Online Non-negative Dictionary Learning via Moment Information for Sparse Poisson Coding

Xiaotian Yu, Haiqin Yang, Irwin King and Michael R. Lyu
Shenzhen Key Laboratory of Rich Media Big Data Analytics and Applications,
Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen, China
Department of Computer Science and Engineering,
The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong
Email: {xtyu, hqyang, king, lyu}cse.cuhk.edu.hk

***Abstract*—Online dictionary learning for sparse coding is an effective tool for data analysis. It incrementally learns a set of basis vectors with sparse linear combinations of these vectors when new samples appear. Previous work assumes that the samples embed Gaussian noises, which weaken the power of these methods in handling real applications with non-negative data (e.g., frequency data in word counts). Differently, in this paper, we concentrate on online learning for non-negative dictionary by using moment information for sparse Poisson coding. We exploit the non-negativity of Poisson models to learn a set of non-negative basis vectors and a non-negative sparse linear combination for the moment information of samples. Specifically, we first formulate the online learning problem via the maximum-a-posteriori (MAP) framework. We then propose a novel online algorithm which alternatively updates the sparse-coefficient vector and the basis vectors with non-negativity constraints when a new sample arrives. More importantly, we present sufficient convergence analyses to guarantee the performance of the proposed algorithm, which leads to convergence of a stable dictionary for characterizing the moment information of samples. We finally conduct a series of experiments on word-counts data and image data to show merits of the proposed online algorithm.**

I. Introduction

Dictionary learning for sparse coding, due to its superiority in signal representation, has been particularly useful for many real applications, such as image processing [1], visual tracking [2], compressed sensing [3], and high-dimensional text data analysis [4]. It has attracted numerous academic researchers and industrial practitioners to develop efficient algorithms [5]. The aim of dictionary learning for sparse coding is to learn a set of basis vectors instead of predefined ones (e.g., wavelets [6]) or sampling-constant ones [4]. Then a sparse linear combination of the learned basis vectors (which are called a learned dictionary) can be adopted to approximate a given signal. Compared with traditional signal-decomposition techniques, the dictionary learning for sparse coding has shown good performance on numerous tasks, such as classifications [7] and visual tracking [2]. Besides, the traditional signal-decomposition techniques have additional harsh constraints, e.g., principle component analysis requiring orthogonal basis vectors. By contrast, the dictionary learning for sparse coding shows a prominent advantage of relatively loose constraints in the basis vectors.

How to learn these basis vectors of the dictionary in the batch-learning mode has been well studied [4], [5], [8]. However, when the samples appear sequentially, learning the dictionary incrementally becomes a challenging problem. Recently, several pieces of work have investigated the online dictionary learning for sparse coding [2], [9], [10]. The previous work on online dictionary learning for sparse coding assumes that the samples embed Gaussian noises in [2], [9], [10]. In other words, it basically assumes that the data follow the Gaussian distribution. Based on this assumption, the dictionary learning problem is then formulated as an optimization problem based on the Euclidean norm of the reconstruction errors for samples. Besides, in [9], it transforms the batch-learning mode to the online-learning mode by constructing matrices for saving previous information. In that case, it is both time-consuming in solving the resultant optimization problem and memory-consuming for storing matrices information.

Although the assumption of Gaussian distribution fits in many real-world applications, it is inappropriate for some domains, such as the frequency data in word counts [4], [8], [11]–[13]. For frequency data, they exhibit the characteristics of non-negativity and unboundedness due to the inherent mechanisms of word counts. In this case, assuming data with Poisson distributions becomes more practical [4], [14]. In practice, by using Poisson distributions instead of Gaussian distributions, we are facing two extremely challenging problems in the online-learning mode. First, how to cumulate previous information of online dictionary learning for sparse Poisson coding, where we cannot use the addition property of errors based on Euclidean distance [9]. Second, how to maintain the non-negativity constraints for both the dictionary's vectors and the sparse-coefficient vector.

The non-negativity constraints of matrices can be solved from a viewpoint of non-negative matrix factorization (NMF) [4], [8], [12]. However, in [12], the authors developed the Itakura-Saito divergence, which cannot be extended to the case for Poisson models. Dikmen et al. [8] provided a new algorithm for NMF based on Kullback-Leibler divergence. But they only discussed the dictionary learning in the batch-learning mode. In [4], the authors considered the sparse coding based on NMF, which was also in the batch-learning mode.

In order to tackle the above two challenging problems in

the online-learning mode, in this paper, we propose a novel online algorithm for dictionary learning with non-negativity constraints, i.e., online non-negative dictionary learning via moment information for sparse Poisson coding. The significant constraints we meet are the non-negativity for both dictionary's basis vectors of and the sparse-coefficient vector of moment information. Specifically, by using the moment information of samples, we firstly present a maximum-a-posteriori (MAP) problem for online dictionary learning. Here, we exploit the assumption of all the dimensions in the sample vector being mutually independent. Then, for solving the MAP problem, we propose an elegant online algorithm to alternatively update the dictionary's vectors and the sparse-coefficient vector while rigorously maintaining their non-negativity. More importantly, we present sufficient convergence analyses to guarantee the efficiency of the proposed algorithm on online non-negative dictionary learning via moment information for sparse Poisson coding. Finally, experimental results on word-counts data and image data show merits of the online algorithm.

In summary, our work contains the following contributions:

- We are the first to develop online non-negative dictionary learning via moment information for sparse Poisson coding. The proposed algorithm can make up the ineffectiveness of traditional methods with the Gaussian assumption and help to solve many real-world applications, especially for the cases of frequency data.
- We provide sufficient theoretical analyses for the convergence of the proposed algorithm, which guarantees that the learned dictionary converges to a stable one for characterizing the moment information of samples.
- We conduct a series of experiments in the applications of word-counts data and image data to further demonstrate the merits of the proposed algorithm.

II. Related Work

We briefly review some closely related work of dictionary learning for sparse coding in the batch-learning mode, dictionary learning for sparse coding based on the Gaussian assumption in the online-learning mode, and the technique of Poisson regression used in dictionary learning.

Dictionary learning is important in representing real data by using sparse linear combinations of the basis vectors [15]. Various methods have been proposed for dictionary learning, such as dictionary learning via optimal directions [16], design of overlapping dictionaries [17], iterative-least-square dictionary learning [18], and dictionary learning by the K-SVD method [19]. It is worth pointing out that all these previous methods focus on the batch-learning mode.

