Received December 5, 2017, accepted January 11, 2018, date of publication January 17, 2018, date of current version February 28, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2794587

A Dictionary Learning Based Automatic Modulation Classification Method

KEZHONG ZHANG[1], (Student Member, IEEE), EASTON LI XU[2,3], (Member, IEEE), ZHIYONG FENG[1], (Senior Member, IEEE), AND PING ZHANG[4], (Senior Member, IEEE)

[1]Key Laboratory of Universal Wireless Communications, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA
[3]School of Science and Engineering, The Chinese University of Hong Kong at Shenzhen, Shenzhen 518172, China
[4]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China

Corresponding author: Zhiyong Feng (fengzy@bupt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61631003 and Grant 61525101 and in part by the Shenzhen Fundamental Research Fund under Grant KQTD2015033114415450.

ABSTRACT As the process of identifying the modulation format of the received signal, automatic modulation classification (AMC) has various applications in spectrum monitoring and signal interception. In this paper, we propose a dictionary learning-based AMC framework, where a dictionary is trained using signals with known modulation formats and the modulation format of the target signal is determined by its sparse representation on the dictionary. We also design a dictionary learning algorithm called block coordinate descent dictionary learning (BCDL). Furthermore, we prove the convergence of BCDL and quantify its convergence speed in a closed form. Simulation results show that our proposed AMC scheme offers superior performance than the existing methods with low complexity.

INDEX TERMS Modulation classification, data driven, dictionary learning, block coordinate descent, sparse representation.

I. INTRODUCTION

Automatic Modulation Classification (AMC) is the process of identifying the modulation format of the received signal [1], [2], which has attracted substantial attention due to its wide applications in civilian and military fields, such as cognitive radio [3], spectrum monitoring [4], soft-define radio [5], and so forth. Various methods have been proposed for AMC in the past years, which could be roughly classified into two categories: Likelihood-Based (LB) methods and Feature-Based (FB) methods.

The LB methods are optimal in the Bayesian sense, which minimize the probability of the wrong classification. In particular, the Average Likelihood Ratio Test (ALRT) method [6] regards the unknown parameters as random variables and calculates the expectation of the likelihood function. The Generalized Likelihood Ratio Test (GLRT) proposed by Hameed *et al.* [7] estimates the unknown parameters by maximizing the likelihood function. Moreover, the Hybrid Likelihood Ratio Test (HLRT) and Quasi-HLRT (QHLRT) algorithms [7] can be viewed as a combination of ALRT and GLRT. However, the performance of LB methods could be severely degraded by model mismatches [2], since they usually suffer from high computational complexity.

On the other hand, FB methods identify the modulation format based on one or more statistical features derived from the training signals. Their performance may be inferior compared with the LB methods [2], but they generally have lower computational complexity. In [8], the maximum value of the spectral power density is utilized for AMC. Moreover, Swami and Sadler [9] adopted the High Order Cumulants (HOC) to identify the modulation format, which is more robust to the noise compared to methods in [8], while it performs poorly on high modulation order, such as M-QAM ($M > 4$).

Apart from the aforementioned methods, some researchers also used Machine Learning (ML) methods to tackle the AMC problem. Aslam *et al.* [10] utilized the Genetic Programming and K-Nearest Neighbors (GP-KNN) algorithms for AMC, which adopted the HOC as features. Han *et al.* [11] adopted Support Vector Machine (SVM) for modulation classification. However, both the GP-KNN and SVM based methods are vulnerable to the frequency and phase offset. Thus, their performance could be dramatically degraded by imperfect synchronization.

Dictionary learning methods have been highly successful utilized in various signal analysis and processing tasks [12]. The idea of dictionary learning is to represent a given

2169-3536 © 2018 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

signal as a linear combination of the codewords in a dictionary, with most coefficients taking zero or very small value [13]. There are various dictionary learning algorithms. In particular, the Method of Optimal Directions (MOD) proposed by Engan *et al.* [14] is a parallel update algorithm, which often needs a large amount of resources (such as memory, cache, and higher bit processor) to execute. Aharon *et al.* [15] proposed the K-SVD method which updates each column of the dictionary sequentially and empirical results prove its performance advantages over MOD [12]. Sahoo and Makur [16] proposed a novel dictionary method called Sequential Generalization of K-means (SGK) which is a combination of MOD and K-SVD. However, the convergences of MOD, K-SVD, and SGK cannot be guaranteed [12]- [16], which indicates that the performance of MOD, K-SVD, and SGK is unstable. To overcome this drawback, Xu and Yin [17] adopted a Block Proximal Gradient (BPG) descent-based method which has convergence guarantee. However, BPG is vulnerable to the frequency offset and phase noise.

In this paper, we first propose a framework adopting the dictionary learning method for AMC, in which we first train the dictionary and then classify the modulation formats with its sparse representation via a specially designed dictionary learning method called Block Coordinate descent Dictionary Learning (BCDL). Different from traditional dictionary learning methods, *e.g.*, MOD, K-SVD, and SGK [12]- [16], the convergence of BCDL can be guaranteed and we could further quantify its convergence speed. The simulation results show that the proposed BCDL method yields better classification accuracy than other methods mentioned previously with shorter training time.

The rest of the paper is organized as follows. We introduced the system model in Section II and Section III proposes a dictionary learning based AMC framework. We presented our dictionary learning method and its theoretical analysis of the convergence and computational complexity in Section IV. We conducted simulation experiments to compare the BCDL method with other existing methods in Section V and Section VI concludes this paper.

II. SYSTEM MODEL

In this section, we introduced the system model. Moreover, the performance of AMC methods will be degraded dramatically because of imperfect synchronization, for instance, frequency offset and initial phase. Therefore, we estimate these channel parameters to improve classification accuracy.

A. SIGNAL MODEL

Let $s(n)$ and $r(n)$ denote the transmitted and received signal at the n-th time slot, respectively. Their relationship is given by

$$r(n) = a_n e^{j(\omega_0 n+\theta_0)} s(n) + v(n), \tag{1}$$

where a_n is the channel gain, ω_0 is the frequency offset, θ_0 is the phase offset, and $v(n)$ is the complex Additive White Gaussian Noise (AWGN) with mean 0 and variance $2\sigma_v^2$.

Let M be the number of possible modulation formats and $\boldsymbol{\Theta} = [\Theta_1, \ldots, \Theta_M]$ denote the candidate set where Θ_m, $m = 1, \ldots, M$, is the m-th modulation format. For the signal generated by the m-th modulation format in the candidate set, AMC methods estimate the modulation format as $\hat{m} = f_{\text{AMC}}(r)$, where f_{AMC} denotes the classification function of the AMC classifier, and $\hat{m}$ is the modulation format estimated by the classifier according to received signal.

For $m = 1, 2, \ldots, M$, let $\mathcal{H}_m$ denote the hypothesis that the transmitted sequence $\mathbf{s} = [s(1), \ldots, s(N_0)]^T$ is generated from the m-th modulation format. In both the LB and FB methods, the probability of correct classification is calculated as

$$P_c = \frac{1}{M} \sum_{m=1}^{M} \mathbf{Pr}(\hat{m} = m|\mathcal{H}_m) \cdot \mathbf{Pr}(\mathcal{H}_m).$$

B. ORDER STATISTICS

The features used by dictionary learning based AMC are simply the real and imaginary parts of the received signals $r(1), \ldots, r(N_0)$ [11]. An assumption here is that each symbol in the constellation is generated with same probability. Therefore, the feature sequences generated by the same modulation format share the same order statistics. For $n = 1, \ldots, N_0$, let $\Re(r(n))$ and $\Im(r(n))$ denote the real and imaginary parts of $r(n)$, respectively. Let $N \triangleq 2N_0$, the extended signal of $\mathbf{r}$, denoted as $\tilde{\mathbf{r}} \in \mathbb{R}^N$, is given as

$$\tilde{\mathbf{r}} = [\Re(r(1)), \ldots, \Re(r(N_0)), \Im(r(1)), \ldots, \Im(r(N_0))]^T.$$

Let $\mathbf{y} \in \mathbb{R}^N$ denote the order statistics of $\mathbf{r}$. The n-th entry of $\mathbf{y}$, denoted as y_n, is the n-th smallest element of $\tilde{\mathbf{r}}$.

For a sequence $\mathbf{r} = [r(1), \ldots, r(N_0)]^T \in \mathbb{C}^{N_0}$, suppose that there are N real-valued features. Let $y_n \in \mathbb{R}$ be the n-th smallest feature of $\mathbf{r}$; thus we have $y_1 \leqslant \cdots \leqslant y_N$. Then the joint Probability Density Function (PDF) of $\mathbf{y} = [y_1, \ldots, y_N]^T$, is given by [11]

$$f_{\text{PDF}}(\mathbf{y}) = \begin{cases} N! \prod_{n=1}^{N} f_{\text{PDF}}(y_n), & \text{if } y_1 \leqslant \cdots \leqslant y_N, \\ 0, & \text{otherwise}, \end{cases}$$

where $f_{\text{PDF}}(y_n)$ denotes the PDF of y_n with $n = 1, \ldots, N$, for which we assume their individual distributions are independent. There are $N!$ possible sequences for y_n, $n = 1, \ldots, N$, with the same order statistics, and it is much easier to classify the modulation format with the help of order statistics as shown in [11].

