Appl. Comput. Harmon. Anal. • • • (• • • •) • • •–• • •

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Convergence radius and sample complexity of ITKM algorithms for dictionary learning

Karin Schnass

Department of Mathematics, University of Innsbruck, Technikerstraße 13, 6020 Innsbruck, Austria

ARTICLE INFO

Article history:
Received 8 June 2015
Received in revised form 5 August 2016
Accepted 15 August 2016
Available online xxxx
Communicated by Guy Nason

Keywords:
Dictionary learning
Sparse coding
Sparse component analysis
Sample complexity
Convergence radius
Alternating optimisation
Thresholding
K-means

ABSTRACT

In this work we show that iterative thresholding and K means (ITKM) algorithms can recover a generating dictionary with K atoms from noisy S sparse signals up to an error $\tilde{\varepsilon}$ as long as the initialisation is within a convergence radius, that is up to a $\log K$ factor inversely proportional to the dynamic range of the signals, and the sample size is proportional to $K \log K \tilde{\varepsilon}^{-2}$. The results are valid for arbitrary target errors if the sparsity level is of the order of the square root of the signal dimension d and for target errors down to $K^{-\ell}$ if S scales as $S \leq d/(\ell \log K)$.

1. Introduction

The goal of dictionary learning is to find a dictionary that will sparsely represent a class of signals. That is given a set of N training signals $y_n \in \mathbb{R}^d$, which are stored as columns in a matrix $Y = (y_1, \dots, y_N)$, one wants to find a collection of K normalised vectors $\phi_k \in \mathbb{R}^d$, called atoms, which are stored as columns in the dictionary matrix $\Phi = (\phi_1, \dots, \phi_K) \in \mathbb{R}^{d \times K}$, and coefficients x_n, which are stored as columns in the coefficient matrix $X = (x_1, \dots, x_N)$ such that

$$Y = \Phi X \quad \text{and} \quad X \text{ sparse.} \tag{1}$$

Research into dictionary learning comes in two flavours corresponding to the two origins of the problem, the slightly older one in the independent component analysis (ICA) and blind source separation (BSS) community, where dictionary learning is also known as sparse component analysis, and the slightly younger

E-mail address: karin.schnass@uibk.ac.at.

http://dx.doi.org/10.1016/j.acha.2016.08.002
1063-5203/© 2016 Elsevier Inc. All rights reserved.

one in the signal processing community, where it is also known as sparse coding. The main motivation for dictionary learning in the ICA/BSS community comes from the assumption that the signals of interest are generated as sparse mixtures – random sparse mixing coefficients X_0 – of several sources or independent components – the dictionary Φ_0 – which can be used to describe or explain a (natural) phenomenon, [15,30,27,26]. For instance in the 1996 paper by Olshausen and Field, [15], which is widely regarded as the mother contribution to dictionary learning, the dictionary is learned on patches of natural images, and the resulting atoms bear a striking similarity to simple cell receptive fields in the visual cortex. A natural question in this context is, when the generating dictionary Φ_0 can be identified from Y, that is, the sources from the mixtures. Therefore the first theoretical insights into dictionary learning came from this community, [18]. Also the first dictionary recovery algorithms with global success guarantees, which are based on finding overlapping clusters in a graph derived from the signal correlation matrix $Y^\star Y$, take the ICA/BSS point of view, [6,2].

The main motivation for dictionary learning in the signal processing community is that sparse signals are immensely practical, as they can be easily stored, denoised, or reconstructed from incomplete information, [13,33,31]. Thus the interest is less in the dictionary itself but in the fact that it will provide sparse representations X. Following the rule 'the sparser – the better' the obvious next step is to look for the dictionary that provides the sparsest representations. So given a budget of K atoms and S non-zero coefficients per signal, one way to concretise the abstract formulation of the dictionary learning problem in (1) is to formulate it as optimisation problem, such as

$$(P_{2,S}) \qquad \min \|Y - \Phi X\|_F \quad \text{s.t.} \quad \|x_n\|_0 \leq S \quad \text{and} \quad \Phi \in \mathcal{D}, \tag{2}$$

where $\|\cdot\|_0$ counts the nonzero elements of a vector or matrix and $\mathcal{D}$ is defined as $\mathcal{D} = \{\Phi = (\phi_1, \ldots, \phi_K) : \|\phi_k\|_2 = 1\}$. While $(P_{2,S})$ is for instance the starting point for the MOD or K-SVD algorithms, [14,3], other definitions of *optimally* sparse lead to other optimisation problems and algorithms, [49,37,48,32,43,38]. The main challenge of optimisation programmes for dictionary learning is finding the global optimum, which is hard because the constraint manifold $\mathcal{D}$ is not convex and the objective function is invariant under sign changes and permutations of the dictionary atoms with corresponding sign changes and permutations of the coefficient rows. In other words for every local optimum there are $2^K K! - 1$ equivalent local optima.

So while in the signal processing setting there is a priori no concept of a generating dictionary, it is often used as auxiliary assumption to get theoretical insights into the optimisation problem. Indeed without the assumption that the signals are sparse in some dictionary the optimisation formulation makes little or no sense. For instance if the signals are uniformly distributed on the sphere in $\mathbb{R}^d$, in asymptotics $(P_{2,S})$ becomes a covering problem and the set of optima is invariant under orthonormal transforms.

Based on a generating model on the other hand it is possible to gain several theoretical insights. For instance, how many training signals are necessary such that the sparse representation properties of a dictionary on the training samples (e.g. the optimiser) will extrapolate to the whole class, [34,47,35,20]. What are the properties of a generating dictionary and the maximum sparsity level of the coefficients and signal noise such that this dictionary is a local optimiser or near a local optimiser given enough training signals, [21,17,39,40,19].

An open problem for overcomplete dictionaries with some first results for bases, [44,45], is whether there are any spurious optimisers which are not equivalent to the generating dictionary, or if any starting point of a descent algorithm will lead to a global optimum? A related question (in case there are spurious optima) is, if the generating dictionary is the global optimiser? If yes, it would justify using one of the graph clustering algorithms for recovering the optimum, [6,2,4,7]. This is important since all dictionary learning algorithms with global success guarantees are computationally very costly, while optimisation approaches are locally very efficient and robust to noise. Knowledge of the convergence properties of a descent algorithm, such as