In the batch learning, given a finite unlabeled sequence of sample set as $\mathcal{X} = \{\mathbf{x}_t\}_{t=1}^T$ (with $\mathbf{x}_t = [x_t^1, x_t^2, \cdots, x_t^m]^{\mathrm{T}} \in \mathbb{R}^m$ and $^{\mathrm{T}}$ being the transpose of a vector) and the assumption of independent and identical distribution (i.i.d.) for the samples, we have

$$\begin{aligned} \mathbf{x}_t =& \mathbf{D}\mathbf{w} + \mathbf{e}, \\ P(\mathcal{X}|\mathbf{D}) =& \prod_{t=1}^{T} P(\mathbf{x}_t|\mathbf{D}), \end{aligned} \tag{1}$$

where $\mathbf{D} \in \mathbb{R}^{m\times k}$ is the dictionary with k being the number of basis vectors, $\mathbf{w} \in \mathbb{R}^k$ is the sparse-coefficient vector, $\mathbf{e} \in \mathbb{R}^m$ is the noise vector, and $P(\mathcal{X}|\mathbf{D})$ is the conditional probability. Note that the conditional probability of sample $\mathbf{x}_t$ can be calculated as $P(\mathbf{x}_t|\mathbf{D}) = \int P(\mathbf{x}_t|\mathbf{D}, \mathbf{w})P(\mathbf{w})\mathrm{d}\mathbf{w}$. However, it is hard to implement the integration over $\mathbf{w}$ [20]. Instead, we solve $\mathbf{w}$ by solving the maximization of $P(\mathbf{x}_t, \mathbf{w}|\mathbf{D})$. Then, based on maximum likelihood, we try to solve

$$\max_{\mathbf{D}} \big(\sum_{t=1}^{T} \max_{\mathbf{w}} \log P(\mathbf{x}_t, \mathbf{w}|\mathbf{D}) \big). \tag{2}$$

By using the assumptions of dimension independence in $\mathbf{x}_t$ and Gaussian distribution in $\mathbf{e}$ of (1), the traditional dictionary learning is to learn a set of basis vectors by minimizing the following objective function

$$\min_{\mathbf{D}} \sum_{t=1}^{T} \min_{\mathbf{w}} \|\mathbf{x}_t - \mathbf{D}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1. \tag{3}$$

Note that, from Eq. (2) to Eq. (3), a regularized term of $\lambda\|\mathbf{w}\|_1$ is added for sparsity and $\lambda \in \mathbb{R}_+$ is a sparsity parameter with $\mathbb{R}_+$ denoting the set of non-negative real values. Besides, we find that, in Eq. (3), the addition property of errors has been used based on the Gaussian assumption and Euclidean distance. Differently, we use the Poisson models here, which means that we cannot learn the dictionary by solving Eq. (3).

In the big data era, the popularity of sequential data motivates the online learning paradigm. The traditional dictionary learning methods are restricted to the online sequential data, where the samples arrive one-by-one. Online learning has been studied for decades. Different from these batch learning algorithms, online learning algorithms receive samples sequentially and update models incrementally. Perceptron is the first and simplest algorithm for online learning [21]. After that, many online learning algorithms have been further developed [22], [23], including online gradient descent algorithms [24], [25], ALMA [26], ROMMA [27], passive aggressive algorithms [28], etc. Recently, online dictionary learning for sparse coding has been accordingly investigated [2], [9], [10]. However, in [2], [9], [10], they all have the Gaussian assumption for real data, which weakens the power of these methods in handling real applications with non-negative data. For example, for word-counts data, the Gaussian assumption restricts the effectiveness of the previous algorithms [4]. It is worth pointing out that most of the previous work on dictionary learning for sparse Poisson coding focuses on the batch-learning mode [4], [8], [14].

For these real data with underlying Poisson models, it is reasonable to employ the technique of Poisson regression for estimating the models' parameters. Poisson regression is

effectively applied to solving the problems of predictions and classifications [4], [29]. Traditional Poisson regression assumes that the mean and variance of Poisson models can be linear regressive based on the past samples' information. Note that, in our work, we adopt the linear-regression assumption for estimating parameters of Poisson models.

To tackle the above issue of online non-negative dictionary learning for sparse Poisson coding, our work concentrates on developing a novel and efficient online algorithm for the dictionary learning via the moment information of samples, along with sufficient convergence analyses to guarantee good performance of the learned dictionary.

III. Preliminaries and Problem Statement

In this section, we give the preliminaries and the corresponding problem statement for the novel online learning algorithm. We firstly present how to obtain the moment information of samples and develop the MAP problem based on Eq. (1) with sequential samples. Furthermore, we give the Poisson regression for the resultant optimization problem.

A. Preliminaries

Inspired by Eq. (1), we focus on online dictionary learning of sample sequences with the non-negativity constraints in Poisson models. Given a finite unlabeled data sequence $\mathcal{X}_+ = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T\}$, where $\mathbf{x}_t \in \mathbb{R}_+^m$ ($t = 1, 2, \cdots, T$ and T is the total number of samples) with $\mathbb{R}_+^m$ denoting the set of non-negative real values in m-dimension space, our goal is to obtain a good dictionary D^* to characterize the sample information of $\{\mathbf{x}_t\}_{t=1}^T$.

In [9], based on Gaussian assumption, Mairal et al. implemented the online dictionary learning by developing two matrices, which cumulate the previous sample information. By contrast, we try to online train a dictionary via the moment information in our work instead of storing previous sample information. By assuming that the samples are all independent and identical distributed with the true mean $\boldsymbol{\mu}^* \in \mathbb{R}_+^m$, we have an unbiased estimator of $\boldsymbol{\mu}^*$ as

$$\widetilde{\mathbf{x}}_t = \frac{1}{t}(\mathbf{x}_1 + \mathbf{x}_2 + \cdots + \mathbf{x}_t). \tag{4}$$

Based on the law of large numbers, we know

$$\lim_{t\to\infty} \widetilde{\mathbf{x}}_t - \boldsymbol{\mu}^* \to \mathbf{0}. \tag{5}$$

Note that, though T is finite in practice, we can satisfy Eq. (5) by setting T sufficiently large in implementations.

Then, in our work, we learn the dictionary via the unbiased estimator of mean, which is shown as Eq. (4). Inspired by [9] and based on problem of Eq. (2), we can obtain the following optimization problem for online dictionary learning via moment information:

$$\mathbf{D}_t = \arg\max_{\mathbf{D}} \big(\max_{\mathbf{w}} \log P(\widetilde{\mathbf{x}}_t, \mathbf{w}|\mathbf{D}) \big). \tag{6}$$

In the batch-learning mode, we can employ the Poisson regression in Eq. (6) so that $\widetilde{\mathbf{x}}_t = \frac{1}{t}\sum_{l=1}^t \mathbf{x}_l = \mathbf{D}_t\mathbf{w}_t$. Based on Eq. (5), we have the result of $\lim_{t\to\infty} \mathbf{D}_t\mathbf{w}_t \to \boldsymbol{\mu}^*$, with more details discussed in Section V.