C. CHANNEL PARAMETER ESTIMATION

Some channel parameters, like the frequency offset and initial phase, can affect the performance of AMC dramatically [18]. These parameters need to be estimated before modulation classification. We could estimate the frequency and initial phase via the methods proposed in [19] and [20], respectively. In particular, let $\arg(r(n))$ and $r^*(n)$ denote the argument and

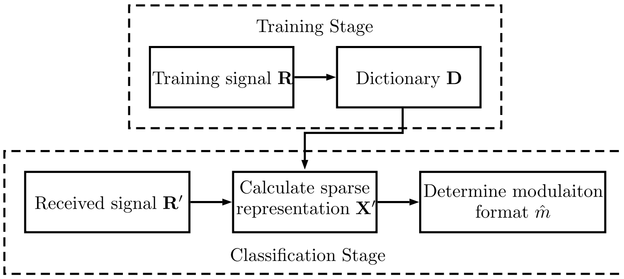


FIGURE 1. Framework of dictionary learning based AMC.

complex conjugate of $r(n)$, respectively. The frequency offset ω_0 is estimated as [19]

$$\omega_0 = -\frac{2}{N_0 - 2} \sum_{n=1}^{N_0/2} \arg\left(r(n) \cdot r^*(n + N_0/2 - 1)\right)$$

with N_0 assumed an even number. We then estimate the initial phase θ_0 as [20]

$$\theta_0 = \arctan\left(\frac{\sum_{n=1}^{N_0/2} \Im\left[\hat{r}(n) \cdot \hat{r}^*(2n)\right]}{\sum_{n=1}^{N_0/2} \Re\left[\hat{r}(n) \cdot \hat{r}^*(2n)\right]}\right),$$

where $\Im$ and $\Re$ denote the imaginary and real parts, respectively, and $\hat{r}(n) = r(n) \cdot e^{-j\omega_0 n}$. For simplicity, we assume the resulting estimation errors are negligible.

III. DICTIONARY LEARNING BASED MODULATION CLASSIFICATION

In this section, we present an AMC framework applying dictionary learning method. Since dictionary learning algorithms can effectively reduce the influence of noise [12], they can be used for AMC. Our dictionary learning based framework for modulation classification is illustrated in Fig. 1. In the training stage, we train the dictionary $\mathbf{D}$. Then, in the classification stage, we classify the received signals through its sparse representation via trained $\mathbf{D}$.

A. TRAINING STAGE

Let $\mathbf{R} = [\mathbf{r}_1, \ldots, \mathbf{r}_L] \in \mathbb{R}^{N_0 \times L}$ be the L training signal sequences, and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_L] \in \mathbb{R}^{N \times L}$ denote the matrix whose l-th column $\mathbf{y}_l$ is the order statistics of $\mathbf{r}_l$, for $1 \leqslant l \leqslant L$. Note that $L \gg M$, such that for each modulation format in the candidate set, there are a sufficient number of observations. According to [13], the order statistics matrix $\mathbf{Y}$ is assumed to be represented as

$$\mathbf{Y} = \mathbf{D}\mathbf{X} + \mathbf{W},$$

where $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_M] \in \mathbb{R}^{N \times M}$ is the dictionary, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_L] \in \mathbb{R}^{M \times L}$ denotes the sparse representation of $\mathbf{Y}$, and $\mathbf{W} \in \mathbb{R}^{N \times L}$ is the approximation error matrix.

There are two main stages in the modulation format classification. First, in the training stage, we use the training signals generated from each modulation format, *i.e.*, $\mathbf{R}$, to train the dictionary $\mathbf{D}$. Then, in the classification stage, we keep the trained dictionary $\mathbf{D}$ constant and identify the modulation format via the sparse representation of the corresponding signal. In the training stage, each dictionary atom $\mathbf{d}_m$, $m = 1, \ldots, M$, is initialized to be the order statistics of a sequence generated from the m-th modulation, and then the dictionary $\mathbf{D}$ is trained by solving the following optimization problem [21]

Problem I:

$$\underset{\mathbf{D},\mathbf{X}}{\text{minimize}}\, f(\mathbf{D}, \mathbf{X}) = \left\|\mathbf{Y} - \mathbf{D}\mathbf{X}\right\|_F^2 + \lambda_1 \left\|\mathbf{X}\right\|_{1,1} + \lambda_2 \left\|\mathbf{X}\right\|_F^2$$
$$\text{subject to } \left\|\mathbf{d}_m\right\|_2 \leqslant 1, \quad m = 1, \ldots, M,$$

where $\|\cdot\|_{1,1}$ and $\|\cdot\|_F$ are the matrix $L_{1,1}$-norm and Frobenius norm [21], respectively; $\|\cdot\|_2$ denotes the vector L_2-norm; and $\lambda_1, \lambda_2 \geqslant 0$ are the tradeoff parameters between two penalty terms. We adopt the objective function with both $\|\cdot\|_{1,1}$ and $\|\cdot\|_F$ terms (called elastic net [21]), since the $\|\cdot\|_{1,1}$ norm can lead to a sparse solution for identifying the modulation format, and the Frobenius norm helps improve the numerical stability at low SNRs [12].

B. CLASSIFICATION STAGE

Let $\mathbf{R}' = [\mathbf{r}'_1, \ldots, \mathbf{r}'_{L'}] \in \mathbb{R}^{N_0 \times L'}$ be the L' received signal sequences generated from one unknown modulation format in the candidate set $\mathbf{\Theta}$. Let $\mathbf{Y}' = [\mathbf{y}'_1, \ldots, \mathbf{y}'_{L'}]$ denote the order statistics of test signals $\mathbf{r}'_1, \ldots, \mathbf{r}'_{L'}$. In the classification stage, we keep the dictionary $\mathbf{D}$ obtained in the training stage constant, and find the sparse representation $\mathbf{X}' = [\mathbf{x}'_1, \ldots, \mathbf{x}'_{L'}]$ via Orthogonal Matching Pursuit (OMP) [22]. If $\mathbf{y}'_l$, $l = 1, \ldots, L'$, is generated by the m-th modulation, the m-th entry in $\mathbf{x}'_l$ is expected to have a larger magnitude than the other entries. By aggregating the contributions of the test signals, we determine the modulation of the test signals as

$$\hat{m} = \arg\max_m \left\|\hat{\mathbf{x}}'_m\right\|_2^2, \tag{2}$$

where $\hat{\mathbf{x}}'^T_m$ is the m-th row of $\mathbf{X}'$ for $1 \leqslant m \leqslant M$.

Algorithm 1 Dictionary-Learning Based AMC

Input Training signals $\mathbf{R}$ and received signals $\mathbf{R}'$.
Output Modulation format $\hat{m}$.
begin *Training Stage*
  Extract the order statistics $\mathbf{Y}$ of $\mathbf{R}$.
  Train the dictionary $\mathbf{D}$ via dictionary learning method.
end
begin *Classification Stage*
  Extract the order statistics $\mathbf{Y}'$ of $\mathbf{R}'$.
  Calculate the sparse representation $\mathbf{X}'$ of $\mathbf{Y}'$.
  Determine the modulation format of $\mathbf{R}'$ by (2).
end

The dictionary learning based AMC framework is summarized in Algorithm 1.

IV. BLOCK COORDINATE DESCENT DICTIONARY LEARNING

In this section, we introduce our proposed method, called Block Coordinate descent Dictionary Learning (BCDL). Then we prove the convergence of BCDL, where we could further quantify its convergence speed. Furthermore, we analysis the computational complexity of our proposed BCDL.

A. TRAINING DICTIONARY

Let $\mathbf{X}_{(k)}$ and $\mathbf{D}_{(k)} = [\mathbf{d}_{(k),1}, \ldots, \mathbf{d}_{(k),M}]$ be the values of $\mathbf{X}$ and $\mathbf{D}$ generated in the k-th iteration, respectively. The m-th column of initialized dictionary $\mathbf{D}_{(0)}$ is set to be the order statistics of noiseless signals generated by the m-th modulation format. We initialize the coefficient $\mathbf{X}_{(0)}$ such that

$$x_{(0),m,l} = \begin{cases} \dfrac{\|\mathbf{y}_l\|_2}{\|\mathbf{d}_{(0),m}\|_2} & \text{if } \mathbf{y}_l \text{ is generated by the} \\ & \quad m\text{-th modulation format}, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

where $x_{(0),m,l}$ denotes the (m, l)-th entry of $\mathbf{X}_{(0)}$.

