabla_\theta q(\mathbf{x}^i,\widehat{\theta}_r^{(i-1)},\Gamma_r)\big\rangle+$ $+\mathrm{E}_{\mathscr{F}_r^{n-1}}(\Delta^i\big\langle\Delta^i,\nabla_\theta q(\mathbf{x}^i,\theta_r^{(i-1)^-},\Gamma_r)-\nabla_\theta q(\mathbf{x}^i,\widehat{\theta}_r^{(i-1)},\Gamma_r)\big\rangle$, here $\theta_r^{(i-1)^\pm}\in\left[\theta_r^{(i-1)},\theta_r^{(i-1)}\pm\beta^i\Delta^i\right]$. The function $q(\mathbf{x},\cdot,\Gamma)$ satisfies the Lipschitz condition and the function $F_r(\cdot,\Gamma)$ is strongly convex, then using valid for any $\varepsilon>0$ inequality $\|\widehat{\theta}_r^{(i-1)}-\theta^\star_r\|\le(\varepsilon^{-1}\beta^i+\varepsilon\beta^{i-1}\|\widehat{\theta}_r^{(i-1)}-\theta^\star_r\|^2)/2$, derive $-\alpha^i\left\langle\widehat{\theta}_r^{(i-1)}-\theta^\star_r,\beta^{i-1}\mathrm{E}_{\mathscr{F}_r^{n-1}}\Delta^i\left(y^i_{r,+}-y^i_{r,-}\right)\right\rangle=$ $=-2\alpha^i\left\langle\widehat{\theta}_r^{(i-1)}-\theta^\star_r,\nabla F_r(\widehat{\theta}_r^{(i-1)},\Gamma_r)\right\rangle-$ $-\alpha^i\big\langle\widehat{\theta}_r^{(i-1)}-\theta^\star_r,\mathrm{E}_{\mathscr{F}_r^{n-1}}\Delta^i\big\langle\Delta^i,\nabla_\theta q(\mathbf{x}^i,\theta_r^{(i-1)^+},\Gamma_r)-$ $-\nabla_\theta q(\mathbf{x}^i,\widehat{\theta}_r^{(i-1)},\Gamma_r)\big\rangle\big\rangle-\alpha^i\big\langle\widehat{\theta}_r^{(i-1)}-\theta^\star_r,\mathrm{E}_{\mathscr{F}_r^{n-1}}\Delta^i\big\langle\Delta^i,$ $\nabla_\theta q(\mathbf{x}^i,\theta_r^{(i-1)^-},\Gamma_r)-\nabla_\theta q(\mathbf{x}^i,\widehat{\theta}_r^{(i-1)},\Gamma_r)\big\rangle\big\rangle\le-2\mu\alpha^i\|\widehat{\theta}_r^{(i-1)}$ $-\theta^\star_r\|^2+Mm^{3/2}\alpha^i\beta^i\left(\varepsilon^{-1}\beta^i+\varepsilon\beta^{i-1}\|\widehat{\theta}_r^{(i-1)}-\theta^\star_r\|^2\right)$.

Hence, we have $\mathrm{E}_{\mathscr{F}_r^{n-1}}\|\widehat{\theta}^i_r-\theta^\star_r\|^2\le\|\widehat{\theta}_r^{(i-1)}-$ $-\theta^\star_r\|^2\left(1-2\mu\alpha^i+Mm^{3/2}\alpha^i\varepsilon+C_2\alpha^{i^2}/4\right)+$ $+Mm^{3/2}\alpha^i\beta^{i^2}\varepsilon^{-1}+\alpha^{i^2}/(4\beta^{i^2})\left(C_1\mathrm{E}_{\mathscr{F}_r^{n-1}}v_r^{i\,2}+o(\beta^{i^2})\right)$. Let ε small enough to $Mm^{3/2}\varepsilon<2\mu$ and i is sufficiently large. By the conditions of Theorem 1 we get $\mathrm{E}_{\mathscr{F}_r^{n-1}}\|\widehat{\theta}^i_r-\theta^\star_r\|^2\le\|\widehat{\theta}_r^{(i-1)}-\theta^\star_r\|^2\left(1-C_3\alpha^i\right)+$ $+C_4\left(\alpha^i\beta^{i2}+\alpha^{i^2}\beta^{i-2}(1+\mathrm{E}_{\mathscr{F}_r^{n-1}}v_r^{i^2}\right)$. Then by Robbins-Sigmund lemma [23] $\widehat{\theta}^i_r\to\theta^\star_r$ at $i\to\infty$ with probability one. $\mathrm{E}\left\{\|\widehat{\theta}^i_r-\theta^\star_r\|^2\right\}\le\mathrm{E}\left\{\|\widehat{\theta}_r^{(i-1)}-\theta^\star_r\|^2\right\}(1-C_5\alpha^n)+$ $+C_6(\alpha^n\beta^{n2}+\alpha^{n2}\beta^{n-2}(1+\sigma^{n2}))$. The convergence in the mean-square sense to a point $\theta^\star_r$ of sequence of estimates $\{\widehat{\theta}^i_r\}$ follows from [23].