By contrast, in the online-learning mode, at time instant t, we only have sample information of $\mathbf{x}_t$, which means that $\widetilde{\mathbf{x}}_t = \frac{1}{t}\sum_{l=1}^t \mathbf{x}_l$ cannot be obtained. Thus we try to use the previous dictionary information for the calculation of $\widetilde{\mathbf{x}}_t$.

Suppose at time instant $t-1$ we have the dictionary and the sparse vector, respectively, as $\mathbf{D}_{t-1}$ and $\mathbf{w}_{t-1}$. We readily have $\lim_{t-1\to\infty} \mathbf{D}_{t-1}\mathbf{w}_{t-1} \to \boldsymbol{\mu}^*$. Then we can obtain the first-order moment information as

$$\widetilde{\mathbf{x}}_t = \begin{cases} \mathbf{x}_t, \text{ iff } t = 1, \\ \big(\mathbf{x}_t + (t-1)\mathbf{D}_{t-1}\mathbf{w}_{t-1}\big)/t, \text{ iff } t > 1. \end{cases} \tag{7}$$

Note that, in Eq. (7), we have $\lim_{t\to\infty} \widetilde{\mathbf{x}}_t \to \boldsymbol{\mu}^*$.

When we consider the noise in Eq. (1) as Poisson distribution, we have $P(\widetilde{\mathbf{x}}_t, \mathbf{w}|\mathbf{D}) = \text{Poisson}(\widetilde{\mathbf{x}}_t, \mathbf{w}|\mathbf{D})$ in Eq. (6), where $\text{Poisson}(\widetilde{\mathbf{x}}_t, \mathbf{w}|\mathbf{D})$ denotes an underlying multi-variate Poisson model. Besides, similar to [4], we assume that the dimensions in $\mathbf{x}_t$ are mutually independent. Then, by using the moment information in Eq. (7), we have

$$\text{Poisson}(\widetilde{\mathbf{x}}_t, \mathbf{w}|\mathbf{D}) = \widetilde{\prod}\Big[\frac{(\mathbf{D}\mathbf{w})^{\cdot\widetilde{\mathbf{x}}_t}.\exp(-\mathbf{D}\mathbf{w})}{\widetilde{\mathbf{x}}_t.!}\Big], \tag{8}$$

where $\widetilde{\prod}$ denotes the product of all elements in a vector. Note that, in Eq. (8), the dot in the power of $(\mathbf{D}\mathbf{w})^{\cdot\widetilde{\mathbf{x}}_t}$ denotes the element-wise operation for the column vector of $\mathbf{D}\mathbf{w}$. The dot in $\mathbf{x}_t.! = [x_t^1!, x_t^2!, \cdots, x_t^m!]^{\mathrm{T}}$ has the same meaning of element-wise operation. In addition, in Eq. (8), the expression of $.\exp(-\mathbf{D}\mathbf{w})$ is also element-wise since $\mathbf{D}\mathbf{w}$ is a column vector. Here, based on Poisson regression, we use the formula of $\mathbf{D}\mathbf{w}$ as the estimation of mean for sample information in the Poisson model. From this viewpoint, we know that a good dictionary $\mathbf{D}^*$ should well characterize the information of true mean $\boldsymbol{\mu}^*$ for samples.

In the online-learning mode, our goal for online dictionary learning is to minimize the reconstruction error of the moment information over the learning process, which is defined as

$$\begin{aligned} \mathcal{L}_t(\mathbf{D}) = \ & \frac{1}{t}\sum_{l=1}^t \ell(\widetilde{\mathbf{x}}_l, \mathbf{D}), \\ & \text{s.t. } \mathbf{D} \geq \mathbf{0}, \end{aligned} \tag{9}$$

where $\ell(\widetilde{\mathbf{x}}_l, \mathbf{D})$ is a loss function and $\mathbf{D} \in \mathbb{R}^{m\times k} \geq \mathbf{0}$ means that each element in $\mathbf{D}$ is no less than 0. By using a column-vector representation, we have $\mathbf{D} = [\mathbf{d}_1, \cdots, \mathbf{d}_k]$. Besides, in order to guarantee the boundedness of $\mathbf{D}$ [9], we can also introduce constraints of $\mathbf{d}_j^{\mathrm{T}}\mathbf{d}_j \leq 1$, with $j = 1, \cdots, k$.

We further consider $\mathbf{w} \geq \mathbf{0}$, and set the loss function as the negative formulation of (6). Thus, we have

$$\begin{aligned} \ell(\widetilde{\mathbf{x}}_l, \mathbf{D}) = & - \min_{\mathbf{D}\in\mathbb{R}_+^{m\times k}} \big(\min_{\mathbf{w}\in\mathbb{R}_+^k} \log P(\widetilde{\mathbf{x}}_l, \mathbf{w}|\mathbf{D}) \big), \\ & \text{s.t. } \mathbf{d}_j^{\mathrm{T}}\mathbf{d}_j \leq 1, \ j = 1, \cdots, k. \end{aligned} \tag{10}$$

In Section V, we will further show that solving Eq. (10) is equivalent to minimizing the original problem of Eq. (9).

B. MAP for Sparse Coding

In the previous work [9], [14], we know that it is difficult to directly solve Eq. (10), because Eq. (10) is not jointly convex for $\mathbf{D}$ and $\mathbf{w}$. Inspired by previous work, we can firstly solve Eq. (10) by applying the information of $\mathbf{D}_{t-1}$. Secondly, we can resolve $\mathbf{D}_t$ by using the sparsity information of $\mathbf{w}_t$. Clearly, given the dictionary $\mathbf{D}_{t-1}$, we know that solving Eq. (10) can be reformulated as the sparse coding for $\widetilde{\mathbf{x}}_t$ by adding a sparsity constraint [20]. That is,

$$\begin{aligned}\mathbf{w}_t = &\arg\min_{\mathbf{w}} \big(-\log P(\widetilde{\mathbf{x}}_t, \mathbf{w}|\mathbf{D}_{t-1})\big) + \lambda\|\mathbf{w}\|_1,\\ &\text{s.t. } \mathbf{w} \geq \mathbf{0},\end{aligned} \tag{11}$$

where $\lambda \in \mathbb{R}_+$ denotes the sparsity parameter.