Let $\Psi_1(\mathbf{D}, \mathbf{X}) = \left\|\mathbf{Y} - \mathbf{D}\mathbf{X}\right\|_F^2 + \lambda_2 \left\|\mathbf{X}\right\|_F^2$ denote the smooth part of the objective function in Problem I. Using the proximal map notation defined in [23], we update the sparse representation $\mathbf{X}$ and dictionary $\mathbf{D}$ in the k-th iteration as follows

$$\mathbf{X}_{(k)} \in \operatorname{prox}_{t_{\mathbf{X}}}^{\|\cdot\|_{1,1}} \left(\mathbf{X}_{(k-1)} - \frac{1}{t_{\mathbf{X}}} \nabla_{\mathbf{X}} \Psi_1(\mathbf{D}_{(k-1)}, \mathbf{X}_{(k-1)})\right), \tag{4a}$$

$$\mathbf{d}_{(k),m} \in \operatorname{prox}_{t_{\mathbf{d}_m}}^{\mathcal{D}} \left(\mathbf{d}_{(k-1),m} - \frac{1}{t_{\mathbf{d}_m}} \nabla_{\mathbf{d}_m} \Psi_1(\mathbf{D}_{(k-1)}, \mathbf{X}_{(k-1)})\right). \tag{4b}$$

where $\mathcal{D} = \{\mathbf{D} : \|\mathbf{d}_m\|_2 \leqslant 1, m = 1, \ldots, M\}$, $t_{\mathbf{X}} > 0$ and $t_{\mathbf{d}_m} > 0$ $(m = 1, \ldots, M)$, are step sizes. Let $\hat{\mathbf{x}}_{(k-1),m}^T$ denote the m-th row of $\mathbf{X}_{(k-1)}$, and for $m = 1. \ldots, M$, we set

$$t_{\mathbf{X}} = 2 \left\|\mathbf{D}_{(k-1)}^T \cdot \mathbf{D}_{(k-1)} + \lambda_2 \mathbf{I}\right\|_F, \tag{5a}$$

$$t_{\mathbf{d}_m} = 2\hat{\mathbf{x}}_{(k-1),m}^T \cdot \hat{\mathbf{x}}_{(k-1),m}, \tag{5b}$$

where $\mathbf{I}$ denotes the identity matrix. In the following, we will prove that $t_{\mathbf{X}}$ and $t_{\mathbf{d}_m}$, $(m = 1, \ldots, M)$, are Lipschitz constants [24], respectively.

For $k \geqslant 1$, (4a) and (4b) can be given as (6a) and (6b), as shown at the top of the next page, respectively, where $\mathbf{tr}(\cdot)$ denotes the trace. In the following, we further derive the closed-forms of $\mathbf{X}_{(k)}$ and $\mathbf{d}_{(k),m}$ in (4a) and (4b), for $m = 1, \ldots, M$. Let $\partial_{\mathbf{U}}$ denote the Frechét subdifferential [24] of $\mathbf{U}$. Setting the Frechét subdifferential of the objective function of (6a) with respect to $\mathbf{U}$, that is,

$$\partial_{\mathbf{U}} \Big\{ \lambda_1 \left\|\mathbf{U}\right\|_{1,1} + \frac{t_{\mathbf{X}}}{2} \cdot \left\|\mathbf{U} - \mathbf{X}_{(k-1)}\right\|_F^2 + \mathbf{tr}\left[(\mathbf{U} - \mathbf{X}_{(k-1)})^T \cdot \nabla_{\mathbf{X}} \Psi_1(\mathbf{D}_{(k-1)}, \mathbf{X}_{(k-1)})\right] \Big\},$$

equal to $\mathbf{0}$ yields

$$\lambda_1 \partial_{\mathbf{U}} \left\|\mathbf{U}\right\|_{1,1} + t_{\mathbf{X}} \mathbf{U} = -2\mathbf{D}_{(k-1)}^T (\mathbf{D}_{(k-1)} \mathbf{X}_{(k-1)} - \mathbf{Y}) + (t_{\mathbf{X}} - 2\lambda_2) \mathbf{X}_{(k-1)}.$$

Let $\mathcal{S}(z, \alpha)$ denote the soft-thresholding operator with $\alpha \geqslant 0$, given as $\mathcal{S}(z, \alpha) = \max(z - \alpha, 0) + \min(z + \alpha, 0)$, Setting the Frechét subdifferential of the objective function of (6a) with respect to $\mathbf{X}$ equal to $\mathbf{0}$, the (m, l)-th entry of the sparse representation $\mathbf{X}_{(k)}$, denoted as $x_{(k),m,l}$, is updated as

$$x_{(k),m,l} = \frac{1}{t_{\mathbf{X}}} \cdot \mathcal{S}(b_{(k-1),m,l}, \lambda_1), \tag{7}$$

where $b_{(k-1),m,l}$ is the (m, l)-th entry of $\mathbf{B}_{(k-1)}$ given as

$$\mathbf{B}_{(k-1)} = (t_{\mathbf{X}} - 2\lambda_2)\mathbf{X}_{(k-1)} - 2\mathbf{D}_{(k-1)}^T (\mathbf{D}_{(k-1)} \mathbf{X}_{(k-1)} - \mathbf{Y}). \tag{8}$$

For any vector $\mathbf{a}$, let $\mathcal{P}_{\mathcal{B}}(\mathbf{a})$ be the Euclidean projection to the closed unit ball $\mathcal{B} \triangleq \{\mathbf{e} : \|\mathbf{e}\|_2 \leqslant 1\}$. We have

$$\mathcal{P}_{\mathcal{B}}(\mathbf{a}) = \frac{\mathbf{a}}{\max\left(1, \|\mathbf{a}\|_2\right)}. \tag{9}$$

Then letting the Frechét subdifferential of the objective function of (6b) with respect to $\mathbf{v}_m$ for $m = 1, \ldots, M$, equal to $\mathbf{0}$, the dictionary atom $\mathbf{d}_m$ is updated by

$$\mathbf{d}_{(k),m} = \mathcal{P}_{\mathcal{B}} \left(\mathbf{d}_{(k-1),m} + \frac{2}{t_{\mathbf{d}_m}} \times \left(\mathbf{Y} - \mathbf{D}_{(k-1)} \mathbf{X}_{(k-1)}\right) \hat{\mathbf{x}}_{(k-1),m} \right). \tag{10}$$

The stopping criterion in the training stage of our proposed BCDL is

$$\left| f\left((\mathbf{D}_{(k+1)}, \mathbf{X}_{(k+1)})\right) - f\left((\mathbf{D}_{(k)}, \mathbf{X}_{(k)})\right) \right| \leqslant \epsilon,$$

where $\epsilon > 0$ denotes the threshold.

Algorithm 2 BCDL Algorithm

Input Training signals $\mathbf{R}$ and received signals $\mathbf{R}'$.
Output Dictionary $\mathbf{D}$.
Initialize $\mathbf{D}_{(0)}$, $\mathbf{X}_{(0)}$, threshold ϵ, and K.
Extract the order statistics $\mathbf{Y}$ of $\mathbf{R}$.
for $k = 1$ **to** K **do**
  Set step sizes $t_{\mathbf{X}}$ and $t_{\mathbf{d}_m}$, $m=1,\ldots,M$, by (5a) and (5b).
  Calculate $\mathbf{B}_{(k-1)}$ by (8).
  Update the sparse representation $\mathbf{X}_{(k)}$ by (7).
  for $m = 1$ **to** M **do**
    Update $\mathbf{d}_{m,(k)}$ by (10).
  end
  if $\left|f(\mathbf{D}_{(k)}, \mathbf{X}_{(k)}) - f(\mathbf{D}_{(k-1)}, \mathbf{X}_{(k-1)})\right| \leqslant \epsilon$ **then**
    break.
  end
end
Set $\mathbf{D} := \mathbf{D}_{(k)}$.

The BCDL algorithm is summarized in Algorithm 2.

$$\mathbf{X}_{(k)} \in \arg\min_{\mathbf{U}} \frac{t_{\mathbf{X}}}{2} \left\|\mathbf{U} - \mathbf{X}_{(k-1)}\right\|_F^2 + \lambda_1 \left\|\mathbf{U}\right\|_{1,1} + \mathbf{tr}\left[\left(\mathbf{U} - \mathbf{X}_{(k-1)}\right)^T \cdot \nabla_{\mathbf{X}} \Psi_1 \left(\mathbf{D}_{(k-1)}, \mathbf{X}_{(k-1)}\right)\right], \tag{6a}$$

$$\mathbf{d}_{(k),m} \in \arg\min_{\mathbf{v}_m \in \mathcal{D}} \frac{t_{\mathbf{d}_m}}{2} \left\|\mathbf{v}_m - \mathbf{d}_{(k-1),m}\right\|_2^2 + \left(\mathbf{v}_m - \mathbf{d}_{(k-1),m}\right)^T \cdot \nabla_{\mathbf{d}_m} \Psi_1 \left(\mathbf{D}_{(k-1)}, \mathbf{X}_{(k-1)}\right), \quad m = 1, \ldots, M. \tag{6b}$$

B. CONVERGENCE OF BCDL

Different from other dictionary learning algorithms, *e.g.*, MOD, K-SVD, and SGK, the proposed BCDL can guarantee the convergence, where we could further quantify its convergence speed.