(2) Let Θ is fixed. Consider estimates $\{\widehat{\Gamma}^n_l=\{\widehat{g}^n_{ij}\}\}$, $l\in 1..k$. Denote $\widehat{s}^n_{ij}=\{(\widehat{\theta}_l-\mathbf{x}^n)(\widehat{\theta}_l-\mathbf{x}^n)^{\mathrm{T}}\}_{ij}$ — i,j-th element of l-th scatter matrix, then $\widehat{g}^n_{ij}=(\lambda+\sum_{r=\lambda+1}^n\omega^r\widehat{s}^r_{ij})n^{-1}$. By the Theorem 1 conditions: $\lambda n^{-1}\to 0$, $\omega^r\to 1$, $\widehat{s}^r_{ij}n^{-1}\to\Gamma_l$ at $n\to\infty$. Thus, $\widehat{\Gamma}^n\xrightarrow{p}\Gamma^\star$ as an estimate obtained be the likelihood maximization.

(3) By assumptions about $\mathbf{q}(\Theta,\Gamma,\mathbf{x})$ and **(1)** and **(2)** we obtain the result of Theorem 1.

■

REFERENCES

[1] H. Robbins and S. Monro, "A stochastic approximation method," *Annal. Mathemat. Statist.*, pp. 400–407, 1951.

[2] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *Annal. Mathemat. Statist.*, vol. 23, no. 3, pp. 462–466, 1952.

[3] J. R. Blum, "Multidimensional stochastic approximation methods," *Annal. Mathemat. Statist.*, vol. 25, no. 4, pp. 737–744, 1954.

[4] J. C. Spall, "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE Trans. Autom. Control*, vol. 37, no. 3, pp. 332–341, Mar. 1992.

[5] O. Granichin, "Procedure of stochastic approximation with disturbances at the input," *Autom. and Remote Control*, vol. 53, no. 2, pp. 232–237, 1992.

[6] O. Granichin, "Randomized algorithms for stochastic approximation under arbitrary disturbances," *Autom. and Remote Control*, vol. 63, no. 2, pp. 209–219, 2002.

[7] A. Vakhitov, O. Granichin, L. Gurevich, "Algorithm for stochastic approximation with trial input perturbation in the nonstationary problem of optimization," *Autom. and Remote Control*, vol. 70, no. 11, pp. 1827–1835, 2009.

[8] O. Granichin, N. Amelina, "Simultaneous perturbation stochastic approximation for tracking under unknown but bounded disturbances," *IEEE Trans. Autom. Control*, vol. 60, no. 6, pp. 1653–1658, June 2015.

[9] N. Granichin, O. Granichina, S. Trapitsin, P. Proskurnikov, "Control of educational processes using SPSA," *In: Proc. of the International Conference on Computational Science and Computational Intelligence*, pp. 293–298, 2016.

[10] S. Lloyd, "Least Squares Quantization in PCM," *IEEE Trans. on Inform. Theory*, vol. 28, no. 2, pp. 129–136, 1982.

[11] M. Shindler, A. Wong and A. Meyerson, "Fast and Accurate *k*-means For Large Datasets," NIPS, 2011.

[12] D. Sculley, "Web Scale K-Means clustering," *Proceedings of the 19th WWW conf.*, 2010.

[13] D. Katselis, C. L. Beck and M. van der Schaar, "Ensemble Online Clustering through Decentralized Observations," *Proc. of the 53rd IEEE CDC*, pp. 910–915, 2014.

[14] L. Kaufman, P. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis." New York: John Wiley & Sons Inc., 1990.

[15] A. Dempster, N. Laird and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[16] C. M. Bishop, "Pattern Recognition and Machine Learning." Springer, 2006.

[17] M. Song and H. Wang, "Highly Efficient Incremental Estimation of GMM for Online Data Stream Clust.," *Proc. of SPIE III*, Mar. 2005.

[18] O. Granichin, O. Izmakova, "A randomized stochastic approximation algorithm for self-learning," *Autom. and Remote Control*, vol. 66, no. 8, pp. 1239–1248, 2005.

[19] P. J. Huber, "Robust Statistics." Wiley, 1981.

[20] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.

[21] O. Granichin, Z. Volkovich, and D. Toledano-Kitai, "Randomized Algorithms in Automatic Control and Data Mining." Springer, 2015.

[22] Y. LeCun, C. Cortes and C. Burges, http://yann.lecun.com/exdb/mnist/

[23] B. T. Polyak, "Introduction to Optimization." New York: Optimization Software, Publications Division, 1987.