By using the Poisson model in Eq. (8), we further have the following problem for sparse Poisson coding:

$$\begin{aligned}\mathbf{w}_t = &\arg\min_{\mathbf{w}} \mathbf{1}_m^T\Big[-\log\Big(\frac{(\mathbf{D}_{t-1}\mathbf{w})^{\cdot\widetilde{\mathbf{x}}_t}.\exp(-\mathbf{D}_{t-1}\mathbf{w})}{\widetilde{\mathbf{x}}_t.!}\Big)\Big]\\ &+\lambda\|\mathbf{w}\|_1,\\ &\text{s.t. } \mathbf{w} \geq \mathbf{0},\end{aligned} \tag{12}$$

where $\mathbf{1}_m \in \mathbb{R}^m$ is a vector with all elements being 1.

C. MAP for Dictionary Learning

Based on the definition of loss function in Eq. (10), at time instant t, the optimization problem for online non-negative dictionary learning can be formulated as

$$\begin{aligned}\mathbf{D}_t = &-\arg\min_{\mathbf{D}} \big(\log P(\widetilde{\mathbf{x}}_t, \mathbf{w}_t|\mathbf{D})\big),\\ &\text{s.t. } \mathbf{D} \geq \mathbf{0},\\ &\quad \mathbf{d}_j^{\mathrm{T}}\mathbf{d}_j \leq 1,\ j = 1, \cdots, k.\end{aligned} \tag{13}$$

Online learning problem of Eq. (13) can be formulated as

$$\begin{aligned}\mathbf{D}_t = &\arg\min_{\mathbf{D}} \mathbf{1}_m^T\Big[-\log\big(\frac{(\mathbf{D}\mathbf{w}_t)^{\cdot\widetilde{\mathbf{x}}_t}.\exp(-\mathbf{D}\mathbf{w}_t)}{\widetilde{\mathbf{x}}_t.!}\big)\Big],\\ &\text{s.t. } \mathbf{D} \geq \mathbf{0},\ \mathbf{d}_j^{\mathrm{T}}\mathbf{d}_j \leq 1,\ j = 1, \cdots, k.\end{aligned} \tag{14}$$

IV. Online Learning Algorithm

In this section, we present a novel and efficient algorithm for solving the optimization problem of Eq. (10) when samples appear sequentially with updated moment information. It aims at alternatively solving the optimization problem of Eq. (12) and then the problem of Eq. (14) when samples arrive.

A. Updating Sparse Weights

By fixing the dictionary at the previous time instant (i.e., $\mathbf{D}_{t-1}$), we can solve the optimization problem of Eq. (12).

Theorem 1. The solution for optimization problem of Eq. (12) is equivalent to solving

$$\min_{\mathbf{w}\in\mathbb{R}_+^k} \sum_{i=1}^m \big(\tilde{x}_t^i \log(\frac{\tilde{x}_t^i}{\mu_t^i}) - \tilde{x}_t^i + \mu_t^i\big) + \lambda\sum_{j=1}^k w^j + c, \tag{15}$$

where $\boldsymbol{\mu}_t = \mathbf{D}_{t-1}\mathbf{w} = [\mu_t^1, \cdots, \mu_t^m]^{\mathrm{T}} \in \mathbb{R}^m$ and $c = \sum_{i=1}^m \big(\log(\tilde{x}_t^i!) - \tilde{x}_t^i\log(\tilde{x}_t^i) + \tilde{x}_t^i\big)$ is a constant.

Algorithm 1 An algorithm of online non-negative dictionary learning via moment information for sparse Poisson coding

Require: $\mathbf{x}_t \in \mathbb{R}^m$, $\lambda \in \mathbb{R}_+$, $\mathbf{D}_0 \in \mathbb{R}_+^{m\times k}$, $\mathbf{w}_0 \in \mathbb{R}_+^k$, T

1: $\mathbf{D}_0 \leftarrow \mathbf{1}_{m\times k}/(mk)$, $\mathbf{w}_0 \leftarrow \mathbf{1}_k/k$

2: **for** $t = 1, \cdots, T$ **do**

3: Calculate the moment information as

$$\widetilde{\mathbf{x}}_t = \begin{cases}\mathbf{x}_t, \text{ iff } t = 1,\\ \big(\mathbf{x}_t + (t-1)\mathbf{D}_{t-1}\mathbf{w}_{t-1}\big)/t, \text{ iff } t > 1.\end{cases}$$

4: Update all the elements in $\mathbf{w}_t$ by using

$$w_t^j = \frac{w_{t-1}^j}{\sum_{p=1}^m d_{t-1,j}^p + \lambda}\sum_{i=1}^m \frac{\tilde{x}_t^i d_{t-1,j}^i}{\sum_{p=1}^k d_{t-1,p}^i w_{t-1}^p},$$

where $j = 1, 2, \cdots, k$.

5: Update all the elements in $\mathbf{D}_t$ by using

$$d_{t,j}^i = \frac{d_{t-1,j}^i \tilde{x}_t^i}{\sum_{p=1}^k d_{t-1,p}^i w_t^p},$$

where $i = 1, 2, \cdots, m$ and $j = 1, 2, \cdots, k$.

6: Update columns in $\mathbf{D}_t$ for satisfying $\mathbf{d}_j^{\mathrm{T}}\mathbf{d}_j \leq 1$ by using

$$\mathbf{d}_j = \frac{1}{\max(\|\mathbf{d}_j\|_2, 1)}\mathbf{d}_j,\ j = 1, 2, \cdots, k.$$

7: **if** $\mathcal{L}_t(\mathbf{D}_t)$ converges **then** Return $\mathbf{D}^* = \mathbf{D}_t$.

8: **end for**

9: $\mathbf{D}^* = \mathbf{D}_T$

10: **return** $\mathbf{D}^*$

Proof: By expanding the formulation of Eq. (12) and applying properties of logarithmic function, we have

$$\begin{aligned}&\min_{\mathbf{w}} \sum_{i=1}^m \big(\log(\tilde{x}_t^i!) + \mu_t^i - \tilde{x}_t^i\log(\mu_t^i)\big) + \lambda\|\mathbf{w}\|_1,\\ &\text{s.t. } \mathbf{w} \geq 0,\end{aligned} \tag{16}$$

where $\boldsymbol{\mu}_t = \mathbf{D}_{t-1}\mathbf{w} = [\mu_t^1, \cdots, \mu_t^m]^{\mathrm{T}} \in \mathbb{R}^m$.