Theorem 1: The sequence $\{(\mathbf{D}_{(k)}, \mathbf{X}_{(k)})\}_{k=1}^{\infty}$ *generated by BCDL is a Cauchy sequence and converges to a point.*

Proof: Note that Problem I is equivalent to

$$\underset{\mathbf{D},\mathbf{X}}{\text{minimize}}\, \tilde{f}(\mathbf{D}, \mathbf{X}) = \left\|\mathbf{Y} - \mathbf{D}\mathbf{X}\right\|_F^2 + \lambda_1 \left\|\mathbf{X}\right\|_{1,1} + \lambda_2 \left\|\mathbf{X}\right\|_F^2 + \delta_{\mathcal{D}}(\mathbf{D}),$$

where $\delta_{\mathcal{D}}$ denotes an indicator function, defined as

$$\delta_{\mathcal{D}}(\mathbf{D}) = \begin{cases} 0 & \text{if } \mathbf{D} \in \mathcal{D}, \\ +\infty & \text{otherwise.} \end{cases}$$

According to [23], $\tilde{f}(\mathbf{D}, \mathbf{X})$ is a semi-algebraic function satisfying the Kurdyka-Łojasiewicz property. Furthermore, (4b) is equivalent to (10) [25]. In addition, the sequence $\left\{\mathbf{D}_{(k)}\right\}_{k=1}^{\infty}$ is in the bounded set $\mathcal{D}$. Tradeoff parameters, $\lambda_1, \lambda_2 > 0$ make $\left\{\mathbf{X}_{(k)}\right\}_{k=1}^{\infty}$ bounded, since otherwise the sequence $\left\{\mathbf{D}_{(k)}\right\}_{k=1}^{\infty}$ and the objective function of Problem I will blow up. Thus, the sequence $\left\{\left(\mathbf{D}_{(k)}, \mathbf{X}_{(k)}\right)\right\}_{k=1}^{\infty}$ is bounded.

Furthermore, in the following we prove that $t_{\mathbf{X}}$ and $t_{\mathbf{d}_m}$, ($m = 1, \ldots, M$), are Lipschitz constants, respectively. Note that for any $\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{R}^{M\times L}$, according to Cauchy-Schwarz inequality, we have

$$\begin{aligned} &\left\|\nabla_{\mathbf{X}}\Psi_1(\mathbf{D}, \mathbf{X}_1) - \nabla_{\mathbf{X}}\Psi_1(\mathbf{D}, \mathbf{X}_2)\right\|_F \\ &\quad = 2\left\|\left(\mathbf{D}^T\mathbf{D} + \lambda_2\mathbf{I}\right)(\mathbf{X}_1 - \mathbf{X}_2)\right\|_F \\ &\quad \leqslant 2\left\|\mathbf{D}^T\mathbf{D} + \lambda_2\mathbf{I}\right\|_F \cdot \left\|\mathbf{X}_1 - \mathbf{X}_2\right\|_F. \end{aligned}$$

For any $\mathbf{d}_m, \mathbf{d}'_m \in \mathbb{R}^N$, $1 \leqslant m \leqslant M$, we have

$$\begin{aligned} &\left\|\nabla_{\mathbf{d}_m}\Psi_1(\mathbf{d}_m, \ldots, \mathbf{X}) - \nabla_{\mathbf{d}_m}\Psi_1\left(\mathbf{d}'_m, \ldots, \mathbf{X}\right)\right\|_F \\ &\quad = 2\left\|\left(\mathbf{d}_m - \mathbf{d}'_m\right)\hat{\mathbf{x}}_m^T\hat{\mathbf{x}}_m\right\|_F \\ &\quad = 2\hat{\mathbf{x}}_m^T\hat{\mathbf{x}}_m\left\|\mathbf{d}_m - \mathbf{d}'_m\right\|_F. \end{aligned}$$

From above, we can find that $t_{\mathbf{X}}$ and $t_{\mathbf{d}_m}$, $m = 1, \ldots, M$, are Lipschitz constants, respectively. Moreover, since the sequence $\left\{\left(\mathbf{D}_{(k)}, \mathbf{X}_{(k)}\right)\right\}_{k=1}^{\infty}$ generated by BCDL is bounded, the Lipschitz constants specified in (5a) and (5b) must be upper-bounded. Therefore, according to [24, Th. 2.8] and [26, Corollary 12], Theorem 1 can be guaranteed. □

Furthermore, we prove the following theorem with the help of Theorem 1.

Theorem 2: The sequence $\{(\mathbf{D}_{(k)}, \mathbf{X}_{(k)})\}_{k=1}^{\infty}$ *generated by BCDL converges to a point (denoted by* $(\mathbf{D}^*, \mathbf{X}^*)$*) and there exist* $\omega, \zeta > 0$ *such that*

$$\left\|(\mathbf{D}_{(k)}, \mathbf{X}_{(k)}) - (\mathbf{D}^*, \mathbf{X}^*)\right\| \leqslant \omega k^{-\zeta}. \tag{11}$$

Proof: Let $\boldsymbol{\Xi} = (\mathbf{D}, \mathbf{X}) \in \mathbb{R}^{N\times M} \times \mathbb{R}^{M\times L}$ and a special norm in $\mathbb{R}^{N\times M} \times \mathbb{R}^{M\times L}$ is defined as

$$\left\|\boldsymbol{\Xi}\right\|_A = \left\|(\mathbf{D}, \mathbf{X})\right\|_A = \sqrt{\left\|\mathbf{D}\right\|_F^2 + \left\|\mathbf{X}\right\|_F^2}.$$

Let $\boldsymbol{\Xi}^* = (\mathbf{D}^*, \mathbf{X}^*)$ denote the optimal point of $\{\boldsymbol{\Xi}_{(k)}\}_{k=1}^{\infty} = \{(\mathbf{D}_{(k)}, \mathbf{X}_{(k)})\}_{k=1}^{\infty}$ generated by BCDL. Since the convergence of BCDL is guaranteed by Theorem 1, we have $\boldsymbol{\Xi}^* = \lim_{k\to\infty} \boldsymbol{\Xi}_{(k)}$. According to [27], there exists a Łojasiewicz exponent $\theta \in (\frac{1}{2}, 1)$ for a semi-algebraic function. That is, there exist $\epsilon, k_0 > 0$, such that for $k > k_0$,

$$\left\|\boldsymbol{\Xi}_{(k)} - \boldsymbol{\Xi}^*\right\|_A < \epsilon,$$

and for any $\hat{\boldsymbol{\Xi}}_{(k)} \in \partial_{\boldsymbol{\Xi}}\tilde{f}(\boldsymbol{\Xi}_{(k)})$, there exist $C > 0$ and $\theta \in (\frac{1}{2}, 1)$, such that

$$\left|\tilde{f}(\boldsymbol{\Xi}_{(k)}) - \tilde{f}(\boldsymbol{\Xi}^*)\right|^{\theta} \leqslant C\left\|\hat{\boldsymbol{\Xi}}_{(k)}\right\|_A. \tag{12}$$

By [27, Th. 5], for $\theta \in (\frac{1}{2}, 1)$, there exists $\omega \geqslant 0$ such that

$$\left\|(\mathbf{D}_{(k)}, \mathbf{X}_{(k)}) - (\mathbf{D}^*, \mathbf{X}^*)\right\|_A \leqslant \omega k^{-\frac{1-\theta}{2\theta-1}}. \tag{13}$$

The proof is completed by choosing $\zeta = \frac{1-\theta}{2\theta-1} > 0$. □

We will further conduct simulation experiments to illustrate the convergence of BCDL and compare with other dictionary learning methods in Section V.

C. COMPLEXITY ANALYSIS OF BCDL

To quantify the complexity of BCDL, we first estimate the number of required iterations and then derive the computational complexity of BCDL via Theorem 2.

1) NUMBER OF ITERATIONS

The stopping criterion of BCDL is

$$\left|f\left((\mathbf{D}_{(k)}, \mathbf{X}_{(k)})\right) - f\left((\mathbf{D}^*, \mathbf{X}^*)\right)\right| \leqslant \epsilon,$$

where ϵ denotes the threshold. By the functional Łojasiewicz inequality [27], we have

$$C'\|(\mathbf{D}_{(k)}, \mathbf{X}_{(k)}) - (\mathbf{D}^*, \mathbf{X}^*)\|^{\frac{1}{1-\theta}} \leqslant \epsilon, \tag{14}$$

for some constant $C' > 0$, where θ is the Łojasiewicz exponent in (12). Substituting (13) into (14), we conclude that the number of iterations holds as

$$k \geqslant \left[\omega \cdot \left(C'/\epsilon\right)^{1-\theta}\right]^{\frac{1}{\zeta}} = \omega^{\frac{1}{\zeta}} \cdot \left(C'/\epsilon\right)^{\frac{1}{1+2\zeta}}, \tag{15}$$

where $\zeta = \frac{1-\theta}{2\theta-1}$. Choosing the lower bound of k which holds inequality (15), the number of iterations of BCDL given $\zeta > 0$ is

$$K_{\text{BCDL}} = \omega^{\frac{1}{\zeta}} \cdot \left(C'/\epsilon\right)^{\frac{1}{1+2\zeta}}. \tag{16}$$

2) COMPUTATIONAL COMPLEXITY OF BCDL

First, we need $\mathcal{T}_{t_{\mathbf{X}}} = O(NM^2)$ and $\mathcal{T}_{t_{\mathbf{d}_m}} = O(L)$ FLoating Point Operations (FLOPs) to update the step sizes $t_{\mathbf{X}} = 2\|\mathbf{D}_{(k)}^T\mathbf{D}_{(k)} + \lambda_2\mathbf{I}\|_F$ and $t_{\mathbf{d}_m} = 2\hat{\mathbf{x}}_{(k),m}^T\hat{\mathbf{x}}_{(k),m}$, $m = 1, \ldots, M$, respectively. According to (7), it requires $\mathcal{T}_{\mathbf{X}} = O(NML)$ FLOPs to update $\mathbf{X}$. By (10), it requires $\mathcal{T}_{\mathbf{d}_1} = O(NML)$ and $\mathcal{T}_{\mathbf{d}_m} = O(NL)$ FLOPs to update $\mathbf{d}_1$ and $\mathbf{d}_m$, $m = 2, \ldots, M$, respectively. Therefore, the total number of FLOPs we need to update the dictionary set $\mathbf{D}$ and the sparse representation $\mathbf{X}$ in each iteration is

$$\begin{aligned}\mathcal{T}_1 &= \mathcal{T}_{t_{\mathbf{X}}} + M\mathcal{T}_{t_{\mathbf{d}_m}} + \mathcal{T}_{\mathbf{X}} + \mathcal{T}_{\mathbf{d}_1} + (M-1)\mathcal{T}_{\mathbf{d}_m} \\ &= O(NML + NM^2).\end{aligned}$$

Therefore, the computational complexity of BCDL is

$$\mathcal{T}_{\text{BCDL}} = \mathcal{T}_1 \cdot K_{\text{BCDL}} = O((NML + NM^2) \cdot \epsilon^{-\frac{1}{1+2\zeta}}),$$

where $\zeta > 0$.

Let N_I denote the number of iterations for training dictionary, the computational complexity of BCDL can also be given as

$$\mathcal{T}_{\text{BCDL}} = \mathcal{T}_1 \cdot N_I = O\left(NN_IML + NN_IM^2\right). \tag{17}$$

In practice, $L \gg M, N_I$ [14]- [16]. Therefore, we have $\mathcal{T}_{\text{BCDL}} = O\left(NML\epsilon^{-\frac{1}{1+2\zeta}}\right)$ or $\mathcal{T}_{\text{BCDL}} = O\left(NN_IML\right)$.

3) SUMMARY OF THE COMPLEXITY OF DICTIONARY-LEARNING BASED METHODS

We calculate the computational complexity of other dictionary-learning algorithms in the Appendix. The computation complexities of the aforementioned dictionary learning methods are summarized in Table 1. We can see that, for the same number of iterations, *i.e.*, $N_I = N_I' = N_I'' = N_I'''$, BCDL has a lower computational complexity than other dictionary learning based classifiers, *e.g.*, K-SVD, SGK, and BPG. If $\zeta = \zeta'$, the computational complexity of BCDL is lower than that of BPG. We will further make simulation experiments to compare their computational complexities in Section V.

V. EXPERIMENT RESULTS

In this section, a variety of experiments are conducted to validate the performance of the proposed method. We consider the following modulation format candidate set $\mathbf{\Theta} = \{\text{QPSK, 8PSK, 8QAM, 16QAM, QASK, 8ASK}\}$, since we can investigate the performance of both the inter-class and intra-class classification [9]. All results are based on 10000 Monte Carlo trials.

We choose $N = 512$, $L = 1024$, and $L' = 10$. The maximum number of iterations for dictionary learning based algorithms, denoted as K, is set as 60. For traditional AMC methods, we adopt the HOC based classifier proposed in [9], and GLRT in [2]. Since GLRT performs better than ALRT and HLRT [2], we analyze the performance of GLRT, instead of ALRT and HLRT. On the other hand, for machine learning based methods, we utilize SVM [11] and GP-KNN [10]. Furthermore, the strategy of SVM we used in the comparison is One Versus One (OVO), since the classification accuracy of OVO is higher than that of One Versus All (OVA) [28]. For the dictionary learning based methods, we choose K-SVD [15], SGK [16], and BPG [17], since they are reported to perform better than MOD [12]. We set a small random initial phase offset for the received signals.

TABLE 1. Summary of the computation complexities of BCDL, BPG, SGK, and K-SVD.

Method	Computational Complexity
BCDL	$O(NML \cdot \epsilon^{-\frac{1}{1+2\zeta}})$ or $O\left(NN_IML\right)$
BPG	$O(M(L+N)L \cdot \epsilon^{-\frac{1}{1+2\zeta'}})$ or $O(N_I'M(L+N)L)$
SGK	$O(NN_I''M^2L)$
K-SVD	$O(L^3 + NN_I'''L^2 + N^2N_I'''L)$

A. COMPARISON WITH EXISTING METHODS

In this subsection, we compare the BCDL with the existing AMC methods, that is, traditional and machine-learning based algorithms. Following [9], we adopt seven different cumulants for HOC: C_{21}, C_{40}, C_{41}, C_{42}, C_{60}, C_{63}, C_{80}, which are defined in [2]. The probability of correct classification of BCDL algorithm and traditional methods, that is, HOC [9], and GLRT [2], over SNR on the candidate set are shown in Fig. 2. The classification accuracy of BCDL reaches 100% even at SNR of 0dB, while HOC and GLRT achieve 100% at SNR of 12dB and 20dB, respectively. At low SNRs, BCDL performs the best while HOC and GLRT produce much significantly worse performance. HOC performs better than GLRT, since we set a small random initial phase offset for the received signals. Therefore, BCDL achieves higher accuracy over SNR than traditional methods.

Then we compare our proposed method with the machine-learning based methods. Fig. 3 illustrates the classification accuracies over SNR of BCDL, SVM [11], and GP-KNN [10] on the candidate set. At the SNR of 0dB, BCDL, SVM, and GP-KNN achieve 100%, 63.7%, and 19.2%, respectively. BCDL performs better than machine learning based algorithms, which indicates that BCDL is more suitable for AMC over SNR than the machine learning based methods.

From Fig. 2 and Fig. 3, we find that BCDL is quite robust over SNR, which can be utilized in low SNR regime.

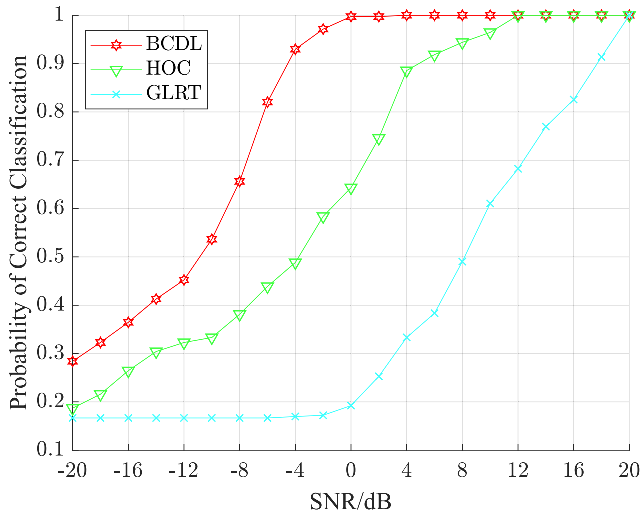


FIGURE 2. Classification accuracy of BCDL and traditional AMC methods.

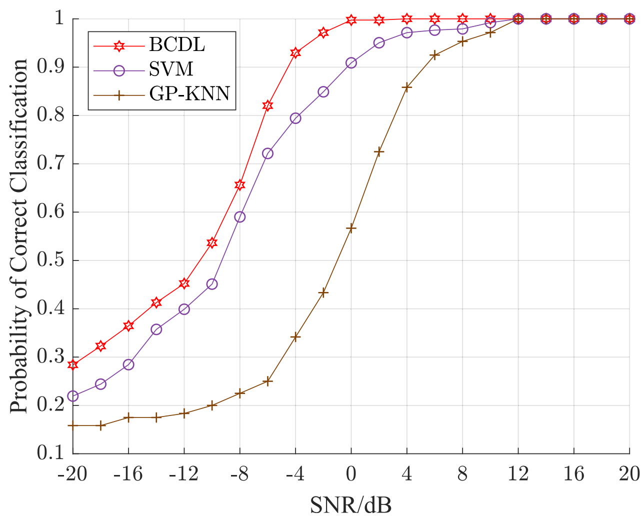


FIGURE 3. Classification accuracy of BCDL and machine learning based AMC methods.

B. COMPARISON WITH OTHER DICTIONARY LEARNING BASED ALGORITHMS

In this subsection, we compare the performance of the proposed method with other dictionary-learning based methods. The classification accuracies over SNR on the candidate set are illustrated in Fig. 4. We can find that BCDL performs better than other dictionary-learning based algorithms, which indicates that BCDL is more suitable for AMC over SNR than other dictionary-learning based algorithms. Thus, BCDL is quite robust over SNR and we can adopt BCDL in low SNR regime.