In Eq. (16), we find a solution in $\mathbf{w} \in \mathbb{R}_+^k$, which guarantees the non-negativity of sparse weights. Besides, by adding constant terms of $\pm\mathbf{x}_t$ and $\pm\mathbf{x}_t.\log(\mathbf{x}_t)$ with the dot being the element-wise operation, we have Eq. (15). □

In order to solve Eq. (15), we can develop the online updating strategy as

$$w_t^j = \frac{w_{t-1}^j}{\sum_{p=1}^m d_{t-1,j}^p + \lambda}\sum_{i=1}^m \frac{\tilde{x}_t^i d_{t-1,j}^i}{\sum_{p=1}^k d_{t-1,p}^i w_{t-1}^p}, \tag{17}$$

where $d_{t-1,p}^i$ denotes the element of $\mathbf{D}_{t-1}$ at the position of the i-th row and the p-th column. w_{t-1}^p denotes the p-th element in $\mathbf{w}_{t-1}$. Once we initialize the weights $\mathbf{w}_0 \geq 0$ and the dictionary $\mathbf{D}_0 \geq 0$, we have the result that $\mathbf{w}_t \geq 0$ for $t = 1, 2, \cdots, T$. Note that more detailed theoretical analyses can be found in the next section.

B. Updating Dictionary

Similar to the transformation of Eq. (12) and Eq. (15), we have the following theorem for solving the online dictionary.

Theorem 2. The solution for optimization problem of Eq. (14) is equivalent to solving

$$\begin{aligned} \min_{\mathbf{D}\in\mathbb{R}_+^{m\times k}} & \sum_{i=1}^{m}\big(\tilde{x}_t^i\log(\frac{\tilde{x}_t^i}{\mu_t^i})-\tilde{x}_t^i+\mu_t^i\big)+c, \\ \text{s.t.} \quad & \mathbf{d}_j^{\mathrm{T}}\mathbf{d}_j\le 1,\ j=1,\cdots,k. \end{aligned} \tag{18}$$

where $\boldsymbol{\mu}_t=\mathbf{D}\mathbf{w}_t=[\mu_t^1,\cdots,\mu_t^m]^{\mathrm{T}}\in\mathbb{R}^m$, and c is a constant with the same definition in optimization problem of Eq. (15).

Proof: The proof is similar to the proof in Theorem 1, which is omitted here. □

Similarly, we can develop the following online updating strategy for dictionary learning

$$d_{t,j}^i=\frac{d_{t-1,j}^i\tilde{x}_t^i}{\sum_{p=1}^k d_{t-1,p}^i w_t^p}. \tag{19}$$

By initializing the weights $\mathbf{w}_0\ge 0$ and the dictionary $\mathbf{D}_0\ge 0$, we can readily have the result that $\mathbf{D}_t\ge 0$ for $t=1,2,\cdots,T$. We give more detailed theoretical analyses in the next section.

Moreover, for illustrations, the specific procedures of the proposed online algorithm are shown in Algorithm 1. In the algorithm, we just update the weights' vector and the dictionary once when a new sample appears, which is more efficient than the traditional algorithms in [9], [10]. In previous work without using the moment information, the method should firstly cumulate the sample information by using matrices. Then, the method updates alternatively the weights' vector $\mathbf{w}$ and the dictionary $\mathbf{D}$ until they both converge.

V. Convergence Analyses

In this section, we give detailed convergence analyses of the proposed algorithm shown in Algorithm 1.

Definition 1. Divergence between two non-negative matrices M and N can be defined as the following formula [30]

$$D(M||N)=\sum_{i,j}\big(M_{i,j}\log\frac{M_{i,j}}{N_{i,j}}-M_{i,j}+N_{i,j}\big). \tag{20}$$

From [30], we know that the divergence is an asymmetric distance measure of two matrices/vectors. More importantly, we have significant properties for the measure, which are $D(M||N)\ge 0$ for any M and N, and $D(M||N)=0$ if and only if $M=N$. Clearly, we know that the optimization problems of Eq. (15) and Eq. (18) have the forms of divergence minimization. This means that the optimization of Eq. (10) is a minimization of divergence based on the Poisson distribution.

Theorem 3. The divergence of $D(\widetilde{\mathbf{x}}_t||\mathbf{D}_t\mathbf{w}_t)$ in Eq. (10) is non-increasing under the updating rules of Eq. (17) and Eq. (19). More importantly, the divergence is invariant under these updating rules if and only if $\mathbf{D}_t$ and $\mathbf{w}_t$ are at a stationary point of the divergence.

Proof: First of all, we give the proof on the updating rule of Eq. (17). Define

$$G(\mathbf{w},\mathbf{w}_t)=g(\mathbf{w},\mathbf{w}_t)+\lambda\sum_j^k w^j, \tag{21}$$

where

$$\begin{aligned} g(\mathbf{w},\mathbf{w}_t)=&\sum_i(\tilde{x}_t^i\log\tilde{x}_t^i-\tilde{x}_t^i)+\sum_{i,j}d_{t,j}^i w^j- \\ &\sum_{i,j}\tilde{x}_t^i\frac{d_{i,j}^i w_t^j}{\sum_p d_{t,p}^i w_t^p}\big(\log d_{t,j}^i w^j-\log\frac{d_{t,j}^i w_t^j}{\sum_p d_{t,p}^i w_t^p}\big). \end{aligned} \tag{22}$$

We further define

$$F(\mathbf{w},\mathbf{w}_t)=f(\mathbf{w},\mathbf{w}_t)+\lambda\sum_j^k w^j, \tag{23}$$

where

$$f(\mathbf{w},\mathbf{w}_t)=\sum_i\tilde{x}_t^i\log\big(\frac{\tilde{x}_t^i}{\sum_p d_{t,p}^i w_t^p}-\tilde{x}_t^i+\sum_i d_{t,j}^i w_t^j\big). \tag{24}$$

Based on the theoretical results in [30], we have $g(\mathbf{w},\mathbf{w})=f(\mathbf{w})$ and $g(\mathbf{w},\mathbf{w}_t)\ge f(\mathbf{w})$. Then, we can easily obtain $G(\mathbf{w},\mathbf{w})=F(\mathbf{w})$ and $G(\mathbf{w},\mathbf{w}_t)\ge F(\mathbf{w})$. This result shows that we can apply the gradient of $G(\mathbf{w},\mathbf{w}_t)$ with respect to $\mathbf{w}$ for obtaining the local minimum of $F(\mathbf{w})$. By taking the derivative of $G(\mathbf{w},\mathbf{w}_t)$ with respect to $\mathbf{w}$ and setting it as zero, we have the updating rule of Eq. (17). One can refer to the similar procedures in [30] for more details. This result shows that, under rule of Eq. (17), the divergence of $D(\widetilde{\mathbf{x}}_t||\mathbf{D}_t\mathbf{w}_t)$ (i.e., the value of Eq. (15)) is non-increasing.