Convergence speed is an important criteria for evaluating dictionary learning algorithms. To measure the convergence speed of dictionary learning algorithms, the metrics we observe, denoted as ϱ_k, and τ_k, are given as

$$\varrho_k = \frac{1}{NM}\left\|\mathbf{D}_{(k)} - \mathbf{D}_{(k-1)}\right\|_F^2,$$
$$\tau_k = \frac{1}{ML}\left\|\mathbf{X}_{(k)} - \mathbf{X}_{(k-1)}\right\|_F^2,$$

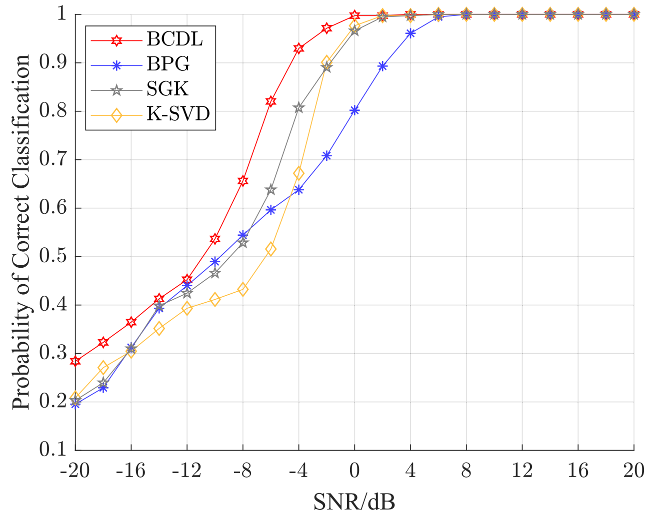


FIGURE 4. Classification accuracy of BCDL and other dictionary learning based AMC methods.

where $\mathbf{D}_{(k)} \in \mathbb{R}^{N\times M}$ and $\mathbf{X}_{(k)} \in \mathbb{R}^{M\times L}$ are the dictionary and sparse representation in the k-th iteration, respectively. We can see that a faster fall of ϱ_k and τ_k indicates faster convergence of the dictionary and sparse representation, respectively. Fig. 5(a) and Fig. 5(b) illustrate the convergence speed of dictionary and sparse representation on the candidate set for SNR $=$ 20dB, respectively. The ϱ_k and τ_k generated by BCDL fall faster than other methods, which shows that both the dictionary and sparse representation generated by BCDL converge faster than other dictionary learning based methods. The simulation results verify the theorems in Section IV that BCDL can guarantee its convergence. Therefore, BCDL can obtain more stable results than other dictionary learning based methods.

In this experiment, we analyze the training time of dictionary learning methods, since shorter training time indicates lower computational complexity. We adopt SNR $=$ 20dB and the results are summarized in Fig. 6. It can be seen that BCDL costs a shorter training time than other dictionary-learning based methods, which verifies our computational complexity analysis in Section III and the Appendix.

C. COMPARISON OF ROBUSTNESS

In this subsection, we compare the robustness of different AMC methods over imperfect synchronization. As mentioned in Section II, there are two main categories of imperfect synchronization, that is, frequency offset and phase offset. We first show how performance is influenced by the frequency offset, which causes the constellation to rotate. Note that although we conduct preprocessing to eliminate the frequency offset in Section II, there may still exist Residual Frequency Offset (RFO) in practice, which could be modeled as a zero-mean Gaussian random variable with variance $\sigma_{\omega_0}^2$ [29]. Fig. 7 illustrates the classification accuracies versus σ_{ω_0} on the candidate set for SNR=20dB, with σ_{ω_0} ranging from 0 to 1.2×10^{-4}. We see that the performance

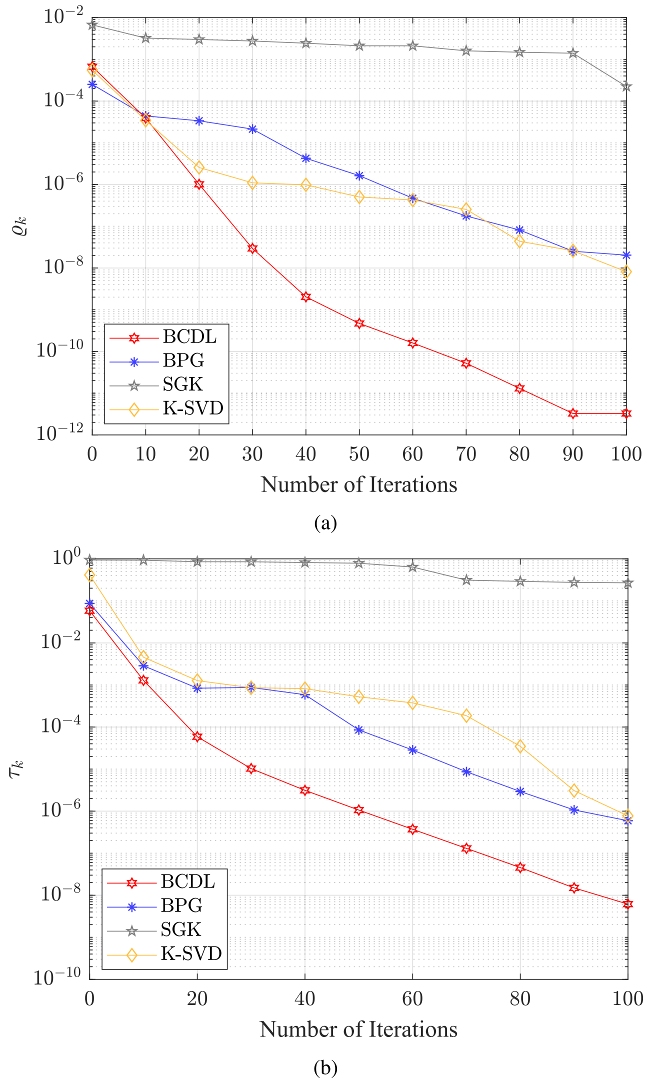


FIGURE 5. Convergence speed of dictionary learning algorithms: (a) for dictionary; (b) for sparse representation.

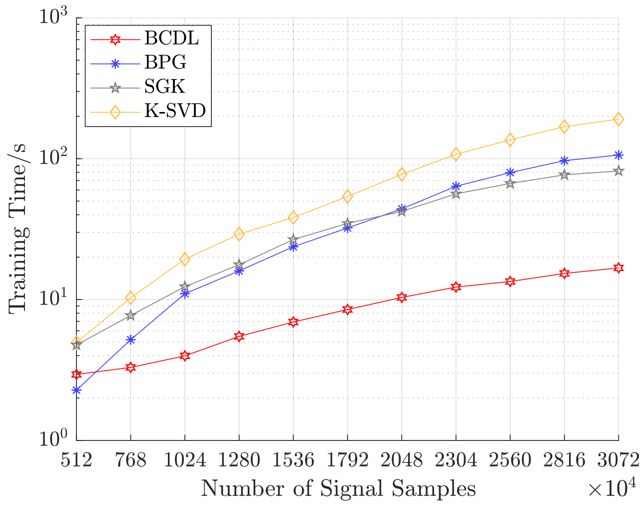


FIGURE 6. Training time of dictionary learning based approaches versus the number of signal samples.

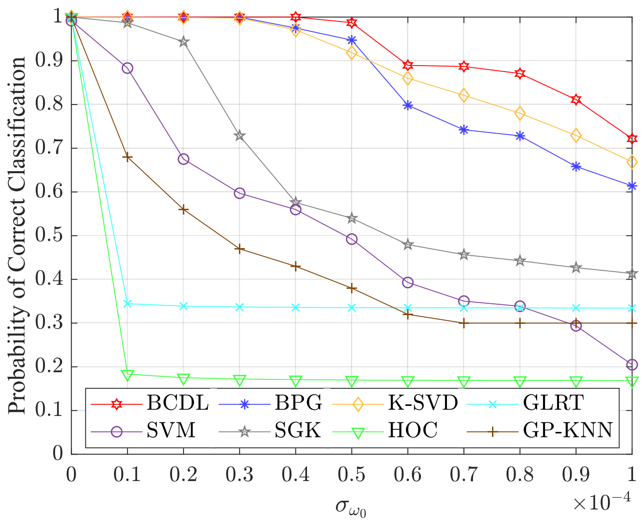


FIGURE 7. Classification accuracy versus residual frequency offset.

of BCDL is better than other methods, which indicates that BCDL is more robust over RFO. Therefore, BCDL can be adopted for signals under imperfect synchronization, such as Doppler Frequency Offset (DFO) in high speed vehicle communication.

Receivers could be influenced by phase noise which is usually caused by imperfect synchronization and channel fading. There are two common kinds of phase noise, that is, uniformly distributed [30] and Wiener phase noise [31]. First, we compare the robustness over uniformly distributed phase noise. We modeled the phase noise as an uniform distributed over $[-\vartheta, \vartheta]$ [30]. The classification accuracies of different methods over ϑ ranging from 0 to π rad are illustrated in Fig. 8(a) for SNR $=$ 20dB. We find that the probability of correct classification of BCDL is higher than other methods over uniformly distributed phase noise.

Then we analyze the robustness of different methods over Wiener phase noise, which usually caused by the oscillator without a Phase Locked Loop (PLL). The Wiener phase noise is modeled as [31]

$$\theta'_{n+1} = \theta'_n + \xi_n,$$

where θ'_n denotes the phase noise of the n-th signal and ξ_n is an independent and identically distributed (i.i.d.) Gaussian random variable following $\xi_n \sim \mathcal{N}\left(0, \sigma_p^2\right)$. Fig. 8(b) illustrates the classification accuracies of different methods over Wiener phase noise for SNR $=$ 20dB, with σ_p ranging from 0 to 1. We can find that BCDL performs better than other methods, which indicates that BCDL is more robust over the Wiener phase noise. From Fig. 8(a) and Fig. 8(b), we see that BCDL is robust over phase noise.

From the above experiments, we can conclude that BCDL can be used in the scenario with imperfect synchronization.

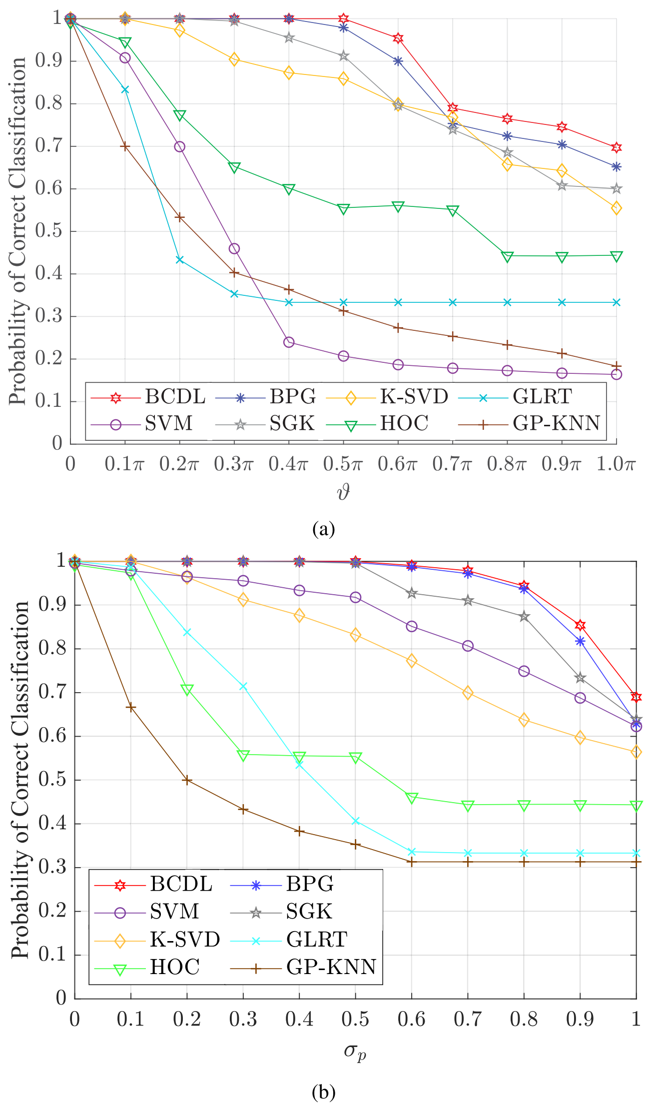


FIGURE 8. Classification accuracy of BCDL and other methods over: (a) uniformly distributed phase; (b) wiener phase noise.

VI. CONCLUSIONS

In this paper, we proposed a dictionary learning based AMC framework, where we first use training signals to train the dictionary and then classify the modulation format for the test signals via their sparse representations. We also proposed a special dictionary learning based method, called Block Coordinate descent Dictionary Learning (BCDL). We proved the convergence of our proposed BCDL where we could further quantify its convergence speed. Moreover, both the theoretical and experimental results show that BCDL can be trained faster than other existing dictionary learning methods. It can be seen from the simulation results that the BCDL method achieves a higher classification accuracy than other methods. Furthermore, simulation results also show that the BCDL can be utilized in the scenario with imperfect synchronization.

APPENDIX

Here, we calculate the computational complexity of dictionary-learning based methods. It can be seen that our proposed BCDL achieves lower computational complexity than other dictionary-learning based methods.

COMPARISON OF COMPUTATIONAL COMPLEXITY AMONG DIFFERENT ALGORITHMS

1) THE COMPLEXITY OF BLOCK PROXIMAL GRADIENT (BPG) DESCENT

In the BPG method, it requires $\mathcal{T}_{t_{\mathbf{X}}} = O(NML)$, $\mathcal{T}_{t_{\mathbf{D}}} = O(ML^2)$, and $\mathcal{T}_{\mathbf{D}} = O(NML)$ FLOPs to update the step sizes $t_{\mathbf{X}} = 2\|\mathbf{D}^T\mathbf{D}\|_F$, $t_{\mathbf{D}} = 2\|\mathbf{X}\mathbf{X}^T\|_F$ and dictionary $\mathbf{D}$, respectively. According to [17], BPG requires $\mathcal{T}_{\mathbf{X}} = O(NML)$ FLOPs to update $\mathbf{X}$. Because $M \ll L$ in practice [17], we need

$$\begin{aligned}\mathcal{T}_2 &= \mathcal{T}_{t_{\mathbf{X}}} + \mathcal{T}_{t_{\mathbf{D}}} + \mathcal{T}_{\mathbf{X}} + \mathcal{T}_{\mathbf{D}} \\ &= O(ML^2 + NML + NM^2) = O(ML^2 + NML)\end{aligned}$$

FLOPs in each iteration. Xu and Yin [17] proved that the number of iteration of BPG holds $K_{\text{BPG}} = O(\epsilon^{-\frac{1}{1+2\zeta'}})$, where $\zeta' > 0$ is a constant and $\epsilon > 0$ denotes the threshold. Thus, the computational complexity of BPG is

$$\mathcal{T}_{\text{BPG}} = \mathcal{T}_2 \cdot K_{\text{BPG}} = O(M(L+N)L \cdot \epsilon^{-\frac{1}{1+2\zeta'}}).$$

For the number of iterations N_I' which holds $N_I' \ll L$ [17], the computational complexity of BPG is

$$\mathcal{T}_{\text{BPG}} = \mathcal{T}_2 \cdot N_I' = O(N_I' M(L+N)L).$$

2) THE COMPLEXITY OF SEQUENTIAL GENERALIZATION OF K-MEANS (SGK)

Let $\hat{\mathbf{x}}_m^T$ denote the m-th row of the sparse representation $\mathbf{X}$. According to [16], SGK requires $\mathcal{T}_{\mathbf{E}_m} = O(NML)$ and $\mathcal{T}'_{\mathbf{d}_m} = O(NL)$ FLOPs to calculate the error matrix $\mathbf{E}_m = \mathbf{Y} - \sum_{j\neq m}\mathbf{d}_j\hat{\mathbf{x}}_j^T$, and $\mathbf{d}_m, m = 1, \ldots, M$, respectively. In each iteration, the computational complexity is

$$\mathcal{T}_3 = M\mathcal{T}_{\mathbf{E}_m} + M\mathcal{T}'_{\mathbf{d}_m} = O(NM^2L).$$

Hence, for the number of iterations N_I'' which holds $N_I'' \ll L$ [16], the computational complexity of SGK is given by

$$\mathcal{T}_{\text{SGK}} = N_I''\mathcal{T}_3 = O(NN_I''M^2L).$$

3) THE COMPLEXITY OF K-SVD

K-SVD requires $\mathcal{T}_{\mathbf{E}_m} = O(NML)$ FLOPs to calculate the error matrix $\mathbf{E}_m = \mathbf{Y} - \sum_{j\neq m}\mathbf{d}_j \cdot \hat{\mathbf{x}}_j^T$. The computational complexity of the Singular Value Decomposition (SVD) of an $N \times L$ matrix is $\mathcal{T}_{\text{SVD}} = O(L^3 + NL^2 + N^2L)$ [15]. Therefore, in each iteration, for $M \ll L$, the number of FLOPs is

$$\mathcal{T}_4 = \mathcal{T}_{\mathbf{E}_m} + \mathcal{T}_{\text{SVD}} = O(L^3 + NL^2 + N^2L).$$

With the assumption that the number of iterations N_I''' holds $N_I''' \ll L$ [15], the computational complexity of K-SVD is given by

$$\mathcal{T}_{\text{K-SVD}} = N_I''' \cdot \mathcal{T}_4 = O(L^3 + NN_I'''L^2 + N^2N_I'''L).$$

The computational complexities of the aforementioned methods are summarized in Table 1.

REFERENCES

[1] Z. Wu, S. Zhou, Z. Yin, B. Ma, and Z. Yang, "Robust automatic modulation classification under varying noise conditions," *IEEE Access*, vol. 5, pp. 19733–19741, Aug. 2017.