Similar to the proof of Eq. (17), we can obtain the updating rule for the dictionary matrix shown as Eq. (19). Therefore, under Eq. (17) and Eq. (19), the optimizations of Eq. (15) and Eq. (18) are alternatively minimized. By repeating the updating rules, we can obtain the local minimum of Eq. (10). Besides, based on the results in [30], the divergence is invariant under these updates if and only if $\mathbf{D}_t$ and $\mathbf{w}_t$ are at a stationary point of the divergence, which means $\widetilde{\mathbf{x}}_t=\mathbf{D}_t\mathbf{w}_t$. □

Theorem 4. By calculating the moment information of

$$\widetilde{\mathbf{x}}_t=\begin{cases}\mathbf{x}_t, \text{ iff } t=1,\\ \big(\mathbf{x}_t+(t-1)\mathbf{D}_{t-1}\mathbf{w}_{t-1}\big)/t, \text{ iff } t>1,\end{cases}$$

as well as the updating rules of Eq. (17) and Eq. (19), the dictionary has the convergence of

$$\lim_{t\to\infty}\mathbf{D}_t\mathbf{w}_t\to\boldsymbol{\mu}^*.$$

Proof: Based on the results in Theorem 3, we have the result that the divergence between $\mathbf{D}_t\mathbf{w}_t$ and $\widetilde{\mathbf{x}}_t$ is less than an error ball δ after an iteration, where δ is an appropriately small enough positive number, e.g., $\delta\le 10^{-1}$. Note that, with the sample information increasing, we know that the error of $\|\mathbf{D}_t\mathbf{w}_t-\widetilde{\mathbf{x}}_t\|_2$ decreases. This also means that the loss function in Eq. (10) is decreasing. Then, we can formulate $\|\mathbf{D}_t\mathbf{w}_t-\widetilde{\mathbf{x}}_t\|_2\le\delta$. Based on the update rule as

$$\widetilde{\mathbf{x}}_t=\begin{cases}\mathbf{x}_t, \text{ iff } t=1,\\ \big(\mathbf{x}_t+(t-1)\mathbf{D}_{t-1}\mathbf{w}_{t-1}\big)/t, \text{ iff } t>1,\end{cases}$$

we have the result as

$$\lim_{\delta\to 0,\ t\to\infty}\mathbf{D}_{t-1}\mathbf{w}_{t-1}\to(\mathbf{x}_1+\mathbf{x}_2+\cdots+\mathbf{x}_{t-1})/(t-1). \tag{25}$$

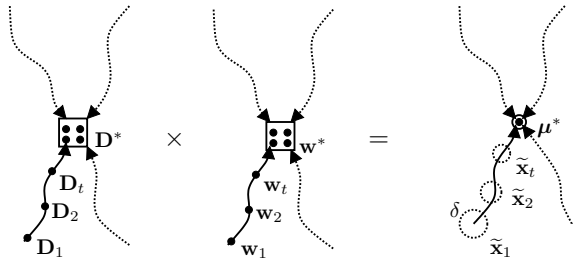

Fig. 1. The schematic diagram of how the algorithm of online dictionary learning via moment information works intuitively.

By updating $\mathbf{D}_t$ and $\mathbf{w}_t$ in Algorithm 1, we have

$$\lim_{\delta\to 0,\ t\to\infty} \mathbf{D}_t\mathbf{w}_t \to (\mathbf{x}_1+\mathbf{x}_2+\cdots+\mathbf{x}_t)/t.$$

Then, based on Eq. (5), we further have $\lim_{t\to\infty,\delta\to 0}\mathbf{D}_t\mathbf{w}_t = \boldsymbol{\mu}^*$. We know that, with the increasing of sample information, the error ball of δ is decreasing, which means that $\widetilde{\mathbf{x}}_t$ characterizes more and more moment information in light of the law of large numbers. Then, we can have the result of $\lim_{t\to\infty}\mathbf{D}_t\mathbf{w}_t$ converging to $\boldsymbol{\mu}^*$, which completes the proof. □

Remarks. The above theoretical result means that the learned dictionary will eventually converge to a stable one and characterize the moment information of the underlying Poisson model for real data. Besides, how does the algorithm of online dictionary learning via moment information work intuitively is shown in Fig. 1. In the figure, we set the stable dictionary and stable weights' vector, respectively, as $\mathbf{D}^*$ and $\mathbf{w}^*$. Note that, since the stable solutions are local minimums, we demonstrate several solid dots (e.g., four dots in the rectangle) to represent the solution set of $\mathbf{D}^*$. The demonstration of $\mathbf{w}^*$ has the similar meaning. When the dictionary converges, we have $\mathbf{D}^*\mathbf{w}^* = \boldsymbol{\mu}^*$. From the figure, we know that, with the sample information increasing, the error ball of δ decreases. When there are enough samples, the error ball of δ becomes zero, which has been guaranteed by the law of large numbers. Finally, the learned dictionary will capture the characteristics of the moment information of the sequential samples. For more illustrations, we also show different initializations of dictionary learning by using dash-line of arrows in Fig. 1.

Lemma 1. Convergence of Cesaro Means [31]. Suppose there is a sequence of $\{a_n\}$ with the Cesaro mean $c_n = \frac{1}{n}\sum_{i=1}^{n} a_i$, if we have $\lim_{n\to\infty} a_n = c$, then we further have $\lim_{n\to\infty} c_n = c$.

Proof: The proof can be found in [31]. □

Theorem 5. By using Algorithm 1, we have the result that the cumulative loss shown in Eq. (9) converges to a non-negative constant.

Proof: Based on Theorem 3 and by using Algorithm 1, we have the result that the loss for moment information shown in Eq. (10) converges to a non-negative constant because the divergence is always no less than zero and non-increasing when samples appear. In other words, we have

$$\lim_{t\to\infty} \ell(\widetilde{\mathbf{x}}_t, \mathbf{D}) = \text{const}, \tag{26}$$

where const $\in \mathbb{R}_+$.

In light of Lemma 1, we can readily get the cumulative loss to be

$$\lim_{t\to\infty} \mathcal{L}_t(\mathbf{D}) = \text{const}. \tag{27}$$

Then, we have the result that the cumulative loss shown in Eq. (9) converges to a non-negative constant. Note that the constant may not be zero, because Eq. (15) and Eq. (18) both have a non-negative constant part. In fact, if the constant parts in Eq. (15) and Eq. (18) are both zero, then $\mathcal{L}_t(\mathbf{D})$ will converge to zero. □

VI. Experimental Results

In order to verify the effectiveness of the proposed online algorithm, we conduct a series of experiments on a personal computer with Inter CUP@3.70GHz and 16GB memory. In the following experiments, we use document corpora [4] and image data [10]. We set the sparsity parameter $\lambda = 0.1$. Note that we have tried other values for the sparsity parameter and find that 0.1 is a relatively good one. The setting for the number of basis vectors k will be discussed in the experiments. Besides, the experiments are conducted via the datasets of TDT2 document corpora, Reuters-21578, MNIST and USPS. We firstly verify the convergence of cumulative loss function shown in Eq. (9), and show the learned stable dictionary $\mathbf{D}^*$ for the datasets. Then, we further show some merits of the proposed algorithm by comparing it with the online dictionary algorithm in [9] based on the sparse coding for classifications.