[2] O. A. Dobre, A. Abdi, Y. Bar-Ness, and W. Su, "Survey of automatic modulation classification techniques: Classical approaches and new trends," *IET Commun.*, vol. 1, no. 2, pp. 137–156, Apr. 2007.

[3] B. Ramkumar, "Automatic modulation classification for cognitive radios using cyclic feature detection," *IEEE Circuits Syst. Mag.*, vol. 9, no. 2, pp. 27–45, Jun. 2009.

[4] C. Dubuc, D. Boudreau, F. Patenaude, and R. Inkol, "An automatic modulation recognition algorithm for spectrum monitoring applications," in *Proc. IEEE Int. Conf. Commun.*, Jun. 1999, pp. 570–574.

[5] J. L. Xu, W. Su, and M. Zhou, "Software-defined radio equipped with rapid modulation recognition," *IEEE Trans. Veh. Technol.*, vol. 59, no. 4, pp. 1659–1667, May 2010.

[6] W. Wei and J. M. Mendel, "Maximum-likelihood classification for digital amplitude-phase modulations," *IEEE Trans. Commun.*, vol. 48, no. 2, pp. 189–193, Feb. 2000.

[7] F. Hameed, O. A. Dobre, and D. Popescu, "On the likelihood-based approach to modulation classification," *IEEE Trans. Wireless Commun.*, vol. 8, no. 12, pp. 5884–5892, Dec. 2009.

[8] E. E. Azzouz and A. K. Nandi, "Automatic identification of digital modulation types," *Signal Process.*, vol. 47, no. 1, pp. 55–69, Nov. 1995.

[9] A. Swami and B. M. Sadler, "Hierarchical digital modulation classification using cumulants," *IEEE Trans. Commun.*, vol. 48, no. 3, pp. 416–429, Mar. 2000.

[10] M. W. Aslam, Z. Zhu, and A. K. Nandi, "Automatic modulation classification using combination of genetic programming and KNN," *IEEE Trans. Wireless Commun.*, vol. 11, no. 8, pp. 2742–2750, Aug. 2012.

[11] L. Han, F. Gao, Z. Li, and O. A. Dobre, "Low complexity automatic modulation classification based on order-statistics," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 400–411, Jan. 2017.

[12] I. Tošić and P. Frossard, "Dictionary learning," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 27–38, Mar. 2011.

[13] M. Sadeghi, M. Babaie-Zadeh, and C. Jutten, "Dictionary learning for sparse representation: A novel approach," *IEEE Signal Process. Lett.*, vol. 20, no. 12, pp. 1195–1198, Dec. 2013.

[14] K. Engan, S. O. Aase, and J. H. Husoy, "Method of optimal directions for frame design," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Mar. 1999, pp. 2443–2446.

[15] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[16] S. K. Sahoo and A. Makur, "Dictionary training for sparse representation as generalization of K-means clustering," *IEEE Signal Process. Lett.*, vol. 20, no. 6, pp. 587–590, Jun. 2013.

[17] Y. Xu and W. Yin, "A fast patch-dictionary method for whole image recovery," *Inverse Problems Imag.*, vol. 10, no. 2, pp. 563–583, May 2014.

[18] X. Yin, J. Liu, Y. Su, X. Xiong, and G. Xiong, "A low-complexity synchronizer for OFDM-UWB-based vehicular communications," *IEEE Access*, vol. 5, pp. 7272–7284, Apr. 2017.

[19] J. J. van de Beek, M. Sandell, and P. O. Börjesson, "ML estimation of time and frequency offset in OFDM systems," *IEEE Trans. Signal Process.*, vol. 45, no. 7, pp. 1800–1805, Nov. 1997.

[20] M. Barkat, "Detection and parameter estimation," in *Signal Detection and Estimation*, 2nd ed. London, U.K.: Artech House, 2005, ch. 10, sec. 5, pp. 576–580.

[21] K. Xiang, B. N. Li, L. Zhang, M. Pang, M. Wang, and X. Li, "Regularized Taylor echo state networks for predictive control of partially observed systems," *IEEE Access*, vol. 4, pp. 3300–3309, Jun. 2016.

[22] C. Liu, Y. Fang, and J. Liu, "Some new results about sufficient conditions for exact support recovery of sparse signals via orthogonal matching pursuit," *IEEE Trans. Signal Process.*, vol. 65, no. 17, pp. 4511–4524, Sep. 2017.

[23] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Math. Program.*, vol. 146, nos. 1–2, pp. 459–494, 2014.

[24] Y. Xu and W. Yin, "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion," *SIAM J. Imag. Sci.*, vol. 6, no. 3, pp. 1758–1789, 2013.

[25] T. Hastie, R. Tibshirani, and M. Wainwright, "Optimization methods," in *Statistical Learning with Sparsity: The Lasso and Generalizations*. New York, NY, USA: CRC Press, 2015, pp. 103–107.

[26] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, "Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality," *Math. Oper. Res.*, vol. 116, no. 1, pp. 5–16, Jan. 2009.

[27] H. Attouch and J. Bolte, "On the convergence of the proximal algorithm for nonsmooth functions involving analytic features," *Math. Program.*, vol. 116, no. 1, pp. 5–16, Jan. 2009.

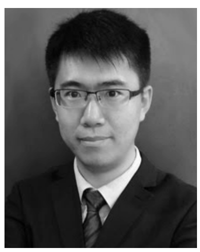
[28] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002.

[29] S. Han, Y. Sung, and Y. H. Lee, "Filter design for generalized frequency-division multiplexing," *IEEE Trans. Signal Process.*, vol. 65, no. 7, pp. 1644–1659, Apr. 2017.

[30] J. G. Proakis and M. Salehi, "Optimum receivers for AWGN channels," in *Digital Communications*, 5th ed. New York, NY, USA: McGraw-Hill, 2007, ch. 4, sec. 5, pp. 210–226.

[31] T.-J. Lee and Y.-C. Ko, "Channel estimation and data detection in the presence of phase noise in MIMO-OFDM systems with independent oscillators," *IEEE Access*, vol. 5, pp. 9647–9662, Jun. 2017.

KEZHONG ZHANG (S'15) received the B.S. degree from the Beijing University of Posts and Telecommunications in 2012. He is currently pursuing the Ph.D. degree in communication and information systems with the Wireless Technology Innovation Institute, BUPT. He is currently a member of the Key Laboratory of Universal Wireless Communications, Ministry of Education, China. His research interests include the sparse representation, signal processing in non-cooperative environment, pattern recognition, and neural networks.



EASTON LI XU (S'13–M'14) received B.S. degree in mathematics and the B.S. degree in computer science from Peking University, Beijing, China, and the Ph.D. degree from the Department of Mathematics, The University of Hong Kong, Hong Kong, in 2007 and 2014, respectively. He was a Post-Doctoral Research Associate with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA. He is currently a Presidential Post-Doctoral Fellow with the Department of Electrical Engineering and Computer Science, University of Michigan, and also with the School of Science and Engineering, The Chinese University of Hong Kong at Shenzhen. His research interests include statistical learning, coding theory, and data analytics. He has served as TPC Member of the IEEE GLOBECOM in 2016 and 2017, the IEEE GlobalSIP in 2016 and 2017, the IEEE WCNC in 2017 and 2018, and the IEEE 5GWF in 2018. He has also served as the Local Arrangement Co-Chair for MIIS in 2016 and the Webmaster for WCI in 2013 and CAM in 2016. He was the 2014 Exemplary Reviewer for the IEEE COMMUNICATIONS LETTERS.

ZHIYONG FENG (M'08–SM'15) received the B.S., M.S., and Ph.D. degrees from the Beijing University of Posts and Telecommunications, China. She is currently a Professor with BUPT and the Director of the Key Laboratory of Universal Wireless Communications, Ministry of Education, China. Her main research interests include the wireless network virtualization in fifth generation mobile networks (5G), spectrum sensing and dynamic spectrum management in cognitive wireless networks, universal signal detection and identification, network information theory,and so on. She is active in standards development, such as ITU-R WP5A/WP5D, IEEE 1900, ETSI, and CCSA.

PING ZHANG (M'05–SM'15) received the Ph.D. degree from the Beijing University of Posts and Telecommunications. He is currently a Professor with the Beijing University of Posts and Telecommunications, the Vice Director of the *Science Bulletin* (Chinese Academy of Science), an Expert of MIIT major special technological programs, a member of the IMT-Advanced (4G) Group, and a Consultancy Expert of the National Science Foundation Committee. He has authored eight books. He was invited to participate in the publishing of the book *Long Term Evolution*, and invited to write the book *Cognitive Radio Networks* (Science Publishing Company, 2010). He has authored over 400 papers at home and abroad publications, such as the *IEEE Communications Magazine*, the IEEE Electronics Letters, the *Chinese Journal of Electronics*. He holds about 170 patents. He acts as the Vice Editor-in-Chief of the History of China's Telecommunication Discipline organized by the China Institute of Communications.