A. Convergence Verifications

The plots of the cumulative loss function in Eq. (9) for using documents in TDT2 and Reuters-21578 are shown in Fig. 2. We find that the cumulative loss converges, which is consistent with the theoretical analyses in Theorem 5. Specifically, we firstly set the number of basis vectors to be 100 in the dictionary. Since there are documents belonging to different categories in TDT2 and Reuters-21578, we use the documents in the same category to satisfy the i.i.d. assumption. In Fig. 2, we use the documents in the largest category. Then, we obtain the documents one by one for dictionary learning with sparse Poisson coding using Algorithm 1. We find that the cumulative loss function $\mathcal{L}_t(\mathbf{D}_t)$ converges to a stable constant after getting 500 documents in TDT2 of the largest category. Note that the constant is larger than zero, because both problems of Eq. (15) and Eq. (18) have a constant part. For the documents in Reuters-21578 of the largest category, we have the similar convergence results after getting 267 documents. For other categories of documents in TDT2 and Reuters-21578, we also have the similar convergence results, which are omitted.

For more illustrations, we further show the convergence of cumulative loss function for the dataset of MNIST in Fig. 3. For clarity, we then show the convergence of the cumulative loss function for ten digits (i.e., from 0 to 9) in Fig. 3. We find that the cumulative loss function of $\mathcal{L}_t(\mathbf{D}_t)$ of all digits converges. Besides, there are differences in the constants for different digits. These results mean that the

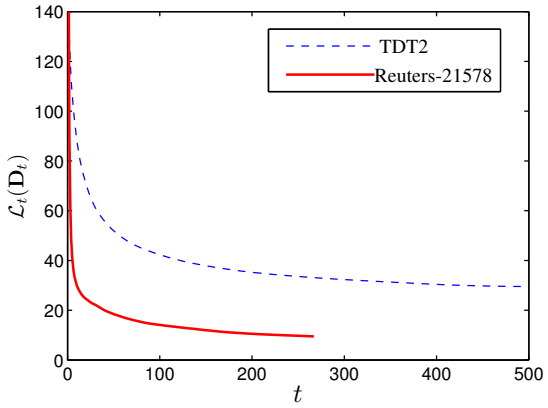

Fig. 2. Convergence of the cumulative loss function $\mathcal{L}_t(\mathbf{D}_t)$ for documents in TDT2 (after 500 samples) and Reuters-21578 (after 267 samples).

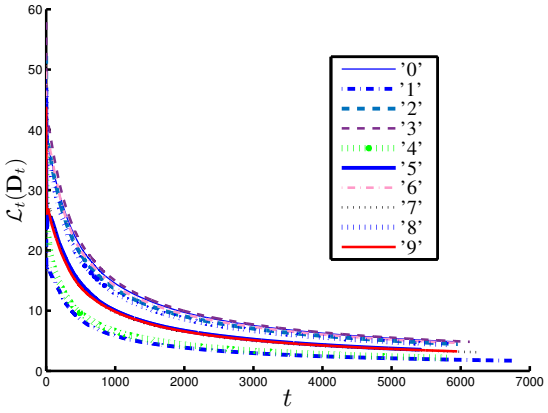

Fig. 3. Convergence of the cumulative loss function $\mathcal{L}_t(\mathbf{D}_t)$ for digits from 0 to 9 in the dataset of MNIST.

moment information varies for different digits. Note that we have the similar results for USPS, which are omitted.

Based on the above results, we know that the convergence of the cumulative loss function in Algorithm 1 is robust in word-counts data and image data.

B. Learned Dictionary

We show the stable learned dictionaries in Figs. 4 and 5 for document corpora and digit data. Specifically, in Fig. 4, we show a basis vector in the learned dictionary of the largest category for documents in Reuters-21578. Note that the basis vector is reshaped to be a matrix for illustrations. From the figure, we find that the learned dictionary is highly sparse, because the black region means zero values in the dictionary. The high sparsity can accelerate the convergence of the learning process for frequency data, which has been verified in Fig. 2. Besides, for the image data, we also show a basis vector in the dictionary for each digit after reshaping the vector. Since there are ten digits in MNIST, we have ten plots in Fig. 5. From the figure, we find that the learned dictionary has well characterized the information of each digit.

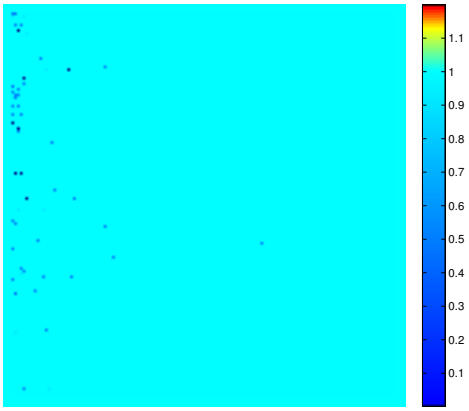

Fig. 4. A basis vector in the learned dictionary for the largest category of documents in Reuters-21578 after reshaping the vector to be a matrix.

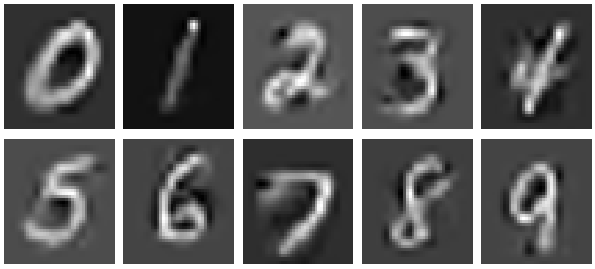

Fig. 5. Ten basis vectors for the corresponding dictionaries in the dataset of MNIST after reshaping the vectors to be matrices.

TABLE I
ACCURACY RATE OF CLASSIFICATIONS VIA SPARSE CODING

Parameter	Dataset	Our Method	Mairal et al. [9]
$k = 50$	TDT2	**82.83%**	80.01%
	Reuters-21578	**86.32%**	85.59%
	MNIST	57.52%	**63.78%**
	USPS	**59.05%**	57.54%
$k = 100$	TDT2	**82.65%**	81.03%
	Reuters-21578	**86.30%**	85.44%
	MNIST	57.10%	**64.71%**
	USPS	**60.23%**	58.28%

C. Classifications by Learned Dictionary

In order to illustrate the merits of the proposed algorithm, we perform the classification test based on datasets of TDT2 document corpora, Reuters-21578 [4], MNIST and USPS [10].

In the following experiments, we conduct classifications of new instances based on sparse Poisson coding with the learned dictionary. We compare the performance of our proposed algorithm with that of the online algorithm in [9]. We also investigate the effect of different number of basis vectors k in classifications via sparse coding.

The detailed results are shown in Table I. From the table, we find that the experimental results show good performance of our proposed algorithm in classifications for documents corpora, where Algorithm 1 beats the online algorithm in [9].

By contrast, for the image data, our algorithm is comparable for the dataset of USPS, but fails for the dataset of MNIST. Specifically, we use the training sets of these datasets for dictionary learning. Then, we use testing sets for sparse coding and then perform classifications based on reconstruction errors. By using the assumption of Poisson distribution for document corpora, a good dictionary is learned based on Algorithm 1. For the image data, the assumption of Poisson distribution may not be appropriate, which leads to the poor performance in MNIST. Besides, we also find that the number of basis vectors does not affect greatly for the performance of the algorithm.

VII. Conclusions

In this paper, we propose a novel algorithm of online non-negative dictionary learning via moment information for sparse Poisson coding. By assuming that real data may follow Poisson distribution, the algorithm incrementally learns a set of non-negative basis vectors and the sparse linear combinations of these vectors for non-negative data. Firstly, we formulate the online dictionary learning problem via moment information of samples in light of the maximum-a-posteriori framework, which has not been investigated. Secondly, we develop the multiplicative updating rules to learn the sparse-coefficient vectors and the basis vectors, which can maintain the non-negativity. Thirdly, we provide sufficient theoretical analyses on the convergence of the proposed online algorithm, which can guarantee its performance for online dictionary learning. Finally, we conduct a series of experimental evaluations on the applications of non-negative data (i.e., word-counts data and image data) to show the advantages of the proposed algorithm.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper. The work described in this paper was partially supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (No. CUHK413213 and No. CUHK14205214 of the General Research Fund) and 2015 Microsoft Research Asia Collaborative Research Program (Project No. FY16-RES-THEME-005).

References

[1] S. Kong and D. Wang, "A dictionary learning approach for classification: separating the particularity and the commonality," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 186–199.

[2] N. Wang, J. Wang, and D.-Y. Yeung, "Online robust non-negative dictionary learning for visual tracking," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 657–664.

[3] V. M. Patel and R. Chellappa, "Sparse representations, compressive sensing and dictionaries for pattern recognition," in *Pattern Recognition (ACPR), 2011 First Asian Conference on*. IEEE, 2011, pp. 325–329.

[4] C. Wu, H. Yang, J. Zhu, J. Zhang, I. King, and M. R. Lyu, "Sparse poisson coding for high dimensional document clustering," in *Big Data, 2013 IEEE International Conference on*. IEEE, 2013, pp. 512–517.

[5] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. R. Bach, "Supervised dictionary learning," in *Advances in neural information processing systems*, 2009, pp. 1033–1040.

[6] S. Mallat, *A wavelet tour of signal processing*. Academic press, 1999.

[7] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3501–3508.

[8] O. Dikmen and C. Févotte, "Maximum marginal likelihood estimation for nonnegative dictionary learning in the gamma-poisson model," *Signal Processing, IEEE Transactions on*, vol. 60, no. 10, pp. 5163–5175, 2012.

[9] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 689–696.

[10] C. Lu, J. Shi, and J. Jia, "Online robust dictionary learning," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 415–422.

[11] S. Schbath, "Compound poisson approximation of word counts in dna sequences," *ESAIM: probability and statistics*, vol. 1, pp. 1–16, 1997.

[12] A. Lefevre, F. Bach, and C. Févotte, "Online algorithms for nonnegative matrix factorization with the itakura-saito divergence," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*. IEEE, 2011, pp. 313–316.

[13] F. Liu and L. Chan, "Causal discovery on discrete data with extensions to mixture model," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 7, no. 2, pp. 21:1–19, 2015.

[14] H. Lee, R. Raina, A. Teichman, and A. Y. Ng, "Exponential family sparse coding with application to self-taught learning." in *IJCAI*, vol. 9. Citeseer, 2009, pp. 1113–1119.

[15] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 4, pp. 791–804, 2012.

[16] K. Engan, S. O. Aase, and J. Hakon Husoy, "Method of optimal directions for frame design," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 5. IEEE, 1999, pp. 2443–2446.

[17] K. Skretting, J. H. Husøy, and S. O. Aase, "A simple design of sparse signal representations using overlapping frames," in *Image and Signal Processing and Analysis, 2001. ISPA 2001. Proceedings of the 2nd International Symposium on*. IEEE, 2001, pp. 424–428.

[18] K. Engan, K. Skretting, and J. H. Husøy, "Family of iterative ls-based dictionary learning algorithms, ils-dla, for sparse signal representation," *Digital Signal Processing*, vol. 17, no. 1, pp. 32–49, 2007.

[19] R. Rubinstein, T. Peleg, and M. Elad, "Analysis k-svd: A dictionary-learning algorithm for the analysis sparse model," *Signal Processing, IEEE Transactions on*, vol. 61, no. 3, pp. 661–677, 2013.

[20] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4311–4322, 2006.

[21] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[22] A. Globerson, S. Shalev-Shwartz, and A. Birnbaum, "Learning the experts for online sequence prediction," 2012.

[23] A. Gonen, S. Sabato, and S. Shalev-Shwartz, "Efficient active learning of halfspaces: an aggressive approach," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 2583–2615, 2013.

[24] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," 2003.

[25] E. Hazan, A. Rakhlin, and P. L. Bartlett, "Adaptive online gradient descent," in *Advances in Neural Information Processing Systems*, 2007, pp. 65–72.

[26] C. Gentile, "A new approximate maximal margin classification algorithm," *The Journal of Machine Learning Research*, vol. 2, pp. 213–242, 2002.

[27] Y. Li and P. M. Long, "The relaxed online maximum margin algorithm," *Machine Learning*, vol. 46, no. 1-3, pp. 361–387, 2002.

[28] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," *The Journal of Machine Learning Research*, vol. 7, pp. 551–585, 2006.

[29] D. W. Osgood, "Poisson-based regression analysis of aggregate crime rates," *Journal of quantitative criminology*, vol. 16, no. 1, pp. 21–43, 2000.

[30] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[31] G. H. Hardy, *Divergent series*. American Mathematical Soc., 2000, vol. 334.